# Stochastic Models for Service Operations Management

Oualid Jouini

HDR 2014 015 Année 2014

# HABILITATION A DIRIGER DES RECHERCHES

présentée devant

l'Institut National des Sciences Appliquées de Lyon
et l'Université Claude Bernard LYON I

**Stochastic Models for Service Operations Management**

SECTION CNU 61 - SPECIALITE PRODUCTIQUE

par
JOUINI Oualid

Soutenue le 5 décembre 2014 devant la Commission d'examen

---

BOCQUET Jean Claude, Professeur, Ecole Centrale Paris, **Examinateur**
CACHARD Christian, Professeur, Université Lyon 1, **Examinateur**
CAMPAGNE Jean Pierre, Professeur, INSA Lyon, **Directeur de recherche**
CHEVALIER Philippe, Professeur, Université Catholique de Louvain, **Rapporteur**
DI MASCOLO Maria, Directrice de Recherche CNRS, Grenoble INP, **Examinateur**
FREIN Yannick, Professeur, Grenoble INP, **Rapporteur**
GOURGAND Michel, Professeur, LIMOS-ISIMA, **Examinateur**
HENNET Jean Claude, Directeur de Recherche CNRS, LSIS Marseille, **Président**
KOOLE Ger, Professeur, VU University Amsterdam, **Examinateur**
XIE Xiaolan, Professeur, Ecole des Mines de Saint-Etienne, **Rapporteur**

Laboratoire Décision et Information pour les Systèmes de Production (DISP)

*A Zouha et Aziz,*

*A ma famille,*

# Contents

**Résumé en Français** **vii**

1 Evolution de Mes Travaux . . . . . viii

2 Motivation et Positionnement de Mes Travaux . . . . . x

3 Synthèse de Mes Résultats . . . . . x

3.1 Centres d'Appels . . . . . x

3.2 Processus Stochastiques et Applications . . . . . xi

4 Travaux en Cours et Perspectives . . . . . xii

4.1 Centres d'Appels . . . . . xii

4.2 Services d'Urgence . . . . . xiii

4.3 Processus Stochastiques et Applications . . . . . xiii

**Presentation of the Dissertation** **2**

**Part I Presentation of the Candidate** **4**

1 Personal Details . . . . . 5

2 Education . . . . . 5

3 Professional Experiences . . . . . 6

4 Research Activities . . . . . 6

4.1 Research Supervision . . . . . 6

4.2 My Publications . . . . . 10

4.3 Evidence of Research Esteem . . . . . 15

5 Teaching Activities . . . . . 19

6 Administrative and Editorial Activities . . . . . 20

6.1 Administrative Responsibilities . . . . . 20

6.2 Editorial Activities . . . . . 21

**Overview on My Research Activities** **22**

**Part II Operations Management in Call Centers** **24**

**Chapter II.1 Modeling of Call Centers** . . . . . 26

1 Introduction . . . . . 26
2 Call center Background . . . . . 26
3 Performance Indicators with Impatience . . . . . 28
3.1 Statistical Analysis and Modeling of Abandonments . . . . . 28
3.2 Analysis of Call Center Metrics . . . . . 29
3.3 Concluding Remarks and Future Research . . . . . 31

**Chapter II.2 Design of Call Centers** . . . . . 32

1 Introduction . . . . . 32
2 Team-Based Organization . . . . . 34
2.1 Context and Motivation . . . . . 34
2.2 Positioning of My Contributions . . . . . 34
2.3 Problem Setting . . . . . 35
2.4 Analysis of the Efficiency of the Team-Based Organization . . . . . 37
3 Flexible Architecture . . . . . 39
3.1 Background and Research Objectives . . . . . 39
3.2 My Main Findings . . . . . 40
3.3 Modeling . . . . . 40
3.4 Approximate Numerical Comparison . . . . . 41
3.5 Synthesis on the Effect of Parameter Asymmetry . . . . . 42
4 Concluding Remarks and Future Research . . . . . 43

**Chapter II.3 Personnel Planning in Call Centers** . . . . . 45

1 Context and Contributions . . . . . 45
1.1 Positioning of My Contributions . . . . . 46
1.2 Problem Formulation . . . . . 47
1.3 Solution Methodologies . . . . . 48
1.4 Insights . . . . . 50
2 Concluding Remarks and Future Research . . . . . 50

**Chapter II.4 Operational Issues: Call Centers with Delay Information** . . 52

1 Introduction . . . . . 52
2 Single Class Setting: Endogenized Customer Reaction . . . . . 53
2.1 Context and Motivation . . . . . 53
2.2 Positioning of My Contributions . . . . . 54
2.3 Modeling . . . . . 55
2.4 Experiments . . . . . 57

3 Multi-Class Setting: A Newsvendor-Like Approach . . . . . 57
3.1 Introduction and Related Literature . . . . . 57
3.2 Choosing What to Announce . . . . . 59
3.3 Predicting Delays . . . . . 59
3.4 Announcing a Delay from the Estimated Delay Distribution . . . . . 60
3.5 Data-Based Validation of Delay Announcements . . . . . 61
4 Concluding Remarks and Future Research . . . . . 61
**Chapter II.5 Operational Issues: Optimal Routing in Call Centers** . . . . . 63
1 Introduction . . . . . 63
2 Online Policies for Impatient Customers . . . . . 64
2.1 Context and Motivation . . . . . 64
2.2 Related Literature . . . . . 65
2.3 Framework . . . . . 66
2.4 Online Scheduling Policies . . . . . 67
2.5 Experiments and Synthesis . . . . . 69
3 Threshold Policies for Calls and Emails . . . . . 69
3.1 Introduction and Related Literature . . . . . 69
3.2 Problem Formulation . . . . . 70
3.3 Constant Arrival Rate . . . . . 71
3.4 Our Adaptive Threshold Policy (ATP) . . . . . 72
3.5 Non-Constant Arrival Rates . . . . . 73
4 Job Routing with Idling Times during the Call Service . . . . . 75
4.1 Introduction . . . . . 75
4.2 My Contributions . . . . . 76
4.3 Positioning of My Contributions . . . . . 76
4.4 Problem Description and Modeling . . . . . 77
4.5 Single Server Analysis . . . . . 79
4.6 Multi-Server Case . . . . . 82
5 Concluding Remarks and Future research . . . . . 83

**Part III Analysis of Stochastic Processes** **85**
**Chapter III.1 Analysis of Markov Chains** . . . . . 87
1 Introduction . . . . . 87
2 Computation of First Passage Times . . . . . 88
2.1 Positioning and Contributions . . . . . 88
2.2 Model Description and Notations . . . . . 88
2.3 Applications . . . . . 91
3 Sums of Erlang Random Variables . . . . . 91

3.1 Introduction . . . 91
3.2 Computation . . . 92
4 Concluding Remarks and Future Research . . . 93
**Chapter III.2 Analysis of Queueing Systems with Impatience** . . . 95
1 Context and Contributions . . . 95
2 Related Literature . . . 96
3 General Abandonments . . . 97
3.1 Introduction . . . 97
3.2 The Result . . . 97
4 Multiple Priority . . . 99
4.1 Context and Contributions . . . 99
4.2 Modeling . . . 99
4.3 Results . . . 100
5 Monotonicity Properties . . . 102
5.1 Introduction . . . 102
5.2 Model Formulation . . . 102
5.3 Monotonicity Results . . . 103
6 Concluding Remarks and Future Research . . . 104
**Chapter III.3 Dynamic Control of Queueing Systems** . . . 106
1 Context and Contributions . . . 106
2 Optimal Routing for the Slow-Server Queue . . . 107
2.1 Introduction and Positioning of the Contributions . . . 107
2.2 Problem Formulation . . . 108
2.3 Optimal Routing . . . 109
2.4 Performance Measures . . . 110
3 Uniformization for Queues with Abandonment . . . 111
3.1 Introduction . . . 111
3.2 Erlang Approximation with Abandonment . . . 112
4 Concluding Remarks and Future Research . . . 113

**Part IV Ongoing Work and Research Perspectives** **114**
1 Call Center Operations . . . 115
1.1 Shift-Scheduling with Uncertain Arrival Rates . . . 115
1.2 Advertisement While Waiting . . . 116
1.3 Multi-Channel Issues . . . 117
2 Emergency Department Operations . . . 118
2.1 Performance indicators . . . 119
2.2 Human Resource Related Issues . . . 120

2.3 Process Related Issues . . . . . 121
3 Stochastic Models and Their Applications to Services . . . . . 121
3.1 Appointment-Driven Arrivals . . . . . 122
3.2 Collaboration in Service Systems . . . . . 123

**Bibliography** **125**

# Résumé en Français

# Modèles Stochastiques pour la Gestion des Opérations de Service

Ce mémoire présente l'avancement de mes travaux de recherche. Dans ce qui suit, je commence par décrire l'évolution de mes travaux dans le temps tout en précisant leurs contextes ainsi que les liens éventuels entre eux. Je résume brièvement ensuite mes résulats de recherche à ce jour. Finalement, je présente mes projets en cours ainsi que mes perspectives de recherche.

## 1 Evolution de Mes Travaux

Mes activités de recherche ont commencé en 2002 avec mon master recherche en génie industriel au Laboratoire Génie Industriel (LGI) de l'Ecole Centrale Paris (ECP). J'ai poursuivi ensuite avec ma thèse, que j'ai soutenue en 2006. Les résultats de recherche de mon master et ma thèse, qui se sont effectués sous la direction de Yves Dallery, portaient sur l'analyse et l'optimisation du centre d'appels de Bouygues Telecom. J'ai travaillé sur des problèmes stratégiques et opérationnels proposés par nos partenaires: Fabrice Chauvet et Rabie Nait-Abdallah (recherche et développement), et Olivier Belma et Thierry Prat (système d'information et de télécommunication). Pendant ma thèse, j'ai aussi travaillé sur l'analyse des files d'attente en général.

Après ma thèse, j'ai été recruté en 2007 en tant qu'assistant au laboratoire LGI à ECP. Au LGI, j'appartiens à l'équipe 2 *Aide à la Décision pour les Systèmes de Production et de Service* et je me positionne au niveau du premier project de recherche *Gestion des Opérations de Service*. J'ai continué après ma thèse à travailler sur les centres d'appels en élargissant les sujets traités. Ceci a pu avoir lieu grâce à mes partenaires industriels (Bluelink, Interact-iv.com, Digiway Consulting, etc.), et mes collègues Ger Koole (VU University Amsterdam) et Zeynep Aksin (Koç University) que j'ai fréquemment et régulièrement rencontrés entre 2007 et 2009. Durant cette période, Ger Koole a aussi passé un séjour sabbatique de plusieurs mois au LGI.

En partant d'une problématique de gestion des opérations, mon approche de recherche consiste à construire premièrement un modèle - souvent stochastique - puis à développer une analyse quantitative dans le but de trouver des éléments de réponse à la question posée au départ. J'ai été ainsi fréquemment confronté à des questions challengeantes liées au processus stochastiques. Ceci m'a motivé à dépasser les modèles spécifiques de centres d'appels et essayer d'aller plus loin dans des analyses théoriques qui soient utiles pour des contextes plus larges de systèmes de service.

En 2007, j'ai obtenu avec Zeynep Aksin un financement de TÜBİTAK (l'agence scientifique et technologique de Turquie). Ceci nous a permis d'avancer considérablement sur nos projets d'analyse de centres d'appels avec annonce de temps d'attente. En même temps, j'ai collaboré étroitement avec Yves Dallery, Ger Koole et Auke Pot de VU University Amsterdam. Nous avons principalement travaillé sur des problématiques de routage d'appels. Avec Ger Koole, on était

en contact avec Bluelink (fournisseur de service de Air France KLM) et Digiway Consulting (société de conseil pour les centres d'appels). Cette dernière nous a fourni des données et nous a motivé Ger et moi à travailler sur les problèmes de planification des agents avec des paramètres d'arrivées incertains. On a ainsi profité à l'époque d'un financement CSC (bourse du gouvernement Chinois) pour lancer la thèse de Shuangqing Liao sur ce sujet en 2008. La thèse a été co-encadrée par Christian van Delft de HEC Paris, étant donné son expertise sur la programmation stochastique. Shuangqing a soutenu sa thèse en 2011.

En 2008, j'ai fait un postdoc à University of Minnesota sous la direction de Saif Benjaafar. J'ai travaillé sur l'analyse de modèles de files d'attente avec un nombre fini d'arrivées hétérogènes. L'application de ces modèles est liée aux systèmes de service générés par des évènements. J'ai continué depuis à travailler sur ce sujet avec Saif. En 2010, son thésard Rowan Wang nous a rejoint pour travailler sur plusieurs extensions. Au cours de la même année, ils m'ont tous les deux rendu visite pendant deux mois à ECP.

Après mon postdoc, je suis revenu à ECP toujours sur mon poste d'assistant, puis j'ai été recruté en 2010 sur un poste de Chef Travaux (contractuel, équivalent Maître de Conférences). En 2010, j'ai obtenu un financement d'Interact-iv.com, une société qui fournit des solutions logicielles et matérielles aux centres d'appels. Le financement m'a servis à recruter le thésard Benjamin Legros pour travailler sur des problématiques de routage dans les centres d'appels multi-compétences et multi-canaux. Ceci a permis d'intensifier considérablement mes résultats de recherche sur les problématiques opérationnelles, et en même temps d'accentuer mes contributions théoriques sur les files d'attente. Benjamin a soutenu sa thèse en 2013. En 2010 et 2011, j'ai aussi travaillé avec Ger Koole et son thésard Alex Roubos sur la modélisation des abandons. Avec Alex, j'ai aussi travaillé sur l'analyse de performance de files d'attente multi-classe.

Plus tard, en 2012, j'ai obtenu avec Céline Gicquel et Abdel Lisser de l'Université Paris Sud un financement doctorale de Digiteo, sur lequel on a recruté Mathilde Excoffier pour continuer à travailler sur la problématique assez riche de planification, en utilisant et comparant plusieurs approches d'optimisation stochastique. En 2012, j'ai également eu la chance de gagner un financement ANR Jeunes Chercheurs qui m'a permis de recruter en 2013 les 2 postdocs Benjamin Legros and Mahdi Fathi. Avec Benjamin et Mahdi, j'ai continué à travailler sur des problématiques opérationnelles de centres d'appels.

En résumé, après ma thèse, j'ai étendu mes travaux et contribué à la littérature des centres d'appels sur plusieurs niveaux. En parallèle, j'ai contribué à la littérature sur les processus stochastiques avec des applications aux systèmes de service qui peuvent être modélisés par des files d'attente. Ceci m'a permis d'avancer pas à pas dans la direction de mon objectif de recherche qui est de contribuer à la littérature sur les modèles stochastiques et la gestion des opérations de service.

Récemment en 2013, j'ai commencé à travailler sur les services d'urgence dans le cadre de la thèse de Karim Ghanes, pour laquelle on a obtenu un financement de l'Agence Régionale Santé Ile-de-France. Nous sommes en collaboration avec quelques services d'urgence de la région Parisienne sur des problématiques d'optimisation des processus et des ressources. Un service

d'urgence est un système de service complexe ayant d'importants facteurs humains et impacts sociétaux. Cela constitue pour moi une excellente opportunité pour continuer à avancer dans mes contributions aux opérations de service.

## 2 Motivation et Positionnement de Mes Travaux

En premier lieu, mes travaux s'intègrent dans la discipline de gestion des opérations. Ils s'intègrent également dans la discipline de recherche opérationnelle. Sur le plan national, mes recherches sont principalement rattachées à la section CNU 61 (génie industriel), mais aussi 26 (mathématiques appliquées). Les communautés de référence à l'échelle nationale sont GDR MACS et ROADEF. Celles à l'échelle internationale sont INFORMS, POMS et IIE.

Mes intérêts de recherche aux systèmes de service sont liés à la croissance importante observée dans ce secteur et le besoin d'améliorer leur organisation en termes de délai d'accès et de coûts d'exploitation engendrés. En France, comme dans les pays les plus développés ayant accédé à l'économie post-industrielle, les services représentent jusqu'à 70% des richesses produites et sont devenus le principal moteur de croissance économique. Depuis plusieurs décennies, les systèmes manufacturiers de biens ont fait l'objet de beaucoup de travaux de recherche au détriment des systèmes de service.

En plus des enjeux économiques, les systèmes de service sont caractérisés par les aspects humains liés aux comportements, et à la satisfaction des usagers et des personnels impliqués. Ces aspects sont au coeur des préoccupations de mes travaux. L'objectif ultime de mes travaux est de développer des recommandations et insights utiles aux managers, i.e., qui leur permettent de rendre leurs systèmes plus réactifs.

## 3 Synthèse de Mes Résultats

Mes principaux résultats de recherche sont classés en deux parties. Une première concerne les centres d'appels, et une suivante qui englobe mes contributions théoriques aux méthodes d'analyse de chaînes de Markov et de files d'attente avec applications aux systèmes de service. Les deux familles de résultats sont détaillées respectivement sur les Parties II et III de ce manuscrit. Dans ce qui suit, je les présente brièvement.

### 3.1 Centres d'Appels

Les centres d'appels, ou en général les centres de contacts, sont des systèmes de service considérés de nos jours comme l'outil de relation clientèle privilégié. Ils remplissent de plus en plus de fonctions; ils interviennent sur toute la chaîne de service clients, depuis l'avant-vente, jusqu'à l'après vente, en passant par l'assistance et la fidélisation des clients, mais aussi sur la qualification de prospect, la télé-vente et l'information. Aujourd'hui, les centres d'appels intéressent tous les secteurs d'activité; les opérateurs téléphoniques, les compagnies d'assurances et les banques à distance, mais également les services publics, les hôpitaux ou les collectivités locales. Grâce à

l'évolution des technologies de couplage de l'informatique et de la téléphonie, nous avons connu un réel décollage de cette activité.

Compte tenu de leurs enjeux économiques importants et de leurs complexités, tout un champ de recherche s'est développé depuis plusieurs années. Mes travaux se situent dans ce contexte. Ils couvrent la modélisation des centres d'appels ainsi que les 3 horizons de gestion des opérations : long terme, moyen terme, et court terme.

- **Modélisation de Centre d'Appels:** En se basant sur des données réelles, j'ai travaillé sur les temps de patience des clients, en proposant deux nouvelles modélisations. J'ai également proposé plusieurs définitions de qualité de services liés aux abandons, calculé leurs expressions, et discuté les avantages et les inconvénients de chacune. J'ai également démontré l'intérêt pour les managers de choisir la bonne métrique à appliquer.

- **Problématiques Long-Terme:** Dans un contexte mono-compétence, j'ai travaillé sur le développement de nouvelles architectures de centres d'appels qui permettent une meilleure gestion des ressources humaines. Dans un contexte multi-compétences, je me suis intéressé au développement d'architectures qui offrent de la flexibilité à travers des répartitions intelligentes des compétences entre les équipes de téléconseillers.

- **Problématiques Moyen-Terme:** J'ai travaillé sur des modèles d'optimisation des emplois de temps des agents tout en considérant un processus d'arrivée des appels doublement stochastique. On permet en effet la possibilité d'avoir des paramètres incertains pour l'arrivée afin de tenir compte des erreurs de prévision. J'ai travaillé sur des approches de résolution issues de la programmation stochastique et la programmation robuste.

- **Problématiques Court-Terme:** Quant aux problématiques opérationnelles sur le court terme, j'ai travaillé sur plusieurs questions de gestion temps-réel telles que l'estimation et l'annonce du temps d'attente tout en intégrant les impacts du phénomène d'abandon, et le routage dynamique des appels entrants vers les agents. J'ai également étudié des problématiques d'optimisation de l'ordonnancement des tâches dans le nouveau contexte de centre d'appels multi-canaux (appels entrants, appels sortants et back-office).

## 3.2 Processus Stochastiques et Applications

Les analyses quantitatives menées dans le cadre des travaux cités ci-haut sont surtout basées sur les chaînes de Markov et les files d'attente. Comme déjà mentionné, cela m'a motivé à aller chercher de nouvelles contributions théoriques génériques qui dépassent le cadre des centres d'appels, et concernant un large spectre d'applications aux systèmes de service. Mes résultats peuvent être structurés comme suit:

- **Analyse de Chaînes de Markov:** J'ai travaillé sur le calcul des moments de plusieurs types de variables aléatoire de temps de premier passage, ordinaires et conditionnels, dans

un processus de naissance et de mort de forme générale. J'ai également travaillé sur le calcul de la distribution d'une somme arbitraire de variables aléatoires Erlang.

- **Files d'Attente avec Impatience:** Pour une file d'attente mono-classe avec des abandons généralement distribués, j'ai travaillé sur des approximations contrôlées des performances. La méthode consiste à approximer la fonction du hasard avec une fonction par palier. J'ai également développé une méthode exacte basée sur les transformées de Laplace pour l'analyse d'une file d'attente multi-classe avec priorité non-préemptive et dont la discipline de service par classe peut être premier arrivé, premier servi ou dernier arrivé, premier servi. Je me suis aussi intéressé à prouver des résulats de monotinicité de premier et deuxième ordres dans une file d'attente avec clients impatients et une capacité du système finie.

- **Contrôle Dynamique:** Je me suis intéressé à une file d'attente connue sous le nom de *modèle du serveur lent*. J'ai développé un nouveau résultat de contrôle optimal sur ce modèle, en ajoutant la possibilité qu'un serveur peut avoir un échec de service. J'ai aussi travaillé sur l'uniformisation des chaînes de Markov à sauts non bornés. En utilisant une approche de modélisation non classique, on aboutit à une représentation du processus qui est exacte et naturellement bornée.

## 4 Travaux en Cours et Perspectives

Tout d'abord, je souhaite continuer à travailler sur les nouvelles problématiques riches de centres d'appels, intensifier mes travaux sur la gestion des opérations des services d'urgence, et poursuivre mes travaux sur les processus stochastiques et leurs applications aux services.

Je voudrais contribuer à ces branches de litérature tout en essayant d'incorporer le plus possible les facteurs clés de comportements humains et d'avancées technologiques. Je voudrais également accentuer l'utilisation de données réelles dans mes modèles. Heureusement, les récents développements dans les systèmes d'information ont rendu disponibles de larges quantités de données. C'est une opportunité, qui permettra sans doute d'enrichir les modèles existants et aboutira à des recommendations plus précises et donc des études plus impactantes pour la pratique.

Dans ce qui suit et en suivant la structure (1) centres d'appels, (2) services d'urgence, et (3) processus stochastiques et applications, je décris mes travaux en cours ainsi que mes perspectives pour les années à venir.

### 4.1 Centres d'Appels

Dans la période à venir, je voudrais surtout travailler sur des problématiques sur les niveaux tactiques et opérationnels de centres d'appels. Les problématiques opérationnelles sont surtout liées au contexte grandissant des centres d'appels multi-canaux.

Au sujet du niveau tactique, je travaille actuellement sur la planification des agents avec des paramètres incertains, dans un contexte multi-vacation. Mon approche est basée sur la programmation par contraintes probabilistes, où le manager doit estimer un niveau de risque acceptable pour le non respect du niveau de service.

En ce qui concerne le niveau opérationnel, nous avons obtenu récemment des données sur des centres d'appels qui passent de la publicité pendant l'attente en ligne. Pour certaines entreprises utilisant des centres d'appels (notamment les annuaristes), le modèle économique basé sur les numéros surtaxés est en train de disparaître pour être remplacé par un nouveau modèle basé sur la publicité en faisant intervenir de tierces parties. Mon objectif ici est d'analyser l'impact de ce nouveau modèle économique sur les performance d'un centre d'appels, en termes d'attente et de comportement d'abandon.

Toujours sur le niveau opérationnel, je voudrais développer et analyser de nouveaux modèles de centres d'appels qui soient adaptés à la multiplication de l'utilisation des canaux (autres que le téléphone). On va se concentrer en particulier sur le système de chat en pleine expansion en ce moment vu son efficacité à priori (un agent peut traiter plusieurs clients à la fois). On va aussi s'intéresser à l'option de rappel et à l'analyse de son utilité pour lisser la variabilité de la demande.

## 4.2 Services d'Urgence

Dans un contexte économique difficile, les managers de services d'urgence essayent continuellement de réduire le gap entre les resources disponibles et la demande, dans le but de satisfaire les patients à moindre coût. Mes travaux s'intègrent dans ce cadre.

Je travaille actuellement sur une analyse critique des indicateurs de performances de congestion dans les services d'urgence. Il s'agit d'identifier les avantages et les inconvénients d'utiliser une métrique au lieu d'une autre. Ceci pourrait faire émerger des propositions de combinaisons d'indicateurs de performances.

Je suis aussi sur l'optimisation, sous contrainte de budget, des ressources humaines (médecins juniors et séniors, infirmières, et brancardiers) et matérielles (box de déchoquage, lits, etc.). L'approche d'optimisation utilisée est basée sur la simulation à évènements discrets.

En plus des ressources, je vise également à travailler sur l'optimisation des processus qui pourraient améliorer la performance du service d'urgence en termes d'attente des patients. On pense en particulier à se focaliser sur les modes de triage et de routage des patients.

## 4.3 Processus Stochastiques et Applications

Je voudrais continuer à contribuer à l'analyse de modèles théoriques. Comme déjà mentionné, l'intérêt de tels modèles provient de leur généricité pour pouvoir être utilisés dans de nombreuses applications. Je suis assez méfiant quant à l'utilisation trop rapide de résultats théoriques dans le cadre d'une application spécifique. Par exemple, l'application de résultats théoriques pour les centres d'appels ou les service d'urgences, sans une bonne appropriation au préalable des

caractéristiques spécifiques de leurs contextes, ne garantit pas des résultats appropriés et utiles.

J'ai commencé à explorer la problématique d'optimisation des rendez-vous de clients arrivants à un système de file d'attente. L'objectif est de trouver le bon compromis entre les deux notions contradictoires de temps d'attente des clients et de temps libre de la capacité. Je vais travailler sur des modèles qui intègrent l'hétérogénéité des clients en termes de ponctualité et d'absence, et je vais analyser l'impact de tels comportements sur les performances du système.

Je voudrais également travailler sur des stratégies de collaboration de partage de ressources entre plusieurs acteurs. Jusqu'à maintenant, tous mes travaux concernent le cas mono-acteur. En pratique, il serait aussi intéressant de penser à des schémas de mutualisation des ressources ainsi que des solutions de contractualisation entre acteurs indépendants. L'outil d'analyse à mobiliser est la théorie des jeux.

Les résultats exposés ci-dessous émergent d'un travail collectif avec des doctorants, postdocs et collègues, sans eux, rien n'aurait pu avoir lieu. Par souci de brièveté, je n'ai pas cité ces personnes dans ce résumé. Par contre, elles sont toutes citées sur les différentes parties du manuscrit là où je décris d'une façon plus détaillée mes travaux.

# Presentation of the Dissertation

My passionate interest in research started during my master of science in industrial engineering at Ecole Centrale Paris in 2003. I can still remember my motivation and enthusiasm for my first research projects. At that time, I realized that an academic career is exactly what I want to pursue. I had thereafter the opportunity to pursue my PhD thesis with the same group (Laboratoire Génie Industriel, LGI). That was a very positive experience. I was a part of a wonderful team: Yves Dallery and Mohamed Salah Aguir from LGI, and the call center research group of Bouygues Telecom. I had the chance to see how research results can be really useful for practice. Starting from discussions with Bouygues Telecom, research questions were raised, and then it was amazing to see how step by step we ended up with a better understanding of the problem, solutions that have been implemented by the company and followed by a step back concretized as a contribution to the scientific literature. My postdoctoral and research visiting experiences with Saif Benjaafar (University of Minnesota), Ger Koole (VU University Amsterdam) and Zeynep Aksin (Koç University) allowed me thereafter to work with the top researchers in my field. I owe all my results to them and to Yves Dallery, and I can never thank them enough. I was then appointed as an assistant professor at LGI where I enjoyed and I am still enjoying a nice equilibrium between my teaching and research activities. Today, based on my experience, and my teaching and research results, I feel myself ready to state my case towards the HDR (*Habilitation à Diriger des Recherches*) qualification. This would allow me to be more autonomous and enlarge my responsibilities, which is an important step forward in my career. This is the subject of this dissertation.

The ultimate goal of my research is to provide academicians and practitioners with useful recommendations and insights. My contributions concern the operations management of call centers, and the analysis of stochastic models with applications to services. These are also the main topics of this dissertation. More recently, I have started to work on the operations management issues of emergency departments as well as other service applications. The related ongoing work is described in the last part of the dissertation. The objective of this dissertation is to synthesize and provide perspectives of my research work. It is divided into four parts as shown in Figure 1.

In Part I, I present my education and qualifications, my teaching and research outputs, my academic and industrial collaborations, as well as my administrative and editorials activities.

In Part II, I summarize my contributions to the literature on the operations management of call centers. The biggest part of my contributions pertains to this literature. In Chapter II.1, I give a background on call centers followed by a summary of my contributions on the modeling of call centers with impatient customers. My remaining results on call centers can be classified according to the standard three decision levels. Chapters II.2 and II.3 deal with my contributions to the strategic and tactical decisions, respectively. Those related to operational decisions are significant and are further classified into two families: delay information issues, and job routing issues. They are described in Chapters II.4 and II.5, respectively.

In Part III, I summarize my theoretical contributions to the analysis of stochastic processes. The studies have applications for the analysis of a wide range of service systems. This part is

divided into three chapters. Chapter III.1 deals with the computation of first passage times in birth-death processes. Chapter III.2 summarizes my contributions to the performance analysis of queueing systems with impatience. Chapter III.3 describes my results related to the dynamic control of queueing systems.

In Part IV, I present my ongoing work and research perspectives. Section 1 describes my future projects on call center operations. I will extend my previous work on planning under uncertain arrival parameters. I will also work on call center settings with advertisement during the waiting in the queue. Another set of studies concern new multi-channel issues. The topic in Section 2 is about emergency departments. I present my ongoing survey on the key performance indicators, also, my ongoing work on the optimization of human resources. I then describe my future project on the analysis of the impact of changing the service process on the emergency department performance. Finally, Section 3 focuses on two future projects related to the analysis of stochastic models and their applications to services. The first project deals with queueing systems with appointment-driven arrivals. The second one concerns collaboration strategies between independent queueing service systems.

It goes without saying that my contributions are the result of my collaborations with doctoral students, postdoctoral students and many colleagues, to which I want to express my heartfelt gratitude. Throughout the dissertation, their names are highlighted. I would like also to express my gratitude to my excellent department LGI and at the first place to its head Jean Claude Bocquet.

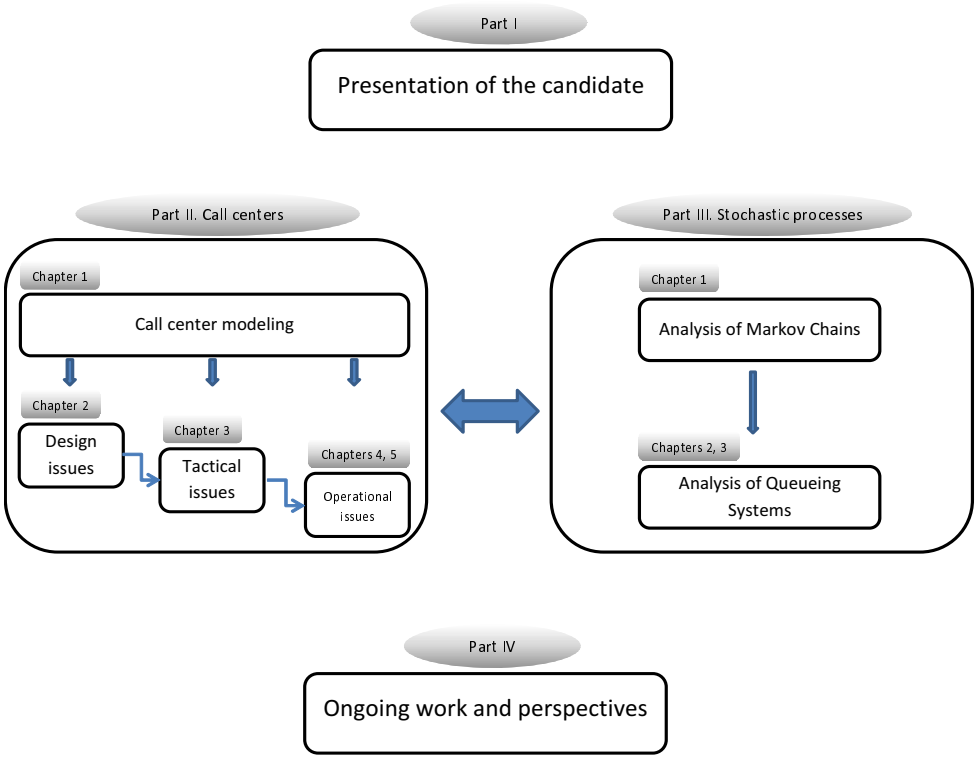



Figure 1: Organization of the dissertation

# Part I

# Presentation of the Candidate

# 1 Personal Details


- Oualid JOUINI
- Born on December 22, 1978 in Tunisia
- Tunisian-French dual citizenship
- Married, 1 child
- Professional Address: Ecole Centrale Paris, Laboratoire Génie Industriel, Grande Voie des Vignes, 92290 Châtenay-Malabry
- Email: `oualid.jouini@ecp.fr`
- Phone: +33 1 41 13 15 02
- Webpage: `http://www.lgi.ecp.fr/~jouini/`


# 2 Education

**PhD in Industrial Engineering, Ecole Centrale Paris**

- Defended on December 11, 2006, obtained with higher distinction
- Subject: Stochastic models for the analysis of call centers
- PhD advisor: Yves Dallery
- PhD Examination committee:

  - Philippe Chevalier (president), Professor, Université Catholique de Louvain, Belgium

  - Yannick Frein (reviewer), Professor, INPG, Grenoble

  - Ger Koole (reviewer), Professor, VU University Amsterdam, The Netherlands

  - Fabrice Chauvet (examiner), HDR, Head of the department of Simulation and Optimization, Gaz de France, Paris

  - Yves Dallery (PhD advisor), Professor, Ecole Centrale Paris, Paris.

**Master of Science in Industrial Engineering, Ecole Centrale Paris**

- Graduated on September 2003
- Ranking: 1 out of 10
- Research intership subject: Optimization of call centers (funded by Bouygues Telecom)
- Supervisors: Yves Dallery and Mohamed Salah Aguir.

**Engineering Diploma in Industrial Engineering, Ecole Nationale d'Ingénieurs de Tunis (Tunisia)**

- Graduated on June 2001
- Ranking: 4 out of 60
- Internship subject: Layout optimization for the manufacturing of electronic equipments in the telecommunication company Omniacom (Tunisia).

# 3 Professional Experiences

- **Since 2010:** Assistant Professor, Laboratoire Génie Industriel, Ecole Centrale Paris

  - Supervision of research projects

  - Teaching in industrial engineering, operations management, stochastic models, and simulation of production systems

  - Co-manager of the professional master PMTI (Purchasing Manager in Technology and Industry), 100% in English

  - Member of the team Decision Aid for Manufacturing and Service Systems, and member of the transverse group Industrial Engineering and Healthcare Management.

- **2008:** Postdoc at the Department of Industrial and System Engineering, University of Minnesota (USA).
- **2007-2009:** Teaching and Research Assistant, Laboratoire Génie Industriel, Ecole Centrale Paris.

# 4 Research Activities

My research interests are in stochastic modeling and service operations management with main applications to call centers and emergency departments. My primary goal is to assist managers to face the challenging task of organizing their processes more effectively and efficiently. I try to do that through deriving useful guidelines and recommendations. This pushes me always to account for the important features, in particular human modeling features, in order to be as close as possible to reality.

## 4.1 Research Supervision

I have participated in the supervision of 2 postdoctoral students, 5 PhD students, and about 25 research projects of master students. The details on the supervised postdoctoral and PhD students are given in Sections 4.1.1 and 4.1.2, respectively. I would like also to mention that, in the coming months, I will participate in the supervision of:

- The PhD student Guillaume Lamé

  - Starting in January 2015

  - Subject: Modeling and optimization of the clinical pathways at Henri-Mondor Hospital

  - Funded by the French ministry of higher education and scientific research

  - Expected supervision rate: 50%.

- A postdoctoral student, recruitment in progress

  - Expected to start in January 2015

  - Subject: Stochastic bilevel optimization

  - Funded by the Digiteo project SUN

  - Expected supervision rate: 40%.

#### 4.1.1 Supervision of Postdoctoral Students

Table 1 provides a summary of the supervised postdoctoral students. The details are given thereafter.

Table 1: Summary of the supervision of postdoctoral students

| **Student** | **Research topic** | **Supervision rate** | **Research outputs** |
|---|---|---|---|
| Mahdi Fathi | Analysis of emergency call centers | 100% | 1 ISI journal paper**;** 1 international conference communication |
| Benjamin Legros | Optimal routing solutions for blended call centers | 100% | 5 submitted or under revision papers for ISI journals**;** 1 international conference paper**;** 2 international conference communications |

**Mahdi Fathi, postdoc**

- From December 2013 to November 2014

- Research topic: Analysis of emergency call centers

- Funded by the JCJC ANR project OPERA

- Publications:

  - 1 accepted paper in an ISI Web of Science journal

  - 1 international conference communication.

**Benjamin Legros, postdoc**

- Ongoing, from January 2014 to August 2015
- Research topic: Optimal routing solutions for blended call centers
- Funded by the JCJC ANR project OPERA
- Publications:

  - 5 submitted or under revision papers for ISI Web of Science journals

  - 1 international conference paper

  - 2 international conference communications.

### 4.1.2 Supervision of PhD Students

Table 2 provides a summary of the supervised PhD students. The details are given thereafter.

Table 2: Summary of the supervision of PhD students

| **Student** | **Research topic** | **Supervision rate** | **Research outputs** |
|---|---|---|---|
| Shuangqing Liao | Staffing and shift-scheduling of call centers under call arrival rate uncertainty | 30% | 1 ISI journal paper**;** 2 international conference papers |
| Benjamin Legros | Optimization of multi-skill and multi-channel call centers | 70% | 2 ISI journal papers**;** 3 international conference communications |
| Mathilde Excoffier | Stochastic programming approaches for workforce scheduling of call centers with uncertain demand forecasts | 40% | 1 book chapter**;** 1 international conference communication |
| Lisa Peng | Collaboration in service systems | 50% | 1 national conference communication |
| Karim Ghanes | Optimization of emergency healthcare systems | 70% | 1 submitted paper to an ISI journal**;** 1 international conference paper**;** 1 national conference communication |

**Shuangqing Liao, PhD student**

- Duration: 3.5 years. Defended on July 1, 2011
- Research topic: Staffing and shift-scheduling of call centers under call arrival rate uncertainty

- Funded by China Scholarship Council
- Supervision rate of 30%. Co-supervised with Yves Dallery, Ger Koole (VU University Amsterdam) and Christian van Delft (HEC Paris)
- Examination Panel:
  - Professor Abdel Lisser, University Paris Sud, president
  - Professor Zeynep Aksin, Koç University, Turkey, reviewer
  - Professor Jean-Philippe Vial, Université de Genève, Switzerland, reviewer
  - Professor Christian van Delft, HEC Paris, co-supervisor
  - Professor Ger Koole, VU University Amsterdam, The Netherlands, co-supervisor
  - Dr. Oualid Jouini, Ecole Centrale Paris, co-supervisor
  - Professor Yves Dallery, Ecole Centrale Paris, supervisor
- Publications:
  - 1 accepted paper in an ISI Web of Science journal
  - 2 international conference papers.

**Benjamin Legros, PhD student**

- Duration: 3.5 years. Defended on December 13, 2013
- Research topic: Optimization of multi-skill and multi-channel call centers
- Funded by Interact-iv.com
- Supervision rate of 70%. Co-supervised with Yves Dallery and Ger Koole (VU University Amsterdam)
- Examination Panel:
  - Professor Stephen Chick, INSEAD, president
  - Professor Zeynep Aksin, Koç University, Turkey, reviewer
  - Professor Raik Stolletz, University of Mannheim, Germany, reviewer
  - Professor Rob van der Mei, CWI, Amsterdam, The Netherlands, examiner
  - Sébastien Thorel, Interact-iv.com, examiner
  - Professor Ger Koole, VU University Amsterdam, The Netherlands, co-supervisor
  - Dr. Oualid Jouini, Ecole Centrale Paris, co-supervisor
  - Professor Yves Dallery, Ecole Centrale Paris, supervisor
- Publications:
  - 2 accepted papers in ISI Web of Science journals
  - 3 international conference communications.

**Mathilde Excoffier, PhD student**

- Ongoing. Started in October 2012
- Research topic: Stochastic programming approaches for workforce scheduling of call centers with uncertain demand forecasts
- Funded by the Digiteo project SPACE
- Supervision rate of 40%. Co-supervised with Céline Gicquel and Abdel Lisser from Université Paris Sud
- Publications:

  - 1 book chapter

  - 1 international conference communication.

**Lisa Peng, PhD student**

- Ongoing, started in February 2013
- Research topic: Collaboration in service systems
- Funded by the Collaboration in service systems
- Supervision rate of 50%. Co-supervised with Zied Jemai and Yves Dallery
- Publications:

  - 1 national conference communication.

**Karim Ghanes, PhD student**

- Ongoing, started in January 2013
- Research topic: Optimization of emergency healthcare systems
- Funded by ARS Ile-de-France
- Supervision rate of 70%. Co-supervised with Zied Jemai and Ger Koole
- Publications:

  - 1 submitted paper to an ISI journal

  - 1 national conference communication.

## 4.2 My Publications

A summary of my research outputs is given in Table 3. The complete lists are detailed thereafter.

Table 3: Summary of my research publications

| | |
|---|---|
| Papers in international peer reviewed journals (ISI Web of Science) | 18 |
| Peer reviewed book chapters | 1 |
| Papers in professional magazines | 2 |
| Papers in international peer reviewed conference proceedings | 13 |
| Communications in international peer reviewed conferences | 23 |
| International patents | 1 |

**Papers in International Peer Reviewed Journals (ISI Web of Science)**

1. B. Legros, O. Jouini and G. Koole. A Flexible Architecture for Call Centers with Skill-Based Routing. ***International Journal of Production Economics***. In Press (DOI: 10.1016/j.ijpe.2014.09.025), 2014.

2. M. Fathi, F. Zandi and O. Jouini. Modeling the Merging Capacity for Two Streams of Product Returns in Remanufacturing Systems. ***Journal of Manufacturing Systems***. In Press (DOI: 10.1016/j.jmsy.2014.08.006), 2014.

3. B. Legros, O. Jouini and G. Koole. Adaptive Threshold Policies for Multi-Channel Call Centers. ***IIE Transactions***. In Press (DOI:10.1080/0740817X.2014.928965), 2014.

4. O. Jouini, Z. Aksin, F. Karaesmen, M.S. Aguir and Y. Dallery. Call center Delay Announcement Using a Newsvendor-Like Performance Criterion. ***Production & Operations Management***. In Press (DOI: 10.1111/poms.12259), 2014.

5. R. Wang, O. Jouini and S. Benjaafar. Service Systems with Finite and Heterogeneous Customer Arrivals. ***Manufacturing & Service Operations Management***, 16:365-380, 2014.

6. O. Jouini and A. Roubos. On Multiple Priority Multi-Server Queues with Impatience. ***Journal of the Operational Research Society***, 65:616-632, 2014.

7. S. Ioannidis, O. Jouini, A.A. Economopoulos and V.S. Kouikoglou. Control Policies for Single-Stage Production Systems with Perishable Inventory and Customer Impatience. ***Annals of Operations Research***, 209:115-138, 2013.

8. M.Z. Babai and O. Jouini. Operations Management in Service Systems (Editorial). ***IMA Journal of Management Mathematics***, 24:135-136, 2013.

9. A. Roubos and O. Jouini. Call Centers with Hyperexponential Patience Modeling. ***International Journal of Production Economics***, 141:307-315, 2013.

10. O. Jouini, G. Koole and A. Roubos. Performance Indicators for Call Centers with Impatience. ***IIE Transactions***, 45:359-372, 2013.

11. O. Jouini. Analysis of a Last Come First Served Queueing System with Customer Abandonment. ***Computers & Operations Research***, 39:3040-3045, 2012.

12. S. Liao, G. Koole, C. van Delft and O. Jouini. Staffing A Call Center with Uncertain Non-Stationary Arrival Rate and Flexibility. ***OR Spectrum***, 34:691-721, 2012.

13. O. Jouini, Z. Aksin and Y. Dallery. Call Centers with Delay Information: Models and Insights. ***Manufacturing & Service Operations Management***, 13:534-548, 2011.

14. O. Jouini, A. Pot, G. Koole and Y. Dallery. Online Scheduling Policies for Multi-class Call Centers with Impatient Customers. ***European Journal of Operational Research***, 207:258-268, 2010.

15. O. Jouini, Y. Dallery and Z. Aksin. Queueing Models for Multi-Class Call Centers with Real-Time Anticipated Delays. ***International Journal of Production Economics***, 120:389-399, 2009.

16. O. Jouini, Y. Dallery and R. Nait-Abdallah. Analysis of the Impact of Team-Based Organizations in Call Centers Management. ***Management Science***, 54:400-414, 2008.

17. O. Jouini and Y. Dallery. Moments of First Passage Times in General Birth-Death Processes. ***Mathematical Methods of Operations Research***, 68:49-76, 2008.

18. O. Jouini and Y. Dallery. Monotonicity Properties for Multi-server Queues with Reneging and Finite Waiting Lines. ***Probability in the Engineering and Informational Sciences***, 21:335-360, 2007.

**Peer Reviewed Book Chapters**

1. M. Excoffier, C. Gicquel, O. Jouini, A. Lisser. A Stochastic Programming Approach for Staffing and Scheduling Call Centers with Uncertain Demand Forecasts. ***Lecture Notes in Communications in Computer and Information Science, SPRINGER-VERLAG***, 2014.

**Papers in Professional Magazines**

1. O. Jouini and G. Koole. Including Abandonments in Call Center Staffing Models. ***Softigator, Business Networking to Call Center professionals***, February 2008.

2. O. Jouini and G. Koole. Team-Based Organizations in Call Centers. ***Softigator, Business Networking to Call Center professionals***, June 2008.

**Papers in International Peer Reviewed Conference Proceedings**

1. K. Ghanes, O. Jouini, Z. Jemai, M. Wargon, G. Koole, R. Hellmann, V. Thomas. A Comprehensive Simulation Modeling of an Emergency Department: A Case Study for Simulation Optimization of Staffing Levels. ***Proceedings of the Winter Simulation Conference***, 2014, Savannah, USA.

2. B. Legros, O. Jouini, G. Koole. Imbricating Tasks in a Multi-channel Contact Center. ***Proceedings of ICMSAO***, 2013, Hammamet, Tunisie.

3. S. Ioannidis, O. Jouini and Y. Dallery. Production and Sales Control in Systems with Flexible Capacity and Perishable Items. ***Proceedings of SMMSO***, 2013, Kloster Seeon, Germany.

4. O. Jouini, Z. Aksin, F. Karaesmen and Y. Dallery. Data-Based Analysis of Delay Estimators and Announcements in a Call Center. ***Proceedings of SMMSO***, 2011, Izmir, Turkey.

5. S. Liao, C.van Delft, G. Koole and O. Jouini. Shift-Scheduling of Call Centers with Uncertain Arrival Parameters. ***Proceedings of MOSIM***, 2010, Hammamet, Tunisie.

6. S. Liao, C. van Delft, G. Koole and O. Jouini. Call center capacity allocation with newsboy model. ***Proceedings of CIE39***, 2009, Troyes, France.

7. O. Jouini and S. Benjaafar. Appointment Scheduling with Non-Punctual Arrivals. ***Proceedings of INCOM 2009***, Moscow, Russia.

8. O. Jouini and Y. Dallery. Stationary Delays for a Two-Class Priority Queue with Impatient Customers. ***Proceedings of VALUETOOLS***, 2007, Nantes, France. "

9. O. Jouini and Y. Dallery. Modeling Multi-class Call Centers with Delay Information. ***Proceedings of IESM***, 2007, Beijing, China.

10. O. Jouini and Y. Dallery. Estimating and Announcing Waiting Times in Multiple Customer Class Call Centers. ***Proceedings of INCOM***, 2006, Saint-Etienne, France.

11. O. Jouini and Y. Dallery. Real-Time Scheduling Policies for Multi-class Call Centers. ***Proceedings of IEEE-SSSM***, 2006, Troyes, France.

12. O. Jouini and Y. Dallery. Predicting Queueing Delays for Multi-class Call Centers. ***Proceedings of VALUETOOLS***, 2006, Pisa, Italy.

13. O. Jouini and Y. Dallery. Stochastic Models of Customer Portfolio Management in Call Centers. ***Proceedings of the German Operations Research Society***, 2004, Tilburg, The Netherlands.

**Communications in International Peer Reviewed Conferences**

1. K. Ghanes, O. Jouini, Z. Jemai, M. Wargon, G. Koole, R. Hellmann, V. Thomas. Simulation-Based Optimization of an Emergency Department Staffing Levels. ***Euro Mini Conference on Stochastic Optimization and Energy applications***, 2014, Paris, France

2. B. Legros, O. Jouini, G. Koole. Optimal Scheduling of Calls in Call Centers with a Call Back Option. ***20th Conference of the International Federation of Operational Research Societies***, 2014, Barcelona, Spain.

3. O. Jouini, B. Legros, G. Koole. On the Scheduling of Jobs in a Contact Center with Idling Times during the Call Service. ***StochMod14***, Mannheim, Germany, 2014.

4. B. Legros, O. Jouini, G. Koole. Threshold Policy for Call Centers with a Call Back Option. ***14th International Conference on Project Management and Scheduling***, 2014, Munchen, Germany.

5. R. Wang, O. Jouini and S. Benjaafar. Service Systems with Finite and Heterogeneous Customer Arrivals. ***44th annual meeting of the Decision Sciences Institute***, 2013, Baltimore, USA.

6. M. Excoffier, A. Lisser, C. Gicquel, O. Jouini. Stochatic Programming Approaches for Staffing in Call Centers with Uncertain Forecasts. ***EURO-INFORMS Joint International Meeting***, 2013, Rome, Italy.

7. B. Legros, O. Jouini, G. Koole. Call Centers with a Call Back Option. ***EURO-INFORMS Joint International Meeting***, 2013, Rome, Italy.

8. B. Legros, O. Jouini, G. Koole. Optimal Routing in a Multi-Channel Call Center. ***Informs Annual Meeting***, 2013, Minneapolis, USA.

9. O. Jouini, Z. Aksin, F. Karaesmen and Y. Dallery. Data-Based Analysis of Delay Estimators and Announcements in a Call Center. ***Informs Annual Meeting***, 2012, Phoenix, USA.

10. R. Wang, O. Jouini and S. Benjaafar. Service Systems with Finite and Heterogeneous Customer Arrivals. ***MSOM Annual Conference***, 2011, Michigan, USA.

11. O. Jouini and S. Benjaafar. Queueing Systems with Appointment-Driven Arrivals, Non-Punctual Customers, and No-Shows. ***Informs Annual Meeting***, 2011, Austin, USA.

12. S. Ioannidis, O. Jouini, A. Economopoulos, and V. Kouikoglou. An (s 1, s) Inventory System with General Product Lifetimes and Customer Impatience. ***EURO Working Group on Stochastic Modeling***, 2010, Nafplio, Greece.

13. Y. Wang, O. Jouini and S. Benjaafar. Queueing Systems with Finite Arrivals. ***Informs Annual Meeting***, 2009, San diego, USA.

14. O. Jouini and G. Koole. Performance Models of a Single-skill Call Center. ***Informs Annual Meeting***, 2009, San diego, USA.

15. O. Jouini and S. Benjaafar. Appointment-driven Queueing Systems with Non-punctual Customers. ***Informs Annual Meeting***, 2009, San diego, USA.

16. Z. Aksin, O. Jouini and Y. Dallery. Call Centers with Delays Announcement. ***Informs Annual Meeting***, 2008, Washington DC, USA.

17. G. Koole and O. Jouini. Call Center Performance Indicators. ***Informs Annual Meeting***, 2008, Washington DC, USA.

18. O. Jouini, K. Haj Youssef and Y. Dallery. Quoting Customer Lead Times in a Make-to-Stock System. ***STOCHMOD08***, 2008, Istanbul, Turkey.

19. G. Koole and O. Jouini. Call Center Performance Indicators Calculations and Simulations. ***STOCHMOD08***, 2008, Istanbul, Turkey.

20. Z. Aksin, O. Jouini and Y. Dallery. Call Centers with Delays Information: Models and Insights. ***Informs Annual Meeting***, 2007, Seattle, USA.

21. O. Jouini, A. Pot, Y. Dallery and Ger Koole. Real-Time Dynamic Scheduling Policies for Multi-class Call Centers with Impatient Customers. ***Informs Annual Meeting***, 2007, Seattle, USA.

22. A. Pot, O. Jouini, G. Koole and Y. Dallery. An Online Policy for Call Centers. ***Applied Probability Society of Informs Conference***, 2007, Eindhoven, The Netherlands.

23. O. Jouini, M.S. Aguir and Y. Dallery. Analysis of a Skill-Based Routing Call Center Model. ***Applied Probability Society of Informs Conference***, 2007, Eindhoven, The Netherlands.

**International Patents**

1. O. Jouini and Y. Dallery (Ecole Centrale Paris). F. Auriol, O. Belma, F. Chauvet, R. Nait-Abdallah and T. Prat (Bouygues Telecom). International Patent. Client Portfolio-Based Call Center Architecture. International publication number WO 2006/003306, ***World Intellectual Property Organization***.

## 4.3 Evidence of Research Esteem

The evidence of my research esteem is described through the following collection of indicators.

### 4.3.1 Research Grants

I have participated to several research projects funded by companies, institutions, and councils. In what follows, I mention those where I was the principal investigator or the principal co-investigator. A summary of these projects are given in Table 4, and the details are given below.

Table 4: Summary of funded research projects

| **Project** | **Topic** | **Grant** |
|---|---|---|
| Digiteo, SUN | Stochastic optimization of uncertain bilevel problems | 102 000 € |
| Agence Régionale de Santé Ile-de-France | Optimization of emergency departments | 150 000 € |
| ANR Jeunes Chercheurs, OPERA | Operations management in call centers | 125 000 € |
| Digiteo, SPACE | Stochastic programming approaches for workforce scheduling of call centers with uncertain demand forecasts | 100 000 € |
| Interact-iv.com | Optimization of multi-channel call centers | 73 000 € |

**Digiteo project SUN: Stochastic optimization of uncertain bilevel problems**

- From January 2015 to December 2016
- Partner: Université Paris Sud
- Co-investigator with Abdel Lisser (Université Paris Sud)
- Funded by Digiteo, postdoctoral funding, total grant: 102 000 €.

**Project with Agence Régionale de Santé Ile-de-France: Optimization of emergency departments**

- From January 2013 to December 2015
- Partner: Hospital Saint Camille
- Principal investigator. Collaboration with Ger Koole and Zied Jemai
- Funded by ARS: 150 000 €.

**ANR project OPERA : Operations management in call centers**

- From October 2012 to September 2015

- Principal investigator and coordinator. Collaboration with Ger Koole, Zeynep Aksin, Zied Jemai and Yves Dallery
- Funded by the ANR Jeunes Chercheurs program: 125 000 €.

**Digiteo project SPACE: Stochastic programming approaches for workforce scheduling of call centers with uncertain demand forecasts**

- From September 2012 to August 2015
- Partner: Université Paris Sud
- Co-investigator with Céline Gicquel and Abdel Lisser (Université Paris Sud)
- Funded by Digiteo, PhD thesis funding, total grant: 100 000 €.

**Interact-iv.com: Optimization of multi-channel call centers**

- During 2011-2012
- Principal investigator
- Funded by the consulting company Interact-iv.com: 73 000 €.

### 4.3.2 Participation to Scientific Conference Committees

- Member of the scientific committee of EURO Working Group on Stochastic Modeling, since 2012.
- Member of the scientific committee of the 3rd International Symposium & 25th National Conference on Operational Research, Volos, Greece, June 2014.
- Member of the program committee of Euro Conference on Stochastic Programming and Energy Applications, Paris, September 2014.

### 4.3.3 Organization of Conference Tracks and Special Sessions

- Chair of the special session "Operations Management in Service Systems" in the international conference on Computers and Industrial Engineering, Troyes, 2009.
- Chair of the special session "Optimization in Service Systems" in MOSIM'10, Tunis, Tunisia, 2010.
- Co-chair of the track "Operations Management" in the 5th International Conference on Modeling, Simulation And Applied Optimization, Hammamet, Tunisia, 2013.
- Co-chair of the track "Service Systems" in IIE Annual Conference, Nashville, USA, 2015.

### 4.3.4 Research Visiting Positions

- **May-June 2014: Singapore University of Technology and Design**, Pillar of Engineering Systems and Design, Singapore. Collaboration with Saif Benjaafar and Rowan Wang on the analysis of service systems with a finite number of non-punctual and heterogeneous customers.

- **Mai-July 2007, June 2009, November 2013: Koç University**, College of Administrative Sciences and Economics, Turkey. Collaboration with Zeynep Aksin and Fikri Karaesmen on various subjects related to delay information models for call centers.

- **July 2010: University of the Aegean**, Department of Mathematics, Greece. Collaboration with Ioannidis Efstratios and Vassilis Kouikoglou on the analysis of perishable inventory systems.

- **3 month from 2007 to 2009: VU University Amsterdam**, Department of Applied Mathematics, The Netherlands. Collaboration with Ger Koole, Alex Roubos, Auke Pot on various subjects related to the optimal routing of jobs in call centers.

### 4.3.5 Academic Collaborators

- Yves Dallery, Ecole Centrale Paris. He is my PhD advisor. Collaboration until now on call centers issues, collaboration on 1 funded project, co-supervision of 3 PhD students, 7 published papers.

- Ger Koole, Auke Pot and Alex Roubos, VU University Amsterdam, The Netherlands. Collaboration on various call centers topics. Co-supervision of 2 PhD students, collaboration on 3 funded projects, 5 published papers, 4 submitted papers.

- Zeynep Aksin, Fikri Karaesmen, Koç University, Turkey. Collaboration on call center models with delay information, 3 published papers, 1 working paper.

- Saif Benjaafar (University of Minnesota, USA) and Rowan Wang (Singapore Management University, Singapore). Collaboration on service systems with finite arrivals, 1 published paper, 1 working paper.

- Stratos Ioannidis Efstratios and Vassilis Kouikoglou, Technical University of Crete, Greece. Collaboration on inventory problems with perishable items and impatient customers, 1 published paper, 1 working paper.

- Céline Gicquel and Abdel Lisser, Université Paris Sud. Collaboration on call center planning and stochastic bilevel optimization, collaboration on 2 funded projects, co-supervision of 1 PhD student and 1 postdoctoral student, 1 book chapter, 1 working paper.

- Zied Jemai, Ecole Nationale d'Ingénieurs de Tunis, Tunisia. Collaboration on cooperation in service systems and on emergency departments, collaboration on 2 funded projects, co-supervision of 2 PhD students, 3 working papers.

- Christian van Delft, HEC Paris. Collaboration on call center staffing with uncertain parameters, co-supervision of 1 PhD student, 1 published paper.

#### 4.3.6 Distinctions

- M. Excoffier, C. Gicquel, O. Jouini, A. Lisser. A Stochastic Programming Approach for Staffing and Scheduling Call Centers with Uncertain Demand Forecasts. **Finalist for the Best Paper Award**. *ICORES*, Angers, France, 2014.

- **Merit based scholarship** from TÜBİTAK, The Scientific & Technological Research Council of Turkey, 2007.

- O. Jouini, Z. Aksin and Y. Dallery. Call Centers with Delay Information. **Honorable mention for the excellence in paper content**. *IESM Conference*, 2007, Beijing, China.

- O. Jouini, M.S. Aguir and Y. Dallery (Ecole Centrale Paris), Z. Aksin and F. Karaesmen (Koç University), F. Chauvet, R. Nait-Abdallah and T. Prat (Bouygues Telecom). Improving Call Center Operations at Bouygues Telecom. **Semi-Finalist for the INFORMS Edelman Award**, 2005.

## 5 Teaching Activities

My teaching and student supervision activities belong to the teaching department *Science de l'Entreprise* at Ecole Centrale Paris. I am involved in the following programs: the engineering program (1st, 2nd and 3rd year levels), the master of science in industrial engineering OSIL, the professional master on supply chain MIPSC, and the professional master on purchasing PMTI. The hourly volume I teach per year at Ecole Centrale Paris is around 135. The details for the courses I am responsible on at Ecole Centrale Paris are given in Table 5. I also regularly give lectures in operations management for undergraduate students at EMLyon and Neoma Business School (around 18 hours per year), and doctoral lectures in stochastic processes at HEC Paris (around 21 hours per year).

I would like to underline the coherence between my teaching and research topics. I think that this coherence has and will have an added value on both activities. On the one hand, it allows to continuously enrich the contents treated in my lectures. On the other hand, it allows me to improve the communication of my research work by popularizing it.

For the different programs at Ecole Centrale Paris, I supervise each year professional projects, innovation projects, national and international internships in companies, and national and international internships in academic institutions. This corresponds to about 100 hours per year.

Table 5: Summary of the courses I am responsible on at Ecole Centrale Paris

| Course | Program | Volume | Period |
|---|---|---|---|
| Processus stochastiques et files d'attente | 2nd year | 36h | Since 2008 |
| Modèles stochastiques et applications | 3rd year - OSIL | 24h | Since 2014 |
| Modélisation et simulation des systèmes de production | 3rd year - OSIL - MIPSC - PMTI | 24h | Since 2009 |
| Prévision de demande | 3rd year - OSIL | 24h | Since 2010 |
| Introduction au génie industriel | 1st year | 6h | Since 2008 |
| Jeu de supply chain | MIPSC | 12h | Since 2009 |
| Decision aid tools | PMTI | 9h | Since 2010 |

# 6 Administrative and Editorial Activities

## 6.1 Administrative Responsibilities

- **From 2010 to 2013:** Co-manager of the professional master PMTI (Purchasing Manager in Technology and Industry, 100% in English) with Eric David at Ecole Centrale Paris:

  - Definition of the program (lecturers, case studies, industrial visits, seminars, etc.)

  - Recruitment of the professors

  - Coordination of the link with the company partners of the master

  - Coordination of the link with the department of studies at Ecole centrale Paris

  - Recruitment of students, population of 20 students

  - Accompanying the student for their choice of projects, internships, career orientation, etc.

  - **Distinctions delivered by SMBG:** The innovation award in 2011; second rank for the launching program in 2012; Second rank among French purchasing programs in 2013.

- **June 2012:** Chair Organizer of the EURO Workshop on Stochastic Modeling at Ecole Centrale Paris, `http://www.lgi.ecp.fr/StochMod2012/pmwiki.php`

  - 70 participants

  - Advertisement for the workshop

  - Managing of the workshop budget, registration, social program, transportation, etc.

  - Organization of the plenary talks

  - Coordination of the abstracts review process.

## 6.2 Editorial Activities

- **Since 2010:** Member of the Editorial Board of the journal *International Journal of Information Systems in the Service Sector*.
- **Since 2012:** Associate Editor for the journal *IMA Journal of Management Mathematics*.
- **2013:** Organization of a special issue on "Operations Management in Service Systems" for the journal *IMA Journal of Management Mathematics*, with Zied Babai (Kedge Business School).
- **Since 2014:** Associate Editor for the journal *Supply Chain Forum, an International Journal*.
- **Reviewing of proposals:** I am involved in the review process of the following research councils:

  - ANR (Agence Nationale de la Recherche), France
  - NWO (The Netherlands Organization for Scientific Research), The Netherlands
  - CNCS (National Research Council), Romania
  - Canadian Network of Centres of Excellence, Mitacs Proposals, Canada.

- **Reviewing of papers.** I regularly review papers for the following journals:

  - Management Science
  - Manufacturing & Service Operations Management
  - Operations Research
  - Productions & Operations Management
  - International Journal of Production Economics
  - European Journal of Operational Research
  - IIE Transactions
  - Journal of the Operational Research Society
  - International Journal of Production Research.

# Overview on My Research Activities

In what follows, I give an overview on my research activities. I describe their progress over time, the link between them, as well as their context (PhD thesis, collaboration with colleagues and companies, research contract, etc.).

My research activities started when I joined Ecole Centrale Paris in 2002 to carry out my master of science in industrial engineering within Laboratoire Génie Industriel (LGI). I started thereafter a PhD and received my degree in 2006. For both of them, my master and PhD, my research results concerned the analysis and optimization of the Bouygues Telecom call center, under the supervision of Yves Dallery. I have worked on strategic and operational issues proposed by our partners from Bouygues Telecom, namely Fabrice Chauvet and Rabie Nait-Abdallah (research and development), and Olivier Belma and Thierry Prat (information and telecommunication systems). I have also contributed to the literature on the analysis of queueing systems. We addressed theoretical questions that are motivated by the issues encountered during the quantitative analysis of call center models.

After my PhD, I was appointed in 2007 as a teaching and research assistant with the same group LGI. I belong to team 2 of LGI *Decision Aid for Production and Service Systems* and my concern is related to the first research project of Team 2 *Service Operations Management*. I have continued after my PhD to work on call centers by considering a broad panel of call center operations management (OM) issues thanks to our industrial partners (Bluelink, Interact-iv.com, Digiway Consulting, etc.), and to the colleagues Ger Koole (VU University Amsterdam) and Zeynep Aksin (Koç University) that I have frequently and regularly visited from 2007 to 2009. During that period, Ger Koole has also spent several months at LGI as a visiting professor.

Starting from an OM issue, my approach consists of first building a stochastic model, and then developing a quantitative analysis to obtain response elements for the addressed question. I was then often confronted to challenges related to the theory of stochastic processes. This have motivated me to attempt to contribute to that literature in order to serve not only my specific call center issues, but also a wide range of service system situations.

In 2007, I obtained with Zeynep Aksin a merit based scholarship from TÜBİTAK (the scientific and technological Research council of Turkey). This helped us to continue our projects on the analysis of call centers with customer delay information. At the same time, I was closely collaborating with Yves Dallery, Ger Koole and also Auke Pot from VU University Amsterdam. We have mainly worked on call routing issues. We have been also in contact, Ger Koole and I, with Bluelink (the service provider of Air France KLM) and Digiway Consulting (a consulting company for call centers). The latter gave us valuable data and motivated us to work on agent planning problems with parameter uncertainty. Thanks to a funding from the China Scholarship Council, we have launched the PhD thesis of Shuangqing Liao on this subject in 2008. We have collaborated for the supervision of the PhD thesis with Christian van Delft from HEC Paris who

is an expert of stochastic programming. Shuangqing has defended her PhD in 2011.

During 2008, I pursued a postdoc at University of Minnesota with Saif Benjaafar, where I have worked on the analysis of queueing systems with a finite number of heterogenous arrivals, with application to event-driven service systems. Since then, I continued to work with Saif on this topic. In 2010, his PhD student Rowan Wang joined us to work on related extensions. During the same year, they both visited me two months at Ecole centrale Paris.

After my postdoc, I came back to Ecole Centrale Paris, and I was appointed there in 2010 as an assistant professor. In 2010, I obtained a funding from Interact-iv.com, which is a company working on software and hardware solutions for call centers. The funding served to launch the PhD thesis of Benjamin Legros on routing problems for multi-skill and multi-channel call centers. This allowed to intensify my results on call center operational issues, and also theoretical results on queueing systems. Benjamin has defended his PhD in 2013. In 2010 and 2011, I have worked with Ger Koole and his PhD student Alex Roubos on the modeling of customer abandonment times. With Alex, I have also worked on the performance analysis of multi-class queueing systems.

Later on, in 2012, I obtained with Céline Gicquel and Abdel Lisser from Université Paris Sud a PhD funding from Digiteo that allowed us to recruit Mathilde Excoffier and continue the work on the rich problem of planning using further stochastic optimization approaches. In 2012, I also obtained a funding from Agence Nationale Recherche, thanks to which, I have recruited in 2013 the two postdocs Benjamin Legros and Mahdi Fathi and I have continued to extend my work on call centers operational problems.

In summary, after my PhD, I have extended my work and contributed to the literature on various call center issues. In parallel, I continued to contribute to the theoretical literature on stochastic processes with application to services that can be modeled as queueing systems. This has allowed me to make a step by step progress towards reaching my research objective which is to contribute to the literature on stochastic modeling and operations management of service systems.

Recently, in 2013, I have started to work on emergency department operations under the PhD thesis of Karim Ghanes. We obtained for it a funding from Agence Régionale Santé Ile-de-France. We are since collaborating with real emergency departments on the optimization of resources and processes. An emergency department is a complex service system with important human features and societal impacts. This constitutes for me an excellent opportunity to carry on my progress on contributing to the OM of service operations.

I described above the timeline of my past research. In the remaining parts of the dissertation, I describe the structure of my research and synthesize the results. Part II summarizes my contributions to the literature on the operations management of call centers. Part III summarizes my theoretical contributions to the analysis of stochastic processes. The chapters of Parts II and III start with the motivation and the positioning of my contributions within the existing literature, and end with short-term avenues for future research. Finally, Part IV summarizes my ongoing work and research perspectives.

# Part II

# Operations Management in Call Centers

Call centers, also known as telephone, customer service, contact or customer interaction centers, have emerged as the primary vehicle for firms to interact with consumers, transforming consumer service jobs once characterized by variety and personal relationships into routinized and high speed operations. Call centers are used to provide services in many areas and industries: banks, insurance companies, emergency centers, information centers, help-desks, tele-marketing and more.

The continued growth of both importance and complexity of modern call centers has been came along an extensive and growing literature. Numerous related academic surveys focusing on various disciplines were published. The main disciplines related to call centers are Mathematics and statistics, operations research, operations management, information technology, human resource management, as well as psychology and sociology.

My contributions, synthesized in this part, are pertaining to the operations management literature on call centers. My general objective is to take a step back and enhance our understanding of the complex environment of call centers, so as we gain useful guidelines for practitioners. To the contrary to traditional work where usually a pure queueing analysis is performed, my research approach consists of developing and analyzing stochastic models that incorporate both customer and agent behavior key features. In call centers, customers and agents are human beings, and the human element is shown to have a significant impact on system performance. Therefore, incorporating human behavior into the quantitative analysis of models would yield more realistic and useful insights.

We distinguish three main issues dealing with the operations management in call centers. The first issue involves strategic or long-term decisions for the design of the facility. The second issue is related to medium-term aggregate planning of services. The third issue deals, in turn, with short-term decisions made on a daily or weekly basis. As illustrated in Figure 2, my research results are related to literature streams on these three decision levels, in addition to the literature on call center modeling.

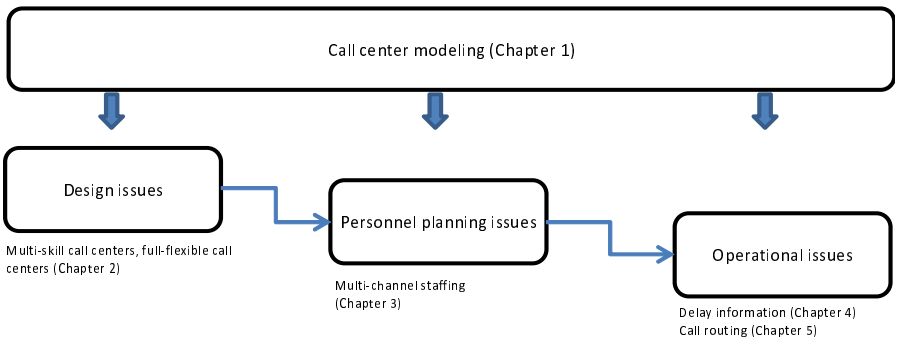



Figure 2: My contributions to the operations management literature of call centers

# Chapter II.1

# Modeling of Call Centers

## 1 Introduction

A call center is a service system. It is a facility designed to support the delivery of some interactive service via communications channels. 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. Due to the uncertainty governing the call center environment (customers and agents behaviors), the literature has standardly addressed its issues using stochastic models, and in particular queueing models.

This chapter summarizes my contributions to the literature on call center queueing modeling. After introducing a brief background, I describe my contributions on the performance indicators for call centers with customer impatience. The importance of modeling abandonments in call centers is emphasized by Garnett et al. (2002), Gans et al. (2003), and Mandelbaum and Zeltyn (2009). Empirical evidence regarding abandonments in call centers can be found in Brown et al. (2005) and Feigin (2005). This work has been done with Ger Koole and the PhD student Alex Roubos. Based on real-life data, we propose new models for the patience time distribution. We also study a number of different service level definitions, including all those used in practice. Through a quantitative analysis, we emphasize for call center managers the importance of choosing the right metrics.

## 2 Call center Background

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 require 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 the 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, it points 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.

Figure II.1.1 depicts an operational scheme of a simple call center as a queuing system. The trunk lines connect calls to the center while a group of agents serve incoming calls. An arriving call that finds all the trunk lines occupied receives a busy signal and is blocked from entering the system. Otherwise it is connected to the call center and occupies one of the free trunk lines. If some of the agents are available, the call is served immediately. Otherwise, it waits in the queue for an agent to become available. Callers who become impatient hang up, or abandon, before getting into service. Some of the blocked and abandoned calls become retrials that attempt to reenter service. The remaining of them are lost. Finally, it is also possible that served caller may return to the system.

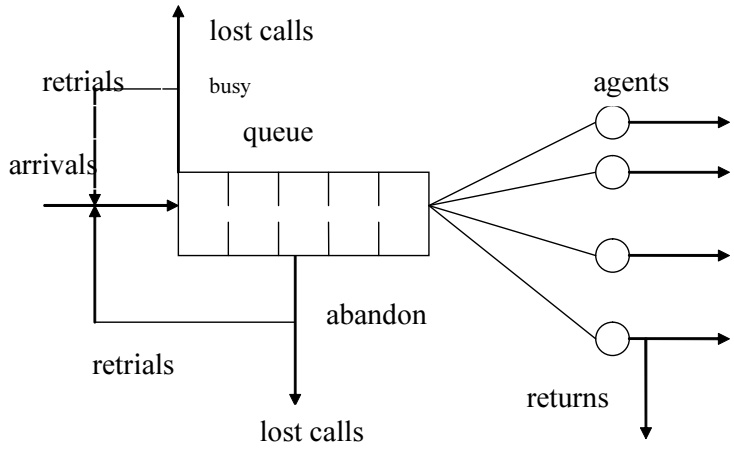



Figure II.1.1: Operational scheme of a simple call center

Key performance indicators (KPIs) are critical for the successful management of call centers. The right metrics identify the causes of problems and generate solutions that change the results. It is almost impossible to develop a universal set of KPIs that will work equally well in every situation in every call center. Every business unit is different, with its unique structure and problems. Still, it is possible to formulate a set of KPIs useful for most call centers. Correct measurement of such KPIs will offer call center managers valuable information.

KPIs can be classified into two families: those that are product related and those that are process related. Product-related metrics are performance indicators mostly related to the content of the call, while process-related metrics are performance indicators that are related to call center operations. The most well-known call center product-related metrics that managers

can use to improve customer experience are: First Call Resolution; Turnover; Attendance and Punctuality; Contact Quality; and Customer Satisfaction. The most familiar process-related metrics used in call centers are: Probability of Blocking; Probability of Abandonment; Short Abandonments; Service Level; Average Speed of Answer; Longest Delay in Queue; and Agent Occupancy.

On the one hand, basing an entire service strategy on the number of calls handled per hour or on the average speed of answer will inevitably lead to shortcomings in the quality. On the other hand, focusing too strongly on quality metrics while disregarding process-related measurements can still have an adverse effect on customer experience.

## 3 Performance Indicators with Impatience

We focus in what follows on metrics related to queueing delays. These are classic process-related metrics that lie at the heart of effective call center and customer relations management. They are the clearest indication of what customers experience when they attempt to reach the call center. We in particular focus on metrics related to the important feature of customer impatience.

One important point has to be clarified before impatience can be included in queueing models. That is, we need additional information concerning the patience, the willingness to wait until service commences. Similarly as for the input of the Erlang C model (simple queue with Markovian assumptions), the patience has to be determined from historical data. However, a number such as the average patience cannot be determined by simply averaging over the abandonment times. Indeed, the time at which other calls got connected tells us something about their patience, which should be taken into account. Statistical techniques exist to deal with these so-called censored data. Not using these methods can lead to a significant underestimation of patience, because the abandonments occur mostly among the very impatient customers. We conduct a statistical analysis on real call center data in order to characterize the statistical distribution of times before abandonments.

By taking abandonments into account, the computations become more difficult. Moreover, even when patience times are assumed to have an exponential distribution (the Erlang A model), there exist only expressions for some metrics, such as the conditional waiting time given service. In this work, we give a comprehensive list of the metrics including abandonments, and explicitly derive the expressions for the probability distributions of these metrics. By doing so, we obtain existing results and derive new ones, such as the conditional waiting time given service of the customers who do not have short patience times.

### 3.1 Statistical Analysis and Modeling of Abandonments

To analyze the patience, we need to know how long customers have spent waiting, and whether an abandonment occurred at the end of the waiting time. From customers that have abandoned, we know exactly what their patience is. However, from customers that did not abandon (but

received service), we only know that their patience is greater than the time they have waited. To be more precise, we observe the minimum of the patience and the virtual waiting time, and we also know which one we observe. This is called right-censored data. Techniques exist to deal with censored data, one of which is the Kaplan-Meier estimator (see Kaplan and Meier, 1958). In our statistical analysis, we use data obtained from several real call centers. The data originate from a large banking call center located in the US, from a bank located in the Netherlands, from a bank located in Israel, and from a Dutch university medical center.

The result of the Kaplan-Meier estimator is the empirical cumulative distribution function $F(t)$ of the patience. By taking the derivative we can obtain the probability density function $f(t)$, and the hazard rate $h(t) = f(t)/(1-F(t))$. The empirical hazard rates are smoothed three times using a moving-average filter with a span of five, to produce better-looking lines. The patience on all four data sets can, for the most part, be characterized in the same way. In the first couple of seconds the hazard rate is high, indicating very impatient customers who are not willing to wait at all. The hazard rate quickly becomes constant thereafter, which suggests that the patience from then on is exponential.

**Model 1:** A way to model this customer behavior is to extend Erlang A by including the possibility of balking. Let $T$ denote the random variable measuring the patience times. The distribution of $T$ consists of a discrete mass at zero corresponding to very impatient customers, and a remaining exponential distribution for customers with a positive patience. We denote by $\alpha$ the probability that a customer, arriving to a busy system, will immediately balk. This feature models a non-negligible portion of the customers who immediately hang up once they know that they have to wait for service. On the other hand, with probability $1-\alpha$, customers who find a busy system will accept to join the queue. For these customers, the patience thresholds are independent and exponentially distributed with rate $\gamma$. Hence, the cumulative distribution function is $F_T(t) = \alpha + (1-\alpha)(1-e^{-\gamma t})$, for $t \geq 0$.

**Model 2:** Another way to model customer patience is by the hyperexponential distribution with two phases. The hyperexponential distribution is a mixture of two exponential distributions such that with probability $p$ it is exponential with parameter $\gamma_1$ and with probability $1-p$ it is exponential with parameter $\gamma_2$. If $T$ is hyperexponential, its cumulative distribution function $F_T$ is given by $F_T(t) = p(1-e^{-\gamma_1 t}) + (1-p)(1-e^{-\gamma_2 t})$, for $t \geq 0$. The statistical analysis shows that the hyperexponential distribution fits the empirical patience very well. The parameters of the random variable $T$ are obtained by minimizing the mean squared error between $F(t)$ and $F_T(t)$.

## 3.2 Analysis of Call Center Metrics

Consider a call center model with a single class of customers and $s$ statistically identical, parallel servers. We assume that arrivals follow a Poisson process with rate $\lambda$, and that service times are exponentially distributed with rate $\mu$. The queueing discipline is first-come first-served

| | |
|---|---|
| $SL_1$ | $\frac{\#\text{ answered} \leq \tau}{\#\text{ offered}}$ |
| $SL_2$ | $\frac{\#\text{ answered} \leq \tau}{\#\text{ offered} - \#\text{ short abandonments}}$ |
| $SL_3$ | $\frac{\#\text{ answered} \leq \tau}{\#\text{ offered} - \#\text{ abandoned} \leq \tau}$ |
| $SL_4$ | $\frac{\#\text{ answered} \leq \tau}{\#\text{ answered}}$ |
| $SL_5$ | $\frac{\#\text{ virtually answered} \leq \tau}{\#\text{ offered}}$ |
| $SL_6$ | $\frac{\#\text{ sojourn in queue} \leq \tau}{\#\text{ offered}}$ |
| $SL_7$ | $\frac{\#\text{ abandoned}}{\#\text{ offered}}$ |

Table II.1.1: Service levels.

(FCFS). In addition, we let customers be impatient. As discussed earlier, we denote by $T$ the random variable measuring patience times, and we consider two different ways to model $T$. Let $\tau$ be the acceptable waiting time and $a$ be the threshold of short abandonments. In practice, reasonable values for $\tau$ and $a$ are for example 20 and 5 seconds, respectively. For some managers, customers who immediately balk or those who enter the queue and quickly abandon before $a$ are not really considered as unsatisfied. Therefore, such customers may not be accounted for in the service-level metric of the call center.

In Table II.1.1, we define seven service levels. We denoted them by $SL_i$, for $i = 1, \ldots, 7$. We present them, as is customary in call centers, in terms of the numbers of calls that arrive in a certain time period.

What should be the right metric? $SL_1$ and $SL_4$ do not give information about abandonments. $SL_5$ is hard to understand by managers and is also not directly measurable using historical data. For this reason it is, according to our experience, never used in call centers. However, this service-level definition dominates the Erlang A literature. $SL_6$ does not differentiate between waiting prior to service or to abandonment. $SL_7$ does not give information about waiting.

$SL_2$ and $SL_3$ exclude short abandonments which is a good aspect. The main drawback of these two metrics, similarly to all other metrics that use the parameter $\tau$, is that they do not give any information on how long callers that have exceeded $\tau$ still have to wait. They entice managers to give priority to callers who have not yet reached the acceptable waiting time, thereby increasing even more the waiting time of callers that have waited longer than $\tau$. Even though they have perverse effects, these metrics are regularly used in practice. One way to avoid unwanted behavior is to add an objective on the performance of the customers who wait more than $\tau$, or to use a different service-level objective. One possibility is to use the time that waiting exceeds $\tau$. In contrast with the expected waiting time (the average speed of answer) it is sensitive to waiting-time variability. Another intuitive and simple solution is to use FCFS in all cases.

Let $V_Q$ be the random variable denoting the virtual waiting time of a tagged, infinitely patient customer. In other words if the tagged customer finds a busy system upon arrival, this customer does not balk, neither abandon while waiting in the queue. Note that "answered" means $V_Q \leq T$ and "abandoned" means $V_Q > T$. Let $W_Q$ be the random variable measuring the sojourn time of a customer in the queue. This sojourn time will end either as a result of an abandonment or a start of service. Thus $W_Q = \min\{V_Q, T\}$. One may give the expressions for the service levels in Table II.1.1 as a function of the random variables $V_Q$, $W_Q$, and $T$. For example, $\mathrm{SL}_1 = \mathbb{P}(V_Q \leq \tau, V_Q \leq T)$. One may then explicitly derive all service level expressions using results from Baccelli and Hebuterne (1981) and Zeltyn and Mandelbaum (2005).

In practice, managers usually use $\mathrm{SL}_1$ which is not appropriate since we are penalized with customers who are very impatient. These customers do not really experience frustration. A better metric would be $\mathrm{SL}_2$ which ignores short abandonments. An even better metric could be $\mathrm{SL}_3$ which ignores abandonments within the acceptable waiting time. An additional benefit from using these last two metrics is shown by the numerical experiments. The required staffing levels are indeed lower than those for $\mathrm{SL}_1$.

To go further and confirm the interest of $\mathrm{SL}_2$ and $\mathrm{SL}_3$, it is worth to look on the behavior of the probability of abandonment. We observe that the performance in terms of abandonments after $\tau$ are acceptable for the metrics $\mathrm{SL}_2$ and $\mathrm{SL}_3$ (while they do need lower staffing levels). This comment is particularly relevant for large call centers, due to the benefit of pooling on performance.

## 3.3 Concluding Remarks and Future Research

We have analyzed various process-related call center metrics that include customer abandonment. We derived new results for new metrics considering short abandonments or abandonments within the acceptable waiting time. In practice, many managers choose not to count short abandonments against the call center performance metrics. Although the models used here are simple, we have shown their robustness using real call center data.

We have presented two models for customer patience that have a very good agreement with reality. The method to derive the call center metrics works for empirical patience distributions as well. The benefit of using our models is that the Markovian property is preserved. This is especially useful when one wants to consider other service-time distributions.

There are several avenues for future research. It would be useful to extend the analysis to the case of more than one customer type with non-identically distributed patience and service times. Another interesting and challenging extension of the current analysis is to consider a non-stationary arrival process.

# Chapter II.2

# Design of Call Centers

## 1 Introduction

This chapter describes my contributions to the literature on the design of call centers. The design of call centers is concerned with structural long-term changes. The strategic decisions involve the allocation of resources (equipments) as well as the layout and location of the facilities. Included in this category of decisions are those specificating how to partition customers into classes and how the different communication channels are to be used for serving the customers: for example, which types of customers are to be answered by automates, internal agents, external agents (outsourcing), etc.

There are two important aspects in the design of call centers. The first one deals with the issue of skills: should agents be cross-trained with all skills (full-flexible call centers) or should the agents only be trained for a subset of skills? In the later case, what are the subsets of skills that will be considered and how many agents will have each subset of skills? A typical example of such multi-skill call centers is an international call center where incoming calls are in different languages (Gans et al., 2003). Related studies include those by Garnett and Mandelbaum (2001), Akşin and Karaesmen (2007) and references therein. The second aspect deals with the issue of the level of pooling in call centers, i.e., are the agents all gathered into a single large team or are they partitioned into a set of independent teams? This issue is encountered in general in multi-skill call centers but in particular in full-flexible call centers, i.e., call centers in which all agents have all skills (all agents are flexible enough to answer all requirements of service).

My research results can be divided into two parts, each of which is related to one of two above families of strategic questions. My focus was on:

- The analysis of the impact of a team-based organization in call center management
- The study of flexibility in the architecture of skill-based routing call centers.

The motivation for the analysis of team-based organization started with my collaboration with the French mobile phone company Bouygues Telecom, Yves Dallery, Rabie Nait-Abdallah and Fabrice Chauvet. The company managers were interested in investigating the benefits of mi-

grating from a call center where all agents are pooled and customers are treated indifferently by any agent, towards a call center where customers are grouped into clusters with dedicated teams of agents. Each cluster is referred to as a portfolio. Customers of the same portfolio are always served by an agent of the corresponding team. The reason for moving to this organization is that dealing with teams of limited size allows a much better workforce management compared to the situation usually encountered in large call centers. Our purpose is then to examine how the benefits of moving to this new organization can outweigh its drawback (less pooling effect). The benefit comes from the better human resource management that results in a higher efficiency of the agents, both in terms of speed and in terms of the quality of the answer they provide to customers. Our analysis is supported by the use of some simple queueing models and provides some interesting insights. In particular, it appears that for some reasonable ranges of parameters, the new organization is attractive in the sense that it can outperform the original organization. The details of the results are given in Section 2.

The application of customer portfolio management had very significant effects in the Bouygues Telecom call center. The quality of answers has been improved reducing call backs by 25%. The proportion of disconnected calls (because of a full queue) was divided by 2. And no supplementary agents were hired in spite of the increase of the total number of customers by 15%. This provides an experimental confirmation of the results and insights presented in this work. Note also that we published an international patent related to our proposed new call center organization (Jouini et al., 2006).

I focus in the second part of my work on the design of call centers with multiple customer types and multiple agent skills. This is the subject of Section 3 of this chapter. In the design of SBR call centers, one of the key questions for an operations manager is to determine the appropriate type and level of flexibility. More specifically, the flexibility design problem investigates skill set design for flexible call center employees, as well as the right mix of flexible and specialized agents. This was exactly the concern of Bluelink, the call center of the airline company Air France KLM, but under the specific context of highly asymmetric parameters. With Florian Grumiller and Bané Jankovic from Bluelink, Yves Dallery and the PhD student Benjamin Legros, our purpose was to develop a novel architecture with limited flexibility for SBR call centers with asymmetric parameters: unbalanced workload, different service requirements, a predominant customer type, unbalanced abandonments and high costs of cross-training. The most well-known architectures with limited flexibility such as chaining fail against such asymmetry. We proposed a new architecture referred to as single pooling with only two skills per agent and we demonstrated its efficiency by conducting a comprehensive comparison between this 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. We also show that these observations are more apparent for situations with larger number of skills, or for those with larger call center size.

# 2 Team-Based Organization

## 2.1 Context and Motivation

The purpose of my work here is to provide some insights into the impact of internal organization of call centers on their performance. As mentioned above, it is the result of a collaboration with Bouygues Telecom. The Bouygues Telecom call center handles an average of 100,000 phone calls daily. Some of the calls are treated by an automated operator. Agents deal with about 60% of these contacts. There are also about one million contacts per year handled by mail, e-mail and fax. We want to investigate the adequacy of migrating from a call center where all agents are pooled and customers are treated indifferently, towards a call center where customers are grouped into portfolios. Managers of Bouygues Telecom believe that the challenge is not only to answer quickly but also to answer customers correctly. In the mobile telephony sector, it is not rare to see customers switching from one company to another as a consequence of low quality responses provided by agents. Agents are the interface between the company and the customers; hence, customer satisfaction is closely linked to agents performance. Managers need to motivate their employees so that the assistance they provide to customers is efficient, both in terms of speed and quality of answers. On the other hand, employees need to feel strongly supported by the company so that the turnover is as low as possible. In fact, turnover means training new employees, and it implies more costs.

The aim of Bouygues Telecom through migrating into customer portfolio management is to make agents more responsible towards their own customers. Moreover, partitioning agents into groups creates competition, which increases agents motivation. These factors result in overall agents efficiency improvement, both quantitatively and qualitatively. We argue that these advantages may outweigh the variability that results from the loss in economy of scale originally associated with the pooled system. Such a managerial approach has been widely and successfully used in industry and is also likely to be of interest in service. It is one of the key success factors of the so-called World Class Manufacturing. For example, Schonberger (1986) refers to it as cellular manufacturing and describes its benefits as follows: "Cells create responsibility centers where non existed before. The cell leader and the work group may be charged with making improvements in quality, cost, delays, etc."

## 2.2 Positioning of My Contributions

Our work is related to two streams of literature, one dealing with pooling and the other with human factors in queueing systems. While it is easy to see that pooled systems are more effective than independent ones, this intuition was for a long time based on experience and numerical data rather than rigorous mathematical proof. Smith and Whitt (1981) are the first to formally prove this result, when combining systems with identical service time distributions. Benjaafar (1995) extends theses results by providing performance bounds on the effectiveness of several pooling scenarios. When we allow service rates in separate systems to become different, combining

queues can be counterproductive (Whitt, 1999b; Tekin et al., 2009).

The above results do not account for the human element. This takes us to the second area of literature close to our work. Human element is the main characteristic of call centers and contact centers. Both customers and agents are people. Even though it is natural to focus on understanding human behavior, few papers integrate this aspect to analyze call centers and, in general, queueing systems. One of the first papers is Rothkopf and Rech (1987), which deals with the question of combining queues. The authors discuss the tradeoff between pooled and separated systems by including customer reaction and jockeying (a customer can move from one queue to another). Moreover, they show how separate systems may lead to servers that are more responsible towards their own customers. It may also allow for a faster service due to the degree of specialization gained through experience. To our knowledge, they were the first to emphasize this issue.

Fischer et al. (1999) conclude that call center management requires a mix of disciplines that are not typically found in organizations. The review of Boudreau et al. (2003) follows through this new area. They propose a framework which is a fertile source of research opportunities. They justify by real examples that operations management itself, without human resource management, can not well analyze systems such as those we are dealing with, and vice versa. In others words, there is a mutual impact between the two fields, and taking into account this fact yields to more realistic and precise insights. In particular, Boudreau et al. (2003) consider that more realistic operations management models need to integrate human factors, such as; turnover, motivation and team structure. In fact, a team setting allows for better communication, and may allow for more responsible and motivated agents. Boudreau (2004) underlines once again the significant opportunities for fruitful research at the boundaries between the traditional topics of operations management and human resource management. We address this issue in a call center context. We explore how managing agents by creating separate pools might lead the agents performing more efficiently.

## 2.3 Problem Setting

In this section we present the general problem. Consider a company operating a fairly large call center. The call center provides assistance to the customers of the company. Customers call the company whenever they need assistance and their request is addressed by a set of agents. We assume that the call center is operated in such a way that all agents have the same skill.

### 2.3.1 Current Organization Mode

The call center is operated in such a way that at any time, any call can be addressed by any agent. So, whenever a call arrives, it is addressed by one of the available agents, if any. If not, the call is placed into a queue and will be addressed as soon as possible. There is a single queue and waiting calls are answered on a first come, first served (FCFS) basis. For simplicity, we assume that the queue has no capacity constraint and that customers do not abandon while

waiting. Under this organization, the agents have a given efficiency. The quantitative efficiency is measured by the distribution of the processing times, which represents the time it takes for an agent to answer a call. Note that the randomness of the processing times comes in particular from the variety of questions asked by the customers. The qualitative efficiency is measured by the probability of successfully answering the question of the customer. We assume that if the call has not been addressed in an adequate manner, the customer will call back to get assistance from another agent. This concept of call resolution probability was argued by de Véricourt and Zhou (2005) in a call routing problem. As for the global efficiency of the call center under the current organization, its positive side comes from the pooling effect. Its negative side is in terms of human resource (HR) management, given that, it is usually very difficult to have an efficient management of a large set of agents in a large call center.

### 2.3.2 New Organization Mode

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

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

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

homogeneous would allow for performance comparisons between the different teams resulting in a "global competition". Incentives could be given to agents based on the global performance of the team.

### 2.3.3 Research Objectives

Our purpose is to study the tradeoff between the pros and cons of moving from the original organization to the team-based organization, also referred to as the portfolio organization. To do that, we consider a simple stochastic model of the original pooled organization (Figure II.2.1). This model captures the original behavior of the call center when all agents are pooled. Under this situation, the call center has a nominal behavior in terms of efficiency (quantitative and qualitative efficiencies). It achieves a given quality of service (QoS). We actually consider two different QoS measures: the average waiting time and the 80/20 rule, which is an industry standard for telephone service (Gans et al., 2003). Under the 80/20 rule, at least 80% of customers must wait no longer than 20 sec.

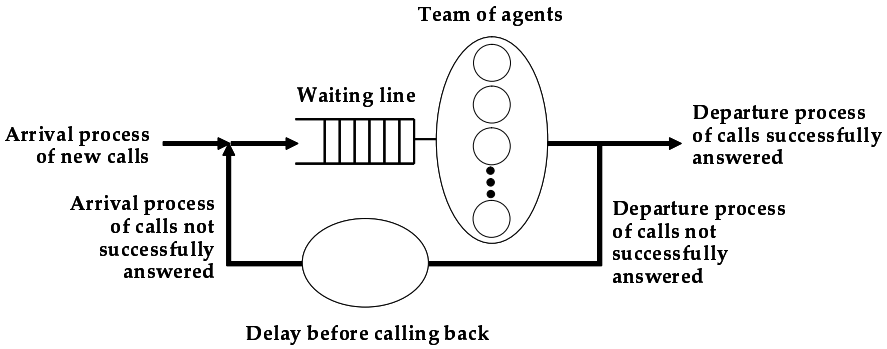



Figure II.2.1: The generic model

## 2.4 Analysis of the Efficiency of the Team-Based Organization

We use simple queueing systems and determine the performance measures of interest. The original call center model is referred to as the Pooled System. The team-based organization is referred to as the Dedicated System. They are shown in Figures II.2.2 and II.2.3, respectively.

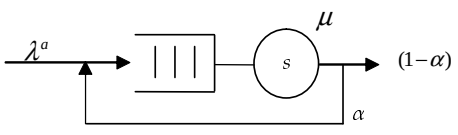



Figure II.2.2: Pooled System

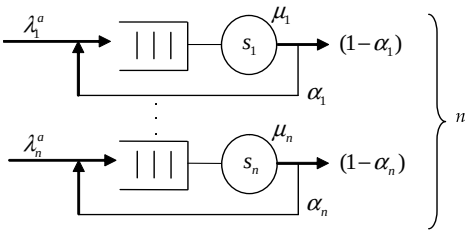
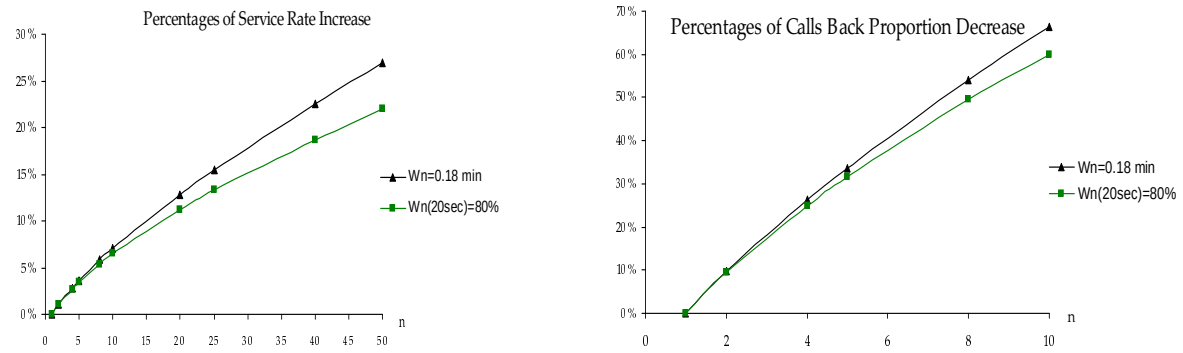



Figure II.2.3: Dedicated System

We start from a Pooled System with a given QoS in terms of the percentage of waiting less than a given threshold $W(t)$, or the expected waiting time $W$. The purpose is to evaluate the required service rate in a Dedicated System with $n$ pools in order to ensure the same QoS ($W_n(t) = W(t)$ or $W_n = W$). The total staffing level, the total arrival rate of first-attempt calls, and the call back proportion are all held constant. Numerical experiments are shown in Figure II.2.4(b). We do the same for the required decrease of the call back proportion. The results are shown in Figure II.2.4. The results show that it is possible to even up the performances of a Pooled System by slightly increasing the service rate, or by slightly decreasing the call back proportion.




(a) Service rate

(b) call back proportion

Figure II.2.4: Experiments

We have performed a more systematic analysis to confirm the robustness of our conclusions. The results show that migrating towards separated call centers may not be as bad an idea as it seems. In addition, it would be realistic to assume that the better team management enabled by the new organization implies an improvement of both parameters. The results show that by having an improvement on both efficiencies, the required performance improvement on each one is not as high as when focusing on them separately. For instance, when we migrate to a Dedicated System with $n = 10$, we need to increase $\mu_n$ by 3% and decrease $\alpha_n$ by about 37%. In such a case, it should come as no surprise that we improve the performance in the dedicated systems rather than deteriorate them. Team management effects may change both parameters and may go beyond the simple fact of outweighing the increase of variability.

Another advantage of the team-based organization is its robustness with respect to errors in the estimation of the arrival rate of primary calls. We observe that the QoS of the Pooled System is much more affected than the one of the Dedicated System by an underestimation of the first-attempt calls arrival rate. Let us give an explanation. Under the original expected first-attempt arrival rate, the server utilization in the Pooled System is closer to 1 than that in the Dedicated System. If the first-attempt calls arrival rate is underestimated, the deterioration of the quality of service is increasing faster when the server utilization is closer to 1, since the queue becomes less and less stable. This is an attractive feature that gives another strong argument in favor of the team-based organization.

# 3 Flexible Architecture

## 3.1 Background and Research Objectives

The concept of flexibility is related to the ability of a company to efficiently match its capacity to an uncertain demand with multiple types. 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 tradeoff between performance and cost.

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. On the one hand, FF would require less agents than any other architecture, in order to reach a given predefined service level. On the other hand, 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! Full flexibility and full dedication, however, are only two extreme situations. A well-known intermediate configuration is *chaining*, first pointed out by Jordan and Graves (1995). Under chaining, 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. 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 (two skills per agent such as in chaining) has the potential to achieve the performance of total flexibility.

Developing intelligent configurations such as chaining is very interesting for practitioners. They allow to 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 on one hand to account for demand asymmetry, and on the other hand to capture the benefits of pooling with only a limited flexibility. This is our purpose in this work.

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 the call center of Bluelink, each agent speaks two languages: her own native language and English. Note that this call center is more interested in agents speaking two languages rather 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 post-sales service call centers of major retailers that are, at the same time, 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 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.

### 3.2 My Main Findings

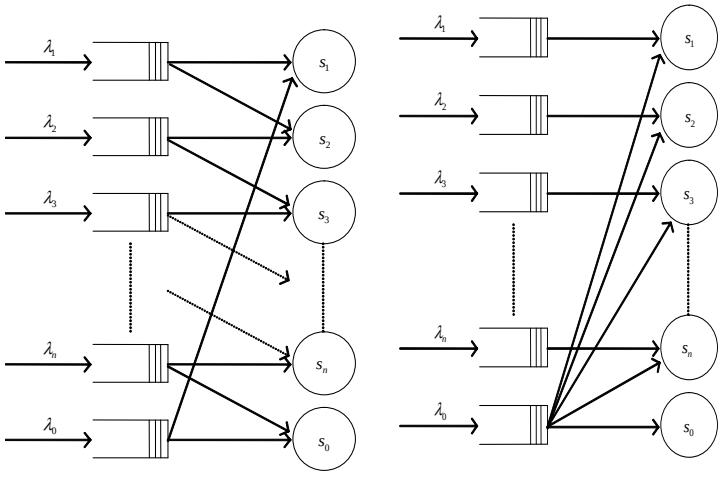
Motivated by the prevalence of the flexibility concept in practice, we propose a new organizational model, referred to as *single pooling* (SP), where we dedicate a team of agents to each difficult type of calls, and the easy type of calls have access to all agents from all teams. Balancing the workload among the agents in this way captures the benefits of pooling without requiring every agent to process every call type. We do not claim that our model is better than chaining in all cases, but only in the particular situations of 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.

Using simulation, we conduct a comprehensive comparison between single pooling and chaining. As a function of the various system parameters, we delimit the regions where either chaining or single pooling is the best. Few of 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 and abandonment times, 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.

### 3.3 Modeling

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 $\lambda_i$, $\mu_i$ and $\gamma_i$, 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. 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 focus on the performance in terms of the steady-state 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 $i = 0, 1, ..., n$. The two models that we compare in this work are chaining and single pooling. They are shown in Figures II.2.5(a) and II.2.5(b), respectively.




(a) Chaining (b) Single pooling

Figure II.2.5: Call center configurations

### 3.4 Approximate Numerical Comparison

We numerically compute approximate expected waiting times for single pooling and chaining. For tractability, we consider Markovian assumptions for inter-arrival and services times, and customer abandonment is ignored. The objective of this analysis is to obtain some sense on the effect of the parameters asymmetry on the comparison between the two architectures. A more comprehensive analysis is thereafter conducted using simulation. We employ a Markov chain method for the performance analysis of each design.

For single pooling, we first compute the steady-state system probabilities, from which we deduce the expected waiting time for each customer type. We use a truncation point in the Markov chain for the numerical computation. Because of the routing mechanism in chaining, a standard Markov chain modeling is not appropriate. Once an agent completes a service, she chooses next to service the oldest customer among those in the head of two queues, if any. A standard modeling only based on the number of customers in the queues can not take this decision into account. We thus propose to discretize the waiting time of the first in line in each queue instead of using the number of agents in each queue. The modeling of the first in line as a tool for analyzing a queueing system was proposed by Koole et al. (2012). We again use a

truncation point for the computation.

**A Real-Life Numerical Illustration:** The real example consists of an airline company call center, located in Australia and handling 4 types of customers: Japanese (type 1), Korean (type 2), Bahasa (type 3) and English (type 0) speaking customers. Customer types are identical in their requests (flight booking and modification, claims, etc.). The expected service time is the same for all types, $\frac{1}{\mu_i}$=6.8 minutes for $i = 0, ..., 3$. An example of the daily arrival rates is given in Figure II.2.6. For the numerical illustration, we consider a given time interval with the parameters $\lambda_0 = 4.6$, $\lambda_1 = 7.7$, $\lambda_2 = 10.1$ and $\lambda_3 = 1.5$. Note that we ignore here several features such as abandonment, retrial, rejection, agent reservation routing rules, back-office tasks, etc.

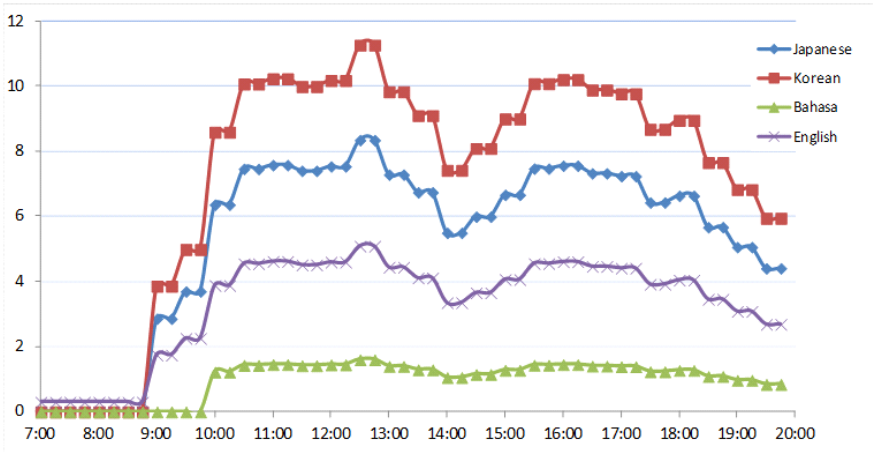



Figure II.2.6: Customer arrival rates

This call center uses the SP architecture, where an agent from a given team has skill 0 and skill $i$, for $i = 0, 1, 2, 3$. Let us compare the costs of using SP and chaining. We know from this call center that the salary per hour of an agent with the easy skill (English) and 1 regular skill (one of the other languages), is 20% higher than that of an agent with only the easy skill. Also, the salary of an agent with the easy skill and 2 regular skills, is 16% higher than that of an agent with the easy skill and 1 regular skill. We then consider that the salary of an agent in SP is 1.2 and that in chaining is either 1.2 or 1.4 according to her set of skills. Under a service level constraint ($W_i^* = 0.2$ for $i = 0, ..., 3$), the total staffing costs are 230.2 and 210 for chaining and SP, respectively. SP behaves better in this example because of the asymmetry in the arrival rates and also the agents salary structure.

### 3.5 Synthesis on the Effect of Parameter Asymmetry

In order to obtain a comprehensive understanding of the comparison, we resort to simulation. As we are interested in the effect of asymmetry of the parameters on performance, we propose various forms of asymmetry. For customers 0, we measure the relative importance in arrivals and service durations. We measure the asymmetry between the arrival rates of regular customers,

and that between service durations. We also consider for customers 0 the asymmetry in the variability of service times, measured by the coefficient of variation of its distribution. We define, in addition, 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 assess 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. Finally note that in order to have a coherent comparison, we optimize for each model the 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.

The numerical analysis shows that single pooling performs better than chaining for various cases of asymmetry. 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.

One may summarize the recommendations and guidelines to call center managers as follows. The manager choice of a flexible call center design should be single pooling under situations of asymmetry in arrival and service rates. This holds even for small differences between the skill costs. This choice more apparently prevails for large call centers and/or in the case of a high number of skills. However, the choice of the design is highly impacted in the context of call centers with customer abandonment. Abandonments may affect the system by either increasing or decreasing the asymmetry of the parameters. In the first case, the preference remains for single pooling, while it is for chaining in the second case.

# 4 Concluding Remarks and Future Research

We focused on fundamental problems in the design and management of call centers. In a first part, we argued how team management benefits, that come from the portfolio/team one-to-one

link, may outweigh the economy of scale associated with the pooled organization. We studied partitioning of a large call center into identical and separated call centers, where agents of a same team are dedicated to one portfolio of customers. We showed that the costs of migrating towards separated systems are not as important as it may appear. In practice, combining the benefits of the team-based organization in terms of both improved service rate efficiency and reduced call back proportion can easily outweigh the loss of the economy of scale. In a future study, we will extend our models by considering abandonments and limited waiting lines. We will also try to improve the approximation models discussed here to get more accurate analyzes.

In a second part, we considered the context of SBR 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. We proposed the 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. In a future research, it would be useful to extend the numerical approximations, of the performance of SP and chaining, in the case of customer abandonment or non-Markovian assumptions. 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 the easy skill.

# Chapter II.3

# Personnel Planning in Call Centers

## 1 Context and Contributions

This chapter describes my contributions to the literature on call center planning. These have been done within the PhD thesis of Shuangqing Liao, and under the collaboration with my colleagues Ger Koole and Christian van Delft. We focus on a call center staffing problem of a given working day where we take into account the feature of uncertainty in the arrival parameters.

The staffing cost is a major component in the operating costs of call centers. Unfortunately, uncertainty plaguing the arrival process and the corresponding workloads usually leads to a complex staffing problem. Traditionally, most call center models in the literature assume known and constant mean arrival rates, mainly for tractability issues. However, in addition to the usual uncertainty captured by a stochastic process modeling, real data show another uncertainty in the process parameters themselves. In this work, we consider the staffing problem of a single shift call center, in which we allow the mean arrival rate of calls to be uncertain. We model the arrival process of calls by a doubly non-stationary stochastic process, with random mean arrival rates. As in the traditional way, a service level constraint limits the waiting time for inbound calls. In addition to the job of calls, our call center has to process back-office jobs, such as answering emails. These additional jobs are assumed to be given at the beginning of the day and have to be processed within the same day, if necessary in overtime. We also allow the workload of back-office jobs to be random. The possibility of delaying back-office jobs introduces some flexibility to the daily workforce management. A typical example of our call center is that of a hospital, or of a government or of a public agency, where inbound calls and back-office operations are handled by agents in a single shift (during administrative hours). The agents can be, in real-time, affected to one job type or another depending on the actual workload and the operating costs.

We model the staffing problem as a cost optimization-based newsboy-type model. The cost criterion function includes the regular and overtime salary cost and a penalty cost for excessive waiting times for inbound calls. Our objective is to find the optimal staffing level which minimizes the total call center operating cost. We consider a multi-period single-shift call center staffing

problem, with the constant staffing level as the single decision variable. We propose two solution methodologies. First, we formulate the problem as a stochastic program, by a discretization of the underlying probability distributions. The second approach relies on the robust optimization theory. We prove a convexity result of the problem, which allows us to find the optimal solution via a relaxed real-valued optimization model. We then conduct a numerical study in order to illustrate the main characteristics of the two approaches and the associated optimal solutions. In the numerical illustration, we use real data gathered from a call center of a Dutch hospital that handles inbound calls and emails.

## 1.1 Positioning of My Contributions

It has become apparent that general queueing systems performance indicators are very sensitive to the fluctuations of the parameters characterizing the arrival process over time (Ingolfsson et al., 2007). As a consequence, a stream of research has begun to address the problem of how call centers can better manage the capacity-demand mismatch that results from arrival rate uncertainty.

First, the pure statistical forecasting issue has been considered in several papers analyzing the probability distribution of arrival rates (Avramidis et al., 2004; Brown et al., 2002, 2005; Weinberg et al., 2007; Shen and Huang, 2008; Aldor-Noiman et al., 2009). Various call center particularities have been pointed out in these studies. As a second step, the analysis of performance measures of queueing systems with fluctuating arrival rates has appeared. The first setting concerns deterministic non-stationarity, i.e., some parameters evolve along time according to a known dynamics. A direct method of accommodating such time-varying parameters consists of numerically solving the complex queueing models associated to the transient system behavior (Ingolfsson et al., 2007; Yoo, 1996). Another intuitive means of accommodating changes in the arrival rate is to consider piecewise stationary measures over successive intervals, while reducing the time length of the intervals over which such stationary measures could be applied. This is the essence of the point-wise stationary approximation (PSA) used in Green et al. (2007). In a different setting, a few papers have considered the issue of random non-stationarity in the arrival process parameters. In Jongbloed and Koole (2001), the authors include arrival parameter uncertainty via a Poisson mixture model for the arrival process, which allows to model the overdispersion associated with random arrival rates. Most of existing methods assume independent intervals, which would lead to inaccurate results particularly in case of systems that are overloaded during a certain number of periods.

The last issue concerns the call center staffing optimization problem under non-stationary parameters. Some models rely on a fixed staffing level methodology: there is no possible flexibility during a daily period and the staffing cannot be updated throughout the day. In Harrison and Zeevi (2005); Whitt (2006), this problem is solved via a static stochastic program using a stochastic fluid model approximation. In Jongbloed and Koole (2001), the standard Erlang formula-type for a fixed staffing approach is generalized through a new Poisson mixture model

for the arrival process. Robbins and Harrison (2010) introduce uncertainty for parameters via a discretization of the underlying parameters probability distribution. The approach has also been applied in the case of a call center with multiple call types in order to investigate the flexibility introduced by adding a proportion of cross trained workforce (Robbins et al., 2007; Robbins et al.). The optimal staffing problems in Gurvich et al. (2010) are solved by a chance-constrained programming approach.

## 1.2 Problem Formulation

We consider a multi-period single-shift call center staffing problem. The call center handles various types of jobs: inbound calls as well as some alternative back-office jobs. The mean arrival rate of inbound calls is allowed to be uncertain. The workload of the back-office jobs is also uncertain. The inbound calls have to be handled as soon as possible, while the back-office jobs, such as emails, can be delayed to some extent within the same day. In this section, we describe the corresponding stochastic minimal cost staffing problem.

**Inbound Calls:** We model the inbound call arrival process by a doubly stochastic Poisson process as follows. We assume that a given working day is divided into $n$ distinct, equal periods of length $T$, so that the overall horizon is of length $nT$. The period length in practice is often 15 or 30 minutes. The mean arrival rate of calls during period $i$ is denoted by $\Lambda_i$ and is random. Using the same modeling as in Avramidis et al. (2004) and in Whitt (1999a), we assume that the arrival rate $\Lambda_i$ is of the form $\Lambda_i = \Theta f_i$, for $i = 1, ..., n$, where $\Theta$ is a positive real-valued random variable. The random variable $\Theta$ can be interpreted as the unpredictable "busyness" of a day. A large (small) outcome of $\Theta$ corresponds to a busy (not busy) day. The constants $f_i$ model the shape of the variation of the arrival rate intensity across the periods of the day. Formally, if a sample value in a given day of the random variable $\Theta$ is denoted by $\theta$, the corresponding outcome of the arrival rate over period $i$ for that day is defined by $\lambda_i = \theta f_i$. The random variable $\Theta$ is assumed to follow a discrete probability distribution, defined by the sequence of outcomes $\theta_l$ and the associated sequence of probabilities $p_{\theta_l}$, with $l = 1, ..., L$. Finally, we assume that service times for inbound calls are independent and exponentially distributed with rate $\mu$. The calls arrive to a single infinite queue working under FCFS.

For period $i$, let the random variable $WT_i$ denote the waiting time of an arbitrary call. The probability distribution of the waiting time of calls, $Pr\{WT_i \leq AWT \,|\, \theta\}(v) = F_{\theta\, i}(v)$, is computed using the classical Erlang C results.

**Back-Office:** We assume that the random back-office workload arrives at the beginning of the day. As an example, one can think of a call center that stores all the emails of a given day and handles them the next day. We denote by $W$ the number of agents required to handle this back-office workload during a single period. The random variable $W$ is characterized by a discrete probability distribution, defined by the sequence of outcomes $w_k$ and the associated sequence of probabilities $p_{w_k}$, with $k = 1, ..., K$.

### 1.2.1 Cost Criterion

We consider a single-shift call center. Let us denote by $y$ the number of agents staffed for the day. All the $y$ agents will be therefore present all day long. We also assume that all agents are able to handle both types of jobs, calls and back-office jobs. We give priority to inbound calls as follows. For each period $i$, if the actual number of agents $y$ is larger than $v_i(\theta f_i)$ (the required number of agents to handle the calls), we assign $v_i(\theta f_i)$ agents to calls and $y - v_i(\theta f_i)$ agents to back-office jobs. If $y < v_i(\theta f_i)$, all the $y$ agents are assigned to calls. If back-office jobs are not yet finished at the end of the regular working periods in that day, they are done in overtime.

We define a risk level $\alpha$ is expressed via an associated under-staffing penalty cost denoted as $u_\alpha$. More concretely for each period $i$, a proportional under-staffing penalty $u_\alpha$ is paid when the actual capacity $y$ is lower than a sample value of the required agents number $v_i(\theta f_i)$. We assume that each agent gets a salary $c$ per period, the overtime salary is $r$ per agent per period. As usual, the cost parameters satisfy the ordering $c < r < u_\alpha$ for all possible values of $\alpha$. The inequality $r < u_\alpha$ ensures that inbound calls have the priority over back-office jobs. The inequality $c < r$ is straightforward.

Since the time-horizon of the considered problematic is significant, the cost criterion of the formulation is the expected daily total cost associated with the staffing level $y$, which is expressed as

$$C(y) = E\Big[C(y,\theta,w)\Big] = \sum_{l=1}^{L}\sum_{k=1}^{K} p_{\theta_l}\, p_{w_k}\, C(y,\theta_l,w_k), \tag{II.3.1}$$

with

$$C(y,\theta,w) = n\,c\,y + u_\alpha \sum_{i=1}^{n}(y - v_i(\theta f_i))^- + r\left[w - \sum_{i=1}^{n}(y - v_i(\theta f_i))^+\right]^+, \tag{II.3.2}$$

where $x^+ = max(0,x)$ and $x^- = max(0,-x)$ for $x \in \mathbb{R}$. In Equation (II.3.2), the first term is the salary of the agents working during regular time. The second term is the under-staffing penalty cost. The third is the overtime salary.

Under this economic framework, our objective consists of deciding on the optimal value of $y$ which minimizes the expected daily total cost given by Equation (II.3.1). We prove that the expected daily total cost function $C(y)$ is convex in $y$. We then deduce an important property of the problem: the integer optimal solution is indeed in the neighborhood of the real-valued relaxed optimal solution.

## 1.3 Solution Methodologies

We develop two different approaches to solve the staffing problem, according to the availability of the probability distributions of the random variables. First, under the assumption that the probability distributions associated with the random variables are known exactly, a direct *stochastic programming approach* is applied to Equation (II.3.1), built on the discrete probability distributions characterizing $\Theta$ and $W$. The second approach referred to as *robust programming*

consists of optimizing the staffing level with respect to (w.r.t) the worst case scenarios in a given uncertainty set.

### 1.3.1 Stochastic Programming Approach

Assuming that we know the exact probability distributions associated with the random variables $\Theta$ and $W$, a common approach consists of expressing Equation (II.3.1) as a linear program via the discrete probability distributions associated with these random variables. For each sample $\theta_l$ of $\Theta$, we use the associated sample arrival rate in each period $i$, $\lambda_{i,l} = \theta_l\, f_i$. The required number of agents is $v_i(\lambda_{i,l})$. The optimization problem from Equation (II.3.1) can be then formulated by the following linear program:

$$Min \quad nc\, y + u_\alpha \sum_{l=1}^{L}\sum_{i=1}^{n} p_{\theta_l}\, M_{i,l}^{-} + r \sum_{k=1}^{K}\sum_{l=1}^{L} p_{\theta_l}\, p_{w_k}\, N_{k,l} \tag{II.3.3}$$

$$\text{s.t.} \quad M_{i,l} = y - v_i(\theta_l\, f_i), \qquad \text{with} \;\; i = 1,...,n, l = 1,...,L, \tag{II.3.4}$$

$$M_{i,l} = M_{i,l}^{+} - M_{i,l}^{-}, \qquad \text{with} \;\; i = 1,...,n, l = 1,...,L, \tag{II.3.5}$$

$$N_{k,l} \geq w_k - \sum_{i=1}^{n} M_{i,l}^{+}, \qquad \text{with} \;\; l = 1,...,L, k = 1,...,K, \tag{II.3.6}$$

$$y, M_{i,l}^{+}, M_{i,l}^{-}, N_{k,l} \geq 0, \qquad \text{with} \;\; i = 1,...,n, l = 1,...,L, k = 1,...,K. \tag{II.3.7}$$

In this problem $M_{i,l}$ represents the difference between the staffing level and the required agent number in period $i$ for scenario $l$. The positive and negative part of $M_{i,l}$ are denoted by $M_{i,l}^{+}$ and $M_{i,l}^{-}$, respectively. $M_{i,l}^{-}$ is associated to under-staffing cost in the objective function. $N_{k,l}$ is the over-time workload required in order to finish back-office jobs in scenario $(k,l)$. This overtime induces overtime cost in the objective function. The unique decision variable in our staffing problem is the staffing level $y$.

### 1.3.2 Robust Programming Approach

Robust programming based formulations are often computationally tractable even for large-scale problems and do not require a probabilistic description of the uncertain parameters.

A main issue of the robust programming implementation is the design of an efficient uncertainty set which fixes the tradeoff between robustness (i.e., protection against the worst case) and average performance. We consider a robust approach associated with uncertainty sets for $\Theta$ and $W$. In order to analyze the above robust formulation, we first study the properties of the optimal value, denoted as $C^*(\theta, w)$, of the purely deterministic optimization problem for given

outcomes $\theta$ and $w$,

$$Min \quad nc\,y + u_\alpha \sum_{i=1}^{n} M_i^- + r\,N \tag{II.3.8}$$

$$\text{s.t.} \quad M_i = y - v_i(\theta\, f_i), \qquad \text{with} \;\; i = 1, ..., n, \tag{II.3.9}$$

$$M_i = M_i^+ - M_i^-, \qquad \text{with} \;\; i = 1, ..., n, \tag{II.3.10}$$

$$N \geq w - \sum_{i=1}^{n} M_i^+, \tag{II.3.11}$$

$$y, M_i^+, M_i^-, N \geq 0, \qquad \text{with} \;\; i = 1, ..., n. \tag{II.3.12}$$

In this formulation, $M_i$ represents the difference between the staffing level and the required agent number in period $i$. The positive and negative part of $M_i$ are denoted by $M_i^+$ and $M_i^-$, respectively. $M_i^-$ is associated to under-staffing cost in the objective function. $N$ is the over-time workload required in order to finish back-office jobs.

### 1.4 Insights

In what follows, we comment on the numerical results and summarize the main insights. Some tradeoff exists between the average cost and the associated standard deviation: Above the threshold which is the optimal staffing level of SP, the average total cost increases while the associated standard deviation decreases in $y$. It is also obvious to see that the under-staffing probability decreases in the under-staffing penalty $u_\alpha$. For large values of $u_\alpha$, this probability becomes negligible. Concerning the average cost, SP is as expected the most efficient. This stems from the fact that for a call center with given distributions of the "busyness factor" $\Theta$ and the back-office workload $W$, we associate an under-staffing penalty cost $u_\alpha$. The gap between the optimal staffing levels of the deterministic and stochastic approaches is significant, particularly when the back-office workload is small. The deterministic approach neither captures the negative impact of the randomness in arrival rates on service quality, nor on the under-staffing cost.

An obvious benefit from adding back-office jobs comes from the fluctuating shape exhibited by the call arrival rate as a function of the periods of the days. Since we are considering a single shift call center, the strongest quality-of-service constraints (corresponding to the period with the highest arrival rates), tend to force to have a typically high staffing level for the whole day. Such a level is in fact required for only some periods. Clearly, this situation leads to over-staffing during the other periods, which can be used without any additional cost in order to handle some back-office jobs. Also, the variability of the call arrival process can be smoothed by increasing back-office workload.

## 2 Concluding Remarks and Future Research

We have developed a single shift call center model with two types of jobs: inbound calls and back-office jobs. We focused on optimizing the staffing level with respect to the total operating

cost of the call center. We modeled this problem as a cost optimization-based newsboy-type model. We then proposed various approaches to numerically solve it. We underline the necessity of taking into account the uncertainty in the call demand parameters, which is not often the case in the majority of existing studies. We also highlighted the pros and cons of the various solutions approaches. Finally, we showed to what extent the flexibility associated with storable back-office jobs helps in absorbing uncertainty in the call process.

In a future research, we intend to extend the analysis of this work to a multi-shift setting, with the possibility of removing or adding agents within the same day. Another interesting extension would be to consider a global service level constraint for the whole day, instead of having a period by period constraints.

# Chapter II.4

# Operational Issues: Call Centers with Delay Information

## 1 Introduction

This chapter synthesizes my contributions to the analysis of call centers with delay information. The interest in prediction and announcement of delays in service systems has intensified as the call center industry has grown and become technologically sophisticated. Managers have several objectives in providing such information; modulating demand by signaling times of high congestion, enhancing satisfaction with inevitable waiting, or both.

Information about anticipated delays is especially important for call centers, because the queues are invisible (Zohar et al., 2002; Bitran et al., 2008). In such systems, the uncertainty involved in waiting is high. Upon arrival and during their wait, customers have no means to estimate queue lengths or progress rate. "Uncertain waits are perceived to be longer than known, finite waits" (Maister (1985) p. 118) and have been related with lower satisfaction (Taylor, 1994). Providing delay information is shown to improve satisfaction (Taylor, 1994; Katz et al., 1991; Hui and Zhou, 1996).

These practical objectives bring with them several challenges, which have motivated my research on the subject. The challenges can be summarized as (i) estimating real-time delays for each customer in a stochastic environment; (ii) deciding on what to announce given customer preferences regarding waiting times and announcements made; and (iii) exploring customer reactions to announcements. In the first part of my work, I have addressed these three issues for a single class setting. In the second part, I have addressed only the first two issues, but under the more general multi-class setting.

The first part of my work is described in Section 2. This work was done with my colleagues Yves Dallery and Zeynep Aksin from Koç University, and has been supported by the Scientific & Technological Research Council of Turkey. We analyze a call center with impatient customers. We study how informing customers about their anticipated delays affects performance. Customers react by balking (immediately leave the system upon arrival) upon hearing the delay

announcement, and may subsequently abandon (leave the system while waiting in the queue), particularly if the realized waiting time exceeds the delay that has originally been announced to them. The balking and abandonment in such a system are functions of the delay announcement. Modeling the call center as an M/M/s+M queue with endogenized customer reactions to announcements, we analytically characterize the performance measures. The analysis allows us to explore the role announcing different percentiles of the waiting time distribution, i.e., *announcement coverage*, plays on subsequent performance in terms of balking and abandonment. We show how managers of a call center with delay announcements can control the tradeoff between balking and abandonment, through their choice of announcements to be made.

The second part of my work deals with a more general setting. It is described in Section 3. The motivation comes from Bouygues Telecom. For their multi-site system, the managers want to develop real-time delay estimation to use them in routing calls to the various sites. This is a joint work with Yves Dallery; Rabie Nait-Abdallah and Fabrice Chauvet (Bouygues Telecom); Mohamed Salah Aguir (ESTI, Tunisia); and Zeynep Aksin and Fikri Karaesmen (Koç University). We consider the problem of estimating delays experienced by customers with different priorities, and the determination of the appropriate delay announcement to these customers, in a multi-class call center with time varying parameters, abandonments and retrials. The system is approximately modeled as an $M(t)/M/s(t)$ queue with priorities, thus ignoring some of the real features like abandonments and retrials. Delay estimators are proposed and tested in a series of simulation experiments. Making use of actual state dependent waiting time data from Bouygues Telecom, the delay announcements from the estimated delay distributions that minimize a newsvendor-like cost function are considered. The performance of these announcements are also compared to announcing the mean delay. We find that an Erlang distribution based estimator performs well for a range of different under-announcement penalty to over-announcement penalty ratios.

# 2 Single Class Setting: Endogenized Customer Reaction

## 2.1 Context and Motivation

In call center settings, satisfaction with waiting experiences affects customers' reactions in terms of balking and abandonment behavior. Delay announcements, through their effects on customers, may further modulate these customer reactions (Katz et al., 1991; Hui and Tse, 1996; Taylor, 1994; Hui and Zhou, 1996; Munichor and Rafaeli, 2007). When we inform a customer about her anticipated delay, she will decide right away, either to hang up immediately because she estimates that her delay is too long, or to start waiting in the queue. For customers who enter the queue, delay announcements, by reducing the uncertainty, may further have the effect of increasing patience. However, since perfect announcements are not possible in reality, some customers may experience longer delays than what has been announced to them. Customers may abandon even if they had chosen to start waiting. It is natural to expect that such customers would abandon

in a different way than in a setting without announcements (Feigin, 2005).

It is possible to provide different types of information regarding delays. A common one is to announce the number of customers ahead. This is not very meaningful in a setting where the number of servers are unknown to customers, service times are random and where customers ahead may abandon the queue. A further possibility is to announce some real-time delay estimators based on recent delay experience by customers, as in Armony et al. (2009) or Ibrahim and Whitt (2009). For settings where the state-dependent waiting time of each new arrival can be derived, it is possible to give the whole distribution of the anticipated waiting time to each new customer or to communicate the expected value of the delay distribution, as in Whitt (1999c). In the type of announcement considered in this work, the call center manager specifies a unique tail probability for everybody, say $1 - \beta$. The parameter $\beta$ is a coverage probability based on which we determine a time $x$ from the waiting time distribution of a newly arriving customer. The actual waiting time of this customer will be less than that $x$ with probability $\beta$. Typically, the announcement will contain information on $x$ as in"You will wait $x$ minutes".

This type of delay announcement allows the service provider to control the desired reliability of the announcement by choosing $\beta$. In making that choice, the manager considers the following tradeoff. Informing the customer of short waiting times, which is likely to underestimate the actual waiting, might lead to less balking but excessive abandonment and reduce the reliability of the service provider in the eyes of the customers. On the other hand, informing the customer of large waiting times increases the number of balking customers, but as a result leading to a system that might allow to serve customers within shorter and reasonable delays. Through a numerical analysis, we investigate how the ideal percentile should be chosen.

## 2.2 Positioning of My Contributions

Although the modeling approaches differ from one work to another, the findings usually confirm the benefits of communicating delays to customers. Guo and Zipkin (2007) is one exception, where conditions are identified under which more information may hurt the customer or the service provider. Allon et al. (2011) show the possibility of a strategic self-interested firm that might choose to provide *intentionally vague* information to strategic customers to induce desired behavior from them. Our modeling approach, allowing us to capture the link between the announced value and resulting system performance explicitly, enables us to illustrate that the service provider could choose delay announcements to induce particular balking and abandonment reactions by customers. The idea of controlling the tradeoff between balking and abandonment through appropriate manipulation of customer reactions, resembles the idea of selecting the size of a finite waiting space in a queueing system to control this tradeoff (Kolesar, 1984; Gans et al., 2003).

The literature on customers influenced by delay begins with Naor (1969). Focusing on customer psychology in waiting situations Maister (1985) proposes a set of hypotheses some of which are tested in the subsequent literature. We rely on the recent review by Bitran et al. (2008)

and references therein to highlight results from the literature on which we draw in formulating our model. It is apparent from this review that modeling waiting experiences and generalizing customer reactions are difficult to do due to the presence of many moderating effects including personal differences and service context. Call center specific evidence is very limited.

One of the first models is by Osuna (1985) constructing a direct relationship between waiting time and dissatisfaction or stress experienced by customers during the wait. Elsewhere it is argued that it is not time but the perception thereof that drives customer satisfaction (Hornik, 1984; Zakay, 1989). It is shown that in settings where customers lack information on the duration of the wait, people tend to overestimate waiting durations (Taylor, 1994). This implies higher dissatisfaction with the wait. Information on delays on the other hand shorten perceived waiting time by customers (Katz et al., 1991; Hui and Tse, 1996). Many have attributed this to the sense of control that the customer feels about the wait once uncertainty about it has been removed through an announcement (Taylor, 1994; Katz et al., 1991; Hui and Tse, 1996). Lab experiments simulating a tele-queue in Munichor and Rafaeli (2007) show that customers prefer queue position announcements over music or apologies, and react by being more persistent in holding the line, providing evidence to the idea that progress matters to customers. While a delay announcement can act like a time guarantee thereby increasing satisfaction as the wait proceeds to the announced time, exceeding this time will have a negative effect on customers, reducing their satisfaction (Katz et al., 1991; Kumar et al., 1997).

## 2.3 Modeling

### 2.3.1 Model 1: No Delay Announcement

Among the customers who find all agents busy, a proportion $\alpha_0$ immediately balks. For the remaining customers, there is a proportion $\alpha_1$ who choose to immediately balk too. We attribute this proportion to uncertainty or ambiguity averse customers. Indirect evidence for this type of balking can be found in Pazgal and Radas (2008), where it is observed that customer balking increases in settings with no information on waiting, due to the higher variability in waiting times estimated by experiment participants in such settings. In Model 1, shown in Figure II.4.1, there is no time difference between the two balking decisions. A customer who finds all servers busy will balk with probability $\alpha_0 + (1 - \alpha_0)\alpha_1$. If she joins, she is willing to wait in queue only a certain amount of time. If service has not begun by this time she will abandon and be lost. We assume that her patience threshold is a realization derived from an exponentially distributed random variable with rate $\gamma''$. Model 1 can be viewed as an M/M/s+M queueing system with balking. abandonment makes the system unconditionally ergodic for any $\gamma'' > 0$.

We derive the performance measures related to Model 1, by first computing the steady-state probabilities of the number of customers present in the system, denoted by $p(i)$, for $i \geq 0$. Making use of the PASTA property (Poisson Arrivals See Time Averages), we then compute the probability of immediate service for a newly arriving customer, defined as $P^I$; the expected number of customers waiting in queue, denoted by $L_q$; the probability of a new arrival to balk,

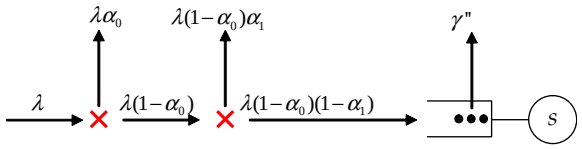



Figure II.4.1: Model without delay announcement, Model 1

denoted by $P^B$; to abandon, denoted by $P^R$; and to enter service, denoted by $P^S$. Using techniques applied in the computation of first passage times in birth-death processes, we also compute the conditional waiting times, given abandonment and given service.

### 2.3.2 Model 2: With Delay Announcement

Upon arrival, if less than $s$ customers are in the system, the new customer gets service immediately. If all agents are busy, we assume that each new arrival has a probability $\alpha_0$ to immediately balk, before even hearing his anticipated delay. The proportion $\alpha_0$ is identical to that in Model 1 since these customers do not experience any difference between these systems.

Contrary to Model 1, we believe that balking stemming from waiting uncertainty (no information) has no reason to be present here, i.e., there is no $\alpha_1$ parameter. Consider a customer, who finds all servers busy and $n$ ($n \geq 0$) waiting customers ahead of her in queue. With probability $1-\alpha_0$, she will accept to hear the information provided to her about her anticipated delay. If she does so, we derive the distribution of her virtual delay which we denote by $D_n$. It is the conditional waiting time of a customer that will wait until service begins, given the queue state $n$. Then, we communicate to her the delay which corresponds to a given coverage probability $\beta$. Define $T$ as the random variable measuring the initial random patience threshold of customers, (with probability $1-\alpha_0$: an exponential distribution with rate $\gamma$). Let $d_n$ be the delay we communicate to our customer. It means that the queueing delay of the new customer does not exceed $d_n$ with a chance $\beta$. The customer balks if her random patience threshold, $T$, does not exceed her anticipated delay $d_n$. Let $p^B(n)$ denote the probability of this event. Assuming that balking decisions of successive customers are independent leads to $p^B(n) = P(T < d_n) = 1 - e^{-\gamma d_n}$, for $n \geq 0$. So, given that a customer does not balk with probability $\alpha_0$, she may balk with probability $p^B(n)$ in response to the delay announcement.

The resulting model referred to as Model 2 is shown in Figure II.4.2. We use a fixed point method to get the new parameter of abandonment $\gamma'$. It is computed by relating each customer's patience to her original one and to the announced delay.

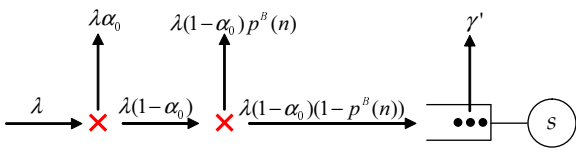



Figure II.4.2: The new model incorporating delay announcement, Model 2

### 2.4 Experiments

We conduct a comprehensive numerical study, in which we analyze the sensitivity of the optimal announcement coverage to various system parameters and compare the performance of the models with and without delay information. We explore the effect of customer behavior (as modeled by $\theta$, and the possibility of misperception), the effect of system size, and the effect of the system load on the announcement coverage in Model 2. We also compare Model 2 when we announce a delay with coverage $\beta$, with Model 2 when we announce the mean delay, as well as Model 1 where we have no delay announcement.

To analyze and compare Models 1 and 2, it is necessary to establish a framework for comparison. We propose a service level framework, similar to those used for waiting times, to account for the tradeoffs between balking and abandonment. The decision variable is $\beta$. We want to determine the best $\beta$, say $\beta^*$, ensuring that no more than $\delta$ (in %) of the customers that enter the queue should abandon. Stated this way, one may choose $\beta^* = 100\%$ because this would lead to no abandonment. However, when this is the case, only customers who find at least one idle server will enter service, thus resulting in very high balking. This is not desirable from a customer service standpoint. We thus formulate the objective under the service level framework as follows:

$$\begin{cases} \min\ P^B \\ \text{subject to } P^R \leq \delta. \end{cases} \tag{II.4.1}$$

The numerical analysis shows that delay announcement, and in particular announcements with higher reliability are more important when customer reaction is high (low $\theta$ or misperception), when systems are small, when the avoidance of customer abandonments is deemed essential (strict service level constraint), and when system congestion is high (overloaded systems under efficiency driven regime). For large systems operating in a quality efficiency driven regime, abandonment decreases, diminishing the importance of announcing delays, or controlling the reliability of the announcements being made.

## 3 Multi-Class Setting: A Newsvendor-Like Approach

### 3.1 Introduction and Related Literature

This work presents an analysis for the estimation of real-time delays in a multi-class real-life setting, and then the decision on what to announce given customer preferences regarding waiting times and announcements made. The setting is that of the large multi-site call center of Bouygues Telecom handling more than 60,000 calls daily. Calls are handled at several sites, differentiated by their size but of identical capability in terms of the types of calls that can be handled. This multi-site system is not equipped with networked routing capabilities implying that each site has its own queue. An important objective of the real-time delay estimation is to use these estimates in routing calls to the various sites. The delay estimators we propose below and subsequently

analyze for the purpose of delay announcement, have indeed been implemented for real-time routing decisions of calls at the Bouygues Telecom call center. The use for routing purposes is not explored any further in this work. At the time of study, the call center was highly congested as manifested by periods with high call retrial. Abandonment probabilities of around 5% were experienced.

Two different types of analysis have been pursued in papers that deal with prediction and announcement in queueing systems: the first predicts and announces delays based on transient queueing analysis (Whitt, 1999c,d; Jouini et al., 2009, 2011) whereas the second considers announcing real-time delay estimators under a fluid model applicable in large and overloaded systems (Armony et al., 2009; Ibrahim and Whitt, 2009). The approach herein is closer to the first. Like in the second approach, it employs a real-time estimation idea, however not directly for the delays but rather for the underlying model parameters. Since model parameters are unknown, an approximation that makes use of real-time estimators for the number of servers is employed. We take the approach of providing simple approximations that are easy to implement in practice. Real state dependent waiting time data is subsequently used to test the quality of the developed delay estimators.

Various announcement forms are considered in the literature: delay announcements of the type "you will wait $x$ minutes", derived from distributions (Whitt, 1999d; Jouini et al., 2009) delays based on real-time estimators (Armony et al., 2009; Ibrahim and Whitt, 2009) state occupancy or length of queue information as indirect waiting time announcements (Guo and Zipkin, 2007; Xu et al., 2007; Aksin et al., 2013), or more general, possibly vague and non-quantitative announcements (Allon et al., 2011). In this work, we focus on delay announcements derived from state-dependent waiting time distributions. More importantly, we propose a new framework making use of a newsvendor-like performance criterion to pick the value to announce from the estimated delay distribution. This framework enables incorporating asymmetric under and over-announcement penalties that, compared to symmetric ones, are more consistent with behavioral evidence. Within this framework, we further propose and test a robust estimator obtained from a robust optimization formulation of the newsvendor problem.

The current work is among the first to study delay announcements in a service setting combining modeling analysis with empirical validation. Most empirical work to date comes from experiments in psychology and marketing that analyze people's reactions to waiting situations, with and without information, in call centers and elsewhere (Munichor and Rafaeli, 2007; Pazgal and Radas, 2008). The papers by Brown et al. (2005) and Feigin (2005) analyze call center data where delay announcements are present, however pursue a more descriptive analysis than the one in this work. The recent paper Aksin et al. (2013) is an exception that combines modeling and empirical analysis in an analysis of delay announcements in a call center.

## 3.2 Choosing What to Announce

In choosing a value to announce from the delay distribution, there seem to be a number of simple options. Labeling the announced delay as $d_a$ (single value), the realized delay as $D_r$ (random variable), one may wish to choose $d_a$ to minimize $E[(D_r - d_a)^2]$. This would result in $d_a^* = E[D_r]$ and estimators for the mean delay can be readily used. Another alternative is to choose $d_a$ to minimize $E[|D_r - d_a|]$. The optimal announcement corresponds then to the median of $D_r$. If the delay distribution is approximated by a symmetrical distribution such as a normal distribution, $d_a^*$ would then be selected as the estimator of mean delay. For non-symmetrical distributions however, the median must be obtained.

The above approaches penalize under announcements and over announcements similarly and ignore the fact that under announcements and over announcements are perceived differently. Our proposed announcement scheme lets the manager choose asymmetric penalties for under announcing ($\alpha$ per unit time), and over announcing ($\beta$ per unit time). In this case, the manager's decision of what to announce to an A-type customer can be formulated as

$$\text{Min } \alpha E[(D_r - d_a)^+] + \beta E[(d_a - D_r)^+]. \tag{II.4.2}$$

Letting $\gamma = \alpha/(\alpha + \beta)$, this leads to the following well-known newsvendor problem's critical fractile solution (Zipkin, 2000) for the optimal announcement, $d_a^* = F_{D_r}^{-1}(\gamma)$, where $F_{D_r}(.)$ is the cumulative distribution function (cdf) of the random variable $D_r$.

## 3.3 Predicting Delays

### 3.3.1 Predicting Delays for Type A

Consider an $M/M/s$ queue where $s$ denotes the number of servers. Under the assumption that the service times are exponential and there are $s$ servers, the delay distribution of a customer who arrives with $n$ waiting customers in front corresponds to the sum of $n+1$ independent exponential random variables with rate $s\mu$. This is an Erlang distribution with $n+1$ stages and rate per stage $s\mu$. Thus, for such an $M/M/s$ system where the number of servers are known, the delay of the high priority customers will have an Erlang distribution.

In our context, the number of active servers is not known. To approximate the delay distribution, we propose to approximate the aggregate service rate $s\mu$ by the total arrival rate of all customers $\lambda(t)$. In relevant applications however, both the arrival rate $\lambda(t)$ and the number of servers $s(t)$ may be time varying. In order to obtain a simple point estimate for the arrival rate at time $t$, we focus on $R(t-\tau)$, the total number of arrivals to service (from all types) in a time window of $(t-\tau, t]$ and propose $\widehat{\lambda}(t) = \frac{R(t-\tau)}{\tau}$. The resulting approximation for the delay distribution is then an Erlang distribution with $n+1$ stages and a rate per stage of $\widehat{\lambda}(t)$. We denote by $\widehat{D}_{erl}$ the resulting random variable. We also use a normal distribution with the same mean and standard deviation in order to obtain a simple formula. The resulting random variable $\widehat{D}_{norm}$ has a normal distribution with mean $(n+1)/\widehat{\lambda}(t)$ and standard deviation $\sqrt{n+1}/\widehat{\lambda}(t)$.

### 3.3.2 Predicting Delays for Types B and C

The delay prediction for type B and C customers is more challenging since those customers not only wait for customers ahead of them at their time of arrival but also have to wait for higher class customers who arrive during their wait. The waiting time of a type B customer is equivalent to the busy period duration in an $M/M/1$ queue with arrival rate $\lambda_A$ and service rate $s\mu$. As in the previous section since the number of servers and the arrival rates are unknown, we approximate $s\mu$ by $\widehat{\lambda}(t)$ and similarly $\lambda_A$ by $\widehat{\lambda}_A(t) = \frac{R_A(t-\tau)}{\tau}$, where $R_A(t-\tau)$ represents the arrivals to the system for type A calls in the time interval $(t-\tau)$. Note that the call center technology allows to compute the number of arrivals from any type to the system or to service. We then adapt the definition of $R(.)$ such that it leads to the best results. We use the number of arrivals that enter service in order to estimate the system capacity $s\mu$, while we use the number to the system in order to estimate the arrival rate $\lambda_A$. We therefore obtain the expectation and the variance of the conditional waiting time denoted by $D_B$ of a new customer B, given $n_1$ and $n_2$ (note that her wait is not affected by the $n_3$). The results for type C customers are similar.

Beyond the moments, the waiting time distribution is difficult to approximate in a simple way. We propose two approximations. One is a normal approximation with the estimated means and standard deviations. The other is for the B-type calls and is an Erlang approximation. In choosing the Erlang distribution we are approximating each busy period by an exponential random variable with rate $(s\mu - \lambda_A)$.

## 3.4 Announcing a Delay from the Estimated Delay Distribution

Recall that the manager's decision of what to announce to an A-type customer is formulated as

$$\text{Min } \alpha E[(D_r - d_a)^+] + \beta E[(d_a - D_r)^+], \tag{II.4.3}$$

leading to the solution for the optimal announcement as $d_a^* = F_{D_r}^{-1}(\gamma)$, where $\gamma = \alpha/(\alpha+\beta)$ and $F_{D_r}(.)$ is the cdf of the random variable $D_r$. Of course, $F_{D_r}$ in the above expression is unknown, and will be replaced by the approximations for A-type customers in Section 3.3.1 to obtain approximately optimal values for $d_a$. In particular, the Erlang approximation then leads to $d_{a,erl}^* = F_{\widehat{D}_{erl}}^{-1}(\gamma)$, and the normal approximation results in $d_{a,norm}^* = \frac{n+1}{\widehat{\lambda}(t)} + z^* \frac{\sqrt{n+1}}{\widehat{\lambda}(t)}$, where $z^* = \Phi^{-1}(\gamma)$ and $\Phi^{-1}(.)$ denotes the inverse cdf of a standard normal random variable.

As another benchmark, we propose a robust estimator that finds the optimal announcement for the worst-case probability distribution with mean $(n+1)/\widehat{\lambda}(t)$ and standard deviation $\sqrt{n+1}/\widehat{\lambda}(t)$. The Erlang and normal delay approximations make distributional assumptions as well as assumptions about the distribution parameters. The distribution free robust estimator which we propose provides a benchmark where the worst case distributional form is found for the given mean and standard deviation.

We first consider the penalty maximizing (worst-case) delay distribution for a given $d_a$ subject to constraints on the expectation and variance values. This is a maximization problem. We then

consider the worst case delay random variable for a given $d_a$. This is a minimization problem. The above robust optimization formulation is known as a min-max distribution-free procedure in the context of the newsvendor problem and leads to a surprisingly simple solution (Scarf, 1958; Gallego and Moon, 1993) for the optimal $d_a$. It is given by $d^*_{a,rob} = \frac{n+1}{\widehat{\lambda}(t)} + \frac{\sqrt{n+1}}{2\widehat{\lambda}(t)}\left(\sqrt{\frac{\alpha}{\beta}} - \sqrt{\frac{\beta}{\alpha}}\right)$. We follow the same approach for the B-type calls.

### 3.5 Data-Based Validation of Delay Announcements

We explore the performance of delay announcements under the two approximations (Erlang and normal) for different values of $\gamma = \alpha/(\alpha+\beta)$, by comparing them to the corresponding announcements for the data on state dependent waiting times. This data based validation allows us to assess the value of the approximations in making delay announcements in a real call center setting. Thus we show that under all complexities of a real operation, the earlier tested simple approximations perform well also when used in making delay announcements. We measure the performance of each estimator with respect to the realized waiting time distribution.

For the A-type calls, we observe from the numerical study that while announcing the mean delay does quite well for a $\gamma$ value that is close to 0.5, its performance deteriorates dramatically as the customers attach a higher penalty to under-announcements. The Erlang approximation performs well across all $\gamma$ values. Comparing the normal approximation based announcements to the robust delay announcement, we observe that once the mean and standard deviation have been estimated, it is better to use the robust delay announcement, which performs particularly well for $\gamma$ values 0.7 and 0.8. For the B-type calls, the relative errors are higher compared to the A-type ones. This is not surprising due to the increasing level of approximations being performed both in the data and models. However, the Erlang-based announcement is still quite good for all $\gamma$ values, particularly as these are getting higher. Announcing the mean appears to be the best option for $\gamma$ values 0.6 and 0.7, but it deteriorates for higher $\gamma$ values. Thus, without a good understanding of these penalties, announcing the mean seems risky. We also observe that the robust delay announcement ensures an average relative error of around 10%. The robust estimator mostly outperforms the mean and the normal approximation based announcements.

## 4 Concluding Remarks and Future Research

My contributions here are related to the formulation and analysis of call center models where anticipated delays are announced to customers upon arrival. In the first part of this work, we considered a single class call center where informed customers may react to delay information through balking or abandonment. Our analysis illustrated the tradeoffs between abandonment and balking that have to be made in choosing the announcement percentile, and demonstrates the role customer and system parameters play on this choice. In future work, it would be interesting to empirically describe customers' reactions in response to delay announcements. Lab experiments that control for everything else and allow a direct comparison between the

models with imperfect, and no delay information would be valuable in supporting assumptions pertaining to balking and abandonment made in the analysis herein.

In the second part of this chapter, I presented my work on estimating delays experienced by customers with different priorities, and the determination of the appropriate delay announcement to these customers, in a multi-class real-life call center. The robust delay announcement that makes use of the moment estimators provides an alternative that protects against the worst case when such queueing analysis is not available. The idea of a robust delay announcement is new, and should be explored further in future practice as well as research, particularly in settings with high complexity and uncertainty like the one we considered.

# Chapter II.5

# Operational Issues: Optimal Routing in Call Centers

## 1 Introduction

This chapter synthesizes my contributions to the literature on the optimal routing of jobs in call centers. A routing policy, or a scheduling policy, or a discipline of service, determines the rule of assigning jobs to the agents, upon arrivals or at service completion times. New technology-driven innovations in call centers 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, emails and chat. However, several issues on the management of call center operations emerged also as a result of advanced technology. In this context, an interesting question for managers is how the real-time match of demand (various job types) and agents should be prescribed?

My research results on the scheduling have been motivated by my collaborations with the mobile phone company Bouygues Telecom and the call center consulting company Interact-iv.com. My overall objective is to answer the question: how to efficiently share the agent time between the available job types in order to improve the call center performance? I have investigated this question for various problem formulations and under various settings of single and multi-channel call centers. A major part of my contributions on call centers are related to routing issues. This may explain why this chapter is longer than the previous ones. My contributions can divided into three parts as described below.

The first part of my contributions is related to the online scheduling for a single channel multi-class call center of Bouygues Telecom, and is described in Section 2. This has been done with my colleagues Auke Pot (PhD student) and Ger Koole from VU University Amsterdam, and Yves Dallery. Modeling our call center as a GI/GI/s+M queue with two classes of impatient customers (premium and regular), we focus on developing scheduling policies that satisfy a target ratio constraint on the abandonment probabilities of premium customers to regular ones. In the Bouygues Telecom call center, managers want to reach some predefined preference between customer classes for any workload condition. The motivation for this constraint comes from the

difficulty of predicting in a quite satisfying way the workload. In such a case, the traditional routing problem formulation with differentiated service levels for different customer classes would be useless. For this new problem formulation, we propose a family of online queue joining policies. The principle of our policies is that we adjust their routing rules by dynamically changing their parameters.

The second and third parts of my contributions deal with the optimal control for multi-channel call centers. I have collaborated in this work with Benjamin Legros (PhD student) and Ger Koole. The issues we addressed are originally initiated by the clients (call centers) of the company Interact-iv.com. In The second part of my work, as described in Section 3, we consider a multi-channel call center with inbound calls and emails. We focus on the analysis of a threshold policy on the reservation of agents for the inbound calls. We study a general non-stationary model where calls arrive according to a non-homogeneous Poisson process. The optimization problem consists of maximizing the throughput of emails under a constraint on the waiting times of inbound calls. We propose an efficient adaptive threshold policy that is easy to implement in the Automatic Call Distributors (ACD). This scheduling policy is evaluated through a comparison with the optimal performance measures found in the case of a constant arrival rate.

In the last part, we consider a blended call center with calls arriving over time and an infinitely backlogged amount of outbound jobs. Inbound calls have a non-preemptive priority over outbound jobs. The inbound 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, not explored yet, to efficiently split the agent time between inbound calls and outbound jobs. We focus on the optimization of the outbound job routing to agents. Our objective is to maximize the expected throughput of outbound jobs subject to a constraint on the waiting times of inbound calls. We develop a general framework with two parameters for the outbounds. 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 control parameters has an extreme value.

## 2 Online Policies for Impatient Customers

### 2.1 Context and Motivation

Inspired by a real-life problem, we consider a single channel multi-class call center. There are two types of impatient customers: premium and regular ones. Given a staffing level, our purpose is to develop control schemes for arrival calls and idle agents, subject to satisfying a constraint related to the probabilities of being lost, i.e., the probabilities to abandon. The reason behind this objective is to translate a desired fairness between customer classes. In Bouygues Telecom,

the abandonment of any class of customers is equivalent to a loss of goodwill. Both classes are indeed valuable for the company with a particular preference to the premium class. Having nearly no abandonments of the premium class and a lot of abandonments of the regular one is not desirable. The call center would instead prefer to have more abandonments of premium calls and fewer abandonments of regular ones. This is captured through a ratio of the abandonment probabilities. The ratio would be typically between 0 and 1. A low value of this ratio would translate to a strict preference of the company to premium calls. A ratio close to 1 would however translate an equal preference between the two classes. An intermediate value would translate an in-between preference.

In practice, managers traditionally handle this problem by separately setting for each customer class a constraint on the probability to abandon. The performance of such a formulation highly depends on the quality of the workload prediction. Several studies (Jongbloed and Koole, 2001; Avramidis et al., 2004) have shown however that the arrival process and the workload are hard to predict in call centers. Once the actual workload deviates from the predicted one, we are no longer able to meet the predefined performance constraints. Existing solutions often rely on static strict priority rules, such as a static strict non-preemptive priority for one class over the other. If the workload is underestimated, most of the capacity of the system will be dedicated to premium calls. We may then satisfy the performance constraint of premium calls, while having a heavily penalized one for regular calls. However, if the workload is overestimated, the performance of premium calls will be very high and that of regular ones will not profit that much from the overcapacity.

A new formulation of the routing problem using a target ratio constraint between the service levels of the two classes would, as a consequence, be a better alternative. It allows to better control the different situations which may occur (under- or overestimation of the workload). Satisfying the constraint ratio enables to share as desired the capacity of the system between the two classes. In addition, this new formulation generalizes the traditional one where we have a target abandonment probability for each class.

## 2.2 Related Literature

The most related literature to this work is that on the control of queueing systems. Scheduling policies have been studied in great depth within the context of queueing systems. A scheduling policy, or a discipline of service, prescribes the order in which customers are served. Randolph (1991) classifies scheduling policies into those using online schedule rules and those using static schedule rules. Each of the above classes of policies can be further classified into two major classes: agent scheduling and customer routing. We refer the reader to Garnett and Mandelbaum (2001) for a background.

In what follows, we briefly review existing results about scheduling policies for V-models, as we consider in this work. Pekoz (2002) addresses the analysis of a multi-server non-preemptive priority queue with exponentially distributed inter-arrival and service times. She finds and

evaluates the performance of an asymptotically optimal policy that minimizes the expected queueing delay for high priority customers. Guérin (1998) presents a model without waiting queues. The model contains a multi-server station, which receives low and high priority arrivals. He develops an admission policy for the low priority customers such that the fraction of blocked high priority customers is bounded and he analyzes the system under that policy. In the context of call centers, Gurvich et al. (2008) consider a large-scale system under the V-design and characterize asymptotically optimal scheduling and staffing schemes (as system load grows to infinity). The optimal scheduling and staffing schemes minimize the staffing costs subject to satisfying quality of service constraints for the different customer classes. Maglaras and Zeevi (2005) consider profit maximization for a loss system two-class V-model with pricing, sizing, and admission control. Milner and Olsen (2008) explore the role that service level constraints in outsourcing contracts play in settings where the contractor firm has both contractual and non-contractual customers. Another paper with a similar idea of contract and non-contract customers is Bhulai and Koole (2003). For a detailed survey of relevant papers considering the optimal control of the V-model, we refer the reader to Gurvich (2004).

## 2.3 Framework

### 2.3.1 Model Description

We model our call center as a queueing system with two customer classes: a premium customer class $A$, and a regular one $B$. The model consists of two infinite queues, say queues 1 and 2, and a set of $s$ parallel, identical servers representing the set of agents. All agents are able to answer all customer classes. The call center is operated in such a way that at any time, any customer can be addressed by any agent. Upon arrival, a customer is addressed by one of the available agents, if any. If not, the call joins one of the queues.

We consider the family of queue joining. A policy determines the rule of assigning customers upon arrival to one of the queues. Upon arrival, a customer of any class can be sent to any queue. That is, each time an arrival $A$ or $B$ enters the system, an individual decision is made as a function of the system state: assign this new arrival to queue 1 or 2. In other words, we want to specify that there is no an a priori fixed rule of assigning for example all customers $A$ to queue 1 and all customers $B$ to queue 2. So queue 1 or 2 may contain a mix of the two customer types. Finally note that customers waiting in queue 1 have a non-preemptive strict priority over those in queue 2. Customers waiting in a given queue, are served in the order of their arrivals, i.e., under FCFS. Also, the priority rule between the queues is non-preemptive because it is not common in call centers to interrupt the service of a customer and serve another one with a higher priority.

Inter-arrival times and service times are assumed to be i.i.d. and follow a general distribution. In certain cases, we shall consider the exponential distribution for successive service times. The mean service time rates of classes $A$ and $B$ are $\mu_A$ and $\mu_B$, respectively. In addition, we let the customers be impatient. After entering the queue, a customer waits a random length of time

for service to begin. If service has not begun by this time she will abandon (leaves the queue). Patience times of classes $A$ and $B$ are assumed to be i.i.d. and exponentially distributed with rates $\gamma_A$ and $\gamma_B$, respectively. Following similar arguments, the behavior of this call center can be viewed as a GI/GI/s+M queueing system.

We denote by $m$ the class of a customer, for $m \in \{A,\ B\}$, and by $\pi$ a scheduling policy. In the long run, the fraction of abandonments (probability to abandon) of class $m$ is denoted by $\mathbf{P}_\pi^m$, and that of all classes by $\mathbf{P}_\pi$. We now define our main performance measure, denoted by $\mathbf{c}_\pi$. It is the ratio of the probability to abandon of class $A$ over that of class $B$, i.e., $\mathbf{c}_\pi = \dfrac{\mathbf{P}_\pi^A}{\mathbf{P}_\pi^B}$.

### 2.3.2 Problem Formulation

Due to the highly uncertain environment of call centers, it is usually hard to estimate the workload within a relative accuracy. We often end up in practice with either an underestimated, or an overestimated workload. Next, the common practice in call centers is to develop routing policies that aim to satisfy differentiated service level constraints. In case of a fixed number of agents, the abandonment probability of each class will be affected by the forecasting error and will thereafter deviate from the predefined one. Furthermore, the actual service level deviations can be very different between the customer classes (for example when using strict priority scheduling policies). This behavior is undesirable for managers because one would lose some given fairness between customer classes. For any work condition, a manager would like to reach a desired fairness between customer classes, through an appropriate share of the available capacity.

Based on this motivation, we formulate the following problem. We assume that staffing has already taken place, such that the number of available agents is known. We aim to develop scheduling policies that satisfy a target ratio constraint, say $\mathbf{c}^*$, of the abandonment probabilities between of the two customer classes. In mathematical terms, we look, if at all possible, for $\pi \in \Pi$ subject to $\mathbf{c}_\pi = \mathbf{c}^*$, where $\Pi$ denotes the class of workconserving non-preemptive scheduling policies.

The target ratio $\mathbf{c}^*$ translates a desired preference between the two customer classes that we want to reach for any actual workload. The $\mathbf{c}^*$ target formulation generalizes the traditional formulation where we have a service level constraint for each class. If the workload is quite correctly estimated, having a target ratio of abandonment probabilities is equivalent to having a target abandonment probability for each class. In addition if the workload is incorrectly estimated (under- or overestimated), the capacity of the system is shared over customer classes in a way that allows to preserve (if possible) the predefined preference between customer classes.

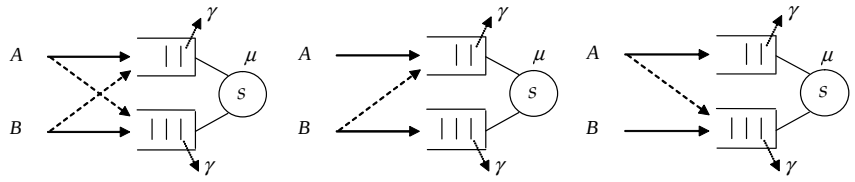
## 2.4 Online Scheduling Policies

We propose three queue joining scheduling policies, denoted by $\pi_1$, $\pi_2$ and $\pi_3$. Customers in queue 1 have a non-preemptive priority over customers in queue 2. Customers of a class can be sent to any queue. Our policies do not anticipate on future events. They just react to the realization of the ratio that is determined by the history of the process.

**Scheduling Policy $\pi_1$:** The scheduling policy $\pi_1$ starts identically as a strict priority policy that gives the higher priority to class $A$. After the epoch at which the first customer $B$ finishes her service, we apply the following assignment rule for any new arrival (denoted by the $k^{th}$ arrival). Let $d_k$ be the epoch of that arrival. Let $\mathbf{P}_k^A$ ($\mathbf{P}_k^B$) be the achieved service level until $d_k$ for class $A$ ($B$). Let $\mathbf{c}_k$ be the achieved ratio starting until $d_k$, $\mathbf{c}_k = \mathbf{P}_k^A/\mathbf{P}_k^B$. If $\mathbf{c}_k < \mathbf{c}^*$, then we give high priority to class $B$, i.e., if the new arrival is class $A$, it is routed to queue 2, otherwise, it is routed to queue 1. However if $\mathbf{c}_k \geq \mathbf{c}^*$, we give high priority to class $A$, i.e., if the new arrival is class $A$, it is routed to queue 1, and if it is class $B$, it is routed to queue 2. An illustration of $\pi_1$ is shown on Figure II.5.1(a).

**Scheduling Policy $\pi_2$:** The scheduling policy $\pi_2$ starts identically as $\pi_1$ until the epoch at which the first customer $B$ finishes service. Let a new arrival enter the system. Under $\pi_2$, a customer $A$ is always routed to queue 1. However, the assignment rule of class $B$ customers is as follows. If $\mathbf{c}_k < \mathbf{c}^*$, a new class $B$ arrival is routed to queue 1, otherwise if $\mathbf{c}_k \geq \mathbf{c}^*$, it is routed to queue 2. An illustration of $\pi_2$ is shown on Figure II.5.1(b).

**Scheduling Policy $\pi_3$:** The scheduling policy $\pi_3$ starts identically as policies $\pi_1$ and $\pi_2$ until the first customer $B$ finishes service. Let a new arrival enter the system. Under $\pi_3$, a customer $B$ is always routed to queue 2. However, the assignment rule of customers of class $A$ is as follows. If $\mathbf{c}_k \geq \mathbf{c}^*$, then a new class $A$ arrival is routed to queue 1, otherwise if $\mathbf{c}_k < \mathbf{c}^*$, it is routed to queue 2. An illustration of $\pi_3$ is shown on Figure II.5.1(c).




(a) $\pi_1$ (b) $\pi_2$ (c) $\pi_3$

Figure II.5.1: Queue joining policies

Policy $\pi_1$ can be immediately obtained intuitively. It allows the achieved ratio to be updated upon each arrival such that it converges in the long run to the objective. The idea behind $\pi_2$ is that we keep always customers of class $A$ in the high priority queue, however when it is necessary, we assign customers of class $B$ to this queue to improve their service level (which deteriorates the service level of customers of class $A$). This allows to increase the transient ratio and keep it close to the objective. As a consequence, the ratio converges in the long run to the desired value. Policy $\pi_3$ can be viewed as another variant. It sometimes allows to penalize class $A$ customers by assigning them to the low priority queue, which again allows to increase the transient ratio.

One may construct several auxiliary policies similar to the ones above. For example, instead of changing the priority rule at each new arrival epoch, we only change it at the arrival epoch of the customer who finds all servers busy and both queues empty. Then, we continue that rule

until the end of the current busy period. With regard to reaching the target ratio, the latter class of policies has the same properties as those of $\pi_1$, $\pi_2$ and $\pi_3$. One drawback could be that they are less reactive to correct the transient ratio.

### 2.5 Experiments and Synthesis

We consider various numerical examples to cover a wide range of real-life settings. For each setting we determine $[\mathbf{c}_{\pi_A}, \mathbf{c}_{\pi_B}]$ and check that $\mathbf{c}^*$ belongs to this interval. The experiments show that the target ratio is always met by policies $\pi_1$, $\pi_2$ and $\pi_3$. For each system, the value of the ratio under policy $\pi_A$ represents a lower bound for the achievable ratio under any workconserving non-preemptive scheduling policy. We can not do better when considering that class of policies.

Experiments for class $A$ show that the expected waiting times in the queue are ordered according to policies $\pi_A$, $\pi_1$, $\pi_3$ and $\pi_2$. The order $\pi_1$ then $\pi_3$ then $\pi_2$ is expected because of the general property that FCFS maximizes the expected waiting time of served customers. Policy $\pi_A$ is the best for the expected waiting time of served customers. An explanation would be related to the small values of waiting times achieved under that policy. From the experiments, we see for class $B$ that the expected waiting times in the queue of served customers are ordered according to policies $\pi_1$, $\pi_2$, $\pi_3$ and $\pi_A$. One may explain these results through the same arguments used above.

## 3 Threshold Policies for Calls and Emails

### 3.1 Introduction and Related Literature

The objective of this work is the analysis of threshold policies in the context of multi-channel call centers with inbound calls and emails (back-office jobs). It is undertaken under our collaboration with call center consulting company Interactiv.com. To limit the necessity to have extremely accurate forecasts, inbound calls are sometimes mixed with other types of customer contacts which have a less strict delay requirements, 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.

In Gans and Zhou (2003), it is shown that an efficient assignment policy is of the following form: outbound jobs should only be scheduled when there are no waiting inbound calls and when the number of idle agents exceeds a certain threshold that depends on the system parameters. But the parameters, especially the arrival rate, are often hard to determine. In this work, 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 call service level by the end of the day, while

maximizing the number of emails that are done.

Only few papers focus on blending. Deslauriers et al. (2007) extend the earlier mentioned papers by having different types of agents. Outbound jobs are served only by multi-channel (blended) agents, whereas inbound calls can be served by either inbound-only or blended agents. They evaluate several performance measures of interest, including the rate of outbound jobs and the proportion of inbound calls waiting more than some fixed number of seconds. Armony and Ward (2010) present an optimization problem; the objective is to minimize steady-state expected customer waiting time subject to a fairness constraint on the workload division. They show that in such a problem, which is close to ours, 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 common use to give priority to contract customers only in off peaks. They show that this choice is a good one under classical assumptions (such as stationarity). They also present 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.

### 3.2 Problem Formulation

We consider a call center modeled as a multi-server queueing system with two types of jobs, foreground jobs (inbound calls) and background jobs (emails). 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 FCFS queue with infinite capacity. There is an infinite supply of background jobs, waiting for treatment in a dedicated FCFS queue. There are $s$ agents. Each agent can handle both types of jobs. We assume that the service times of foreground and background jobs are exponentially distributed with rates $\mu$ and $\mu_0$, respectively. 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.

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 and we restrict ourselves to the case of threshold policies. 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, 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, so, there are at least $u$ agents working at any time.

We propose an adaptive threshold policy which adjusts the threshold as a function of the process of the call service level. We divide the working day into $N$ identical intervals, each with length $\theta$. The total working duration in a day is $D$, $D = N\theta$. 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 $D$. Let also $SL$ be the proportion, for the whole day, of calls that have waited less than $\tau$. In summary, our optimization problem can be formulated as

$$\begin{cases} \text{Maximize } T \\ \text{subject to } SL \geq \alpha, \end{cases} \tag{II.5.1}$$

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 $t \geq 0$, $\mu$ and $s$ are such that $SL \geq \alpha$ for $u_i = 0$ ($i = 1, ..., N$).

### 3.3 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. 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 $\tau$ ($\tau > 0$).

**Equal Service Rates:** We consider the case $\mu = \mu_0$. Let us define the stochastic process $\{x(t), t \geq 0\}$, where $x(t) \in \{u, u+1, u+2, \cdots\}$ 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. It is similar to that of an M/M/s queue without the states $\{0, 1, \cdots, u-1\}$. The transition rate from state $x$ to state $x-1$ is $\min\{x, s\}\mu$, for $x > u$, and that from state $x$ to state $x+1$ is $\lambda$, for $x \geq u$. We denote by $a$ the ratio $\frac{\lambda}{\mu}$. Also, under the stability condition $\frac{\lambda}{s\mu} < 1$, we denote by $p_x$ the steady-state probability to be in state $x \in \mathbb{N}$. We are then able to compute explicitly the email throughput, $T(s, u, a)$, and the probability that the call waiting time is less than $\tau$, $SL = P(W < \tau)$. In Proposition II.5.1, we prove monotonicity results of the system performance measures as a function of the threshold.

**Proposition II.5.1** *For $a > 0$, the following holds:*

1. *The email throughput $T$ is strictly increasing and neither convex nor concave in $u$, for $0 \leq u \leq s$. The end of the email throughput, for $0 \leq s - 2 \leq u \leq s$, is concave in $u$.*

2. *The call service level $P(W < \tau)$ is strictly decreasing and concave in $u$, for $0 \leq u \leq s$.*

**Unequal Service Rates:** 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 are however still useful for the numerical experiments. Our approach consists in using a Markov chain analysis to derive the steady-state probabilities of the system, from which the performance measures are characterized thereafter. We only focus on the particular case $u = s$, where the most of the job is the manipulation of Erlang and hypoexponential distributions. 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.

**Numerical Observations:** We use the performance evaluation 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.

### 3.4 Our Adaptive Threshold Policy (ATP)

We propose for Problem (II.5.1) an adaptive threshold policy which adjusts the threshold as a function of the call workload. This policy is based on the the first and second order monotonicity properties of the performance measures as a function of the threshold $u$, and on the numerical observation. 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 $\alpha$ at the beginning of an interval $i$ ($i = 2, ..., N$) then the policy increases (decreases) the threshold. To update 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 (in case if $SL_i > \alpha$), then we consider $c_i = c_{i-1} + 1 - c_{i-1}/s$. If we need to decrease the threshold (in case $SL_i < \alpha$), then $c_i = c_{i-1} - c_{i-1}/s$. In the remaining case ($SL_i = \alpha$), we consider $c_i = c_{i-1}$. This policy is refereed to as ATP.

In what follows, we discuss the efficiency of how ATP updates the threshold. The main two characteristics of ATP are:

- An Increasing (decreasing) of the threshold in case the measured call service level is better (worse) than the target service level,

- A decreasing speed in the increasing (decreasing) of the threshold when this threshold increases (decreases).

From Proposition II.5.1, we know that the throughput increases and the call service level decreases in $u$. Thus, the threshold should be increased when the measured service level is better than the target service level, and vice versa. This justifies the first characteristic of ATP.

The second characteristic of ATP is justified by the convexity of the performance measures and the correlation between them.

**Evaluation of the Adaptive Threshold Policy:** 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 see that the threshold should vary between two or more values. The reason is that we need to exactly satisfy the constraint on calls in Problem (II.5.1) in order to maximize the email throughput. From Bhulai and Koole (2003), we know that to exactly satisfy the constraint on calls, randomization is optimal for threshold policies. A randomized threshold policy, between two thresholds $u_1$ and $u_2$ and with a randomization parameter $p \in [0, 1]$, works as follows. At each event (an inbound call arrival or a service completion), the value of the threshold value changes from $u_1$ to $u_2$ with probability $p$, stays in $u_1$ with probability $1 - p$, changes from $u_2$ to $u_1$ with probability $1 - p$, stays in $u_2$ with probability $p$.

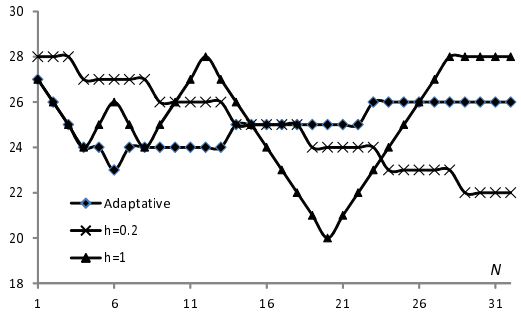
The randomization between two thresholds allows for the constraint on calls to be met exactly. For our system with constant parameters, the randomization is between two successive thresholds. Since the throughput is neither convex nor concave it is difficult to rigorously prove this result. For all the 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 $s$. We run experiments with constant arrival rates and compare the optimal throughput with the one found under ATP. Although ATP is not optimal, the difference with the optimum is quite small. This shows the advantage of ATP in the case of constant arrival rates.

## 3.5 Non-Constant Arrival Rates

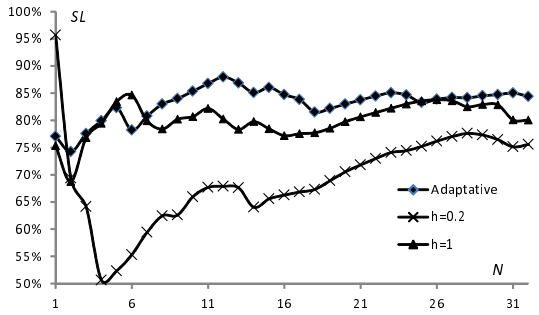
We compare ATP with methods that use constant step sizes. We consider cases where the length of the working day equals eight hours ($D = 8$h) and a frequent possibility of reevaluating the real threshold $c$, at the beginning of each time interval. We consider various interval lengths. We use simulation to obtain the performance measures.

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 $c_0 = u_0 = s$. From the numerical experiments, we observe that

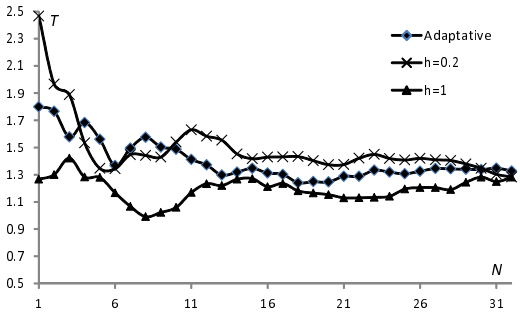
ATP performs better or at least similarly to the constant step methods with an aversion of risk equal to 100.




(a) Threshold




(b) Service level




(c) Email throughput

Figure II.5.2: Evolution of the threshold, the service level and the throughput (Scenario 2)

Figures II.5.2(a), II.5.2(b) and II.5.2(c) give 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 a given scenario. 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 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 summary, the analysis of the performance of ATP under the cases of constant and non-constant arrival rates shows its efficiency. 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.

# 4 Job Routing with Idling Times during the Call Service

## 4.1 Introduction

As in the previous section, the work presented here deals with the analysis of multi-channel call centers. It is also motivated by our collaboration with Interact-iv.com. We consider a blended call center setting where inbound calls and outbound jobs are combined. As in the previous section, we want to address the routing problem of outbound jobs to agents (Bernett et al., 2002; Bhulai and Koole, 2003; Legros et al., 2014a), i.e., as a function of the systems parameters and the service level constraints, when should we ask the agent to treat outbound jobs between the call conversations?

The routing question is further important in the context 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 job if needed.

For an efficient use of the agent time, one would think about the routing of the less urgent jobs 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 payment 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 for example), which may take some minutes. For such situations, it is natural that the system manager thinks about using the opportunity to route outbound jobs to an agent during the break of an undergoing call conversation, and not only when no calls are waiting in the queue. The advantage is an efficient use of the agent time and therefore a better call center performance. Also, agents become less bored because of the diversity of activities, and therefore, they are kept from falling into a rut.

In this work, we consider a call center with an infinite amount of outbound jobs. Inbound calls arrive over time, and in the middle of an inbound call conversation, a break is required. Given this type of call centers, we are interested in optimizing its functioning by controlling how the resource should be shared between the two types of jobs. Calls are more important

than outbound jobs in the sense that calls request a quasi-instantaneous answer (waiting time in the order of some minutes), however outbound jobs are more flexible and could be delayed for several hours. An appropriate functioning is therefore that the agent works on inbound calls as long as there is work to do for inbound calls. The agent can then work on outbound jobs 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.

## 4.2 My Contributions

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 outbound job routing between call conversations but not inside a call conversation.

To answer this question, we develop a general framework with two parameters for the outbound job routing to agents. One parameter controls the routing between calls, and the other does the control inside a call conversation. Although this modeling is not optimal, its performance measures as shown later are closed to the optimal ones and its routing policy is easier to implement than the complex optimal routing. 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 outbound job expected throughput under a constraint on the service level of the call waiting time. As a function of the system parameters (the server utilization, the outbound job 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 outbound job routing parameters has an extreme value. An extreme value means that the agent should do all the time outbound jobs 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.

We then focus on the routing optimization problem for the multi-server case, using simulation and approximations developed under the light and heavy traffic regimes. We found that most of the observations of the single server case are still valid, in particular the result stating that at least one control parameter has an extreme value.

## 4.3 Positioning of My Contributions

There are three related streams of literature to this work. 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 is already given in Section 3.1. We next review some of the literature on the analysis of queueing systems with phase type service time distributions. This mainly involves the steady-state analysis of Markov chains and is usually addressed using numerical methods (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 (Seelen, 1986; 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 well known matrix-geometric method (Neuts, 1995). 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 jobs is not efficient even with two easy jobs because of the possible interferences. In our models, we are not considering simultaneous tasks in the sense that an agent can not talk to a customer and at the same time treats an outbound job. Charron and Koechlin (2010) studied the capacity of the frontal lobe to deal with different tasks by alternation (as here for calls and outbound jobs). They develop the notion of *branching*: capacity of the brain to remember information while doing something else. They show that the number of jobs 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).

## 4.4 Problem Description and Modeling

We consider a call center modeling with $s$ identical agents and two types of jobs: inbound calls and outbound jobs. The arrival process of inbound calls is assumed to be Poisson with mean arrival rate $\lambda$. We assume to have an infinite amount of outbound jobs that are waiting to be treated in a dedicated first come, first served (FCFS) queue with an infinite capacity.

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 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. We assume that the durations of the three stages

are jointly independent. The service rates of the first, second and third stages are denoted by $\mu_1$, $\mu_2$ and $\mu_3$, respectively. An agent handles an outbound job within one single step without interruption. The time duration of an outbound job treatment is random and assumed to be exponentially distributed with rate $\mu_0$. Upon arrival, a call is immediately handled by an available agent, if any. If not, the call waits for service in an infinite FCFS dedicated queue. Inbound calls have a non-preemptive priority over outbound jobs. We are interested in an efficient use of the agent time between inbound calls and outbound jobs. More concretely, we want to answer the question when should we treat outbound jobs for the following optimization problem

$$\begin{cases} \text{Maximize the expected throughput of outbound jobs} \\ \text{subject to a service level constraint on the call waiting time in the queue.} \end{cases} \tag{II.5.2}$$

We propose a simpler model for the routing of outbound jobs to agents. It is referred to as *probabilistic model* or *Model PM* and is described below. Although this model is not optimal, we numerically show its efficiency through a comparison with the optimal policy (Legros et al., 2014b). It is moreover easy to implement and to understand by a system manager.

**Probabilistic Model (Model PM):** We distinguish the two situations when an agent is available to handle outbound jobs 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 outbound jobs with probability $p$ (independently of any other event), or does not work on outbound jobs 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 outbound job to work on. After finishing the treatment of this outbound job, 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 outbound job, and so on. At some point in time, a new call would arrive while the agent is working on an outbound job. The agent will then handle the call as soon as she finishes the outbound job treatment.
*Inside a call*: Just after the end of the first stage of an underway call service (regardless whether there are other waiting calls in the queue or not), the agent treats one or more outbound jobs with probability $q$ (independently of any other event), or does not work on outbound jobs 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 by 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 outbound job to work on. After finishing the treatment of this outbound job, 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 outbound job, and so on. At some point in time, the currently served

Table II.5.1: Particular cases of Model PM

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

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

We further consider next 4 particular cases of Model PM as shown in Table II.5.1. Although these models might appear to be too restrictive to solve Problem (II.5.2), we show later their merit 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 expected throughput of outbound jobs 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.

## 4.5 Single Server Analysis

We provide an exact method to characterize the call waiting time in the queue and the outbound job expected throughput for Model PM and its extreme cases for a single-server model. We also develop various structural results for the optimization problem, which allows to enhance our understanding of the system behavior. This would not be possible to do directly for the multi-server case since an exact analysis is very complex. We extend the analysis to multi-server case using approximate asymptotic regimes.

### 4.5.1 Performance Evaluation

Our approach consists of 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.

We use 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. Using the underlying Markov chain, we explicitly compute the probability of delay of a call (probability of waiting) denoted by $P_D$, and the expected throughput of outbound jobs denoted by $T$. Note that the stability condition of Model PM is $\lambda < \left(\frac{q}{\mu_0} + \frac{1}{\mu_1} + \frac{1}{\mu_2} + \frac{1}{\mu_3}\right)^{-1}$. Let us now define $W$, a random variable, as the steady-state 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 and using PASTA, we numerically obtain $\mathrm{P}(W < t)$.

We then may compute the expected call waiting time in Model PM, denoted by $E(W)$. Consider first a model similar to Model PM except that outbound jobs can only be treated inside a call conversation. We denote this model by Model PM', and its call expected waiting time by $E(W')$. We prove that he expected waiting time in PM is delayed by $\frac{p}{\mu_0}$ compared to that in PM', for $p \in [0,1]$. We then compute $E(W')$ using the Pollaczeck-Kinchin result for an M/G/1 queue, from which we finally derive $E(W)$. For the 4 extreme cases of Model PM (Models 1,...,4), one may simply apply the previous analysis with the corresponding extreme values of $p$ and $q$.

### 4.5.2 Comparison Analysis and Insights

We start by a comparison analysis between the extreme cases Models 1,...,4. The comparison is based on the optimization problem (II.5.2). 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. We then 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 outbound job treatment of outbound jobs either between calls or inside a call conversation, which gives an interest in practice for Models 1,...,4.

**Comparison between the Extreme Cases:** We first compare between Models 1,...,4 based on their performance in terms of the outbound job expected throughput, denoted by $T_1$,..., $T_4$, respectively. It is obvious that Model 4 is the best and Model 1 is the worst (no outbound jobs at all). Let us now compare between Models 2 and 3. 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 $\frac{1}{\frac{1}{\mu}+\frac{1}{\mu_2}} < \lambda < \mu$. The condition under which $T_3 > T_2$ is then $R = \frac{1}{\frac{1}{\mu_0}+\frac{1}{\mu_1}+\frac{1}{\mu_3}+\frac{2}{\mu_2}} < \lambda < \mu$.

Treating outbound jobs 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, idle 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 outbound job 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 words, there is no sufficient time to treat outbound jobs inside the call conversation.

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

following problem

$$\begin{cases} \text{Maximize } T \\ \text{subject to } E(W) \leq w^*, \end{cases} \tag{II.5.3}$$

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 $\lambda$, the manager is likely to choose Model 1 (no outbound jobs), and for other periods with a very low $\lambda$, she is likely to choose Model 4 (outbound jobs between calls and inside a call). However for intermediate values of $\lambda$, the optimal choice is not clear. This is what we next investigate.

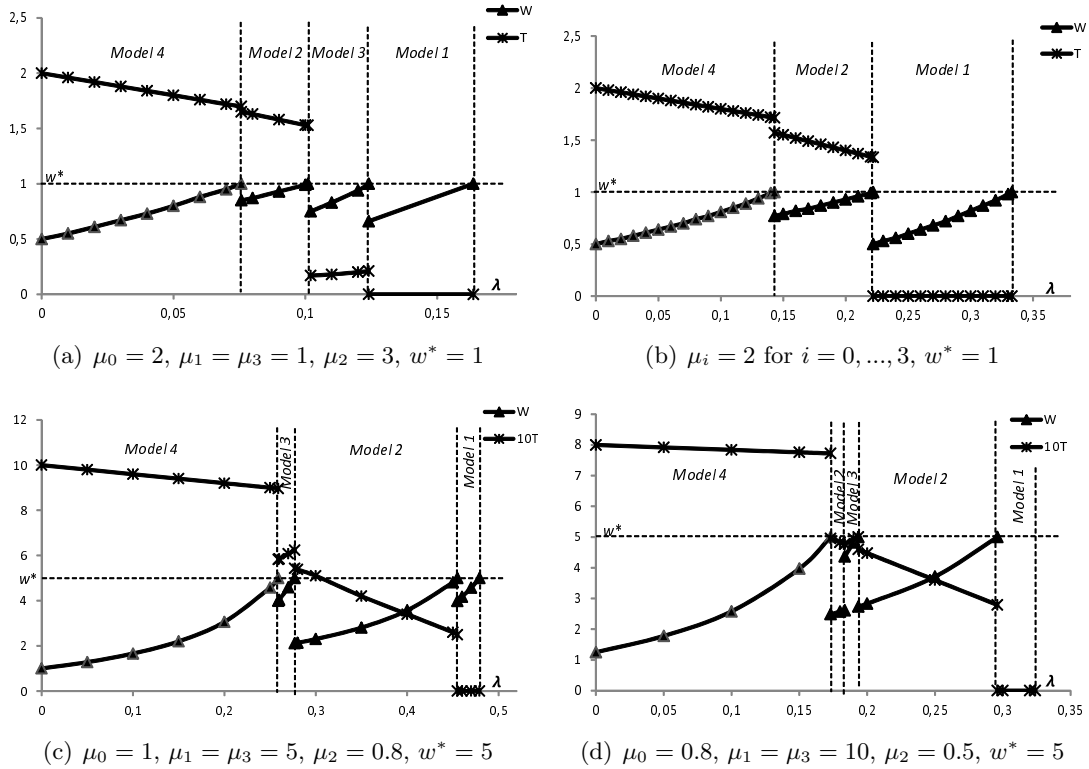
Under the condition of stability of Model $i$, $E(W_i)$ is continuous and strictly increasing in $\lambda$, for $i = 1, ..., 4$. The constraint $E(W_i) \leq w^*$ is then equivalent to $\lambda \leq \overline{\lambda}_i$, for $i = 1, ..., 4$, where the $\overline{\lambda}_i$ can be easily computed. For a given $\lambda$ 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. We have explicitly identified the conditions under which an optimal choice of Model 2 or Model 3 may happen.

We numerically illustrate the analysis above. For various system parameters, Figure II.5.3 gives the optimal model choice as a function of the mean arrival rate of calls, $\lambda$. An intuitive reasoning of a manager would choose the ordering Model 4 (outbound jobs between calls and inside a call), then 2 (outbound jobs only between calls), then 3 (outbound jobs only inside a call), then 1 (no outbound jobs) as $\lambda$ increases.

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 $\lambda$ if $R \leq \overline{\lambda}_4$ and $\overline{\lambda}_3 < \overline{\lambda}_2$, see Figure II.5.3(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 $\lambda$ increases, is Model 2, then Model 3, then again Model 2 (see Figure II.5.3(d)) which happens for system parameters such that $\overline{\lambda}_4 < R < \overline{\lambda}_3 < \overline{\lambda}_2$.

**Optimization of Model PM:** We are interested in the optimization of the parameters $p$ and $q$ in Model PM for Problem (II.5.2). Concretely, we want to find the optimal routing parameters of Model PM that allows the manager to maximize the outbound job expected throughput while respecting a call service level constraint.

From the expression of the outbound job expected throughput $T$ for Model PM, 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 is maximized (worst) for $p = q = 1$. Therefore in order to solve Problem (II.5.2), one would be interested analyzing the sensitivity of $T$ with respect to $p$ and $q$. We have proved that $\frac{\partial T}{\partial p} > \frac{\partial T}{\partial q}$ if and only




(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 II.5.3: Comparison between Models 1 to 4 with a constraint on $E(W)$

if $0 < \rho_0 < \overline{\rho}_0$, where $\overline{\rho}_0$ can be explicitly expressed. In what follows we prove that the optimal policy is such that $p \in \{0, 1\}$ or $q \in \{0, 1\}$.

**Theorem II.5.1** *For $p, q \in [0, 1]$, the optimal values of $p$ and $q$ of the optimization problem*

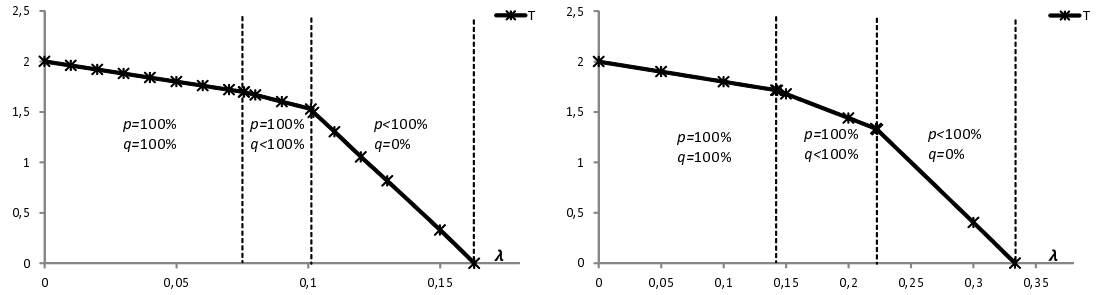
$$\begin{cases} \textit{Maximize } T \\ \textit{subject to } E(W) \leq w^*, \end{cases} \tag{II.5.4}$$

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

Figure II.5.4 provides a numerical illustration of Theorem II.5.1. 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.

## 4.6 Multi-Server Case

An exact analysis as that done for the single server case is too complex. We use simulation to optimize the $(p, q)$ couple. The results provide a numerical evidence that Theorem II.5.1 still holds for $s > 1$. 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. While increasing the workload, we again observe that the choice is first for $p = q = 1$; then $p = 1$ and $0 < q < 1$; then $0 < p < 1$ and $q = 0$. Thus the two questions




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

Figure II.5.4: Optimal $p$ and $q$ with $w^* = 1$

of routing outbound jobs (between calls or during the break) are not considered together at the same time. Simulation results also reveal that the interval of workload values for which the solution $p = q = 1$ answers Problem (II.5.2) enlarges in the system size. The explanation is related to the pooling effect. The larger is the system, the better are the performance for inbound calls. We then may profit from the two opportunities for the routing of outbound jobs (inside and between calls).

One can make use of the light traffic approximation to address the routing optimization problem. Under the light traffic regime, the presence of calls in the system can be neglected. The parameter $q$ does not thus impact the results. The only parameter to focus on for Problem (II.5.2) is $p$. For a choice limited to the extreme cases, we should choose Model 4 if $\frac{1}{s\mu_0} \leq w^*$ (or $1 - e^{-\mu_0 sAWT} \geq SL$). Otherwise Model 3 is the best. The optimal value of $p$ with the constraint $P(W < AWT) \geq SL$ is $p = SLe^{s\mu_0 AWT}$. The optimal expected throughput is then $(s-1)\mu_0 + \mu_0 SLe^{s\mu_0 AWT}$. The optimal value of $p$ with a the constraint $E(W) \leq w^*$ is $p = s\mu_0 w^*$. The optimal expected throughput is then $(s-1)\mu_0 + s\mu_0^2 w^*$.

We also use Heavy Traffic approximations. The numerical experiments show that simulation results converge to the approximate ones as the workload increases ($q$ increases). The only parameter to consider here is $q$. A simple analytical analysis, as that under a light traffic regime, is not possible. One can only then numerically optimize the parameter $q$. For a choice limited to the extreme cases as the workload increases, we should first choose Model 4 then Model 2.

## 5 Concluding Remarks and Future research

I described in this chapter my contributions to the literature on job routing in single and multi-channel call centers. We first considered a two-class call center and developed online scheduling policies subject to satisfying a target ratio constraint of the abandonment probabilities of the two customer classes. This new formulation of the control problem is robust with respect to the system workload. The analysis focused on a given period of the day. In a future research, we would like to focus on many intervals of a day. Then it would be interesting to find a method that translates the whole day objective into a set of objectives per period of the day. It would

be also interesting to extend the analysis to more than two customer classes and agents with different skill sets.

We then considered call centers with inbound calls and an infinite supply of emails. We proposed a scheduling policy, refereed to as ATP, where the objective is to do as much emails as possible while satisfying a service level constraint on the call waiting time. 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, an analytical analysis for the adaptive blending might be useful to better understand ATP. Second, it would be interesting to account for different types of inbound calls but this would considerably complicate the analysis.

Finally, we focused on the analysis of a blended call center with calls and outbound jobs. The call conversation is characterized by a natural break. We focused on the optimization of the outbound job routing given that calls have a non-preemptive priority over outbound jobs. Our objective was to maximize the expected throughput of outbound jobs subject to a constraint on the call waiting time. 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 to extend the analysis to cases with an additional channel, such as chat. 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.

# Part III

# Analysis of Stochastic Processes

This part of the dissertation focuses on my theoretical contributions to the analysis of stochastic processes. The story about my interest for such results is as follows. Let me first recall that my major research interests are in operations management issues for service systems in general, and call centers and healthcare in particular. My approach relies on first developing stochastic models and then analyzing them using stochastic methods. While working on these issues, I was with my colleagues regularly brought to the literature on stochastic processes in order to apply or adapt some existing results. For some situations, we have noticed that the specific result we need does not exist. For other situations, we do find the result we need, but, we also get ideas about extending it to more general settings. Both situation types motivated us to investigate new results for the analysis of stochastic processes that could be applied for a wide range of situations, and not necessarily only related to our specific OM issue.

My research results are mainly related to the analysis of queueing systems and Markov chains. My contributions on the analysis of these two stochastic models can be divided into three parts. The first part, described in Chapter III.1, deals with the computation of first passage times in birth-death processes. A birth-death process is a special case of a Markov chain where the state transitions are of only two types: either a birth which increases the state by one, or a death which decreases the state by one. The second part of my contributions, described in Chapter III.2, focuses on queueing systems with impatient customers. We investigate the computation of various performance measures related to queueing delays. We also investigate their monotonicity properties as a function of the system parameters. Finally, Chapter III.3 describes my contributions related to the dynamic control of queueing systems using Markov decision processes. A summary of my contributions is depicted in Figure 5.

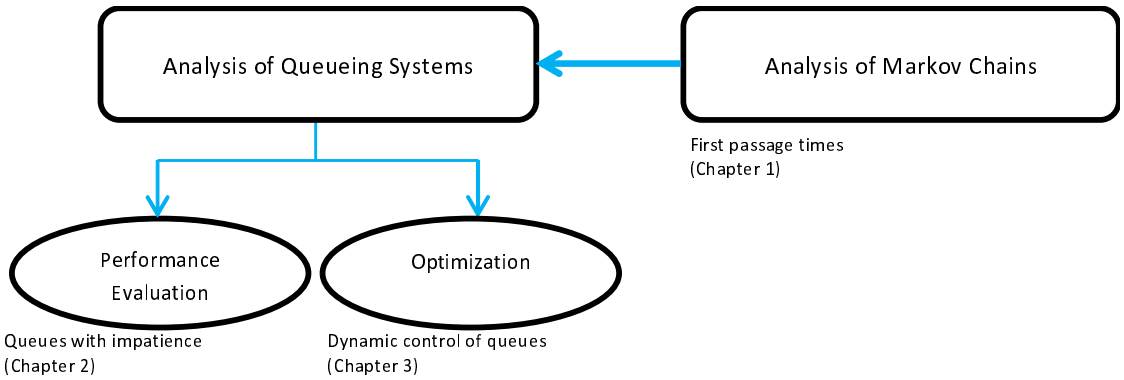



Figure 5: My contributions to the literature on stochastic processes

# Chapter III.1

# Analysis of Markov Chains

## 1 Introduction

This chapter summarizes my contributions to the literature on the computation of passage times in Markov chains. Markov chains, are broadly used in the field of queueing theory. They are a rich and important class in modeling numerous phenomena. My contributions deal with the transient analysis of birth-death processes. They can be divided into two parts. A first part focuses on the characterization of ordinary and conditional first passage times in general birth-death processes. A second part focuses on the computation of the cumulative distribution function of the sum of Erlang random variables with arbitrary parameters. This summation can be modeled as a first passage time in a pure death process.

The first part is the subject of Section 2. Under existence conditions, we derive closed-form expressions for any moment of ordinary and conditional first passage times. We also give an explicit condition for a birth-death process to be ergodic degree 3. These results are useful for the computation of the moments of busy and idle periods for non-standard Markovian queues. They are also helpful for the computation of state-dependent queueing delays.

The second part of the results, described in Section 3, were developed during the postdoc of Benjamin Legros. We propose a matrix analysis approach for the characterization of the summation of the Erlang random variables. This reduces to the computation of the exponential of the involved generator matrix. We propose a particular basis of vectors in which we write the generator matrix. We find, in the new basis, a Jordan-Chevalley decomposition allowing to simplify the computation of the exponential of the generator matrix. This is a simpler alternative approach to the existing ones in the literature, where the complex computation of high order derivatives or integrals may arise. The results are of value for the analysis of service systems where the involved processes can be modeled as a succession of stages in series, for the analysis of the reliability of systems with exponentially distributed components lifetimes, just to name a few.

# 2 Computation of First Passage Times

## 2.1 Positioning and Contributions

We consider in this work the transient behavior of general birth-death processes, and its application to time-dependent solutions for queueing systems. We mean by "general" that transition rates are arbitrary and need not to satisfy some special structure.

The literature related to birth-death processes is extensive and growing. References include Karlin and McGregor (1957a); Keilson (1979, 1981); Sumita (1984); Mao (2004); Guillemin (2005); Coolen-Schrijner and van Doorn (2002). We also refer the reader to Keilson (1964a), Keilson (1964b), and Kijima (1997) for an overview on the subject. Guillemin and Pinchon (1999) revisit the solving of the forward Chapman-Kolmogorov equations associated with a birth-death process through the spectral theory. Their work is based on the connection between probability theory and continued fractions addressed first by Karlin and McGregor (1957a). They investigated, specifically, how Laplace Transforms of different transient characteristics related to excursions in a general birth-death process can be expressed by means of the basic orthogonal polynomials system and the spectral measure. Flajolet and Guillemin (2000) develop a formal calculus of basic events described by lattice paths associated with birth-death processes. They express several basic events in terms of continued fractions and their associated orthogonal polynomials. An extension is developed by Ball and Stefanov (2001), where the authors use an approach based on viewing birth-death processes as exponential families.

Using Chapman-Kolmogorov equations and via Laplace transforms, we derive closed-form expressions for the moments of ordinary and conditional downcrossing and upcrossing times between pairs of states. Some of our results about the ordinary downcrossing and upcrossing times are already derived in the literature using a different approach. The existing results lead to a representation of Laplace transforms of transient characteristics in terms of continued fractions and orthogonal polynomials. Although continued fractions are known to be useful especially for numerical issues, few of their closed-form expressions are available. The known expressions deal with simple models such as the M/M/1 and M/M/$\infty$ queues. The analysis in this work allows us however to address several further applications. We show the equivalence between the analysis of various characteristics in some examples of queueing systems and that of hitting and return times. Also, we recover in a simple way some classical results such as the busy period and busy cycle durations in some standard Markovian systems.

## 2.2 Model Description and Notations

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

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

$$q_{m,m+1} = \lambda_m > 0,\ q_{m,m-1} = \mu_m,\ q_{m,m} = -(\lambda_m + \mu_m) \text{ for } m \geq 0, \text{ and } q_{m,n} = 0 \text{ otherwise.} \tag{III.1.1}$$

The rate $\mu_0$ is equal to 0, and $\mu_m > 0$ for $m > 0$. For $m \geq 0$, we define the quantities $\pi_m$ by

$$\pi_0 = 1, \text{ and } \pi_m = \frac{\lambda_0 ... \lambda_{m-1}}{\mu_1 ... \mu_m} \text{ for } m \geq 1. \tag{III.1.2}$$

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

$$p_m = \frac{\pi_m}{\sum_{i=0}^{\infty} \pi_i} > 0, \text{ for } m \geq 0. \tag{III.1.3}$$

### 2.2.1 First Passage Times

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

$$\theta_m = Inf\{t > 0 :\ E(t) = m - 1 \ \mid \ E(0) = m\}. \tag{III.1.4}$$

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

$$\tau_m = Inf\{t > 0 :\ E(t) = m \ \mid \ E(0) = m - 1\}. \tag{III.1.5}$$

Let us discuss their conditions of existence. For the upcrossing time, $\tau_m$, it is clear that no specific conditions are required. However, this is not necessarily the case for the downcrossing time, $\theta_m$. The following set of conditions are required.

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

Roughly speaking, the ergodic degree gives the number of finite moments possessed by the time of the first passage at a given state $i$ starting from any state $j \neq i$. We refer the reader to Mao (2004) for more details. In particular, Condition $C^1$ simply reflects the classical ergodicity assumption: the condition under which the process settles into equilibrium (the birth rates are not too large relative to the death rates). This is equivalent to say that Condition $C^1$ is the necessary and sufficient condition for the expectation of the first passage time from any state $i$ to any state $j \neq i$ to be finite. Conditions $C^1$ and $C^2$ are given by Karlin and McGregor (1957b) and Coolen-Schrijner and van Doorn (2002), respectively. However, no explicit expressions for, $k \geq 3$, exist in the literature. We provide Condition $C^3$ in an explicit form.

The analysis done for first passage times is as follows. We first compute the $k^{th}$ order moment expression of the random variable $\theta_m$. Thereafter, we deduce the expectation and the variance of $\theta_m$. A similar analysis is done for the random variable $\tau_m$. We Notice that the expressions of $E(\theta_m^k)$, for $k \geq 3$, and $E(\tau_m^k)$, for $k \geq 2$, are new.

### 2.2.2 Conditional First Passage Times

In what follows, we define some random variables associated to conditional first passage times. Let ${}^r\theta_m$ be the first passage time of the process $\{E(t),\, t \geq 0\}$ from state $m$ to state $m-1$, given that the process does not visit state $r$ in between, $1 \leq m < r$, defined by

$$ {}^r\theta_m = Inf\{t > 0 :\ E(t) = m-1 \ \mid\ E(0) = m \text{ and no visit to } r\}. \tag{III.1.6}$$

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

$$ {}^r\tau_m = Inf\{t > 0 :\ E(t) = m \ \mid\ E(0) = m-1 \text{ and no visit to } r\}. \tag{III.1.7}$$

One may also define the conditional downcrossing and upcrossing times between two different states, given no visit to a third state. Our results allow to cover such analysis.

To the contrary to ordinary first passage times, the conditional ones are not very known. However, one may find several situations where these conditional random variables are useful. One interesting application would be for a make-to-stock system. Consider an inventory system with finite capacity ($K_1$ items at most) in which demands are backlogged if no items are available for them upon arrival. There is a single machine with a finite queue size. A maximum of $K_2$ customers can be accepted in queue. A customer who finds a full waiting line is lost. In case of Markovian inter-arrival demand and production processing times, this system can be modeled as a birth-death process. In practice, it is useful to determine the time from an idle system (with no items in stock and no waiting customers) until a full inventory (with $K_1$ items in stock) given no backlogged demands in between. This is equivalent to compute the conditional first passage time of a particle (in the associated birth-death process) from the "idle" state up to the "full" inventory state, given that it does not visit the state with one backlogged customer. The latter state is the one just before the "idle" state. Another interesting performance measure is to compute the time from a full waiting line ($K_2$ waiting customers) until all customers are served given no lost customers. This is again equivalent to compute a conditional first passage time in the associated birth death process. We notice in addition that evaluating these performances would allow to optimize the system parameters such as the inventory capacity, $K_1$, and the waiting line size, $K_2$.

We compute the $k^{th}$ order moment, $k \geq 1$, of the conditional first passage times. The results are new except for a special case for ${}^r\tau_m$. Note that no existence conditions are required for the computation of their moments. We need to introduce some notations. These preliminaries are

specifically related to the notion of ruin probabilities. Consider again the birth-death process defined in Section 2.2. Let ${}^r\eta_m$ be the ruin probability that the particle, starting at $m$, reaches $m-1$ first before $r$, $1 \leq m < r$. It is clear that the ruin probability ${}^r\eta_{r-1}$ to reach $r-2$ starting at $r-1$, without visiting $r$, is given by $\frac{\mu_{r-1}}{\lambda_{r-1}+\mu_{r-1}}$. We also have the following recursive relation ${}^r\eta_m = \frac{\mu_m}{\mu_m+\lambda_m(1-{}^r\eta_{m+1})}$, for $1 \leq m < r-1$, starting with ${}^r\eta_{r-1} = \frac{\mu_{r-1}}{\lambda_{r-1}+\mu_{r-1}}$. The above ruin probabilities are useful for the computation of the moments of ${}^r\theta_m$.

For ${}^r\tau_m$, as above, we first introduce some notations. Let ${}^r\nu_m$ be the ruin probability that the process, starting at $m-1$, reaches $m$ first before $r$, $m \geq r+2$. It is clear that ${}^r\nu_{r+2} = \frac{\lambda_{r+1}}{\lambda_{r+1}+\mu_{r+1}}$. With a similar explanation as for the ruin probability ${}^r\eta_m$, we give the following recursive relation, for $m > r+2$, ${}^r\nu_m = \frac{\lambda_{m-1}}{\lambda_{m-1}+\mu_{m-1}(1-{}^r\nu_{m-1})}$. Again the results for ${}^r\tau_m$ are given as a function of the ruin probabilities.

### 2.3 Applications

The results obtained above have various applications for the analysis of queueing systems. One application is the computation of busy periods, idle periods and busy cycles for single and multi-server queueing systems. Another application is the computation of state-dependent queueing delays for non-standard queueing systems, such as multi-class priority systems, systems with impatient customers (linear growth death rates), or state-dependent arrival rates, or in general, systems with state-dependent transition rates.

The analysis leads to computing the moments of the random variables. These moments are in turn helpful for the characterization of the probability distributions. Having only the first and second moments is important but often not sufficiently accurate to approximate an exact distribution. For instance, the third moment of a random variable allows to compute its skewness, and the fourth moment allows to compute its kurtosis. The skewness is a measure of the "asymmetry" of a probability distribution, and the kurtosis is a measure of its "peakedness". For further details about these notions, we refer the reader to Joanes and Gill (1998). In general, such an analysis is related to the well known Moment Problem. The Moment Problem is the problem of finding a distribution whose moments have specified values, or of determining whether such a distribution exists. This area was first started by the works of Chebyshev through the well known Chebyshev inequalities. A literature and historical perspective is given in Bertsimas and Popescu (2005).

## 3 Sums of Erlang Random Variables

### 3.1 Introduction

Many situations in service and manufacturing service systems involves the computation of the sum of independent exponential random variables. Examples include healthcare or production systems with different stages in series, system reliability with exponentially distributed components lifetimes, and wireless mobile systems with cooperative diversity schemes. This summation

arises also in the transient analysis of Markovian queueing systems, and in general, semi-Markov processes.

We consider the general case of a hypoexponential distribution defined as the sum of $n$ independent Erlang distributions, for $n \in \mathbb{N}$. An Erlang distribution is defined by two parameters, a number of i.i.d. exponential stages and a rate per stage. Thus, the general hypoexponential distribution is completely defined by the couples of parameters $(\lambda_i,\ k_i)$ for $i = 1, ..., n$. Each couple $(\lambda_i,\ k_i)$ defines an Erlang distribution ($\lambda_i \in \mathbb{R}$, $k_i \in \mathbb{N}$), and the rates $\lambda_i$ for $i = 1, ..., n$ are all distinct. We denote by $K_i = k_1 + k_2 + ... + k_i$ for $i = 1, ..., n$ and use the convention $K_0 = k_0 = 0$. The cumulative distribution function (cdf) of the hypoexponential distribution is then given by

$$F(x) = 1 - \boldsymbol{\alpha} e^{xM} \mathbf{1}, \tag{III.1.8}$$

for $x \geq 0$, where $\mathbf{1}$ is a column vector of size $K_n$ with ones everywhere, $\boldsymbol{\alpha}$ is a line vector of size $K_n$ and is given by $\boldsymbol{\alpha} = (1, 0, \ldots, 0)$, and $e^{(.)}$ denotes the exponential operator. The generator square matrix $M$ of size $K_n \times K_n$ is defined by the coefficients $m_{i,j}$ for $i, j \in \{1, ..., K_n\}$. We have $m_{j,j} = -\lambda_i$ and $m_{j,j+1} = \lambda_i$, for $K_{i-1} + 1 \leq j \leq K_i$ and $i = 1, ..., n$. All remaining coefficients of $M$ are zero.

Scheuer (1988) provides a formula for $F(.)$ that involves high order derivatives of products of multiple functions. The formula is however hard to compute numerically. Amari and Misra (1997) proposes a simplification of Scheuer (1988)'s formula using Laplace transforms and multi-function generalization of the Lebnitz rule for higher order derivatives of products of two functions. For a particular case with constraints on the values of the $\lambda_i$s, van Khuong and Kong (2006) provide the probability distribution function by inverting its Fourier transform. Using the Wilk's integral representation of the distribution of the product of independent beta random variables, Favaro and Walker (2010) provides an alternative formula for $F(.)$. We also refer the reader for more details to the review by Nadarajah (2008).

We propose an alternative simple approach to compute the cdf of $F(.)$. The approach is based on a linear algebraic matrix analysis, which avoids the numerical complexities that may arise in the computation of high order derivatives or integrals. The structure of the approach is as follows. We first obtain some particular eigenvectors of the generator matrix $M$. These are next used to construct a new basis of vectors. The new basis allows to find the Jordan-Chevalley decomposition of $M$ into a sum of two commutative linear operators, a diagonal one and a nilpotent one. The exponential of the matrix $M$ then simply follows by inverting the new basis matrix using the Cayley-Hamilton theorem, which leads to the cdf of $F(.)$.

## 3.2 Computation

Lemma III.1.1 provides the eigenvalues of the matrix $M$, and one eigenvector associated to each eigenvalue.

**Lemma III.1.1** *The eigenvalues of $M$ are $-\lambda_i$ for $i = 1, ..., n$. An eigenvector of size $K_n$ associated to $-\lambda_i$ is the column vector $u_i$, where the coefficients of $u_i$, denoted by $u_{i,l}$ for $1 \leq l \leq K_n$, are given by*

$$\begin{cases} u_{i,l} = 1 & , l = K_{i-1} + 1, \\ u_{i,l} = 0 & , l > K_{i-1} + 1, \\ u_{i,l} = \left(\frac{\lambda_{i-1}}{\lambda_{i-1}-\lambda_i}\right)^{K_{i-1}-l+1} & , K_{i-2} + 1 \leq l \leq K_{i-1}, \\ u_{i,l} = \prod_{j=1}^{i-(m+1)} \left(\frac{\lambda_{i-j}}{\lambda_{i-j}-\lambda_i}\right)^{k_{i-j}} \left(\frac{\lambda_m}{\lambda_m-\lambda_i}\right)^{K_m-l+1} & , K_{m-1} + 1 \leq l \leq K_m \text{ and } 0 \leq m < i-1. \end{cases}$$

Let us denote by $\mathcal{B}$ the standard basis composed by the family of column vectors $e_l$, for $1 \leq l \leq K_n$. The coefficients of $e_l$ $(1 \leq l \leq K_n)$ are all zero except the coefficient in line $l$ which is equal to one. Consider now a new family of vectors denoted by $\mathcal{B}'$ and composed by the vectors $e'_l$ for $1 \leq l \leq K_n$, where $e'_{K_{i-1}+1} = u_i$, and $e'_l = e_l$ for $l \neq K_{i-1} + 1$ and $i = 1, ...n$. In Theorem III.1.1, we prove that $\mathcal{B}'$ is a basis.

**Theorem III.1.1** *The family of vectors $B'$ is a basis.*

Let us now proceed to a change of basis from $\mathcal{B}$ to $\mathcal{B}'$. We want to write $M$ in the new basis $\mathcal{B}'$, which leads to a matrix denoted by $M'$. We denote by $P$ the new basis matrix allowing to move from $\mathcal{B}$ to basis $\mathcal{B}'$. This means that $P$ is given by the vectors of the old basis $\mathcal{B}$ but written in the new basis $\mathcal{B}'$. We have $M = PM'P^{-1}$. The matrix $P^{-1}$ is computed using the Cayley-Hamilton theorem. We then prove that $M' = D + N$, where $N$ is nilpotent matrix and $D$ is a diagonal matrix. Since $D$ commutes with any other matrix, in particular $N$, the decomposition of $M'$ into $M' = N + D$ is the unique Jordan-Chevalley decomposition of $M'$ into a summation of two commutative matrices, a nilpotent one and a diagonalisable one. Using the new basis matrix $P$, we have $M = PM'P^{-1} = P(N + D)P^{-1}$, so, $e^{xM} = Pe^{x(N+D)}P^{-1}$ (Gourdon, 1994), for $x \geq 0$. Since $N$ and $D$ commute, we deduce that $e^{x(N+D)} = e^{xD} \times e^{xN}$. The computation of $e^{xD}$ is simple. Also, the computation of $e^{xN}$ has no difficulty. It is done within a finite number of summations, because $N$ is nilpotent. The cdf of the hypoexponential distribution follows then from the coefficients of the first line of the matrix $e^{xM}$, for $x \geq 0$.

# 4 Concluding Remarks and Future Research

We focused on the transient behavior analysis of a general birth-death process. We gave closed-form expressions for the moments of important state-dependent characteristics. The characteristics deal with the random variables of ordinary and conditional first passage times. We derived several new expressions of the moments of the defined hitting and return times. We also discussed the condition under which a birth-death process is said to be ergodic degree $k$. In particular, we gave a new explicit expression for the condition of ergodicity degree 3.

Several further applications could be also possible. For instance, deriving the stationary waiting time moments for some Markovian model where the arrival rate depend on the system

state. Concretely, for example in a system where a new customer has a state-dependent probability to join the queue. To do so, we may compute the state-dependent waiting times as shown in this work. Thereafter, we derive the desired stationary $k^{th}$ order moment of queueing delays, by averaging on all states seen by arrivals. It would be also interesting in practice to investigate approximations or numerical methods to further simplify the computation effort.

# Chapter III.2

# Analysis of Queueing Systems with Impatience

## 1 Context and Contributions

This chapter synthesizes my contributions to the literature on the analysis of queueing systems. The literature dealing with the study of queueing systems is huge. The analytical studies were intended to obtain useful information for the decision making process, basically related to the design, the control, and the measurement of effectiveness of the systems.

My contributions concern in particular the performance analysis of queueing system with impatient customers. Impatience (abandonment) is an important feature for a wide variety of situations that may be encountered in practice. Examples include telecommunication systems, manufacturing systems with perishable items, and service systems such as call centers and health care systems. Theoretical models incorporating abandonment are therefore necessary to obtain more accurate analysis.

My contributions here can be divided into three parts as follows. In the first part, which is a joint work with the postdoc Benjamin Legros, we consider the single type queue M/M/s+GI, with generally distributed abandonment times. We extend the existing results by proposing a controlled approximation of the performance measures that could be as accurate as preferred. Our approach is based on Riemann integration. It consists of approximating the hazard rate function of the patience distribution by a step function, i.e., a finite linear combination of indicator functions. The step function can be chosen as close as preferred to the real hazard rate function.

The results of the second part are obtained under a collaboration with the PhD student Alex Roubos. We consider Markovian multi-server queues with two types of impatient customers: high- and low-priority ones. The first type of customers has a non-preemptive priority over the other type. We consider two cases where the discipline of service within each customer type is FCFS or LCFS. For each type of customers, we characterize various performance measures related to queueing delays: unconditional waiting times, and conditional waiting times given

service and given abandonment.

In the third part, we consider a queueing system with limited buffer size and impatient customers. We investigate monotonicity properties of first and second order of the probability of service with respect to the buffer size. Under the stationary regime, we prove that our service level is strictly increasing and concave in the buffer size, whereas we prove under the transient regime that it is only increasing. Such results are helpful, in general, for the system understanding, and in particular, they are useful for the system design, i.e., the optimal choice of the system parameters.

## 2 Related Literature

The literature on queueing models with abandonments focuses especially on performance evaluation. A number of approximations for the probability to abandon are developed by Boxma et al. (1994). The authors have considered a multi-server queue with generally distributed service times and patience times. The impact of the patience distribution on performance is studied by Mandelbaum and Zeltyn (2004). They observe an approximate linearity between the abandonment probability and the average waiting time. To analyze multi-server queues with generally distributed service times and patience times, Whitt (2005a) develops an algorithm to compute approximations for standard steady-state performance measures. One of his conclusions is that the behavior of the patience distribution near the origin primarily affects the performance. Iravani and Balcıoğlu (2008) propose two approximations that are based on scaling the single-server queue to obtain estimates for the waiting-time distributions. Other papers have treated the impatience phenomenon under various assumptions. Related studies include those by Baccelli and Hebuterne (1981), Altman and Borovkov (1997), Ward and Glynn (2003), Jouini (2012).

Although the two features of abandonment and priority have each received attention separately, there is limited literature that deals with both of them. We refer the reader to Choi et al. (2001), where the authors derive several performance measures for an M/M/1 queue with two types of impatient customers in which type 1 customers have impatience of constant duration, and type 2 customers have no impatience and low priority level. An extension of the latter model is addressed by Brandt and Brandt (2005) for general distributed patience times. Wang (2004) considers a single-server non-preemptive priority queue with two classes of impatient customers. He proposes an approximation for the probability to have an idle server, which allows to compute the expected values of the queue lengths and the unconditional waiting times. Rozenshmidt (2007) considers a similar model to ours (under FCFS) and derives expressions for the unconditional expected waiting times of all types. Here we extend that analysis by considering additional performance measures, by considering also LCFS, and by computing all moments of the random variables.

# 3 General Abandonments

## 3.1 Introduction

We focus on the performance analysis of the M/M/s+GI queue where "+GI" indicates generally and identically distributed customer abandonment times. We consider an exponential distribution for service times. Although it has been shown in the literature that service times are in general not exponentially distributed (Mandelbaum and Schwartz, 2002; Brown et al., 2005), this Markovian assumption leads to appropriate approximate performance measures for the multi-server setting. Whitt (1993) shows that the dependence of the performance on the service time distribution reduces with the number of servers. For large systems, the limited impact of the service time distribution on performance in an M/G/s queue is also shown through simulation by Mandelbaum and Schwartz (2002). In the call center context, Whitt (2005b) demonstrates the efficiency of this approximation for an M/GI/s+GI. For more details, we refer the reader to Section 5 in Whitt (2005b).

In this work, we extend the existing results by proposing a controlled approximation to evaluate the performance measures of the M/M/s+GI queue. We propose closed-form expressions that are functions of the well known and tabulated gamma function. We mean by "controlled" an approximation that can be designed to reach the exact results as accurate as preferred. The method consists of approximating the hazard rate function of the patience distribution by a step function, i.e., a finite linear combination of indicator functions. The considered step function can be chosen as close as preferred to the real hazard rate function. This result is based on Riemann integration (Rudin, 1987). It is proven in this theory that any non-negative Riemann integrable function is the point-wise limit of a monotonic increasing sequence of non-negative step functions. Using this approach, we derive closed-form expressions for the building blocks in terms of integrals given in Baccelli and Hebuterne (1981). These building blocks then directly lead to the performance measures as shown in Baccelli and Hebuterne (1981) and summarized by Zeltyn and Mandelbaum (2005).

## 3.2 The Result

We consider an M/M/s+GI queue with a single type of customers. The arrival process is Poisson with rate $\lambda$. Service times are i.i.d. and exponentially distributed with rate $\mu$. There are $s$ parallel, identical servers. Customers are served in the order of their arrivals. We assume that the hazard rate function of the patience is known, denoted by the positive function $h(t)$, for $t \geq 0$.

From Rudin (1987), the function $h$ can be approximated as close as preferred by a step function. We consider the step function $h_n$, defined by $h_n = \sum_{k=1}^{n} \alpha_k \mathbb{I}_{[t_{k-1},t_k)}$, with the parameters $0 = t_0 < t_1 < \cdots < t_n$, $\alpha_1, \alpha_2, \cdots, \alpha_n \in \mathbb{R}^+$; $n$ a strictly positive integer; and $\mathbb{I}_A$ an indicator function of a given set $A$. Note that we can choose $t_n$ as high as preferred such that the number of customers who experience a waiting time higher that $t_n$ is negligible, or even $t_n = \infty$ if the

patience behavior is exponential beyond a given threshold ($t_{n-1}$).

From now we assume that the hazard rate function of the patience is $h_n$. The value of having a step function hazard rate is that the patience distribution is exponential with a time-dependent parameter. This time-dependent parameter is constant and equals to $\alpha_k$ on each interval $[t_{k-1}, t_k)$, for $k = 1, 2, \cdots, n$. For $t \geq t_n$, since the hazard rate function is zero, the patience is infinite. We denote by $X$ the random variable measuring the customer patience times. Consider the cumulative distribution function, $G(t) = P(X < t)$ ($1 - G(t)$ is the survival function). For $t \in [t_{k-1}, t_k)$ and $1 \leq k \leq n$, we have

$$G(t) = 1 - e^{-\alpha_k(t-t_{k-1}) - \sum_{i=1}^{k-1} \alpha_i (t_i - t_{i-1})}. \tag{III.2.1}$$

For $t \geq t_n$, similarly we may write $G(t) = 1 - e^{-\sum_{i=1}^{n} \alpha_i (t_i - t_{i-1})}$. Recall from Baccelli and Hebuterne (1981) that the performance measures of the M/M/s+GI queue are functions of the building blocks defined by $H(x) = \int_0^x (1 - G(t))\,\mathrm{d}t$; $J(t) = \int_t^\infty \exp(\lambda H(x) - s\mu x)\,\mathrm{d}x$; $J_1(t) = \int_t^\infty x.\exp(\lambda H(x) - s\mu x)\,\mathrm{d}x$; and $J_H(t) = \int_t^\infty H(x).\exp(\lambda H(x) - s\mu x)\,\mathrm{d}x$.

The computation of $H(x)$ is easy. Our contribution is the computation of the expressions of the building blocks $J(t)$, $J_1(t)$ and $J_H(t)$, as a function of the incomplete gamma function defined by $\gamma(x, y) = \int_0^y t^{x-1} e^{-t}\,\mathrm{d}t$, and its derivative $\gamma'(x, y) = \frac{\partial \gamma(x,y)}{\partial x} = \int_0^y \ln(t) t^{x-1} e^{-t}\,\mathrm{d}t$. The approach is as follows. First, we divide the integration interval $[t, \infty)$ into the intervals $[t, t_k)$, $[t_k, t_{k+1})$, $\cdots$, $[t_n, \infty)$ (for $1 \leq k \leq n$ and $t_{k-1} \leq t < t_k$). Next, we compute the integrals on each interval using variable substitution. A special attention is given to the last interval $[t_n, \infty)$, since the expression of $H(t)$ is a constant on this interval. Also, there is one more step for $J_1(t)$. It consists of computing the integral $\int_{t_n}^\infty x.\exp(\lambda H(x) - s\mu x)\,\mathrm{d}x$ using integration by parts.

When the hazard rate function $h$ is known but can not be easily identified as a step function as for the exponential or deterministic cases, a useful method (Rudin, 1987) to choose a step function sufficiently close to the hazard rate function is as follows. If we define the intervals $A_{n,k} = \left[\frac{k-1}{2^n}, \frac{k}{2^n}\right)$ for $n \geq 0$ and $k = 1, 2, \cdots, 2^{2n}$, and $h_n = \sum_{k=1}^{2^{2n}} h\left(\frac{k-1}{2^n}\right) \mathbb{I}_{A_{n,k}}$, then as $n$ goes to infinity $h_n$ converges point-wise to $h$.

As previously mentioned in Chapter II.1, the knowledge of the patience is often only given through observed data. From observed customers we know the minimum between the customer patience time and the customer virtual waiting time, and we also know which one we observe. This is called right-censored data. Techniques exist to deal with censored data, one of which is the Kaplan-Meier estimator (Kaplan and Meier, 1958). With this technique we can estimate the hazard rate function empirically. Thus, we can numerically build an appropriate step function to approach the estimated hazard rate function.

# 4 Multiple Priority

## 4.1 Context and Contributions

We analyze queueing systems with multiple types of impatient customers. As already explained in previous chapters, customer abandonment is an important feature in a wide variety of situations. Another important feature in practice is the differentiation in the service given to different customer types. A priority mechanism is a useful scheduling method that allows different customer types to receive differentiated performance levels. Priority queueing comes up in many applications such as communication networks with differentiated services, call centers with VIP and less important customers, and more. Priority schemes are additionally known for their ease of implementation, explaining their prevalence in practice.

We consider a Markovian multi-server queueing system with two types of impatient customers: high- and low-priority ones. The high-priority type has non-preemptive priority over the other type. We assume common exponential distributions for service times as well as patience times for both customer types. We analyze two different systems by considering different disciplines of service within each queue. The most common discipline that can be observed in everyday life is FCFS. Some other in common usage are random order of service (ROS) and last-come first-served (LCFS), which is applicable to many inventory systems when it is easier to reach the nearest stored items which are the last in. In this work, we consider FCFS and LCFS policies and derive various performance measures related to queueing delays. Our approach is based on the use of Laplace-Stieltjes transforms and on the characterization of the virtual waiting time of a "virtual" infinitely patient customer.

Our main contributions can be summarized as follows.

- We compute the Laplace-Stieltjes transforms of various random variables related to queueing delays: unconditional waiting times, and conditional waiting times given service and given abandonment. We do so for both high- and low-priority customers. Our approach is based on the computation of virtual waiting times. One can then easily numerically invert the Laplace-Stieltjes transforms in order to obtain the cumulative distribution functions of these random variables at any point of time (Abate and Whitt, 2006).

- The analysis is given for two different non-preemptive priority models. One where the discipline of service within each class is FCFS, and another one working under LCFS. Moreover, the analysis we develop holds for a priority queue with mixed policies, i.e., FCFS for the first type and LCFS for the second one, and vice versa.

## 4.2 Modeling

Consider a queueing model with two types of customers: important customers denoted by type 1, and less important ones denoted by type 2. The model consists of two infinite-buffer queues for types 1 and 2, and a set of $s$ parallel, identical servers. All servers are able to handle all types

of customers. The system is work conserving. Upon arrival, a customer is addressed by one of the available servers, if any. If not, the customer must join one of the queues. Newly arriving customers of types 1 and 2 are assigned to queues 1 and 2, respectively. Customers of type 1 (waiting in queue 1) have a non-preemptive priority over customers of type 2 (waiting in queue 2). Within each queue, we consider two cases for the discipline of service: FCFS and LCFS. Arrival processes of types 1 and 2 follow a Poisson process with rates $\lambda_1$ and $\lambda_2$, respectively. Successive service times are assumed to be independent and identically distributed (i.i.d.), and follow a common exponential distribution with rate $\mu$ for both customer types.

In addition, we let customers be impatient. Times before abandonment, for both customer types, are assumed to be i.i.d. and exponentially distributed with a common rate denoted by $\gamma$. We describe patience times by the random variable $T$. Finally, retrials are ignored, and abandonment is not allowed once a customer starts service. Following similar arguments, the behavior of the system can be viewed as a two-class M/M/s+M queueing system. The resulting model where the policy for each queue is FCFS (LCFS) is referred to as $\text{Model}_{\text{FCFS}}$ ($\text{Model}_{\text{LCFS}}$). Note that owing to abandonments, $\text{Model}_{\text{FCFS}}$ and $\text{Model}_{\text{LCFS}}$ are unconditionally ergodic.

### 4.3 Results

We denote by $m$ the type of a customer, $m \in \{1, 2\}$. During the stationary regime, we define the following performance measures for $\text{Model}_{\text{FCFS}}$ and $\text{Model}_{\text{LCFS}}$.

- $W$ is the unconditional queueing delay of an arbitrary customer (regardless of her type).
- $W_m$ is the unconditional queueing delay of a type $m$ customer.
- $W_{m,s}$ is the conditional queueing delay of a type $m$ customer, given that she will enter service.
- $P_{m,s}$ is the probability that a type $m$ customer enters service.
- $W_{m,r}$ is the conditional queueing delay of a type $m$ customer, given that she will abandon.
- $P_{m,r}$ is the probability that a type $m$ customer abandons.
- $W_{m,d}$ is the conditional queueing delay of a type $m$ customer, given that she has to wait.
- $P_d$ is the probability of delay, i.e., the probability that a new arrival has to wait. Since $\text{Model}_{\text{FCFS}}$ and $\text{Model}_{\text{LCFS}}$ are work conserving, $P_d$ is independent of the customer type.
- $W_{m,d,s}$ is the conditional queueing delay of a type $m$ customer, given that she was queued and that she will enter service. (We do not define a similar quantity for abandoned customers, since an abandoned customer is necessarily a delayed customer.)
- $P_{m,d,s}$ is the probability that a type $m$ customer waiting in the queue will enter service.

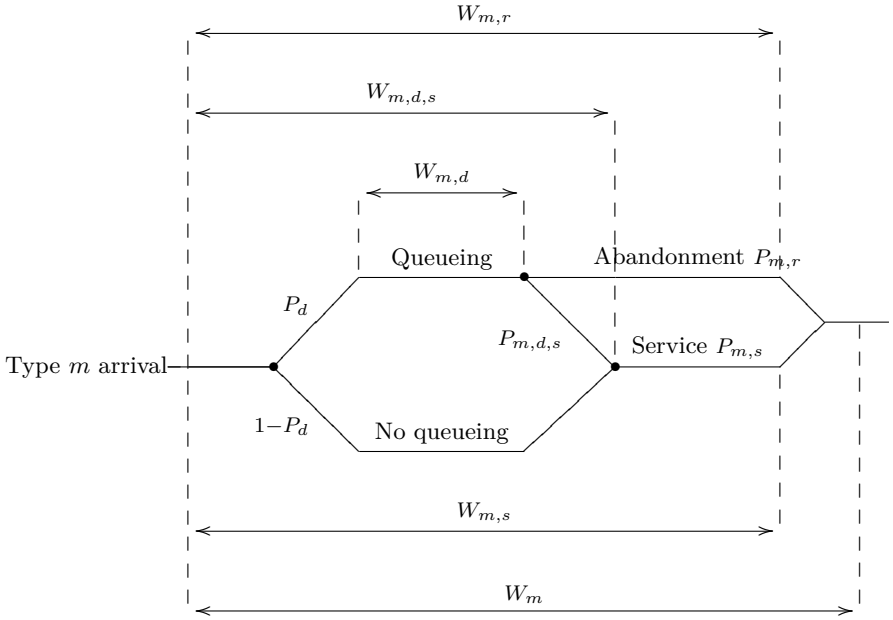



Figure III.2.1: Performance measures for a type $m$ customer.

To clarify the numerous definitions, we depicted in Figure III.2.1 a schema of the performance measures of interest.

The approach to compute the various performance measures is as follows. We start by computing the stationary probability distributions of the system states for $\text{Model}_{\text{FCFS}}$ and $\text{Model}_{\text{LCFS}}$. At a given instant $t$, we denote by $n_1(t)$, $n_2(t)$, and $n(t) = n_1(t) + n_2(t)$ the number of type 1 customers in queue 1, that of type 2 in queue 2, and the total in both queues, respectively. Computing the stationary distribution of the process $\{n_2(t), t \geq 0\}$ or $\{(n_1(t), n_2(t)), t \geq 0\}$ is a complicated task. We only consider the processes $\{n_1(t), t \geq 0\}$ and $\{n(t), t \geq 0\}$ which are sufficient for the derivation of the performance measures. Patience times are memoryless. Thus, as long as the scheduling policy within each queue is work conserving, the number of type 1 customers and type 2 customers in the system remain unchanged. Moreover, since patience as well as service times are identically distributed for both customer types, a work-conserving policy (priority between the queues or not) does not affect the total number of customers in the system.

From the stationary probabilities of the system states, we deduce $P_d$, $P_{m,r}$, $P_{m,s}$ and $P_{m,d,s}$. For the remaining delay performance measures, the method to derive their Laplace-Stieltjes transforms rely on the two following components:

- The computation of an $n$-busy period duration, for $n \geq 0$. For $n \geq 1$, an $n$-busy period is defined as the elapsed time from the arrival of a customer to a busy M/M/s+M system with $n-1$ waiting customers in the queue ($n$ customers in the queue including the new arrival) until the epoch at which one server becomes idle. The 0-busy period reduces to the classical busy-period definition defined to begin with the arrival of a customer to a system with $s-1$ busy servers and to end when again one server becomes idle.

- The computation of the virtual waiting time, i.e., the waiting time of an infinitely patient customer.

Our approach easily extends to that of a mixed model in which we allow the discipline of service in one queue to be different from the one in the other queue. All the expressions for the stationary probabilities hold again for the mixed system. For a given type that is served under FCFS (LCLS), it suffices to apply the same analysis as in the non-mixed system $\text{Model}_{\text{FCFS}}$ ($\text{Model}_{\text{LCFS}}$).

# 5 Monotonicity Properties

## 5.1 Introduction

Monotonicity properties of performance measures are useful for understanding and solving optimization problems of queueing systems. Optimization models are being used increasingly in the design of a variety of systems where queueing phenomena arise. Examples include flexible manufacturing systems, as well as service systems and telecommunications networks. For such problems, it is important to know the convexity properties of the performance measures with respect to the design variables. These properties may enable us to reduce the performance optimization problem to a convex programming problem which is easier to solve. Using a convexity result, Yao and Shanthikumar (1987) accelerate their computation procedure to design a loss queueing system subject to constraints on the loss probability. Koole and Pot (2011) consider an optimization problem for an M/M/s/K+M queue. The objective function is a profit function of the number of servers and the buffer size. They derive some monotonicity properties about the defined performance measure. Based on these properties, they develop a fast algorithm which avoids the research of all possible solutions to get the global optimum.

Several convexity properties about various performance measures have been investigated in the queueing literature. The major performance measures for delay systems are the average waiting time, the average queue length and the probability of delay. Those for pure loss systems include basically the probability for a new arrival to be lost. In general, the loss probability is related to systems involving finite buffers or systems with abandonment. In this work, we consider a queueing system with impatient customers and finite waiting line. The performance measure of interest is the probability for a new arrival customer to enter service, or equivalently, the probability to not be lost. We investigate first and second order monotonicity properties of our performance measure as a function of the queue size, which is useful for the design of a limited buffer size.

## 5.2 Model Formulation

Consider a multi-server queueing system with a single class of customers. The model consists of a set of $s$ parallel, identical servers and a finite queue (waiting line). There is a maximum number of customers that may be simultaneously present, we assume that the system can hold

at most a total of $K$ customers including those in service. Clearly $K \geq s$, and we denote the queue capacity by $k = K - s$, $k \geq 0$. Upon arrival, a customer is addressed by one of the available servers, if any. If not, the customer joins the queue if less than $K$ customers are present in system. If not, the customer is refused entry and departs immediately without service. He is blocked and considered lost. In addition, we assume that customers are impatient. After entering the queue, a customer will wait a random length of time for service to begin. If service has not begun by this time, he will abadon, and again considered to be lost. Finally, retrials are ignored, and abandonment is not allowed once a customer starts his service.

The arrival of customers is assumed to follow a Poisson process. Inter-arrival times are i.i.d. and exponentially distributed with rate $\lambda$. Successive service times are assumed to be i.i.d., independent from the arrival process, and follow an exponential distribution with rate $\mu$. Times before abandonment are assumed to be i.i.d., and exponentially distributed with rate $\gamma$. Following similar arguments, the system can be modeled as an M/M/s/K+M queue. The system is unconditionally ergodic because of abandonment and also its limited capacity. Finally, we need not to specify a scheduling policy for our results, except that it is workconserving.

We focus on characterizing the performance measure of interest. It is defined in terms of the stationary fraction of customers who get service, i.e., the fraction of customers who are not blocked and who do not abandon. We denote this performance as $Q$, and we explicitly compute it using the system state stationary probability and the PASTA property. We also define the probability of being served under the transient regime, $Q(t)$.

## 5.3 Monotonicity Results

One may intuitively state that the performance measures $Q(t)$ and $Q$ increase with respect to the queue capacity $k$, keeping the parameters $\lambda$, $\mu$, $\gamma$ and $s$ constant. The idea is that, although adding more places in the waiting line may increase abandonments, it is clear that it could not deteriorate the performances we consider here. On the contrary, it allows for more customers to enter service. If not, it will at worst achieve an equal fraction of successful departures comparing to a system with less queue capacity. We rigorously prove this result using two different approaches. The first approach is coupling arguments. We prove that the transient and stationary probabilities of service increase in the buffer size $k$. We do so for a more general setting, namely, the GI/M/s/K+M queue. The approach is analytical. For our original Markovian system, we prove that the stationary probability of service increases in $k$.

The main results for the first approach are given next.

**Lemma III.2.1** *Consider a GI/GI/s/K +M queue. Times before abandonment are assumed to be i.i.d. and exponentially distributed. Then, the probability of being served $Q$ is constant for any workconserving non-preemptive scheduling policy.*

Although the probability of being served is independent of the scheduling policy, the mean waiting time in queue for the served customers does depend on the scheduling policy. Jouini

et al. (2010) have proved the latter result when considering the particular case of a GI/GI/s + M queue. They have also characterized the policies under which upper and lower bounds of the mean waiting time are achieved.

We should note however that the result in Lemma (III.2.1) does not hold if times before abandonment are not i.i.d. and exponentially distributed, or if service times at any point during an arbitrary busy period are order of service dependent, we need to assume that no service needs are created or destroyed within the system: no abandonment in the midst of service, no forced idleness of servers, and so on.

In Lemma (III.2.2), we show that $Q$ is still unchanged for any workconserving scheduling policy (with preemption or not) if we further assume that service times are i.i.d. and exponentially distributed.

**Lemma III.2.2** *Consider a GI/M/s/K +M queue. Times before abandonment are assumed to be i.i.d. and exponentially distributed. Then, the probability of being served $Q$ is constant for any workconserving scheduling policy.*

Using the above results, we state the following theorem:

**Theorem III.2.1** *Consider a GI/M/s/K +M queue. Times before abandonment are assumed to be i.i.d. and exponentially distributed. Then, probability of being served $Q$ is strictly increasing in the buffer size $k$.*

Similarly to the proof of Theorem (III.2.1), we also state that $Q(t)$ is an increasing function of $k$. Note that it is not necessarily strictly increasing in $k$ as it is the case for $Q$.

We now investigate the second order property of monotonicity, of the probability of being served, in the queue capacity. It is easy to prove by coupling arguments that $Q(t)$ is not concave in $k$. As for the stationary probability of service, we state the following result.

**Theorem III.2.2** *Consider an M/M/s/K +M queue. Times before abandonment are assumed to be i.i.d. and exponentially distributed. Then, $Q$ is a strictly concave function in the buffer size $k$.*

## 6 Concluding Remarks and Future Research

We considered the analysis of various queueing systems with customer abandonment. We first extended existing results for the M/M/s+GI queue. We proposed a controlled approximation of the performance measures that could be as accurate as preferred. The approach consists of approximating the hazard rate function of the patience distribution by a step function.

We then considered multi-server non-preemptive priority queueing systems, working under FCFS and LCFS. For each customer type, we explicitly derived the Laplace-Stieltjes transforms of the unconditional waiting times, the conditional waiting times given service, and the conditional waiting times given abandonment. A challenging and interesting step is to extend our approach to the case of many customer types with different mean service and patience

times. Another useful extension would be to consider protocols with mixed priorities, i.e., both preemptive and non-preemptive priorities.

Finally, we considered a queueing system with abandonment and finite buffer size. We investigated monotonicity results of the probability of being served with respect to the buffer size. These results are helpful when addressing optimizations issues. We considered both transient and stationary quantities of the performance of interest. As a topic for future research, it would be interesting to investigate the convexity properties of the performance measure as a function of other parameters such as the arrival rate, service rate, and in particular the number of servers.

# Chapter III.3

# Dynamic Control of Queueing Systems

## 1 Context and Contributions

This chapter summarizes my contributions on the dynamic control of queueing systems. My contributions mainly deal with Markov decision processes (MDPs). An MDP is defined as a discrete time stochastic control process. It is an extension of Markov chains; the difference is the addition of actions and rewards. MDPs provide a mathematical framework for modeling decision making in situations where outcomes are partly random and partly under the control of a decision maker. They are useful for studying a wide range of queueing optimization problems.

My contributions can be divided into two parts. They are undertaken in collaboration with Ger Koole and the postdoc Benjamin Legros. In a first part, we focus on the two-server slow-server problem with service failure. We consider manufacturing and service systems that are concerned about two conflicting goals: minimizing waiting time and maximizing the number of satisfied customers. For a Markovian two-server queueing model, we formulate this problem as a dynamic control problem with the objective of minimizing a weighted sum of the expected waiting time and the rate of unsatisfied customers. Using an MDP approach, we prove under finite and infinite horizons that the optimal routing is of threshold type. The result is proven for a large class of performance indicators. Under infinite horizon, there is one preferred server that should be always used, and the other one should be only used when the former is busy and the number of customers waiting in the queue exceeds a certain threshold. Using a Markov chain approach, we also provide closed-form expressions for the stationary performance measures in terms of the expected waiting time, and the production rate for each server. Finally, we use the performance results to identify the threshold of the optimal policy, as well as the server that should be prioritized.

The second part of my contributions deal the uniformization for jump Markov processes. This is useful for the analysis of unbounded Markov processes, for which major numerical difficulties are identified, and the existing literature have failed to provide appropriate solutions. We

consider multi-server queueing systems with generally distributed abandonment times. Using a non-standard Markov chain modeling, we obtain a natural bounded jump Markov process. The idea consists of explicitly modeling the waiting time of the first customer in line. This avoids to obtain unbounded abandonment rates, as it is the case with the traditional modeling using the number of customers waiting in the queue. In addition to having a uniformizable system, the new approach allows for the policies that are based on the actual waiting time, and not simply its expected value. Our approach can be applied to a wide range of open queueing optimization problems. This allows us to believe that we have provided cutting edge results.

# 2 Optimal Routing for the Slow-Server Queue

## 2.1 Introduction and Positioning of the Contributions

The operation speed usually interacts with the quality of the provided good or service. In some cases, a high speed means a hurry and no enough attention, which leads to a poor quality. In other cases, high speed may be related to a well trained and experienced human capacity, which implies high quality. Managers are then worried about the customer waiting time and, at the same time, about the quality of the provided service. An important question is how to match between customers and servers so as to optimize the system performance?

This work is most closely related to the slow-server problem literature. In the two-server slow-server problem, there is a single Poisson arrival process, and two exponential servers with different service rates (speeds). The objective is to find a non-preemptive scheduling rule that minimizes the customer expected waiting time in the queue. There is a rich literature dealing with two-server slow-server problem. The major drawback in this literature is that it ignores the heterogeneity in the quality of the provided service. Using different approaches, Larsen and Agrawala (1983), Lin and Kumar (1984), Walrand (1984), and Koole (1995) prove that the fast server should be always used, and the slow server should be only used when the fast server is busy and the number of customers waiting in the queue exceeds a given threshold. The exact analysis for the general setting with more than two servers is however still open (Weber, 1993; Rykov, 2001; Cabral, 2005; de Véricourt and Zhou, 2006).

Unfortunately, the literature has rarely addressed the slow-server routing problem by including service quality related factors. Two exceptions, belonging to the call center operations management literature, are de Véricourt and Zhou (2005), and Zhan and Ward (2014). Using an MDP formulation, de Véricourt and Zhou (2005) address a dynamic control problem under the objective of minimizing the expected total time of call resolution. For the two-server case, they prove that the optimal policy is of a threshold type. A call should be routed to the server with the highest resolution rate (resolution probability times service rate) whenever possible. The resolution rate policy is however shown to perform poorly under an objective that involves the call back probability (Mehrotra et al., 2012). Under the asymptotic many server quality and efficiency driven regime, Zhan and Ward (2014) extend the analysis of de Véricourt and Zhou

(2005), by considering similar modeling and assumptions, but a more general objective measured as a weighted sum of the expected waiting time and the call back rate. They approximate this asymptotic problem by a diffusion control problem.

In this work, we consider a Markovian two-server queueing model with one stream of arrivals. Each server has its own service rate and resolution probability. Using an exact MDP approach, we address the optimal routing decision problem under the objective of a weighted sum of the expected waiting time and the unsatisfied customer rate.

This is an important result as pointed out by de Véricourt and Zhou (2005). We prove under finite and infinite horizon that the optimal routing is of threshold type. We prove this result for a large class of performance measures including the expected waiting time (time in the system, other moments of the waiting time, etc.). Under infinite horizon, there is one preferred server that should be always used, and the other one should be only used when the former is busy and the number of customers waiting in the queue exceeds a certain threshold. Using a Markov chain approach, we provide closed-form expressions for the system stationary performance measures in terms of the expected waiting time, and the production rate for each server. Finally, we use the performance measures to identify the threshold, and also the server that should be prioritized under the optimal policy.

## 2.2 Problem Formulation

Consider a queueing system with a single customer type and two parallel servers, servers 1 and 2. Customers arrive, at a dedicated first come first served (FCFS) infinite queue, according to a Poisson process with rate $\lambda$. Service times are independent and exponentially distributed with rate $\mu_i$ for server $i$, $i \in \{1, 2\}$. Once server $i$ completes a service, the customer is either satisfied with probability $1 - \alpha_i$, or unsatisfied with probability $\alpha_i$, $i \in \{1, 2\}$. An unsatisfied customer defects, and this is considered as a loss of goodwill. To ensure stability, we assume that $\lambda < \mu_1 + \mu_2$. The stationary performance measures of interest are the customer expected waiting time, denoted by $E(W)$, and production rate (throughput) of server $i$, denoted by $T_i$, $i \in \{1, 2\}$.

Consider now the set of all non-preemptive non-anticipating FCFS routing policies. At any point of time, we want to decide for the first customer in the queue (if any) whether to keep her in the queue, or to serve her by an available server (if any). If we choose to serve her, we want to decide also to which server she should be routed. We combine two objectives to have a tradeoff between minimizing waiting times and maximizing customer satisfaction about the provided service. Concretely, the goal is find the optimal routing to minimize the following weighted sum

$$\alpha_1 T_1 + \alpha_2 T_2 + \alpha_W E(W), \tag{III.3.1}$$

where the coefficient $\alpha_W$ ($\alpha_W \geq 0$) translates the relative importance given, by the system manager, to the expected waiting time compared to the throughput of unsatisfied customers.

## 2.3 Optimal Routing

We formulate the routing problem as an MDP, and apply discrete-time dynamic programming to characterize the optimal routing policy. Let us denote by $x$ the number of customers in the queue, $x \geq 0$. The state of the servers is described through the symbols 0, $A_1$, $A_2$ and $A_1 + A_2$. State 0 is a situation where the two servers are idle. State $A_i$ is a situation where only server $i$ is working, $i \in \{1, 2\}$. State $A_1 + A_2$ is a situation where the two servers are working.

The possible actions for an idle server, just after a service completion or an arrival, is either to remain idle, or to serve a waiting customer, if any. Let us denote by $V_n(.,.)$ the value function over $n$ steps depending on the state of the system (the first variable is the state of the servers, and the second variable is the number of customers in the queue), for $n \geq 0$. We also consider the cost function $c(.,.)$ that also depends on the state of the system. This is a more general framework for the last term in the objective function in (III.3.1), and for which we derive the optimal policy. The cost has however to belong to a specific class of functions. In the case of only considering the expected waiting time, this cost function is defined as proportional to the number of customers in the queue. If we are interested in the expected number of customers in the system, we define $c(.,.)$ by $c(0, x) = x$, $c(A_i, x) = x + 1$ and $c(A_1 + A_2, x) = x + 2$, for $x \geq 0$ and $i \in \{1, 2\}$. For the expected number of customers in the queue, we define $c(.,.)$ by $c(0, x) = c(A_i, x) = c(A_1 + A_2, x) = x$, for $x \geq 0$ and $i \in \{1, 2\}$. For higher moments of the number of customers in the queue, we define $c(.,.)$ by $c(0, x) = c(A_i, x) = c(A_1 + A_2, x) = x^k$, for $x \geq 0$, $i \in \{1, 2\}$ and $k \geq 0$. This is also true for higher moments of the number of customers in the system.

In Theorem III.3.1, we prove by induction on the value functions that the optimal policy is of threshold type. More concretely, under finite or infinite horizon, there exists two thresholds on the queue length, a first one strictly under which both servers should be idle, and a second one above or equal to which both servers should work. In the remaining cases (above or equal to the first threshold, or strictly under the second one) only one of the two servers should work and the other one should remain idle.

**Theorem III.3.1** *The optimal policy is of threshold type.*

The main difference in the proof of Theorem III.3.1 compared to the proofs on the optimality of a threshold policy from Lin and Kumar (1984) and Koole (1995) is that we do not take into account that a server can be better than the other one and thus should be prioritized. Our proof does not provide even conditions to known which server should be preferred. This will be done in the next section thanks to the performance measure results. The difficulty in the choice for server 1 or server 2 comes from the service quality feature in our formulation, which is not captured by the previous works in the literature.

Note also that, under the finite horizon, forcing the two servers to be idle could be optimal under some situations. Yet, under an infinite horizon having the two servers idling at the same time can not be optimal, as long as a waiting customer represents a strictly positive cost for the system. Consider the first customer in the queue. If the two servers are idle, the decision not

to serve this customer simply delays the decision to the next event. The probability not to be successful in the customer treatment remains identical at the next event, however the waiting cost has increased. Thus, idling the two servers at the same time can not be optimal in the long run. The first threshold, under the infinite horizon, is simply 0.

### 2.4 Performance Measures

Consider the two-server queueing model working under the infinite horizon optimal policy. Using a Markov chain approach, we derive the stationary performance measures, in terms of the expected waiting times, $E(W)$ and the server throughputs $T_1$ and $T_2$. We assume without loss of generality that server 1 is always used and server 2 is only used when the number in the queue exceeds a certain threshold $u$. These results allow also to identify the threshold on the queue length, as well as the preferred server.

**Optimal Threshold:** It is clear that the expected waiting time $E(W)$ is strictly increasing in $u$. It is also easy to see that the throughput from server 2 decreases in $u$. The overall bad throughput from the two servers is $\alpha_1 T_1 + \alpha_2 T_2 = \alpha_1(\lambda - T_2) + \alpha_2 T_2$. We have $\alpha_1(\lambda - T_2(u+1)) + \alpha_2 T_2(u+1) - \alpha_1(\lambda - T_2(u)) - \alpha_2 T_2(u) = (\alpha_2 - \alpha_1)(T_2(u+1) - T_2(u))$. Since $T_2$ is decreasing in $u$, $T_2(u+1) - T_2(u) < 0$. Therefore the bad throughput strictly increases in $u$ if $\alpha_2 < \alpha_1$, and it strictly decreases in $u$ if $\alpha_2 > \alpha_1$. A consequence is that if $\alpha_1 \geq \alpha_2$ and the priority is given to server 1, then the optimal value of the threshold is $u = 0$, because $\alpha_1 T_1 + \alpha_2 T_2 + \alpha_W E(W)$ is strictly increasing in $u$. In the case $\alpha_2 \geq \alpha_1$, $\alpha_1 T_1 + \alpha_2 T_2 + \alpha_W E(W)$ is positive, therefore, there exists a value of the threshold $u^*$ which minimizes this expression, $u^* = \arg\min_{u \geq 0} (\alpha_1 T_1 + \alpha_2 T_2 + \alpha_W E(W))$. Note that we can have the case $u^* = +\infty$, for which server 2 is never used. This corresponds to a case with an extreme small value for $\alpha_W$.

**Prioritized Server:** The coefficient $\alpha_W$ defines the relative importance given to the expected waiting time. For extreme values of $\alpha_W$ which corresponds to situations where the manager only cares about the rate of satisfied customers or the expected waiting time exclusively, the priority should be given to the more efficient or to the fastest server, respectively. If server 1 is at the same time faster ($\mu_1 > \mu_2$) and more efficient ($\alpha_1 \leq \alpha_2$) than server 2, thus the priority for server 1 is clear and we choose the optimal threshold $u^*$ such that $u^* = \arg\min_{u \geq 0} (\alpha_1 T_1 + \alpha_2 T_2 + \alpha_W E(W))$. Consider the case when server 1 is faster ($\mu_1 > \mu_2$) but less efficient ($\alpha_1 > \alpha_2$). Therefore, it is no longer obvious that we should always prioritize server 1. Since $\alpha_1 T_1 + \alpha_2 T_2 + \alpha_W E(W)$ strictly increases in $u$ if $\alpha_1 > \alpha_2$, the choice of the preferred server is as follows. We compare between the values of $\alpha_1 T_1 + \alpha_2 T_2 + \alpha_W E(W)$ in two cases: the first case is when server 1 is prioritized and $u = 0$, and the second case is when server 2 is prioritized with the corresponding optimal threshold. The case with the best objective function leads then to the choice of the preferred server. Assume now that $\alpha_1 > \alpha_2$ and $u^* = \arg\min_{u \geq 0} (\alpha_1 T_1 + \alpha_2 T_2 + \alpha_W E(W))$ when server 2 is prioritized over server 1. Based on the above explanations, we identify a sufficient

and necessary condition for server 1 to be prioritized.

# 3 Uniformization for Queues with Abandonment

## 3.1 Introduction

Existing applications of Markov decision processes often fail in addressing the dynamic control questions for queueing models with abandonments. Such questions are important and arise in numerous applications. Examples include the optimal scheduling of jobs in call centers, patients in healthcare systems, perishable products in manufacturing systems, just to name a few. The reason is that the available approaches require uniformization (Down et al., 2011), while in the considered queueing models, the jump rates are generally unbounded functions of actions and states.

To overcome the limitations of the standard techniques, Bhulai et al. (2013) propose for a single server model a method that modifies the system rates by linearly smoothing them. However, this method only works under some conditions that guaranty an appropriate approximation of the original Markov decision process by the smoothed one (Blok and Spieksma, 2013). In this work, we propose a non-standard definition for the system states that allows to obtain a natural uniformized system with no rate modification, or state truncation. The idea is to explicitly model the customer waiting in the system state, instead of the traditional modeling using the number of customers. This idea has been first proposed by Koole et al. (2012) in order to analyze the performance measures of queueing systems with no abandonments. The approach consists of discretizing the customer waiting time using successive exponential phases, and report the waiting phase in the Markov process. We only need this information for the first customer in line. The difficulty of applying this method in the case of abandonment comes from the fact that the next customer first in line, if any, is no longer necessarily the customer arrived after the customer who just left the queue. The former might actually have abandoned.

In this work, we consider queueing systems with generally distributed abandonment times. Using an explicit modeling of the waiting time of the first customer in line, we obtain a bounded jump Markov process and illustrate its usefulness to solve dynamic control problems. This method is expected to be applicable to a wide range of performance analysis and optimization problems.

An additional advantage of the first in line modeling is that it allows for the use of policies that are based on the actual waiting time and not simply on its expected value (based on the number of customers in the queue as it is usually done in previous work). This is especially important for wide range of objective functions for systems with customer abandonment. It is for instance clear that scheduling policies according to the actual waiting time are efficient to deal with increasing or decreasing failure rate abandonment times.

### 3.2 Erlang Approximation with Abandonment

Consider a queueing system with one infinite FCFS queue. Customers arrive according to a Poisson process with parameter $\lambda$. We let customers be impatient while waiting in the queue. Times before abandonment are i.i.d. and follow a general distribution. There are no specific assumptions on the service process.

We use a non-traditional approach for the modeling of the queue, as proposed in Koole et al. (2012). The idea is to use a continuous time Markov chain approach in which we discretize the waiting time of the first customer in line (FIL) by an Erlang random variable with rate $\gamma$ per stage. The higher is $\gamma$, the better is the approximation. The system states are defined by the waiting time stage denoted by $i$ ($i > 0$) of the customer FIL, if any. State 0 represents an empty queue. The transition rate from the waiting stage $i$ to $i + 1$ is $\gamma$, for $i > 0$. The transition rate from state 0 to state 1 is $\lambda$.

Once the current FIL leaves the queue from state $i$ ($i > 0$) to start service or because she does abandon, the next state is $i - h$, $i > 0$ and $0 \leq h \leq i$. The main difficulty is to find the transition probabilities from state $i$ to $i - h$, $i > 0$ and $0 \leq h \leq i$, after an abandonment or a start of a service. The next first in line, if any, is no longer necessarily the customer arrived after the FIL who just left. The former might actually have abandoned.

We approximate times before abandonment by a Coxian random variable. The value of the Coxian modeling, with identical or different phase rates, comes from its universality. It is dense in the field of all positive-valued distributions (Schassberger, 1973). We consider a Coxian distribution with the following parameters: we denote by $b$ the probability for an arbitrary customer to accept waiting. All phases durations are exponentially distributed with the same rate $\gamma$. The conditional probability for a given customer to move from phase $i$ to $i+1$ is $\dfrac{\gamma}{\gamma + \beta_i}$, with $\beta_i \in \mathbb{R}^+$ (Figure III.3.1).

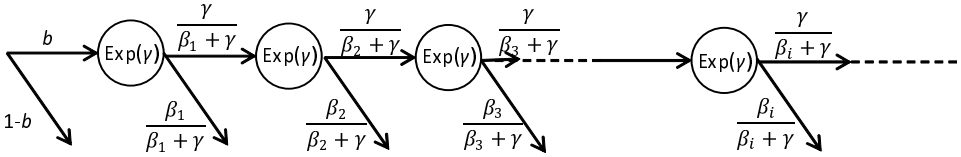



Figure III.3.1: Coxian distribution for abandonment times

We denote by $p_{i,i-h}$the transition probability to move from stage $i$ to stage $i - h$ in the Markov chain, for $i > 0$ and $0 \leq h \leq i$. Our main result is that these probabilities are given by $p_{i,i-h} = \prod_{k=1}^{i} q_k$, for $i = h$ and $i > 0$; and $p_{i,i-h} = (1 - q_{i-h}) \prod_{k=i-h+1}^{i} q_k$, for $0 \leq h < i$; where $q_k = (1 + \frac{b\lambda}{\gamma} \prod_{j=1}^{k} \frac{\gamma}{\beta_j + \gamma})^{-1}$, for $k > 0$.

The value of the FIL modeling is that it allows to obtain a natural uniformizable system, which is new in the literature. This avoids the major existing numerical issues for the dynamic control of queueing systems with abandonments. Moreover, the FIL approach allow for the use

of policies that are based on the actual waiting time and not simply on its expected value (based on the number of customers in the queue as it is usually done).

# 4 Concluding Remarks and Future Research

We considered dynamic control problems for queueing systems. We first focused on a two-server queueing problem with service failure. We formulated and solved a dynamic control problem with the dual performance objectives of minimizing the expected waiting time and maximizing the number of satisfied customers. An interesting future research is to extend the results to the multi-server case. It would be also interesting to consider the control problem in a more general context with multiple customer types and heterogeneous servers.

We then considered queueing systems with general abandonments. We proposed a Markov process that explicitly model the waiting time of the first customer in line. This has led to a bounded jump Markov process that allows for policies based on the actual waiting time. A direction for future research is to extend the analysis to the case with more than one type of impatient customers.

# Part IV

# Ongoing Work and Research Perspectives

The service sector is the largest sector of the economy in most industrialized nations, and is fast becoming the largest sector in developing nations as well. Driven by today's new business environment including advanced telecommunications, accelerated business globalization, increased automation, and highly on-demand and competitive innovations, the complexity of the operations management of service systems is continuously increasing. Managers of service systems are wrestling to deliver both traditional conflicting objectives of low operating costs and high service quality.

I want to contribute to the service operations literature while accounting as much as possible for the key features of advanced technology and human factors. This is what I have done in my previous work, through quantitative approaches that are mainly based on stochastic processes. This is also exactly what I want to do during the coming years. What feeds my motivation is the need in practice for relevant research results accounting for the complex real-life features, in order to derive useful recommendations and insights to practitioners.

My objective is then to continue my work on call center operations, intensify my ongoing work on emergency departments, and continue my work on stochastic methods with applications to services. For the last objective, the considered models are rather theoretical and deal with fairly general settings. On the one hand, this allows us to obtain results that are helpful for a wide range of applications. On the other hand, we are also aware that this genericity may ignore some specific and important features when applied to some specific contexts.

Finally, I would like to point out that I will accentuate the use of real data in my models. Fortunately, the recent information system advances have allowed to obtain large data sets. This is a significant opportunity that will, with no doubt, enrich the existing models and lead to more accurate and high impact studies.

The details on my ongoing and future research projects are described in the sections below.

# 1 Call Center Operations

## 1.1 Shift-Scheduling with Uncertain Arrival Rates

In Chapter II.3, I summarized my contributions to call center planning with uncertain arrival rates and a single shift setting. Within the PhD thesis of Mathilde Excoffier, supervised also by Abdel Lisser and Céline Gicquel, we would like to consider a setting of workforce scheduling with multiple shifts.

We will focus on computing the required number of agents to be assigned to a set of predefined shifts. Our approach relies on developing stochastic programming approaches under demand forecasts uncertainty. The main scientific challenge ahead is to devise an approach where the optimization problem is modeled with a sufficient degree of accuracy to ensure the practical relevancy of the obtained schedule, while keeping the mathematical formulation computationally tractable.

We first aim at proposing a chance-constraint programming approach, where the manager

estimates an acceptable risk level for the potential shortfall in the quality of service. Moreover, we plan to explicitly use a continuous probability distribution of the forecast errors to represent the uncertainty on call arrival rates as this should lead to schedules which provide an actual risk level closer to the expected one. Solving the resulting optimization problem will be a challenging task. For that reason, we will in a first step focus on a simplified version of the problem. We will consider for instance a single-class single-skill call center and a period-by-period measure of the quality of service. This should allow us to gain a better understanding of the fundamental tradeoff encountered by call center managers, namely minimizing staffing costs while ensuring with an acceptable risk level that the target quality of service will be provided to customers. We will then try to gradually add some complicating but realistic features such as more flexible shift patterns, a multi-period measure of the quality of service or a multi-class multi-skill layout of the call center.

Our second objective will be to develop a multi-stage stochastic programming approach for the problem. We will thus consider a multi-day scheduling horizon with the managerial possibility to resort to some recourse actions to adjust the agent schedules once part of the actual demand is realized and the call arrival forecasts have been adjusted accordingly. The cost criterion functions for this problem include the regular salary costs, adjustment costs and penalty costs for under-staffing. Our objective is to find the optimal initial shift scheduling and update policy which minimizes the total call center operating cost. To deal with the resulting difficult optimization problem, we plan to use an approximate representation of the uncertainty by discretizing the probability distribution and constructing event-trees with scenarios.

We would like also to study the general impact of the original feature considered in our models, i.e., the randomness in call arrival rates, on the call center shift scheduling problem. From a managerial point of view, does it really matter to explicitly take into account the uncertainty on arrival parameters while building the agent schedules? What are the practical consequences of not considering it? What is the additional staffing cost needed to hedge against this uncertainty? We will also try to evaluate to which extent it might be profitable to have the flexibility to update the staffing decisions one or several times during the scheduling horizon. It is of course expected that more flexibility in staffing should lead to lower costs and better quality of service. But it would be interesting for a call center manager to be able to estimate the extent of this improvement as well as to obtain some insights about the number and timing of the adjustments to be carried out.

## 1.2 Advertisement While Waiting

This work will be undertaken under our collaboration with Interact-iv.com. With Zeynep Aksin, Ger Koole and the postdoc Benjamin Legros, our objective is to understand and study queueing systems with advertisement. This is expected to be the basis of the new economic model for some types of financial call centers.

Consider, for instance, the case of a call center offering the service of directory sellers. It

consists of helping callers to get contact information on a variety of service providers (a plumber, an electrician, a restaurant, etc.). The current economic model, in France, of such a call center is mainly based on an overtaxed waiting time. In the near future, it is likely expected that a new law will force direct sellers to substitute their overtaxed numbers by cheap or free numbers. The new economic model that has been started to be adopted by direct sellers is to broadcast advertisements for other parties. The idea is that advertisement revenues would compensate the loss from removing overtaxed numbers.

We recently obtained data on directory sellers under both situations: the current and the new economic models. We want to analyze the data in order to study the customer waiting experience. First results and discussions with the company indicate that the abandonment customer behavior has changed with the new economic model, i.e., with advertisement. We then aim at characterizing the new abandonment behavior. We want also to optimize the advertisement parameters: number, length, time during the wait, etc. Another interesting issue is the study of the impact of advertisements on the customer loyalty and retention. Actually, this is a new and interesting framework for which most issues are not solved. Our goal in short and mid terms is to address various related performance evaluation and optimization issues, that would help managers to better understand their systems and to make them more efficient.

## 1.3 Multi-Channel Issues

New advances in telecommunication technology are revolutionizing the way call centers interact with customers. Customers preferences are also evolving rapidly toward the use of new technology. As a result, call centers are currently increasing the use of new channels, which has in turn pointed out a large number of new challenging issues.

For this reason, multi-channel call centers have recently emerged as a fertile ground for academic research. With Ger Koole and Benjmain Legros (postdoc), we have initiated a collaboration with Rob van der Mei and Sihan Ding (PhD student) to work on multi-channel issues. Our objective is to contribute significantly to this new stream of literature, given its impact on the real-life call center practice and at the same time the involved challenging theoretical studies.

There are many interesting OM issues that are, roughly speaking, related to the study of the impact of new technologies on the call center performance. I in particular mention two issues, one on chat (instant messaging) and the other on call backs.

**Chat systems:** Chat systems allow customers to access an instant messaging system built into the call center website to interact with agents online. From an operational point of view, the main difference between call and chat channel is that while an agent can only serve a single call at once, she can serve multiple customers simultaneously using chat (Tezcan and Zhang, 2014). Other advantages of chat systems are the features such as screen sharing and the ability to share files and data, which are particularly useful to computer companies, software companies,

and e-retailers (Cui and Tezcan, 2014). There are however drawback from using chat systems, including a longer service time due to extra typing and reading time and frustration caused by technological barriers (Shae et al., 2007).

From an OM perspective, we want to address the following questions. What are the appropriate metrics for a chat system? What should be the maximum number of opened chat sessions? What should be the optimal routing decisions? How to optimize the staffing level under a given objective service level? Due to the specificity of the service process, new queueing methods should be used to address these questions. Some first ideas suggest to consider the analogy with processor sharing queues.

**Call Back Option:** In the context of highly congested call centers, the use of alternative service channels can be proposed to customers so as to better match demand and capacity. Alternative channels could be email, chat, blog, postponed call back service, etc. We focus on this last alternative. The idea is that customers, who are expected to experience long waiting times, receive the option to be called back later. This leads to a contact center with two channels, one for inbound calls, and another for outbound calls.

The flexibility of the call back option comes from the willingness of some customers to accept future processing. The call center can then make use of this opportunity to better manage arrival uncertainty, which in turn would improve the system performance. One important question for managers in this setting is how should be the routing rule of jobs that would ensure non-excessive waiting times for both job types, i.e., upon service completion, should the agent handle an inbound or an outbound call?

We want to address this question under a queueing modeling framework in which we capture the customer reaction to the call back option. The key distinction of call center problems with blending comes from the fact that outbound calls have less urgency relative to inbound calls. The existing blending models in the literature mainly consider outbound calls as back-office jobs that are already stored and are infinite. In a call center with a call back option, the number of customers waiting to be called back is finite in order to avoid excessive waiting. The routing policy would then depend on the length of the call back queue. Another difference, compared to cases with classical outbound tasks, is that inbound and outbound calls are negatively correlated. This implies a different analysis, and also leads to different managerial recommendations.

Our goal is to study the impact of using the call back option on the system performance. We also want to derive the optimal scheduling policy of jobs that minimizes an objective function involving queueing delays for inbounds and outbounds. Another interesting study is to optimize the threshold at which the system transforms an inbound call to an outbound one.

## 2 Emergency Department Operations

One year ago, I have started with Zied Jemai, Ger Koole and Karim Ghanes (PhD student) to work on the optimization of emergency departments. This is done under a project funded

by Agence Régionale Santé Ile-de-France. We are closely collaborating with the urban French hospital Saint-Camille.

An emergency departments (ED) is a service system. It is the main entrance to a hospital for emergency incidents, offering non-stop services for any kind of patients. The continuous increase in demand combined with austerity measurements have led to extensive congestion (Hoot and Aronsky, 2008). Under a difficult economic context, ED managers are trying to improve performance by minimizing the mismatch between patient demand and supply. However, an ED is a complex environment with various types of heterogeneous patients and resources where most of the parameters are uncertain. Healthcare practitioners have therefore resorted to researchers in operations management and operations research in order to develop scientific approaches for the performance optimization of EDs.

For the coming years, I am planning to extensively work on various operations management issues of EDs. My motivations comes from the important societal impact of EDs where costs and profits are not the sole elements. The employee well being and the quality of service offered to the patient are in the heart of the manager concerns. As for call centers, the human element is a central features that makes the study of EDs interesting, but at the same time challenging due to the human complex factors. In what follows, I give a description of my ongoing and future research works on this subject.

## 2.1 Performance indicators

Here, I describe an ongoing work on the analysis of key performance indicators for an ED. The performance of emergency departments is facing a recurrent worldwide problem nowadays, namely overcrowding. Overcrowding or congestion in EDs occurs when the available caring capacity cannot meet the demand represented by patient flow, and it can manifest itself through different ways. For instance, an excessive number of patients present in the ED, long patient stays and waiting times, and treatment in hallways, are all overcrowding signs. Congestion in emergency departments leads to negative effects such as decreased physician productivity, miscommunication between working staff, diversion of ambulances (Paul et al., 2010), and dissatisfaction of patients who may sometimes leave without treatment (Saghafian et al., 2012). Moreover, it leads to high levels of stress, violence, decreased morals among ED staff, increased medical errors, higher mortality rates, high staff turnovers and unnecessarily high costs (Trzeciak and Rivers, 2003; Kuo et al., 2012; Spirivulis et al., 2006).

ED managers often evaluate their system through the use of Key Performance Indicators (KPIs) that are related to overcrowding, such as the total length of stay, the time to first treatment or the rate of patients that leave without being seen by a physician. In this work, we want to produce a survey on the existing KPIs from an Operations Research/management perspective.

The motivation of our work is as follows. The selection of KPIs for EDs has always been a controversial subject, and the whys and wherefores of this choice remain unclear. ED is a large

and complex system and each of these metrics measures something different (Hwang et al., 2011). Neither the scientific community nor practitioners are able to decide on the most appropriate KPI, as each indicator presents at the same time benefits and drawbacks. We want to highlight and discuss these issues for all existing KPIs. For instance, ambulance diversion and the rate of patients left without being seen cannot be used as a reference to compare different EDs since they depend of the ED environment. Time to first treatment is a crucial KPI for critical cases but it does not give any information about the ED state in other important stages of the process, and the length of stay (LOS) gives an overview of the entire system performance but does not allow to figure out local strengths and weaknesses. We want also to underline eventually some relevant combinations of KPIs.

## 2.2 Human Resource Related Issues

We have recently initiated a work on human resource issues, i.e., the effect of staffing levels and allocations on ED performance. As a first step, we proposed a simulation model based on a comprehensive understanding of the real-world functioning of emergency departments. A field study was conducted for this purpose through a close collaboration with the ED of Saint Camille hospital. Real data and expert judgments are both used for the construction of the model. For the validation, the model outputs were compared to historical data and judged by experts. In order to alleviate congestion, ED managers and the general management of Saint Camille hospital intend to invest in human staffing. Their objective is to improve the ED performance by investing in human resources. The question we are facing here is: By how much should the current staffing budget be increased and how should this additional budget be used in the allocation of human resources?

Two performance metrics are involved in this study: the expected length of stay (LOS, sum of sojourn times in all ED subsections), and the expected time to first treatment (TTFT, time between the patient's arrival and the first handling by a physician). LOS allows to approach the ED in a holistic way, however, focusing only on LOS could have important drawbacks. The impact could be in the non-urgent cases, or worst, the non-urgent cases could be benefited on behalf of prolonging the waiting time of the urgent ones. We then consider TTFT, because it allows to measure the most crucial element for severe incidents, as this waiting time affects the mortality rate of very ill patients (Spirivulis et al., 2006).

We will adopt a simulation-based approach for the optimization of staffing levels of the various human resource types involved in the ED (senior physicians, junior physicians, nurses, etc.). We will study the effect of the staffing budget on LOS. We want also to understand the effect of including the TTFT constraint, and investigate how this additional constraint may affect the optimal planning solution. Because of the correlation between the two metrics, there is an important tradeoff that one should understand and be aware of. We expect that the takeaways and key conclusions of this study will be useful for our partner but also for most other ED frameworks.

### 2.3 Process Related Issues

In the near future, we will focus on process-related ED issues, namely, we will assess the impact of modifying the process or changing some protocols and organizational rules on ED performance. As underlined by our partners, ED problems can stem from the process itself, not the staffing levels. There is a growing literature dealing with such issues. For instance, using MDP, Saghafian et al. (2012) assess the effect of using a complexity-augmented triage on the performance of the ED, while EDs typically use triage systems that classify and prioritize patients almost exclusively in terms of urgency. Huang et al. (2012) address the problem of patient flow control in EDs. They investigate the optimal decision for the physician at some point in the process: either to handle a new patient coming from triage or an "in-process" patient. Other related literature include Pallin and Kittell (1992) and García et al. (1995).

We plan to investigate the benefits of changing various process related procedures. For instance, we want to assess the effect of some "anticipation methods" like allowing the triage nurse to order tests and treatments (currently, the triage nurse only categorize the severity index). This would reduce queueing delays for the first consultation, however it would also imply an error because nurses are less experts than physicians. Such error may further create congestion, since the exam equipments are a significant bottleneck in EDs. Another example of process related issues is the controversial same patient same physician (SPSP) rule. From the one hand, applying SPSP would deteriorate performance (less pooling effect). From the other hand, ignoring SPSP would create a non-negligible duration for a physician to understand the health situation of patient that has been first seen by another physician. Also, there is a human link between the patient and the physician that is lost. Our objective is to quantify the comparison between applying or not the SPSP rule. Another interesting question to address is related to diagnostic tests (blood, urine, imaging, etc.) that are performed by a common service to all hospital departments (the ED is one of them). Often, ED practitioners complain about too long diagnostic delays. Point-of-care testing (POCT) may be a solution to this problem. It consists of performing biological tests and simple imaging inside the ED with the use of ED devices. We want to investigate whether the performance improvement may outperform the investment in diagnostic equipments or not. Many other ideas could be proposed, and assessing their benefits is worthwhile.

## 3 Stochastic Models and Their Applications to Services

In addition to call centers and emergency departments, I would like to contribute to the literature of other service systems. My goal is to develop quantitative stochastic methods that would be useful for a wide range of service applications. This is already the case as shown in my theoretical contributions to the analysis of stochastic processes (Part III), for which the applications goes beyond call centers and emergency departments. In addition to the future research directions mentioned in the conclusions of the chapters of Part III, I want to work in the near future on two

particular subjects. One deals with the study of queueing systems with finite and appointment-driven arrivals, and the other deals with collaboration strategies between service systems that are modeled as queueing systems. They are described below.

## 3.1 Appointment-Driven Arrivals

Under a collaboration with Saif Benjaafar and the PhD student Rowan Wang, we have worked on the analysis of the effect of heterogeneity in inter-arrival and service times on the performance of queueing systems with finite number of arrivals (Wang et al., 2014). We have examined various settings of patterns where inter-arrival and service times increase, decrease, increase and then decrease, or decrease and then increase. Applications include systems where arrivals are trigged by the start of an event or a service. An example is the arrival of passengers to check-in for or to board a flight. Passengers may belong to different classes (e.g., early, on-time, and late) or are assigned to different groups (e.g., priority boarding zones), so that arrivals occur in waves with each wave drawing from the population of the corresponding class or group.

With same team, we are planning in the near future to extend this work for queueing systems where the arrival of customers is driven by appointments, with a scheduled appointment time associated with each customer. However, customers are not necessarily punctual and may arrive either earlier or later than their scheduled appointment times. Customers may also not show up altogether. The arrival times of customers (relative to their scheduled appointments) and their service times are both stochastic. We will consider case where customers are not homogeneous in their punctuality, show-up probabilities, and time between previous and subsequent appointments, which may vary from customer to customer.

There are numerous service systems where the arrivals of customers are driven by scheduled appointments (Mondschein and Weintraub, 2003; Cayirli and Veral, 2003; Gupta and Denton, 2008). Examples include arrivals to healthcare facilities, government agencies (e.g., immigration, social services, and internal revenue), the offices of tax and financial service providers, academic advising offices at universities, restaurants and spa treatment facilities, just to name a few. Despite this prevalence, analytical tools for the performance evaluation of these systems are relatively limited. Existing approaches from queueing theory cannot be readily applied because of several important differences between standard queueing systems and systems with appointment-driven arrivals (ADA). Systems with ADA are characterized by (1) a finite number of customers (e.g., the set of patients that have been scheduled at a clinic in a given day), so that steady state analysis cannot be applied, (2) arrivals that are in part determined by known scheduled appointment times, (3) appointment times that may not be equally spaced, and (4) the possibility of customer non-punctuality and no-shows. The difficulty of the analysis can be further compounded in settings in which customers are heterogeneous in their service time requirements, punctuality, and no-show probabilities. To our knowledge there are no existing results that consider simultaneously appointment driven arrivals, non-punctuality, and no-shows, and do so for a setting as general as we want to investigate.

Our objective in as first step is to develop an approach to obtain various performance measures related to the customer waiting time. We want then to examine the impact of not accounting for non-punctuality and no-shows and see whether doing so may or not lead to significant errors. We also aim at developing an optimization method that can be used to support individualized appointment scheduling (scheduling that takes into account the punctuality, no-show behavior, service time distribution, and service level requirement of each customer).

## 3.2 Collaboration in Service Systems

Within the PhD thesis of Lisa Peng, co-supervised by Zied Jemai, we want to work on collaboration strategies between queueing service systems. Up to now, all my work only focus on a single actor. In practice, one may have several actors that collaborate under various architectures, which may affect considerably the system performance of each actor. There are vertical collaborations (between a supplier and a company), and horizontal ones (between companies with the same type of products). In particular, our goal is to work on queueing pooling strategies.

Resource pooling is an efficient strategy for dealing with uncertainty. It refers to an arrangement in which a group of common resources or servers is held for multiple customer streams rather than dedicated, separate resources for each individual customer stream. The main benefit of resource pooling is reduced congestion, as measured by the time spent by customers waiting to be served.

The efficiency benefits of resource pooling are commonly exploited in case multiple customer streams are served by one common service provider (Tekin et al., 2009). But these benefits can also be obtained if the customer streams belong to several independent service providers (Guo et al., 2013). There are numerous real-life examples in various sectors of independent service providers who may collaborate by pooling their resources into a joint service system. For instance, several manufacturers of advanced technical equipment may employ a number of non-branded repairmen to maintain and repair machines at their customers sites. Similarly, business units of a large insurance firm may operate a common call center with cross-trained telephone agents. One can also think of airline companies pooling check-in counters. Further, a hospital is often comprised of clinical departments that share operating rooms, hospital beds, and medical staff.

In general, collaboration among service providers enables more efficient use of their resources, offers the opportunity to benefit from large economies of scale, and enhances their negotiation power (Anily and Haviv, 2010). But how should be a good collaboration arrangement? How should the independent entities allocate the total costs/profits of the pooled strategy among them? A fair cost division is an essential prerequisite for a successful cooperation, but the construction of such an allocation tends to be challenging.

We will use the concept of the cooperative game theory, which offers a natural paradigm to tackle the above questions. We will consider short-term as well as long-term collaborations. In the former case, each entity (queueing system) has a fixed known number of servers, which she

brings to any coalition. In the latter case, which is much more challenging, each coalition picks a cost-minimizing number of servers.

This finishes the description of the projects I am planning to work on, and finishes also my HDR dissertation. At the end of this dissertation, I would like to again thank all my colleagues to which I owe all my results as well as my passionate interest in research.

# Bibliography


J. Abate and W. Whitt. A unified framework for numerically inverting Laplace transforms. *Informs Journal on Computing*, 18:408–421, 2006.

O.Z. Akşin and F. Karaesmen. Characterizing the performance of process flexibility structures. *Operations Research Letters*, 35:477–484, 2007.

O.Z. Aksin, B. Ata, S. Emadi, and C.L. Su. Impact of delay announcements in call centers: An empirical approach. Working paper, Koç University, 2013.

S. Aldor-Noiman, P.D. Feigin, and A. Mandelbaum. Workload forcasting for a call center: Methodology and a case study. *Annals of Applied Statistics*, 3:1403–1447, 2009.

G. Allon, A. Bassamboo, and I. Gurvich. We will be right with you: Managing customers with vague promises. *Operations Research*, 59:1382–1394, 2011.

E. Altman and A.A. Borovkov. On the stability of retrial queues. *Queueing Systems*, 26:343–363, 1997.

S.V. Amari and R.B. Misra. Closed-form expressions for distribution of sum of exponential random variables. *IEEE Transactions on Reliability*, 46:519–522, 1997.

Shoshana Anily and Moshe Haviv. Cooperation in service systems. *Operations Research*, 58(3): 660–673, 2010.

M. Armony and A.R. Ward. Fair dynamic routing in large-scale heterogeneous-server systems. *Operations Research*, 58(3):624–637, 2010.

M. Armony, N. Shimkin, and W. Whitt. The impact of delay announcements in many-server queues with abandonment. *Operations Research*, 57(1):66–81, 2009.

A.N. Avramidis, A. Deslauriers, and P. L'Ecuyer. Modeling daily arrivals to a telephone call center. *Management Science*, 50:896–908, 2004.

F. Baccelli and G. Hebuterne. On queues with impatient customers. *Performance'81 North-Holland Publishing Company*, pages 159–179, 1981.

F. Ball and V.T. Stefanov. Further approaches to computing fondamental characteristics of birth-death processes. *Journal of Applied Probability*, 38:995–1005, 2001.

S. Benjaafar. Performance bounds for the effectiveness of pooling in multi-processing systems. *European Journal of Operational Research*, 87:375–388, 1995.

H.G. Bernett, M.J. Fischer, and D.M.B. Masi. Blended call center performance analysis. *IT Professional*, 4(2):33–38, 2002.

D. Bertsimas and I. Popescu. Optimal inequalities in probability theory: A convex optimization approach. *SIAM Journal on Optimization*, 15:780–804, 2005.

S. Bhulai and G. Koole. A queueing model for call blending in call centers. *IEEE Transactions on Automatic Control*, 48:1434–1438, 2003.

S. Bhulai, A.C. Brooms, and F.M. Spieksma. On structural properties of the value function for an unbounded jump Markov process with an application to a processor sharing retrial queue. *Queueing Systems*, pages 1–22, 2013.

G.R. Bitran, J-C. Ferrer, and P. Rocha e Oliviera. Managing customer experiences: Perspectives on the temporal aspects of service encounters. *Manufacturing & Service Operations Management*, 10:61–83, 2008.

H. Blok and F.M. Spieksma. Continuity and ergodicity properties of a parametrised collection of countable state Markov processes. 2013. Working paper. University of Leiden.

V.A. Bolotin. Telephone circuit holding time distributions. *Proceedings of the 14th International Teletraffic Conference. Labetoulle J.and Roberts J.W., editors*, pages 125–134, 1994.

J. Boudreau. Organizational behavior, strategy, performance, and design in management science. *Management Science*, 50:1463–1476, 2004.

J. Boudreau, W. Hopp, J.O. McClain, and L.J Thomas. On the interface between operations and human ressources management. *Manufacturing & Service Operations Management*, 5: 179–202, 2003.

O. Boxma, G. Koole, and Z. Liu. Queueing-theoretic solution methods for models of parallel and distributed systems. *Performance Evaluation of Parallel and Distributed Systems - Solution Methods, CWI Tract 105 & 106, Amsterdam*, 1994.

A. Brandt and M. Brandt. On the two-class M/M/1 system under preemptive resume and impatience of the prioritized customers. *Queueing Systems*, 47:147–168, 2005.

L. Brown, N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L. Zhao. Statistical analysis of a telephone call center: A queueing-science perspective. *Journal of the American Statistical Association*, 100:36–50, 2005.

L.D. Brown, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L. Zhao. Multifactor poisson and gamma-poisson models for call center arrival times. Technical report, University of Pennsylvania, 2002.

F.B. Cabral. The slow server problem for uninformed customers. *Queueing systems*, 50(4): 353–370, 2005.

T. Cayirli and E. Veral. Outpatient scheduling in health care: A review of literature. *Production and Operations Management*, 12:519–549, 2003.

S. Charron and E. Koechlin. Divided representation of concurrent goals in the huma frontal lobes. *Science*, 328:360–363, 2010.

B.D. Choi, B. Kim, and J. Chung. M/M/1 queue with impatient customers of higher priority. *Queueing Systems*, 38:49–66, 2001.

G.L. Choudhury, D.M. Lucantoni, and W. Whitt. Numerical solution of Mt/Gt/1 queues. *Operations Research*, 45:451–463, 1995.

P. Coolen-Schrijner and E.A. van Doorn. The deviation matrix of a continuous-time Markov chain. *Probability in the Engineering and Informational Sciences*, 16:351–366, 2002.

L. Cui and T. Tezcan. Approximations for chat service systems using many-server diffusion limits. 2014. Working paper. University of Rochester.

J.N. Daigle and D.M. Lucantoni. Queueing systems having phase-dependant arrival and service rates. 1991. Chapter 10 of Numerical Solutions of Markov Chains, Editor: W.J. Stewart, Marcel Dekker, INC., 161-202.

F. de Véricourt and Y.-P. Zhou. Managing response time in a call routing problem with service failure. *Operations Research*, 53:968–981, 2005.

F. de Véricourt and Y.P. Zhou. On the incomplete results for the heterogeneous server problem. *Queueing Systems*, 52:189–191, 2006.

A. Deslauriers, P. L'Ecuyer, J. Pichitlamken, A. Ingolfsson, and A.N. Avramidis. Markov chain models of a telephone call center with call blending. *Computers & operations research*, 34: 1616–1645, 2007.

D. G. Down, G. Koole, and M. E. Lewis. Dynamic control of a single-server system with abandonments. *Queueing Systems*, 67:63–90, 2011.

P.E. Dux, M.N. Tombu, S. Harrison, B.P. Rogers, F. Tong, and R. Marois. Training improves multitasking performance by increasing the speed of information precessing in human prefontal cortex. *Neuron*, 63:127–138, 2009.

S. Favaro and S.G. Walker. On the distribution of sums of independent exponential random variables via wilks' integral representation. *Acta applicandae mathematicae*, 109(3):1035–1042, 2010.

P. Feigin. Analysis of customer patience in a bank call center. 2005. Working Paper, The Technion.

M. Fischer, D. Garbin, A. Gharakhanian, and D. Masi. Traffic engineering of distributed call centers: Not as straight forward as it may seem. 1999. Mitretek Systems.

P. Flajolet and F. Guillemin. The formal teory of birth-and-death processes, lattice path combinatorics and continued fractions. *Advances in Applied Probability*, 32:750–778, 2000.

G. Gallego and I. Moon. The distribution free newsboy problem: Review and extensions. *The Journal of the Operational Research Society*, 44:825–834, 1993.

N. Gans and Y.-P. Zhou. A call-routing problem with service-level constraints. *Operations Research*, 51:255–271, 2003.

N. Gans, G. Koole, and A. Mandelbaum. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management*, 5:73–141, 2003.

Marelys L García, Martha A Centeno, Camille Rivera, and Nina DeCario. Reducing time in an emergency room via a fast-track. In *Simulation Conference Proceedings, 1995. Winter*, pages 1048–1053. IEEE, 1995.

O. Garnett and A. Mandelbaum. An introduction to skills-based routing and its operational complexities. 2001. Teaching notes, Technion.

O. Garnett, A. Mandelbaum, and M. Reiman. Designing a call center with impatient customers. *Manufacturing & Service Operations Management*, 4:208–227, 2002.

W.H. Gladstones, M.A. Regan, and R.B. Lee. Division of attention: The single-channel hypothesis revisited. *Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology*, 41:1–17, 1989.

X. Gourdon. *Les maths en tête : Algèbre*. Ellipses, Paris, 1994.

L.V. Green, P.J. Kolesar, and W. Whitt. Coping with time-varying demand when setting staffing requirements for a service system. *Production and Operations Management*, 16:13–39, 2007.

R. Guérin. Queueing-blocking system with two arrival streams and guard channels. *IEEE Transactions on Communications*, 36:153–163, 1998.

F. Guillemin. Spectral analysis of birth and death processes. 2005. Working paper, submitted to Journal of Applied Probability.

F. Guillemin and D. Pinchon. Excursions of birth and death processes, orthogonal polynomials, and continued fractions. *Journal of Applied Probability*, 36:752–770, 1999.

P. Guo and P. Zipkin. Analysis and comparison of queues with different levels of delay information. *Management Science*, 53:962–970, 2007.

P. Guo and P. Zipkin. The effects of information on a queue with balking and phase-type service times. *Naval Research Logistics*, 55:406–411, 2008.

P. Guo, M. Leng, and Y. Wang. A fair staff allocation rule for the capacity pooling of multiple call centers. *Operations Research Letters*, 41(5):490–493, 2013.

D. Gupta and B. Denton. Appointment Scheduling in Health care: Challenges and Opportunities. *IIE Transactions*, 40:800–819, 2008.

S. Gurumurthi and S. Benjaafar. Modeling and analysis of flexible queueing systems. *Naval Research Logistics*, 51:755–782, 2004.

I. Gurvich. Design and control of the M/M/N queue with multi-class customers and many servers. 2004. Masters thesis, Technion.

I. Gurvich, M. Armony, and A. Mandelbaum. Service level differentiation in call centers with fully flexible servers. *Management Science*, 54:279–294, 2008.

I. Gurvich, J. Luedtke, and T. Tezcan. Staffing call-centers with uncertain demand forecasts: A chance-constraints approach. *Management Science*, 56:1093–1115, 2010.

J.M. Harrison and A. Zeevi. A method for staffing large call centers based on stochastic fluid models. *Manufacturing and Service Operations Management*, 7:20–36, 2005.

D. Holman, R. Batt, and Holtgrewe U. The global call center report: International perspectives on management and employment. 2007. Global Call Centre Research Network.

N.R. Hoot and A. Aronsky. Systematic review of emergency department crowding: Causes, effects, and solutions. *Annals of Emergency Medicine*, 52(2):126–136, 2008.

J. Hornik. Subjective vs. objective time measures: A note on the perception of time in consumer behavior. *Journal of Consumer Research*, pages 615–618, 1984.

J. Huang, B. Carmeli, and A. Mandelbaum. Control of patient flow in emergency departments, or multiclass queues with deadlines and feedback. 2012. Working Paper.

M. Hui and D. Tse. What to tell customer in waits of different lengths: an integrative model of service evaluation. *Journal of Marketing*, 60:81–90, 1996.

M. Hui and L. Zhou. How does waiting duration information influence customers' reactions to waiting for services? *Journal of Applied Social Psychology*, 26:1702–1717, 1996.

U. Hwang, M. L. McCarthy, and D. Aronsky et al. Measures of crowding in the emergency department: A systematic review. *Academic Emergency Medicine*, 18:527–538, 2011.

R. Ibrahim and W. Whitt. Real-time delay estimation based on delay history. *Manufacturing & Service Operations Management*, 11:397–415, 2009.

ICMI. Extreme engagement in the multichannel contact center: Leveraging the emerging channels research Report and best practices guide. 2013. ICMI Research Report.

A. Ingolfsson, A. Akhmetshina, S. Budge, Y. Li, and X.A. Wu. Survey and experimental comparison of service level approximation methods for non-stationary M(t)/M/s(t) queueing systems. *INFORMS Journal on Computing*, 19:201–214, 2007.

F. Iravani and B. Balcıoğlu. Approximations for the $M/GI/N + GI$ type call center. *Queueing Systems*, 58(2):137–153, 2008.

D.N. Joanes and C.A. Gill. Comparing measures of sample skewness and kurtosis. *Journal of the Royal Statistical Society (Series D): The Statistician*, 47:183–189, 1998.

G. Jongbloed and G.M. Koole. Managing uncertainty in call centers using poisson mixtures. *Applied Stochastic Models in Business and Industry*, 17:307–318, 2001.

W.C. Jordan and S.C. Graves. Principles on the benefits of manufacturing process flexibility. *Management Science*, 41:577–594, 1995.

W.C. Jordan, R.R. Inman, and D.E. Blumenfeld. Chained cross-training of workers for robust performance. *IIE Transactions*, 36:953–967, 2004.

O. Jouini. Analysis of a last come first served queueing system with customer abandonment. *Computers & Operations Research*, 39:3040–3045, 2012.

O. Jouini, Y. Dallery, F. Auriol, O. Belma, R. Chauvet, F. Nait-Abdallah, and T. Prat. Client portfolio-based call center architecture. 2006. International Patent, publication number WO 2006/003306, World Intellectual Property Organization.

O. Jouini, Y. Dallery, and O.Z. Aksin. Queueing models for full-flexible multi-class call centers with real-time anticipated delays. *International Journal of Production Economics*, 120:389–399, 2009.

O. Jouini, A. Pot, G. Koole, and Y. Dallery. Online scheduling policies for multiclass call centers with impatient customers. *European Journal of Operational Research*, 207:258–268, 2010.

O. Jouini, Y. Dallery, and O.Z. Akşin. Call centers with delay information: Models and insights. *Manufacturing & Service Operations Management*, 13:534–548, 2011.

E.L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481, 1958.

S. Karlin and J. McGregor. The differential equation of birth and death processes, and the setieltjes moment problem. *Trans. Amer. Math. Soc.*, 85:489–546, 1957a.

S. Karlin and J. McGregor. The classification of birth and death processes. *Trans. Amer. Math. Soc.*, 86:366–401, 1957b.

K. Katz, B. Larson, and R. Larson. Prescription for the waiting-in-line blues: Entertain, enlighten, and engage. *Sloan Management Review*, pages 44–53, 1991.

J. Keilson. A review of transient behavior in regular diffusion and birth-death processes. part I. *Journal of Applied Probability*, 1:247–266, 1964a.

J. Keilson. A review of transient behavior in regular diffusion and birth-death processes. part II. *Journal of Applied Probability*, 1:247–266, 1964b.

J. Keilson. *Markov Chain Models - Rarity and Exponentiality.* Springer-Verlag, New York, 1979.

J. Keilson. On the unimodality of passage time densities in birth-death processes. *Statist. Neerlandica*, 35:49–55, 1981.

J. Keilson, U. Sumita, and M. Zachmann. Row-continuous finite Markov chains: Structure and algorithms. *Journal of the Operations Research Society of Japan*, 30:291–314, 1987.

M. Kijima. *Markov Processes for Stochastic Modeling.* Chapman & Hall, 1997. First Edition.

P. Kolesar. Stalking the endangered cat: A queueing analysis of congestion at automatic teller machines. *Interfaces*, 14(6):16–26, 1984.

G. Koole. A simple proof of the optimality of a threshold policy in a two-server queueing system. *Systems and Control Letters*, 26:301–303, 1995.

G. Koole and A. Pot. Technical note-a note on profit maximization and monotonicity for inbound call centers. *Operations research*, 59(5):1304–1308, 2011.

G. Koole, B.F. Nielson, and T.B. Nielson. First in line waiting times as a tool for analysing queueing systems. *Operations Research*, 60:1258–1266, 2012.

P. Kumar, M.U. Kalwani, and M. Dada. The impact of waiting time guarantees on customers waiting experiences. *Marketing Science*, 16:295–314, 1997.

Y.H. Kuo, J.M.Y. Leung, and C.A. Graham. Simulation with data scarcity: Developing a simulation model of a hospital emergency department. *Proceedings of the 2012 Winter Simulation Conference*, pages 1–12, 2012.

R.L. Larsen and A.K. Agrawala. Control of a heterogeneous two-server exponential queueing system. *Software Engineering, IEEE Transactions on*, (4):522–526, 1983.

B. Legros, O. Jouini, and G. Koole. Adaptive threshold policies for multi-channel call centers. *IIE Transactions*, 2014a. To appear.

B. Legros, O. Jouini, and G. Koole. On the scheduling of jobs in a contact center with idling times during the call service. *Working paper. Ecole Centrale Paris*, 2014b.

W. Lin and P.R. Kumar. Optimal control of a queueing system with two heterogeneous servers. *Automatic Control, IEEE Transactions on*, 29(8):696–703, 1984.

C. Maglaras and A. Zeevi. Pricing and design of differentiated services: Approximate analysis and structural insights. *Operations Research*, 53:242–262, 2005.

D. Maister. *The Psychology of Waiting Lines.* Lexington Books, Lexington MA, 113-123, 1985. Czepiel, J., Solomon, M., Suprenant, C. eds. The Service Encounter: Managing

A. Mandelbaum and R. Schwartz. Simulation experiments with M/G/100 queues in the Halfin-Whitt (QED) regime. 2002. Technical report, The Technion, Israel.

A. Mandelbaum and S. Zeltyn. The impact of customers patience on delay and abandonment: Some empirically-driven experiments with the M/M/N+G queue. *OR Spectrum*, 26:377–411, 2004.

A. Mandelbaum and S. Zeltyn. Staffing many-server queues with impatient customers: Constraint satisfaction in call centers. *Operations Research*, 57(5):1189–1205, 2009.

Y. H. Mao. Ergodic degrees for continuous-time Markov chains. *Science in China Ser. A*, 47: 161–174, 2004.

W. Marengo. Skill based routing in multi-skill call center. 2004. Working Paper. Vrije universiteit, The Netherlands.

V. Mehrotra, K. Ross, G. Ryder, and Y.P. Zhou. Routing to manage resolution and waiting time in call centers with heterogeneous servers. *Manufacturing & service operations management*, 14(1):66–81, 2012.

J. Milner and T.L. Olsen. Service-level agreements in call centers: Perils and prescriptions. *Management Science*, 54:238–252, 2008.

I. Mitrani and R. Chakka. Spectral expansian solution of a class of Markov models: Application and comparison with the matrix-geometric method. *Performance Evaluation*, 23:241–260, 1995.

S. Mondschein and G.H. Weintraub. Appointment policies in service operations: A critical analysis of the economic framework. *Production and Operations Management*, 12:266–286, 2003.

N. Munichor and A. Rafaeli. Numbers or apologies? customer reactions to telephone waiting time fillers. *Journal of Applied Psychology*, 92:511–518, 2007.

S. Nadarajah. A review of results on sums of random variables. *Acta Applicandae Mathematicae*, 103(2):131–140, 2008.

P. Naor. The regulation of queue size by levying tolls. *Econometrica*, 37:15–24, 1969.

M.F. Neuts. *Matrix-geometric solutions in stochastic models: an algorithmic approach*. Dover Publications, Revised edition, 1995.

E.E. Osuna. The psychological cost of waiting. *Journal of Mathematical Psychology*, 29(1): 82–105, 1985.

A. Pallin and R.P. Kittell. Mercy hospital: simulation techniques for ER processes. *Industrial Engineering*, 24(2):35–37, 1992.

S.A. Paul, M.C. Reddy, and C.J. DeFlitch. A systematic review of simulation studies investigating emergency department overcrowding. *Simulation*, 86:559–571, 2010.

A.I. Pazgal and S. Radas. Comparison of customer balking and reneging behavior to queueing theory predictions: An experimental study. *Computers and Operations Research*, 35:2537 – 2548, 2008.

E. Pekoz. Optimal policies for multi-server non-preemptive priority queues. *Queueing Systems*, 42:91–101, 2002.

W. H. Randolph. *Queueing Methods for Services and Manufacturing*. Prentice Hall, 1991.

T.R. Robbins and T.P. Harrison. A stochastic programming model for scheduling call centers with global service level agreements. *European Journal of Operational Research*, 207(3):1608–1619, 2010.

T.R. Robbins, D.J. Medeiros, and T.P. Harrison. Optimal cross training in call centers with uncertain arrivals and global service level agreements.

T.R. Robbins, D.J. Medeiros, and T.P. Harrison. Partial cross training in callcenters with uncertrain arrivals and global service level agreements. In *Proceedings of the 2007 Winter Simulation Conference*, 2007.

M.H. Rothkopf and P. Rech. Perspectives on queues: Combining queues is not always beneficial. *Operations Research*, 35:906–909, 1987.

L. Rozenshmidt. On priority queues with impatient customers: Stationary and time-varying analysis. Master's thesis, Technion, Israel Institute of Technology, 2007.

W. Rudin. *Real and Complex Analysis*. McGraw-Hill, 1987.

V.V. Rykov. Monotone control of queueing systems with heterogeneous servers. *Queueing Systems*, 37:391–403, 2001.

S. Saghafian, W.J. Hopp, M.P. van Oyen, J.S. Desmond, and S.L. Kronick. Patient streaming as a mechanism for improving responsiveness in emergency departments. *Operations Research*, 60(5):1080–1097, 2012.

H. Scarf. A min-max solution of an inventory problem. *In Studies in The Mathematical Theory of Inventory and Production. (K. ARROW, S. KARLIN and H. SCARF, Eds) pp 201-209. Stanford University Press, California.*, 1958.

Rolf Schassberger. *Warteschlangen*. Springer-Verlag Vienna, 1973.

E.M. Scheuer. Reliability of an m-out of-n system when component failure induces higher failure rates in survivors. *IEEE Transactions on Reliability*, 37(1):73–74, 1988.

R. Schonberger. *World Class Manufacturing: The Lessons of Simplicity Applied*. Free Press, New York, 1986. 10-11.

L.P. Seelen. An algorithm for Ph/Ph/c queues. *European Journal of Operational Reseearch*, 23: 118–127, 1986.

Z.Y. Shae, D. Garg, R. Bhose, R. Mukherjee, S. Guven, and G. Pingali. Efficient internet chat services for help desk agents. In *Services Computing, 2007. SCC 2007. IEEE International Conference on*, pages 589–596, 2007.

M. Sheikhzadeh, S. Benjaafar, and D. Gupta. Machine sharing in manufacturing systems: Total flexibility versus chaining. *International Journal of Flexible Manufacturing Systems*, 10:351–378, 1998.

H. Shen and J.Z. Huang. Interday forecasting and intraday updating of call center arrivals. *Manufacturing and Service Operations Management*, 10:391–410, 2008.

D.R. Smith and W. Whitt. Resource sharing for efficiency in traffic systems. *The Bell System Technical Journal*, 60:39–55, 1981.

P.C. Spirivulis, J.A. Da Silva, I.G. Jacobs, A.R. Frazer, and G.A. Jelinek. The association between hospital overcrowding and mortality among patients admitted via western australian emergency departments. *Medical Journal of Australia*, 184(5):208–212, 2006.

U. Sumita. On conditional passage time structure of birth-death processes. *Journal of Applied Probability*, 21:10–21, 1984.

S. Taylor. Waiting for service: The relationship between delays and evaluations of service. *Journal of Marketing*, 58:56–69, 1994.

E. Tekin, W.J. Hopp, and M.P. van Oyen. Pooling strategies for call center agent cross-training. *IIE Transactions*, 41:546–561, 2009.

T. Tezcan and J. Zhang. Routing and staffing in customer service chat systems with impatient customers. 2014. Working paper. University of Rochester.

S. Trzeciak and E.P. Rivers. Emergency department overcrowding in the united states: An emerging threat to patient safety and public health. *Emergency Medicine Journal*, 20(5): 402–405, 2003.

H. van Khuong and H.Y. Kong. General expression for pdf of a sum of independent exponential random variables. *IEEE Communications Letters*, 10(3):159–161, 2006.

R.B. Wallace and W. Whitt. A staffing algorithm for call centers with skill-based routing. *Manufacturing & Service Operations Management*, 7:276–294, 2005.

J. Walrand. A note on optimal control of a queueing system with two heterogeneous servers. *Systems and Control Letters*, 4:131–134, 1984.

Q. Wang. Modeling and analysis of high risk patient queues. *European Journal of Operational Research*, 155:502–515, 2004.

R. Wang, O. Jouini, and S. Benjaafar. Service systems with finite and heterogeneous customer arrivals. *Manufacturing & Service Operations Management*, 16:365–380, 2014.

A.R. Ward and P.W. Glynn. A Diffusion approximation for a Markovian queue with reneging. *Queueing Systems*, 43:103–128, 2003.

R. Weber. On a conjecture about assigning jobs to processors of differing speeds. *Automatic Control, IEEE Transactions on*, 38:166–170, 1993.

J. Weinberg, L.D. Brown, and J.R. Stroud. Bayesian forecasting of an inhomogeneous poisson process with applications to call center data. *Journal of the American Statistical Association*, 102:1186–1199, 2007.

W. Whitt. Approximations for the GI/G/m Queue. *Production Operation Management*, 2: 114–161, 1993.

W. Whitt. Dynamic staffng in a telephone call center aiming to immediately answer all calls. *Operations Research Letters*, 24:205–212, 1999a.

W. Whitt. Partitioning customers into service groups. *Management Science*, 45:1579–1592, 1999b.

W. Whitt. Improving service by informing customers about anticipated delays. *Management Science*, 45:192–207, 1999c.

W. Whitt. Predicting queueing delays. *Management Science*, 45:870–888, 1999d.

W. Whitt. A multi-class fluid model for a contact center with skill-based routing. *International Journal of Electronics and Communications*, 2005a. Forthcoming.

W. Whitt. Engineering solution of a basic call-center model. *Management Science*, 51:221–235, 2005b.

W. Whitt. Staffing a call center with uncertain arrival rate and absenteeism. *Production and Operations Management*, 15:88–102, 2006.

S.H. Xu, L. Gao, and J. Ou. Service performance analysis and improvement for a ticket queue with balking customers. *Management Science*, 53:971–990, 2007.

D. D. Yao and J. G. Shanthikumar. The optimal input rate to a system of manufacturing cells. *Information Systems and Operational Research*, 25:57–65, 1987.

J. Yoo. Queueing models for staffing service operations. 1996. Ph.D. Dissertation, University of Maryland.

D. Zakay. An integrated model of time estimation. *Times and Human Cognition: A Life Span Perspective*, 1989. Iris Levin and Dan Zakay, eds, Amsterdam: North Holland.

S. Zeltyn and A. Mandelbaum. Call centers with impatient customers: many-servers asymptotics of the M/M/n+G queue. *Queueing Systems*, 51:361–402, 2005.

D. Zhan and A.R. Ward. Threshold routing to trade off waiting and call resolution in call centers. *Manufacturing & Service Operations Management*, 16(2):220–237, 2014.

P.H. Zipkin, editor. *Foundations of Inventory Management*. McGraw-Hill Singapore, 2000.

E. Zohar, A. Mandelbaum, and N. Shimkin. Adaptive behavior of impatient customers in tele-queues: Theory and empirical support. *Management Science*, 48:566–583, 2002.