Modèles stochastiques pour l'aide à la décision dans les centres d'appels

Mohamed Salah Aguir

ÉCOLE CENTRALE
PARIS

THÈSE DE DOCTORAT

Spécialité
Génie Industriel

Présentée par

Mohamed Salah AGUIR

Le 11 février 2004
Pour l'obtention du

GRADE DE DOCTEUR

MODÈLES STOCHASTIQUES POUR L'AIDE À LA DÉCISION DANS LES CENTRES D'APPELS

Composition du jury

Président	Philippe Chevalier – *Professeur à l'Institut d'Administration et de Gestion*
Rapporteur	Gérard Hébuterne – *Professeur à l'Institut National des Télécommunications*
Rapporteur	Ger Koole – *Professeur à la Vrije Universiteit*
Examinateur	Fabrice Chauvet – *Responsable Optimisation et Aide à la Décision - Bouygues Telecom*
Invitée	Zeynep Akşin – *Maître de conférences à Koç University*
Codirecteur de thèse	Fikri Karaesmen – *Maître de conférences à Koç University*
Directeur de thèse	Yves Dallery – *Professeur à l'École Centrale Paris*

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

2004 - 03

À ma famille

Remerciements

Que Gérard Hébuterne et Ger Koole trouvent ici l'expression de ma gratitude pour avoir bien voulu évaluer cette thèse en acceptant d'en être les rapporteurs. Je remercie, également, Philippe Chevalier pour l'honneur qu'il me fait en présidant le jury de cette thèse.

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

Je tiens à exprimer ma vive reconnaissance à Fikri Karaesmen qui n'a pas cessé de m'aider pendant ces années et qui, avec sa culture scientifique, m'a toujours donné des conseils bénéfiques à l'avancement du travail. Je le remercie, également, pour sa joie de vivre qui m'a encouragé à progresser.

Je remercie particulièrement Zeynep Akşin qui m'a été d'une aide inestimable avec ses conseils judicieux. Je lui exprime toute ma gratitude pour le temps qu'elle m'a consacré et pour avoir suivi de près l'évolution de cette thèse, notamment dans l'épreuve difficile de la rédaction. Avec elle, je dispose d'une directrice de thèse exceptionnelle.

Je tiens à remercier la société Bouygues Telecom dont la direction Recherche a été à l'origine des problématiques étudiées dans cette thèse.

Je suis fier d'avoir travaillé avec Fabrice Chauvet dont l'intérêt pour cette thèse a contribué sensiblement à une excellente collaboration, facilitée par ses qualités humaines. Sa visibilité du fonctionnement du système, ajoutée à ses connaissances théoriques, ont fait que cette collaboration soit renforcée.

Je remercie également Rabie Nait Abdallah, Francis de Véricourt et Alexandre Meyrignac qui ont facilité le déroulement du projet. Je tiens également à remercier Eric Bouzou, Thierry Prat et Christophe Dupouy qui ont manifesté leur intérêt à ce travail.

Je garderai un excellent souvenir des discussions diverses que j'ai eues avec mes collègues du bureau Mohamed Zied Babaï, Khaled Hadj Youssef et Zied Jemai. Ils n'ont cessé de m'encourager lors des périodes difficiles. Je les remercie pour ceci et je tiens à ce qu'ils sachent que leur amitié compte, et comptera, beaucoup pour moi.

Je remercie Evren Sahin qui m'a beaucoup soutenu. Je la soutiens, aussi, pour son amitié et j'espère qu'elle terminera bientôt sa thèse et qu'elle sera la prochaine à soutenir.

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

Salah

Table des matières

Liste des figures

Liste des tableaux

Chapitre 1 : Introduction Générale

1.1 Introduction

Dans un monde industriel de plus en plus soucieux de satisfaire les attentes des clients, les centres d'appels constituent une vitrine remarquable qui permet aux entreprises de garder un lien direct avec leurs clients et forment, de cette manière, un moyen de communication idéal avec eux. En outre, la mondialisation, ainsi que la diminution de l'importance des frontières entre les pays, ont contribué à l'augmentation de la concurrence internationale. Il est primordial pour l'entreprise de conserver ses clients et d'en acquérir de nouveaux. Désormais, ce besoin n'est plus une tâche facile à effectuer étant donnée la concurrence accrue. Les clients deviennent de plus en plus exigeants et les entreprises se montrent attentives en essayant de satisfaire à leurs besoins. Dans ce contexte, les centres d'appels apportent un plus indéniable aux entreprises en leur offrant un moyen de contact rapide et convivial avec leurs clients. Les entreprises disposent, ainsi, d'un outil ayant une multitude de possibilités. En effet, les centres d'appels servent aux entreprises pour leurs pratiques commerciales. Ceci se traduit, dans certains secteurs, par la proposition, aux clients, d'une partie de l'offre par téléphone. Les avantages sont nombreux, à commencer par la réduction des coûts puisque le coût d'un centre d'appels revient moins cher qu'une agence commerciale de contact direct et ce, pour un service équivalent. Ici, il faut noter que, dans de nombreux secteurs, les clients peuvent formuler un certain nombre de requêtes, auprès des entreprises, à tout moment de la journée. Ceci se fait, notamment, grâce à des requêtes

préenregistrées qui permettent d'automatiser la réponse et qui fournissent, donc, le service aux clients à tout moment et ce, avec un coût minime. La satisfaction immédiate des demandes des clients améliore nettement leur appréciation du service fourni par l'entreprise. Ceci facilite, à la fois, l'acquisition de nouveaux clients et la fidélisation des anciens. Pour ce faire, il faut noter qu'il est indispensable que le service offert par le centre d'appels soit d'une bonne qualité. En effet, afin d'améliorer la relation entre l'entreprise et ses clients, il faut fournir un bon niveau de service qui rend les clients satisfaits. Autrement, il est possible d'engendrer la non satisfaction des clients, ce qui est contradictoire avec la raison pour laquelle le centre d'appels a vu le jour. Ceci peut être à l'origine de la migration des clients vers les entreprises concurrentes. Le mécontentement des clients résulte, généralement, de la mauvaise qualité du service fourni par les conseillers et qui ne s'aligne pas sur leurs attentes, ou encore de l'encombrement du centre d'appels, ce qui rend difficile l'obtention d'un service immédiat. Le travail décrit dans ce mémoire s'intéresse à cette deuxième problématique et s'inscrit, donc, dans le cadre de l'amélioration du service perçu par les clients dans le centre d'appels avant d'être répondus par des conseillers. Cette amélioration s'effectue sur la structure interne du centre d'appels qui correspond à la façon dont les appels des clients arrivent aux conseillers. Elle influe, également, sur les ressources du système qui, n'étant pas illimitées, doivent correspondre à un compromis entre le niveau de service donné et le coût engendré. À noter, ici, l'importance de ce compromis puisque les clients ressentent directement le niveau de service fourni et que, parallèlement, la maîtrise des ressources reste cruciale étant donnée la proportion des coûts salariaux dans un centre d'appels, et qui est estimée entre 60 % et 70 % du coût total.

Le secteur des centres d'appels a enregistré une importante croissance ces dernières années. Cette croissance est de l'ordre de 8 % par an aux États Unis. En 1999, et toujours aux États Unis, les centres d'appels emploient 1,4 % de l'ensemble des salariés du secteur privé, ce qui représente 1,55 millions de salariés. En 1998, AT&T, un opérateur téléphonique majeur aux États Unis, a déclaré que 40 % des appels qui passent par son réseau étaient des appels gratuits, passés vers des numéros spéciaux représentatifs de beaucoup de centres d'appels. Ces statistiques ont été rapportées par Gans, Koole et Mandelbaum [27] qui constitue une référence bibliographique remarquable sur les centres d'appels. En France, l'éclosion des centres d'appels fut

moins rapide que celle outre-Atlantique, en témoignent les 30400 numéros spéciaux recensés par France Télécom, et cités par Caïazzo [18], en 1998 contre les 13 millions de lignes aux États Unis à la même époque. Cependant, comme cela a été rapporté par le cabinet Cesmo, la croissance, en France, est devenue importante puisqu'elle atteint les 8,4 % en 2002 pour atteindre un total de 183000 salariés et 2900 centres d'appels de plus de 10 positions. En Europe, d'après le site Internet http://callcenternews.com/resources/stats_size.shtml, le nombre de centres d'appels devrait passer de 12750 en 1999 à 28289 en 2006.

Le centre d'appels, call center en Anglais, comme il a été défini par Babaï [9], est une entité dont la vocation est de gérer à distance la relation que les entreprises souhaitent entretenir avec leurs clients et prospects. C'est un ensemble de moyens humains, organisationnels et techniques mis en place afin d'apporter à la demande et aux besoins de chaque client une réponse adaptée. Il ne s'agit pas d'un simple plateau téléphonique, installé dans un local ergonomique, mais bien d'un pôle d'activité qui cristallise les problèmes, et enjeux de l'entreprise.

Avec la définition précédente, il est clair que le centre d'appels vise la satisfaction et, donc, la fidélisation de ses clients, tout en acquérant de nouveaux. Ceci est valable pour des centres d'appels qui opèrent dans des secteurs diversifiés. En effet, cela est applicable pour les banques, les compagnies d'assurances et les services après-vente où la relation client par téléphone a facilité, nettement, l'accès aux services pour les clients, ce qui augmente, en toute logique, leur satisfaction et leur fidélité. Ceci est également applicable pour les compagnies des chemins de fer, les agences de voyage et les hôtels. Dans ce cas, il est plus question d'acquérir de nouveaux clients qui contactent le centre d'appels dans le but de se renseigner ou de réserver, directement, un billet ou bien une chambre d'hôtel. Il est évident qu'une mauvaise qualité de service, pour ces futurs clients, entraîne, systématiquement, un manque à gagner pour l'entreprise d'où l'importance de bien concevoir le centre d'appels. Si, auparavant, le centre d'appels était réservé à des secteurs particuliers, maintenant il est devenu nécessaire dans la plupart des domaines de l'industrie. En effet, il n'est pas possible, par exemple, pour un fabriquant d'électroménager de ne pas avoir de centre d'appels pour le service après vente. Autrement, le nombre de ses clients va diminuer d'une façon sensible.

Toujours à partir de la précédente définition du centre d'appels, nous mettons l'accent sur le fait que le centre d'appels est un pôle d'activité important au sein de l'entreprise. Nous pouvons constater ceci en suivant le même raisonnement que celui exposé par Caïazzo [18]. Étant donné le contact étroit que le centre d'appels permet d'avoir avec les clients, il est logique qu'il dépende de la direction marketing. Or, le centre d'appels effectue des ventes par téléphone, chose qui le rend dépendant de la direction commerciale. Parallèlement, il y a un retour d'informations de la part des clients à propos des produits de l'entreprise, et ceci intéresse plus particulièrement la direction de la production. En étendant le raisonnement, nous pouvons trouver que la question touche, presque, toutes les directions de l'entreprise, et ceci engendre un réel problème organisationnel. Caïazzo [18] cite un exemple pour montrer l'implication de toutes les directions dans la mise en place d'un centre d'appels. Cet exemple concerne Automobiles Peugeot qui a mené, en 1999, une étude dont le but est de mettre en place un centre d'appels pour améliorer la relation client. Suite à la décision de la direction générale de l'entreprise de rendre ce projet prioritaire, cela avait impliqué les directions informatique, marketing, commerciale, financière, internationale, régionales, le service après vente et la direction des ressources humaines. Cet exemple n'est pas un cas isolé puisqu'aux États Unis, surtout, c'est assez fréquent de voir une telle importance accordée au centre d'appels au sein de l'entreprise.

Le travail réalisé dans le cadre de cette thèse a été effectué en étroite collaboration avec l'équipe de Recherche et Développement de l'opérateur de téléphonie mobile *Bouygues Telecom*. Les problématiques traitées sont la conséquence des besoins observés par le centre d'appels de l'entreprise. Elles sont toutes axées sur l'amélioration de la qualité de service perçue par les clients et ce, tout en maîtrisant les coûts. Les outils utilisés lors de ce travail se composent, nécessairement, de la théorie des files d'attente, des chaînes de Markov et de la simulation.

1.2 Description du mémoire

Le travail décrit par ce mémoire se compose, essentiellement, de deux parties. La première partie traite le phénomène de renouvellement d'appels, et la deuxième partie traite l'estimation des temps d'attente avec une application au routage et une analyse des priorités entre les classes. Dans les deux parties, ce sont les clients qui contactent le

centre d'appels et non l'inverse. Dans la réalité, les conseillers de clientèle d'un même centre d'appels peuvent recevoir des appels entrants tout en ayant la possibilité d'effectuer, eux, les appels sortants.

Dans la première partie, nous considérons que les clients du centre d'appels sont traités de la même façon dans le système. Nous étudions, dans cette partie, un phénomène bien connu qui est le renouvellement d'appels ou, tout simplement, les rappels. Les rappels sont dus aux clients qui ont essayé d'entrer en contact avec un conseiller de clientèle sans y parvenir à cause de l'encombrement du système. Dans ce cas, ces clients décident de rappeler plus tard afin d'être servis. Cette partie est divisée en deux chapitres. Dans le premier, nous montrons l'importance de la considération du phénomène de rappels puisque beaucoup de centres d'appels ne le prennent pas en compte. Nous abordons, également, le dimensionnement du système en vue de satisfaire un taux de prise en charge objectif. Ceci nous permet de constater la différence entre un modèle qui considère l'existence des rappels, et un autre qui ne le fait pas. Cette étude est réalisée au régime stationnaire. Le deuxième chapitre peut être considéré comme une extension du chapitre qui le précède. Nous y analysons le régime transitoire en plus du régime stationnaire.

Dans la deuxième partie nous considérons plusieurs classes de clients. Nous effectuons une comparaison entre un centre d'appels possédant une seule file d'attente qui conduit aux différents sites, et un système qui possède une file d'attente par site. Nous commençons cette partie par un chapitre qui aspire à estimer les temps d'attente des clients à leurs arrivées. Par la suite, cette estimation nous servira, dans le deuxième chapitre, à router les clients entre les sites qui composent le centre d'appels. En fonction d'objectifs de niveaux de service à atteindre pour chaque classe de clients, nous comparons ce système avec le système possédant une seule file d'attente. Dans le troisième chapitre de cette partie, nous nous intéressons aux règles de priorité qui différencient les classes entre elles. Nous étudions des règles de priorité probabiliste et nous les comparons avec un système où la priorité entre les classes est stricte. La comparaison s'effectue, à nouveau, en termes de conseillers nécessaires à la satisfaction des qualités de service objectif de chaque classe de clients.

1.3 Principales contributions

Dans la première partie, nous étudions le phénomène de rappels et ses répercussions sur le dimensionnement du centre d'appels. Nous avons commencé par expliquer l'importance du phénomène de rappels en montrant que les rappels peuvent constituer une partie importante des appels observés par le système. Cette étude a été menée par l'intermédiaire d'une chaîne de Markov à deux dimensions dont la résolution aboutit aux probabilités des états au régime stationnaire. Ainsi, nous avons pu expliquer l'influence de tous les paramètres sur le comportement du système. Cette étude nous permet, aussi, de dimensionner le centre d'appels selon un taux de prise en charge objectif et ce, suite au calcul préalable de la demande réelle en fonction de la demande observée. Le dimensionnement réalisé nous a permis de conclure que la non considération des renouvellements d'appels peut aboutir à un sur-dimensionnement ou bien à un sous-dimensionnement du centre d'appels traduits par des performances éloignées des objectifs. Par la suite, grâce à la méthode de l'approximation fluide, nous étudions un modèle plus générique dans lequel une estimation du temps d'attente est annoncée aux clients à leurs arrivées. Nous aboutissons à une expression robuste du taux de rappels au régime stationnaire. Cette expression est insensible à plusieurs paramètres du système, en particulier la fonction de renoncement des clients et qui peut être assez compliquée. Nous avons, également, utilisé l'approximation fluide pour analyser une journée entière pendant laquelle tous les paramètres peuvent varier en fonction du temps. En décomposant la journée en plusieurs périodes, nous déterminons l'évolution de la demande réelle au-cours du temps.

Dans la deuxième partie, nous avons étudié les temps d'attente des clients dans le centre d'appels. Désormais, nous considérons un système multi-classe dans lequel les priorités entre classes de clients sont non-préemptives. Nous avons, en premier lieu, proposé un estimateur du temps d'attente, noté ASA^2, des clients à leurs arrivées. Nous montrons que la précision de cet estimateur augmente avec la charge du système, spécialement lorsque tous les clients passent par une seule et unique file. En second lieu, pour un système composé par plusieurs files d'attente, nous analysons plusieurs règles de routage basées sur l'estimateur ASA^2. Nous montrons qu'avec un routage dynamique adéquat, la performance du système peut s'avérer meilleure que celle d'une file unique. Ceci nous amène à l'analyse des règles de priorités et leurs influences sur le

comportement du système. Dans le cas où chaque classe de clients dispose d'un objectif propre à elle, nous montrons que le fait de changer les priorités de façon statique et non dynamique, permet d'améliorer le rendement du centre d'appels. En effet, nous montrons que la performance issue des priorités probabilistes est meilleure que celle relative à la priorité stricte. Nous avons, en particulier, montré que ce genre de priorités aboutit à la capacité de service optimale, lorsqu'il s'agit de satisfaire des objectifs de service exprimés en termes de temps moyens d'attente pour chaque classe de clients. Nous avons également calculé cette capacité de service optimale ainsi que les paramètres de la priorité probabiliste dans le cas d'un système mono-serveur.

1.4 Plan du mémoire

Le mémoire est organisé comme suit :

Dans le chapitre 2, nous analysons un centre d'appels où les clients peuvent perdre patience s'ils attendent longtemps dans la file et, dans ce cas, ils abandonnent. La file d'attente est limitée en capacité. Si elle est pleine à l'arrivée d'un client, celui-ci est déconnecté du système tout en ayant la possibilité de rappeler plus tard. Notre analyse se base sur une chaîne de Markov à deux dimensions. Suite à la résolution de la chaîne de Markov au régime stationnaire, nous analysons l'évolution du comportement du système en fonction de tous les paramètres. Nous prouvons l'importance du phénomène de rappels en montrant qu'ils peuvent avoir un ordre de grandeur comparable aux appels de première intention. Par la suite, nous regardons l'effet des rappels sur le dimensionnement du système et nous prouvons que, pour des niveaux de service objectif fixés, la non considération du phénomène induit à des erreurs qui se traduisent par un sur-dimensionnement, ou par un sous-dimensionnement, du système.

Dans la chapitre 3, nous continuons à analyser le phénomène de rappels, mais avec un modèle plus compliqué que le précédent. Désormais, les clients qui abandonnent peuvent rappeler. La file d'attente possède une capacité infinie. Et, surtout, une annonce du temps d'attente est communiquée aux clients à leurs arrivées, ce qui a pour résultat la diminution du taux d'abandon. Toutefois, les clients peuvent renoncer dès qu'ils savent que leur attente n'est pas nulle et ce, suivant une fonction de renoncement

qui dépend de l'état du système. Dans cette analyse, nous nous appuyons sur la méthode de l'approximation fluide. Elle nous permet d'avoir une expression simple et robuste du taux de rappel au régime stationnaire. Cette expression est insensible à la variation de quelques paramètres du système, parmi lesquels figure la fonction de renoncement ainsi que le taux d'abandon. L'approche utilisée nous permet, également, d'avoir l'évolution du système au-cours du temps. Ceci est particulièrement intéressant lorsqu'il s'agit de déterminer l'évolution de la vraie demande au-cours de la journée à partir de la demande observée.

Le chapitre 4 est, lui, consacré à l'estimation du temps d'attente des clients à leurs arrivées. Cette estimation peut servir, notamment, au routage des clients entre les différents sites du système. Dans ce chapitre, nous introduisons un estimateur du temps d'attente issu de formules analytiques et dont la précision est validée par simulation. Nous entamons, une comparaison du système constitué d'une seule file d'attente avec le système qui en possède plusieurs.

La comparaison du système à une file unique ave le système multi-file se poursuit dans le chapitre 5. Dans ce chapitre, nous nous intéressons à l'étude du routage des appels dans le centre d'appels. La règle de routage est basée sur l'estimateur du temps d'attente mis au point précédemment. La comparaison des deux systèmes s'effectue en termes de conseillers de clientèle nécessaires à la satisfaction des objectifs de qualité de service. Nous montrons que, dans certains cas, le système multi-file peut être plus performant que le système de la file unique, en particulier si le routage dispose de plus d'informations sur l'état du système en temps réel.

Dans le chapitre 6, nous étudions deux règles de priorités probabilistes. Nous fixons toujours des objectifs de qualité de service pour chaque classe de clients. Nous prouvons que les règles de priorités probabilistes sont meilleures que la priorité stricte, en particulier lorsque les objectifs s'expriment en temps moyens d'attente. Dans ce cas, nous pouvons calculer, analytiquement, la capacité de service optimale ainsi que les paramètres de la première des deux règles de priorités. L'avantage de ces règles de priorité vient du fait que toutes les classes atteignent la qualité de service objectif ensemble, ce qui n'est pas le cas de la priorité stricte où il y a, nécessairement, une classe contraignante.

Dans le chapitre 7, nous présentons nos conclusions et nos perspectives pour le travail réalisé dans le cadre de la thèse.

Chapitre 2 : Dimensionnement dans un centre d'appels avec Abandons et Rappels

2.1 Introduction

Le travail décrit dans ce chapitre correspond à une problématique énoncée par l'opérateur de téléphonie mobile *Bouygues Telecom*. Au sein du centre d'appels de l'entreprise, il a été constaté que les demandes des clients arrivent d'une façon pas facilement compréhensible. Plus tard, ils ont penché sur le problème de renouvellement d'appels après avoir remarqué qu'un client peut essayer de contacter le centre d'appels plusieurs fois au cours d'une même journée. Après une analyse de la base de données, ils ont trouvé que le nombre d'appels est supérieur au nombre de clients qui essayent d'entrer en contact avec un conseiller de clientèle. La différence entre les deux peut même être importante et elle est due, presque exclusivement, aux clients qui ne peuvent rejoindre un conseiller et qui sont obligés de rappeler plus tard pour voir leur demande satisfaite. Bien entendu, l'arrivée des clients ainsi que les temps de service sont aléatoires ce qui a pour conséquence de ne pas toujours trouver un conseiller libre au moment de l'arrivée d'un appel. Les rappels qui vont être générés par ces clients sont, donc, dépendants du nombre de conseillers présents. Ainsi, l'utilisation de l'historique des appels donne une information insuffisante qui engendre un dimensionnement erroné. Ce dimensionnement se base sur le nombre d'appels (observés) et non sur la vraie demande représentative du nombre de clients qui

appellent. L'un des objectif de ce chapitre est d'estimer le nombre de clients à partir du nombre total des appels ce qui permettra d'avoir une information plus précise qui peut être exploitée plus tard lors du dimensionnement du système.

Notre premier objectif, dans ce chapitre, sera de montrer l'importance du phénomène de renouvellement d'appels. Pour y parvenir, nous allons démontrer qu'il peut y avoir autant de rappels que d'appels frais – les appels frais[1] étant la vraie demande des clients alors que les rappels représentent les clients non encore satisfaits et qui sont obligés de renouveler leurs appels afin de pouvoir consulter un conseiller de clientèle. Notre deuxième objectif est de dimensionner un centre d'appels en maximisant le taux de prise en charge des clients. Ce taux correspond à la proportion des appels qui ont pu joindre un conseiller de clientèle. Nous allons montrer que le fait de ne pas considérer que les appels reçus se composent de rappels en plus des nouveaux appels, peut engendrer des erreurs importantes lors du dimensionnement entraînant, ainsi, la non détermination du nombre optimal de conseillers à planifier. Ces erreurs ont une importance considérable dans un centre d'appels vu que le personnel représente entre 60 % et 70 % de son coût total.

Le centre d'appels que nous allons étudier dans ce chapitre emploie C conseillers de clientèle pour satisfaire la demande des clients. Le système traite les appels de façon équitable ce qui veut dire que les clients appartiennent tous à une seule et unique classe. La capacité totale de la file étant fixée à K. Ceci permettra d'éviter des temps d'attente très longs aux clients qui arrivent au moment où il y a un grand nombre de conseillers en attente. En contre partie, la limitation de la taille de la file va être à l'origine des déconnexions que vont subir les clients lorsqu'ils appellent à l'instant où la file est déjà pleine. Lorsqu'un conseiller termine le service d'un client, il répond au client qui a attendu le plus longtemps dans la file qui fonctionne, donc, suivant la politique "FIFO". La patience des clients est supposée limitée ce qui veut dire qu'ils peuvent abandonner après un certain temps d'attente passé dans le système. Nous faisons l'hypothèse que le système atteint le régime stationnaire assez rapidement pour pouvoir négliger l'importance du régime transitoire. Grâce à cette hypothèse, nous pourrons utiliser les probabilités stationnaires des états (de la chaîne de Markov décrivant le système). Nous supposons également que les paramètres du système (taux

[1] Également appelés appels primaires ou appels de première intention

d'arrivée des clients, temps de service, etc.) ne varient pas dans le temps. Ceci implique que l'analyse proposée dans ce chapitre, doit être effectuée plusieurs fois au cours de la journée si nécessaire et ce, pour tenir compte de la variation de la valeur des paramètres dans une même journée. Ainsi, une même journée peut être découpée en trois paliers (matin, après-midi et début de soirée) ou encore, comme c'est souvent le cas, en plusieurs paliers d'une demi-heure chacun. Comme indiqué précédemment, les paramètres seront considérés comme inchangés dans un même palier, et le temps nécessaire pour atteindre l'état stationnaire est supposé être négligeable devant la durée totale du palier.

Dans la section 2 de ce chapitre, nous présenterons un état de l'art des travaux antérieurs qui traitent du phénomène du renouvellement d'appels, des abandons ainsi que des déconnexions. Dans cette section, nous allons également aborder des publications qui concernent l'approximation du fonctionnement du système au régime transitoire. Ceci nous sera d'une grande utilité dans le prochain chapitre.

Dans la section 3, nous allons décrire le problème de renouvellement des appels. Nous y analyserons aussi la chaîne de Markov qui modélise le système et y expliciterons également une méthode pour la résolution de la chaîne obtenue. Cette analyse aboutira au calcul des probabilités des états au régime stationnaire.

Dans la 4ème section, nous passerons à l'étude du phénomène de rappel en validant, par simulation, le calcul numérique correspondant à la chaîne de Markov étudiée et en montrant que les rappels peuvent être aussi importants que les arrivées des nouveaux clients.

Dans la 5ème section, nous allons analyser l'effet des différents paramètres du système sur le taux de rappel que le centre d'appels reçoit.

Dans la section 6, nous décrirons le dimensionnement du centre d'appels en nous basant sur la chaîne de Markov déjà étudiée. Nous y comparons deux systèmes, le premier effectue le dimensionnement selon un critère de service en prenant en compte les rappels alors que le deuxième ne les considère pas. Nous y comparons également

notre méthode de dimensionnement avec une approximation connue et qui ne tient pas compte des rappels.

La section 7 illustre nos conclusions concernant l'étude du phénomène de rappel au régime stationnaire ainsi que ses répercutions sur le dimensionnement du système.

2.2 Étude bibliographique

Étant donné l'importance grandissante des centres d'appels, beaucoup de chercheurs se sont intéressés de près à la problématique. Gans, Koole et Mandelbaum [27] ont produit une excellente référence qui constitue un état de l'art détaillé autour des centres d'appels. Ils ont traité différents modèles mathématiques qui ont servi, à plusieurs reprises à modéliser les centres d'appels téléphoniques. Ces modèles font souvent recours à des modélisations stochastiques exploitant, par exemple, la théorie des files d'attente, les chaînes de Markov et les approximations fluides. Dans cette section, nous allons décrire uniquement les travaux qui ont un rapport avec le phénomène de renouvellement d'appels ainsi que les modèles qui vont être introduits dans ces deux premiers chapitres. Nous commencerons par une présentation des travaux servant à une étude du régime stationnaire et qui s'intéressent aux rappels ou aux abandons (ce qui concerne nécessairement ce chapitre). Nous allons passer ensuite aux travaux qui traitent des modèles permettant l'étude du régime transitoire et qui intègrent également la notion de renoncement des clients dès qu'ils savent qu'ils vont devoir attendre (ces travaux vont servir dans le chapitre suivant).

2.2.1 Renouvellement d'appels et régime stationnaire

Comme mentionné dans la section précédente, nous allons nous intéresser dans ce chapitre à un modèle de file d'attente qui représente un centre d'appels dans lequel les clients peuvent abandonner et rappeler plus tard. Dans ce but, nous avons abordé le problème de renouvellement d'appels téléphoniques en nous basant sur un modèle stochastique de chaîne de Markov à deux dimensions. Ce modèle s'approche d'un autre étudié par Tran-Gia et Mandjes [64] dans les réseaux de téléphonie mobile mais, contrairement à eux, le nombre de clients dans notre modèle n'est pas limité, sachant qu'ils peuvent patienter dans une file d'attente de capacité non nulle.

Plusieurs travaux ont traité des files d'attente avec abandons. En particulier, concernant les centres d'appels, Baccelli et Hebuterne [10] ainsi que Brandt et Brandt [17] ont traité le cas où les temps d'impatience suivent une loi générale et ils ont analysé les performances de systèmes qui fonctionnent de cette manière. Les temps d'impatience qui suivent des lois générales ont également été analysés dans le contexte des systèmes de télécommunications dans Boxma et de Waal [15]. Akşin et Harker [6] ainsi que Garnett, Mandelbaum et Reiman [29] ont traité l'impatience des clients pour des centres d'appels particuliers et ce, pour des temps d'impatience exponentiels.

Les modèles de files d'attente avec rappels (Retrial queues) ont été abondamment étudiés. Nous pouvons, par exemple, citer Yang et Templeton [73], Falin [23], Falin et Templeton [24], ou encore Choi et Chang [19]. De Véricourt et Zhou [22] considèrent dans leur article les rappels générés par une qualité de réponse non satisfaisante pour les clients qui sont obligés de rappeler pour avoir un meilleur service. Dans ce travail, nous allons nous restreindre aux rappels des clients qui n'arrivent pas à joindre un conseiller de clientèle.

Neuts et Rao [59] ont étudié un modèle dans lequel les clients ne passent pas par une file d'attente et sont déconnectés directement du système. Ces clients rappellent plus tard suivant une loi exponentielle. Puisqu'il n'y a pas de file d'attente, il ne peut y avoir d'abandons. Dans ce modèle, les clients ne peuvent quitter définitivement le système que lorsqu'ils sont servis. Ils finissent tous par être servis. Les auteurs de l'article ont effectué une approximation afin de résoudre numériquement la chaîne de Markov relative à leur modèle. Cette approximation consiste à limiter le nombre de clients qui génèrent les rappels. Avec cette approximation, ils obtiennent une forme particulière du "générateur infinitésimal" correspondant à la chaîne de Markov. Cette forme leur permet d'utiliser la méthode de la "matrice géométrique" pour aboutir à des résolutions numériques performantes. Les auteurs montrent que la qualité de l'approximation dépend de la charge du système et de la rapidité avec laquelle les clients renouvellent leurs appels. Généralement, la méthode de la "matrice géométrique" ne peut pas être utilisée pour une résolution exacte des files d'attente avec rappels.

Le modèle que Neuts et Rao [59] ont analysé ne tient pas compte des abandons. En fait, la plupart des modèles qui existent dans la littérature ne considèrent pas le phénomène d'abandons. Hoffman et Harris [35] ont combiné rappels et abandons dans un modèle destiné, comme dans notre cas, à estimer la demande réelle à partir de l'historique contenant les données enregistrées auparavant. Dans le prochain chapitre, nous visons une estimation plus précise des appels frais à partir de cet historique tout en analysant l'impact des rappels sur la performance du centre d'appels comme cela a été fait pour le cas des abandons par Garnett, Mandelbaum et Reiman [29]. L'étude menée dans ce chapitre fait partie du travail fait par Aguir *et al.* [5].

2.2.2 Renouvellement d'appels et régime transitoire

Dans le prochain chapitre, nous allons étudier un centre d'appels qui intègre, en plus des rappels et des abandons analysés dans ce chapitre, le renoncement des clients à joindre un conseiller de clientèle dès qu'ils savent qu'ils sont obligés d'attendre avant d'être répondus. Ainsi, ces clients mettent fin à leurs appels par leur propre initiative. L'étude de ce centre d'appels va être menée dans le régime stationnaire ainsi que dans le régime transitoire.

Dans cette section, nous allons présenter les travaux complémentaires à ceux qui ont été présentés dans la section précédente pour l'étude du centre d'appels décrit ci-dessus. Ces travaux concernerons donc, essentiellement, les renoncements et le régime transitoire.

Un nombre important de travaux ont tenu compte des renoncements dans leurs modélisations des centres d'appels. Artalejo [8] a considéré un système multi-serveur avec renoncements et rappels. Les clients qui trouvent tous les serveurs occupés sont supposés renoncer à joindre le système et ce, avec une probabilité qui dépend du nombre de clients qui attendent dans la file. Un client ayant renoncé rappelle plus tard après un délai exponentiel. L'auteur analyse la performance du système en se basant sur l'approximation RTA (Retrials see Time Averages). Avec cette approximation, les clients qui rappellent sont supposés voir "des temps moyens" ou, en d'autres termes, le nombre moyen de clients ayant déjà renoncé mais qui n'ont pas encore rappelé, ne dépend pas du nombre de clients en attente. Cette approximation facilite la résolution

de la chaîne de Markov résultante au régime stationnaire. Le modèle que l'auteur a utilisé ne considère pas les abandons tout comme Greenberg et Wolff [32] qui ont introduit l'approximation. Ces derniers ont utilisé ladite approximation dans un autre système intégrant les renoncements et ce, dans le but de trouver une borne supérieure à la performance du système. Les systèmes modélisés par Whitt [70] combinent, eux, renoncements et abandons. L'objectif de l'article est de comparer la performance de deux systèmes : dans le premier une information sur l'état du système est communiquée aux clients à leur arrivée, et dans le deuxième aucune information n'est donnée. Dans le système où aucune information n'est annoncée aux clients, une proportion des clients renoncent au service alors que ceux qui décident de rester peuvent abandonner plus tard. Dans le système où l'on communique une information sur l'état, les clients renoncent avec une probabilité plus importante de telle sorte à ce que tous les abandons sont remplacés par des renoncements. Notre modèle ressemble au cas où l'on fournit une information aux clients à leurs arrivées. Toutefois, dans notre système les clients peuvent renoncer, abandonner et rappeler plus tard. Hui et Tse [39] ont réalisé une étude expérimentale afin d'analyser l'effet de l'annonce aux clients (de centres d'appels ou d'autres systèmes de services) d'une quelconque forme de temps d'attente ou d'une information liée à l'état du système. Ils ont montré que l'information appropriée dépend du temps d'attente. Ainsi, pour une longue attente il vaut mieux annoncer la position dans la file pour une meilleure perception du service. Pour une attente moyenne, il vaut mieux annoncer la durée estimée. Et pour une faible attente, il est inutile d'annoncer la moindre information. Dans notre étude, nous supposons qu'une estimation du temps moyen d'attente est annoncée aux clients. Maglaras et Armony [52] ont considéré l'impact d'une annonce de l'estimation du temps d'attente sur l'option d'appel reporté ou de service de rappel de la part d'un conseiller de clientèle au lieu d'un rappel fait par le client lui même.

Jusqu'à présent, tous les modèles décrits étudient des systèmes dont les paramètres sont supposés stationnaires ou encore des systèmes pouvant être analysés comme étant des systèmes stationnaires mono-période. En fait, il n'est même pas certain d'arriver à un régime stationnaire compte tenu de la variabilité de plusieurs paramètres au cours du temps. Mandelbaum *et al.* [53] considèrent un système multi-serveur avec rappels et abandons pour lequel ils proposent une approximation fluide qui leur sert à étudier le comportement du modèle au régime transitoire. Leur modèle diffère du nôtre

(présenté au chapitre prochain) par le fait qu'il n'intègre pas les renoncements et que les files dans leur modèle sont toujours de capacité infinie. Justement, grâce à une propriété d'insensibilité que nous allons démontrer dans le deuxième chapitre, nous montrons que l'approximation que nous avons développée pour des systèmes de capacité infinie marche bien également lorsqu'il n'y a pas de renoncements et lorsque la capacité de la file d'attente est limitée.

Green, Kolesar et Svoronos [31] ont étudié l'effet de la non-stationnarité sur les performances d'un système multi-serveur sans abandons ni rappels et ce, lorsque les arrivées suivent la loi de Poisson avec un taux qui varie de façon sinusoïdale au cours du temps. Ils ont discuté de la pertinence de l'estimation d'un modèle non-stationnaire par un modèle stationnaire. Là, il faut savoir à quel degrés les arrivées peuvent être non-stationnaires. Puisque le taux d'arrivée varie d'une manière sinusoïdale dans le temps, ils ont estimé que le "degrés de non-stationnarité" des arrivées dépend de l'amplitude de la fonction sinusoïdale ainsi que de la fréquence des évènements par cycle. En plus de ces deux facteurs, ils ont étudié la pertinence de l'estimation en fonction de la charge du système et de sa taille. Toujours dans le cadre d'arrivées non stationnaires, Green et Kolesar [30] ont proposé "The Lagged PSA" pour l'estimation du retard du pic de congestion dans un système multi-serveur avec un taux d'arrivée cyclique dans le temps. Pour un tel système, le pic des arrivées et le pic relatif au nombre de clients dans le système (ou à la probabilité d'attente) n'arrivent pas au même moment. Il y a un certain retard entre les deux. Les auteurs utilisent notamment cette méthode pour le dimensionnement du système en minimisant la probabilité d'attente. Jennings *et al.* [41] se sont également intéressés aux arrivées non-stationnaires dans le but de déterminer, au cours du temps, le nombre de serveurs qu'il faut avoir pour atteindre une probabilité d'attente objectif. Pour cela, ils se sont basés sur des approximations avec un nombre de serveurs infini. Dans le prochain chapitre, nous allons considérer des arrivées non-stationnaires dans notre analyse de l'impact des rappels. Ainsi, en ce qui concerne le régime transitoire, nous allons nous orienter plutôt vers une approximation fluide et nous montrerons sous quelles conditions cela représente une estimation précise de la performance réelle.

2.3 Le problème de renouvellement des appels

2.3.1 Description du problème

Considérons un centre d'appels téléphoniques disposant de *C* conseillers de clientèle. Nous supposons que l'arrivée des clients est poissonienne et que les temps de service des conseillers, qui se composent d'un temps de conversation et d'un temps de traitement supplémentaire pour la clôture du dossier, suivent des lois exponentielles indépendantes. Les clients sont tous traités de la même façon et ils appartiennent à une seule et unique classe. Ce centre d'appels peut donc être modélisé par une file d'attente M/M/C de taux d'arrivée λ et de taux de service μ par conseiller de clientèle. En pratique, les clients qui veulent accéder au centre d'appels peuvent abandonner en mettant fin à leur appel par leur propre initiative après avoir attendu relativement longtemps. Afin d'éviter de très longues attentes, certains centres d'appels limitent la taille de leur file d'attente ce qui évite le mécontentement des clients ayant à attendre longtemps. Ainsi, s'il y a déjà beaucoup de clients en attente lors de l'arrivée d'un nouveau client, alors celui-ci va être déconnecté immédiatement au lieu d'attendre le service de toutes les personnes venues avant lui. Le système ainsi modélisé peut être représenté par une file M/M/C/K + M où *K* désigne la taille totale de la file (*K-C* serait donc la capacité de la file d'attente) et le M supplémentaire désignant la loi markovienne qui représente les abandons des clients de la file. Nous supposons qu'un client abandonne la file d'attente par impatience s'il n'accède pas à un serveur avant une certaine durée correspondante à sa patience. Cette durée est supposée suivre une loi exponentielle de taux θ que nous désignerons, dans la suite, par taux d'abandon.

Dans cette modélisation intégrant abandons et rappels, les clients non satisfaits auront toujours la possibilité de renouveler leurs appels plus tard. Ceci va affecter les appels qui arrivent au système (appels observés) qui seront, désormais, composés de nouveaux appels (ou appels "frais") et de rappels. Nous supposons que les clients ayant abandonné ne veulent pas rappeler par la suite alors que les clients déconnectés vont le faire avec une certaine probabilité *p* (appelée probabilité de rappel) et ce, après un délai exponentiel de taux δ. La Figure 2.1 illustre le fonctionnement du centre d'appels comme nous venons de le décrire.

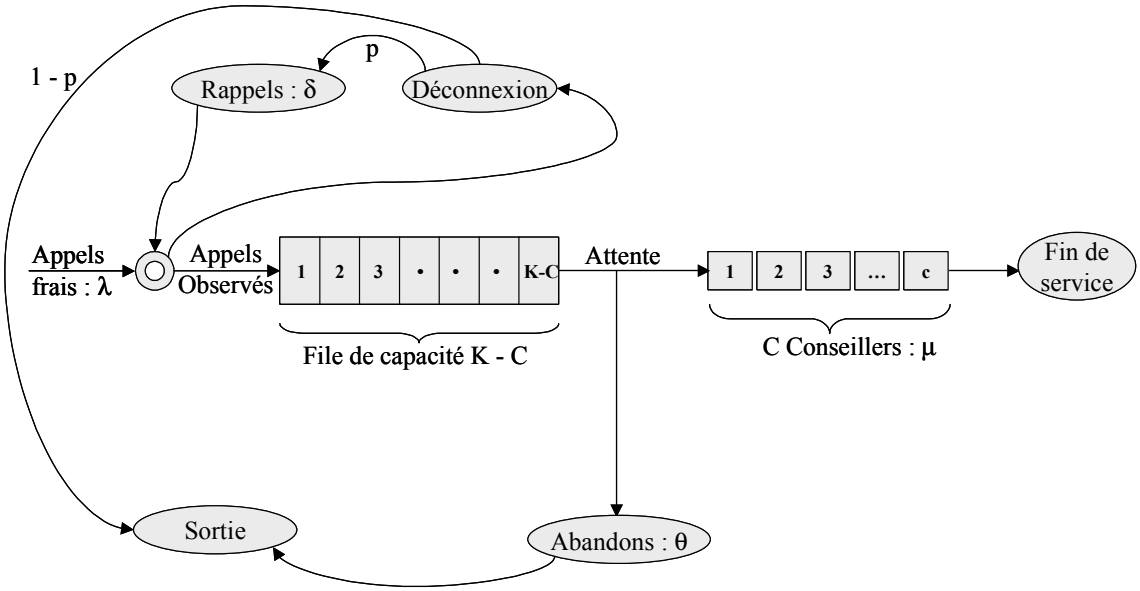


Figure 2.1: Schéma d'un centre d'appels avec abandons, déconnexions et rappels

2.3.2 Modélisation sous forme de Chaîne de Markov à Temps Continu (CMTC)

Le problème de renouvellement des appels illustré par la Figure 2.1 peut être modélisé sous forme de chaîne de Markov à temps continu. Dans cette modélisation (voir Figure 2.2), nous avons représenté l'ensemble des états en deux dimensions. La première dimension correspond à la file réelle (qui se compose des C serveurs et de la file d'attente) où le nombre de clients présents ne peut dépasser K_1, capacité de la file. La deuxième dimension correspond, elle, aux clients déconnectés et décidés à rappeler plus tard. Cette dimension est appelée en littérature "file fictive" ou encore "orbite". Ainsi, l'état (m,n), m=0,1,2,…,K_1, n=0,1,2,…,∞, implique l'existence de m clients dans la file réelle et de n clients en orbite et qui vont rappeler plus tard suivant une loi exponentielle de taux δ.

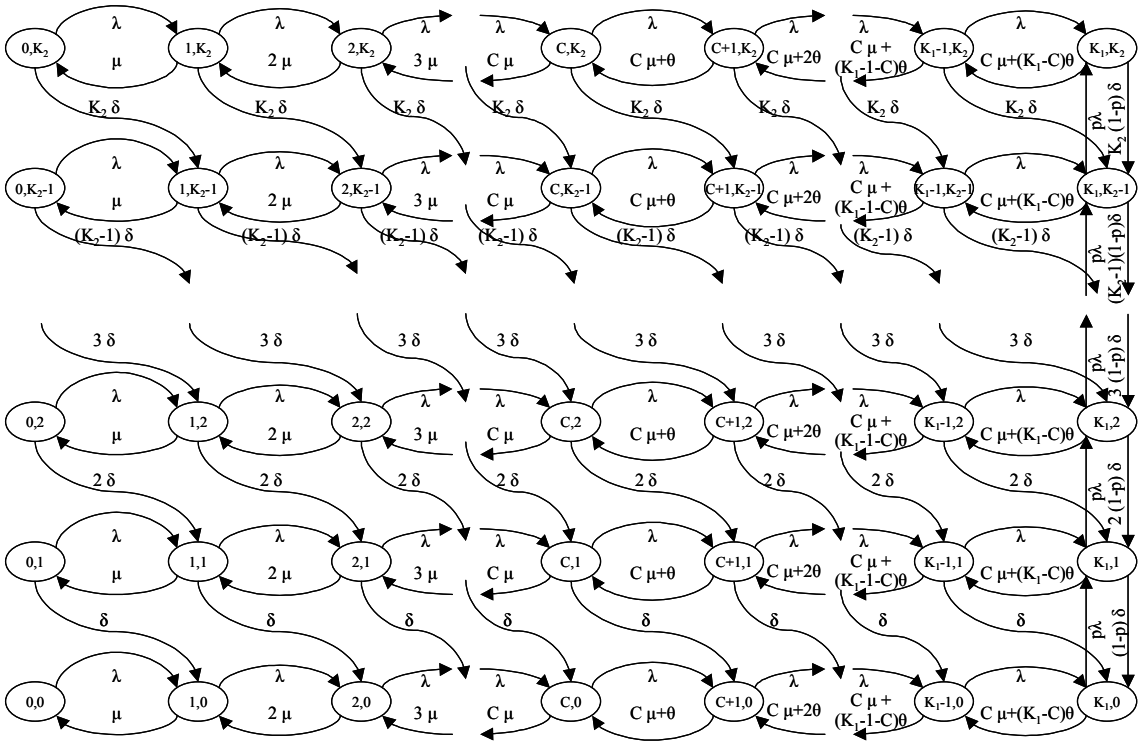


Figure 2.2: Modélisation du problème sous forme d'une CMTC

La chaîne de Markov obtenue ci-dessus est toujours stable pour $p < 1$, même si la charge du système ($\rho = \lambda/(C\mu)$) est supérieure à 1. Cette stabilité vient du fait que plus il y a des clients en orbite plus le nombre de clients qui abandonnent ou quittent le système après déconnexion est grand. À noter aussi que la chaîne est finie dans la dimension de la file réelle (elle est limitée à K_1 places) et infinie dans celle de l'orbite.

2.3.3 Analyse de la chaîne de Markov en régime stationnaire

Étant donné la dimension infinie de l'orbite, la chaîne de Markov obtenue est tronquée à une taille égale à K_2 prise suffisamment grande pour pouvoir négliger les états correspondants à une orbite supérieure à cette valeur. Cette CMTC obéit aux taux de transition $P_{(m,n)(i,j)}$ relatifs aux passages des états (m,n) aux états (i,j). Les taux de transition non nuls sont représentés par les formules suivantes :

– Si $n = 0$:

$$\begin{cases} P_{(m,n)(m+1,n)} = \lambda & si\ m < K_1 \\ P_{(m,n)(m-1,n)} = m\ \mu & si\ 0 < m \le C \\ P_{(m,n)(m-1,n)} = C\ \mu + (m - C)\theta & si\ C < m \le K_1 \\ P_{(m,n)(m,n+1)} = p\ \lambda & si\ m = K_1 \end{cases} \quad (2.1)$$

– Si $0 < n \le K_2$:

$$\begin{cases} P_{(m,n)(m+1,n)} = \lambda & si\ m < K_1 \\ P_{(m,n)(m+1,n-1)} = n\ \delta & si\ m < K_1 \\ P_{(m,n)(m-1,n)} = m\ \mu & si\ 0 < m \le C \\ P_{(m,n)(m-1,n)} = C\ \mu + (m - C)\theta & si\ C < m \le K_1 \\ P_{(m,n)(m,n+1)} = p\ \lambda & si\ m = K_1 \\ P_{(m,n)(m,n-1)} = n\ (1 - p)\ \delta & si\ m = K_1 \end{cases} \quad (2.2)$$

– Si $n = K_2$:

$$\begin{cases} P_{(m,n)(m+1,n)} = \lambda & si\ m < K_1 \\ P_{(m,n)(m+1,n-1)} = n\ \delta & si\ m < K_1 \\ P_{(m,n)(m-1,n)} = m\ \mu & si\ 0 < m \le C \\ P_{(m,n)(m-1,n)} = C\ \mu + (m - C)\theta & si\ C < m \le K_1 \\ P_{(m,n)(m,n-1)} = n\ (1 - p)\ \delta & si\ m = K_1 \end{cases} \quad (2.3)$$

Le calcul des probabilités stationnaires $\Pi_{m,n}$ des états (m,n) peut être effectué par plusieurs méthodes. Nous avons utilisé une méthode dérivée de celle qui a été décrite par Tran-Gia et Mandjes [64]. Ces derniers ont analysé un système pour lequel le nombre total de clients est constant (le taux d'arrivée est donc dépendant du nombre de clients dans le système). Une autre différence entre les deux modèles vient de l'existence d'une file d'attente dans le nôtre. Pour étudier le système au régime stationnaire, ils ont développé un algorithme récursif qui calcule les probabilités d'états pour un nombre n fixé de clients dans l'orbite. Dans l'algorithme qu'ils ont utilisé, n varie de K_2 à 0 et, pour chaque valeur de n, la probabilité de l'état (m,n), $1 \le m \le K_1$, est calculée en fonction de tous les états (i,n), $0 \le i < m$, et également de tous les états (i,n+1) pour $n < K_2$. Finalement, le régime stationnaire sera déterminé en fonction de l'état $(0,K_2)$.

L'algorithme, comme nous l'avons utilisé est décrit ci-dessous:

- Nous commençons par calculer les probabilités des états (m,n) pour $n = K_2$. Pour ce faire, il faut remarquer qu'aucun état de la ligne supérieure de la Figure 2.2 n'admet de

flèche rentrante à partir de ligne du dessous. De cette manière, nous calculons toutes les probabilités de cette ligne grâce aux quatre formules suivantes:

$$\Pi_{1,K_2} = \frac{1}{\mu}\left(\lambda + K_2\delta\right)\ \Pi_{0,K_2} \tag{2.4}$$

$$\Pi_{m,K_2} = \frac{1}{m\ \mu}\left(\left(\lambda + K_2\delta + (m\text{-}1)\mu\right)\ \Pi_{m\text{-}1,K_2} - \lambda\ \Pi_{m\text{-}2,K_2}\right),\ 2 \le \mathrm{m} \le \mathrm{C} \tag{2.5}$$

$$\Pi_{C+1,K_2} = \frac{1}{C\mu + \theta}\left(\left(\lambda + K_2\delta + C\mu\right)\ \Pi_{C,K_2} - \lambda\ \Pi_{C-1,K_2}\right) \tag{2.6}$$

$$\Pi_{m,K_2} = \frac{1}{C\mu + (m-C)\theta}\left(\left(\lambda + K_2\delta + C\mu + (m-C-1)\theta\right)\ \Pi_{m\text{-}1,K_2} - \lambda\ \Pi_{m\text{-}2,K_2}\right),$$

$C+2 \le m \le K_1$ (2.7)

Avec les formules (2.4) à (2.7), nous pouvons déterminer la ligne supérieure de la Figure 2.2 (et qui correspond à une orbite de K_2 clients) en fonction de l'état $(0,K_2)$.

▪ Nous passons maintenant au calcul des probabilités des états correspondant à *n* clients en orbite, $0 \le n < K_2$. Nous allons répéter ce calcul en faisant varier *n* de $K_2 - 1$ jusqu'à 0.

Le calcul des probabilités des états se fait à l'aide des formules suivantes:

$$\Pi_{1,n} = \frac{1}{\mu}\left(\lambda + n\delta\right)\ \Pi_{0,n} \tag{2.8}$$

$$\Pi_{m,n} = \frac{1}{m\ \mu}\left(\left(\lambda + n\delta + (m\text{-}1)\mu\right)\ \Pi_{m\text{-}1,n} - \lambda\ \Pi_{m\text{-}2,n} - (n+1)\delta\ \Pi_{m\text{-}2,n+1}\right),\quad 2 \le \mathrm{m} \le \mathrm{C}$$

(2.9)

$$\Pi_{C+1,n} = \frac{1}{C\mu + \theta}\left(\left(\lambda + n\delta + C\mu\right)\ \Pi_{C,n} - \lambda\ \Pi_{C-1,n} - (n+1)\delta\ \Pi_{C-1,n+1}\right) \tag{2.10}$$

$$\Pi_{m,n} = \frac{\left(\lambda + n\delta + C\mu + (m-C-1)\theta\right)\ \Pi_{m\text{-}1,n} - \lambda\ \Pi_{m\text{-}2,n} - (n+1)\delta\ \Pi_{m\text{-}2,n}}{C\mu + (m-C)\theta},$$

$C+2 \le m \le K_1$ (2.11)

Comme nous pouvons le voir dans les formules (2.8) à (2.11), le calcul des probabilités relatives à *n* clients dans l'orbite dépend également de la probabilité de

l'état (0,n). Ceci est, en particulier, valable pour le calcul de la probabilité de l'état (K_1,n). Or, le calcul relatif à cet état peut être effectué à partir de :

$$\Pi_{K_1,n}=\begin{cases}\dfrac{(C\mu+(K_1-C)\theta+K_2(1-p)\delta)\,\Pi_{K_1,K_2}-\lambda\,\Pi_{K_1-1,K_2}}{p\,\lambda} & si\ n=K_2-1\\ \dfrac{(C\mu+(K_1-C)\theta+(n+1)(1-p)\delta+p\lambda)\,\Pi_{K_1,n+1}-\lambda\,\Pi_{K_1-1,n+1}-(n+2)(1-p)\delta\,\Pi_{K_1,n+2}-(n+2)\delta\,\Pi_{K_1-1,n+2}}{p\,\lambda} & \text{sinon}\end{cases}$$

(2.12)

À partir de cette dernière équation, nous pouvons remonter à la valeur correspondante à l'état (0,n) à l'aide d'une simple division. Cette dernière dépend encore de l'état (0,K_2).

- L'étape précédente est répétée jusqu'à la valeur n = 0 ce qui nous permet d'avoir le fonctionnement du système en fonction de l'état (0,K_2). Maintenant, avec la condition de normalisation qui spécifie que la somme de toute les probabilités est égale à 1, nous obtenons les vraies valeurs des probabilités des états en divisant chaque probabilité par la somme de toutes les valeurs. Ainsi, la probabilité de l'état (m,n) au régime stationnaire devient égale à :

$$\frac{\Pi_{m,n}}{\sum_{\substack{0\le i\le K_1\\ 0\le j\le K_2}}\Pi_{i,j}},\ 0 \le m \le K_1,\ 0 \le n \le K_2,$$

où $\Pi_{m,n}$ se calcule à l'aide des formules (2.4) à (2.12).

L'algorithme décrit précédemment nous permet de déterminer toutes les probabilités d'états au régime stationnaire. Sa rapidité dépend, bien entendu, de la taille de la chaîne de Markov que l'on cherche à résoudre. Il reste, tout de même, assez rapide pour des tailles de centres d'appels réels. Sa complexité est proportionnelle au produit $K_1.K_2$ ce qui explique sa performance.

2.4 Étude numérique du phénomène de rappel

Dans cette section, nous validons, par simulation, la troncation de la CMTC faite dans le but d'avoir une bonne estimation des paramètres du système. Nous montrons par la

suite l'importance que peut avoir le phénomène de rappel à l'aide du même exemple utilisé pour la validation.

Afin de valider le modèle stochastique introduit précédemment, une série de simulations a été effectuée. La Figure 2.3 et la Figure 2.4 montrent l'évolution du taux stationnaire de rappel et représentent le nombre moyen de renouvellements d'appels par unité de temps en fonction du taux d'appel frais. Le taux stationnaire de rappel étant défini par :

$$\text{Taux stationnaire de rappel} = \sum_{n=1}^{K_2} n \, \delta \sum_{m=0}^{K_1} \Pi_{m,n} \qquad (2.13)$$

Dans ces deux figures, nous avons étudié deux centres d'appels semblables. L'un (100 conseillers) est quatre fois plus grand que l'autre (25 conseillers). Dans cette étude, nous avons fait varier le taux d'appel frais λ pour couvrir une charge allant de 70 % à 130 %. La capacité K_1 de la file d'attente a été choisie pour chaque exemple de façon à ce que l'espace d'attente soit réaliste par rapport au nombre de serveurs. Les autres paramètres utilisés pour ces deux exemples sont issus de données réelles. Nous constatons que les courbes obtenues avec le modèle stochastique coïncident avec celles que donne la simulation d'où la fiabilité des calculs effectués et de la troncation effectuée.

Ces mêmes figures nous montrent que pour des charges supérieures à 1, nous obtenons des taux stationnaires de rappel très proches du taux d'appel frais (taux de rappel de l'ordre de 80 % du taux d'appel frais pour une charge de 130 %). Ainsi, pour le premier exemple, le taux de rappel est de 31 pour un taux d'appel frais égal à 39 ce qui fait un taux d'appel observé de 70. Pour le deuxième exemple, le taux de rappel est de 7,8 pour un taux d'appel primaires de 9,8 (d'où un taux d'appel observé de 17,6). L'allure de la Figure 2.3 est similaire à celle de la Figure 2.4. La différence majeure entre les deux vient de la partie qui illustre le comportement pour une faible charge. Nous remarquons que, pour le premier exemple, la courbe reste quasiment collée à l'axe des abscisses jusqu'à atteindre une charge de 93 % (pour laquelle $\lambda = 28$). Concernant le deuxième exemple, la courbe commence à remonter dès que le taux d'appel frais atteint 5,8 ce qui donne une charge de 77 %. Cette différence de comportement due à la taille

du centre d'appels joue un rôle important dans l'étude que nous menons dans le prochain chapitre. Elle s'explique par une diminution de la variabilité du système en fonction de sa taille. Ainsi, dans un gros centre d'appels, la variabilité peut être suffisamment faible pour considérer son comportement comme continue (et non plus aléatoire). De plus, lorsque la charge augmente, l'effet de la variabilité diminue et ce, quelque soit la taille du centre d'appels.

Plus tard, lors du dimensionnement du centre d'appels, nous allons voir que le fait d'ignorer le phénomène de rappel induit des erreurs importantes. En effet, si les clients arrivent avec un taux d'appel égal à 70 alors il faudra mettre beaucoup plus de conseillers qu'il ne le faut pour satisfaire cette demande puisque la demande réelle arrive avec un taux de 39 appels par unité de temps. Ces erreurs impliqueront donc un surcoût dans le fonctionnement du centre d'appels dû aux conseillers supplémentaires.

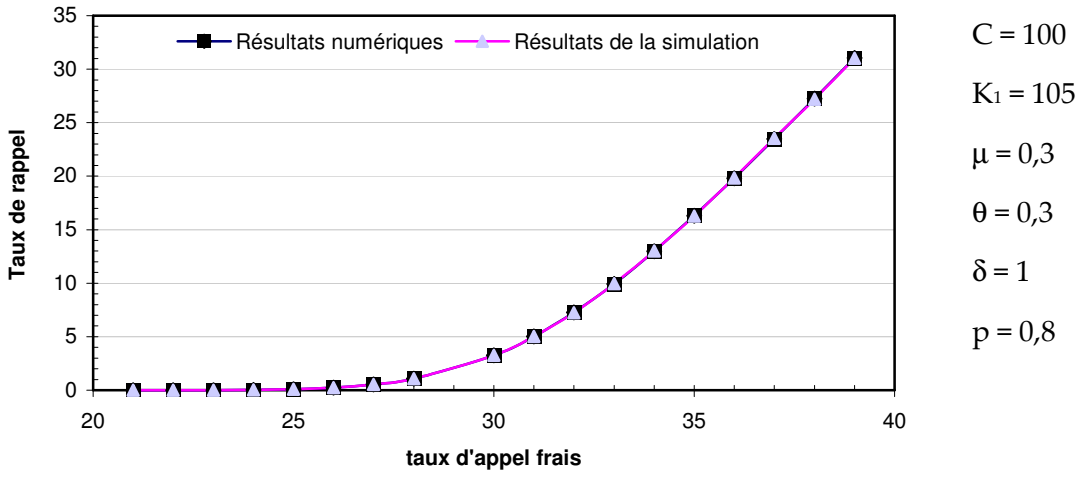
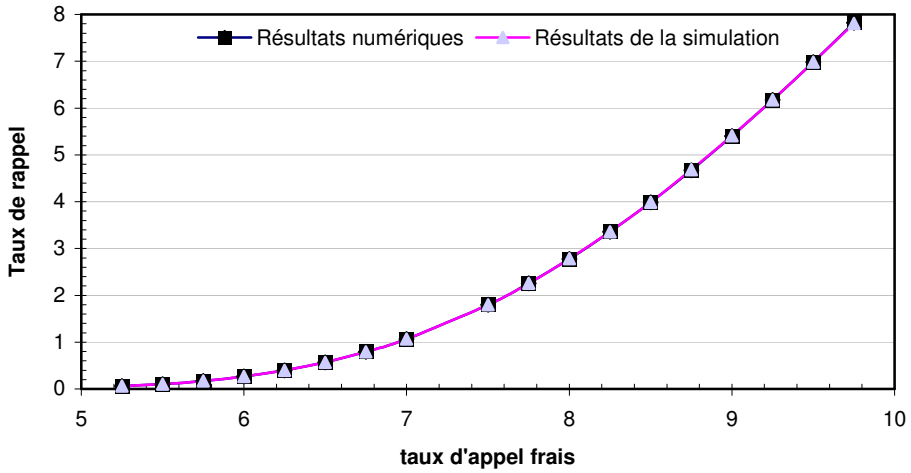


Figure 2.3: Évolution du taux de rappel en fonction du taux d'appel frais



C = 25

K_1 = 27

μ = 0,3

θ = 0,3

δ = 1

p = 0,8

Figure 2.4: Évolution du taux de rappel en fonction du taux d'appel frais

2.5 Effet des paramètres sur le comportement du système

Après avoir validé la méthode de calcul utilisée pour la résolution de la chaîne de Markov, nous étudions maintenant les effets des paramètres du système sur son comportement. Ces paramètres peuvent être séparés en deux catégories: paramètres de type capacitaire (que le manager peut changer directement), et paramètres de type comportemental (qui ne peuvent pas être changés directement parce qu'ils sont relatifs au comportement des clients). La première catégorie se compose du nombre de serveurs *C* et de la capacité de la file d'attente $K_1 - C$. Les paramètres qui représentent le comportement des clients sont: la probabilité statique de rappel *p*, le taux de service μ, le taux d'abandon θ, la taux de rappel individuel δ et le taux d'appel frais λ. Nous n'allons pas regarder le rôle de ce dernier paramètre puisque nous l'avons déjà vu dans la Figure 2.3 et la Figure 2.4. Nous n'allons pas, non plus, nous préoccuper du taux de service μ que nous allons fixer pour le reste de l'étude à 0,3 (ce qui correspond à une valeur réaliste d'un peu plus de 3 minutes de service par appel). En fait, fixer μ n'a pas beaucoup d'importance dans l'étude puisque ce qui compte le plus c'est l'importance relative de μ par rapport aux autres paramètres. En d'autres termes, ce qui est important ce sont les rapports λ / μ, θ / μ et δ / μ.

Dans la Figure 2.5, nous affichons l'évolution du taux de rappel au régime stationnaire en fonction de la probabilité de rappel statique pour un système où le

nombre de serveurs $C = 100$ et l'espace d'attente est limité à 5 places. Dans cette figure, nous constatons que, pour toutes les valeurs de λ considérées, le taux de rappel est strictement croissant en fonction de p – ce qui est logique puisque les clients déconnectés rappellent avec cette probabilité – et de plus sa progression est de plus en plus importante à mesure que p augmente. Ceci implique qu'il est très important pour un centre d'appels d'estimer la valeur de p avec précision dans le cas où cette dernière est élevée parce qu'une petite erreur dans ce cas engendre une imprécision assez importante dans l'estimation de p. Pour de faibles valeurs de p, la précision lors de l'estimation de ce paramètre n'est pas aussi importante, et de plus le taux de rappel n'est pas élevé pour ces valeurs.

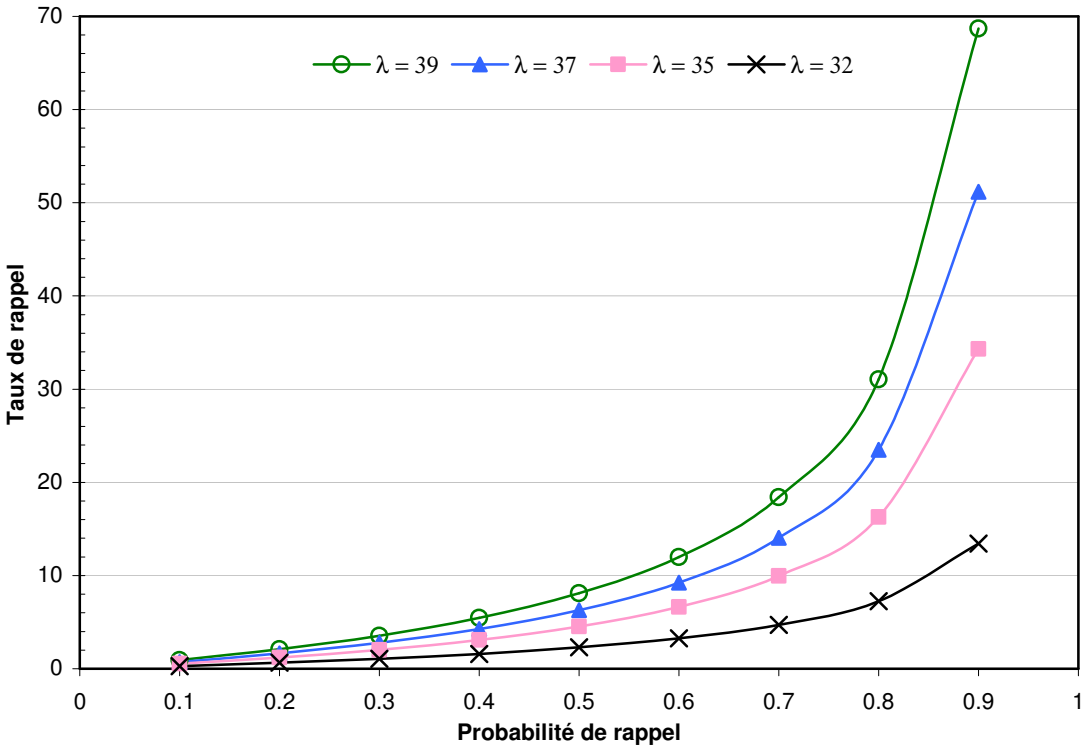


Figure 2.5: Effet de la probabilité de rappel sur le taux de rappel

La Figure 2.6 montre l'influence du taux de rappel individuel δ sur le taux de rappel stationnaire. Nous y constatons que ce dernier augmente très légèrement en fonction de δ. En observant le modèle stochastique de la Figure 2.2, nous constatons que δ joue un double rôle. En effet, plus δ est élevé, plus courte est la taille de l'orbite. En contre partie, les clients qui se trouvent dans l'orbite vont rappeler plus vite, ce qui

compense la diminution de la taille de la file fictive. Ces deux effets combinés nous donnent l'allure "presque constante" de la Figure 2.6. Il est intéressant de noter ici que le taux de rappel est moins sensible à une variation de δ lorsque le taux d'appel frais est important. Dans le prochain chapitre, nous allons voir que lorsque la charge est élevée, le taux de rappel stationnaire est insensible à plusieurs paramètres, parmi lesquels figure δ.

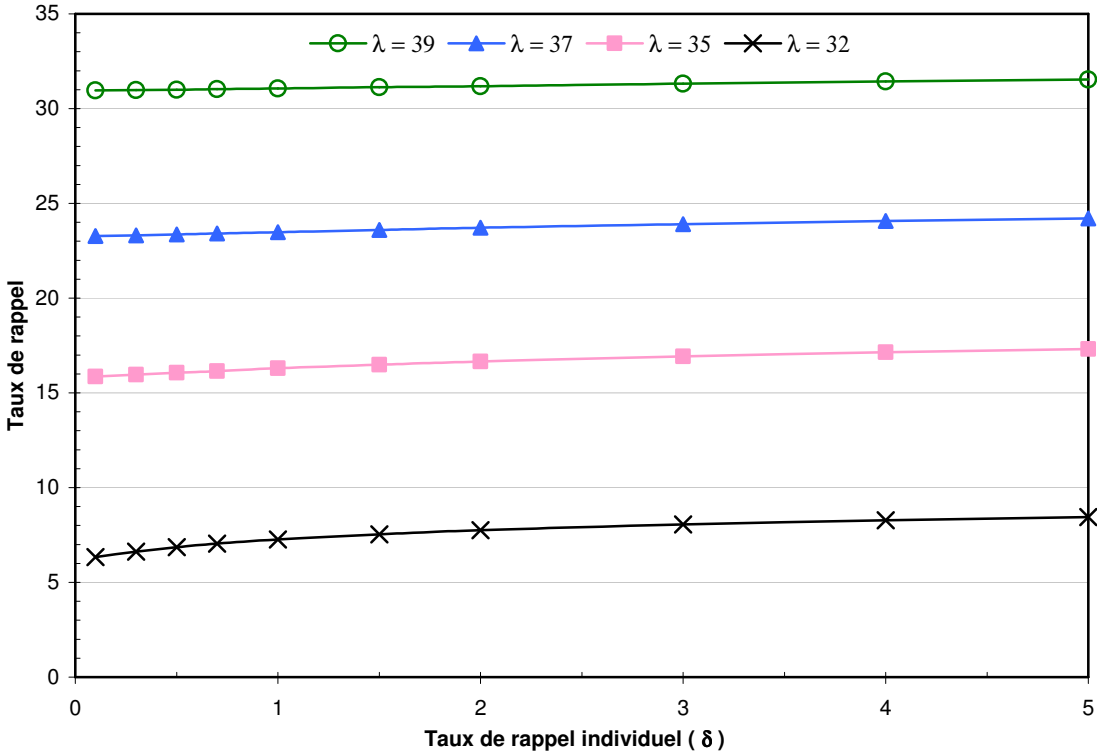


Figure 2.6: Effet de δ sur le taux de rappel

Comme nous pouvons le constater à partir de la Figure 2.7, le taux de rappel décroît en fonction de θ. En fait, le rôle que joue θ est prévisible. Plus il y a d'abandons dans notre modèle, moins il y a de rappels puisque ceux-ci proviennent uniquement des déconnexions qui ne peuvent avoir lieu qu'une fois la file réelle est pleine ce que ne le permet pas un taux d'abandon élevé. Ainsi, plus θ est grand, moins il y aura de rappels et c'est ce que donne la Figure 2.7.

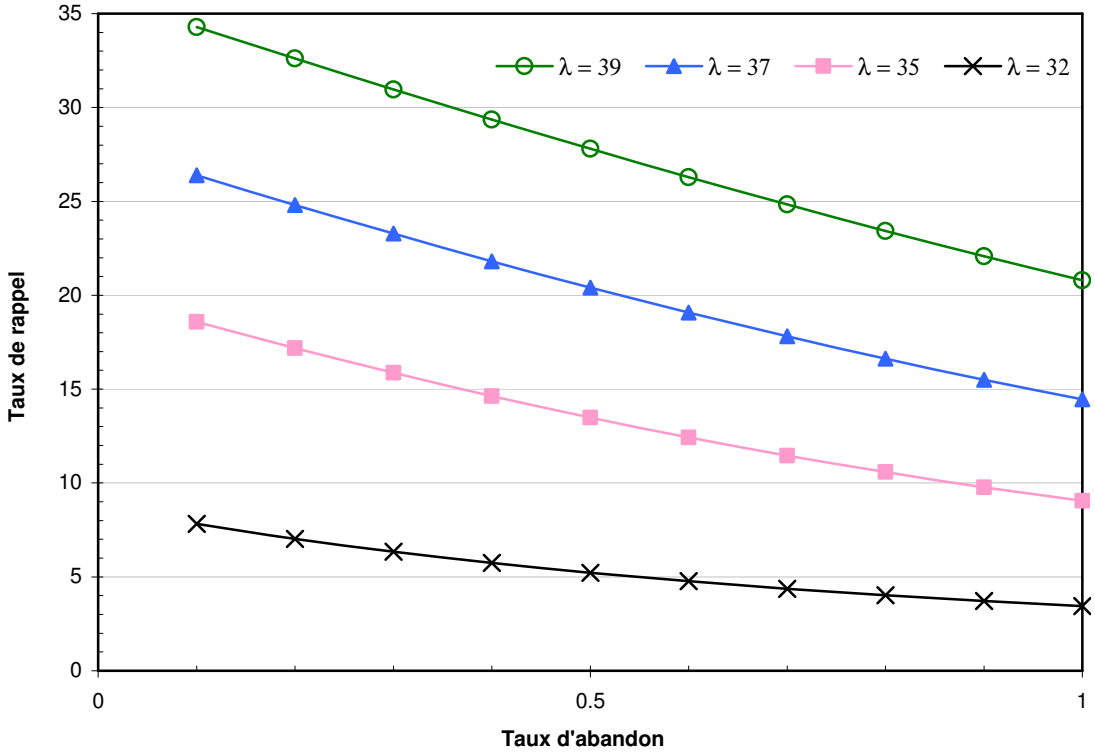


Figure 2.7: Effet de θ sur le taux de rappel

La Figure 2.8 illustre l'évolution du taux de rappel due à un changement de la taille de l'espace d'attente choisie par les managers du centre d'appels. Dans le modèle que nous étudions, les rappels ne peuvent provenir qu'à partir des clients ayant été déconnectés. Or, la probabilité de trouver tous les serveurs occupés (au moment de l'arrivée d'un client) devient plus faible lorsque l'espace d'attente grandit. En particulier, lorsque la file d'attente est illimitée, il n'y a pas le moindre rappel puisqu'il n'y a pas de déconnectés. Ceci explique l'allure décroissante du taux de rappel en fonction de la taille de la file d'attente $K_1 - C$. Pour la figure mentionnée, nous avons utilisé comme paramètres (à part les paramètres communs du début de la section) un nombre de serveurs $C = 100$ et une probabilité de rappel $p = 0{,}8$.

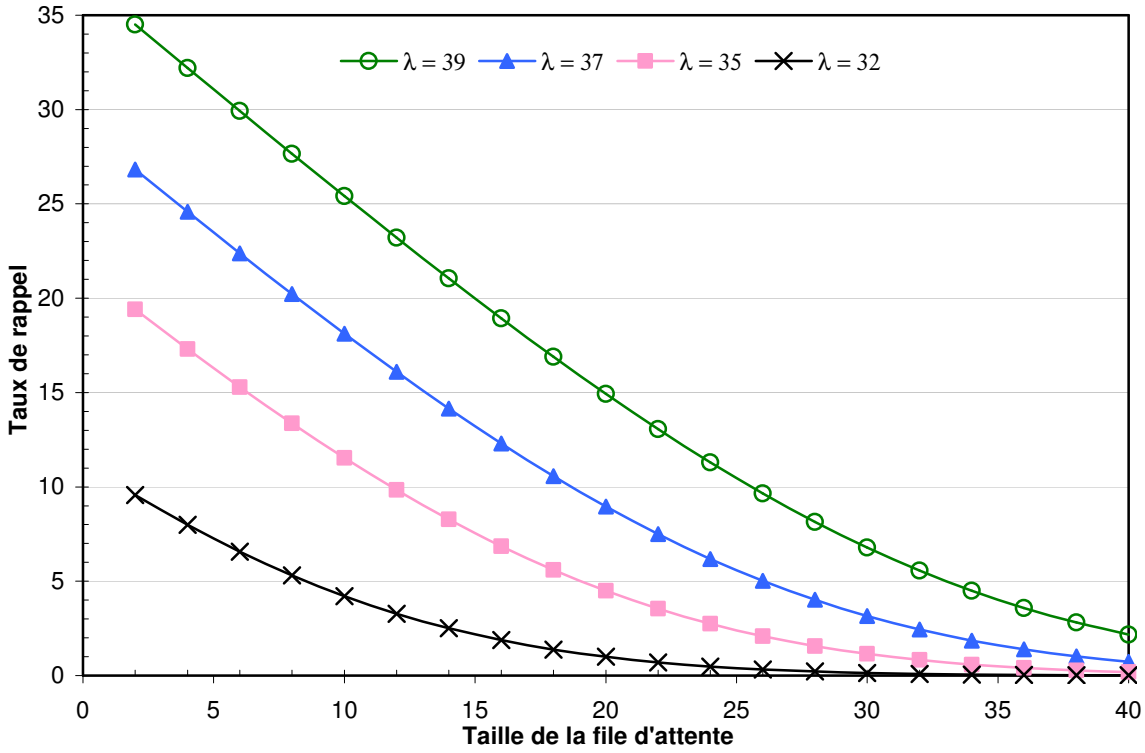


Figure 2.8: Effet de la limitation de l'espace d'attente sur le taux de rappel

Regardons maintenant l'effet du nombre de serveurs. En fait, nous ne pouvons pas faire varier le nombre de serveurs en gardant tous les autres paramètres inchangés. Si nous faisions ceci alors une variation de C impliquerait une variation de la charge ρ ainsi qu'un changement sur l'importance relative de la file d'attente $K - C$ par rapport au nombre de serveurs C et, dans ce cas, nous allons avoir une combinaison d'effets que nous ne pourrions interpréter. Ainsi, pour isoler l'effet du nombre de serveurs C (ou plutôt de la taille du système), nous allons garder la charge $\rho = \lambda / (C\,\mu)$ constante ce qui revient à garder le taux d'arrivée λ proportionnel à C ($\lambda = \rho\, C\, \mu$). Nous fixerons également le rapport $(K - C) / C$ à une valeur constante. Dans la Figure 2.9, ce rapport a une valeur de 0,05 ce qui veut dire que l'espace d'attente est limité à 5 places lorsqu'il y a 100 serveurs dans le système. La charge ρ est fixée à 0,9 et la valeur de la probabilité de rappel p utilisée est égale à 0,8. La Figure 2.9 affiche l'évolution du taux de rappel comme étant un pourcentage du taux d'appel frais et ce, en fonction du nombre de serveurs C. Nous avons préféré montrer l'évolution du rapport taux de rappel / taux d'appel frais parce que cela est plus approprié pour une étude de l'effet de la taille du

système où le mot "taille" implique le nombre de serveurs, le taux d'appel frais et la file d'attente. Dans cette figure, nous remarquons que le rapport étudié décroît en fonction de *C*. Nous savons que plus le système a une taille importante, plus l'influence des effets stochastiques sur son comportement diminue. Avec des tailles très importantes, la diminution de variabilité due à la taille fait que le comportement du système s'approche d'un comportement déterministe continu. Et puisque la charge dans la Figure 2.9 est inférieure à 1 alors plus la taille du système est grande, plus le taux de rappel s'approche de 0 (et donc le rapport taux de rappel / taux d'appel frais s'approche de 0).

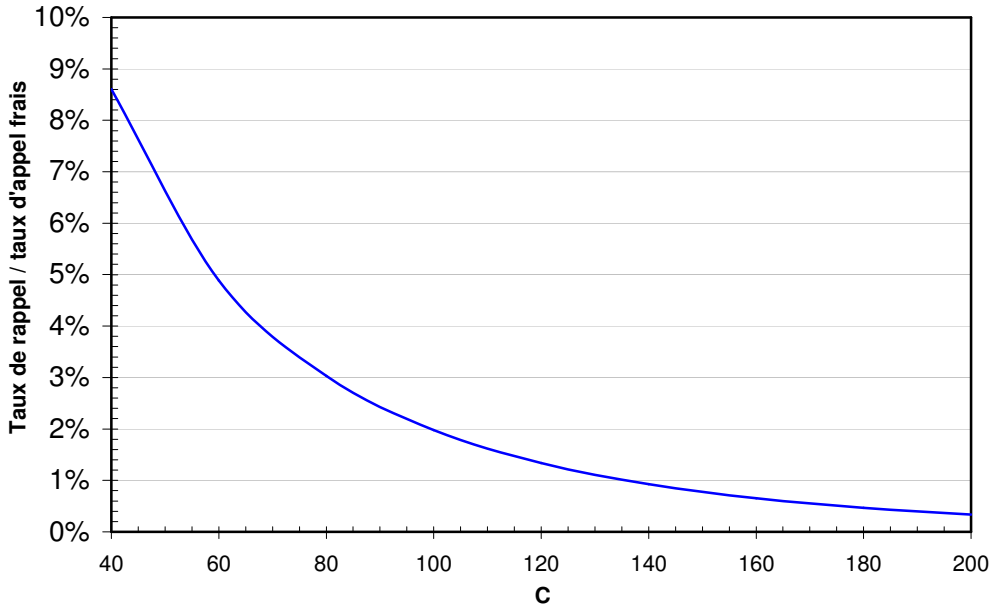


Figure 2.9: Effet de la taille du système pour ρ = 0,9 et $K - C = 0{,}05\ C$

Dans la Figure 2.10 nous avons utilisé les mêmes paramètres que ceux qui ont servi pour la Figure 2.9 à part la charge du système qui passe de 0,9 à 1,3. Comme nous avons expliqué précédemment, une augmentation de la taille du système réduit sa variabilité. Or, précisément, la variabilité est responsable d'une partie des rappels perçus et c'est ce qui implique une diminution de l'importance des rappels (par rapport aux appels frais) en fonction de la taille du système même pour une charge supérieure à 1.

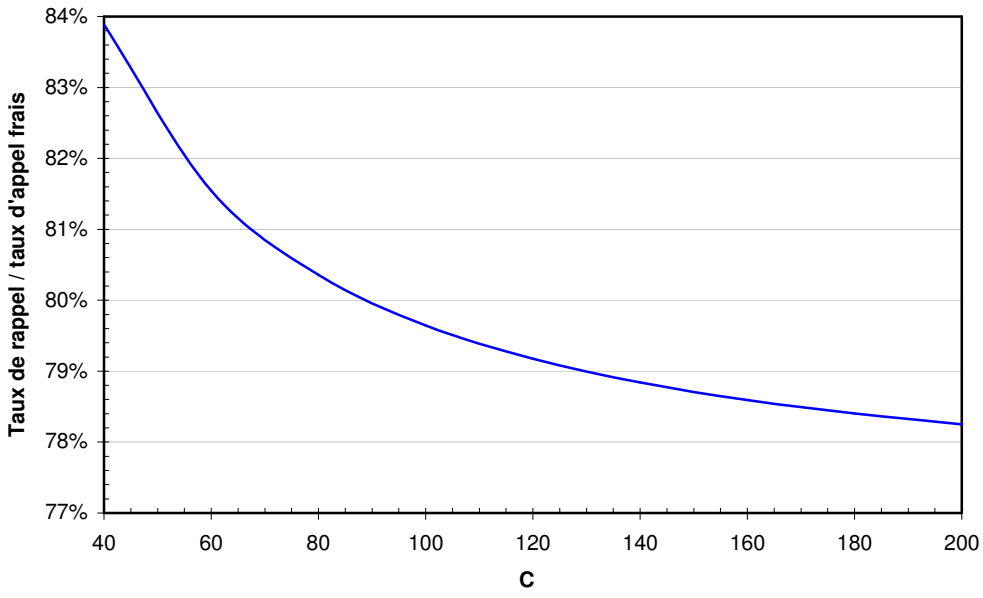


Figure 2.10: Effet de la taille du système pour ρ = 1,3 et $K - C = 0{,}05\ C$

2.6 Dimensionnement d'un centre d'appels

Dans cette section, nous allons traiter l'exploitation de l'information supplémentaire sur le taux de rappel dans le but de bien dimensionner le centre d'appels. Pour cela, nous fixons un critère de qualité de service à satisfaire. Ceci va être le taux de prise en charge des appels, c'est à dire la proportion des appels qui ont pu rejoindre un conseiller. Cette proportion correspond aux appels non déconnectés et non abandonnés. Une comparaison des résultats obtenus avec un système de file d'attente M/M/C/K + M sans rappel est effectuée afin de voir l'intérêt de considérer les rappels dans le dimensionnement. Les données nécessaires pour effectuer la comparaison sont prises dans la base de données de l'historique sur laquelle se base, habituellement, les prévisions. Nous supposons donc que le centre d'appels dispose du taux d'arrivée enregistré sur une journée comparable de point de vue charge de travail ainsi que du nombre de conseillers présents ce jour-là et des autres paramètres déjà rencontrés.

Sans considération des rappels, nous pouvons écrire le taux de prise en charge sous la forme suivante :

$$\text{Prise en charge } (\text{sans rappels}) = 1 \; - \; \frac{(\lambda \, \Pi_K + \; \theta \, Q_W)}{\lambda} \qquad (2.14)$$

Π_K désignant la probabilité stationnaire de trouver K clients dans le système et Q_W étant la taille moyenne de la file d'attente.

Pour ce système sans rappels, le calcul numérique des probabilités stationnaires n'est pas très compliqué puisqu'il s'agit d'un processus de naissance et de mort. Cela a été analysé par Garnett, Mandelbaum et Reiman [29]. Suite à ce calcul, et en fonction du nombre de conseillers que nous fixons, nous pouvons déterminer le taux de prise en charge des appels à l'aide de la formule (2.14) pour laquelle Π_K et Q_W sont issus de la résolution de la file M/M/C/K + M. Une procédure récursive inverse permet de déduire le nombre de conseillers nécessaires pour atteindre l'objectif de qualité de service. Ceci va être comparé avec ce que l'on obtient après la considération des rappels et ce pour quatre différentes qualités de service.

En considérant l'existence des rappels, nous savons que le taux d'appel enregistré dans l'historique est relatif aux appels observés qui se composent d'appels frais et de renouvellements d'appels. Pour dimensionner le système correctement, nous devons commencer par la détermination du taux d'appel frais étant donné le taux d'appel observé ainsi que le nombre de conseillers présents le jour de l'extraction des données. Les autres paramètres sont supposés connus par des analyses antérieures de l'historique. Nous calculons le taux d'appel frais à l'aide d'une procédure numérique récursive, telle la dichotomie, que nous avons appliquée à plusieurs exemples (dont le résultat est affiché par le Tableau 2.1). La différence entre ces exemples se situe au niveau du nombre de conseillers se trouvant le jour où les données ont été enregistrées. Si par exemple 25 conseillers travaillaient le jour de la récolte des données alors le taux d'appels frais serait de 9,79 pour un taux d'appel observé de 15.

C	λ
25	9,79
30	10,81
35	11,78
40	12,69
45	13,48
50	14,15
55	14,62

$K = C + 5$

$\lambda_{observé} = 15$

$\mu = 0,3$

$\theta = 0,3$

$\delta = 1$

$p = 0,8$

Tableau 2.1: Évolution de λ en fonction de C

Une fois le taux d'appel frais déterminé, la prise en charge des appels est déterminée par l'approximation suivante :

$$\text{Prise en charge (avec rappels)} = 1 - \frac{\left(\lambda_{obs} \sum_{n=0}^{K_2} \Pi_{K_1,n} + \theta\, Q_W \right)}{\lambda_{obs}} \qquad (2.15)$$

Dans cette formule, les probabilités $\Pi_{K1,n}$, $n = 0, 1, \ldots, K_2$ sont relatives à la chaîne de Markov de la Figure 2.2. Pour la calcul de ces probabilités, la connaissance du taux d'appel frais est primordiale et c'est pour cela que nous commençons d'abord par le déterminer à partir du taux d'appel observé λ_{obs}. Dans la même formule, Q_w représente la taille moyenne de la file d'attente relative au modèle que nous étudions. Ce paramètre se calcule à partir des probabilités stationnaires à l'aide de la formule suivante:

$$Q_w = \sum_{i=C+1}^{K_1} \left((i - C) \sum_{n=0}^{K_2} \Pi_{i,n} \right)$$

Nous allons maintenant comparer le dimensionnement du centre d'appels à l'aide d'une inversion numérique de la formule (2.15) (et en fonction d'un taux de prise en charge objectif) avec une méthode de dimensionnement connue sous le nom de "Square Root Staffing". Cette méthode a été désignée au départ pour un système multi-serveur M/M/c. Elle a été introduite par Whitt [69] et elle sert à déterminer un

dimensionnement nécessaire en fonction de la charge offerte λ / μ et d'un niveau de service β objectif. Étant donnée la difficulté lors de l'expression de β, cette méthode est spécialement utile pour maintenir un même niveau de service lorsque le taux d'arrivée λ change. Cette méthode constitue également une excellente approximation du nombre optimal de serveurs nécessaires à la satisfaction d'un objectif de service particulier (probabilité d'attendre, probabilité d'abandon, …) grâce à des formules déterminant β (voir Halfin et Whitt [34]). Garnett, Mandelbaum et Reiman [29] ont étendu la méthode du "Square Root Rule" pour dimensionner un système avec abandons modélisé avec la file M/M/C + M. Nous limitons notre intérêt ici à la formule utilisée par Garnett, Mandelbaum et Reiman [29] et qui est à la base de la dite méthode. En effet, cette méthode commence avec une quantification du niveau de service objectif par un nombre réel β que nous n'allons pas chercher à exprimer directement en fonction des paramètres du système (pour atteindre une qualité de service quelconque). Le nombre optimal de serveurs C^* (nécessaires à atteindre la qualité de service objectif) est lié avec β suivant l'équation :

$$\beta = \left(C^* - \frac{\lambda}{\mu} \right) \sqrt{\frac{\mu}{\lambda}} \tag{2.16}$$

Dans l'approximation mentionnée ci-dessus, nous pouvons remarquer que, si le terme du côté droit (exprimé en fonction de C, λ et μ) ne varie pas, alors les performances du système restent quasiment constants. Une fois le niveau de service β déterminé, si le taux d'arrivée change, pour garder le même niveau de service il suffit que C soit l'arrondi (à une valeur entière) de l'expression :

$$C = \frac{\lambda}{\mu} + \beta \sqrt{\frac{\lambda}{\mu}} \tag{2.17}$$

Étudions maintenant un exemple où l'on peut comparer la méthode du "Square Root Staffing" avec la nôtre (qui est basée sur la formule (2.15) et qui tient compte des rappels). Nous devons d'abord fixer les paramètres: $\mu = 0{,}3$, $\theta = 0{,}3$, $\delta = 1$ et $p = 0{,}8$. La taille de la file d'attente est limitée à 5 places. Supposons que le taux d'appel frais est $\lambda = 30$ (appels / minute) lorsque le taux de prise en charge enregistré est égal à 80 %. En inversant numériquement la formule (2.15), nous trouvons que, pour le taux d'appels mentionné, il faut un nombre de serveurs $C^* = 95$ afin d'atteindre les 80 % de prise en

charge. Maintenant, grâce à l'équation (2.16), nous pouvons quantifier le niveau de service à $\beta = -0,5$. Dans le Tableau 2.2, nous avons illustré les résultats obtenus pour le dimensionnement du centre d'appels par la méthode "square root" et par la méthode que nous avons utilisée précédemment et ce, en fonction de λ. La ligne *C* (square root) est calculée à partir de l'équation (2.17). Pour obtenir cette ligne du tableau, nous supposons que le manager du centre d'appels connaît l'évolution du taux d'appel frais λ et que, pour chaque valeur de λ, il détermine avec l'approximation "Square Root Staffing" le nombre de serveurs nécessaires pour le maintien du même niveau de service obtenu avec 95 serveurs (les 80 % de prise en charge). Nous remarquons dans le Tableau 2.2 que les deux méthodes de calcul donnent des résultats assez proches. Cependant, l'erreur croît avec la taille du centre d'appels (elle passe de 0,8 % pour $\lambda = 40$ à 2,2 % pour $\lambda = 100$). La méthode "square root" sur-dimensionne toujours le système puisqu'elle ne tient pas compte des rappels qui vont avoir lieu.

λ	40	50	60	70	80	90	100
C (square root)	128	160	193	226	259	291	324
*C** (avec rappels)	127	159	191	222	254	286	317

Tableau 2.2: Comparaison avec le dimensionnement par la méthode "square root"

Dans l'exemple précédent, les erreurs de la méthode "square root" ne sont pas importants. Toutefois, il faut signaler que la charge du système ne peut pas être très importante pour aboutir à un taux de prise en charge de 80 %. Et comme nous pouvons le voir dans les figures Figure 2.3 et Figure 2.4, le taux de rappel n'est pas important lorsque la charge du système est assez faible. Or, si le taux de rappel est faible alors le système se comporte comme une file M/M/C/K + M qui, elle même, s'approche d'une file M/M/C + M lorsque la charge est faible. Ainsi, pour mieux regarder les différences entre les deux méthodes de dimensionnement, il est plus intéressant de regarder un système où il y a plus de rappels. Nous allons donc diminuer le taux de prise en charge objectif dans un deuxième exemple qui garde les mêmes paramètres que le premier à part le taux de prise en charge objectif qui passe à 60 % (au lieu de 80 %) pour $\lambda = 30$. Nous calculons (de la même façon que précédemment) le niveau de service $\beta = -1,5$ (à partir d'un nombre optimal de serveurs $C^* = 85$. Le Tableau 2.3 montre les résultats obtenus par une étude semblable à ce qui a été fait pour l'exemple précédent. L'erreur

varie entre 0,9 % (pour $\lambda = 40$) et 5,2 % (pour $\lambda = 100$) ce qui est, comme attendu, supérieur à ce qui a été trouvé pour l'exemple précédent.

λ	40	50	60	70	80	90	100
C (square root)	116	147	179	210	242	274	306
*C** (avec rappels)	115	144	173	203	232	261	291

Tableau 2.3: Comparaison avec le dimensionnement par la méthode "square root"

D'après les deux derniers exemples, nous pouvons constater que plus la charge du système est importante, plus le dimensionnement sans tenir compte des rappels devient imprécis. Par contre, lorsque la charge est faible, le comportement du modèle s'approche beaucoup d'un système où il n'y a pas de rappels. Signalons, toutefois, que, pour les deux tableaux précédents, l'approximation du nombre optimal de serveurs tient compte du taux d'appel frais. Or, si l'utilisateur de cette méthode ignore les rappels dans son modèle, il n'est pas évident pour lui d'utiliser λ pour effectuer l'approximation. Pour cet utilisateur, il est plus logique de réagir à une évolution du taux d'appel observé λ_o et non à une variation de λ. Pour comparer les deux modèles sous ces nouvelles conditions, nous reprenons l'exemple précédent pour un taux de prise en charge objectif de 80 %. Pour $\lambda = 30$, le nombre de serveurs nécessaires à la satisfaction de l'objectif est $C^* = 95$. Ceci donne, à l'aide d'une inversion de la formule (2.13), $\lambda_o = 35{,}93$. En remplaçant λ par λ_o dans la formule (2.16), nous obtenons un niveau de service de: $\beta = -2{,}26$. Maintenant, en remplaçant λ par λ_o dans la formule (2.17), nous pouvons estimer le nombre optimal de serveurs par l'approximation "square root". Le résultat de la comparaison des deux méthodes est donné par le Tableau 2.4.

λ_o	40	50	60	70	80	90	100
C (square root)	107	137	168	199	230	261	292
*C** (avec rappels)	99	107	114	120	127	133	140

Tableau 2.4: Comparaison avec le dimensionnement par la méthode "square root"

Le Tableau 2.4 nous montre que le fait de ne pas tenir compte des rappels et de n'exploiter que le taux d'appel observé amène la méthode "square root" à des erreurs qui peuvent être très importantes lors du dimensionnement. Ces erreurs augmentent, comme précédemment, en fonction de la taille du système. Et, compte tenu de la non modélisation des rappels, l'approximation prévoit toujours plus de conseillers qu'il ne le faut.

Nous allons à présent voir l'effet de la non considération du phénomène de rappel sur le dimensionnement du système avec un calcul exact. Cet effet est illustré par la Figure 2.11. Nous y avons tracé l'évolution du nombre optimal de conseillers en fonction de la qualité de service objectif et ce, pour trois systèmes différents. Pour chacun des trois systèmes, λ_{obs} est égal à 15. Le premier système correspond au système qui n'intègre pas la notion de rappel parce qu'il ne connaît pas ou encore parce qu'il suppose que cela n'existe pas. Pour ce système, le taux d'appel frais coïncide, donc, avec le taux d'appel observé. Le nombre de conseillers nécessaires est alors directement dépendant du taux d'arrivée d'appels observés prévus. À partir d'une procédure de calcul inverse, et en utilisant la formule (2.14), nous pouvons déterminer le nombre optimal de serveurs qui permet d'atteindre la qualité de service objectif. Le deuxième système représente ce que nous obtenons en tenant compte des rappels et en partant de l'hypothèse d'un nombre de conseillers égal à 25 dans l'historique. Ainsi, pour ce système le calcul du taux d'appel frais en fonction du taux d'appel observé est faisable. À partir de la valeur de $\lambda_{obs} = 15$ et $C = 25$, la valeur de λ que le système obtient est égale à 9,79 (voir Tableau 2.1). Ensuite, sur la base d'un algorithme d'inversion et à l'aide de la formule (2.15), les managers du centre d'appels peuvent déterminer le nombre optimal de serveurs. Le troisième système est identique au deuxième sauf en ce qui concerne le nombre de conseillers qui passe de 25 à 55.

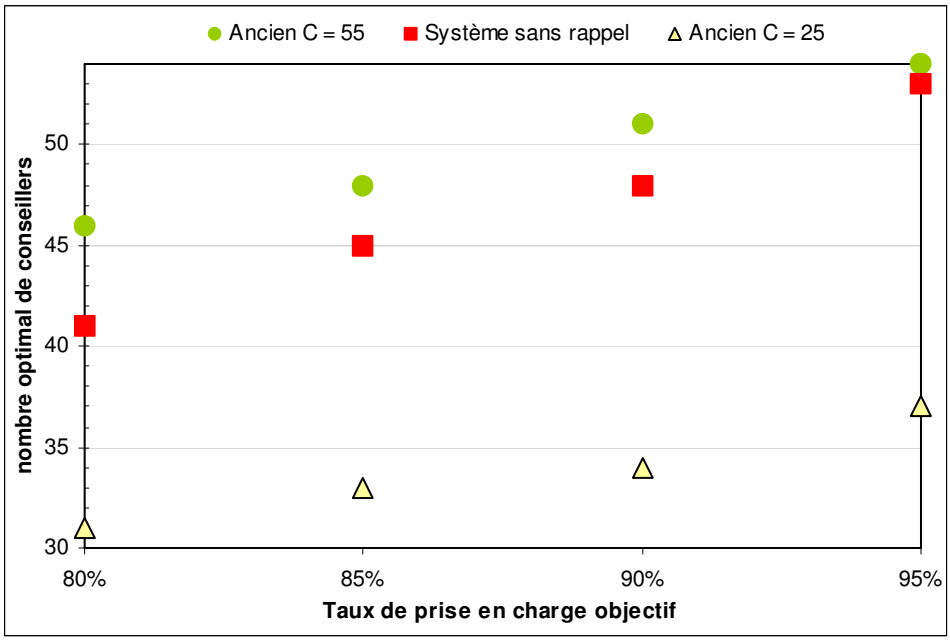


Figure 2.11: Effet de la considération des rappels lors du dimensionnement

Dans la Figure 2.11, nous remarquons que si nous ne considérons pas les rappels, nous risquons de sous-dimensionner le système (ne pas atteindre l'objectif) ou de le sur-dimensionner (atteindre l'objectif largement). En effet, nous observons dans cette figure une nette différence entre les systèmes étudiés. Si par exemple nous fixions comme objectif un taux de prise en charge supérieur à 90 %, alors nous pouvons passer de 48 conseillers (sans rappel) à seulement 34 (ou aussi à 51, cela dépend de ce que donne l'historique pour le nombre de serveurs). Cela implique parfois un surdimensionnement, et parfois un sous-dimensionnement du centre d'appels. Le surdimensionnement se passe lorsque le nombre de serveurs (présents le jour où les données ont été enregistrées) est faible ce qui engendre une différence importante entre λ et λ_{obs} et, de ce fait, en ignorant cette différence, le système qui ne tient pas compte des rappels va sur-dimensionner son staff afin de satisfaire λ_{obs} qui est nettement supérieur à la vraie demande λ. Lorsque le nombre de serveurs est important, λ et λ_{obs} vont être très proches (car $\lambda_{obs} / C\,\mu < 1$ lorsque les données ont été enregistrées) et dans ce cas, la vraie demande λ du modèle qui tient compte des rappels est à peu près égale à celle du modèle qui n'en tient pas compte (et qui est en réalité égale à λ_{obs}). Or, le modèle avec rappels va observer plus de demande que λ puisque, suivant le staff qu'il choisit, il va y avoir un certain nombre de rappels qui s'ajouteront aux appels frais.

Ceci a pour conséquence une demande globale plus importante que celle du modèle sans rappels. D'où le sous-dimensionnement. À noter que l'écart entre les deux modèles lorsqu'il y a sous-dimensionnement augmente lorsque la qualité de service objectif diminue. Nous expliquons ceci par le fait que, pour des objectifs modestes, le nombre de serveurs est faible. Les rappels sont donc assez nombreux et par conséquent, la demande globale est importante, ce qui est ignoré par le modèle sans rappels.

2.7 Conclusions

Dans ce chapitre, nous avons mis l'accent sur la nécessité de considérer le phénomène de rappels lors du dimensionnement d'un centre d'appels. Nous avons proposé pour analyser cette problématique un modèle stochastique de chaîne de Markov à deux dimensions qui fonctionne bien mais qui devient de plus en plus délicat à utiliser pour des systèmes de tailles importantes (à cause de l'augmentation du nombre d'états, ce qui nécessite un temps de calcul plus important). Il faut tout de même signaler que l'algorithme utilisé pour la résolution de la CMTC relative au modèle est assez rapide étant donnée sa complexité réduite en comparaison avec des méthodes classiques de résolution.

Partant de la chaîne de Markov obtenue, nous avons en premier lieu, illustré l'importance des rappels en montrant que le taux de rappel peut être de taille comparable au taux d'appel frais qui l'a engendré. Par la suite, nous avons étudié l'effet des paramètres du système sur l'évolution du taux stationnaire de rappel. Nous avons, finalement, montré les erreurs que la non considération des rappels peut induire lors du dimensionnement du centre d'appels. Et comme le montre bien la Figure 2.11, cela peut provoquer ou bien un surdimensionnement ou bien un sous dimensionnement par rapport au niveau de service cible du centre d'appels. Ces erreurs conduisent à des objectifs non atteints ou atteints largement ce qui signifie des coûts de fonctionnement du centre plus importants que nécessaire. Pour rattraper les erreurs observées, les responsables peuvent toujours faire augmenter ou diminuer le nombre de conseillers, ce qui contribue à l'ajout d'erreurs supplémentaires. Nous avons également regardé une approximation connue du dimensionnement d'une file d'attente multi-serveur avec abandons (mais sans rappels). Nous avons montré que les erreurs de cette

approximation peuvent être importants dans un système qui admet des rappels et que cette méthode conduit toujours à un surdimensionnement du système.

Le travail effectué dans ce chapitre exploite des paramètres constants dans le temps. En réalité, la plupart des paramètres sont variables : le taux d'arrivée varie au cours d'une même journée, les conseillers de clientèle sont programmés par tranches horaires, etc. Il est intéressant d'étudier le phénomène de rappel en multi-période et, dans ce cas, il faut utiliser des modèles qui tiennent en compte l'enchaînement des périodes de la journée ainsi que le régime transitoire du système qui devient alors non négligeable par rapport au régime stationnaire.

Chapitre 3 : Impact des rappels sur le dimensionnement dans un centre d'appels

3.1 Introduction

Comme nous l'avons mentionné dans le chapitre précédent, les coûts salariaux pour un centre d'appels représentent la majeure partie du coût total de son fonctionnement. La réduction de ces coûts est donc d'une importance stratégique au centre d'appels. Parallèlement, pour un opérateur de téléphonie mobile, être compétitif se traduit par la fourniture d'un bon niveau de service aux clients, notamment par le centre d'appels de l'entreprise. La qualité de service d'un centre d'appels se mesure de plusieurs manières mais vise globalement la meilleure satisfaction des besoins des clients.

Dans le chapitre précédent, nous avons analysé le centre d'appels au régime stationnaire. L'utilisation de ce modèle se justifie pour les systèmes qui atteignent rapidement l'état stationnaire ou encore pour ceux dont les paramètres ne varient pas de façon significative au cours du temps. Dans ce chapitre, nous allons analyser un modèle plus complet de centres d'appels dont le fonctionnement diffère (de ce qui a été introduit précédemment) surtout par l'ajout des renoncements, et la non limitation de l'espace d'attente. En outre, les paramètres peuvent varier dans le temps puisque nous allons également étudier le régime transitoire. Le centre d'appels que nous allons analyser dans ce chapitre obéit donc au fonctionnement qui suit. Les appels des clients

sont répondus par un certain nombre de serveurs. Lorsqu'un client essaye de joindre un conseiller, il est immédiatement servi s'il y a au moins un serveur de libre. Si, par contre, tous les serveurs sont occupés (avec d'autres clients), alors l'appel du client sera mis en attente jusqu'à ce qu'un conseiller puisse s'occuper de lui. La file d'attente fonctionne suivant la politique fifo ce qui veut dire qu'un serveur qui termine le service d'un client doit prendre l'appel ayant attendu le plus longtemps. Le centre d'appels peut choisir d'annoncer une estimation du temps d'attente aux clients à leur arrivée si ces derniers ne peuvent pas être servis immédiatement. Certains clients sont suffisamment patients pour attendre jusqu'à ce qu'un conseiller s'occupe d'eux, alors que d'autres vont raccrocher ou abandonner dès qu'ils savent qu'ils doivent attendre ou bien après un certain temps d'attente passé dans la file. Dans le but de donner un meilleur service aux clients, les managers veulent limiter le temps que passent les clients à attendre leur service. Dans beaucoup de centres d'appels, pour atteindre ce résultat, dès que le nombre de clients en attente dépasse un certain seuil, les nouveaux arrivants sont immédiatement déconnectés et sont priés de rappeler plus tard. Une partie de ces clients vont effectivement rappeler pour joindre un conseiller. Bien entendu, les clients ne veulent pas attendre, être déconnectés ou essayer d'accéder au service à plusieurs reprises. Tenant compte des exigences des clients, les managers essayent de déterminer le nombre de serveurs ainsi que le seuil de déconnexion (ou capacité de la file) qui minimisent les coûts tout en permettant d'atteindre certaines qualités de service objectif. L'utilisation de modèles de files d'attente est commode pour ce genre d'analyse dans les centres d'appels. Cependant, avant la moindre optimisation avec ce genre de modèles, des données relatives aux arrivées et à la caractérisation des paramètres du système sont nécessaires dans l'analyse des performances du système. Le contenu de ce chapitre fait partie du travail de Aguir *et al.* [4].

Les centres d'appels peuvent obtenir des informations détaillées relatives aux appels. Ainsi, le nombre total d'appels observés, déconnectés ou abandonnés peuvent être illustrés dans une base de données enregistrant des historiques sur plusieurs années et ce, suivant des intervalles d'une demi-heure. Ces données, en plus d'informations relatives aux durées des appels, des temps avant abandons et des temps de service peuvent être utilisées pour estimer les processus d'arrivée, d'abandon, et de service. Seul un type de données n'est pas disponible et c'est ce qui constitue la vraie

motivation de ce chapitre: en analysant la base de données représentant l'historique des appels, l'entreprise est incapable de connaître, au cours d'une certaine demi-heure, la proportion des appels primaires ou frais ainsi que la proportion constituée de clients qui rappellent et qui essayent d'obtenir un service en ayant abandonné ou en étant déconnectés auparavant. Ceci implique des erreurs lors de l'estimation des taux d'arrivée primaires. Le résultat qu'engendre l'utilisation de ces données dans un modèle de file d'attente qui tente d'optimiser le dimensionnement du système est donc, logiquement, erroné à cause des erreurs de dimensionnement qui existent dans l'historique. Si dans une période (de trente minutes) donnée le dimensionnement du centre d'appels a été effectué telle que la demande des clients dépasse largement la capacité de service, alors ceci génère des rappels lors des périodes suivantes. Toutefois, la non différenciation entre les appels de première intention et les rappels entraîne la méthode de prévision à traiter ce cas de figure comme une augmentation de la demande plutôt qu'un résultat du mauvais dimensionnement du système lors d'une période antérieure. Étant donné l'importance des coûts de dimensionnement, le centre d'appels cherche à déterminer comment extraire les appels de première intention à partir du nombre total d'appels observés en utilisant l'historique mis à disposition.

Nous modéliserons le système par une file M/M/C + M avec rappels et abandons, le + M désignant les abandons exponentiels. Nous allons supposer qu'une estimation du temps d'attente est communiquée aux clients ayant à attendre. Dans la section qui suit, nous allons formuler le modèle avec plus de détails.

Dans la section 3, nous considérons des systèmes où des paramètres comme le taux d'arrivée ne dépendent pas du temps. Nous allons analyser ce genre de systèmes par une étude en mono-période (avec une analyse au régime stationnaire). Dans cette analyse, nous utilisons à la fois une chaîne de Markov à temps continu et une approximation fluide. Cette dernière approche va nous aider à surmonter les difficultés numériques relatives à l'analyse de la chaîne de Markov. Nous allons montrer comment calculer la proportion des appels qui sont des appels frais (appels de première intention) à partir du taux d'appels observés et ce, en utilisant l'approche fluide.

Par la suite, nous ne supposons plus que les paramètres sont indépendants du temps dans la section 4. Ainsi, nous allons considérer le système en multi-période et nous l'analyserons à l'aide de l'approximation fluide. Nous montrons que

l'approximation fonctionne bien en comparant les performances du systèmes obtenues par simulation avec celles données par le modèle fluide. L'interaction entre dimensionnement et rappels est abordée à l'aide d'exemples numériques basés en partie sur des données réelles.

La section 5 illustre nos conclusions sur l'étude menée au cours de ce chapitre.

3.2 Le modèle

3.2.1 Le système où aucune estimation du temps d'attente n'est annoncée aux clients

Considérons le modèle suivant d'un centre d'appels possédant C conseillers de clientèle (CDC). Les temps de service (incluant les temps de réponse et les temps de traitement après l'appel nécessaires à la fermeture des dossiers) sont supposés suivre une loi exponentielle de taux μ. Les appels frais (ou primaires) arrivent au système suivant un processus de Poisson de taux λ. Les clients qui peuvent accéder directement à un CDC au moment de leur arrivée sont immédiatement servis et quittent le système après leur service. Les clients qui ne peuvent pas joindre un CDC libre à leur arrivée peuvent renoncer immédiatement avec la probabilité β. Ces derniers tentent à nouveau d'accéder au système avec une probabilité p et dans ce cas, le rappel se passe au bout d'un temps suivant une loi exponentielle de taux δ. Les clients qui décident de joindre la file (avec la probabilité $1 - \beta$) abandonnent s'ils ne sont toujours pas servis après un temps distribué suivant une loi exponentielle de taux θ_1. Les clients qui abandonnent sont supposés avoir le même comportement quant aux rappels que les clients qui renoncent: ils rappellent avec la probabilité p et s'ils le font, alors le délai avant rappel suit une loi exponentielle de taux δ. Conformément au chapitre précédent et à ce qui se trouve dans la littérature spécialisée, nous désignons l'espace d'attente virtuel qui comporte les clients ayant déjà décidé à rappeler mais qui ne l'ont pas encore fait par orbite ou file fictive. Notons ici que suivant les hypothèses décrites précédemment, le temps qu'un client passe dans l'orbite suit une loi exponentielle de taux δ. Les paramètres énoncés ici sont résumés à la suite:

λ : taux d'arrivée des appels frais
C : nombre de CDCs

μ : taux de service

β : probabilité de renoncement au service par les clients qui trouvent tous les conseillers occupés

θ_1 : taux d'abandon ou d'impatience des clients qui accèdent à la file

p : probabilité de rappel d'un client ayant abandonné ou renoncé auparavant

δ : taux de rappel par client ayant décidé à rappeler

En réalité, tous ces paramètres peuvent changer au cours du temps. Nous abordons les fluctuations dans le taux d'arrivée et le nombre de CDCs dans la deuxième partie de ce chapitre. Les autres paramètres ne changent, généralement, pas de façon significative et ne sont pas traités ici.

Notons, finalement, que le flux total des arrivées au système est constitué de deux flux séparés: appels frais (de taux λ) et renouvellements d'appels (rappels). Nous désignons le taux total d'appels (également appelé taux d'appel "observé") par λ_o (avec $\lambda_o \geq \lambda$).

3.2.2 Le système où une estimation du temps d'attente est annoncée aux clients

Dans cette section, nous allons nous baser sur le modèle décrit dans la section précédente pour traiter le cas où les clients sont informés d'une estimation de leurs temps d'attente au moment de leur arrivée. Whitt [70] a effectué une étude complète du nouveau modèle avec toutes les hypothèses utilisées.

Dans le nouveau modèle, le système communique une estimation du temps d'attente aux clients dès leur arrivée. L'information annoncée aux clients peut être une estimation du temps moyen d'attente ou une autre mesure qui s'y apparente. Dans tous les cas, l'information communiquée se base sur le nombre réel de clients en attente dans la file devant le nouvel arrivant. Comme expliqué par Whitt [70], le renoncement des clients dépend, dans ce cas, du nombre de clients qui attendent dans la file. Soit $r(k)$ ($k \geq C$) la probabilité qu'un nouvel arrivant renonce à joindre le système lorsqu'il y a déjà $k - C$ clients en attente dans la file. Cette probabilité représente le fait que le temps d'attente annoncé aux clients soit supérieur à ce qu'ils peuvent envisager comme attente dans le système. Elle s'écrit sous la forme suivante :

$$r(k) = \beta + (1 - \beta) P(T < S_k) \quad (3.1)$$

Dans cette équation, T est la variable aléatoire décrivant le seuil de patience du client et S_k est celle qui représente le temps entre l'arrivée d'un client et le moment où il réussit à joindre un CDC. Étant donné que les temps de service suivent une loi exponentielle, la variable aléatoire S_k a une distribution d'Erlang avec $k - C + 1$ étages, chacun de taux C μ. Comme dans Whitt [70], il est utile d'approcher $P(T < S_k)$ dans l'équation (3.1) par $P(T < E[S_k])$. $E[S_k]$ étant l'espérance de S_k, elle est égale à :

$$E[S_k] = \frac{k - C + 1}{C \mu}$$

En supposant que le seuil de patience T est distribué exponentiellement avec le taux θ_1, nous obtenons:

$$r(k) = 1 - (1 - \beta) e^{-\theta_1 \frac{k - C + 1}{C \mu}} \quad (3.2)$$

L'analyse effectuée dans le paragraphe précédent suppose qu'un client qui rejoint la file d'attente n'abandonne pas plus tard. Nous pouvons supposer que les clients rejoignant la file peuvent toujours abandonner avec un taux d'abandon θ inférieur au taux d'abandon initial θ_1 (lorsque les clients ne sont pas informés de leurs attentes). L'analyse précédente reste, approximativement, valide si θ est suffisamment faible pour ne pas modifier S_k sensiblement.

3.3 Analyse du régime stationnaire

3.3.1 Le modèle stochastique

L'analyse du système au régime stationnaire permet de déduire son comportement en mono-période. En effet, il est inutile de connaître l'état initial du système ou son évolution avant d'atteindre le régime permanent. Ce genre d'étude correspond à un système où il n'y a quasiment pas de variation des paramètres. Cela peut également être le cas d'un système qui converge rapidement vers son état stationnaire.

Suivant les hypothèses déjà énoncées, une chaîne de Markov peut être utilisée afin de modéliser le centre d'appels avec les renouvellements d'appels. Comme nous l'avons fait au chapitre précédent, la chaîne de Markov que nous utilisons ici possède deux dimensions: la première représente la file réelle (les *C* serveurs et l'espace d'attente) et la deuxième est relative à l'orbite. Suivant cette description, considérons l'état du système (m,n), (m,n ≥ 0), où *m* désigne le nombre de clients dans le système réel et *n* représente le nombre de clients en orbite et qui renouvellent leurs appels de façon exponentielle avec le taux δ. Du moment où la file réelle possède un espace d'attente limité, à la fois *m* et *n* ne sont pas bornés. La stabilité de la chaîne de Markov est assurée par les abandons et rappels qui nous assurent un système stable même pour une charge ρ = λ/Cμ supérieure à 1.

Compte tenu de la dimension infinie des deux files, nous tronquons, respectivement, la file réelle et l'orbite, à des tailles finies K_1 (donc $m \le K_1$) et K_2 ($n \le K_2$). Ces deux limitations sont prises suffisamment élevées pour pouvoir approcher la chaîne de Markov infinie sans, toutefois, fausser l'évaluation des performances du système.

Soit $Q_{(m,n)(m',n')}$ le taux de transition de l'état (m,n) à l'état (m',n'), $m, m' = 0, 1, \ldots, K_1$, $n, n' = 0, 1, \ldots, K_2$. Les taux de transition décrivant la chaîne de Markov, suffisent à déterminer le régime stationnaire du système. Les taux de transition non nuls sont les suivants :

$$Q_{(m,n)\,(m+1,n)} = \begin{cases} \lambda \;\; pour \;\; 0 \le m < C, \;\; 0 \le n \le K_2 \\ \lambda(1 - r(m)) \;\; pour \;\; C \le m < K_1, \;\; 0 \le n \le K_2 \end{cases}$$

$$Q_{(m,n)\,(m-1,n)} = \begin{cases} m\,\mu \;\; pour \;\; 0 < m \le C, \;\; 0 \le n \le K_2 \\ C\,\mu + (m - C)(1 - p)\,\theta \;\; pour \;\; C < m \le K_1, \;\; 0 \le n \le K_2 \end{cases}$$

$$Q_{(m,n)(m+1,n-1)} = \begin{cases} n\,\delta \;\; pour \;\; 0 \le m < C, \;\; 0 < n \le K_2 \\ n\,\delta\,(1 - r(m)) \;\; pour \;\; C \le m < K_1, \;\; 0 < n \le K_2 \end{cases}$$

$$Q_{(m,n)(m-1,n+1)} = (m - C)\,p\,\theta \;\; pour \;\; C < m \le K_1, \;\; 0 \le n < K_2$$

$$Q_{(m,n)\,(m,n+1)} = p\,r(m)\,\lambda \;\; pour \;\; C \le m \le K_1, \;\; 0 \le n < K_2$$

$$Q_{(m,n)\,(m,n\text{-}1)} = n\,(1-p)\,r(m)\,\delta \quad pour \quad C \le m \le K_1,\ 0 < n \le K_2$$

La Figure 3.1 illustre les différentes transitions entre les états de la chaîne de Markov décrite ci-dessus.

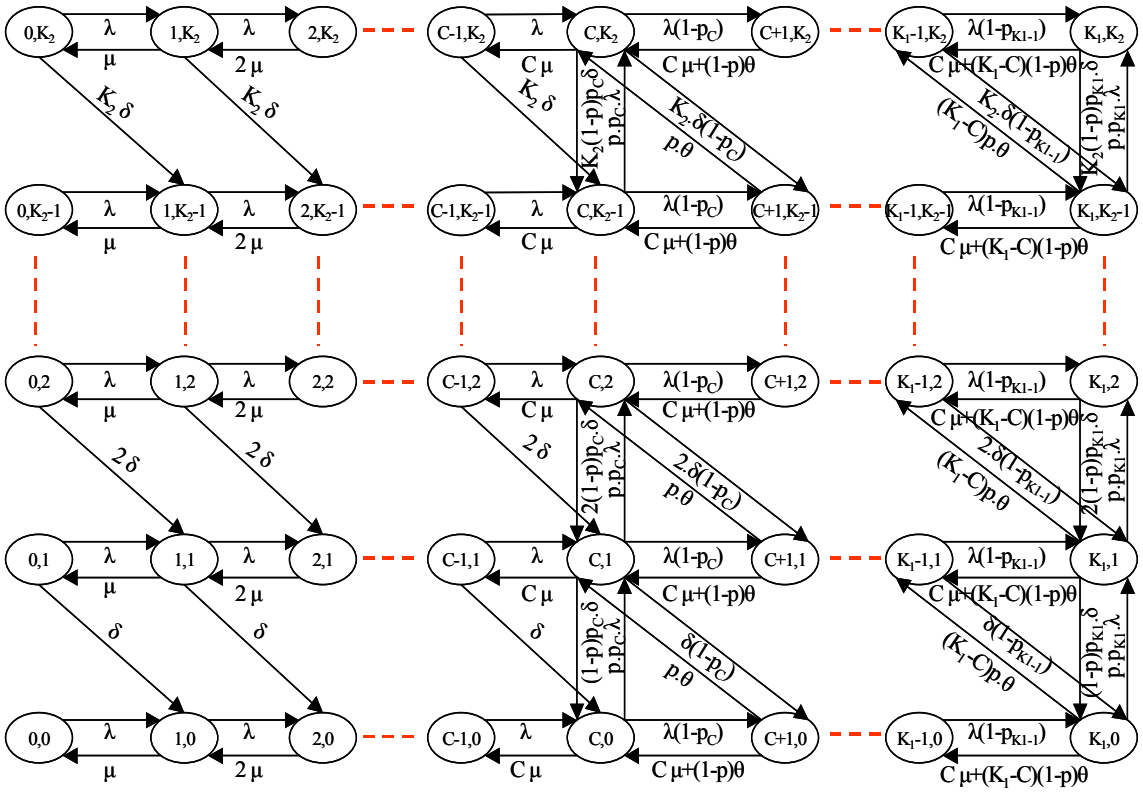


Figure 3.1: Diagramme de transitions de la chaîne de Markov

La complexité de la chaîne de Markov obtenue ci-dessus est telle que, même sa résolution numérique devient difficile. Nous ne pouvons, par exemple, pas utiliser la méthode de matrice géométrique à cause de la non existence de séquence répétitive dans les taux de transition. Le fait que les taux de transition dépendent de la probabilité de renoncement complique encore plus la tâche. Nous avons utilisé des méthodes numériques standards pour obtenir une solution numérique au régime stationnaire après troncation de l'espace d'états. Le choix des valeurs de K_1 et de K_2 doit nous permettre une évaluation précise des performances du système. Pour cela, nous avons fait augmenter progressivement les valeurs de K_1 et de K_2 jusqu'à ce que de nouvelles augmentations n'influent plus sur la distribution des états au régime

stationnaire. Soit $\Pi_{m,n}$, $m = 0, 1, \ldots, K_1$, $n = 0, 1, \ldots, K_2$, la probabilité au régime stationnaire de se trouver à l'état (m,n) (correspondant à m clients dans la file réelle et n clients en orbite). Posons par Tr le taux total de rappel au régime stationnaire calculé à partir de la chaîne de Markov (nombre de rappels effectué au régime stationnaire par unité de temps). Le taux de rappel au régime stationnaire est donné par la formule suivante :

$$Tr = \sum_{n=1}^{K_2} n \, \delta \sum_{m=0}^{K_1} \Pi_{m,n} \tag{3.3}$$

Ainsi, le taux total de rappel au régime stationnaire peut être obtenu en utilisant une solution numérique de la chaîne de Markov de la Figure 3.1 et la formule (3.3). Cependant, cette méthode est numériquement très exigeante (de point de vue ressources informatiques) et son temps de calcul croît rapidement à mesure que la taille du système augmente. Ci-dessous, nous proposons une expression alternative à Tr basée sur un équilibre stationnaire des flux (similaire à Hoffman et Harris [35]). Le résultat est résumé dans la proposition suivante:

Proposition 1 *Soit E[B] le nombre moyen de serveurs occupés. Le taux de rappel, Tr, du système peut être écrit sous la forme :*

$$Tr = \frac{p}{1 - p} \left(\lambda - E[B] \, \mu \right) \tag{3.4}$$

Preuve: Considérons les différents flux dans le système au régime stationnaire. Le flux total, λ_o, pénétrant le système comporte deux composantes: les appels frais et les rappels. Nous avons donc $\lambda_o = \lambda + Tr$. Le flux sortant du système comporte, lui, trois composantes. Notons par A, le flux sortant dû aux abandons et par R, le flux sortant dû aux renoncements. Le flux sortant dû aux services accomplis est donné par le taux de service effectif moyen $E[B]\,\mu$.

L'équilibre au régime stationnaire fait que le flux sortant soit égal au flux entrant ce qui donne:

$$\lambda_o = A + R + E[B] \, \mu \tag{3.5}$$

En plus de l'équilibre mentionné ci-dessus, il existe un équilibre de l'orbite qui fait que le flux moyen entrant à l'orbite et le flux moyen qui en sort soient égaux. Un appel qui rejoint l'orbite est dû soit à un abandon, soit à un renoncement. Le flux entrant à l'orbite est donc $p(A + R)$. Puisque le taux de départ de l'orbite n'est autre que Tr, nous obtenons alors :

$$Tr = p\left(A + R\right) \tag{3.6}$$

En utilisant l'équation (3.5), nous pouvons écrire :

$$\lambda_o = \frac{Tr}{p} + E\left[B\right]\mu \tag{3.7}$$

Puisque $\lambda_o = \lambda + Tr$, l'équation (3.7) nous permet d'avoir :

$$\lambda + Tr = \frac{Tr}{p} + E\left[B\right]\mu$$

d'où le résultat de la proposition.

La Figure 3.2 illustre les flux d'entrée et de sortie utilisés dans la démonstration précédente. Bien que la proposition 1 ne facilite pas le calcul exact du taux de rappel (car $E[B]$ doit toujours être calculé), elle suggère une méthode pour son estimation à partir de données réelles. Si p, λ, μ et $E[B]$ sont connus, la proposition 1 fournit un estimateur simple à utiliser pour Tr. Elle peut également être la base de techniques d'approximation comme celles dans Hoffman et Harris [35]. Nous discuterons d'une approximation basée sur cette proposition dans la prochaine section où une approche différente est utilisée.

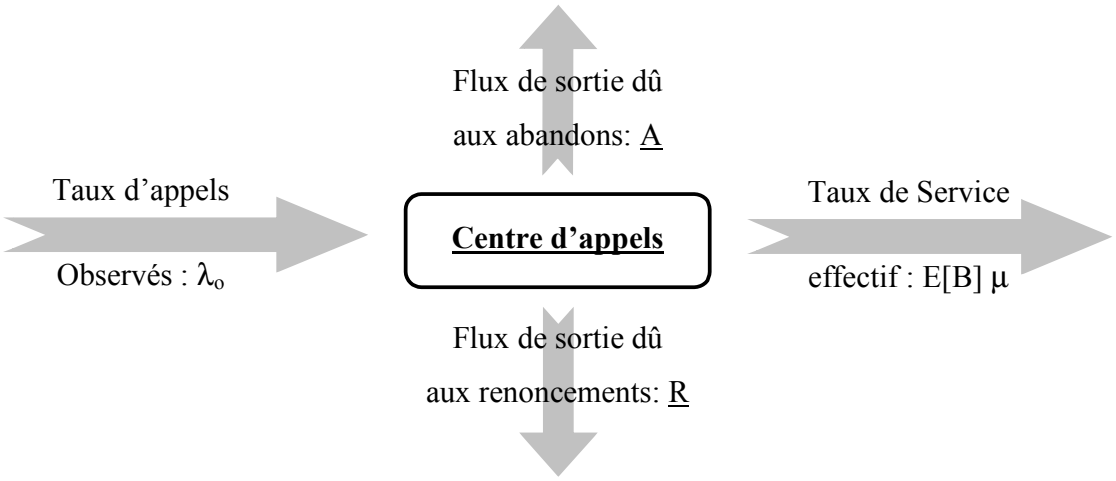


Figure 3.2: Les flux entrants et les flux sortants

3.3.2 L'approximation fluide

Bien que le modèle stochastique de la section précédente peut être utilisé pour calculer numériquement les mesures de performances relatives aux rappels, il n'est pas simple d'utilisation étant donné les ressources qu'il nécessite. Nous introduisons ici une simple approximation qui remplace les taux de transition de la chaîne de Markov par des flux déterministes. Le modèle résultant exploite des flux continus déterministes d'où l'origine de l'appellation de l'approximation par *l'approximation fluide*. Mandelbaum *et al.* [53] ont prouvé qu'une telle approximation est une limite formelle d'un modèle stochastique correspondant et ils ont montré à l'aide d'exemples numériques que l'approximation est précise.

Avec le modèle fluide, nous supposons que tous les paramètres nécessaires à la description du système sont continus dans le temps. Nous commençons par remplacer l'espace d'états discret du modèle stochastique original par un espace d'états continu. Dans le modèle approché, un état est représenté par $(x_1(t),x_2(t))$, où $x_1(t)$ représente la taille (continue) de la file réelle et $x_2(t)$ la taille (continue) de l'orbite à l'instant t. Le taux d'arrivée observé par le système se compose du taux de rappel instantané ainsi que du taux d'appels primaires. Le taux de rappels à l'instant t est proportionnel à la taille de l'orbite $x_2(t)$ puisque chaque client qui s'y trouve rappelle avec un taux δ. Ainsi, avec

l'approximation continue déterministe, nous pouvons écrire le taux d'appels observés comme suit :

$$\lambda_o(t) = \delta\, x_2(t) + \lambda(t) \tag{3.8}$$

Le flux qui s'ajoute à $x_1(t)$ dépend de $\lambda_o(t)$ et du taux de renoncement à l'instant t. En utilisant l'approximation déterministe, le taux avec lequel $x_1(t)$ augmente est donc: $(1 - r(x_1(t)))\ \lambda_o(t)$, où $r(x)$ représente la probabilité de renoncement lorsque la taille de la file réelle est égale à x. Afin de définir cette probabilité, nous nous basons sur ce que nous avions déjà défini à l'équation (3.2) et nous l'adaptons à un espace d'états continu comme suit :

$$r(x) = \begin{cases} 1 - (1 - \beta)\, e^{-\theta_1 \frac{x - C + 1}{C\mu}} & \text{si } x \geq C \\ 0 & \text{sinon} \end{cases} \tag{3.9}$$

La taille de la file réelle $x_1(t)$ diminue à travers les fins de service et les abandons. Avec l'approximation, le taux de diminution dans $x_1(t)$ dû aux services accomplis et aux abandons est égal à :

$$\mu\, Min\left(x_1(t), C\right) + \theta\, Max\left(x_1(t) - C, 0\right)$$

Nous pouvons maintenant exprimer le taux de variation de $x_1(t)$ de la façon suivante :

$$\frac{dx_1}{dt} = \left(1 - r\left(x_1(t)\right)\right) \lambda_o - \mu\, Min\left(x_1(t), C\right) - \theta\, Max\left(x_1(t) - C, 0\right) \tag{3.10}$$

Le remplissage de l'orbite se fait, lui, à travers les abandons et les renoncements et ce, en conditionnant sur l'intention de rappeler plus tard. Le taux de remplissage de l'orbite est donc le suivant :

$$p\, \left(r\left(x_1(t)\right) \lambda_o + \theta\, Max\left(x_1(t) - C, 0\right)\right)$$

L'orbite se vide à mesure que les clients finissent par rappeler. Elle se vide, donc, avec le taux: $\delta\, x_2(t)$.

Ceci nous permet de déduire le taux de variation du niveau $x_2(t)$ de l'orbite :

$$\frac{dx_2}{dt} = p\ \left(r\left(x_1(t)\right)\ \lambda_o\ +\ \theta\ Max\left(x_1(t)\ -\ C,0\right)\right)\ -\ \delta\ x_2(t) \tag{3.11}$$

En remplaçant λ_o par sa valeur de l'équation (3.8), nous pouvons obtenir ($x_1(t)$,$x_2(t)$) comme étant la solution du système différentiel suivant :

$$\begin{cases} \dfrac{dx_1}{dt} = \left(1\ -\ r\left(x_1(t)\right)\right)\ \left(\delta\ x_2(t)\ +\ \lambda\right)\ -\ \mu\ Min\left(x_1(t),C\right)\ -\ \theta\ Max\left(x_1(t)\ -\ C,0\right) \\ \dfrac{dx_2}{dt} = p\ r\left(x_1(t)\right)\ \left(\delta\ x_2(t)\ +\ \lambda\right)\ -\ \delta\ x_2(t)\ +\ \theta\ p\ Max\left(x_1(t)\ -\ C,0\right) \end{cases} \tag{3.12}$$

En particulier, au régime stationnaire nous avons :

$$\lim_{t\to\infty}\frac{dx_1(t)}{dt} = \lim_{t\to\infty}\frac{dx_2(t)}{dt} = 0 \tag{3.13}$$

ce qui nous permet d'obtenir les niveaux des files au régime stationnaire x_1 et x_2 comme suit :

$$\begin{cases} \lim_{t\to\infty} x_1(t) = x_1 \\ \lim_{t\to\infty} x_2(t) = x_2 \end{cases}$$

Il est facile de voir que, si la charge moyenne ρ (= λ / C μ) est inférieure ou égale à 1, alors le niveau stationnaire x_2 de l'orbite est nul. Dans ce cas, l'approximation ne peut fournir aucune information sur le taux de rappel du système. Concentrons nous maintenant sur le cas le plus intéressant où $\rho > 1$. Dans ce cas, il est facile de vérifier que $x_1 \geq C$. Ceci nous permet d'écrire: Min(x_1,C) = C et Max(x_1 - C,0) = x_1 - C. Au régime stationnaire, le système (3.12) devient alors :

$$\begin{cases} \left(1\ -\ r(x_1)\right)\ \left(\delta\ x_2\ +\ \lambda\right) = C\ \mu\ +\ \theta\ \left(x_1\ -\ C\right) \\ \delta\ x_2 = p\ r(x_1)\ \left(\delta\ x_2\ +\ \lambda\right)\ +\ \theta\ p\ \left(x_1\ -\ C\right) \end{cases}$$

$$\Leftrightarrow \begin{cases} p\ \left(\delta\ x_2\ +\ \lambda\right) = p\ C\ \mu\ +\ \theta\ p\ \left(x_1\ -\ C\right)\ +\ p\ r(x_1)\ \left(\delta\ x_2\ +\ \lambda\right) \\ \delta\ x_2 = p\ r(x_1)\ \left(\delta\ x_2\ +\ \lambda\right)\ +\ \theta\ p\ \left(x_1\ -\ C\right) \end{cases}$$

ce qui aboutit finalement à :

$$\begin{cases} p\ (\delta\ x_2\ +\ \lambda)\ -\ \delta\ x_2\ =\ p\ C\ \mu \\ \delta\ x_2\ =\ p\ r(x_1)\ (\delta\ x_2\ +\ \lambda)\ +\ \theta\ p\ (x_1\ -\ C) \end{cases} \tag{3.14}$$

Cette dernière équation nous permet d'obtenir l'état su système au régime stationnaire en déterminant le niveau stationnaire des files (réelle et fictive). La taille de la file réelle est la solution de :

$$(\lambda\ -\ p\ C\ \mu)\ r(x_1)\ +\ \theta\ (1\ -\ p)\ (x_1\ -\ C)\ =\ \lambda\ -\ C\ \mu \tag{3.15}$$

L'équation (3.15) peut être résolue en utilisant des méthodes numériques standards pour déterminer la valeur de x_1. D'un autre coté, il s'avère qu'il y a une solution plus simple pour x_2. Ce qui est intéressant c'est que x_2 ne dépend pas de θ, le taux représentant le seuil de patience des clients. Ceci représente une propriété utile puisque les autres paramètres (C, μ, λ, δ et p) sont relativement plus faciles à estimer en pratique. Une autre conséquence utile de la propriété précédente (insensibilité à θ) est que la même insensibilité est également vraie pour le taux stationnaire de rappel. De plus, cette propriété est également valable pour le modèle stochastique de la section précédente. Ce résultat est présenté par la proposition suivante :

Proposition 2 *i. Le taux de rappel, Tr (fluide), obtenu à travers le modèle fluide, ne dépend pas de θ.*
ii. L'approximation fluide du taux de rappel est asymptotiquement correcte pour le modèle stochastique quand λ croît.

Preuve: À partir du système d'équations (3.14), la taille stationnaire de l'orbite, x_2, est donnée par :

$$x_2\ =\ \frac{p}{1\ -\ p}\ \frac{\lambda\ -\ C\ \mu}{\delta} \tag{3.16}$$

Maintenant, en notant que Tr(fluide) = δ x_2, nous obtenons :

$$Tr(fluide)\ =\ \frac{p}{1\ -\ p}\ (\lambda\ -\ C\ \mu) \tag{3.17}$$

ce qui ne dépend pas de θ.

Afin de démontrer la partie ii. de la proposition 2, considérons l'expression de Tr donnée dans la proposition 1 (équation (3.4)) pour le modèle stochastique. En faisant tendre λ vers l'infini (pour un nombre de serveurs C fixe), nous obtenons $E[B] = C$ (tous les serveurs sont tout le temps occupés). En remplaçant $E[B]$ par C dans l'équation (3.4) nous obtenons l'expression de l'équation (3.17).

La proposition 2 suggère ceci: l'approximation du taux de rappel donnée par l'équation (3.17) doit être précise pour des centres d'appels surchargés. En fait, l'erreur de l'approximation est uniquement due au remplacement de $E[B]$ par C dans la formule correspondante pour le modèle stochastique. Pour les systèmes surchargés, $E[B]$ est raisonnablement proche de C. En notant $\lambda_o = Tr + \lambda$, nous pouvons transformer l'équation (3.17) de telle sorte à ce qu'elle nous donne le taux de rappel en fonction de p, C, μ et λ_o (et non λ) ce qui peut s'avérer très utile pour les managers de centres d'appels puisque, justement, c'est de λ_o qu'ils disposent et non de λ. L'équation résultante est la suivante :

$$Tr\ (fluide) = p\ (\lambda_o - C\ \mu) \tag{3.18}$$

Nous pouvons également, à partir de l'équation (3.17) obtenir λ en fonction des mêmes paramètres ce qui est également très pratique puisque cela représente la vraie demande que reçoit le centre d'appels :

$$\lambda = (1 - p)\ \lambda_o + p\ C\ \mu \tag{3.19}$$

Comme précisé auparavant, l'approximation ne fournit pas d'information utile lorsque $\lambda \leq C\ \mu$. Néanmoins, pour de gros centres d'appels, le phénomène de rappel est essentiellement dû à la surcharge. Dans la prochaine section, nous évaluons la précision de cette approximation suivant plusieurs conditions.

3.3.3 Évaluation numérique de l'approximation fluide

Dans le but d'évaluer la précision de l'approximation fluide au régime stationnaire, nous montrons, à partir d'exemples numériques, que l'erreur entre le taux de rappel δx_2 (donné par l'équation (3.17)) et le taux de rappel stationnaire exact (donné par

l'équation (3.3)) diminue rapidement en fonction du nombre de serveurs et de la charge du système. Autrement dit, le taux de rappel (exact) obtenu à l'aide du modèle stochastique va être comparé avec le taux de rappel (approché) fourni par le modèle fluide.

Considérons d'abord le cas où la charge du système est fixée à 133 % et le nombre *C* de serveurs est variable. Puisque $\rho = \lambda/C\mu$, λ doit également varier proportionnellement à *C* afin d'assurer une charge constante du système. Les résultats de la comparaison des deux modèles (stochastique et fluide) au régime stationnaire est donné par le Tableau 3.1. Dans ce tableau et dans le reste du chapitre, l'unité de temps considérée est la minute. Ainsi, $\mu = 0{,}3$ implique un taux de service égal à 0,3 appels par minute. En utilisant les résultats du Tableau 3.1, nous pouvons calculer l'erreur due à l'approximation fluide comme étant un pourcentage relatif au taux de rappel exact *Tr* (c'est à dire: pourcentage d'erreur = (|*Tr*(fluide)-*Tr*(stochastique)|/ *Tr*(stochastique))*100). Cette erreur est illustrée par la Figure 3.3 en fonction du nombre de serveurs.

C	*Tr*	δx_2
5	0,69	0,50
10	1,20	1,00
15	1,70	1,50
20	2,19	2,00
25	2,68	2,50
30	3,17	3,00
35	3,66	3,50
40	4,15	4,00
45	4,64	4,50
50	5,13	5,00

Tableau 3.1: Comparaison de l'approximation du taux de rappel stationnaire avec la valeur exacte pour $\rho = 1{,}33$

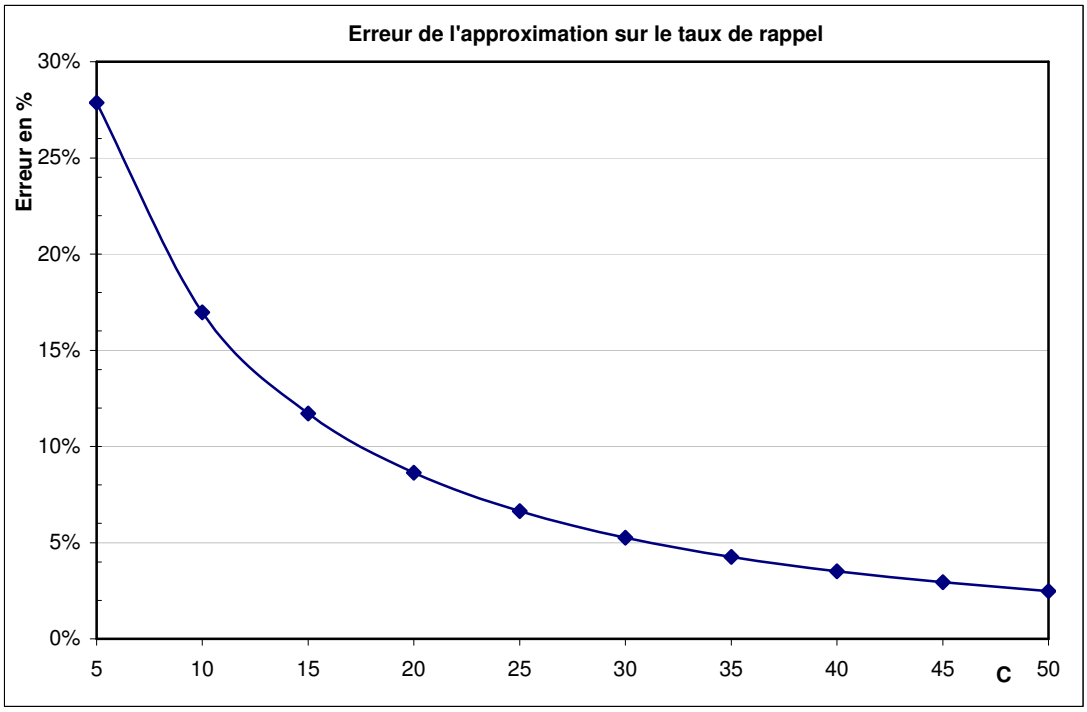


Figure 3.3: Évolution de l'erreur de l'approximation par le modèle fluide du taux de rappel en fonction du nombre de serveurs C pour $\rho = 133$ %

Nous constatons que pour une charge fixée à 133 %, la précision de l'approximation fluide croît en fonction du nombre de serveurs. Pour cinquante serveurs, l'erreur est déjà inférieure à 2,5 %. Pour un faible nombre de serveurs, l'erreur peut être supérieure à 25 %. Compte tenu de ces résultats, nous pouvons dire que pour un centre d'appels de grande taille, l'erreur est négligeable et le modèle fluide devient alors incontournable vue la simplicité de la formule analytique donnant le taux de rappel.

Observons maintenant un exemple où le nombre de serveurs C est fixé à 40, et faisons varier la charge du système en faisant augmenter le taux d'appels primaires λ. Le Tableau 3.2 compare les taux de rappel stationnaires obtenus par les modèles fluide et stochastique. Comme précédemment, la Figure 3.4 montre l'évolution de l'erreur en fonction de la charge du système. Cette figure confirme le fait que la précision de l'approximation fluide croît rapidement en fonction de la charge du système. Ceci n'est pas surprenant compte tenu de la proposition 2 ii qui précise que l'approximation fluide fournit un résultat asymptotiquement exact. Signalons ici que le nombre de serveurs dans cet exemple représente un petit centre d'appels en réalité. Une

observation supplémentaire concerne le cas où la charge ρ est égale à 100 %. Pour ce cas, l'erreur trouvée de l'approximation fluide est égale à 100 %. Ceci n'est pas surprenant étant donné que l'approximation fluide ignore complètement les effets stochastiques dans le système ce qui la conduit à ne pouvoir prédire aucun rappel. Ce qui relativise l'importance de cette erreur est le fait que, pour ce cas particulier, le taux de rappel exact (stochastique) du système est très faible (0,98 par minute) comme observé dans le Tableau 3.2.

ρ	*Tr*	δx_2
100%	0,98	0,00
110%	1,76	1,20
120%	2,72	2,40
130%	3,78	3,60
140%	4,90	4,80
150%	6,05	6,00
160%	7,23	7,20
170%	8,42	8,40
180%	9,61	9,60
190%	10,80	10,80
200%	12,00	12,00

Tableau 3.2: Comparaison de l'approximation du taux de rappel stationnaire avec la valeur exacte pour $C = 40$

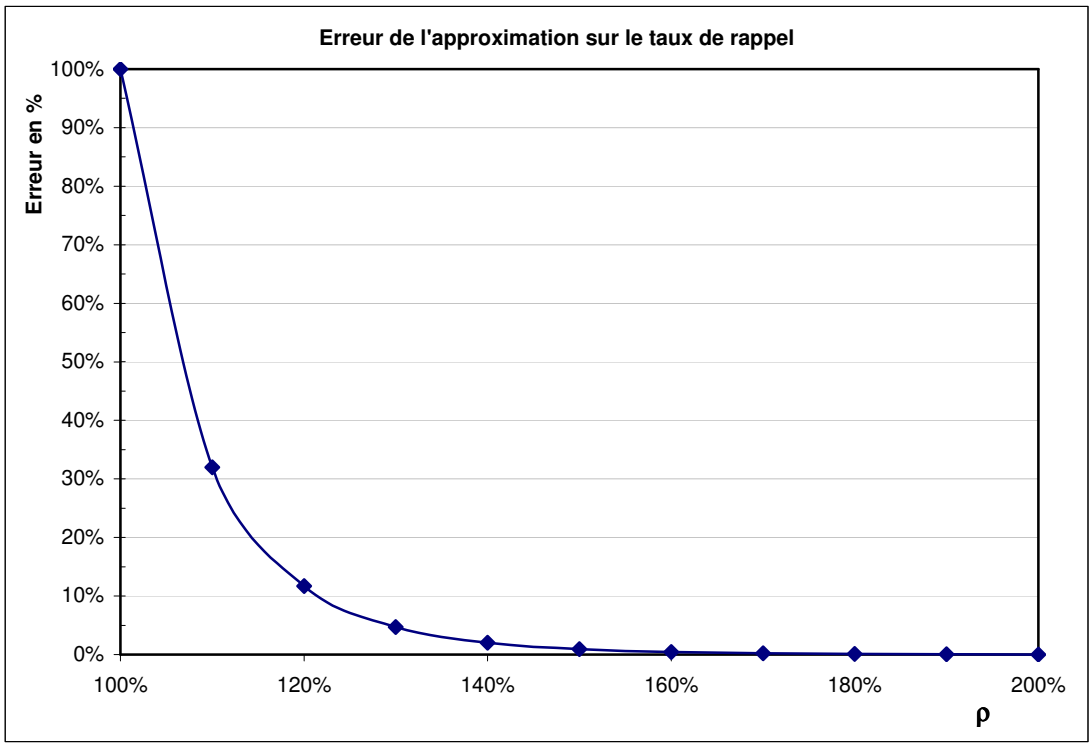


Figure 3.4: Évolution de l'erreur de l'approximation par le modèle fluide du taux de rappel en fonction de la charge ρ pour $C = 40$

Les deux exemples présentés montrent que la précision de l'approximation fluide croît rapidement avec la charge du système et avec le nombre de serveurs. Dans un cas où les deux effets sont combinés, une convergence vers de hauts niveaux de précision a lieu de façon beaucoup plus significative. Ainsi, pour un centre d'appels de grande ou de moyenne taille, et avec une charge réaliste pour de nombreux systèmes, l'approximation fluide du taux de rappel constitue un excellent moyen pour évaluer ce qui se passe réellement et ce, avec une grande précision.

La proposition 2 (équation (3.17)) décrit une autre caractéristique intéressante de l'approximation fluide. En effet, nous pouvons y constater que l'approximation du taux de rappel ne dépend pas de la fonction de renoncement $r(k)$. Cette fonction peut donc prendre une autre expression que celle définie par l'équation (3.2). En utilisant cette propriété, nous allons analyser la performance de l'approximation pour un autre système pour lequel la fonction de renoncement $r_1(k)$ s'écrit sous la forme :

$$r_1(k) = \begin{cases} 0 & \text{si } k < K_1 \\ 1 & \text{sinon} \end{cases},$$

avec K_1 un entier strictement supérieur à C. En pratique, beaucoup de centres d'appels décident de limiter la taille de leur file d'attente à une longueur finie pour que les clients n'attendent pas longtemps dans la file, préférant ainsi déconnecter les clients plutôt que de les voir abandonner. La probabilité de renoncement "artificielle" $r_1(k)$ nous permet de modéliser indirectement un centre d'appels M/M/C/K_1 + M qui intègre l'impatience des clients avec une file de capacité K_1 finie et dans lequel les clients qui abandonnent ou qui sont déconnectés peuvent rappeler plus tard avec une probabilité p et ce, suivant une loi exponentielle de taux δ. Ce nouveau système n'admet plus de renoncements puisque la fonction de renoncement modélise, désormais, une taille finie de la file réelle. De plus, c'est un cas particulier du modèle générique étudié, le modèle fluide devrait, donc, bien marcher avec ce système également.

Pour vérifier cette intuition, nous comparons l'approximation fluide avec le modèle stochastique pour ce nouveau système. Nous considérons deux exemples, un dans lequel l'espace d'attente est limité à 5 clients ($K_1 = C + 5$), et pour l'autre la limitation atteint 10 clients ($K_1 = C + 10$). Les paramètres et les résultats obtenus pour les deux exemples sont illustrés par le Tableau 3.3 pour une charge égale à 133 %. La Figure 3.5 illustre l'erreur des deux systèmes. Nous constatons que, même pour un système avec une file d'attente finie, la précision de l'approximation fluide croît rapidement en fonction de la taille du système. En fait, la précision obtenue est plus importante que celle observée dans le cas d'un espace d'attente illimité et avec le renoncement modélisé comme précédemment. Des résultats similaires à ce que nous avions déjà montré auparavant sont obtenus pour le cas où nous fixons le nombre de serveurs C à 40 serveurs et nous faisons varier la charge du système ρ. Ces résultats sont résumés par le Tableau 3.4 et la Figure 3.6.

	$K_1 = C + 5$		$K_1 = C + 10$	
C	Tr	δx_2	Tr	δx_2
5	0,59	0,50	0,58	0,50
10	1,09	1,00	1,07	1,00
15	1,59	1,50	1,55	1,50
20	2,08	2,00	2,04	2,00
25	2,58	2,50	2,54	2,50
30	3,08	3,00	3,03	3,00
35	3,57	3,50	3,53	3,50
40	4,07	4,00	4,03	4,00
45	4,57	4,50	4,52	4,50
50	5,07	5,00	5,02	5,00

Tableau 3.3: Comparaison de l'approximation du taux de rappel stationnaire avec la valeur exacte suivant la taille de la file d'attente pour ρ = 1,33

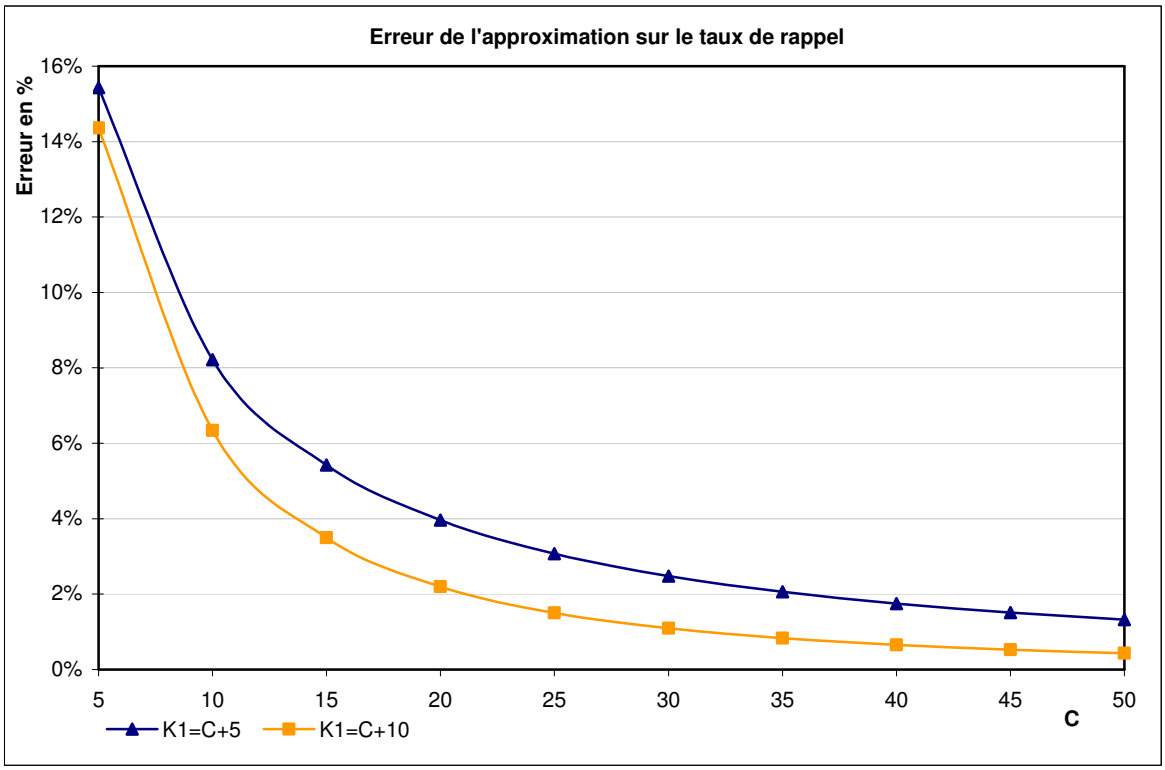


Figure 3.5: Évolution de l'erreur de l'approximation par le modèle fluide du taux de rappel en fonction du nombre de serveurs C pour ρ = 133 % dans un système avec une file d'attente finie

	$K_1 = C + 5$		$K_1 = C + 10$	
ρ	*Tr*	δx_2	*Tr*	δx_2
100%	0,81	0,00	0,72	0,00
110%	1,60	1,20	1,50	1,20
120%	2,59	2,40	2,51	2,40
130%	3,69	3,60	3,64	3,60
140%	4,84	4,80	4,81	4,80
150%	6,02	6,00	6,00	6,00
160%	7,21	7,20	7,20	7,20
170%	8,41	8,40	8,40	8,40
180%	9,60	9,60	9,60	9,60
190%	10,80	10,80	10,80	10,80
200%	12,00	12,00	12,00	12,00

Tableau 3.4: Comparaison de l'approximation du taux de rappel stationnaire avec la valeur exacte suivant la taille de la file d'attente pour $C = 40$

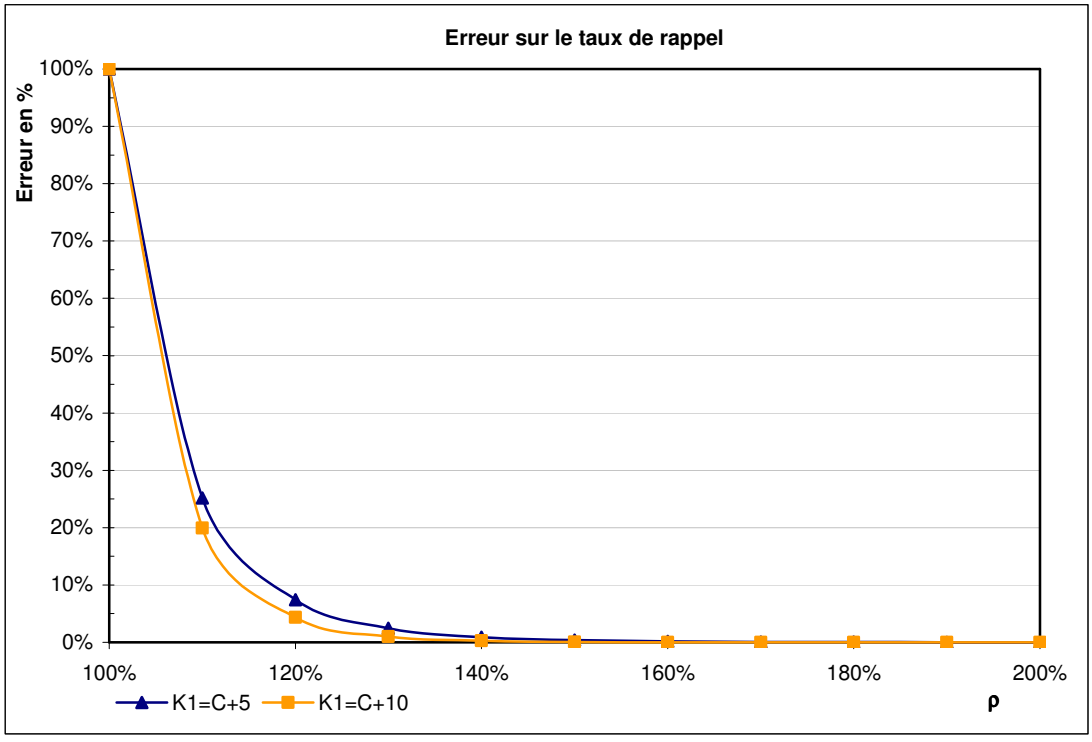


Figure 3.6: Évolution de l'erreur de l'approximation par le modèle fluide du taux de rappel en fonction de la charge ρ pour $C = 40$ dans un système avec une file d'attente finie

3.4 Analyse du système au régime transitoire

Dans la plupart des cas, les paramètres que nous utilisons pour modéliser le centre d'appels, varient en fonction du temps. Dans le cas des arrivées, ceci représente un changement d'envie ou de besoin des clients pour contacter le système durant une période particulière de la journée et ces arrivées sont souvent modélisées comme étant des arrivées non stationnaires. Les clients qui veulent rejoindre le système au début de la journée ne sont pas nécessairement aussi nombreux que ceux qui veulent le rejoindre pendant l'après-midi ou à la fin de la journée. Il est tout à fait banal d'avoir des pics d'arrivées au début et à la fin de la journée pour les centres d'appels qui ne fonctionnent pas 24 heures par jour. Même pour les centres d'appels fonctionnant sans arrêt, les taux d'appels varient clairement entre le jour et la nuit. Le nombre de serveurs varie souvent, lui aussi, au cours d'une même journée, partiellement en réponse à la variation du taux d'appels. Une autre raison de sa variation vient des contraintes contractuelles et sociales de planification. Le dimensionnement des centres d'appels traite, généralement, une journée de travail comme étant une succession de plusieurs périodes durant lesquelles le nombre de serveurs demeure constant. Dans cette section, nous allons considérer qu'une telle période de temps a une durée de trente minutes, ce qui représente la réalité de la plupart des centres d'appels.

Considérons maintenant une période (faisant partie des périodes décomposant la journée) quelconque de durée T de la journée. Pendant cette période, les paramètres du système λ, C, μ, θ, p, δ et $r(x)$ restent constants. Nous supposons que la période commence avec une file réelle de taille x_1^0 et une orbite de taille x_2^0. Le système d'équations différentielles

$$\begin{cases} \dfrac{dx_1}{dt} = \left(1 - r\left(x_1(t)\right)\right)\left(\delta\, x_2(t) + \lambda\right) - \mu\, Min\left(x_1(t), C\right) - \theta\, Max\left(x_1(t) - C, 0\right) \\ \dfrac{dx_2}{dt} = p\, r\left(x_1(t)\right)\left(\delta\, x_2(t) + \lambda\right) - \delta\, x_2(t) + \theta\, p\, Max\left(x_1(t) - C, 0\right) \end{cases} \tag{3.20}$$

nous permet, en le résolvant, de déterminer l'évolution des deux files (réelle et fictive) au cours du temps mais uniquement à l'intérieur de la période de temps considérée. Avec ceci, nous pouvons en particulier calculer la taille de la file réelle x_1^T ainsi que la taille de l'orbite x_2^T à la fin de la période de durée T.

La résolution analytique du système différentiel précédent n'étant pas possible, nous avons recours à une résolution numérique facile à effectuer. Cette résolution nous fournira, à partir des paramètres du système pendant la période choisie et à partir de l'état initial (x_1^0,x_2^0), l'évolution des files réelle et fictive au-cours de la période. En pratique, pour aboutir à la description du système pendant toute la journée, nous commençons par la première période de la journée avec généralement un état initial $(x_1^0,x_2^0) = (0,0)$. La résolution du système pendant cette première période aboutit, entre autres, à la détermination de l'état final (x_1^T,x_2^T). Cet état final va constituer l'état initial de la deuxième période pendant laquelle les paramètres peuvent changer de valeurs par rapport à la période précédente. Nous répétons cette procédure jusqu'à la dernière période de la journée et, de cette manière, nous aboutissons à la détermination de l'évolution de l'état du système au-cours de la journée entière. La procédure qui permet d'évaluer le comportement du système (à partir des équations (3.20)) en multi-période est décrite par l'algorithme suivant :

Algorithme d'analyse en Multi-période

- Étape 1 : Initialisation : $x_1^0 = x_2^0 = 0$. En utilisant le système différentiel (3.20), déterminer $x_1(t)$ et $x_2(t)$ pour $0 \le t \le T$. Lors de la résolution du système différentiel, il faut faire attention à utiliser les paramètres (C, λ, …) correspondants à la période en cours. Poser $x_1^T = x_1(T)$ et $x_2^T = x_2(T)$.

- Étape 2 : Initialisation : $x_1^0 = x_1^T$ et $x_2^0 = x_1^T$. En utilisant le système différentiel (3.20), déterminer $x_1(t)$ et $x_2(t)$ pour $T \le t \le 2.T$. Poser $x_1^T = x_1(2.T)$ et $x_2^T = x_2(2.T)$.

- …

- Étape i, $i \ge 3$: Initialisation : $x_1^0 = x_1^T$ et $x_2^0 = x_1^T$. En utilisant le système différentiel (3.20), déterminer $x_1(t)$ et $x_2(t)$ pour $(i\text{-}1)T \le t \le i\ T$. Poser $x_1^T = x_1(i\ T)$ et $x_2^T = x_2(i\ T)$.
- Répéter l'étape précédente jusqu'à $i = n$: nombre de périodes constituant la journée.

3.4.1 Exemples numériques

En se basant sur des paramètres et des données issus d'un centre d'appels réel, nous voulons afficher l'évolution du taux de rappel que reçoit le système au-cours de la journée. Le taux de rappel durant une journée de fonctionnement de ce centre d'appels est déterminé en utilisant l'algorithme proposé précédemment ainsi que l'expression $\delta x_2(t)$ pour le taux de rappel. Ceci est comparé avec le résultat de la simulation du système en question dans le but de valider l'approximation fluide. La comparaison a été effectuée pour trois systèmes que nous pouvons décrire comme étant le système actuel (Système 1), le système actuel soumis à des taux d'arrivée plus importants (Système 2), et un système où le nombre total de serveurs a été reparti équitablement entre toutes les périodes de la journée (Système 3). Les paramètres communs aux trois systèmes sont présentés par le Tableau 3.5.

Le Tableau 3.6 affiche les taux d'arrivée et le nombre de serveurs pour chaque système et suivant la période de la journée considérée. Le nombre total cumulé de serveurs pour chacun des trois systèmes est égal à 3135 serveurs par jour.

μ	θ	δ	p	β	θ_1
0,3	0,5	0,1	0,6	0,2	1

Tableau 3.5: Les paramètres constants au-cours de la journée pour les systèmes analysés

Période		Système 1		Système 2		Système 3	
Début	Fin	*C*	λ	*C*	λ	*C*	λ
09:00	09:30	86	68	86	110	175	68
09:30	10:00	114	75	114	110	175	75
10:00	10:30	177	101	177	115	175	101
10:30	11:00	180	87	180	100	174	87
11:00	11:30	197	82	197	94	174	82
11:30	12:00	192	80	192	89	174	80
12:00	12:30	169	73	169	75	174	73
12:30	13:00	155	74	155	78	174	74
13:00	13:30	169	67	169	72	174	67
13:30	14:00	124	74	124	80	174	74
14:00	14:30	140	70	140	80	174	70
14:30	15:00	238	68	238	71	174	68
15:00	15:30	231	72	231	70	174	72
15:30	16:00	235	69	235	67	174	69
16:00	16:30	215	67	215	64	174	67
16:30	17:00	214	69	214	73	174	69
17:00	17:30	163	69	163	74	174	69
17:30	18:00	136	73	136	88	174	73

Tableau 3.6: Les paramètres variables au-cours de la journée pour les systèmes analysés

Dans la Figure 3.7, nous constatons que les courbes des systèmes 1 et 2 présentent à peu près la même allure. La différence majeure vient de l'échelle plus importante pour le système 2 – ce qui est normal puisqu'il présente justement un taux d'appel λ plus important que celui du système 1. Nous observons pour ces deux systèmes que l'approximation fluide donne lieu à un résultat très précis pendant la majeure partie de la journée. La seule partie qui pose un petit problème se situe entre 15:30 et 16:30. Si l'on regarde l'évolution de la charge au-cours du temps, nous remarquons que ρ ($\lambda / C.\mu$) se situe pour les deux systèmes entre 0,95 et 1,04 de 14:30 à 16:30 ce qui correspond à ses plus bas niveaux de la journée. Ceci explique le manque de précision par rapport au reste de la journée. Toutefois, l'imprécision de l'approximation admet un retard par rapport à la diminution de la charge. En effet, entre 14:30 et 15:30 la courbe de l'approximation coïncide toujours avec celle donnée par la simulation. Pour expliquer ceci, il faut mentionner que la charge réelle du système est égale à :

$$\rho_{réelle} = \frac{\lambda_o}{C\,\mu}$$

Cette charge peut être supérieure à 1 même lorsque ρ ne l'est pas. C'est ce qui se passe à 14:30 puisque l'orbite du système n'est pas encore vide et donc elle va participer à l'alimentation du système en appels et à hausser, ainsi, la charge réelle du centre d'appels. Au bout d'une longue période où ρ est inférieure à 1, l'orbite (qui s'est vidée progressivement) ne suffit plus à maintenir $\rho_{réelle}$ au dessus de 1 et dans ce cas l'approximation fluide va donner un taux de rappel nul (puisqu'elle néglige les effets stochastiques). De toute façon, lorsque $\rho_{réelle}$ est inférieure à 1, le taux de rappel exact n'est pas important, surtout pour un grand centre d'appels, ce qui minimise l'importance de l'erreur de l'approximation dans cette zone.

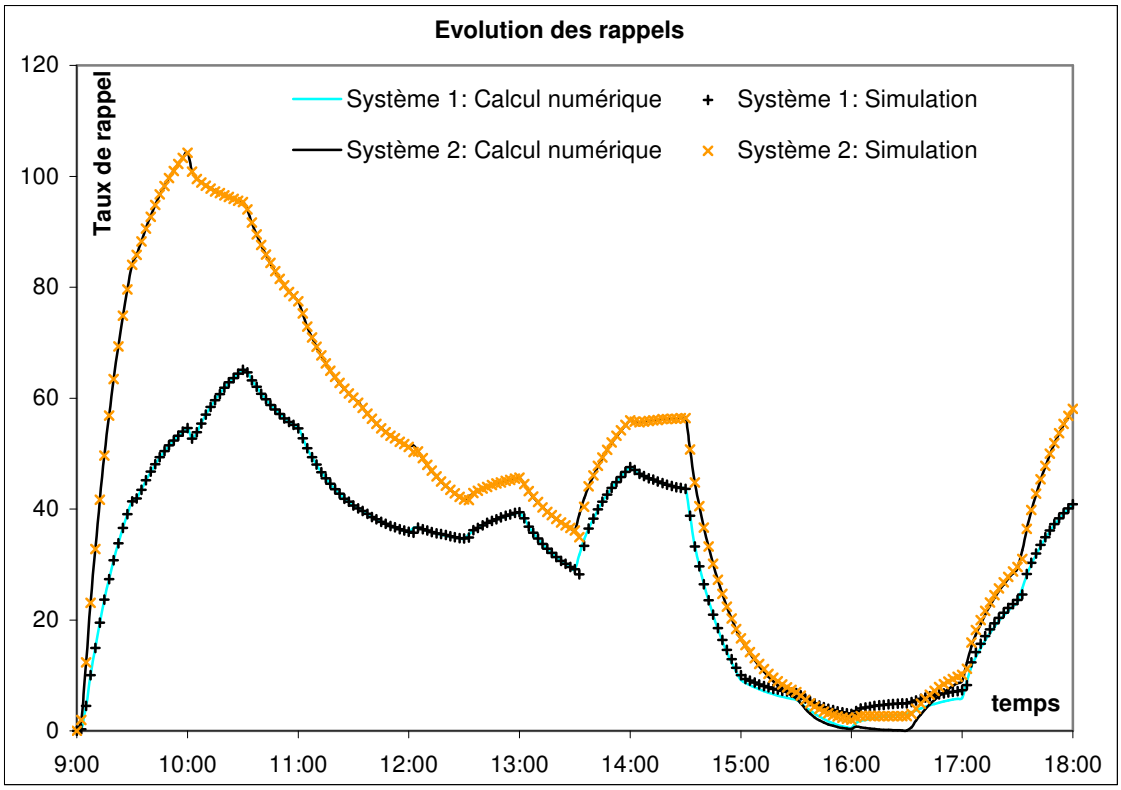


Figure 3.7: Comparaison des résultats de la simulation avec les résultats de l'analyse numérique pour les systèmes 1 et 2

Dans la Figure 3.8, nous comparons les systèmes 1 et 3. Signalons à nouveau que ces deux systèmes possèdent les mêmes taux d'arrivée et que la différence entre eux vient de la distribution du nombre total de serveurs au-cours de la journée. Une fois encore, nous constatons qu'il y a une très bonne similitude entre les courbes que donnent l'analyse numérique et la simulation. En fait, puisque la redistribution des

serveurs pour le système 3 implique des charges supérieures à 1 pendant chaque période, nous observons que les résultats de l'approximation (pour le système 3) coïncident parfaitement avec la simulation pour toutes les périodes. Cette fois, l'allure des courbes de rappel n'est pas la même pour les systèmes 1 et 3. L'exemple de la Figure 3.8 montre l'importance des décisions prises lors de la planification des conseillers. Cet exemple nous montre également les effets des dites décisions sur le phénomène de rappel. En effet, une amélioration de la planification des CDCs peut induire moins de rappels avec une meilleure (plus faible) variabilité.

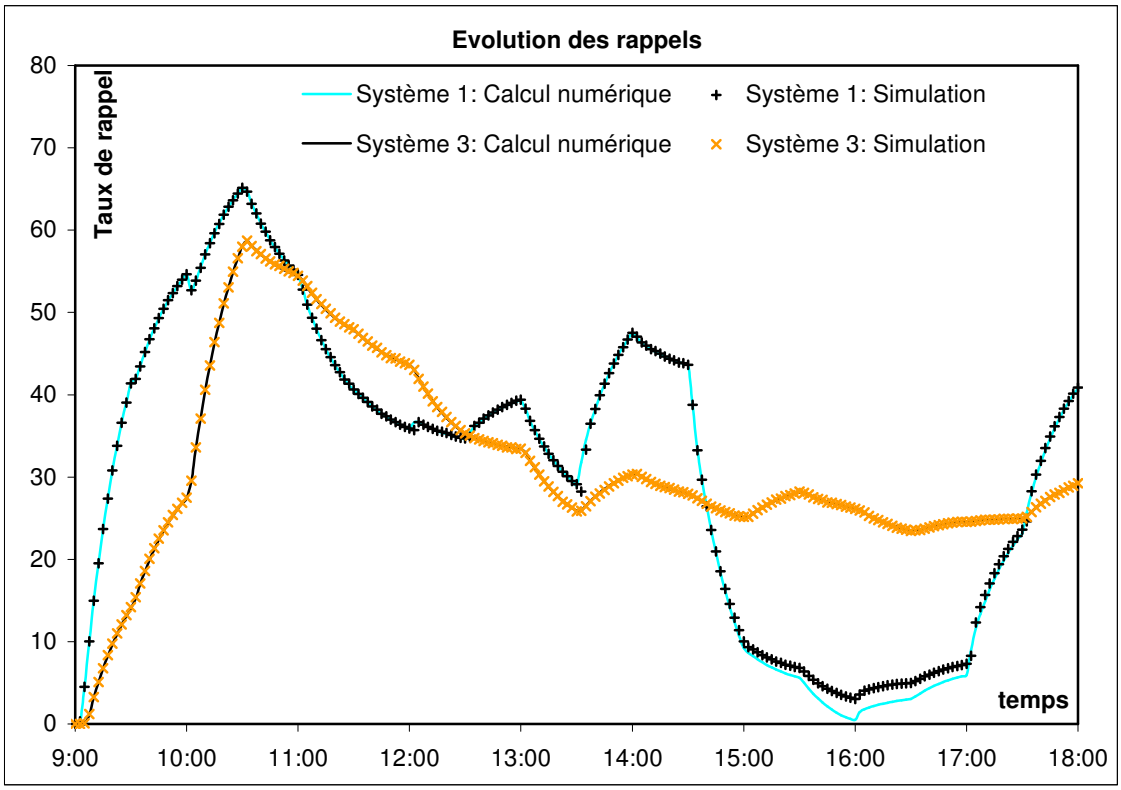


Figure 3.8: Comparaison des résultats de la simulation avec les résultats de l'analyse numérique pour les systèmes 1 et 3

3.4.2 Estimation des appels frais à partir des appels observés

Une fois le calcul de $x_1(t)$ et de $x_2(t)$, $0 \le t \le T$, effectué, l'évolution du taux d'appels observés devient connue pendant toute la période à l'aide de la formule

$$\lambda_o(t) = \lambda + \delta\, x_2(t) \tag{3.21}$$

Le taux moyen d'appels observés pendant la période a donc comme valeur

$$\lambda_o = \frac{1}{T} \int_0^T \lambda_o(t)\, dt$$

et il se calcule finalement en utilisant l'expression

$$\lambda_o(\lambda) = \lambda + \frac{\delta}{T} \int_0^T x_2(t)\, dt \tag{3.22}$$

Les managers de centres d'appels disposent, le plus souvent, d'une information concernant le taux d'appel observé λ_o et non le taux d'appels primaires λ. Ceci rend particulièrement intéressant la recherche de λ à partir de λ_o. Pour retrouver ce taux, il suffit de faire une inversion numérique de l'équation (3.22) pour retrouver λ en fonction de λ_o et ce, en utilisant toujours le système différentiel (3.20). Afin de déterminer λ en fonction de λ_o au-cours de toute la journée, il suffit d'enchaîner les périodes en utilisant l'état final d'une période comme point de départ pour la période suivante.

La Figure 3.9 montre un exemple dans lequel nous avons déterminé λ à partir de λ_o et ce, avec des données représentatives d'un centre d'appels réel. En particulier, les paramètres affichés par le Tableau 3.5 ont été utilisés dans cet exemple ainsi que la distribution du nombre de serveurs des systèmes 1 et 2 du Tableau 3.6. Les arrivées que nous avons considérées comme appels primaires pour le système 2 du Tableau 3.6 sont prises maintenant comme appels observés. En utilisant ces paramètres, l'évolution du taux d'appels primaires résultant (en inversant numériquement l'équation (3.22), avec la méthode de dichotomie, par exemple) est illustrée par la Figure 3.9. Dans cette figure, nous pouvons constater une différence importante entre les deux taux d'appels (les frais et les observés) au début de la journée. Cette période de la journée correspond à des arrivées primaires qui dépassent la capacité de service du système, ce qui implique une accumulation des renouvellements d'appels. À mesure que la capacité de service augmente, les deux courbes des arrivées se rapprochent l'une de l'autre jusqu'à coïncider lorsque la capacité du système devient suffisante à la satisfaction des appels primaires et des rappels issus des périodes précédentes. Pour cet exemple en

particulier, nous observons que la courbe des appels frais est relativement plate. Si les appels observés étaient la base de la planification du système, alors il est clair que le résultat aurait été un système où la capacité varie beaucoup plus qu'il ne le faut au-cours de la journée. Dans ce cas, la capacité ne serait pas adaptée à la demande des clients ce qui devrait, certainement, mener à des clients non satisfaits (puisque dans l'obligation de rappeler) alors qu'une autre planification avec les mêmes ressources réduirait cette proportion de clients insatisfaits.

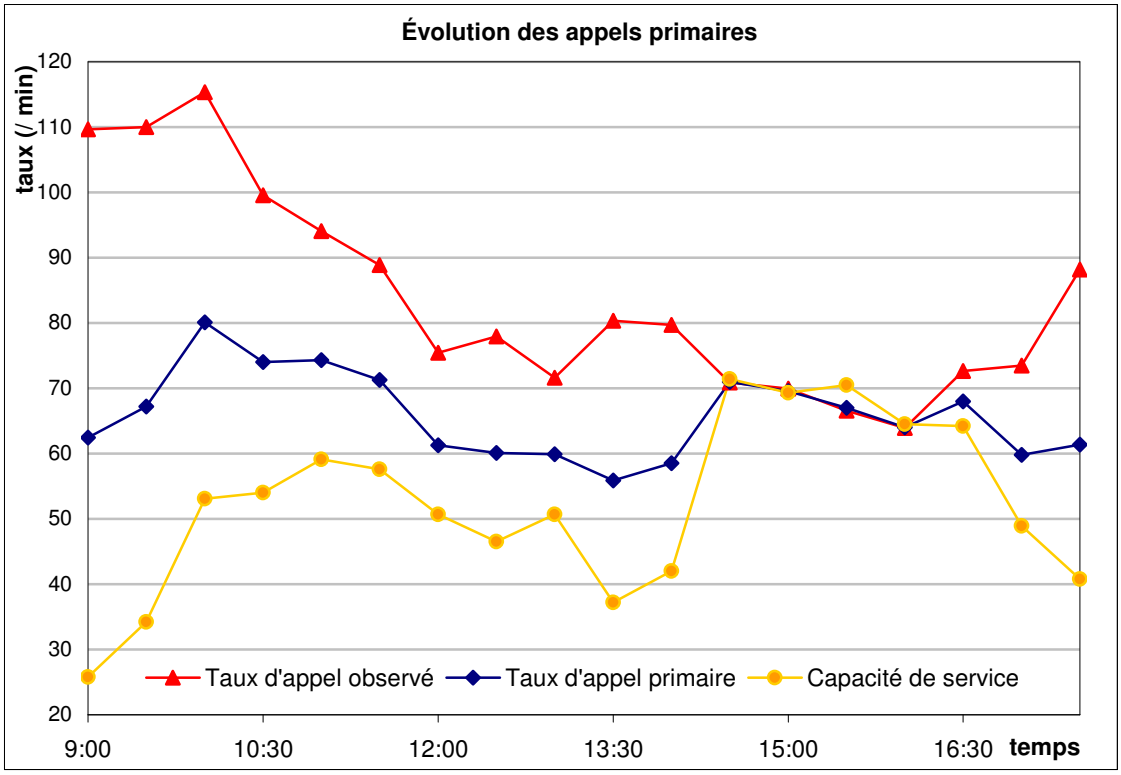


Figure 3.9: Évolution du taux d'appels primaires λ en fonction du taux d'appels observés λ_o

3.5 Conclusions

Nous avons analysé dans ce chapitre le phénomène de rappel dans un centre d'appels avec abandons et renoncements. Nous avons abordé le système dans deux modes de fonctionnement: en mono-période et en multi-période. Pour l'étude mono-périodique, une approximation fluide a été proposée pour estimer le taux de rappel dans le système. Cette approximation fournit une expression analytique du taux de rappel qui

est facile à utiliser. En utilisant cette expression analytique, nous avons montré une insensibilité du taux de rappel de plusieurs paramètres importants du système comme le taux d'abandon, le taux unitaire de rappel, et la distribution de la probabilité de renoncement. Cette propriété d'insensibilité nous a permis d'utiliser la même approximation pour des modèles possédant des files d'attente de tailles finies. Nous avons montré par l'intermédiaire d'exemples numériques que l'approximations fluide fonctionne très bien pour de gros centres d'appels qui ont une charge supérieure à 1. Pour ρ inférieure à 1, l'approximation n'est pas convenable. Il serait intéressant de coupler la méthode avec une bonne approximation du nombre moyen de serveurs occupés. Ceci représente une bonne évolution future de l'étude menée dans ce chapitre. Toutefois, il faut noter que nous sommes moins intéressés par des systèmes avec des charges inférieures à 1 puisque ces systèmes ne présentent pas assez de rappels pour voir leurs performances inquiétées.

Nous avons également utilisé l'approximation fluide pour aborder le système en multi-période. Dans cette étude, le taux de rappel ne peut plus être estimé à travers une formule analytique. Pour cela, une analyse numérique est requise. Une comparaison avec la simulation montre l'efficacité de la méthode. En utilisant des données proches du cas d'un centre d'appels réel, des exemples numériques montrent que le phénomène de rappel peut être considérable et qu'il a un effet important sur l'évaluation des performances et sur l'optimisation ultérieure du système s'il n'est pas pris en compte. En comparant trois systèmes avec des taux d'arrivée différents et avec plusieurs distributions des serveurs à travers les périodes de la journée, nous avons montré que les taux de rappels ainsi que la distribution des serveurs ont un impact significatif sur le taux de rappel. Ainsi, disposer de la mauvaise estimation du taux d'appels (c'est à dire avoir le taux d'appels observés au lieu du taux d'appels frais) ou d'une planification incorrecte des serveurs au-cours de la journée, affecte les renouvellements d'appels générés. De plus, nous avons constaté que l'allure de la courbe représentant l'évolution du taux d'appels observés dans la journée peut différer sensiblement de celle qui illustre le taux d'appels frais, constat qui prouve l'importance d'analyser le phénomène de rappel dans les systèmes non stationnaires. Ceci constitue une conclusion importante pour les managers de centres d'appels puisque les modèles et les logiciels utilisés généralement dans les centres d'appels ignorent le phénomène de rappel. Pour les centres d'appels fonctionnant avec des charges faibles ou modérées,

ceci ne pause pas de problème. Cependant, ceci conduit à des erreurs qui font augmenter les coûts opérationnels pour les centres d'appels qui fonctionnent sous des charges importantes. Nous avons également montré comment le taux d'appels frais peut être estimé à partir de l'historique représentant le volume total des appels. Si les futurs logiciels peuvent permettre aux centres d'appels de distinguer un rappel d'un appel frais au moment de son arrivée, ceci reste utile pour nettoyer les données historiques.

Nous avons démontré dans ce chapitre qu'il y a une forte interaction entre le dimensionnement et les renouvellements d'appels. Comme résultat de ceci, nous avons vu que les rappels peuvent avoir un impact important sur la performance du centre d'appels. Dans le chapitre précédent, nous avons montré comment dimensionner le centre d'appels en présence des rappels dans un système mono-périodique. Dans des travaux futurs, nous pouvons regarder comment considérer la planification dans un système multi-périodes qui prend en compte le phénomène de rappels.

Chapitre 4 : Estimation des temps d'attente dans un centre d'appels

4.1 Introduction

Dans le cadre de notre collaboration avec *Bouygues Telecom*, nous avons mené plusieurs études articulées autour de l'analyse du temps d'attente des clients dans leur centre d'appels. L'objectif de ces études consiste, en premier lieu, à fournir des informations qui facilitent la prise de décisions d'ordre stratégique quant au fonctionnement du centre d'appels. En second lieu, il s'agit de se préoccuper de l'amélioration de la qualité de service perçue par les clients de *Bouygues Telecom*.

Le centre d'appels exploité actuellement par *Bouygues Telecom* se compose de plusieurs sites répartis sur des lieus géographiques distincts. Chaque site comporte un certain nombre de Conseillers de Clientèle (CDC) pour satisfaire la demande des clients. La taille des sites, mesurée en nombre de conseillers, change d'un site à un autre. Chaque site fonctionne comme un "petit" centre d'appels qui possède sa propre file d'attente. Les clients sont routés vers l'un des sites existants au moment de leur appel. Une fois routé, l'appel ne pourra plus être déplacé vers un autre site. Le traitement des appels de la part des conseillers est le même pour tous les clients. Toutefois, le système distingue plusieurs classes de clients. Chaque classe se caractérise par sa priorité. Cette priorité étant non-préemptive, l'arrivée d'un client prioritaire

n'influe pas sur les clients déjà en service, elle se traduit uniquement par un dépassement des clients de moindre priorité dans la file d'attente. De cette manière, une fois un client routé vers un site donné, s'il n'y a pas de conseillers disponibles pour répondre à son appel immédiatement alors il va dépasser tous les clients de moindre priorité dans la file d'attente pour se positionner derrière les clients qui possèdent une priorité supérieure ou égale à la sienne.

Dans ce chapitre, nous abordons le problème de l'estimation du temps d'attente des clients dans la file. Cette estimation peut être exploitée pour mieux analyser, au sein du centre d'appels, l'attente des clients avant service. Elle peut, également, être à l'origine du routage des clients à leurs arrivées, ceci sera traité dans le prochain chapitre. L'estimation du temps d'attente doit être effectuée en exploitant uniquement les informations disponibles afin d'être facilement utilisable dans le centre d'appels de *Bouygues Telecom*. En fait, comme dans beaucoup de centres d'appels, une estimation du temps d'attente est déjà intégrée au système puisqu'elle est incorporée dans plusieurs logiciels standards. L'estimateur utilisé est appelé ASA (Average Speed Approximation). Son manque de fiabilité est à l'origine de l'étude visant à le remplacer. Lors de cette étude, nous allons proposer un nouvel estimateur du temps d'attente. L'élaboration de cet estimateur est issue d'une estimation exacte du temps d'attente et ce, en fonction de la classe du client qui arrive au système. Une comparaison de notre estimateur sera effectuée avec l'ASA pour montrer qu'il est meilleur. En même temps, nous allons analyser la précision du nouvel estimateur pour connaître sa performance réelle.

L'estimateur ASA peut, également, être utilisé pour router les appels aux différents sites qui composent le système. Lors de l'arrivée d'un nouvel appel, une estimation du temps d'attente sur chaque site est effectuée et l'appel est routé vers le site offrant, a priori, le temps d'attente le moins élevé. En plus de cette minimisation du temps d'attente, le routage utilisé doit garantir un équilibrage du flux des arrivées entre les sites.

Dans ce chapitre, nous allons également analyser la performance du système actuel qui consiste à partager le centre d'appels en plusieurs sites. Une comparaison avec un système où l'ensemble des conseillers forment un seul et unique site est menée afin de

voir s'il est judicieux de migrer vers un tel système en fusionnant les différents sites. Pour ce genre de systèmes, nous nous passons de règle de routage puisque tous les clients doivent passer par la même file d'attente. Nous allons désigner ce système par "file logique" en référence à la file unique qu'il comporte et qui n'est pas gérée sur un site en particulier comme c'est le cas actuellement (où chaque site possède sa propre file d'attente). Dans le prochain chapitre, nous allons continuer la comparaison des deux systèmes dont les performances se calculent en nombre de conseillers nécessaires à la satisfaction d'une qualité de service objectif, laquelle concerne la distribution ou la moyenne du temps d'attente des clients avant service.

Nous avons suivi deux démarches pour aborder les problèmes expliqués précédemment. La première s'appuie sur des modélisations analytiques qui prennent en compte les effets stochastiques, tels que les taux des différentes arrivées et les temps de service, mais qui ne considèrent pas les aspects fins tels que les politiques de routage. Cette démarche a comme objectif de nous éclairer sur le comportement qualitatif du système. La deuxième démarche, quant à elle, se base sur des modèles de simulations à évènements discrets. Nous pouvons, ainsi, prendre en compte les aspects de routage en plus des effets stochastiques. Les résultats de ce chapitre sont à l'origine de deux brevets déposés auprès de l'inpi. Il s'agit de Aguir *et al.* [3] et de Aguir *et al.* [2].

Après cette introduction, nous allons présenter une étude bibliographique des travaux traitant, essentiellement, l'estimation des temps d'attente. En fait, dans ce chapitre, l'un des objectifs consiste à estimer le temps d'attente des clients à leurs arrivées. Nous allons également nous intéresser à une comparaison d'un système constitué de plusieurs files d'attente avec un système à une file unique. Cette comparaison est effectuée en termes de temps d'attente. Le routage des clients vers les différents sites dans un système à plusieurs files va se baser sur une estimation du temps d'attente.

Dans la troisième section, nous allons présenter le système actuel ainsi que le système de file logique. Nous allons, également, expliquer le pourquoi de l'intuition du système à une file unique.

Dans la section 4, nous allons analyser le temps d'attente des clients. Cette analyse consiste en une estimation de l'attente de chaque client au moment de son arrivée au centre d'appel. Un routage basé sur cette estimation sera analysé dans le prochain chapitre. Ceci nous donnera l'occasion de comparer le système de file logique avec le système à plusieurs files d'attente. Ici, plusieurs estimateurs sont analysés.

Nous terminons le chapitre par les conclusions et les perspectives de cette étude.

4.2 Étude bibliographique

Dans cette section, nous présentons les travaux relatifs à l'estimation du temps. Nous pouvons distinguer deux grandes classes de travaux: ceux qui concernent l'estimation du temps d'attente d'un client au moment de son arrivée, et ceux qui étudient les temps d'attente au régime stationnaire. Nous allons également présenter des travaux qui correspondent à des files d'attente avec priorités puisque les modèles que nous allons étudier dans ce chapitre comportent plusieurs classes de clients.

L'estimation des temps d'attente au régime stationnaire a fait l'objet de plusieurs livres. Parmi ces livres, nous pouvons citer Gross et Harris [33] ainsi que Medhi [55]. Dans ces deux ouvrages, plusieurs modèles de files d'attente sont étudiés et les principaux paramètres de performance sont déterminés. Ces paramètres comportent, par exemple, la taille moyenne de la file d'attente, le temps moyen de séjour dans le système ainsi que d'autres mesures de performance. Nous pouvons également y trouver le temps moyen d'attente dans un modèle de file d'attente particulier. Kleinrock [48] représente un autre ouvrage "classique" où des modèles de files d'attente de base sont abordés.

L'analyse du temps d'attente a suscité l'intérêt de beaucoup d'autres travaux. Nous pouvons commencer par citer Bertsimas [11] où une analyse d'un modèle de file d'attente G/G/c a été effectuée. L'auteur propose des expressions analytiques et asymptotiques pour la distribution du temps d'attente au régime stationnaire. Il propose également un algorithme pour résoudre le système numérique. Cet algorithme est moins complexe que la méthode de la Matrice Géométrique appliquée au même problème. Ceci étant, cette complexité reste, tout de même, importante et c'est pour

cela que la méthode en question est plus adaptée à une étude qualitative. Un autre travail qui s'intéresse à la distribution des temps d'attente est celui de Davis [21]. Il y a traité une file d'attente multi-serveur dans laquelle la demande arrive de la part de plusieurs classes de clients. La priorité étant non préemptive. Pour le même système précédent, Kella et Yechiali [45] déterminent les deux premiers moments du temps d'attente pour chaque classe de clients. Une approximation asymptotique du temps moyen d'attente ainsi qu'une approximation de la distribution du temps d'attente au régime stationnaire sont proposées par Whitt [69] pour un modèle de file d'attente mono-classe et mono-serveur. Dans ce travail, l'auteur montre également l'effet de la variabilité des arrivées et des temps de service sur le comportement du système. Ce travail fait partie d'une large gamme de travaux qui proposent différentes approximations pour manipuler plus facilement les modèles qu'ils étudient. Ces approximations traitent, le plus souvent, des systèmes intégrant des temps inter-arrivées ou bien des temps de service qui suivent une loi Générale. Nous pouvons citer comme exemple Kimura [46] et Kimura [47] pour des temps de service de distribution Générale.

Concernant l'estimation du temps d'attente lors de l'arrivée d'un client (dans le but de la lui annoncer), nous pouvons citer le travail de Whitt [71]. Dans cet article, l'auteur traite l'estimation du temps d'attente lors de l'arrivée des clients à un système multi-serveur. Il étudie également les avantages que procure la disposition d'une information supplémentaire sur les besoins des clients en termes de temps de service (le temps de service est distribué suivant une loi Générale). Dans Whitt [70], l'auteur analyse les avantages d'annoncer le temps d'attente aux clients à leurs arrivées. Selon lui, il y a moins d'abandons dans un système avec annonce du temps d'attente puisqu'un client averti de son attente a tendance à patienter du moment où il ne raccroche pas immédiatement lorsqu'il sait qu'il doit attendre. Hui et Tse [39] ont analysé diverses informations qui peuvent être communiquées aux clients à leurs arrivées, suivant le temps d'attente qu'ils vont devoir passer dans la file. Nakibly [57] a étudié l'estimation du temps d'attente à l'arrivée des clients. Il a regardé plusieurs modèles avec différentes lois de service et avec la possibilité d'avoir un taux de service dépendant de la classe de client.

Dans ce chapitre, nous allons étudier les conséquences d'une éventuelle migration vers un système à une file d'attente commune. À première vue, et comme nous allons le mentionner dans le reste du chapitre, il est intuitif de passer à ce genre de systèmes. Nous effectuons, dans la suite, une comparaison de ce système avec un système où chaque site possède sa propre file d'attente, et dans lequel un routage des clients est nécessaire. Rothkopf et Rech [60] ont suggéré que le fait de fusionner les files d'attente n'améliore pas nécessairement la performance du système. Ceci concerne spécialement des systèmes où les clients peuvent changer de file lorsqu'ils s'aperçoivent, par exemple, qu'il y a un autre serveur qui se libère. Ce n'est pas le cas pour les centres d'appels où les clients n'ont pas cette possibilité. Whitt [68] a étudié les avantages de partager les clients en plusieurs classes suivant leurs temps de service. En faisant ainsi, les clients qui ont des temps de service importants ne sont pas traités par les mêmes conseillers que les clients qui n'ont pas une demande aussi longue. Ceci rappelle les caisses rapides que l'on rencontre en faisant les courses. Bien entendu, ce partage suppose que les clients ayant à passer beaucoup de temps sont reconnus à l'avance. Ce n'est pas le cas ici puisque nous supposons que le temps de service des clients suit la même loi, indépendamment de la classe du client. Néanmoins, cela pourrait être considéré puisque le type de question peut être connu à l'avance, ce qui donne une idée sur le temps de service.

Passons maintenant aux travaux traitant les files avec priorités. Nous avons déjà cité Davis [21] et Kella et Yechiali [45] pour des files d'attente avec priorité non-préemptive. Le régime stationnaire du modèle multi-serveur avec priorité non-préemptive, a été étudié par Gail, Hantler et Taylor [26] ainsi que par Kao et Narayanan [43]. Miller [56] a étudié le calcul des probabilités au régime stationnaire dans une file d'attente mono-serveur avec deux disciplines de priorité: préemptive et non-préemptive. Il a développé une approche numérique basée sur la méthode de Matrice Géométrique développée par Neuts [58]. Choi, Kim et Chung [20] ont analysé une file d'attente dans laquelle ils ont intégré les abandons des clients de classe prioritaire. Une étude bibliographique plus détaillée des modèles consacrés aux files d'attente avec priorités, sera menée dans le chapitre 6.

4.3 Modélisation du centre d'appels

Dans cette section, nous allons expliquer le fonctionnement du centre d'appels tel qu'il est à l'heure actuelle. Nous allons voir que le système se compose de plusieurs files d'attente séparées. Chaque file d'attente alimente un seul groupe de conseillers en appels. Nous allons introduire la notion de "fusion" des files d'attente pour ne constituer qu'une seule que l'on désignera par *file logique*. Nous allons présenter le modèle de la file logique tout en expliquant l'intuition de ce choix.

4.3.1 Le système actuel

Le système dont nous expliquons le fonctionnement ici correspond, en fait, à une approximation du comportement du centre d'appels réel. En effet, nous ne comptons pas, par exemple, modéliser les abandons ni les renouvellements des appels que nous avions déjà utilisés dans les deux chapitres précédents. Ceci étant, l'approche qualitative et, dans une moindre mesure, l'approche quantitative restent sensiblement les mêmes.

Comme mentionné précédemment, le système réel comporte plusieurs sites. Dans l'étude que nous présentons dans ce chapitre, nous nous limiterons à quatre sites situés sur des lieux géographiques distincts. Dans la suite, nous désignerons ce système par "système actuel". Plusieurs conseillers de clientèle sont présents sur chaque site. Le nombre de conseillers par site n'est pas nécessairement le même : il peut varier d'un site à un autre. Si tous les sites possèdent le même nombre de conseillers, alors nous allons appeler ce système par "équilibré". Dans le cas contraire, nous l'appellerons "déséquilibré".

Le système actuel est illustré par la Figure 4.1. Dans cette figure, nous pouvons constater l'existence de trois classes de clients. La classe la plus prioritaire étant la classe *A*. La classe *B* est celle qui possède une priorité intermédiaire alors que les clients de classe *C* sont les clients les moins prioritaires. La priorité entre classes de clients est stricte, ce qui veut dire que le manager du centre d'appels ne se permet pas de modifier la priorité des clients en cours de la journée et, donc, un client *A* reste prioritaire par rapport à un client *B* qui, lui même, reste prioritaire par rapport à un client *C*. Cette

priorité est non-préemptive. Ceci veut dire qu'un nouvel arrivant dépasse tous les clients de moindre priorité dans la file d'attente vers laquelle il est routé, et se place derrière tous les clients de priorité supérieure ou égale à la sienne [2]. La non-préemption vient du fait qu'une fois un client est entré en communication avec un conseiller, il ne peut plus être mis en attente à cause de la venue d'un client plus prioritaire[3].

Nous pouvons également voir dans la Figure 4.1 qu'une fois arrivé au système, l'appel du client est routé vers l'un des quatre sites formant le centre d'appels. S'il y a un conseiller libre à cet instant sur le site vers lequel l'appel est routé, alors le client est immédiatement satisfait. Dans le cas contraire, le client doit patienter dans la file d'attente appartenant au site en question. Et comme nous l'avons déjà expliqué, il est possible que, dans le cas où le client doit attendre, la position dans la file d'attente soit rétrogradée à cause de l'arrivée d'un appel plus prioritaire.

Nous supposons que l'arrivée totale des appels suit une loi de Poisson de paramètre λ. Les appels de la classe *i* sont, également, supposés suivre une loi de Poisson de paramètre λ_i (λ_A pour la classe *A*, λ_B pour la classe *B* et λ_C pour la classe *C*). Le nombre de CDCs présents sur le site 1 est égal à C_1 (C_2 pour le site 2, C_3 pour le site 3 et C_4 pour le site 4). Chaque conseiller traite les appels suivant une loi exponentielle de taux μ, indépendant de la classe du client à qui il répond. L'ensemble des paramètres est résumé à la suite.

λ : taux d'arrivée des appels toutes classes confondues

λ_A : taux d'arrivée des appels de classe A

λ_B : taux d'arrivée des appels de classe B

λ_C : taux d'arrivée des appels de classe C

C : nombre total de CDCs tous sites confondus

C_1 : nombre de conseillers sur le site 1

C_2 : nombre de conseillers sur le site 2

C_3 : nombre de conseillers sur le site 3

C_4 : nombre de conseillers sur le site 4

μ : taux de service

---

[2] Au sein d'une même classe de clients, la règle choisie est *fcfs* ou encore, *premier arrivé premier servi*

[3] Avec la préemption, c'est le contraire qui se passe

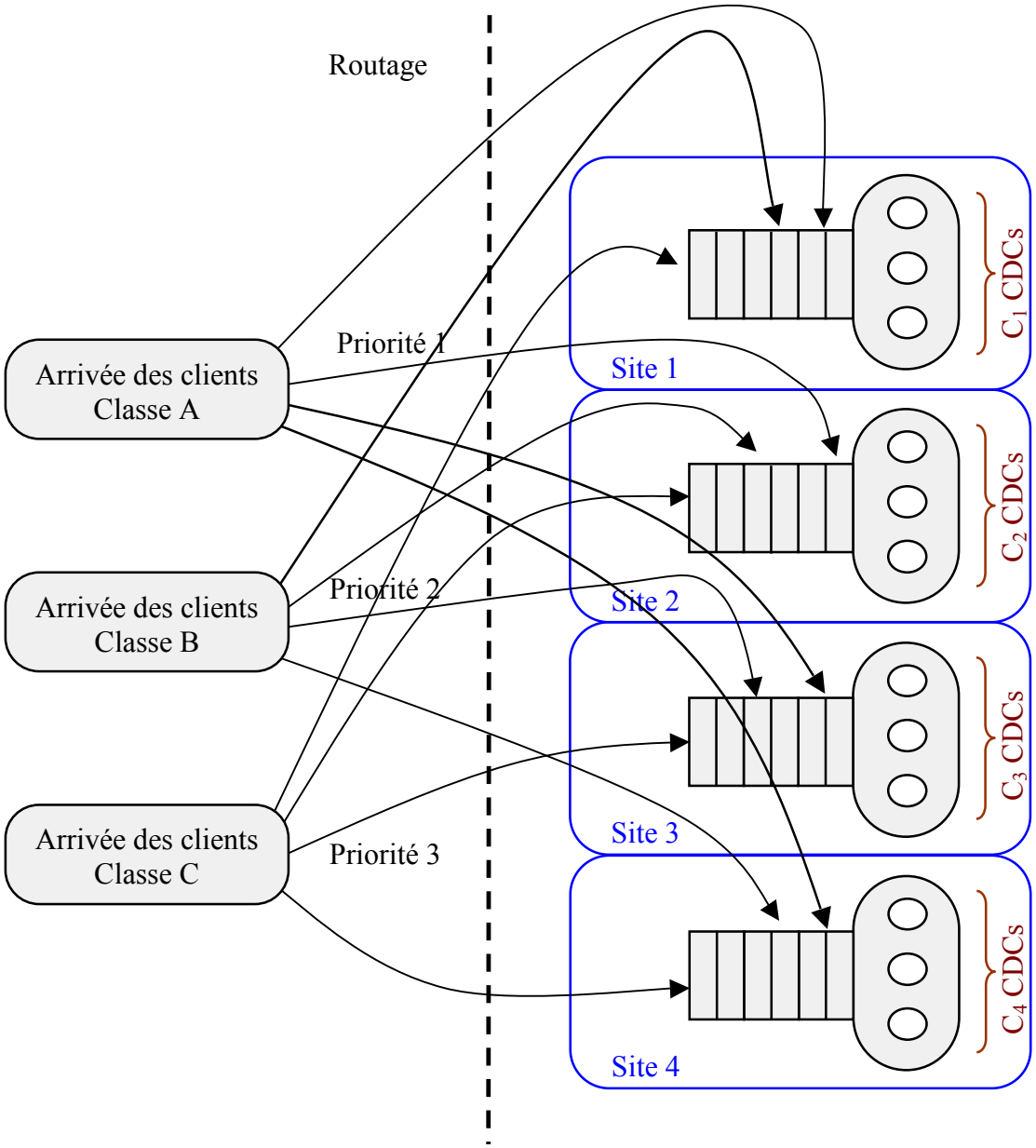


Figure 4.1: Système actuel

4.3.2 La file logique

La file logique, comme expliquée auparavant, consiste à fusionner toutes les files d'attente de telle sorte à ce qu'elles ne forment plus qu'une seule et unique file. Ceci revient à étudier un système M/M/C avec trois classes de clients et une priorité non-préemptive. La file logique est illustrée par la Figure 4.2. Dans cette figure nous

pouvons remarquer qu'il n'y a plus de routage puisque tous les clients passent par une seule et unique file.

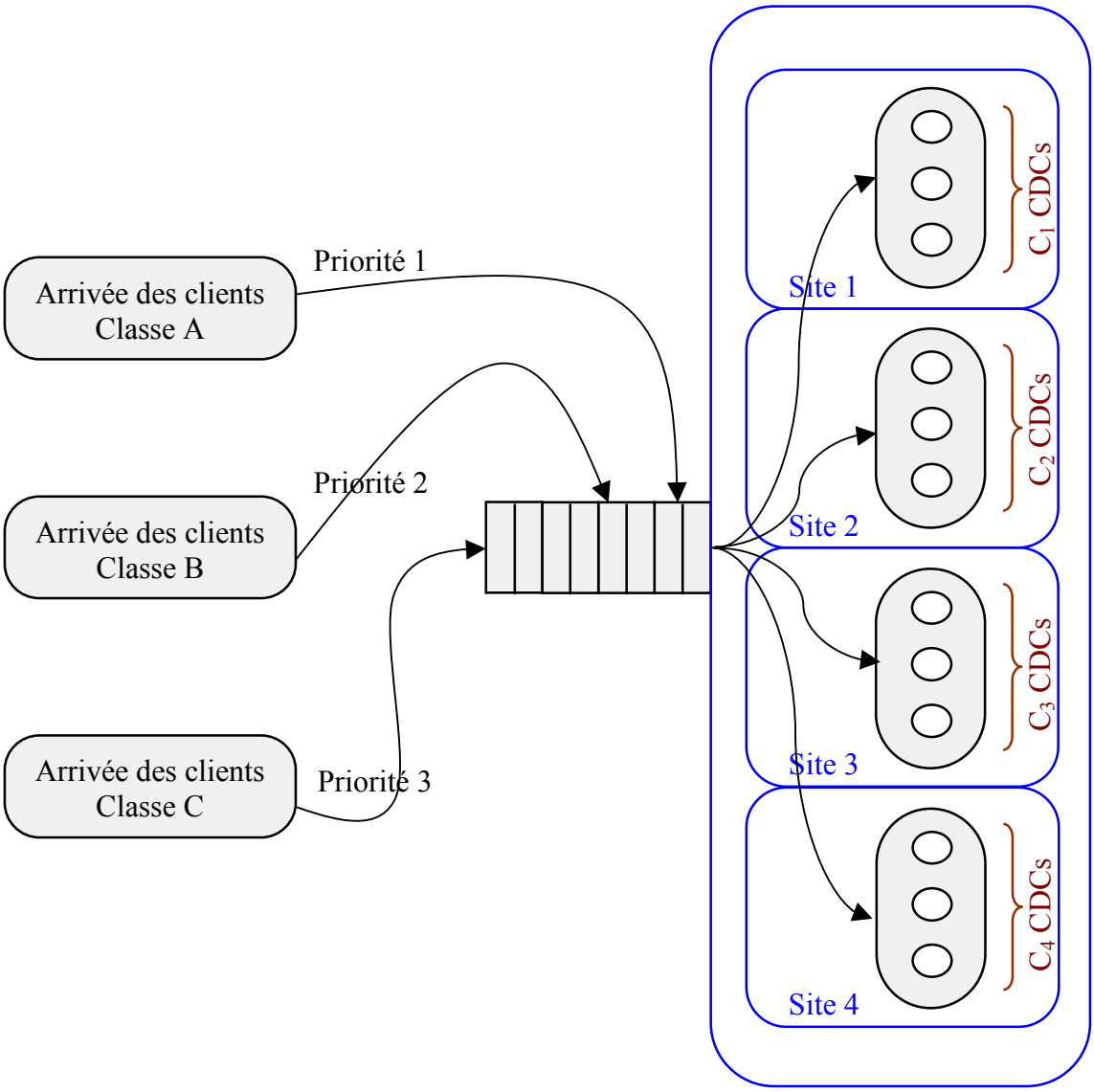


Figure 4.2: File logique

Nous allons dans le paragraphe suivant expliquer l'origine de l'intuition qui nous a poussés à étudier la file logique en vue d'un éventuel remplacement du système actuel.

4.3.2.1 Effets de la taille du système

Intuitivement, nous pouvons nous attendre à ce que la performance d'un système croît avec sa taille. Ceci vient du fait de la réduction de la variabilité avec la taille du

système comme nous l'avons vu dans les deux chapitres précédents où le système s'approche d'un comportement continu lorsque sa taille devient importante. Afin d'étudier l'effet de la taille du système sur sa performance, nous allons analyser le comportement du système en fonction du nombre de conseillers *C* qu'il comporte tout en gardant la charge du système $\rho = \lambda / C\,\mu$ constante. De cette manière, les trois systèmes de la Figure 4.3 possèdent la même charge.

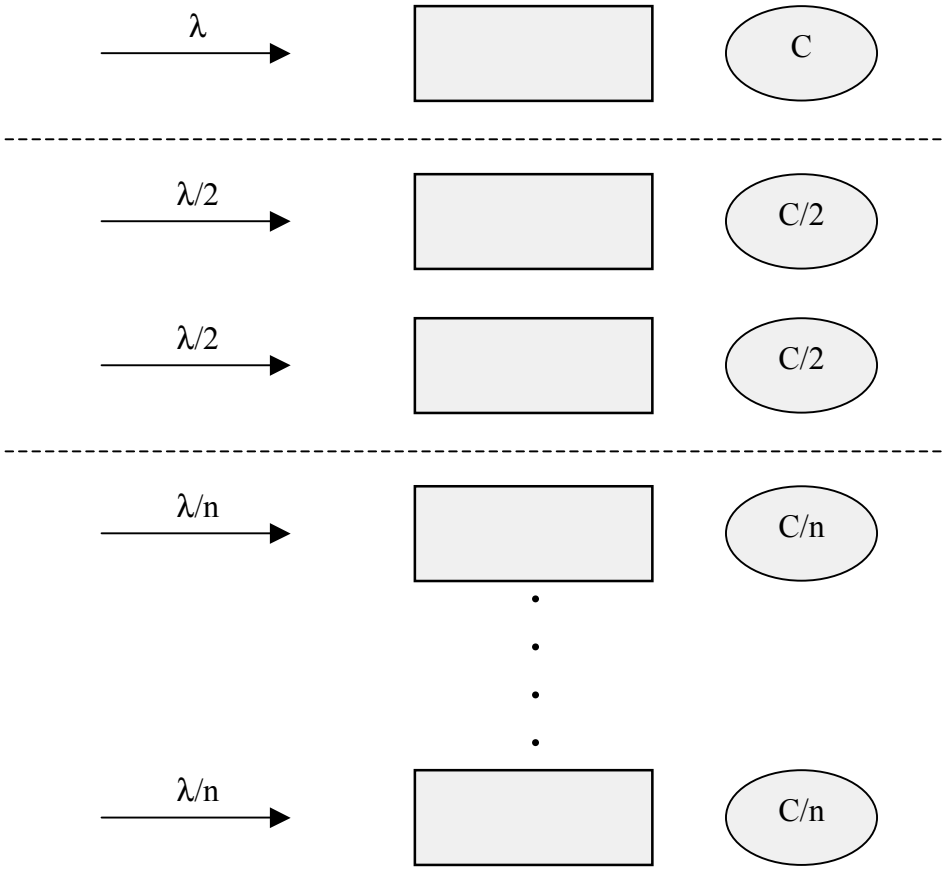


Figure 4.3: Trois systèmes à charge équivalente

Dans la Figure 4.3, nous avons illustré trois systèmes avec la même charge $\rho = \lambda / C\,\mu$. Chaque système se compose d'un ou de plusieurs sous-systèmes identiques et indépendants. De cette manière, le comportement du système global, pour chaque exemple, est exactement celui de n'importe lequel des sous-systèmes qui le composent. Nous allons donc considérer que la taille du système est égale à celle de chaque sous-système qu'il comporte. Si l'on regarde le premier système nous trouvons, ainsi, que c'est celui qui a la taille la plus importante. Nous allons comparer les systèmes suivant un seul critère de performance qui illustre bien, dans notre cas, le comportement du

système. Le critère considéré est le temps moyen d'attente. Ce dernier, s'écrit (voir Gross et Harris [33] sous la forme :

$$W_q = \left(\frac{(\rho\ C)^C}{C!\ (C\ \mu)\ (1-\rho)^2} \right) p_0 \tag{4.1}$$

où p_0 désigne la probabilité d'avoir zéro clients dans le système. Elle s'écrit :

$$p_0 = \left(\sum_{n=0}^{C-1} \frac{(\rho\ C)^n}{n!} + \frac{(\rho\ C)^C}{C!(1-\rho)} \right)^{-1} \tag{4.2}$$

La Figure 4.4 montre l'évolution du temps moyen d'attente en fonction de la taille du centre d'appels. Dans cette figure, nous avons utilisé un temps moyen de service égal à 3 minutes (et donc μ = 0,33 services / minute) et une charge ρ = 95 %. Le nombre de sites *N* est tel que le nombre total de conseillers soit égal à 200. Comme nous pouvons le voir dans cette figure, le temps moyen d'attente dans le système décroît rapidement en fonction du nombre de serveurs par site. Ainsi, il passe de plus de 50 minutes lorsque le site comporte un seul serveur à uniquement 0,1 minutes lorsque celui-ci en comporte 200.

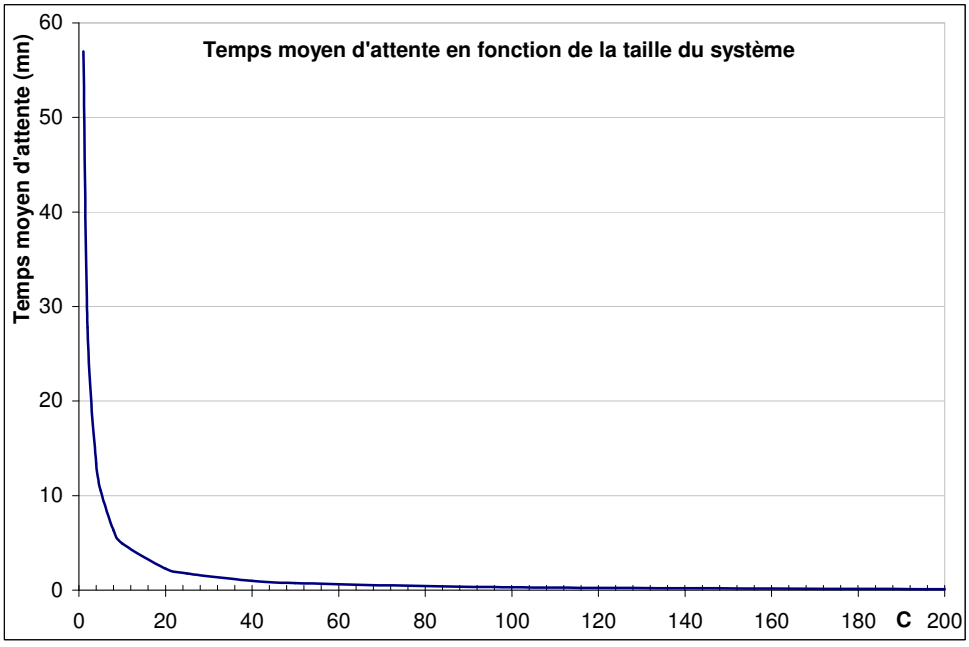


Figure 4.4: Temps moyen d'attente en fonction du nombre de serveurs par site

L'exemple précédent nous montre l'importance de la taille du système sur sa performance. Plus le système est grand, plus il neutralise les effets dus à la variabilité, ce qui le rend plus performant.

4.3.2.2 Effets du routage

Nous pouvons déjà avoir une première idée des conséquences du routage à partir de la Figure 4.4. En effet, nous pouvons interpréter cette figure comme une décomposition du système en plusieurs sous-systèmes semblables avec un routage statique des clients. Prenons le cas de figure où le système initial de 200 conseillers est composé de deux sous-systèmes recevant, chacun, la moitié du total des arrivées. Ceci revient à dire que lors de l'arrivée d'un client, son appel est routé avec une probabilité statique 0,5 au premier site et avec une probabilité de 0,5, également, au deuxième site. Il est possible qu'un conseiller se libère sur un autre site dans lequel il n'y a aucun client en attente alors que sur d'autres sites, des clients attendent toujours d'être servis. Ceci n'est pas possible lorsque tous les clients passent par une seule et unique file d'attente. Notons, tout de même, que la règle du routage statique est loin d'être la plus performante puisqu'elle n'exploite aucune information sur l'état du système au moment du routage de l'appel. En effet, il est possible, avec cette règle, de router un appel vers un site particulier pour qu'il y patiente un moment alors que sur d'autres sites son attente aurait été nulle. Avec une règle de routage dynamique, le routage exploite des informations sur l'état du système. Parmi ces informations nous pouvons citer le nombre de clients en attente dans chaque file ou, même, le temps de service restant dans chaque site pour des systèmes particuliers.

4.4 Estimation du temps d'attente

Dans cette section, nous étudions différents estimateurs du temps d'attente de chaque client à son arrivée au système. Ces estimateurs doivent exploiter les informations disponibles au moment de l'arrivée d'un client afin d'être utilisables. Dans la suite, nous allons distinguer les cas de la file logique et du système actuel pour tester les différents estimateurs. Nous allons, donc, commencer par définir ces estimateurs.

4.4.1 Les estimateurs du temps d'attente

– L'ASA (Average Speed Approximation): c'est l'estimateur proposé par plusieurs logiciels standards pour les centres d'appels. Il se calcule à partir de deux paramètres. Le premier paramètre représente le nombre de clients en attente et qui appartiennent à la même classe que le client qui vient d'appeler. Le deuxième paramètre est le débit moyen d'appels parvenus aux conseillers pendant les dix dernières minutes, suivant la classe de client.

La Figure 4.5 montre une variante de la Figure 4.2 pour représenter la file logique. En effet, dans la Figure 4.2 tous les clients passent par la même file d'attente indépendamment de leurs classes. Toutefois, tout nouveau client dépasse les clients déjà dans la file et qui sont moins prioritaires que lui. Dans la Figure 4.5, chaque classe possède sa propre file d'attente qui fonctionne suivant la politique fcfs (premier – arrivé – premier – servi). Lorsqu'un serveur se libère, il regarde s'il y a des clients en attente en commençant par la première file et en terminant par la troisième. Ceci nous conduit au même système pour les deux figures. Dans la Figure 4.5, X_A, X_B et X_C représentent les débits moyens d'attribution des appels des classes *A*, *B* et *C* aux conseillers pendant les dix dernières minutes. n_A, n_B et n_C représentent, respectivement, le nombre de clients de classes *A*, *B* et *C* en attente dans le système. Avec ces paramètres, l'ASA s'exprime avec la formule suivante :

$$ASA_{Classe} = \frac{n_{Classe}}{X_{Classe}} \tag{4.3}$$

où l'indice *Classe* représente la classe à laquelle appartient le nouveau client à qui le temps d'attente est estimé par ASA_{Classe} (ce paramètre admet, donc, comme valeurs possibles *A*, *B* ou *C*).

Pour le système actuel (avec plusieurs files d'attente et avec routage), à l'arrivée d'un client, une estimation de son temps d'attente est effectuée pour chaque site suivant la formule (4.3), et dans laquelle n_{Classe} et X_{Classe} deviendraient relatifs au site considéré. En fait, chaque site peut être représenté par la Figure 4.5. La plus grande différence vient de l'existence d'un routage en amont.

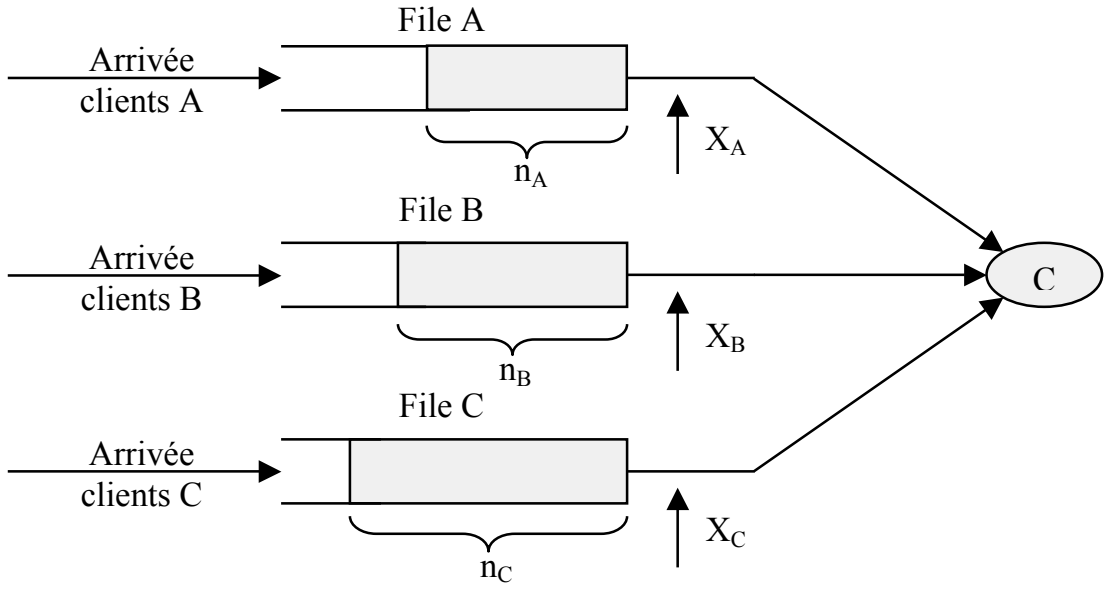


Figure 4.5: File logique

– L'ASA' (formule améliorée de l'ASA): le numérateur dans la formule (4.3) de l'ASA correspond au nombre de clients présents dans la file d'attente. Avec ce raisonnement, lorsqu'un nouveau client arrive et qu'il y a *n* clients dans la file, alors il devra patienter le temps que ces *n* clients soient servis. Or, après *n* services, le nouveau client passe en première position dans la file et il devra encore attendre qu'un conseiller se libère pour le servir. En tout, il doit, donc, patienter jusqu'à ce que *n + 1* clients soient servis avant d'être répondu par un conseiller. Ceci nous a amenés à modifier, dans un premier temps, la formule de l'ASA sous la forme :

$$ASA'_{Classe} = \frac{n_{Classe} + 1}{X_{Classe}} \tag{4.4}$$

– L'estimateur théorique: dans un système de file d'attente M/M/C avec plusieurs classes de clients et une priorité non-préemptive (comme cela a été expliqué pour la file logique), un client de classe *A* qui arrive en trouvant n_A clients, de la même classe que lui, en attente, doit patienter pendant le service de $n_A + 1$ clients. Ceci revient à attendre pendant un temps distribué suivant une loi d'Erlang de type $n_A + 1$ et de taux $C\,\mu$. Le temps moyen d'attente d'un tel client est alors :

$$W_A(n_A) = \frac{n_A + 1}{C\ \mu} \tag{4.5}$$

Concernant la classe *B*, supposons qu'un client rejoint le système de file logique en première position. Ce client doit, donc, attendre que tous les clients de classe *A* qui vont le dépasser soient répondus, avant d'être servi à son tour. La proposition suivante donne le temps moyen d'attente du client considéré.

Proposition 1 *Le temps moyen d'attente d'un client de classe B qui arrive en première position de la file logique, et trouve donc 0 clients de classe A et 0 clients de classe B en attente, est le suivant :*

$$W_B(0,0) = \frac{1}{C\ \mu - \lambda_A} \tag{4.6}$$

Preuve : Pour calculer le temps moyen d'attente du client de classe *B* qui arrive en première position dans la file, nous allons faire une analogie avec un autre système. Il s'agit d'une file semblable M/M/C qui diffère de la file logique, uniquement, par le fait qu'elle ne reçoit que les appels de la classe *A* et ce, avec le même taux d'appel λ_A. Pour ce système, une période d'occupation (Busy Period) est définie par le délai entre le début de l'occupation de l'ensemble des *C* serveurs, et la libération du premier serveur. Puisque les temps de service sont exponentiels, en utilisant la propriété qui stipule que la loi exponentielle n'a pas de mémoire, nous pouvons affirmer qu'à tout moment où les *C* serveurs sont occupés et que la file d'attente est vide, le délai nécessaire pour obtenir pour la première fois un serveur libre, est égale à une période d'occupation. Cette période d'occupation est exactement le temps qu'il faut au client *B* qui arrive en première position de la file logique pour qu'il soit répondu. Ce délai a été analysé par Kleinrock [48] pour une file M/G/1. Pour ce système, la transformée de Laplace relative à la période d'occupation est donnée par :

$$G^*(s) = B^*\left(s + \lambda_A - \lambda_A\ G^*(s)\right) \tag{4.7}$$

Dans cette dernière équation, B^* représente la transformée de Laplace du temps de service. Nous pouvons dire que cette équation est valable pour une file M/M/C puisque, du moment où la période d'occupation a commencé, la durée va être déterminée, uniquement, par la loi d'arrivée, qui reste inchangée, et par la loi de service qui a sa propre transformée de Laplace. Nous observons, également, que

l'équation est récurrente. Toutefois, nous pouvons, toujours, calculer les moments d'ordre k, $1 \leq k$, en utilisant la formule suivante :

$$g_k = (-1)^k \; G^{*(k)}(0) \tag{4.8}$$

À partir de cette dernière formule, il suffit de fixer $k = 1$ pour obtenir, grâce à l'équation (4.7) le résultat suivant :

$$g_1 = \frac{b_1}{1-\rho} \tag{4.9}$$

avec b_1, le moment d'ordre 1 du temps de service et $\rho = \lambda_A / C\mu$. b_1 est égal à la capacité moyenne de service, $C\,\mu$. Le résultat de la proposition en découle.

Pour un client de classe *B* qui trouve en arrivant, n_A clients, de classe *A* et n_B clients de classe *B*, le temps d'attente sera l'équivalent de $(n_A + n_B + 1)$ périodes d'occupation de classe *A*. Son temps moyen d'attente est, donc, donné par la formule qui suit :

$$W_B(n_A, n_B) = \frac{n_A + n_B + 1}{C\,\mu - \lambda_A} \tag{4.10}$$

Pour un client de classe *C*, le raisonnement est similaire. Maintenant, deux classes sont prioritaires. Toutefois, nous pouvons les considérer comme étant une seule classe puisque le taux d'arrivée de chaque classe suit une loi de Poisson. Nous obtenons de cette manière la formule suivante :

$$W_C(n_A, n_B, n_C) = \frac{n_A + n_B + n_C + 1}{C\,\mu - \lambda_A - \lambda_B} \tag{4.11}$$

En ce qui concerne le système actuel, il n'est pas possible d'avoir une expression analytique du temps d'attente pour les clients de classes *B* et *C* à cause des arrivées prioritaires qui ne suivent plus une loi de Poisson après le routage. Cependant, les clients de classe *A* vont garder la même expression de leur temps d'attente moyen en fonction de leur position dans la file puisqu'il n'y a pas de clients qui vont les dépasser.

– L'ASA²: c'est l'estimateur que nous proposons pour l'approximation du temps d'attente dans le centre d'appels. En fait, il n'est pas possible actuellement d'utiliser l'estimateur théorique à cause de la non disponibilité, en temps réel, du nombre de conseillers connectés à leurs postes. Ainsi, il est important d'utiliser un estimateur qui exploite uniquement les informations disponibles. Ces informations ont déjà été utilisées par l'ASA. Il s'agit du nombre de clients en attente par classe en plus des taux moyens d'attribution des appels aux conseillers sur les dix dernières minutes.

Dans cette estimation, nous allons nous baser sur les formules exactes du temps d'attente présentées précédemment. Dans ces formules, les taux d'arrivée λ_A et λ_B peuvent être remplacés par les débits moyens sur les dix dernières minutes X_A et X_B. Plus ces débits sont mesurés sur une période importante, plus ils s'approchent des valeurs exactes des taux λ_A et λ_B. En réalité, les taux d'arrivée varient au-cours de la journée. Si la fenêtre de temps, sur laquelle les débits moyens X_A et X_B sont observés, est très grande, alors les débits mesurés peuvent être éloignés des taux moyens relatifs au moment de la mesure. Si, par contre, la fenêtre de temps est très courte, alors les débits moyens peuvent enregistrer une forte variabilité. Les expérimentations, effectuées dans le cadre des projets menés avec *Bouygues Telecom*, montrent qu'une fenêtre de dix minutes est acceptable. Nous allons également approcher $C\mu$ par le débit moyen X. Ceci revient à dire que la capacité de service est égale au taux d'arrivée, et que la précision de cette approximation augmente avec la charge ρ du système. En remarquant que $X = X_A + X_B + X_C$, nous obtenons l'estimateur ASA² à partir des formules suivantes :

$$ASA_A^2(n_A) = \frac{n_A + 1}{X_A + X_B + X_C} \tag{4.12}$$

$$ASA_A^2(n_A, n_B) = \frac{n_A + n_B + 1}{X_B + X_C} \tag{4.13}$$

$$ASA_A^2(n_A, n_B, n_C) = \frac{n_A + n_B + n_C + 1}{X_C} \tag{4.14}$$

Les formules de l'ASA² peuvent être étendues, comme l'ASA et l'ASA', à une utilisation dans le système actuel et ce, en calculant les débits d'attribution des appels sur chaque site.

4.4.2 Validation de l'ASA^2 dans le cas de la file logique

Nous n'allons pas exposer le résultat de notre investigation de la qualité de l'ASA ni de l'ASA'. Toutefois, il faut dire que l'ASA donne des erreurs importantes puisque, même lorsque tous les serveurs sont occupés, il peut estimer le temps d'attente à zéro. Avec l'ASA', le gain en précision est très important. Nous allons maintenant présenter le résultat de l'analyse de la précision de l'ASA^2. Cette analyse est effectuée avec le logiciel de simulation *ARENA*. Lors de cette analyse, nous avons considéré les paramètres suivants :

- taux de service : $\mu = 0{,}25$ appels / minute
- Nombre total de conseillers : $C = 200$
- proportions d'appels reçus de la classe A : 43 %
- proportions d'appels reçus de la classe B : 17 %
- proportions d'appels reçus de la classe C : 40 %

Avec ces paramètres, nous faisons varier la charge ρ du système en changeant la valeur du taux d'appels λ. Le Tableau 4.1 et le Tableau 4.2 montrent les résultats issus de simulations sur des durées importantes. Nous y avons comparé le temps d'attente réel avec l'estimation donnée par l'ASA (l'unité de temps étant la seconde) suivant la position d'arrivée dans la file. Le Tableau 4.1 est relatif à une charge $\rho = 99$ %, et le Tableau 4.2 correspond à une charge de 85 %.

Classe	Position	1	2	3
A	Temps réel	1,20	2,40	3,60
	ASA^2	1,21	2,42	3,64
	Écart	0,83 %	0,83 %	1,11 %
B	Temps réel	2,09	4,18	6,26
	ASA^2	2,13	4,28	6,45
	Écart	1,91 %	2,39 %	3,04 %
C	Temps réel	2,96	5,92	8,85
	ASA^2	2,93	5,85	8,79
	Écart	-1,01 %	-1,18 %	-0,68 %

Tableau 4.1: Validation de l'ASA^2 pour $\rho = 99$ %

Classe	Position	1	2	3
A	Temps réel	1,200	2,404	3,599
	ASA[2]	1,329	2,662	3,999
	Écart	10,75 %	10,73 %	11,11 %
B	Temps réel	1,890	3,770	5,700
	ASA[2]	2,346	4,709	7,089
	Écart	24,13 %	24,91 %	24,37 %
C	Temps réel	2,442	4,914	7,389
	ASA[2]	3,329	6,667	10,021
	Écart	36,32 %	35,67 %	35,62 %

Tableau 4.2: Validation de l'ASA[2] pour ρ = 85 %

Dans le Tableau 4.1, nous pouvons constater que l'erreur de l'estimation reste limitée pour toutes les classes de clients. Le Tableau 4.2 montre, en revanche, que l'erreur devient importante et que l'ASA[2] surestime, toujours, le temps d'attente réel. Ceci était prévisible puisque, comme nous l'avons mentionné, la précision augmente en fonction de la charge du système. En effet, plus la charge ρ diminue, plus la différence entre la capacité de service $C\,\mu$ et le débit X devient importante. Le débit étant inférieur à la capacité de service réelle, ceci nous donne une surestimation du temps d'attente, préférable à une sous-estimation pour une éventuelle annonce aux clients.

Il faut signaler que les erreurs de l'estimateur n'ont pas une grande importance dans la pratique. En fait, pour les paramètres mentionnés dans cette section, et pour une charge ρ = 85 %, la probabilité d'attente est égale à 1,53 % [4]. En d'autres termes, l'erreur lors de l'estimation est assez grande pour une charge faible mais, en même temps, la proportion de clients pour laquelle l'erreur est significative, reste limitée.

Dans le Tableau 4.3, nous avons affiché le coefficient de variation de l'ASA[2] suivant la classe et la position dans la file et ce, pour une charge ρ = 99 %. À partir de ce tableau, nous pouvons déduire que l'estimateur prédit des temps très proches des moyennes affichées par le Tableau 4.1.

[4] Elle est définie par $P_d = \left(1 + \frac{C!(1-\rho)}{(\rho\,C)^C} \sum_{n=0}^{C-1} \frac{(\rho\,C)^n}{n!}\right)^{-1}$. Voir Gross et Harris [33]

Position	1	2	3
A	0,002	0,002	0,002
B	0,008	0,008	0,008
C	0,008	0,008	0,008

Tableau 4.3: Coefficient de variation CV^2 de l'ASA^2 pour ρ = 99 %

Il faut remarquer que le Tableau 4.3 n'affiche pas le coefficient de variation du temps d'attente réel, mais plutôt le coefficient de variation de l'estimateur proposé. La faiblesse des valeurs obtenues nous montre que la prédiction est presque toujours la même pour une position et une classe données. Pour étudier la variabilité exacte du temps d'attente en fonction de la position lors de l'arrivée, nous pouvons, à nouveau, exploiter l'étude de la période d'occupation pour aboutir aux valeurs suivantes des variances suivant le nombre de clients de chaque classe qui sont déjà en attente au moment de l'arrivée. Comme lors de l'étude des temps moyens, en utilisant les équations (4.7) et (4.8), nous pouvons aboutir aux variances des temps d'attente :

$$\sigma_A^2 = \frac{n_A + 1}{\left(C.\mu\right)^2} \tag{4.15}$$

$$\sigma_B^2 = \left(n_A + n_B + 1\right)\frac{C.\mu + \lambda_A}{\left(C.\mu - \lambda_A\right)^3} \tag{4.16}$$

$$\sigma_C^2 = \left(n_A + n_B + n_C + 1\right)\frac{C.\mu + \lambda_A + \lambda_B}{\left(C.\mu - \lambda_A - \lambda_B\right)^3} \tag{4.17}$$

Si le calcul de la variance de l'attente d'un client de classe *A* est facile, puisqu'il suit une loi d'Erlang de type $n_A + 1$ et de taux $C\,\mu$, pour la classe *B*, il faut utiliser les équations (4.7) et (4.8) pour avoir le moment d'ordre 2 pour un client qui arrive en 1ère position. Ceci nous permet de calculer la variance relative à cette position. Par la suite, il suffit de multiplier la variance par la position réelle pour aboutir à la formule (4.16). Pour la classe *C*, il suffit de voir les classes *A* et *B* comme étant une seule classe prioritaire.

L'étude des variances sans la moindre information sur les moyennes n'apporte pas d'informations pertinentes quant à la loi du temps d'attente. Il est, ainsi, préférable de regarder le coefficient de variation qui est le rapport entre l'écart-type et la moyenne. À partir des trois formules précédentes ainsi que des formules (4.5), (4.10) et (4.11), nous déterminons les coefficients de variation relatifs aux trois classes de clients :

$$CV_A^2 = \frac{1}{n_A + 1} \tag{4.18}$$

$$CV_B^2 = \left(\frac{1}{n_A + n_B + 1} \right) \frac{C.\mu + \lambda_A}{C.\mu - \lambda_A} \tag{4.19}$$

$$CV_C^2 = \left(\frac{1}{n_A + n_B + n_C + 1} \right) \frac{C.\mu + \lambda_A + \lambda_B}{C.\mu - \lambda_A - \lambda_B} \tag{4.20}$$

Ces trois dernières équations nous permettent d'affirmer que, plus la position du client dans la file d'attente est éloignée au moment de son arrivée, plus la moyenne du temps d'attente donne une estimation précise de l'attente réelle. Il n'est pas important de voir, à partir des équations (4.15) à (4.17) que la variance augmente en fonction du nombre de clients déjà en attente. En effet, c'est la variabilité relative donnée par le coefficient de variation qui constitue l'information la plus pertinente.

4.4.3 Validation de l'ASA^2 dans le cas du système actuel

Nous étudions maintenant la précision de l'ASA^2 pour le système de file logique *déséquilibré*. La répartition des conseillers de clientèle entre les sites est la suivante :

- C_1 = 80 serveurs dans le premier site
- C_2 = 50 serveurs dans le premier site
- C_3 = 40 serveurs dans le premier site
- C_4 = 30 serveurs dans le premier site

Le Tableau 4.4 et le Tableau 4.5 montrent les résultats des simulations du système actuel en fonction de la charge du système. Les temps d'attente indiqués sont exprimés en secondes.

		Position	1	2	3
Site 1	A	Temps réel	3,04	6,01	8,98
		ASA²	3,04	6,01	8,98
		Écart	0 %	0 %	0 %
	B	Temps réel	7,48	13,70	20,09
		ASA²	7,32	14,74	21,33
		Écart	-2,14 %	7,59 %	6,17 %
	C	Temps réel	9,02	18,92	29,53
		ASA²	9,25	19,23	29,55
		Écart	2,55 %	1,64 %	0,07 %
Site 2	A	Temps réel	4,80	9,58	14,36
		ASA²	4,84	9,47	14,05
		Écart	0,83 %	-1,15 %	-2,16 %
	B	Temps réel	9,23	17,12	26,02
		ASA²	8,68	17,60	26,15
		Écart	-5,96 %	2,80 %	0,50 %
	C	Temps réel	10,50	22,13	34,22
		ASA²	11,84	24,01	36,16
		Écart	12,76 %	8,50 %	5,67 %
Site 3	A	Temps réel	6,00	11,98	17,78
		ASA²	5,96	11,76	17,01
		Écart	-0,67 %	-1,84 %	-4,33 %
	B	Temps réel	8,51	17,37	26,85
		ASA²	8,36	16,72	25,02
		Écart	-1,76 %	-3,74 %	-6,82 %
	C	Temps réel	10,75	23,05	35,64
		ASA²	13,34	26,47	39,57
		Écart	24,09 %	14,84 %	11,03 %
Site 4	A	Temps réel	8,01	16,08	24,69
		ASA²	7,91	14,88	22,12
		Écart	-1,25 %	-7,46 %	-10,41 %
	B	Temps réel	9,39	19,73	30,22
		ASA²	9,25	18,04	27,09
		Écart	-1,49 %	-8,57 %	-10,36 %
	C	Temps réel	11,38	24,39	37,63
		ASA²	15,51	29,59	43,27
		Écart	36,29 %	21,32 %	14,99 %

Tableau 4.4: Validation de l'ASA² pour ρ = 99 %

		Position	1	2	3
Site 1	A	Temps réel	3,008	6,015	9,137
		ASA²	3,341	6,591	9,841
		Écart	11,07 %	9,58 %	7,70 %
	B	Temps réel	5,603	11,003	16,659
		ASA²	6,145	11,967	18,089
		Écart	9,67 %	8,76 %	8,58 %
	C	Temps réel	6,836	14,409	22,984
		ASA²	8,749	17,351	26,785
		Écart	27,98 %	20,42 %	16,54 %
Site 2	A	Temps réel	4,795	9,650	14,664
		ASA²	5,310	10,298	15,203
		Écart	10,74 %	6,72 %	3,68 %
	B	Temps réel	7,651	14,824	23,552
		ASA²	9,321	18,356	27,033
		Écart	21,83 %	23,83 %	14,78 %
	C	Temps réel	8,680	18,706	28,954
		ASA²	13,387	26,842	40,370
		Écart	54,23 %	43,49 %	39,43 %
Site 3	A	Temps réel	5,939	11,676	18,993
		ASA²	6,483	12,679	17,859
		Écart	9,16 %	8,59 %	-5,97 %
	B	Temps réel	7,933	16,533	25,401
		ASA²	10,985	21,377	30,416
		Écart	38,47 %	29,30 %	19,74 %
	C	Temps réel	8,748	18,680	29,421
		ASA²	16,114	31,644	46,490
		Écart	84,20 %	69,40 %	58,02 %
Site 4	A	Temps réel	8,092	14,755	24,645
		ASA²	8,631	16,465	24,265
		Écart	6,66 %	11,59 %	-1,54 %
	B	Temps réel	9,456	18,404	32,350
		ASA²	13,804	24,528	35,870
		Écart	45,98 %	33,28 %	10,88 %
	C	Temps réel	9,608	20,491	31,388
		ASA²	20,642	38,652	55,247
		Écart	114,84 %	88,63 %	76,01 %

Tableau 4.5: Validation de l'ASA² pour ρ = 85 %

À partir des deux tableaux précédents, nous constatons que la précision de l'estimateur ASA^2 se dégrade significativement par rapport à ce qu'elle était pour la file logique. Désormais, cette précision dépend du site pour lequel l'estimation est effectuée. Et comme le montrent les tableaux, la précision de l'estimateur est meilleure pour le premier site. Ses plus mauvais résultats sont issus du dernier site. Ceci est, essentiellement, dû à la différence de taille entre les sites. Ainsi, la taille du premier site lui procure une meilleure "absorption" de la variabilité des arrivées alors que la taille réduite du quatrième site expose clairement ses performances à cette variabilité. Nous remarquons également que la précision de l'approximation pour la classe *A* reste convaincante alors qu'elle se dégrade de façon importante pour la classe C.

Une fois encore, la précision de l'estimation du temps d'attente croît avec la charge du système. Ainsi, même l'estimation de l'attente de la classe C reste raisonnable pour une charge de 99 %. Signalons finalement que la règle de routage influe également sur l'estimation du temps d'attente parce que c'est bien elle qui décide des arrivées sur chaque site. La variabilité des arrivées à chaque site dépend, donc, en grande partie du routage utilisé. Si, par exemple, ce routage ne partage pas équitablement les appels entre les sites, alors la variabilité des arrivées qui en résultent va être importante et elle aura un effet particulièrement important sur les sites dont la petite taille n'arrive pas à réduire l'effet de cette variabilité.

Dans le Tableau 4.6, nous avons affiché l'évolution du coefficient de variation pour une charge $\rho = 99\ \%$. Nous constatons ainsi que, comme pour la file logique, l'estimation du temps d'attente ne se disperse pas autour de sa moyenne et qu'elle ressemble de ce fait à une estimation constante lorsqu'il s'agit de la même position et de la même classe de clients. À noter, finalement, qu'une charge égale à 85 % nous donne les mêmes résultats en ce qui concerne le coefficient de variation.

	Position	1	2	3
Site 1	A	0,005	0,005	0,005
	B	0,023	0,027	0,030
	C	0,029	0,034	0,034
Site 2	A	0,008	0,008	0,007
	B	0,023	0,021	0,021
	C	0,028	0,030	0,031
Site 3	A	0,010	0,009	0,009
	B	0,020	0,019	0,018
	C	0,034	0,033	0,033
Site 4	A	0,012	0,011	0,013
	B	0,025	0,019	0,017
	C	0,061	0,052	0,047

Tableau 4.6: Coefficient de variation CV2 de l'ASA^2 pour $\rho = 99$ %

Dans le prochain chapitre, les deux estimateurs du temps d'attente ASA et ASA^2 vont être exploités dans le routage des clients aux différents sites du système actuel. Ainsi, lors de l'arrivée d'un nouvel appel, le système effectue une estimation de son temps d'attente sur chaque site par l'intermédiaire des deux estimateurs mentionnés. L'appel sera routé vers le site offrant l'estimation du temps d'attente la moins importante.

4.5 Conclusions

Nous avons traité, dans ce chapitre, l'estimation du temps d'attente des clients à leurs arrivées au centre d'appels. Nous avons, également, analysé l'apport de combiner les files d'attente conduisant aux sites pour qu'elles n'en forment qu'une seule, appelée file logique.

Lors de l'analyse des estimateurs du temps d'attente, nous avons montré que, plus la charge du système est élevée, plus la précision de l'estimateur ASA^2 est importante. Toutefois, pour des charges faibles, il faut noter que même si la précision n'est pas importante, le nombre de clients concernés est limité puisque la majeure partie accède directement au service sans avoir à attendre. Nous pouvons citer, ici, que l'un des avantages de la file logique est que la précision de l'estimateur y est beaucoup plus

importante que lorsqu'il est exploité dans un système à plusieurs files d'attente. Ceci est particulièrement intéressant pour une bonne analyse du temps d'attente que les clients doivent passer dans la file avant d'être servis. En effet, l'information disponible, dans ce cas, est beaucoup plus pertinente et donne une meilleure idée de l'attente réelle des clients, ce qui pourrait permettre une amélioration du service fourni en vue de diminuer le temps d'attente des clients. Notons, également, que la précision de l'estimation du temps d'attente, dans un système composé de plusieurs files d'attente, peut permettre un meilleur routage des appels des clients. Ainsi, plus le système dispose d'informations au moment de l'arrivée des clients, plus le routage s'améliore. Une étude du routage sera menée dans le prochain chapitre.

Outre la précision de l'ASA², nous avons montré que, pour une classe fixée et pour une position donnée lors de l'arrivée, il varie très peu autour de sa moyenne. Ceci rappelle l'estimateur théorique du temps d'attente qui, lui, est constant.

Il serait intéressant de trouver une méthode de détermination de la charge du système. En effet, en disposant de la charge réelle du système, nous pouvons estimer la différence entre le débit mesuré X et la capacité de service réelle $C\mu$. Avec cette estimation, nous pouvons améliorer, nettement, l'ASA² dans le cas où la charge est très faible même si, rappelons-le, la proportion de clients qui attendent, dans ce cas, n'est pas importante.

Chapitre 5 : Affectation des clients dans un centre d'appels multi-site et multi-classe

5.1 Introduction

Dans ce chapitre, nous abordons le problème de routage des clients entre les files d'attente du centre d'appels. Ce travail fait suite au travail décrit dans le chapitre précédent. Nous y avons abordé le problème d'estimation du temps d'attente aux clients lors de leurs arrivées. Nous allons nous baser, dans ce chapitre, sur les estimateurs fournis précédemment et ce, dans le but de router les clients vers les sites qui leurs minimisent leurs temps d'attente.

Le système que nous allons considérer sera, comme lors du chapitre précédent, composé de plusieurs sites avec des nombres de serveurs variables suivant les sites. Chaque site possède sa propre file d'attente. Les clients qui appellent le système sont répartis en trois classes de priorité stricte. Les clients de classe A seraient les plus prioritaires. La classe B aurait la classe intermédiaire alors que la classe C représenterait la classe la moins prioritaire. La priorité dans le système est non-préemptive. À leurs arrivées, les clients sont routés vers l'un des sites afin d'y être servis. Le service peut s'effectuer immédiatement s'il y a un serveur disponible. S'il n'y en a aucun, alors le client doit attendre un moment avant d'être servi. Nous supposons

que les clients n'abandonnent pas et que, de cette façon, tous les clients ayant appelé finissent par être servis.

Nous supposons que les informations dont dispose le système en temps réel, sont les mêmes qui ont servis, lors du chapitre précédent, à l'élaboration de l'estimateur du temps d'attente ASA[2]. Ces informations se composent, donc, en deux catégories. La première catégorie constitue le débit moyen d'envoi des appels aux conseillers d'un site particulier et ce, pour chaque classe de clients. La deuxième catégorie représente, elle, le nombre de clients de chaque classe qui sont en attente sur un site.

Les routages que nous allons étudier dans ce chapitre seront du type dynamique, ce qui veut dire qu'ils exploitent les informations disponibles sur l'état du système en temps réel. Ici, ces informations appartiendront à l'une des deux catégories que nous avons définies.

Nous supposons que les managers du centre d'appels se fixent des objectifs de niveau de service pour chaque classe de clients. Chaque classe se voit attribuer son propre objectif de niveau de service. La mesure du niveau de service que nous allons regarder sera la proportion de clients servis après moins de 20 secondes d'attente. Ceci constitue une mesure très répandue dans les centres d'appels. Une autre mesure fréquemment utilisée correspond au temps moyen d'attente. L'origine des 20 secondes que nous fixons vient du système mono-classe où il est fréquent de fixer l'objectif de satisfaire 80 % des clients en moins de 20 secondes d'attente. Cela provient, probablement, de la fameuse règle des 80/20.

Les règles de routage seront, toutes, basées sur une minimisation du temps d'attente du client à son arrivée. Ainsi, lorsqu'un client arrive, une estimation de son temps d'attente est effectué pour chaque site. L'appel sera routé vers le site qui offre le temps d'attente le moins élevé. Une fois les objectifs fixés, nous pourrons comparer plusieurs règles de routage. La comparaison sera effectuée en termes de serveurs nécessaires à atteindre l'objectif visé. Étant donné la complexité de l'étude avec le routage dynamique, l'étude menée se base, essentiellement, sur la simulation à évènements discrets. En contraste avec la plupart des travaux dans la littérature, nous

nous préoccupons de la satisfaction d'un niveau de service objectif pour chaque classe de clients.

Pour un objectif donné, en plus de la comparaison des règles de routage entre elles, nous allons, également, les comparer avec le système de la *file logique* défini dans le chapitre précédent[5]. À première vue, la file logique semble être la plus performante. Nous allons voir que, suivant certains niveaux de service objectif, ce n'est pas toujours le cas. Cela nous aidera à continuer la comparaison entamée dans le chapitre précédent.

Après cette introduction, nous allons passer à une illustration de quelques travaux qui ont traité le routage.

Dans la troisième section, nous passons à l'étude de performance du *système actuel*[6] tel qu'il a été défini dans le chapitre précédent. Nous étudions différentes règles de routage et nous comparons le système avec la file logique.

La quatrième section expose nos conclusions sur l'étude du routage et termine l'étude de la performance de la file logique commencée au-cours du chapitre précédent.

5.2 Étude bibliographique

Nous allons, dans cette section, exposer quelques travaux ayant traité les modèles de files d'attente avec routage. Beaucoup de travaux ont abordé cette problématique. Le lecteur peut s'orienter vers Boxma, Koole et Liu [13] pour un état de l'art très détaillé sur le domaine. Les auteurs y illustrent, en particulier, les travaux ayant abordé les six modèles clefs qu'ils trouvent représentatifs des modèles traités lors de l'analyse de performance des systèmes parallèles et distribués. Les auteurs mettent l'accent sur le routage et sur la politique de la file la plus courte. Adan, Boxma et Resing [1] ont traité les modèles avec de multiples files d'attente. Ils les ont décomposés en deux classes.

[5] La file logique correspond à la fusion de toutes les files de telle sorte à ce que tous les clients passe par une seule file

[6] Le système actuel comporte quatre sites et nécessite une règle de routage

Dans la première classe, le client est routé vers une file d'attente pour accéder au service. Dans la deuxième classe c'est les serveurs qui choisissent le client à servir suivant une règle de priorité. Concernant la première classe de modèles, les auteurs se sont intéressés à deux disciplines avec une seule classe de clients. Il s'agit d'un routage statique et d'un routage vers la file la plus courte.

Beaucoup de travaux ont traité le routage vers la file la plus courte, connu pour sa simplicité et pour le fait qu'il optimise un système composé de deux files d'attente, chacune intégrant un serveur avec un temps de service exponentiel (Winston [72]). Whitt [68] a montré que, pour certaines distributions du temps de service, router vers la file où l'attente devrait être la plus courte n'optimise pas le système, de point de vue temps moyen d'attente au régime stationnaire, par exemple. Hordijk et Koole [36] ont traité l'optimalité du routage de la file la plus courte en l'étendant à des arrivées par lots et à des files d'attente de capacités finies.

Houck [37] a effectué une comparaison de plusieurs règles de routage dans le cas d'une seule classe de clients. Les temps de service sont exponentiels et les arrivées suivent une loi de Poisson. Les clients appartiennent à une seule et unique classe. L'auteur a donné, pour ce système, deux bornes intuitives du routage optimal. Pour un système composé de deux files, il montre, en particulier, que le routage vers la file qui minimise le temps d'attente est, quasiment, aussi performant que le système issu de la fusion des deux files d'attente. Ce dernier système offre une borne inférieure au nombre optimal de serveurs dans un système avec routage.

Boxma [14] a traité une règle de routage statique basée sur des probabilités avec lesquelles les clients sont routés vers une file en particulier. Il détermine les probabilités qui lui permettent d'optimiser sa fonction objectif qui peut être le nombre moyen de clients en attente. Bhulai et Koole [12] se sont intéressés à un système multi-serveur qui traite deux types d'appels. Ce sont, typiquement, des appels rentrants et des appels sortants. Pour les appels rentrants suivant une loi de Poisson, il y a une contrainte sur le temps d'attente qui doit être satisfaite. Les appels sortants, eux, sont supposés être illimités. L'objectif étant d'effectuer le plus grand nombre d'appels sortants tout en satisfaisant la contrainte sur le niveau de service des appels rentrants. Les auteurs ont montré, dans le cas où les deux types d'appels sont traités avec le

même taux de service exponentiel, que la discipline optimale correspond à ne pas effectuer d'appels sortants si le nombre de serveurs occupés dépasse un certain seuil. Gans et Zhou [28] ont traité le même problème. Ils ont montré que l'optimalité, lorsque les temps de service suivent la même loi pour les deux types d'appels, correspond à étendre la discipline de seuil, décrite précédemment, avec un deuxième seuil. Le seuil effectif est obtenu après un tirage aléatoire.

Masi, Fischer et Harris [54] ont étudié deux règles de routage entre deux groupes de serveurs. La première règle consiste à router l'appel du client à son arrivée vers l'un des sites. C'est ce que nous allons considérer dans ce chapitre. Leur deuxième règle, elle, route, dans un premier temps, l'appel à un site particulier lors de l'arrivée. Plus tard, s'il y a un serveur qui se libère sur l'autre site, une décision de transfert de l'appel vers ce site peut être prise s'il y a une ligne de connexion entre les deux sites qui n'est pas encore utilisée. Le nombre de lignes de connexion est limité. Plus ce nombre est important, plus le système s'approche de la file logique traitée dans le chapitre précédent. L'analyse numérique effectuée se base sur la méthode de matrice géométrique. Grâce à elle, les auteurs montrent que le système avec le routage retardé est plus performant puisqu'il y est moins fréquent de trouver un serveur libre sur un site alors que, sur l'autre site il y a des clients en attente.

La plupart des travaux cités, traitent une seule classe de clients. Et même lorsqu'ils en traitent plusieurs, ils ne s'intéressent qu'à atteindre un niveau de service objectif pour l'une d'entre elles. Dans ce chapitre, nous traitons le cas de trois classes de clients, ce qui change la comportement du système. En outre, nous fixons un objectif de niveau de service pour chaque classe. Notre but n'est pas de trouver la règle de routage optimale mais plutôt de discuter des différentes règles que nous allons introduire et de comparer leurs performances avec celles d'un système à une file unique. D'une manière générale, lorsqu'une discipline de routage est traitée, c'est l'étude des temps d'attente qui est visée en premier lieu. Il est rare d'avoir une analyse de la performance du système en termes de nombre de conseillers. Notre travail s'inscrit, justement dans cette voie. Ainsi, nous allons étudier la performance du système suivant le nombre de conseillers nécessaires à atteindre les objectifs fixés pour chaque classe de clients.

5.3 Routage des clients dans le centre d'appels

Dans cette section, nous allons nous intéresser à l'exploitation des estimateurs du temps d'attente, introduits dans le chapitre précédent, dans le routage des appels entre les différents sites du système actuel. Pour mesurer la performance d'une règle de routage, nous allons considérer des objectifs de niveau de service à atteindre. Il s'agit du pourcentage de clients de chaque classe répondus en moins de 20 secondes. Dans la suite, par la notation "$QoS_A|QoS_B|QoS_C$" nous allons désigner la proportion des clients de classe *A* répondus en moins de 20 secondes (QoS_A), la proportion des clients de classe *B* répondus en moins de 20 secondes (QoS_B) et proportion des clients de classe *C* répondus en moins de 20 secondes (QoS_C).

Les règles de routage analysées seront, toutes, basées sur la minimisation du temps d'attente donné par l'un des estimateurs. Ceci nous permet, dans le cas où l'estimation est exacte, de router l'appel au site sur lequel il va attendre pendant un temps minimum.

Grâce à un logiciel de simulation à évènements discrets, nous allons analyser plusieurs règles de routage basées sur les estimateurs du temps d'attente déjà définis. Pour chaque règle de routage, nous déterminons le nombre de serveurs nécessaires et ce, suivant la qualité de service objectif. Nous allons mesurer la performance des règles de routage en fonction du nombre optimal de serveurs, ce qui nous permet de comparer les règles de routage entre elles. Nous allons également comparer les règles de routage analysées avec un système de file logique (sans routage) pour lequel nous cherchons également le nombre minimum de serveurs qui permettent la satisfaction du niveau de service objectif. Cependant, pour la file logique nous n'allons pas devoir utiliser un logiciel de simulation puisque nous pouvons déterminer le nombre optimal par calcul numérique.

5.3.1 Évaluation de performances de la file logique

Nous allons dans cette section étudier les performances de la file logique en nous intéressant aux critères généralement retenus pour évaluer la performance d'un centre d'appels. Il s'agit, en fait, du temps moyen d'attente et de la proportion des appels pris

en moins de 20 secondes. Ces deux critères dépendent, bien sûr, de la classe de clients considérée. Rappelons que la file logique correspond à un système M/M/C multi-classe avec des priorités non-préemptives.

5.3.1.1 Temps moyens d'attente

Supposons que notre système reçoit les appels de clients répartis en n classes de priorités strictes. Les taux d'arrivée de la classe i, $1 \le i \le n$, est égal à λ_i. La classe la plus prioritaire étant la classe 1. Le temps moyen d'attente de la classe k, $1 \le k \le n$, est donné par l'expression (5.1) (voir Kella et Yechiali [45] pour les détails).

$$Wq_k = \frac{C\mu\, P_d}{\left(C\mu - \sum_{i=1}^{k} \lambda_i\right)\left(C\mu - \sum_{i=1}^{k-1} \lambda_i\right)} \tag{5.1}$$

où $\lambda_0 = 0$ et P_d représente la probabilité d'attendre dans une file M/M/C classique (sans priorités). P_d peut être calculée par la formule (5.2) où ρ représente la charge du système (Gross et Harris [33]).

$$P_d = \left(1 + \frac{C!(1-\rho)}{(\rho\, C)^C} \sum_{n=0}^{C-1} \frac{(\rho\, C)^n}{n!}\right)^{-1} \tag{5.2}$$

En tenant compte des trois classes de clients de la file logique, nous obtenons les formules suivantes pour les temps d'attente moyens dans la file :

$$Wq_A = \frac{P_d}{C\mu - \lambda_A} \tag{5.3}$$

$$Wq_B = \frac{C\mu\, P_d}{(C\mu - \lambda_A)(C\mu - \lambda_A - \lambda_B)} \tag{5.4}$$

$$Wq_C = \frac{C\mu\, P_d}{(C\mu - \lambda)(C\mu - \lambda_A - \lambda_B)} \tag{5.5}$$

Si l'objectif des niveaux de service était exprimé en temps moyens d'attente, alors les trois dernières formules formeraient le noyau d'une procédure numérique qui détermine le nombre optimal de serveurs. Étant donné la simplicité des formules, ce calcul devrait être facile à effectuer.

5.3.1.2 Proportions de clients répondus en moins de vingt secondes

Considérons le même système du paragraphe précédent (*n* classes de priorités avec différents taux d'arrivées). Davis [21] a déterminé la distribution des temps d'attente pour chaque classe. Pour la classe de priorité 1, la distribution du temps d'attente est assez connue puisqu'elle correspond au temps d'attente de la classe la plus prioritaire et que de ce fait, en supposant que l'attente est non nulle, alors elle ne dépend pas des autres classes. La distribution est donnée par :

$$Wq_1(t) = 1 - P_d \; e^{(\lambda_1 - C\mu)\,t} \tag{5.6}$$

avec P_d, la probabilité d'attente définie précédemment.

La fonction de répartition du temps d'attente d'un client de classe *k*, $1 < k \leq n$, est, par contre, beaucoup moins connue puisqu'elle est plus difficile à obtenir. Davis [21] a réussi, grâce à une intégration sur contour, à inverser la transformée de Laplace de la distribution du temps d'attente. La formule résultante, qui reste assez compliquée, est donnée par :

$$Wq_k(t) = \begin{cases} 1 - \dfrac{P_d(1-\rho_k)}{2\,\pi\,\rho_k} \displaystyle\int_{\alpha_k}^{\beta_k} f_k(r)\,dr & si\ \rho_k^2 < \rho_{k-1} \\ 1 - \dfrac{P_d\left(\rho_k^2 - \rho_{k-1}\right)}{\rho_k\left(\rho_k - \rho_{k-1}\right)} e^{-\gamma_k . t} - \dfrac{P_d(1-\rho_k)}{2\,\pi\,\rho_k} \displaystyle\int_{\alpha_k}^{\beta_k} f_k(r)\,dr & \sin on \end{cases} \tag{5.7}$$

où

$$\rho_k = \frac{1}{C\,\mu} \sum_{i=1}^{k} \lambda_i \tag{5.8}$$

$$\alpha_k = \left(1 - \sqrt{\rho_{k-1}}\right)^2 C\mu \tag{5.9}$$

$$\beta_k = \left(1 + \sqrt{\rho_{k-1}}\right)^2 C\mu \tag{5.10}$$

$$\gamma_k = \frac{(\rho_k - \rho_{k-1})(1-\rho_k)}{\rho_k} C\mu \tag{5.11}$$

$$f_k(r) = \frac{\sqrt{(\beta_k - r)(r-\alpha_k)}}{r(r-\gamma_k)} e^{-r.t} \tag{5.12}$$

Les équations (5.7) et (5.6) nous permettent de déterminer la proportion de clients répondus en un temps inférieur à t. Il faut fixer t à 20 secondes pour avoir la qualité de service relative à chaque classe. Par la suite, il suffit de prendre $n = 3$ et d'utiliser les paramètres de notre système pour analyser la file logique. Le nombre optimal de serveurs peut être déterminé en exploitant une procédure numérique classique et qui est, légèrement, plus compliquée que celle de la section précédente puisque les formules sont moins simples pour les critères de service regardés.

5.3.2 Évaluation de performances du système actuel

Dans le système actuel, il n'est pas possible d'évaluer analytiquement (ou numériquement) les performances. Ceci est dû au routage dynamique qui alimente les sites en appels. Nous avons donc opté pour la simulation afin de déterminer le nombre optimal de serveurs. En fonction du niveau de service objectif, nous déterminons le nombre minimal de conseillers qu'il faut prévoir pour atteindre les objectifs de qualité de service.

Dans la suite, nous allons nous intéresser, uniquement, à la qualité de service concernant les proportions de clients répondus en moins de 20 secondes. Sur cette base, nous allons comparer plusieurs règles de routage avec ce que la file logique permet d'avoir et ce, suivant la qualité de service envisagée.

5.3.3 Optimisation du nombre de conseillers

Nous allons considérer plusieurs qualités de service objectif dans la suite. Pour chaque qualité de service, nous comparons plusieurs règles de routage basées sur les estimateurs du temps d'attente décrits auparavant. Ces règles de routage seront, également, comparées avec la file logique qui, rappelons le, ne nécessite pas de routage. Les paramètres utilisés sont ceux de la section 4.4.2 du chapitre précédent, mis

à part le nombre de serveurs qui ne va plus être un paramètre du système mais plutôt un résultat. Le taux d'appels sera fixé à 55 par minute.

– Dimensionnement pour la qualité de service (99,9|99|95)

À partir des formules (5.6) et (5.7), nous déterminons pour la file logique, numériquement, le nombre de serveurs nécessaire pour la satisfaction de 99,9 % des clients A, 99 % des clients B et de 95 % des clients C en moins de 20 secondes. Ces qualités de service constituent le haut de la plage des niveaux de service qu'un centre d'appels peut envisager d'atteindre. Nous obtenons, ainsi, un nombre optimal de serveurs égal à 241 pour ce système composé de trois classes de priorité.

Avec un logiciel de simulation, nous déterminons le nombre optimal de serveurs qui, dans le système actuel, permet d'atteindre le niveau de service mentionné. Nous commençons par regarder la version *équilibrée* du système (tous les sites sont équivalents, il y a au plus une différence de 1 serveur entre deux sites). Avec un routage basé sur une minimisation du temps d'attente par l'ASA, dans lequel les appels sont routés vers le site qui offre la meilleure estimation du temps d'attente suivant l'ASA, le nombre minimum de serveurs à utiliser est égal à 448. Ce nombre contraste fortement avec les 241 serveurs de la file logique. Cette différence importante vient de la règle de routage utilisée. En optant pour une règle de routage basée sur l'ASA^2, nous obtenons un nombre optimal de serveurs égal à 400 qui reste très élevé. Il est, cependant, plus performant que l'ASA. La différence de performance entre les deux vient du fait qu'avec l'ASA, un client de classe *B*, par exemple, peut être routé vers un site pour y attendre juste parce qu'il n'y a pas de clients de classe *B* en attente et ce, même lorsque son attente aurait été nulle sur un autre site. Avec l'ASA^2, ce même client n'est routé vers le site où il va attendre que lorsqu'il n'y a ni des clients de classe *B* ni des clients de classe *A* en attente, ce qui améliore, légèrement, la performance. Toutefois, même avec l'ASA^2 la performance n'est pas du niveau de la file logique. Ce manque de performance est dû à des informations limitées. En effet, que nous utilisons l'ASA ou bien l'ASA^2, nous ne pourrons pas savoir si le site, pour lequel le temps d'attente estimé est nul, offre réellement une possibilité de service immédiat ou non, faute d'informations supplémentaires. Ceci n'est pas très contraignant pour les clients de classe *A* parce que, au pire des cas, ils devront attendre qu'un conseiller se libère.

Pour les clients de classes *B* et *C*, en revanche, c'est plus délicat puisqu'une période d'occupation dure plus longtemps. La qualité de service se détériore d'autant plus que lorsque l'estimation donne un temps d'attente nul sur plusieurs sites, c'est le même site qui soit choisi. En effet, dans ce cas, la probabilité pour qu'un client, à qui est estimé une attente nulle, soit routé vers un "mauvais" site augmente considérablement. Les résultats de la simulation montrent, justement, que l'utilisation du premier site dépasse les 99 % alors que le quatrième site n'est, quasiment, pas exploité. Pour remédier, donc, à ces problèmes, nous décidons, dans le cas où plusieurs sites offrent une attente nulle, de router aléatoirement entre eux avec une probabilité statique dépendante du nombre de serveurs sur chacun de ces sites. Ceci améliore nettement les performances puisque nous obtenons un nombre optimal de 247 serveurs avec l'ASA, et de 246 serveurs avec l'ASA^2. Dans la suite, sans informations supplémentaires sur l'état du système, nous allons toujours incorporer ces routages aléatoires dans le cas où un temps d'attente nul est estimé.

Dans la réalité, les sites sont *déséquilibrés* puisqu'ils présentent, souvent, des nombres de serveurs différents. Nous allons considérer que les serveurs sont répartis de la façon suivante entre les sites :

- 40 % des serveurs se retrouvent sur le premier site
- 25 % des serveurs se retrouvent sur le deuxième site
- 20 % des serveurs se retrouvent sur le troisième site
- 15 % des serveurs se retrouvent sur le quatrième site

Cela représente une répartition assez réelle. Bien entendu, pour un nombre total de serveurs *C*, cette répartition ne sera pas exactement respectée à cause des valeurs réelles des serveurs par site.

Pour le système *déséquilibré*, le routage basé sur l'ASA impose un nombre minimum de serveurs égal à 257. L'ASA^2 en impose 256, ce qui reste en deçà des performances du système *équilibré*. Cette détérioration des performances résulte de l'augmentation de la variabilité. Dans la suite, nous n'allons plus traiter le système *équilibré* puisqu'il n'est pas très représentatif de la réalité.

Toujours pour le système *déséquilibré*, nous nous intéressons à connaître l'étendu de l'amélioration que peut apporter une information supplémentaire. En fait, nous voulons juste pouvoir distinguer les sites où un ou plusieurs conseillers sont libres. S'il n'y en a pas, alors le routage se base sur une minimisation du temps d'attente estimé sur chaque site. Pour ce routage que l'on combinera avec l'ASA[2], le nombre optimal de serveurs est égal à 240, ce qui est encore plus performant que la file logique.

L'ensemble des résultats cités est résumé par le Tableau 5.1.

Qualités de service à atteindre (A|B|C)		99,9|99|95	Classe saturée
file logique		241	C
Système équilibré	ASA + tirage aléatoire	247	B
	ASA$^{(2)}$ + tirage aléatoire	246	A
Système déséquilibré	ASA + tirage aléatoire	257	A
	ASA$^{(2)}$ + tirage aléatoire	256	A
	ASA$^{(2)}$ + Connaissance de la disponibilité	240	B

Tableau 5.1: Dimensionnement pour la qualité de service 99,9|99|95

Dans le Tableau 5.1, nous avons illustré le nombre de serveurs nécessaires pour atteindre la qualité de service 99,9|99|95 et ce, suivant la nature du système étudié. Nous avons également noté la classe saturé lors de l'atteinte du niveau de service objectif. Cette classe représente les clients dont le niveau de service empêche le système d'être plus performant (si le nombre de serveurs diminue alors l'objectif fixé pour cette classe ne sera pas atteint).

À partir de ce tableau, nous pouvons constater que la classe contraignante pour la file logique est la classe *C*. En effet, il s'agit du système qui handicape le plus les clients de cette classe. Pour ce système, tous les clients de classes *A* et *B* venus avant le début de service d'un client de classe *C*, déjà en attente à leurs arrivées, seront servis avant lui. Pour le système actuel, il est possible qu'un client de classe *C* soit servi avant un client de classe *A*, arrivé avant le début de son service mais routé vers un autre site. Il est même possible, pour des raisons stochastiques, qu'un client de classe *C* soit servi

avant un client de classe *A* arrivé avant lui mais routé vers un autre site. Nous ne pouvons, donc, pas affirmer que le système actuel respecte la priorité stricte entre classes, au contraire de la file logique. C'est ce qui rend la classe *C* la plus difficile à satisfaire pour la file logique.

Dans le même tableau, nous remarquons que le système actuel avec une information, lors du routage, sur la disponibilité d'au moins un conseiller sur un site, est légèrement plus performant que la file logique. Cette performance s'explique par la difficulté de la file logique à satisfaire les clients de classe *C*. Ceci conforte notre opinion sur la nécessité de disposer de plus d'informations lors du routage, sans cela, la performance de la file logique est la meilleure.

– Dimensionnement pour la qualité de service (99,9|95|80)

Il s'agit d'un qualité de service plus raisonnable et qui reste, tout de même, assez élevée, spécialement pour la classe *A*. Les résultats sont résumés par le Tableau 5.2.

Qualités de service à atteindre (A|B|C)		99,9|95|80	Classe saturée
file logique		232	C
Système déséquilibré	ASA + tirage aléatoire	257	A
	ASA(2) + tirage aléatoire	256	A
	ASA(2) + Connaissance de la disponibilité	238	A

Tableau 5.2: Dimensionnement pour la qualité de service 99,9|95|80

Dans le Tableau 5.2, nous constatons que la file logique devient le système le plus performant. Elle est toujours handicapée par la classe *C* qui a, pourtant, un objectif moins ambitieux. Désormais, le système actuel, même avec une information sur la disponibilité des serveurs, est loin d'avoir la même performance, lorsque aucune information n'est fournie. Dans ce cas, l'objectif élevé de niveau de service pour la classe *A* fait que le système nécessite le même nombre de conseillers que pour le cas précédent où la qualité de service est 99,9|99|95 (puisque c'est toujours le même niveau de service pour les clients *A*).

– Dimensionnement pour la qualité de service (94|74|54)

Ce niveau de qualité de service peut être jugé plus raisonnable que les deux niveaux précédents. Les résultats sont résumés par le Tableau 5.3.

Qualités de service à atteindre (A|B|C)		94|74|54	Classe saturée
file logique		227	C
Système déséquilibré	ASA + tirage aléatoire	226	C
	$ASA^{(2)}$ + tirage aléatoire	226	C
	$ASA^{(2)}$ + Connaissance de la disponibilité	226	C

Tableau 5.3: Dimensionnement pour la qualité de service 94|74|54

Comme illustré par le Tableau 5.3, la file logique n'est plus aussi performante que le système actuel. À nouveau, elle privilégie nettement la classe *A* en lui satisfaisant largement ses objectifs. Nous remarquons également que la connaissance de la disponibilité sur les sites n'ajoute rien par rapport à un routage basé uniquement sur l'ASA ou sur l'ASA^2. Ceci s'explique par le fait que la charge ρ pour ce résultat est assez élevée (puisque l'objectif n'est plus aussi ambitieux qu'auparavant). Elle est égale à 97,3 % (loin devant les 90 % des exemples précédents). Pour cette charge, la probabilité de trouver un serveur disponible est devenue faible, ce qui diminue l'importance de l'avantage que procure l'information supplémentaire sur la disponibilité. Notons également que, même pour le système actuel, la classe contraignante est la classe *C* puisque la charge a augmenté et qu'il est plus fréquent de se voir dépasser dans les files d'attente.

– Dimensionnement suivant plusieurs niveaux de service

Dans la Figure 5.1, nous avons affiché l'évolution du nombre optimal de conseillers de clientèle en fonction de la qualité de service objectif et ce, pour trois systèmes différents: la file logique, le système actuel avec un routage basé sur l'ASA^2 et le système actuel avec routage ASA^2 et connaissance de la disponibilité. Dans cette figure,

nous avons ajouté d'autres objectifs de qualité de service à ceux déjà analysés. Nous avons également noté la classe contraignante au dessus des histogrammes.

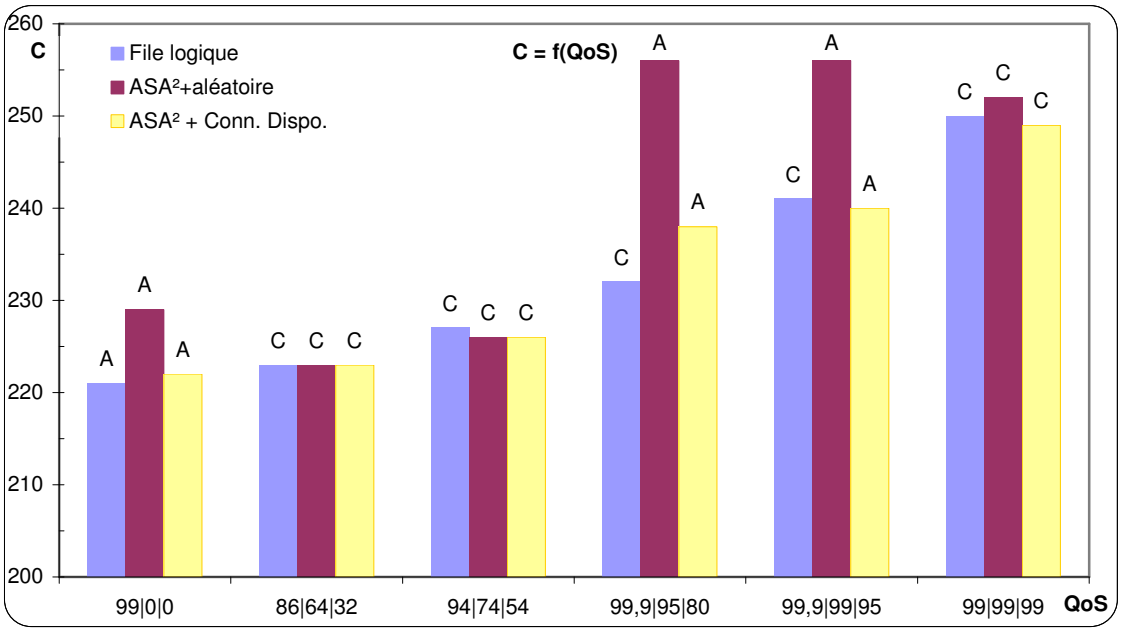


Figure 5.1: Dimensionnement suivant la Qualité de Service (QoS) objectif

À partir de la Figure 5.1, nous pouvons affirmer ce que nous avions déjà constaté: la file logique peine à satisfaire les clients de classe *C* puisque c'est le seul système où l'on peut vraiment parler de priorité stricte. En effet, à part le cas fictif d'un niveau objectif de 99|0|0 (où l'on s'intéresse uniquement au niveau de service des clients *A*), c'est toujours la classe *C* qui l'handicape. Signalons ici que pour ce cas précis, la performance de la file logique pourrait être supérieure. Nous avons choisi de ne pas aller en dessous de 221 serveurs parce que le système serait instable ($\rho \geq 1$), même si ce sont les clients *B* et *C* qui en seront les seuls atteints.

Nous constatons aussi que la file logique est la plus performante lorsque le niveau de service des clients *A* est très important alors que celui des clients *C* ne l'est pas. Dans le cas où le système actuel est meilleur, il ne l'est, que d'un seul serveur. À noter également qu'avec l'information sur la disponibilité des serveurs, le système actuel est au moins aussi performant qu'avec un routage basé sur l'ASA² uniquement.

5.4 Conclusions

L'étude du routage avec des règles basées sur les estimateurs issus du chapitre précédent, a démontré que le meilleur routage est celui qui dispose d'un minimum d'informations. L'information supplémentaire que nous avons exploitée est la connaissance de la disponibilité d'au moins un serveur libre sur les sites. Avec cette information, le système à plusieurs files peut même s'avérer plus performant que la file logique si les objectifs de service ne sont pas très ambitieux pour les clients de classe *A*. Dans ce cas, la différence ne dépasse jamais un serveur, ce qui nous permet d'opter pour la file logique. En effet, lorsque la qualité de service objectif des clients *A* est élevée, la différence en faveur de la file logique peut être chiffrée en plusieurs serveurs.

S'il n'y a pas la moindre information sur la disponibilité des serveurs, alors la performance du système à plusieurs files d'attente peut se détériorer nettement. Tout ceci nous conforte dans notre préférence pour la file logique, quelque soit l'objectif, elle reste très proche du meilleur système, si ce n'est pas elle. De plus, le jour où les objectifs changent, le système à plusieurs files ne va pas nécessairement rester performant.

Un dernier argument qui plaide en faveur de la file logique consiste en sa simplicité. La visibilité du système est meilleure et sa compréhension l'est également. Même la façon de gérer le système à plusieurs files d'attente n'est pas évidente puisque, suivant la règle de routage, la charge de travail pour un groupe de conseillers peut être largement différente de ce qu'un autre groupe de conseillers effectue comme travail.

Malgré le fait que la file logique reste souvent meilleure qu'un système composé de plusieurs files avec un routage des clients, il faut voir que, contrairement à l'idée reçue, il est possible que la décomposition du système en plusieurs sites peut améliorer sa performance. Ceci vient de l'existence de plusieurs classes de clients avec un objectif propre à chaque classe. Le fait que ces objectifs existent peut paraître contradictoire avec les priorités strictes entre classes de clients. Et, justement, nous pouvons voir que la règle de routage dans le système décomposé peut casser la priorité stricte puisqu'il est, même, possible qu'un client de classe *C* soit servi avant un client de classe *A* arrivé,

pourtant, avant lui. Dans le prochain chapitre, nous allons étudier des disciplines de priorité dans le but de satisfaire, au mieux, les objectifs fixés.

Chapitre 6 : Étude de la priorité non-préemptive dans un centre d'appels multi-classe

6.1 Introduction

L'étude que nous menons dans ce chapitre s'inscrit, à nouveau, dans le cadre de l'intérêt que portent les managers de centres d'appels à la satisfaction de l'attente de leurs clients et ce, avec le soucis habituel de limitation des coûts. Cette étude concerne les priorités attribuées aux clients dans la file d'attente.

Nous avons déjà modélisé le centre d'appels avec des priorités entre classes de clients dans le chapitre précédent. Ces priorités – que nous désignerons par *priorités strictes* parce que la priorité d'un client est un acquis qui ne change pas – ne représentent pas forcement le meilleur choix pour le centre d'appels. En effet, comme nous l'avions déjà constaté avec des priorités strictes entre clients, les objectifs de niveau de service fixés peuvent être largement atteints pour une classe alors que ceux correspondant à une autre classe ne sont pas satisfaits. Nous avons même vu que le centre d'appels peut, dans certains cas, atteindre ses objectifs plus facilement lorsqu'il est composé par plusieurs files d'attente au lieu d'une seule. Nous expliquons ceci par le changement des priorités dû au routage et que, de ce fait, le système ne fonctionne

plus réellement avec des priorités strictes entre classes de clients. Ceci est d'autant plus compréhensible lorsque nous remarquons que certains clients de classe *C* sont servis avant quelques clients de classe *A*.

Dans ce chapitre, nous allons nous focaliser sur l'analyse d'une alternative simple aux priorités strictes entre clients et ce, dans le but d'analyser les améliorations que peuvent induire des priorités plus sophistiquées que la priorité stricte. Pour ce faire, nous considérons un centre d'appels composé d'un seul site similaire à la file logique, ce qui a pour avantages d'éliminer les effets indésirables dus au routage dans le cas multi-site. Dans ce dernier cas, la compréhension et l'analyse des priorités auraient été faussées par le routage. Toujours dans un but de compréhension, nous allons nous restreindre à deux classes de clients au lieu de trois. La généralisation à plusieurs classes de clients n'est pas, nécessairement, évidente mais cela nous permet, au moins, d'avoir des éléments qui nous permettent d'acquérir une meilleure visibilité du comportement du système suivant les différents paramètres et, en même temps, d'avoir une idée qualitative sur l'évolution qu'engendre une extension à plusieurs classes de priorités.

Notre objectif reste toujours de satisfaire des niveaux de service cibles, différenciés suivant la classe de clients. Ces objectifs sont fixés au préalable. Il est commode dans les centres d'appels de fixer, également, des priorités strictes entre les clients et ce, suivant leurs importances pour l'entreprise. Comme nous l'avions déjà mentionné, ceci peut nuire à la performance du centre d'appels. Il faut, surtout, remarquer que le fait de fixer, à la fois, le type de priorité ainsi que les objectifs de service pour chaque classe peut être contradictoire. Pour cela, nous allons garder comme objectif un niveau de service minimum à atteindre, pour lequel nous ne nous imposons pas de priorité stricte. Bien entendu, cette recherche de la qualité de service objectif doit s'accompagner d'une minimisation des coûts. Ces coûts sont supposés proportionnels au nombre de conseillers nécessaire à l'atteinte des objectifs. Le travail consistera, donc, à améliorer la règle de priorité considérée jusqu'à présent de telle sorte à diminuer le nombre de serveurs garantissant les qualités de service requises pour chaque classe de clients.

Les qualités de service objectif que nous allons considérer, consistent en deux manières de représentation des niveaux de service. Nous les avons déjà utilisées dans le chapitre précédent. Il s'agit du temps moyen d'attente et de la proportion de clients répondus en moins de vingt secondes, toutes les deux dépendent de la classe de clients. Les deux classes de clients que nous allons utiliser seront la classe *A* et la classe *B*. La classe *A* étant plus exigeante de point de vue qualité de service. Avec des priorités strictes, la classe *A* aurait eu la priorité par rapport à la classe *B*.

Au-cours de ce chapitre, nous allons analyser deux règles de priorités qui font partie des priorités statiques et ne dépendent, donc pas, de l'état du système. Elles seront, fortement basées sur les priorités strictes déjà rencontrées. La première règle attribue aux clients la possibilité de passer par la file prioritaire et ce, avec une certaine probabilité statique qui dépend uniquement de la classe de client. En ce qui concerne la deuxième règle, lorsqu'un serveur termine un service, la classe de clients à qui il va répondre est déterminée par une probabilité statique. Les performances des deux règles seront comparées avec celles de la priorité stricte. La performance étant mesurée, cette fois, en termes de conseillers nécessaires pour atteindre les objectifs de niveaux de service pour le cas multi-serveur, ou en capacité de service minimale dans le cas mono-serveur. Lorsqu'il s'agit de dimensionner le système de telle sorte à ne pas dépasser des temps moyens d'attente objectif pour chaque classe, nous allons prouver que les disciplines utilisées offrent les performances optimales. Lorsque le niveau de service objectif s'exprime en proportions de clients répondus en moins d'un certain temps, nous ne cherchons pas à montrer l'optimalité des disciplines étudiées puisqu'elles ne sont pas optimales, mais nous voulons, juste, montrer que nous pouvons gagner en performances en apportant de simples améliorations à la discipline de priorité la plus répandue.

Après cette introduction, nous allons exposer les travaux antérieurs qui touchent, essentiellement, aux files d'attente avec priorités. Nous y verrons quelques travaux concernant des règles de priorité particulières ainsi que leur mise en œuvre.

Dans la troisième section, nous commençons par introduire la première règle de priorité probabiliste. Nous comparons les gains en performances de cette règle avec la règle de priorité stricte. Nous montrons, dans cette section, que la règle offre la capacité

de service minimale lorsque les objectifs fixés s'expriment en temps moyens d'attente. Nous proposerons, également, une méthode pour étendre la règle de priorité probabiliste à plusieurs classes de clients.

Dans la quatrième section, nous traitons le deuxième type de priorité probabiliste et nous analyserons ses résultats en les comparant avec la priorité stricte et avec la première priorité étudiée.

Nos conclusions sur l'étude des priorités probabilistes seront exposées dans la cinquième section.

6.2 Étude bibliographique

Nous présentons, dans cette section, les travaux ayant traité les files d'attente multi-classe. Étant donné le fonctionnement des centres d'appels, nous restreignons notre intérêt aux publications qui ont étudié les priorités non-préemptives. En fait, dans les centres d'appels, une fois le client accède à un serveur, il ne peut plus être mis en attente à cause de l'arrivée d'un appel plus prioritaire. Ceci représenterait une très mauvaise image du centre d'appels.

Les modèles de files d'attente multi-classes représentent une littérature abondante. Plusieurs livres traitants la théorie des files d'attente, d'une manière générale, ont évoqué les systèmes avec priorités. Nous pouvons citer, par exemple, Gross et Harris [33]. D'autres livres se sont intéressés, plus spécialement, aux files d'attente avec priorités. Le travail de Jaiswal [40] s'inscrit dans ce cadre. Il est dédié, en exclusivité, à l'étude des files d'attente avec priorités. Il y introduit plusieurs disciplines de priorités, outre les deux plus connues et qui sont: la priorité préemptive et la priorité non-premptive. Dans la première discipline, un client peut voir sont service interrompu suite à l'arrivée d'un client plus prioritaire, alors que, dans la deuxième discipline, un service ne peut être interrompue et la différence de priorité se manifeste, uniquement, dans la position dans la file d'attente avant le service. Un autre livre où l'on peut distinguer des disciplines de priorité est Takagi [61].

D'autres travaux se sont intéressés aux files d'attente avec priorités. Dans leur étude des systèmes avec des files d'attente multiples, Adan, Boxma et Resing [1] se sont, notamment, intéressés aux priorités que le serveur engendre en choisissant, lui même, la file d'attente qu'il va servir. Une autre règle de priorité consiste à servir la file d'attente la plus longue qui est, en quelque sorte, le dual du routage à la file la plus courte puisque les deux tendent à équilibrer les tailles des files d'attente. Boxma, Koole et Liu [13] ont pour leur part, proposé un état de l'art détaillé articulé autour des systèmes parallèles et distribués. Ils y évoquent, notamment, la discipline où c'est le serveur qui choisit la prochaine file à servir et ce, de façon dynamique ou statique.

Le modèle auquel nous portons notre premier intérêt est le modèle multi-serveur avec plusieurs classes de clients différenciées par des priorités non-préemptives. Comme nous l'avons mentionné dans le chapitre précédent, ce modèle a été analysé par Davis [21]. Il a déterminé la distribution des temps d'attente de chaque classe de clients, pour un modèle M/M/C. Kella et Yechiali [45] ont déterminé les premiers moments du temps d'attente pour le même modèle et ce, suivant la classe de clients. En se limitant à deux classes de clients, Gail, Hantler et Taylor [26] ainsi que Kao et Narayanan [43] ont analysé le régime stationnaire du modèle M/M/C en se basant sur des approches matricielles. Pour des temps de service déterministes, Altinkemer, Bose et Pal [7] ont étudié un modèle M/D/C avec des priorités non-préemptives et lui ont approché les temps moyens d'attente par classe. Dans le cas d'une priorité préemptive et pour des temps de service suivant une loi générale, van-der-Mei *et al.* [65] ont proposé des approximations du temps de séjour moyen, suivant la classe de clients.

Concernant le cas mono-serveur, Miller [56] a étudié le calcul des probabilités au régime stationnaire avec les deux types de priorité les plus fréquemment utilisés – la priorité préemptive et la non-préemptive. Son modèle s'intéresse à l'existence de deux classes de clients, et l'approche numérique qu'il a développée est basée sur la méthode de Matrice Géométrique issue de Neuts [58]. Le résultat est un algorithme qui calcule les probabilités des états au régime stationnaire. L'auteur arrive également à aboutir à des résultats analytiques comme la détermination de la probabilité d'avoir un certain nombre de clients de la classe prioritaire dans le cas de la non-préemption.

L'analyse du temps d'attente a, souvent, fait l'objet de travaux dans le domaine des files d'attente à priorités. Wagner [66] a traité un modèle multi-serveur à priorité non-préemptive pour lequel l'espace d'attente est limité. Il a développé une procédure numérique dans le but de déterminer la distribution du nombre de clients dans le système ainsi que la distribution du nombre de clients en attente lorsque le nombre de classes est limité à deux. Il a également développé une méthode de détermination des moments du temps d'attente pour chaque classe de clients et ce, pour un nombre de classes quelconque. Kao et Wilson [44] ont étudié l'implémentation d'une méthode avec des séries dans le but d'obtenir des mesures de performances d'une file à priorités non-préemptives.

Plusieurs modèles particuliers ont été analysés. Leemans [51] a analysé un modèle à deux serveurs et deux classes de clients. La priorité étant non-préemptive, une classe prioritaire pour un serveur ne l'est plus pour l'autre serveur. Il détermine la distribution des clients en attente ainsi que des temps d'attente au régime stationnaire avec l'approche de la matrice géométrique. Wang, Gong et Lee [67] ont combiné la règle *premier – arrivé – premier – servi* avec celle du *temps - le - plus - court* dans un modèle M/G/1 et ce, pour diminuer les temps d'attente moyens tout en permettant aux clients ayant des temps de service importants, de ne pas attendre longtemps. Beaucoup de travaux ont incorporé des seuils à leurs approches des files à priorités. Bhulai et Koole [12] ainsi que Gans et Zhou [28] ont étudié un système à deux classes de clients. La classe la plus prioritaire arrive suivant une loi de Poisson. La classe la moins prioritaire est toujours disponible ce qui revient à supposer que son taux d'arrivée est infini. Ils ont montré qu'avec des seuils appropriés, il est possible de maximiser le nombre de clients de moindre priorité tout en satisfaisant une contrainte au niveau du temps d'attente de la classe prioritaire. Nous pouvons citer, également, le travail de Huang [38] dans lequel il traite l'utilisation de deux seuils H et L dans la file d'attente de la classe la moins prioritaire. La priorité étant non-préemptive. Le but est de contrôler le temps d'attente de chaque classe et, en particulier, de satisfaire les besoins de la classe prioritaire, sans pour autant défavoriser plus qu'il ne le faut l'autre classe. L'analyse est effectuée grâce à un modèle de chaîne de Markov à trois dimensions. Avec les deux seuils, l'auteur définit deux états : un *heavy state* et un *light state*. À tout instant, le système se trouve nécessairement dans l'un des deux états et ce, en fonction des deux seuils mentionnés. Si pour le *light state* ce sont les clients les plus prioritaires

qui sont servis, dans le *heavy state* le serveur alterne entre les deux classes de clients. Boxma et Down [16] se sont également intéressés à un modèle de priorité avec un seuil. Ils ont considéré un système mono-serveur de temps de service Général et avec des arrivées suivant la loi de Poisson. Dans leur système, deux files d'attente existent. Lorsque le serveur se trouve à la première file, il continue de servir tous les clients de cette file jusqu'à ce qu'elle se vide, avant de passer à l'autre file. Si le serveur se trouve à la deuxième file, alors il continue de servir les clients de cette file jusqu'à ce qu'elle se vide ou bien jusqu'à ce que la première file atteint un seuil prédéfini. Les auteurs ont, en particulier, proposé une approximations aux tailles moyennes des files d'attente. D'autres modèles sont analysés par Lee, Park et Song [49], Lee, Seo et Yoon [50], ou encore Feng, Kowada et Adachi [25].

Toujours dans le but de différencier les qualités de service perçues par les classes de clients constituant le système, Tham, Yao et Jiang [62] ont analysé une discipline de priorité probabiliste pour un système mono-serveur. Dans cette discipline, chaque classe de clients possède sa propre file d'attente et un poids est affecté à chaque file. Le serveur regarde les files d'attente dans leur ordre de priorité avant de leur affecter des probabilités de service, suivant un calcul dépendant des poids déjà affectés. Un algorithme a été mis au point pour le calcul des probabilité de servir une file en particulier. Une étude des temps moyens d'attente a également été effectuée. Dans cette étude, les auteurs s'appuient sur des bornes inférieures et supérieures, établies dans un autre article, pour montrer que les temps moyens d'attente pour chaque classe se situe bien entre ces bornes pour des arrivées avec une loi de Poisson. Par contre, avec une loi de Pareto, ce n'est pas toujours valable. La même discipline de priorité est analysée par Tham, Yao et Jiang [63] et également par Jiang, Tham et Ko [42] où ils montrent aussi qu'il y a une différenciation du service reçu par les différentes classes de clients. Dans le cas où l'on considère deux classes de clients, l'une des disciplines que nous allons modéliser correspond au cas où les auteurs de ces articles considèrent deux classes de clients également. Cependant, nos approches ne sont pas les mêmes puisqu'ils cherchent à montrer qu'il peut y avoir une différence dans le service fourni à chaque classe de clients alors que nous allons essayer de quantifier le gain, en nombre de serveurs, suite à la satisfaction d'objectifs de qualité de service. Un autre travail qui a abordé une discipline de priorité probabiliste est celui de Boxma [14]. Il y a traité un modèle multi-classe où le serveur choisit la prochaine classe à servir avec une

probabilité statique qui dépend de la classe en question. Chaque classe possède sa propre file d'attente. Dans son modèle, il est possible d'avoir des temps non nuls lors du changement de files. Les temps de service sont dépendants de la classe de clients. Lorsque le serveur commence à servir une classe particulière, en fonction de cette classe il a deux possibilités : il sert tous les clients de la classe avant de passer à une autre file, ou bien il sert tous les clients de cette classe qui étaient déjà en attente lorsqu'il a commencé à servir leur file. L'auteur détermine les probabilités qui permettent de minimiser la charge de travail moyenne du système. Pour le système que nous allons traiter dans notre travail, sa fonction objectif serait constante puisque nous n'avons pas de temps de changements de files et, en même temps, toutes les classes possèdent la même distribution du temps de service. De plus, nous avons un objectif distinct par classe de clients.

À noter, enfin, que dans la plupart des travaux antérieurs, les disciplines de priorités "non classiques" sont abordées dans le but de montrer qu'elles peuvent changer le comportement du système. Dans ce chapitre, nous allons, plutôt, nous focaliser sur le gain que l'on peut escompter en termes de nombre de serveurs suite à l'exploitation d'une certaine discipline de priorités entre classes.

6.3 Premier modèle de la priorité probabiliste

Cette section sera consacrée à l'étude du premier modèle dans lequel nous avons traité la priorité probabiliste. Dans ce modèle, deux degrés de priorité sont possibles. Ils sont affectés aux clients à leurs arrivées au système suivant leurs classes. Nous allons distinguer dans la suite deux cas. Dans le premier cas, il s'agit de déterminer le taux de service optimal lorsque le système se compose d'un seul et unique serveur. Ce cas de figure, non réaliste dans les centres d'appels, nous sert à simplifier le système, dans un premier temps. Dans le deuxième cas, nous cherchons à déterminer le nombre optimal de serveurs pour un taux de service fixé.

6.3.1 Système mono-serveur

Nous considérons l'existence de deux classes de clients, et nous supposons que le centre d'appels n'exploite qu'un seul serveur dont le taux de service peut être choisi à

l'avance. Le centre d'appels est supposé ne pas fixer de priorités strictes à ces clients. Cependant, suivant l'importance relative de ses clients, il se fixe des objectifs de qualité de service qui varient en fonction de la classe de clients. Nous désignerons par *A* la classe des clients à qui est fixée l'objectif le plus ambitieux. La deuxième classe sera appelée classe *B*.

Dans le modèle de priorité probabiliste que nous considérons ici, les clients de classe *A* passent par la file prioritaire, à leurs arrivées, avec une probabilité statique p_A. Les clients de la classe *B* passent, de la même façon, par la file prioritaire avec une probabilité p_B. Les clients de classes *A* et *B* passent par la file la moins prioritaire, respectivement, avec les probabilités complémentaires $1 - p_A$ et $1 - p_B$. Le taux d'arrivée total étant égal à λ composé par des arrivées de classe *A* avec un taux λ_A et des arrivées de clients *B* avec un taux λ_B. Dans chacune des deux files, l'ordre des arrivées est respecté. Le modèle décrit est illustré par la Figure 6.1.

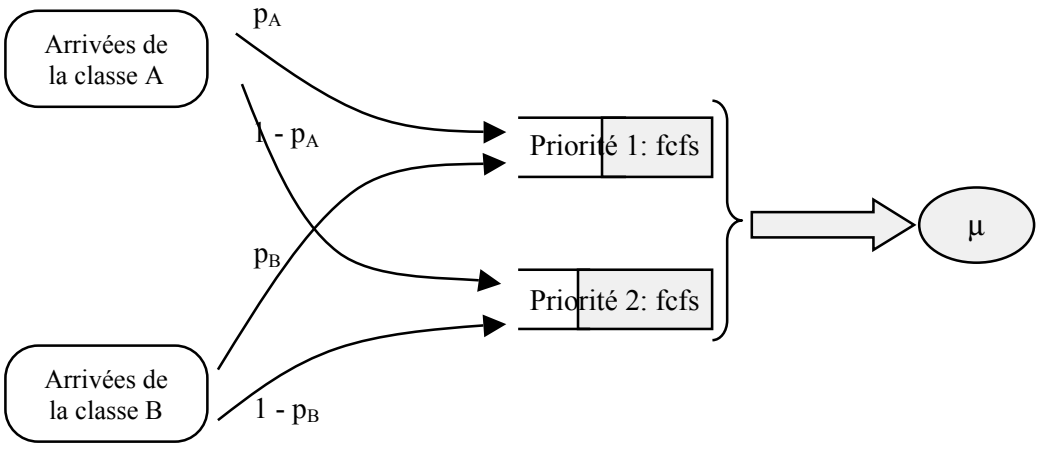


Figure 6.1: File d'attente avec la première priorité probabiliste

Les objectifs de qualités de service de chaque classe de clients se divise en deux catégories. Dans la première, ils s'expriment en proportion de clients répondus en moins de vingt secondes. Dans le deuxième catégorie, il s'agit de fixer un temps moyen d'attente par classe à ne pas dépasser. Nous allons traiter ces deux catégories dans la suite.

6.3.1.1 Proportion d'appels répondus en moins de vingt secondes

Le système modélisé par la Figure 6.1 peut être représenté d'une manière différente. Nous pouvons dire qu'il se compose de deux classe de clients notées classe 1 et classe 2. Chaque classe de clients passe impérativement par une seule file d'attente, la classe 1 passant par la file prioritaire. Les arrivées des nouvelles classes sont les suivantes :

- $\lambda_1 = p_A \lambda_A + p_B \lambda_B$
- $\lambda_2 = (1 - p_A) \lambda_A + (1 - p_B) \lambda_B$

où λ_A et λ_B représentent, respectivement, les taux d'arrivée des classes *A* et *B*.

La nouvelle représentation du système, qui nous rappelle les priorités strictes, est illustrée par la Figure 6.2.

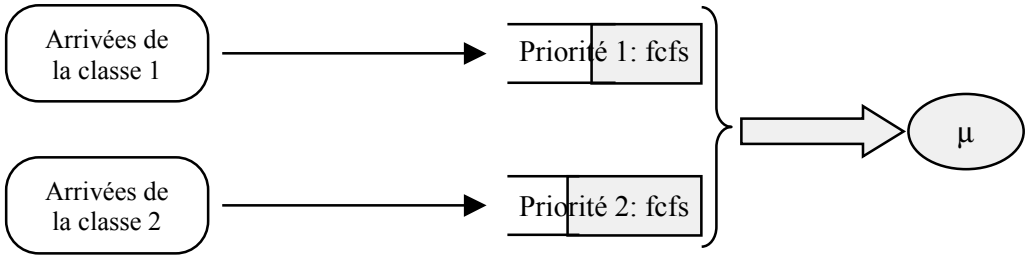


Figure 6.2: File d'attente avec priorité stricte

La distribution des temps d'attente du système de la Figure 6.2 a été étudiée par Davis [21]. Pour la première classe de clients, la probabilité pour que le temps d'attente soit inférieur à *t* est donnée par la formule suivante :

$$Wq_1(t) = 1 - \rho\, e^{(\lambda_1 - \mu)\, t} \tag{6.1}$$

avec la charge du système $\rho = \lambda / \mu$ et $\lambda = \lambda_A + \lambda_B$.

La distribution du temps d'attente pour la deuxième classe est déterminée grâce à la formule qui suit :

$$Wq_2(t) = \begin{cases} 1 - \dfrac{1-\rho}{2\,\pi} \int_{\alpha_2}^{\beta_2} f_2(r)\,dr & si\ \rho^2 < \rho_1 \\ 1 - \dfrac{\rho^2 - \rho_1}{\rho - \rho_1} e^{-\gamma_2 \cdot t} - \dfrac{1-\rho}{2\,\pi} \int_{\alpha_2}^{\beta_2} f_2(r)\,dr & \sin on \end{cases} \tag{6.2}$$

Les paramètres de cette équation ont été introduits dans le chapitre précédent. Il suffit de remplacer le nombre de serveurs *C* par 1 et le nombre de classes *n* par 2.

À partir de cette nouvelle représentation du système, nous pouvons revenir à celle où l'on peut distinguer les priorités probabilistes. Nous pouvons constater, facilement, que les qualités de service des classes *A* et *B* peuvent s'exprimer en fonction de celles des classes 1 et 2, qui dépendent elles mêmes des probabilités p_A et p_B, par l'intermédiaire des équations suivantes :

$$Wq_A(t) = p_A\ \mathrm{Wq_1}(t) + (1 - p_A)\ \mathrm{Wq_2}(t) \tag{6.3}$$

$$Wq_B(t) = p_B\ \mathrm{Wq_1}(t) + (1 - p_B)\ \mathrm{Wq_2}(t) \tag{6.4}$$

L'optimisation du système consiste à trouver le meilleur taux de service pour lequel la proportion d'appels répondus de chaque classe en moins de 20 secondes est supérieure à une valeur cible. Avec la priorité stricte, correspondant à $p_A = 1$ et $p_B = 0$, le taux de service optimal implique que la qualité de service d'une seule classe soit exactement égale à l'objectif fixé. Pour l'autre classe, l'objectif devrait être dépassé d'où une "sur-qualité" pour cette classe de clients. La priorité probabiliste doit nous permettre d'atteindre, exactement, la qualité de service objectif et ce, pour chaque classe de clients.

Nous déterminons à présent, pour divers exemples, le taux de service optimal qui conduit à la satisfaction, au moindre coût, des objectifs de niveaux de service. Nous fixons le taux d'arrivée λ à 100 appels par minute ce qui est réaliste pour un grand centre d'appels. La proportion d'appels issus de la classe *A* est un paramètre dont il faut tenir compte puisqu'il représente la composition des appels en clients de classes *A* et *B*. Pour chaque niveau de service, nous déterminons le taux de service optimal μ* avec la discipline de priorité probabiliste. Ceci est effectué par l'intermédiaire de

procédures de calcul numériques basées sur les formules (6.3) et (6.4). La détermination du taux de service optimal μ* implique la connaissance des probabilités statiques p_A et p_B. Lors des calculs numériques, nous cherchons à déterminer le taux μ* avec la méthode de dichotomie. À chaque étape du calcul, si nous n'arrivons pas à trouver un couple (p_A, p_B) qui nous permet d'atteindre la qualité de service objectif, alors le taux de service μ étudié est inférieur à μ*. Pour chaque valeur μ, afin de restreindre le champ de recherche des probabilités p_A et p_B, il faut remarquer que, généralement, une augmentation de p_A (ou de p_B) entraîne une augmentation de la qualité de service de la classe *A* (ou de la classe *B*). Lors des calculs effectués, nous faisons en sorte que des changements de valeur d'une (voire des deux) probabilité(s), impliquent une dégradation du taux de service assurant la satisfaction des niveaux de service objectifs. Ceci nous permet d'affirmer que le μ* trouvé correspond à la meilleure solution que l'on peut trouver avec cette discipline de priorité probabiliste. Le résultat final est comparé avec ce qu'une discipline de priorité stricte permet d'obtenir. Le Tableau 6.1 illustre les résultats des expérimentations numériques suivant les objectifs fixés et ce, en fonction de trois compositions différentes des appels.

λ_A/λ	Objectifs		Priorité stricte	Priorité probabiliste			Gain
	Classe A	Classe B	μ*	μ*	p_A	p_B	
0,3	90 %	80 %	106.51	100	1	0,850	6,10 %
	95 %	80 %	106.51	100	1	0,850	6,10 %
	99 %	80 %	106.51	100,2	1	0,805	5,92 %
	95 %	85 %	107.68	100,7	0,98	0,870	6,48 %
0,4	90 %	80 %	107.44	100	1	0,850	6,92 %
	95 %	80 %	107.44	100	1	0,850	6,92 %
	99 %	80 %	107.44	101,1	1	0,787	5,90 %
	95 %	85 %	108.77	101,3	0,98	0,864	6,87 %
0,5	90 %	80 %	108.63	100	0,98	0,850	7,94 %
	95 %	80 %	108.63	100,5	0,99	0,825	7,48 %
	99 %	80 %	108.63	103,4	0,99	0,727	4,81 %
	95 %	85 %	110.17	102	0,98	0,855	7,42 %

Tableau 6.1: Comparaison de la priorité stricte avec la priorité probabiliste

Le Tableau 6.1 nous permet d'observer l'évolution du taux de service optimal permettant d'atteindre les objectifs de proportions de clients répondus en moins de 20 secondes. Ceci nous permet de comparer la discipline de la priorité probabiliste avec celle de la priorité stricte et d'en mesurer le gain obtenu grâce à la première discipline.

Dans ce même tableau, nous observons à plusieurs reprises que le taux de service optimal relatif à la priorité probabiliste a une valeur égale à 100. Cette valeur coïncide avec le taux d'arrivée, ce qui nous donne une charge $\rho = 1$ et implique un système instable. En fait, comme le dit Jaiswal [40], un équilibre partiel relatif à la première file, représentative des clients prioritaires, peut être trouvé si $\lambda_1 = p_A \lambda_A + p_B \lambda_B$ est inférieur à μ. Ici, nous nous sommes arrêtés à une charge égale à 1. En réalité, nous aurions pu l'augmenter encore plus. Pour expliquer ceci, prenons l'exemple de la première ligne du tableau. La solution donnée par la priorité probabiliste implique le passage de tous les clients de classe *A* ainsi que de 85 % des clients de classe *B* par la file prioritaire. Ceci nous donne un taux d'arrivée à la file prioritaire $\lambda_1 = 89{,}5$ ce qui est inférieur à la valeur de taux de service obtenue. En prenant ces valeurs, nous savons que les 15 % des clients de classe *B* envoyés vers la file la moins prioritaire ne vont pas être servis dans les temps, s'ils seront servis. Toutefois, les 85 % des clients de classe *B* envoyés vers la file prioritaire vont augmenter la qualité de service de telle sorte que l'objectif soit atteint. Cela revient à dire que le système fait l'impasse sur le service de 15 % des clients de classe *B* pour que les clients de classe *A* et les clients restants de classe *B* aient un système peu chargé.

Nous constatons, également, que pour trois qualités de service différentes (90|80, 95|80 et 95|85), le gain obtenu par la priorité probabiliste augmente en fonction de la proportion d'appels reçus de la part des clients de classe *A*. En effet, à part la qualité de service 99|80, à chaque fois que la proportion des arrivées de classe *A* augmente, le gain augmente. Ceci s'explique par le fait que la priorité stricte a du mal à satisfaire les clients de classe *B* et que, dans ce cas, plus il y a des clients de classe *A*, plus les clients *B* seraient désavantagés. En revanche, la priorité probabiliste satisfait les clients suivant les objectifs, donc s'il y a plus de clients *A*, alors ceci se traduit par l'augmentation du nombre de clients à qui sont donnés des objectifs élevés, mais, en contre partie, il y a diminution du nombre de clients qui ont des objectifs moins élevés. Cette explication n'est plus valable pour des objectifs très élevés de la classe *A* car la priorité stricte les

privilégie déjà et que, avec une augmentation du nombre de clients *A*, la priorité probabiliste a plus de clients dont les objectifs sont très élevés.

Nous devons signaler que plusieurs couples (p_A, p_B) peuvent aboutir à la même valeur de μ^*. Dans le tableau, nous affichons $p_A = 1$ à chaque fois que cela correspond à un optimum. Ainsi, les lignes du tableau dans lesquelles p_A est différente de 1 veulent dire que ce n'est pas possible d'obtenir l'optimum avec $p_A = 1$. D'après le Tableau 6.1, nous constatons que, p_A ne peut pas être égale à 1 lorsque l'objectif des clients *B* est élevé ou encore, lorsque la proportion des clients *A* est importante. Ceci veut dire que, dans ces deux cas, il faut que certains clients de classe *A* soient moins prioritaires pour permettre aux clients de classe *B* d'atteindre les objectifs.

Après ces interprétations des résultats, nous pouvons constater le caractère "non égalitaire" de la discipline de priorité probabiliste. En effet, pour atteindre les objectifs de qualité de service, il suffit parfois de faire passer des clients par la file la moins prioritaire pour qu'ils y attendent longtemps. Un autre point qui confirme ce caractère vient du non respect de l'ordre de l'arrivée au sein d'une même classe de clients. Ainsi, un client *A* peut être servi avant un autre de la même classe et qui est arrivée avant lui. Pour les clients *B*, il est possible d'avoir le même phénomène. Pour un critère de niveau de service qui concerne la proportion de clients répondus en moins d'un certain délai, ce non respect de l'ordre d'arrivée facilite l'amélioration du système. Pour un critère de service basé sur des temps moyens d'attente, comme nous le verrons plus tard, ceci n'aura aucune influence.

Nous fixons, à présent, les niveaux de service objectifs des deux classes de clients. La proportion de clients de classe *A* satisfaits en moins de 20 secondes, doit être supérieure à 95 %. Celle des clients de classe *B* ne doit pas descendre en dessous de 85 %. Le Tableau 6.2 compare la priorité probabiliste avec la priorité stricte pour la qualité de service mentionnée.

λ_A/λ	Priorité stricte	Priorité probabiliste			Gain
	μ^*	μ^*	p_A	p_B	
0,1	106,11	100	1	0,9	5,75%
0,2	106,81	100,2	0,98	0,87	6,19%
0,3	107,68	100,7	0,98	0,87	6,48%
0,4	108,77	101,3	0,98	0,864	6,87%
0,5	110,17	102	0,98	0,855	7,42%
0,6	112,00	102,7	0,97	0,841	8,30%
0,7	114,27	103,5	0,97	0,83	9,43%
0,8	117,15	104,3	0,96	0,814	10,97%
0,9	120,63	105,3	0,95	0,79	12,71%

Tableau 6.2: Comparaison de la priorité stricte avec la priorité probabiliste pour la qualité de service objectif 95|85

Dans le Tableau 6.2, la qualité de service a été fixée pour chaque classe de clients. Le paramètre variable est la répartition des arrivées en clients de classes *A* et *B*. Ainsi, en faisant varier le rapport λ_A / λ entre 0,1 et 0,9, nous comparons les taux de service optimaux qui nous assurent la satisfaction des objectifs de niveau de service déjà fixés. En réalité, ils est rare que les centres d'appels aient plus d'arrivées de la part de clients de classe *A* que de clients de classe *B*. Cependant, ce cas de figure n'est pas complètement fictif puisqu'il y a des entreprises qui doivent externaliser une partie des appels et qu'elles préfèrent garder les clients de classe *A* dans leurs centres d'appels. Ceci peut aboutir à une proportion des clients de classe *A* supérieure à 50 %.

Comme ce que nous avons fait dans le Tableau 6.1, le μ^* relatif à la priorité stricte implique une sur-qualité pour l'une des deux classes alors que pour l'autre classe, la qualité de service est obtenue sans être dépassée. Pour la priorité probabiliste, le μ^* assure la satisfaction des objectifs sans la moindre sur-qualité puisque le niveau de service atteint pour chaque classe est exactement égal à ce qui est recherché. Le gain obtenu grâce à la probabilité probabiliste est donné par la dernière colonne du tableau. L'évolution de ce gain en fonction de la proportion des arrivées de classe *A* est illustrée par la Figure 6.3 qui confirme ce que nous avions déjà constaté auparavant.

Effectivement, la compétitivité de la priorité probabiliste augmente en fonction du taux d'arrivée des clients de classe *A* pour un taux d'arrivée total qui reste constant. Nous remarquons, en particulier, que l'amélioration de la performance du système est très nette lorsque les appels de classe *A* sont importants. Ainsi, le μ qui nous permet de satisfaire les objectifs fixés dans un système de priorité probabiliste peut avoir une valeur de 13 % inférieure au taux de service satisfaisant les mêmes objectifs dans un système à priorité stricte.

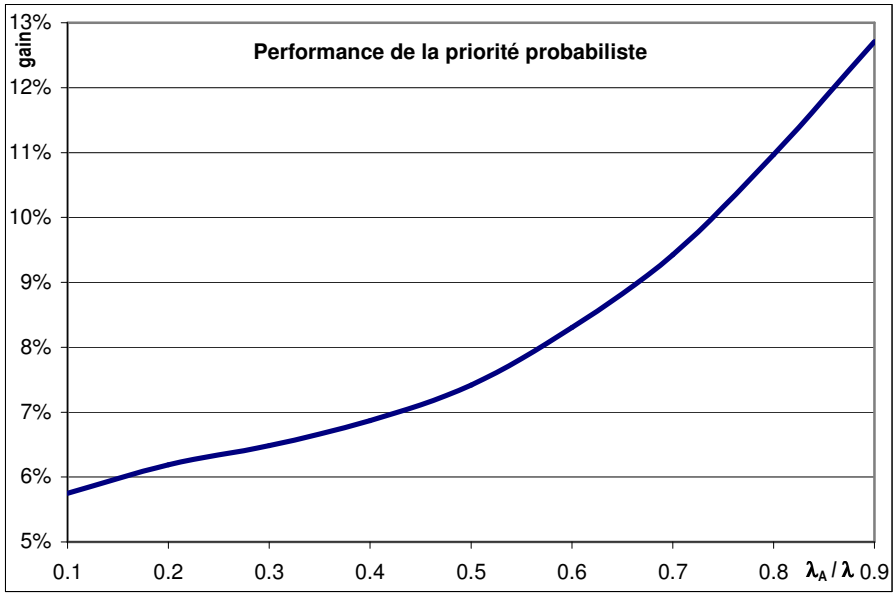


Figure 6.3: Comparaison de la priorité probabiliste avec la probabilité stricte pour la qualité de service objectif 95|85

6.3.1.2 Temps moyens d'attente

Nous allons à présent continuer à traiter la même discipline de priorité probabiliste de la section précédente. Toutefois, les objectifs à satisfaire ne sont plus exprimés en proportions de clients à servir en moins de vingt secondes mais, plutôt, en temps moyens d'attente à ne pas dépasser. Bien entendu, le temps moyen d'attente objectif de la classe *A* est inférieur à celui de la classe *B*.

Pour les objectifs considérés maintenant, nous allons proposer un calcul analytique du taux de service optimal de la priorité probabiliste ainsi que des probabilités p_A et p_B

qui le permettent. Nous montrons, également, que la priorité probabiliste procure la meilleure performance possible et que, de ce fait, un changement de la discipline de priorité ne peut plus améliorer le taux de service optimal.

En faisant l'analogie du système de la priorité probabiliste avec le système modélisé par la Figure 6.2, nous commençons par écrire les temps moyens d'attente des classes 1 et 2 relatives à cette figure. La calcul de ces temps moyens est un résultat connu, nous pouvons le trouver, par exemple, dans Gross et Harris [33]. Les temps moyens d'attente sont donnés par les formules suivante :

$$Wq_1 = \frac{\rho}{\mu - \lambda_1} \tag{6.5}$$

$$Wq_2 = \frac{\rho}{(1-\rho)(\mu - \lambda_1)} \tag{6.6}$$

avec $\lambda_1 = p_A \lambda_A + p_B \lambda_B$ et $\rho = (\lambda_A + \lambda_B) / \mu$.

À partir des équations (6.5) et (6.6), nous pouvons déduire les temps moyens d'attente des classes *A* et *B*. Le principe du calcul est semblable à ce qui a été effectué pour la détermination de la distribution du temps d'attente de chaque classe dans la section précédente. Ce calcul est réalisé grâce aux deux formules suivantes :

$$Wq_A = p_A \ \mathrm{Wq}_1 + (1 - p_A) \ \mathrm{Wq}_2 \tag{6.7}$$

$$Wq_B = p_B \ \mathrm{Wq}_1 + (1 - p_B) \ \mathrm{Wq}_2 \tag{6.8}$$

Les deux formules précédentes nous permettent de déterminer le temps moyen d'attente de chaque classe de clients. En exploitant les formules (6.5) et (6.6), nous pouvons les transformer de la manière suivante :

$$Wq_A = \frac{\rho(\mu - p_A\lambda)}{(\mu - p_A\lambda_A - p_B\lambda_B)(\mu - \lambda)} \tag{6.9}$$

$$Wq_B = \frac{\rho(\mu - p_B\lambda)}{(\mu - p_A\lambda_A - p_B\lambda_B)(\mu - \lambda)} \tag{6.10}$$

Ces deux dernières équations assurent le calcul des temps moyens d'attente suivant les paramètres du système p_A, p_B, λ_A, λ_B et μ. Le taux de service optimal μ^* de la discipline de priorité probabiliste peut être déterminé, comme lors de la section précédente, de façon numérique. Ainsi, il correspond au taux de service le plus faible qui permet de ne pas dépasser les temps d'attente objectif Wq_A^{obj} et Wq_B^{obj}. Les probabilités statiques p_A et p_B peuvent, à nouveau, déterminées numériquement.

Avant de commenter une quelconque procédure de calcul numérique, nous commençons par limiter notre champ d'investigation. La proposition suivante donne une borne inférieure au taux de service optimal. Signalons, ici, que pour chaque discipline de priorité, un taux de service optimal correspond à un critère de service donné. Une discipline de priorité optimale pour ce critère est une discipline qui garantit le meilleur taux de service optimal.

Proposition 1 *Dans un système mono-serveur à deux classes de clients et avec des objectifs sur les temps moyens d'attente tels que décrits dans la section précédente, une borne inférieure au taux de service optimal est la suivante :*

$$\mu_{\inf}^{*} = \frac{\lambda}{2}\left(1+\sqrt{1+\frac{4}{\lambda_A Wq_A^{obj} + \lambda_B Wq_B^{obj}}}\right) \tag{6.11}$$

Preuve: Considérons une discipline de priorité quelconque, de préférence conservatrice de travail puisque c'est plus réaliste. Nous pouvons alors écrire le temps moyen d'attente, toutes classes confondues avec la relation suivante :

$$Wq = \frac{\lambda_A}{\lambda} Wq_A + \frac{\lambda_B}{\lambda} Wq_B \tag{6.12}$$

Or, le temps d'attente moyen de chacune des deux classes ne doit pas dépasser l'objectif correspondant qui est égal à Wq_A^{obj} pour la classe *A* et Wq_B^{obj} pour la classe *B* :

$$\begin{cases} Wq_A \le Wq_A^{obj} \\ Wq_B \le Wq_B^{obj} \end{cases}$$

Ceci nous permet de déduire que le temps moyen d'attente toutes classes confondues ne doit pas dépasser Wq^{obj} :

$$Wq \leq Wq^{obj} \tag{6.13}$$

avec

$$Wq^{obj} = \frac{\lambda_A}{\lambda} Wq_A^{obj} + \frac{\lambda_B}{\lambda} Wq_B^{obj} \tag{6.14}$$

Or, sans distinguer les deux classes de clients, le système peut être considéré comme un modèle M/M/1 où le temps moyen d'attente avant service peut être écrit comme suit :

$$Wq = \frac{\rho}{\mu - \lambda} \tag{6.15}$$

En écrivant $\rho = \lambda / \mu$, ceci revient à écrire :

$$Wq = \frac{\lambda}{\mu^2 - \lambda\mu}$$

ce qui aboutit à l'équation du second degré en μ :

$$Wq\ \mu^2 - \lambda\ Wq\ \mu\ -\ \lambda = 0 \tag{6.16}$$

dont la résolution donne :

$$\mu = \frac{\lambda}{2} + \frac{1}{2}\sqrt{\lambda^2 + 4\frac{\lambda}{Wq}} \tag{6.17}$$

Sachant que plus le taux de service μ diminue, plus le temps moyen d'attente Wq augmente. En obéissant à l'inégalité (6.13), nous pouvons dire que Wq ne peut pas dépasser Wq^{obj} et que, inversement, μ ne peut descendre en dessous de la valeur limite qui aboutit à Wq^{obj}. Cette valeur étant donnée par l'équation (6.17), nous pouvons aboutir au résultat de la proposition donné par l'équation (6.11).

La proposition précédente permet d'avoir une borne inférieure au taux de service optimal. Même si cette borne n'est pas nécessairement atteinte, elle nous donne, néanmoins, une idée de la performance de la discipline de priorité que nous avons considérée. Ainsi, si l'écart entre le taux optimal μ^* de la discipline considérée et la borne inférieure est faible, alors la priorité traitée peut être jugée comme efficace.

Le théorème qui suit démontre que la borne inférieure donnée par la proposition 1 représente le taux de service optimal de la discipline de la priorité probabiliste.

Théorème 1 *La discipline de la priorité probabiliste traitée dans cette section offre la meilleure capacité de service optimale lorsque les critères de service objectif pour chaque classe de clients sont exprimés en termes de temps moyens d'attente. Le taux de service optimal* μ^* *correspondant ainsi que les paramètres de la discipline de priorité sont les suivants :*

$$\mu^* = \frac{\lambda}{2}\left(1+\sqrt{1+\frac{4}{\lambda_A Wq_A^{obj}+\lambda_B Wq_B^{obj}}}\right) \tag{6.18}$$

$$p_A = 1 \tag{6.19}$$

$$p_B = \left(1-\frac{\mu^*}{\lambda}\right)\frac{Wq_B^{obj}}{Wq_A^{obj}}+\frac{\mu^*}{\lambda} \tag{6.20}$$

Preuve: Considérons l'équation (6.9) qui exprime le temps moyens d'attente de la classe *A*. Dans cette équation, nous fixons les paramètres μ, p_A et p_B aux valeurs données par les équations (6.18) à (6.20). Ceci nous donne :

$$Wq_A = \frac{\lambda}{\mu^2-\lambda_A\mu-\lambda_B\left(\mu-\frac{\mu^2}{\lambda}\right)\frac{Wq_B^{obj}}{Wq_A^{obj}}-\frac{\mu^2}{\lambda}\lambda_B}$$

cette équation se transforme en :

$$Wq_A = \frac{\lambda\ Wq_A^{obj}}{\mu^2\left(Wq_A^{obj}+\frac{\lambda_B}{\lambda}Wq_B^{obj}-\frac{\lambda_B}{\lambda}Wq_A^{obj}\right)-\mu\left(\lambda_A Wq_A^{obj}+\lambda_B Wq_B^{obj}\right)}$$

En écrivant $\lambda = \lambda_A + \lambda_B$, cette dernière équation devient :

$$Wq_A = \frac{\lambda\ Wq_A^{obj}}{\mu^2\ Wq^{obj}-\mu\ \lambda\ Wq^{obj}}$$

Finalement, en remarquant que si μ est donné par la formule (6.18), alors il est, nécessairement, solution de l'équation (6.16), nous obtenons :

$$Wq_A = Wq_A^{obj}$$

Avec un raisonnement similaire, nous obtenons :

$$Wq_B = Wq_B^{obj}$$

Étant donné que les objectifs de niveau de service sont atteints lorsque le taux de service est égal à la borne inférieure de la Proposition 1, nous pouvons affirmer que cette borne représente une solution réalisable et que la priorité probabiliste avec les paramètres du Théorème 1 est une priorité optimale pour des objectifs basés sur les temps d'attente puisqu'il n'est pas possible d'atteindre ces objectifs avec un taux de service plus bas et ce, quelque soit la priorité considérée.

Le théorème précédent nous confirme qu'avec les paramètres adéquats, la priorité probabiliste aboutit à une priorité optimale, lorsque les objectifs fixés sont exprimés en temps moyens d'attente. En fait, les couples (p_A, p_B) qui engendrent le taux de service optimal μ* sont innombrables. Nous avons opté pour celui qui nous semble le plus logique et suivant lequel, tous les clients de classe *A* passent par la file la plus prioritaire. Ce cas de figure n'est pas exact lorsque les objectifs s'expriment en proportions de clients ayant attendu moins de 20 secondes. En effet, pour certains niveaux de service, il serait préférable de faire passer quelques clients de classe *A* par la file la moins prioritaire pour que le reste des clients de cette classe soit plus privilégié.

Nous passons, à présent, à la comparaison de la priorité probabiliste avec la priorité stricte, à l'instar de l'étude que nous avons menée dans la section précédente. Nous allons fixer le temps moyen objectif de la classe *A* à 5 secondes et celui de la classe *B* à 10 secondes. Afin d'effectuer la comparaison, nous devons, d'abord, déterminer le valeur du taux de service optimal qui correspond à la discipline de la priorité stricte. En fait, lorsque le taux de service est égal à cette valeur, cela signifie que l'une des deux classes possède un temps moyen d'attente égal à l'objectif. Le niveau de service fourni à l'autre classe est supérieur à l'objectif. Grâce aux équations (6.5) et (6.6), nous pouvons déduire le taux de service optimal de la priorité stricte :

$$\mu^*_{stricte} = Max\left[\frac{1}{2}\left(\lambda_A + \sqrt{\lambda_A^2 + \frac{4\lambda}{Wq_A^{obj}}}\right), \frac{1}{2}\left(\lambda + \lambda_A + \sqrt{\lambda_B^2 + \frac{4\lambda}{Wq_B^{obj}}}\right)\right] \qquad (6.21)$$

Le Tableau 6.3 affiche l'évolution des taux de service optimaux pour chaque type de priorité pour des objectifs de 5 secondes pour la classe *A* et 10 secondes pour la classe *B*.

λ_A/λ	Priorité stricte	Priorité probabiliste			Gain
	μ^*	μ^*	p_A	p_B	
0,1	106,24	105,96	1	0,94	0,26 %
0,2	106,90	106,27	1	0,94	0,59 %
0,3	107,72	106,62	1	0,93	1,02 %
0,4	108,73	107,01	1	0,93	1,58 %
0,5	110	107,45	1	0,93	2,32 %
0,6	111,62	107,94	1	0,92	3,30 %
0,7	113,72	108,51	1	0,91	4,59 %
0,8	116,46	109,16	1	0,91	6,27 %
0,9	120	109,92	1	0,90	8,40 %

Tableau 6.3: Comparaison de la priorité stricte avec la priorité probabiliste pour la qualité de service objectif 5 s|10 s

Dans le Tableau 6.3, nous pouvons constater que le gain de la priorité probabiliste par rapport à la priorité stricte croît en fonction de la proportion de clients de classe *A* dans le système. Ce gain est quasiment nul pour un système qui reçoit, uniquement, 10 % des appels de la part de la classe *A*. Lorsque le système reçoit plus de 70 % d'appels de la part des clients *A*, nous pouvons nous attendre à un gain supérieur à 5 % et qui peut dépasser les 8 %.

Dans la Figure 6.4, nous avons comparé les deux types de priorité suivant plusieurs objectifs de niveau de service. Nous pouvons y constater que la priorité probabiliste se distingue, spécialement, lorsque la qualité de service objectif qui correspond à la classe *B* est élevée. Ceci s'explique par le fait que la priorité stricte satisfait en premier lieu la

classe *A* et que, dans ce cas, il ne lui est pas facile d'atteindre des niveaux de service élevés pour la classe *B*. Ce comportement du système sous la priorité stricte rend la migration vers un type de priorité probabiliste particulièrement intéressante dans le cas où, les qualités de service objectif des deux classes sont assez proches l'une de l'autre. Cet intérêt s'accentue encore plus lorsque la proportion des clients de classe *A* dans le système est plus importante que celle des clients de classe *B*.

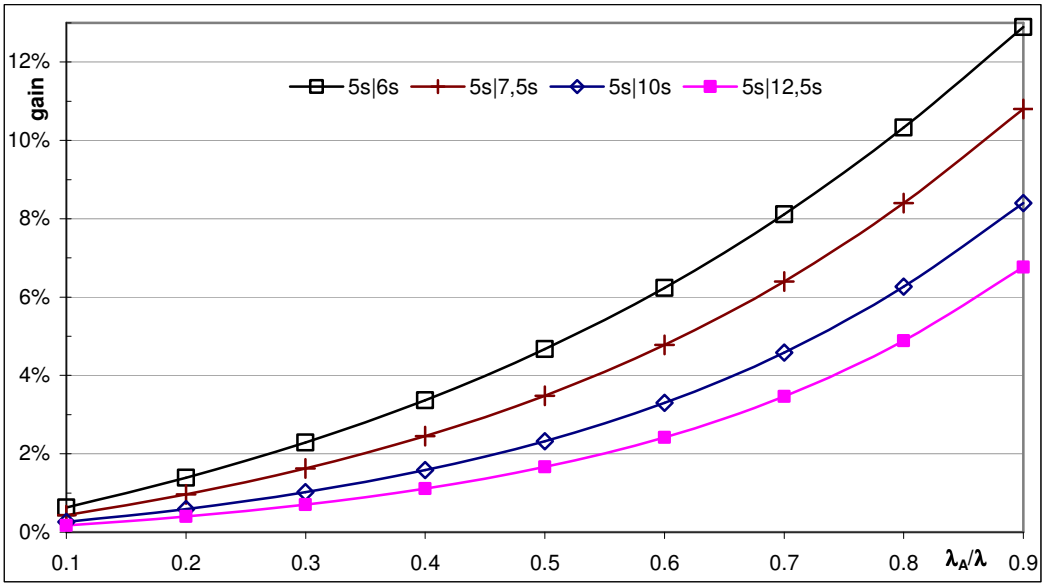


Figure 6.4: Comparaison de la priorité probabiliste avec la probabilité stricte pour différentes qualités de service objectif

Il est intéressant de remarquer que, même si la priorité probabiliste ne conserve pas l'ordre d'arrivée au sein d'une même classe, il est, néanmoins, possible d'aboutir à la même performance qu'elle offre tout en conservant l'ordre des arrivées pour une classe donnée. Le problème ne se pose pas pour les clients de la classe *A* puisqu'ils sont servis suivant la discipline fcfs. Par contre, en ce qui concerne la classe *B*, pour conserver l'ordre d'arrivée des clients sans altérer les performances du système, il suffit d'imposer à tous les clients *B* de passer par la file la moins prioritaire. Ainsi, lorsqu'un client de classe *B* arrive et que la priorité probabiliste l'affecte à la file la plus prioritaire, le système doit imposer son routage par la file la moins prioritaire si celle-ci n'est pas vide. Dans ce cas, le client de classe *B* en tête de la file de 2ème priorité doit être déplacé en dernière position de la file 1. Si la priorité affecte le client à la file la moins prioritaire

alors rien ne sera changé. Avec ces modifications de la priorité probabiliste, le système devient plus juste vis à vis des clients. Étant donné qu'un changement de positions entre des clients de même classe dans la file n'influe pas sur le temps moyen d'attente de cette classe, nous pouvons conclure que les modifications effectuées gardent les mêmes performances que la discipline de priorité probabiliste.

6.3.1.3 Généralisation au cas multi-classe

Dans cette section, nous allons considérer, uniquement, des objectifs de niveau de service exprimés en temps moyens d'attente. L'extension aux distributions du temps d'attente est semblable du point de vue méthodologie. Toutefois, la méthode se base de nouveau sur un calcul numérique comme cela l'a déjà été dans le cas de deux classes de clients.

Par le terme "multi-classe", nous désignons des systèmes intégrant plus de 2 classes de clients puisque le cas de 2 classes a déjà été traité. Nous allons nous intéresser, surtout, à l'étude du système mono-serveur avec trois classes de clients. Plus tard, la généralisation, pour plusieurs classes, peut être effectuée de façon similaire. La Figure 6.5 représente le système avec trois classes de clients. Nous allons montrer par la suite que la discipline de la priorité probabiliste offre, encore, la meilleure capacité de service lorsqu'il s'agit d'une optimisation multicritère basée sur des temps moyens d'attente objectif.

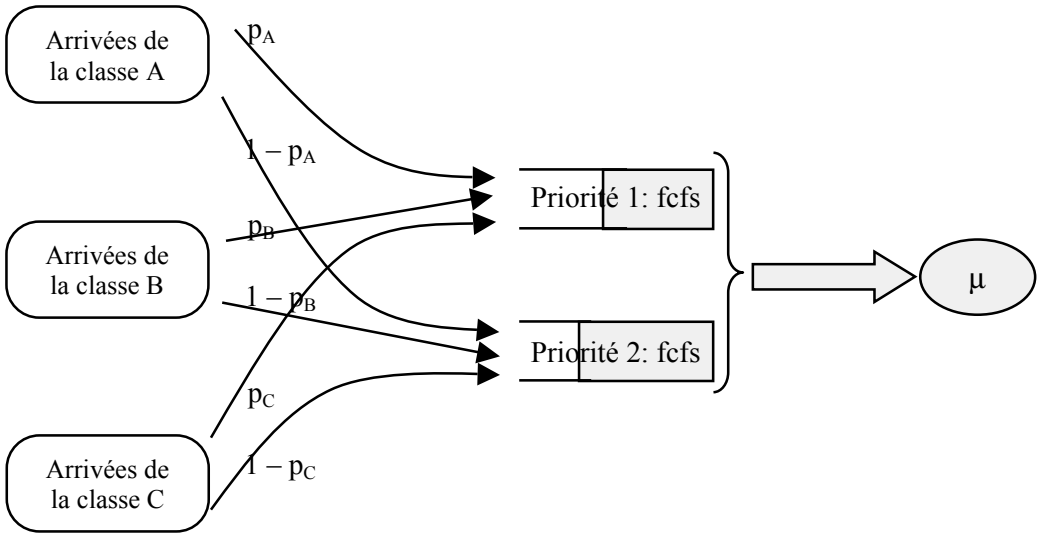


Figure 6.5: Priorité probabiliste avec trois classes de clients

Théorème 2 *En notant les objectifs du temps moyen d'attente des classes A, B et C respectivement par W_{qA}^{obj}, W_{qB}^{obj} et W_{qC}^{obj}, la priorité probabiliste offre la meilleure capacité de service pour le système de la Figure 6.5, si cette discipline est régie par les paramètres suivants :*

$$p_A = 1 \tag{6.22}$$

$$p_B = \frac{p_{B2}\left(\lambda_B + \lambda_C\right)\left(Wq_A^{obj} - Wq_B^{obj}\right) + \lambda_C\left(Wq_B^{obj} - Wq_C^{obj}\right)}{\left(\lambda_B + \lambda_C\right)Wq_A^{obj} - \lambda_B\ Wq_B^{obj} - \lambda_C\ Wq_C^{obj}} \tag{6.23}$$

$$p_C = \frac{p_{B2}\left(\lambda_B + \lambda_C\right)\left(Wq_A^{obj} - Wq_C^{obj}\right) + \lambda_B\left(Wq_C^{obj} - Wq_B^{obj}\right)}{\left(\lambda_B + \lambda_C\right)Wq_A^{obj} - \lambda_B\ Wq_B^{obj} - \lambda_C\ Wq_C^{obj}} \tag{6.24}$$

Avec ces paramètres, le taux de service optimal est donné par :

$$\mu^* = \frac{\lambda}{2}\left(1 + \sqrt{1 + \frac{4}{\lambda_A Wq_A^{obj} + \lambda_B Wq_B^{obj} + \lambda_C Wq_C^{obj}}}\right) \tag{6.25}$$

Le paramètre p_{B2} est donné par la formule suivante :

$$p_{B2} = \left(1 - \frac{\mu^*}{\lambda}\right)\frac{\frac{\lambda_B}{\lambda}Wq_B^{obj} + \frac{\lambda_C}{\lambda}Wq_C^{obj}}{Wq_A^{obj}} + \frac{\mu^*}{\lambda} \tag{6.26}$$

Preuve: Le taux de service optimal de l'équation (6.25) peut être montré comme étant la borne inférieure avec un raisonnement similaire à celui de la Proposition 1. Par la suite, en utilisant les paramètres dans le Théorème 2, nous pouvons montrer que le μ* est atteint. Ceci aboutit à la démonstration du théorème.

Nous allons maintenant expliquer le raisonnement qui a conduit aux paramètres des formules (6.22) à (6.26). Le même raisonnement peut être utilisé pour déterminer les paramètres qui généralisent la discipline de priorité à un système comportant plus de 3 classes de clients.

La probabilité p_A peut toujours être fixée à 1. En fixant le taux de service μ à la valeur μ^* donnée par l'équation (6.25), nous nous assurons que le temps moyen d'attente, toutes classes confondues, est égal à la moyenne pondérée des objectifs fixés avec des poids proportionnels aux arrivées de chaque classe. Ensuite, les deux paramètres p_B et p_C vont imposer le temps moyen d'attente de chaque classe. S'il y a une sur-qualité pour une classe donnée, ceci se traduit, nécessairement, par un objectif non atteint pour, au moins, une classe. Cependant, le fait que le comportement du temps moyen d'attente de chaque classe en fonction de p_B et de p_C soit monotone implique l'existence de couples (p_B, p_C) qui assurent la satisfaction des objectifs sans la moindre sur-qualité.

Si nous fusionnons les classes *B* et *C*, nous savons que le p_B de l'équation (6.20) implique la satisfaction d'un temps moyen d'attente égal à Wq_A^{obj} pour la classe *A* et égal à la moyenne pondérée des objectifs Wq_B^{obj} et Wq_C^{obj} pour la classe fusionnée. Le p_B de l'équation (6.20) sera noté, ainsi, p_{B2} et il s'écrit suivant la formule (6.26).

Une première relation entre p_B et p_C peut être trouvée en remarquant que le taux d'arrivée à la file de priorité 1 doit être le même avec et sans fusion des classes *B* et *C* pour que le niveau de service de la classe *A* ne soit pas modifié. Ceci ce traduit par l'équation suivante :

$$p_B\lambda_B + p_C\lambda_C = p_{B2}\left(\lambda_B + \lambda_C\right) \tag{6.27}$$

Les temps moyens d'attente des classes *A*, *B* et *C* s'écrivent en fonction des temps moyens d'attente des clients qui passent par les files de priorités 1 et 2 et ce, suivant les équations suivante :

$$Wq_A = Wq_1 \tag{6.28}$$

$$Wq_B = p_B\ Wq_1 + \left(1 - p_B\right)\ Wq_2 \tag{6.29}$$

$$Wq_C = p_C\ Wq_1 + \left(1 - p_C\right)\ Wq_2 \tag{6.30}$$

À partir des équations (6.27) à (6.30), nous déduisons les valeurs des paramètres telles que énoncées par le théorème. La détermination des paramètres dans un système composé de plus de 3 classes de clients s'effectue de la même manière et aboutit, également, à la meilleure capacité de service.

Avec le Théorème 2, nous avons les paramètres qui assurent la capacité de service minimale pour un système à trois classes de clients. Comme cela a été le cas pour deux classes de clients, la probabilité pour qu'un client de classe *A* rejoigne la file la plus prioritaire a été fixée à 1.

6.3.2 Système multi-serveur

Nous considérons à présent un système composé de *C* serveurs. Les clients qui contactent le centre d'appels sont divisés en deux classes. La classe *A* sera celle à qui est attribuée la qualité de service objectif la plus élevée. L'autre classe est noté, comme précédemment, classe *B*. Le taux de service moyen μ est supposé constant, ce qui est représentatif de la réalité. Dans cette section, nous cherchons à étudier le nombre optimal de serveurs *C** qui assure la satisfaction des objectifs fixés préalablement. Les objectifs vont, de nouveau, être distingués en deux catégories.

6.3.2.1 Proportion d'appels répondus en moins de vingt secondes

L'étude que nous menons ici est similaire à ce qui a été réalisé dans le cas mono-serveur. Nous allons considérer un temps de service moyen de 3 minutes, ce qui correspond à un taux de service moyen $\mu = 0{,}33$ appels / minute, représentatif de la réalité. La Figure 6.6 schématise le système que nous analysons ici.

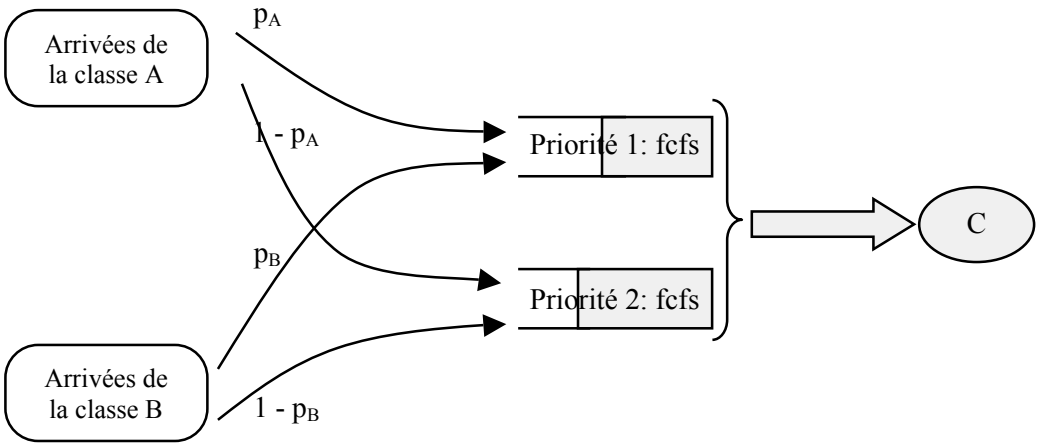


Figure 6.6: Priorité probabiliste en multi-serveur

La distribution des temps d'attente des classes *A* et *B* est donnée par les formules suivantes :

$$Wq_A(t) = p_A \ \mathrm{Wq}_1(t) + (1 - p_A) \ \mathrm{Wq}_2(t) \tag{6.31}$$

$$Wq_B(t) = p_B \ \mathrm{Wq}_1(t) + (1 - p_B) \ \mathrm{Wq}_2(t) \tag{6.32}$$

où $Wq_1(t)$ et $Wq_2(t)$ représentent les distributions des classes dans une file de priorité stricte :

$$Wq_1(t) = 1 - P_d \ e^{(\lambda_1 - C\mu)\, t} \tag{6.33}$$

$$Wq_2(t) = \begin{cases} 1 - \dfrac{P_d(1-\rho)}{2\ \pi\ \rho} \displaystyle\int_{\alpha_2}^{\beta_2} f_2(r)\ dr & si\ \rho^2 < \rho_1 \\ 1 - \dfrac{P_d(\rho^2 - \rho_1)}{\rho(\rho - \rho_1)} e^{-\gamma_2 . t} - \dfrac{P_d(1-\rho)}{2\ \pi\ \rho} \displaystyle\int_{\alpha_2}^{\beta_2} f_2(r)\ dr & \sin on \end{cases} \tag{6.34}$$

avec P_d la probabilité d'attente dans une file M/M/C définie par :

$$P_d = \left(1 + \frac{C!(1-\rho)}{(\rho\ C)^C} \sum_{n=0}^{C-1} \frac{(\rho\ C)^n}{n!} \right)^{-1} \tag{6.35}$$

Le reste des paramètres de l'équation (6.34) sont donnés par le chapitre précédent. Il faut juste fixer le nombre de classes *n* à 2.

Les équations (6.31) et (6.32) nous permettent de déterminer, pour toutes valeurs p_A et p_B, la proportion d'appels ayant attendu moins d'un certain temps, que nous fixons à 20 secondes. Par une procédure numérique simple, nous déterminons le nombre optimal de serveurs qui nous permet d'atteindre des objectifs préalablement définis. Ce calcul est plus facile à effectuer que pour un système mono-serveur. En effet, le nombre de serveurs est discret, ce qui diminue les itérations nécessaires.

Le Tableau 6.4 compare, dans le cas multi-serveur, la discipline de la priorité probabiliste avec celle de la priorité stricte. Dans ce tableau, le gain de la priorité probabiliste par rapport à la priorité stricte ne varie pas de façon importante en fonction de la proportion des clients de classe *A* dans le système.

λ_A/λ	Priorité stricte	Priorité probabiliste			Gain
	C*	C*	p_A	p_B	
0,1	311	301	1	0,9	3,22 %
0,2	312	301	1	0,85	3,53 %
0,3	312	301	0,99	0,87	3,53 %
0,4	313	302	0,99	0,86	3,51 %
0,5	314	303	0,99	0,85	3,50 %
0,6	315	304	0,99	0,83	3,49 %
0,7	316	305	0,98	0,8	3,48 %
0,8	317	306	0,97	0,78	3,47 %
0,9	318	307	0,95	0,75	3,46 %

Tableau 6.4: Comparaison de la priorité stricte avec la priorité probabiliste pour la qualité de service objectif 95|85

À partir du Tableau 6.4, nous remarquons, également, que le gain de la priorité probabiliste en multi-serveur reste plus faible que ce que nous obtenons dans le cas mono-serveur. En effet, plus le nombre de serveurs augmente, plus les performances de la priorité stricte s'améliorent. Nous pouvons expliquer ceci par le fait que la probabilité d'attente diminue lorsque le nombre de serveurs augmente. Dans ce cas, étant donné que la plupart des clients arrivent à joindre directement un conseiller sans la moindre attente et ce, quelque soit leurs classes, alors les objectifs deviennent plus facilement accessibles même avec une discipline de priorité stricte. La Figure 6.7 montre l'allure de la probabilité d'attente en fonction du nombre de serveurs. Pour cette figure, le taux d'arrivée λ a été fixé à 100 appels par minute. La charge ρ du système est, elle, fixée à 95 %. Ainsi, pour chaque valeur de *C*, le taux de service est calculé de telle sorte à avoir une charge constante. La probabilité d'attente P_d est calculée à partir de l'équation (6.35) relative à une file M/M/C. La Figure 6.7 nous montre que, pour un nombre important de serveurs et à une utilisation moyenne constante, il est plus facile d'accéder directement au service que lorsque le système est mono-serveur. Ceci nous renseigne de l'étroitesse de la marge de manœuvre qui nous reste pour améliorer la priorité stricte.

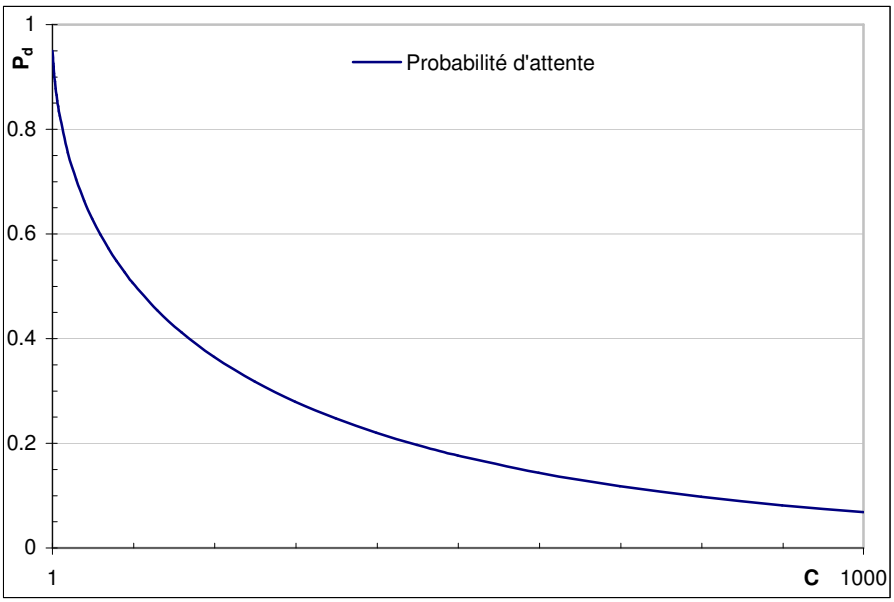


Figure 6.7: Évolution de la probabilité d'attente en fonction du nombre de serveurs

6.3.2.2 Temps moyens d'attente

Comme cela a été montré dans le cas mono-serveur, nous pouvons démontrer, également, que la discipline de la priorité probabiliste offre les meilleures performances lorsqu'elle est bien paramétrée. Ceci est relatif à des objectifs exprimés en temps moyens d'attente. Cela étant, il n'est plus possible d'avoir de formules analytiques ici parce que tout s'exprime en fonction de la probabilité d'attente P_d dont l'expression écarte presque tout calcul analytique exact. Ainsi, le calcul que nous avons mené dans le but de déterminer le nombre optimal de serveurs qui nous assure la satisfaction des objectifs, est basé sur des procédures numériques. Néanmoins, nous sommes toujours assurés qu'à l'optimum, il est possible de fixer la probabilité p_A à 1. Les procédures numériques nécessaires à la détermination de la probabilité p_B et du nombre optimal de serveurs C*, sont basées sur les formules qui suivent :

$$Wq_A = \mathrm{Wq}_1 \tag{6.36}$$

$$Wq_B = p_B \; Wq_1 + (1 - p_B) \; Wq_2 \tag{6.37}$$

Wq_A et Wq_B représentent, respectivement, les temps moyens d'attente des clients de classes *A* et *B*. Dans ces formules, Wq_1 et Wq_2 sont les temps moyens d'attente des

clients qui passent, respectivement, par les files de priorités 1 et 2. Les valeurs de ces paramètres ont été calculées par Kella et Yechiali [45]. Elles s'écrivent en fonction de la probabilité d'attente P_d de la manière suivante :

$$Wq_1 = \frac{P_d}{C\,\mu - \lambda_1} \tag{6.38}$$

$$Wq_2 = \frac{P_d}{(1-\rho)(C\,\mu - \lambda_1)} \tag{6.39}$$

où λ_1 et λ_2 représentent les taux d'arrivée respectifs aux files de priorités 1 et 2.

Le Tableau 6.5 compare la priorité probabiliste avec la priorité stricte pour des objectifs de temps moyens d'attente de 5 secondes pour la classe *A* et de 10 secondes pour la classe *B*. Ce tableau nous montre, à nouveau, que la discipline de la priorité stricte se comporte mieux que pour le cas d'objectifs exprimés en proportions d'appels répondus en moins de 20 secondes. L'explication vient du fait que la non conservation de l'ordre des arrivées au sein d'une même classe, n'améliore pas les performances de la priorité probabiliste. Le gain de la discipline reste inférieur au cas mono-serveur à cause d'une probabilité d'attente qui diminue en fonction du nombre de serveurs et qui, de ce fait, réduit la proportion de clients affectés par un quelconque changement de la discipline de priorité. À noter, finalement, que pour la première ligne du tableau, les deux discipline aboutissent au même nombre optimal de serveurs. Ceci est dû, uniquement, à l'effet discret puisque le C* ne peut prendre que des valeurs entières.

λ_A/λ	Priorité stricte	Priorité probabiliste			Gain
	C*	C*	p_A	p_B	
0,1	310	310	1	0	0,00%
0,2	311	310	1	0,9	0,32%
0,3	311	310	1	0,96	0,32%
0,4	312	311	1	0,9	0,32%
0,5	313	311	1	0,94	0,64%
0,6	314	311	1	0,96	0,96%
0,7	316	312	1	0,95	1,27%
0,8	318	312	1	0,96	1,89%
0,9	321	313	1	0,95	2,49%

Tableau 6.5: Comparaison de la priorité stricte avec la priorité probabiliste pour la qualité de service objectif 5 s|10 s

6.4 Deuxième modèle de la priorité probabiliste

Dans cette partie, nous étudions un deuxième modèle de priorité probabiliste. Dans ce modèle, chaque classe de clients admet sa propre file d'attente. Nous allons considérer, uniquement, le cas de deux classes de clients. Lorsqu'un client arrive, s'il n'a pas la possibilité de joindre immédiatement un conseiller, alors il doit patienter dans la file qui correspond à sa classe. Lorsqu'un conseiller se libère, si une seule file est non vide, alors il répond au premier client de cette file qui fonctionne selon la règle fcfs. Si par, contre, les deux files sont non vides, alors le conseiller répond au premier client de la file *A* avec une probabilité statique *p*. Le premier client de la file *B* est répondu avec la probabilité 1-p.

6.4.1 Système mono-serveur

Nous considérons, dans cette section, que le système se compose d'un seul serveur dont le taux de service constitue un paramètre que l'on peut contrôler et choisir. À chaque fin de service, le serveur regarde en premier la file d'attente de la classe *A* avec une probabilité *p*. Avec la probabilité complémentaire, de serveur regarde d'abord la file de la classe *B*. Dans ce système, l'ordre des arrivées au sein d'une même classe est respecté puisque chaque classe possède sa propre file d'attente fonctionnant suivant la règle fcfs. Le système ainsi modélisé, est illustré par la Figure 6.8.

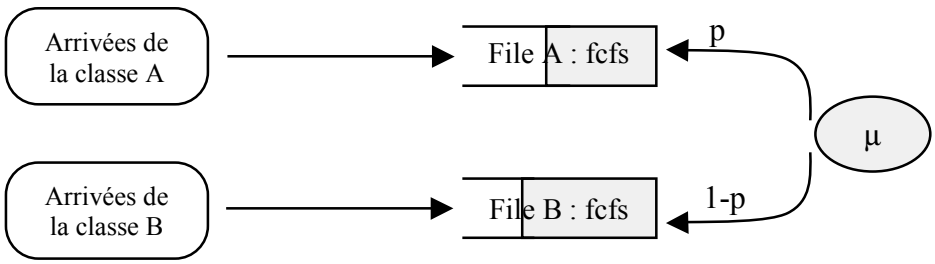


Figure 6.8: File d'attente avec la deuxième priorité probabiliste

Dans la suite, nous allons nous focaliser sur les deux types d'objectifs que nous avons considérés auparavant: les proportions d'appels répondus en moins de 20 secondes et les temps moyens d'attente pour chaque classe de clients. Étant donnée la

complexité du système actuel, l'étude sera, cette fois-ci, basée sur des simulations. Une étude numérique avec une chaîne de Markov à deux dimensions aboutirait au même résultat que la simulation lorsque l'on cherche à satisfaire un objectif exprimé en temps moyens d'attente.

6.4.1.1 Proportion d'appels répondus en moins de vingt secondes

Les niveaux de service objectif sont exprimés en termes de proportions de clients répondus en moins de 20 secondes. Nous cherchons à déterminer le taux de service optimal μ^* qui nous assure la satisfaction des objectifs fixés. Le Tableau 6.6 compare la deuxième discipline de la priorité probabiliste avec la discipline de la priorité stricte. Dans ce tableau, l'objectif fixé pour la classe *A* est de satisfaire, au moins, 95 % des clients en moins de 20 secondes. Pour la classe *B*, la proportion à satisfaire en moins de 20 secondes est égale à 85 %. Le taux d'arrivée des clients au système est $\lambda = 100$ appels par minute.

λ_A/λ	Priorité stricte	Priorité probabiliste		Gain
	μ^*	μ^*	p	
0,1	106,11	106	0,18	0,10%
0,2	106,81	106,2	0,26	0,57%
0,3	107,68	106,5	0,35	1,10%
0,4	108,77	106,9	0,44	1,72%
0,5	110,17	107,4	0,53	2,51%
0,6	112	107,9	0,62	3,66%
0,7	114,27	108,4	0,71	5,14%
0,8	117,15	109	0,8	6,96%
0,9	120,63	109,3	0,88	9,39%

Tableau 6.6: Comparaison de la priorité stricte avec la priorité probabiliste pour la qualité de service objectif 95|85

Dans le Tableau 6.6, la compétitivité de la discipline de la priorité probabiliste augmente par rapport à la priorité stricte en fonction de la proportion des clients de

classe *A* dans le système. Cependant, en comparant les gains obtenus avec ce que donne le Tableau 6.2, nous constatons que la deuxième discipline de la priorité probabiliste est moins performante que la première lorsqu'il s'agit de regarder la proportion des clients servis en moins de 20 secondes. Ceci s'explique par la simple raison que la première discipline ne conserve pas l'ordre des arrivées appartenant à une même classe. De ce point de vue, la deuxième discipline est plus réaliste.

La Figure 6.9 affiche l'évolution de la probabilité *p* qui conduit au taux de service optimal μ* et ce, en fonction de la proportion de clients de classe *A* dans le système. Nous remarquons que l'allure de la courbe ainsi obtenue est linéaire, ce qui peut s'avérer particulièrement intéressant si la pente peut être déterminée.

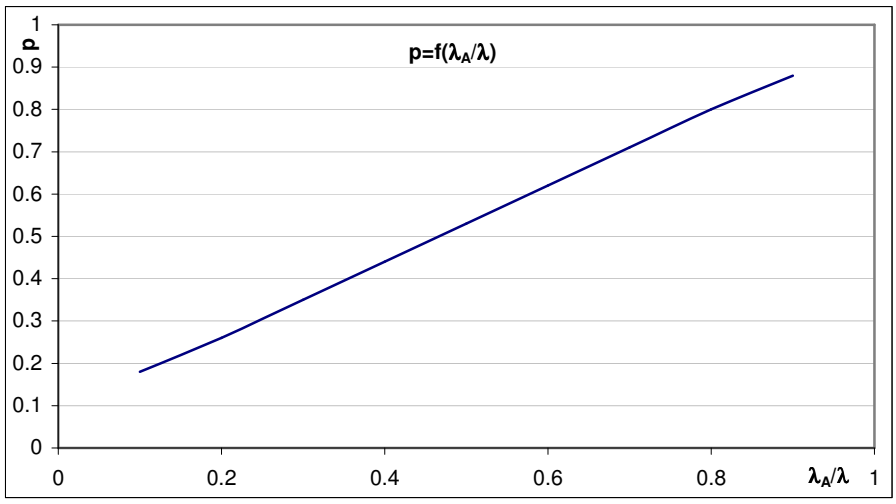


Figure 6.9: Évolution de la probabilité *p* en fonction de la proportion des clients de classe *A*

Il est clair que, pour les critères de qualité de service étudiés dans cette section, la deuxième priorité probabiliste ne peut pas aboutir à la capacité de service minimale. Il est intéressant de savoir si elle s'en approche ou pas. La proposition suivante donne une borne inférieure au taux de service minimal qui peut aboutir à la satisfaction des objectifs.

Proposition 2 *Lorsque les objectifs sont exprimés en termes de proportions de clients répondus en moins d'un temps t, une borne inférieure* μ *au taux de service optimal, dans un*

système qui conserve l'ordre des arrivées au sein d'une même classe, est la solution de l'équation suivante :

$$\frac{e^{(\lambda-\mu)\,t}}{\mu} = \frac{1-Wq_B^{obj}(t)}{\lambda} \tag{6.40}$$

Preuve : Soit une discipline de priorité quelconque. Commençons par fixer le taux d'arrivée λ au système. Notons par μ_P le taux de service optimal correspondant à la discipline considérée. Avec ce taux de service, les niveaux de service des classes *A* et *B* sont, au moins, égaux aux objectifs fixés. Puisque la classe *A* admet un objectif de niveau de service plus élevé que la classe *B*, nous pouvons dire que, si la proportion des arrivées de la classe *A* augmente, alors le taux de service optimal va augmenter à cause du remplacement d'une partie des clients non exigeants (clients *B*) par des clients qui le sont (clients *A*).

De cette manière, le taux de service optimal dans le cas où la proportion des appels de classe *A* est égale à zéro, est inférieur ou égal au taux de service optimal pour n'importe quelle proportion. Pour calculer cette borne inférieure, nous devons donc déterminer le taux de service optimal dans un système recevant des appels de classe *B* exclusivement. Puisque nous limitons notre recherche aux disciplines qui conservent l'ordre d'arrivée au sein d'une même classe, le système en question est nécessairement une file M/M/1. Pour cette file, la distribution du temps d'attente est donnée par Gross et Harris [33] :

$$Wq(t) = 1 - \rho\, e^{(\lambda-\mu)\,t} \tag{6.41}$$

Il suffit, par la suite, de mettre $Wq(t) = Wq_B^{obj}(t)$ pour aboutir au résultat.

Dans la Figure 6.10, nous utilisons la méthode de calcul de la borne inférieure et nous comparons l'écart entre son résultat et le taux de service optimal donné par la priorité probabiliste que nous analysons. Nous constatons que l'écart maximum entre les deux atteint 3,5 %. En fait, la performance optimale se situe entre les deux. Ceci implique un écart maximum, entre la priorité probabiliste et la borne inférieure, qui soit moins important que 3,5 %. À noter ici que la même méthode de calcul de la borne inférieure pour des objectifs exprimés en temps moyens d'attente donne, quasiment, les mêmes écarts par rapport à la 1ère priorité probabiliste alors que nous avons prouvé l'optimalité de la priorité pour ce type de mesure de niveau de service. Ceci laisse

envisager, dans le cas de cette section, que la performance optimale est beaucoup plus proche de la 2ème priorité probabiliste que de la borne inférieure calculée.

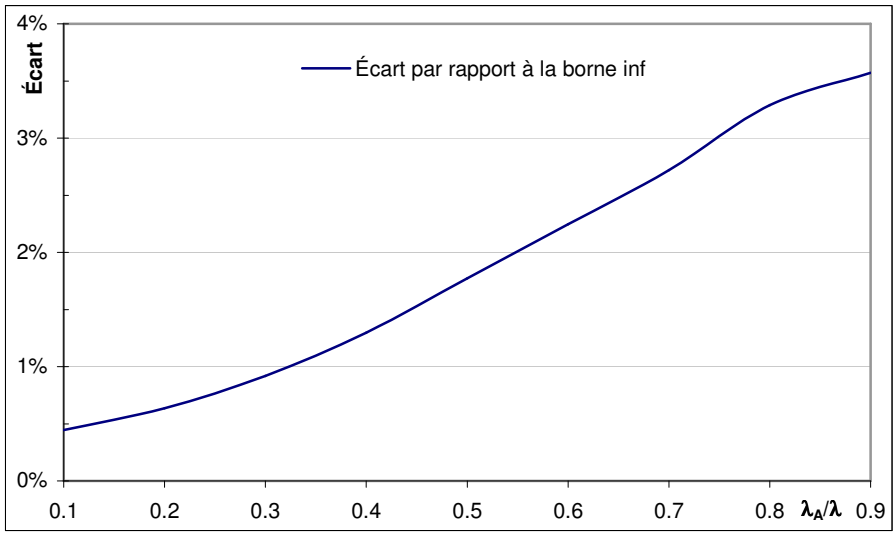


Figure 6.10: Écart de la 2ème priorité probabiliste par rapport à la borne inférieure

6.4.1.2 Temps moyens d'attente

Nous passons à présent à des objectifs qui s'expriment, pour chaque classe de clients, en temps moyens d'attente avant service. Nous avons pu, précédemment dans ce chapitre, déterminer complètement la première priorité probabiliste qui amène au taux de service optimal et ce, de façon analytique. Nous avons, également, montré que cette priorité correspond à la meilleure performance. Maintenant, étant donné que notre analyse est basé sur des simulations, nous ne comptons pas déterminer analytiquement la probabilité p qui optimise le taux de service. Toutefois, nous allons montrer que la deuxième priorité probabiliste offre, également, la meilleure capacité de service pour les critères de service objectif.

Théorème 3 *La deuxième priorité probabiliste correspond à la capacité de service minimale lorsque les objectifs sont exprimés en temps moyens d'attente. Le taux de service optimal est donné par la formule (6.18) issue du premier type de priorité probabiliste.*

Preuve: Commençons par fixer le taux de service à la valeur donnée par la formule (6.18). Puisque la discipline de priorité traitée ici est du type "conservatrice de travail", nous pouvons dire que le taux de service considéré permet d'avoir un temps moyen d'attente Wq, tous clients confondus, suivant l'équation :

$$Wq = \frac{\lambda_A}{\lambda} Wq_A^{obj} + \frac{\lambda_B}{\lambda} Wq_B^{obj} \tag{6.42}$$

Ceci est vrai quelque soit la valeur de la probabilité *p*. Cependant, c'est cette probabilité qui répartie les temps moyens d'attente entre les classes *A* et *B*. En effet, si p = 0, alors cela revient à affecter une priorité stricte à la classe *B*. Ceci nous donne une première borne des temps moyens d'attente des deux classes. Pour p = 1, c'est la classe *A* qui se voit attribuer la priorité stricte, d'où la deuxième borne. Étant donnée la continuité des temps moyen d'attente en fonction de *p*, nous pouvons déduire que le temps moyen d'attente de chaque classe peut prendre n'importe quelle valeur située entre les 2 bornes. Grâce à cette continuité, il existe une valeur particulière de *p* qui engendre un temps moyen d'attente pour la classe *A* égal à Wq_A^{obj}. Et puisque la somme pondérée des temps moyens d'attente de chaque classe obéit, nécessairement à l'équation (6.42), nous pouvons affirmer que, pour cette valeur de *p*, le temps moyen d'attente pour la classe B est égal à Wq_B^{obj}. D'où le résultat du théorème. Le taux de service optimal est, donc, donné par l'équation (6.18).

Avec le Théorème 3, nous pouvons déterminer le taux de service optimal pour le deuxième type de probabilité probabiliste, même si la probabilité *p* n'est pas définie. En reprenant l'exemple du Tableau 6.3, nous pouvons, cette fois, déduire que le gain est du même ordre que celui de la section précédente où les objectifs s'expriment en proportion de clients servis en moins de 20 secondes. Ceci s'explique par le fait que la discipline de priorité étudiée conserve, désormais, l'ordre des arrivées au sein d'une même classe.

6.4.2 Système multi-serveur

Nous revenons maintenant à l'étude du système multi-serveur. Ce qui nous intéresse ce n'est plus la détermination du taux de service optimal mais plutôt, celle du nombre

de serveurs optimal. Nous distinguons, à nouveau, deux mesures de niveaux de service.

6.4.2.1 Proportion d'appels répondus en moins de vingt secondes

Le système que nous étudions est illustré par la Figure 6.11. Les objectifs de chaque classe sont fixés en termes de probabilité d'être répondu en moins de 20 secondes.

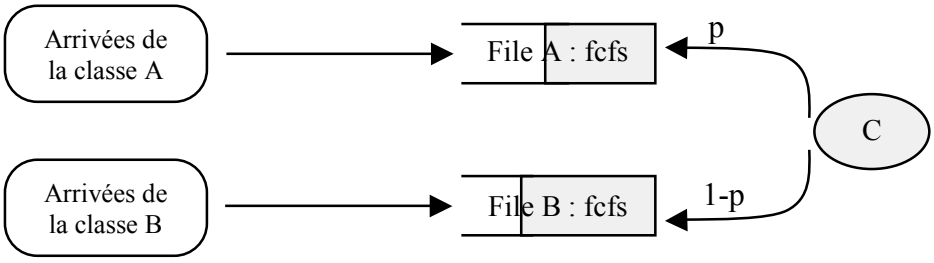


Figure 6.11: File d'attente avec la 2ème priorité probabiliste en multi-serveur

Les résultats obtenus par simulation sont affiché par le Tableau 6.7. Nous y remarquons que le gain devient très faible. Dans le meilleur des cas, il dépasse à peine 1 %. Ceci est dû, essentiellement, à la diminution de la probabilité d'attente P_d. Avec une étude similaire à celle de la Proposition 2, nous obtenons des écarts entre la priorité probabiliste et la borne obtenue, qui restent inférieurs à 1 %. Ceci confirme que la diminution de la proportion des clients qui attendent, améliore sensiblement la performance du système. Notons, finalement, l'allure de nouveau linéaire de p en fonction de la proportion des arrivées de la classe A.

λ_A/λ	Priorité stricte	Priorité probabiliste		Gain
	C*	C*	p	
0,1	311	311	0,15	0,00%
0,2	312	311	0,25	0,32%
0,3	312	311	0,34	0,32%
0,4	313	311	0,44	0,64%
0,5	314	312	0,53	0,64%
0,6	315	312	0,63	0,95%
0,7	316	313	0,72	0,95%
0,8	317	314	0,81	0,95%
0,9	318	314	0,91	1,26%

Tableau 6.7: Comparaison de la priorité stricte avec la 2ème priorité probabiliste pour la qualité de service objectif 95|85

6.4.2.2 Temps moyens d'attente

Lorsque l'optimisation du système part de niveaux de service cibles mesurés en temps moyens d'attente, nous pouvons, via une démonstration similaire à celle du Théorème 3, prouver que la 2ème priorité probabiliste correspond à des performances optimales. Ainsi, le nombre optimal de serveurs est égal à ce qui a été trouvé avec la première priorité probabiliste. Cependant, le paramètre p de la discipline ne peut être trouvé que par des calculs numériques basés sur une chaîne de Markov à deux dimensions ou encore à l'aide de simulations. L'allure de la courbe représentative de p en fonction de la proportion des appels de clients de classe A est, encore une fois, linéaire.

6.5 Conclusions

Nous avons analysé, au-cours de ce chapitre, deux disciplines de priorités alternatives à la priorité stricte qui est la discipline la plus répandue. Les deux disciplines analysées sont basées sur l'utilisation de probabilités statiques dans le but de ne plus privilégier complètement une classe au détriment des autres clients. En faisant de la sorte, le système offre un service différencié à chaque classe de clients. Notre but a été d'étudier et de montrer le gain que procurent les disciplines de priorité probabiliste en termes de capacité de service nécessaire à la satisfaction d'objectifs de niveaux de service pour chaque classe.

Lors de notre analyse du premier type de priorité probabiliste, nous avons montré que, malgré le fait que les probabilités utilisées soient statiques et que la discipline ne tient pas compte de l'état du système, la règle de priorité correspond à la capacité de service minimale lorsqu'il s'agit de ne pas dépasser des temps moyens d'attente fixés pour chaque classe. Nous avons, également, montré que le gain que l'on peut escompter, diminue en fonction du nombre de serveurs. Ainsi, indépendamment du taux de service, moins le système comporte de serveurs, plus la priorité probabiliste s'avère compétitive face à la discipline stricte. En ce qui concerne des objectifs exprimés en pourcentages de clients servis en moins d'un certain temps, la priorité étudiée n'offre pas la meilleure performance. Toutefois, le gain obtenu est encore plus important qu'avec la précédente mesure de niveau de service. Ceci est dû, surtout, au

non respect, par la discipline, de l'ordre d'arrivée au sein d'une même classe. Et c'est là que réside le point faible de la méthode. Pour conserver les mêmes performances lors de la satisfaction de moyennes du temps d'attente, nous avons proposé une variante de la discipline qui conserve, cette fois, l'ordre d'arrivée des clients appartenant à une même classe.

Le deuxième type de priorité probabiliste nous a permis d'avoir une discipline qui conserve l'ordre d'arrivée dans une même classe. Nous avons montré, pour des objectifs qui s'expriment en temps moyens d'attente, que la discipline correspond, elle aussi, à la meilleure performance. Ainsi, la capacité de service optimale peut être déterminée à partir des études effectuées avec la première priorité puisqu'elle est plus facile à analyser analytiquement et numériquement. Cependant, ceci ne nous permet pas de connaître la probabilité p qui définit la discipline. Nous avons constaté, par des exemples numériques, que cette probabilité varie d'une façon linéaire en fonction de la proportion des arrivées des clients de classe A. Cela peut permettre, dans l'avenir, de déterminer la probabilité qui aboutit à la capacité de service optimale. Dans ce cas, l'étude du deuxième type de priorité pourra être étendue au cas multi-classe en se basant, comme lors de l'extension de la première priorité, sur les paramètres correspondants au cas où le système ne comporte que deux classes de clients.

Chapitre 7 : Conclusions et Perspectives

7.1 Conclusions

Dans ce mémoire, nous avons commencé par étudier le phénomène de rappels et ses répercussions sur le dimensionnement du centre d'appels. Nous avons, d'abord, expliqué l'importance que peut avoir ce phénomène en montrant que le taux d'appels observé par le système peut comporter une grande proportion de renouvellements d'appels. Ceci a été effectué par l'intermédiaire d'une chaîne de Markov à deux dimensions. Le calcul des probabilités des états au régime stationnaire nous a permis, en particulier, d'observer et de comprendre le comportement du système en fonction de tous les paramètres qui le définissent. La résolution de la chaîne de Markov au régime stationnaire nous a permis, également, de développer une procédure récursive qui détermine le taux d'appels frais en fonction des appels observés. Ceci nous a aidés à réaliser le dimensionnement du système qui nous permet d'atteindre un taux de prise en charge objectif. En comparant les résultats avec un système qui ne tient pas compte des rappels, nous avons constaté que cette non-considération peut impliquer un sur-dimensionnement ou bien un sous-dimensionnement du système, ce qui engendre, naturellement, des objectifs non réalisés, dans certains cas, et largement satisfaits, dans d'autres cas. Dans tous les cas, la non-considération du phénomène de rappels induit à la non satisfaction des managers du centre d'appels, que ce soit à cause des objectifs qui ne sont pas satisfaits, ou à cause du surcoût de fonctionnement. L'approximation "*square root rule*", connue pour ses bonnes performances lors du dimensionnement du

système lorsqu'il n'y a pas de rappels, a été comparée avec notre modèle. Le résultat montre que, pour un système avec rappels, elle peut engendrer des erreurs de surdimensionnement importantes. Par la suite, nous sommes passés à l'étude d'un système où une annonce du temps d'attente est fournie aux clients à chaque arrivée. Avec une telle annonce, nous supposons, à l'instar de Whitt [70], que le renoncement immédiat des clients augmente et que, parallèlement, le taux d'abandon de la file diminue. Ce modèle, plus générique, est résolu grâce à la méthode de l'approximation fluide. Nous aboutissons, au régime stationnaire, à une formule, à la fois, simple et robuste. Elle est insensible à la variation de plusieurs paramètres du système, en particulier la fonction de renoncement qui peut être assez compliquée. La précision de la méthode augmente avec la charge du système et sa taille. L'approximation fluide nous a, également, permis d'aborder le régime transitoire pour étudier une journée entière où les paramètres varient en fonction du temps et ce, en considérant qu'elle se compose de plusieurs périodes enchaînées. La résolution de ce système aboutit à la détermination de l'évolution des appels frais en fonction du temps.

Nous avons ensuite étudié les temps d'attente des clients dans le centre d'appels. Maintenant, nous considérons un système composé de plusieurs classes de clients avec des priorités strictes non-préemptives. En premier lieu, nous avons proposé un estimateur, appelé ASA^2, du temps d'attente des clients à leurs arrivées. Nous avons montré que la précision de cet estimateur augmente en fonction de la charge du système. De toute façon, lorsque la charge du système n'est pas importante, même si la précision diminue, la proportion des clients concernés est faible puisque la plupart d'eux sont servis immédiatement. À noter, également, que la précision de l'estimateur proposé est meilleure pour un système où toutes les files sont fusionnées. En second lieu, nous avons analysé plusieurs règles de routage basées sur l'ASA^2, dans le cas où chaque site possède sa propre file d'attente. En fixant un objectif de niveau de service à atteindre pour chaque classe de clients, nous montrons qu'un routage dynamique peut, parfois, permettre d'avoir un système plus performant que celui de la file logique, la performance étant mesurée en termes de conseillers nécessaires à atteindre les objectifs fixés. Nous expliquons ceci par le fait que la file logique, où la règle de priorité est stricte, n'atteint pas facilement les objectifs fixés pour la classe la moins prioritaire. Parallèlement, la règle de priorité dans un système de plusieurs files d'attente ne peut plus être considérée comme stricte à cause du routage. L'influence du routage sur la

règle de priorité améliore la qualité de service relative à la classe la moins prioritaire, ce qui peut améliorer la performance globale du système. Ceci est à l'origine de l'étude sur les priorités probabilistes. En effet, nous avons montré que, même avec des priorités qui changent de façon statique et non dynamique, il est possible d'améliorer le rendement du système et d'aboutir à une capacité de service plus intéressante. Nous avons montré que la capacité de service, ainsi obtenue, est la capacité de service optimale lorsqu'il s'agit de mesurer le niveau de service en termes de temps moyens d'attente. Nous avons calculé cette capacité de service optimale dans le cas d'un serveur unique. Toutefois, il faut signaler que le passage d'une priorité stricte à une priorité probabiliste est plus avantageux dans le cas d'un système mono-serveur. Le gain croît également en fonction de la proportion de la classe la plus prioritaire dans le système.

7.2 Perspectives

Nous avons démontré l'existence d'une forte interaction entre le dimensionnement du centre d'appels et les rappels qu'il enregistre. Ainsi, si le système est sous-dimensionné, alors les rappels vont augmenter. Nous avons déjà traité le dimensionnement du système dans un cas mono-période, il est intéressant d'étendre l'utilisation de l'approximation fluide au système multi-périodes afin de trouver une planification optimale des conseillers sur l'ensemble de la journée dans un modèle avec rappels et abandons. Notons que la formule qui donne le taux de rappels au régime stationnaire n'est pas utilisable lorsque la charge est inférieure à 1, même si, dans ce cas, le taux stationnaire de rappel n'est pas important, surtout pour une taille de système importante. La seule approximation de la formule réside dans le remplacement du nombre moyen de serveurs occupés par le nombre total de serveurs. Il serait intéressant de coupler la formule avec une procédure d'approximation du nombre moyen de serveurs occupés.

Lors de la mise au point de l'ASA^2, nous avons tenu compte des paramètres disponibles en temps réel. Nous avons également vu que la précision de l'estimateur croît en fonction de la charge du système. La coordination entre l'ASA^2 et une éventuelle méthode d'approximation de la charge du système améliorerait, certainement, la précision de l'estimation du temps d'attente. Cette amélioration

pourrait faire profiter la règle de routage basée sur l'ASA[2], même s'il vaut mieux disposer d'une information en temps réel sur la disponibilité d'un conseiller libre sur chaque site.

Dans l'étude de la deuxième priorité probabiliste, nous avons montré que la capacité de service fournie par la discipline est optimale, lorsque les objectifs s'expriment en temps moyens d'attente. Cette capacité de service peut être calculée d'une manière analytique. Cependant, nous n'avons pas pu calculer, analytiquement, la probabilité qui caractérise la discipline et ce, tout en observant son évolution linaire en fonction de la proportion des clients les plus prioritaires dans le système. La détermination d'une procédure de calcul simplifiée, pour cette probabilité, permet d'étendre la méthode, comme ce fut le cas pour la première règle de priorité probabiliste, à plus de deux classes de clients. Il est à signaler que les paramètres du système changent, généralement, au-cours du temps. Pour les deux disciplines étudiées, nous déterminons les paramètres au régime stationnaire. Une possible extension au cas multi-périodes mérite d'être étudiée.

Bibliographie

[1] Adan, I. J. B. F., Boxma, O. J. et Resing, J. A. C. Queueing models with multiple waiting lines. *Queueing Systems*, vol. 37, 65-98, 2001.

[2] Aguir, M. S., Chauvet, F., Dallery, Y., Karaesmen, F., Nait-Abdallah, R. M. et Prat, T. Solution pour l'estimation avec intervalle de confiance du temps d'attente en vue de l'annoncer au client. *Institut National de la Propriété Industrielle*, France, 03/50965, 2003.

[3] Aguir, M. S., Chauvet, F., Dallery, Y., Karaesmen, F., Nait-Abdallah, R. M. et Prat, T. Solution pour l'estimation du temps d'attente moyen dans un centre d'appels. *Institut National de la Propriété Industrielle*, France, 03/01270, 2003.

[4] Aguir, M. S., Karaesmen, F., Akşin, O. Z. et Chauvet, F. The impact of retrials on call center performance. *OR Spectrum*, à paraitre, 2004.

[5] Aguir, M. S., Karaesmen, F., Akşin, Z. et Chauvet, F. Analyse du problème des rappels et dimensionnement dans un centre d'appels. *MOSIM'03*, 2003.

[6] Akşin, O. Z. et Harker, P. T. Analysis of a processor shared loss system. *Management Science*, vol. 47, 324-336, 2001.

[7] Altinkemer, K., Bose, I. et Pal, R. Average waiting time of customers in an M/D/k queue with nonpreemptive priorities. *Computers Ops Res.*, vol. 25, 317-328, 1998.

[8] Artalejo, J. R. A queueing system with returning customers and waiting line. *Operations Research Letters*, vol. 17, 191-199, 1995.

[9] Babaï, M. Z. *Modélisation et évaluation des performances de centres d'appels téléphoniques*, Mémoire de DEA, Labo LGI, École Centrale Paris, 2002.

[10] Baccelli, F. et Hebuterne, G. On queues with impatient customers. *Performance'81*, 159-179, 1981.

[11] Bertsimas, D. An exact FCFS waiting time analysis for a general class of G/G/s queueing systems. *Queueing Systems*, vol. 3, 305-320, 1988.

[12] Bhulai, S. et Koole, G. A queueing model for call blending in call centers. *IEEE Transactions on Automatic Control*, vol. 48, 1434-1438, 2003.

[13] Boxma, O., Koole, G. et Liu, Z. *Queueing-theoretic solution methods for models of parallel and distributed systems*, Performance Evaluation of Parallel and Distributed Systems - Solution Methods, CWI, 1994.

[14] Boxma, O. J. in *Recent Trends in Optimization Theory and Applications* (ed. Agarwal, R. P.) (World Scientific Publishing Company, 1995).

[15] Boxma, O. J. et de Waal, P. R. Multiserver queues with impatient customers. *The Fundamental Role of Teletraffic in the Evolution of Telecommunications Networks (Proc. ITC-14)*, 743-756, 1994.

[16] Boxma, O. J. et Down, D. G. Dynamic server assignment in a two-queue model. *European Journal of Operational Research*, vol. 103, 595-609, 1997.

[17] Brandt, A. et Brandt, M. On the M(n)/M(n)/s queue with impatient calls. *Performance Evaluation*, vol. 35, 1-18, 1999.

[18] Caïazzo, B. *Les centres d'appels*, edt. Dunod, 2001.

[19] Choi, B. D. et Chang, Y. Single server retrial queues with priority calls. *Mathematical and Computer Modelling*, vol. 30, 7-32, 1999.

[20] Choi, B. D., Kim, B. et Chung, J. M/M/1 Queue with impatient customers of higher priority. *Queueing Systems*, vol. 38, 49-66, 2001.

[21] Davis, R. H. Waiting-time distribution of a multi-serve, priority queueing system. *Operations Research*, vol. 14, 133-136, 1966.

[22] De Véricourt, F. et Zhou, Y. Managing response time and service quality in a call routing problem. *Working Paper, Fuqua School of Business, Duke University*, 2003.

[23] Falin, G. Estimation of retrial rate in a retrial queue. *Queueing Systems*, vol. 19, 231-246, 1995.

[24] Falin, G. I. et Templeton, J. G. C. *Retrial Queues*, Chapman & Hall, 1997.

[25] Feng, W., Kowada, M. et Adachi, K. Analysis of a multi-server queue with two priority classes and (M,N)-threshold service schedule I: non-preemptive priority. *Intl. Trans. in Op. Res.*, vol. 7, 653-671, 2000.

[26] Gail, H. R., Hantler, S. L. et Taylor, B. A. Analysis of a non-preemtive priority multiserver queue. *Adv. Appl. Prob.*, vol. 20, 852-879, 1988.

[27] Gans, N., Koole, G. M. et Mandelbaum, A. Telephone Call Centers: Tutorial, Review and Research Prospects. *Manufacturing & Service Operations Management*, vol. 5, 79-141, 2003.

[28] Gans, N. et Zhou, Y.-P. A call-routing problem with service-level constraints. *Operations Research*, vol. 51, 255-271, 2003.

[29] Garnett, O., Mandelbaum, A. et Reiman, M. Designing a call center with impatient customers. *Manufacturing & Service Operations Management*, vol. 4, 208-227, 2002.

[30] Green, L. V. et Kolesar, P. J. The Lagged PSA for estimating peak congestion in multiserver markovian queues with periodic arrival rates. *Management Science*, vol. 43, 80-87, 1997.

[31] Green, L. V., Kolesar, P. J. et Svoronos, A. Some effects of nonstationarity on multiserver markovian queueing systems. *Operations Research*, vol. 39, 502-511, 1991.

[32] Greenberg, B. S. et Wolff, R. W. An upper bound on the performance of queues with returning customers. *J. Appl. Prob*, vol. 24, 466-475, 1987.

[33] Gross, D. et Harris, C. M. *Fundamentals of Queueing Theory*, edt. Wiley, New York, 1998.

[34] Halfin, S. et Whitt, W. Heavy-traffic limits for queues with many exponentiel servers. *Operations Research*, vol. 29, 567-587, 1981.

[35] Hoffman, K. L. et Harris, C. M. Estimation of a caller retrial rate for a telephone information system. *European Journal of Operational Research*, vol. 27, 207-214, 1986.

[36] Hordijk, A. et Koole, G. On the optimality of the generalized shortest queue policy. *Probability in the Engineering and Informational Sciences*, vol. 4, 477-487, 1990.

[37] Houck, D. J. Comparaison of policies for routing customers to parallel queueing systems. *Operations Research*, vol. 35, 306-310, 1987.

[38] Huang, T.-Y. Analysis and modeling of a threshold based priority queueing system. *Computer Communications*, vol. 24, 284-291, 2001.

[39] Hui, M. K. et Tse, D. K. What to tell consumers in waits of different lengths: an integrative model of service evaluation. *Journal of Marketing*, vol. 60, 81-90, 1996.

[40] Jaiswal, N. K. *Priority queues*, edt. Press, A., New York, 1968.

[41] Jennings, O. B., Mandelbaum, A., Massey, W. A. et Whitt, W. Server staffing to meet time-varying demand. *Management Science*, vol. 42, 1383-1394, 1996.

[42] Jiang, Y., Tham, C.-K. et Ko, C.-C. A probabilistic priority scheduling discipline for multi-service networks. *Computer Communications*, vol. 25, 1243-1254, 2002.

[43] Kao, E. P. C. et Narayanan, K. S. Computing steady-state probabilities of a nonpreemptive priority multiserver queue. *ORSA Journal on Computing*, vol. 2, 211-218, 1990.

[44] Kao, E. P. C. et Wilson, S. D. Analysis of nonpreemptive priority queues with multiple servers and two priority classes. *European Journal of Operational Research*, vol. 118, 181-193, 1999.

[45] Kella, O. et Yechiali, U. Waiting times in the non-preemptive priority M/M/c queue. *Commun. Statist.-Stochastic Models*, vol. 1, 257-262, 1985.

[46] Kimura, T. Approximations for the delay probability in the M/G/s queue. *Mathl. Comut. Modelling*, vol. 22, 157-165, 1995.

[47] Kimura, T. Equivalence relations in the approximations for the M/G/s/s+r queue. *Mathematical and Computer Modelling*, vol. 31, 215-224, 2000.

[48] Kleinrock, L. *Queueing Systems, Volume I: Theory*, edt. Wiley, 1975.

[49] Lee, D. C., Park, S. J. et Song, J. S. Performance analysis of queueing strategies for multiple priority calls in multiserver personal communications services. *Computer Communications*, vol. 23, 1069-1083, 2000.

[50] Lee, H. W., Seo, W. J. et Yoon, S. H. An analysis of multiple-class vacation queues with individual thresholds. *Operations Research Letters*, vol. 28, 35-49, 2001.

[51] Leemans, H. Waiting time distribution in a two-class two-server heterogeneous priority queue. *Perfomance Evaluation*, vol. 43, 133-150, 2001.

[52] Maglaras, C. et Armony, M. On customer contact centers with a call-back option: customer decisions, routing rules, and system design. *to appear in Operations Research*, 2003.

[53] Mandelbaum, A., Massey, W. A., Reiman, M. I. et Rider, B. Time varying multiserver queues with abandonments and retrials. *The 16th International Teletraffic Conference*, 1999.

[54] Masi, D. M. B., Fischer, M. J. et Harris, C. M. Computation of steady state probabilities for resource sharing call center queueing systems. *Stochastic Models*, vol. 17, 191-214, 2001.

[55] Medhi, J. *Stochastic models in queueing theory*, edt. Press, A., Boston, 1991.

[56] Miller, D. R. Computation of steady-state probabilities for M/M/1 priority queues. *Operations Research*, vol. 29, 945-958, 1981.

[57] Nakibly, E. Predicting waiting times in telephone service systems. Technion Institute of Technology, Haifa, 2002.

[58] Neuts, M. F. *Matrix-Geometric Solutions in Stochastic Models - An Algorithmic Approach*, ed. Press, T. J. H. U., Baltimore, Maryland, 1981.

[59] Neuts, M. F. et Rao, B. M. Numerical investigation of a multiserver retrial model. *Queueing Systems*, vol. 7, 169-190, 1990.

[60] Rothkopf, M. H. et Rech, P. Perspectives on queues: combining queues is not always benefical. *Operations Research*, vol. 35, 906-909, 1987.

[61] Takagi, H. *Queueing analysis, vol. 1*, , North-Holland, 1991.

[62] Tham, C.-K., Yao, Q. et Jiang, Y. Achieving differentiated services through multi-class probabilistic priority scheduling. *Computer Networks*, vol. 40, 577-593, 2002.

[63] Tham, C.-K., Yao, Q. et Jiang, Y. A multi-class probabilistic priority scheduling discipline for differentiated services networks. *Computer Communications*, vol. 25, 1487-1496, 2002.

[64] Tran-Gia, P. et Mandjes, M. Modeling of customer retrial phenomenon in cellular mobile networks. *IEEE Journal on selected areas in communications*, vol. 15, 1406-1414, 1997.

[65] van-der-Mei, R. D., van-den-Berg, J. L., Vranken, R. et Gijsen, B. M. M. Sojourn-time approximations for a multi-server processor sharing system with priorities. *Performance Evaluation*, vol. 54, 249-261, 2003.

[66] Wagner, D. Waiting times of a finite-capacity multi-server model with non-preemptive priorities. *European Journal of Operational Research*, vol. 102, 227-241, 1997.

[67] Wang, P. P., Gong, M. et Lee, T. Waiting time in a combined first-come-first-served and shortest-time-first queue. *Comuters & Operations Research*, vol. 30, 1727-1743, 2003.

[68] Whitt, W. Deciding which queue to join: some counterexamples. *Operations Research*, vol. 34, 55-62, 1986.

[69] Whitt, W. Understanding the efficiency of multi-server service systems. *Management Science*, vol. 38, 708-723, 1992.

[70] Whitt, W. Improving service by informing customers about anticipated delays. *Management Science*, vol. 45, 192-207, 1999.

[71] Whitt, W. Predicting queueing delays. *Management Science*, vol. 45, 870-888, 1999.

[72] Winston, W. Optimality of the shortest line discipline. *Journal Of Applied Prob.*, vol. 14, 181-189, 1977.

[73] Yang, T. et Templeton, J. G. C. A survey on retrial queues. *Queueing Systems*, vol. 2, 201-233, 1987.

Résumé

Depuis quelques années, les centres d'appels enregistrent une forte croissance dans le monde. Les entreprises s'orientent de plus en plus vers ce choix qui leur offre une relation privilégiée avec leurs clients. Ainsi, ils disposent d'un moyen convivial et peu coûteux pour fidéliser leurs clients tout en essayant d'en acquérir de nouveaux. Ce travail fait suite à des problématiques réelles exprimées par l'entreprise *Bouygues Telecom*. Les données utilisées sont dans le même ordre de grandeur que ce qui se passe dans la réalité.

Dans la première partie de ce travail, nous étudions le phénomène de rappels. Nous commençons, d'abord, par montrer son importance dans le système. Par la suite, nous montrons que le fait d'ignorer ce phénomène engendre beaucoup d'erreurs lors du dimensionnement du centre d'appels et se traduit par des objectifs non satisfaits ou par un système surdimensionné sujet à des coûts supplémentaires. Cette analyse concerne le régime stationnaire. L'étude du système multi-période aboutit, elle, à la détermination de l'évolution de la demande réelle en fonction du temps et ce, à partir des arrivées observées.

Dans la deuxième partie, nous commençons par proposer des estimateurs du temps d'attente que les clients doivent passer dans la file avant d'être servis et ce, pour des clients répartis en plusieurs classes de priorité non-préemptive. En fonction de ces estimateurs, nous étudions plusieurs règles de routage. Nous comparons le système composé de plusieurs files d'attente, et nécessitant donc un routage, avec le système où les files d'attente sont fusionnées. Nous terminons cette partie par une analyse de deux disciplines de priorités probabilistes et nous les comparons avec la discipline de la priorité stricte. Comme dans le cas du routage, la comparaison s'effectue en fonction du nombre de conseillers nécessaires à la satisfaction de qualités de service objectif pour chaque classe de clients.

Mots-clés Centres d'appels, chaînes de Markov, files d'attente, approximation fluide, simulation, priorité non-préemptive, routage, qualité de service, rappels, abandons.

Abstract

In the recent years, the call center industry has been rapidly expanding. The companies are more and more frequently selecting this solution which offers them a privileged relation with their customers. This provides these companies a convivial way to better serve their clients and acquire new ones. This work is motivated by problems of the company *Bouygues Telecom*.

In the first part of this work, we study the phenomenon of retrials. We start, initially, by proving its importance. Thereafter, we show that the fact of ignoring this phenomenon generates errors when staffing the call center, resulting in unsatisfied target service levels or an overstaffed system with huge costs. This analysis is performed in the stationary regime. In the study of the multi-period system, we determine the evolution of the real time demand, based on observed arrivals.

In the second part, we study a multi-class system with non-preemptive priority and start by proposing estimators of customer waiting times. We analyse different routing policies based on these estimators. We compare the system with several queues with the single queue one. We finish this part by comparing two probabilistic priority policies with the well-known strict priority rule. The performance of each system is measured by the number of service representatives needed to achieve the target service level.

Key words Call centers, Markov chains, Queueing Theory, fluid approximation, simulation, non-preemptive priority, routing, quality of service, retrials, abandonments.