Méthode des martingales dans les problèmes de files d'attente

HOURIA OUKID

29 janvier 2019

DÉDICACES

A la mémoire de mon père et de mes grands parents,

A ma mère,

A mon mari Kamel et sa famille,

A mes enfants Celia et Walid à qui je souhaite une longue vie pleine de bonheur et de réussite,

A mon frère Abderrezak et sa femme Karima,

A mon frère Sadek,

A ma sœur Salima,

A mes oncles et tantes, en particulier à ma tante Nadia,

A mes cousins et cousines,

A tout le personnel de la faculté des Sciences.

REMERCIEMENTS

Je tiens tout d'abord à remercier le professeur Amar Aïssani d'avoir accepté d'encadrer cette thèse. Je le remercie pour sa disponibilité, ses précieux conseils et cela depuis ma thèse de Magister, dont j'ai eu l'honneur de l'avoir aussi comme encadreur, jusqu'à l'aboutissement de cette thèse de Doctorat.

Je remercie également le professeur M. Blidia d'avoir accepté de présider le jury de ma thèse.

Je souhaite également remercier Mesdames Nadia Oukid, Hafida Saggou, Messieurs Hamid Ould rouis et Noureddine Ikhlef-Eschouf de m'avoir fait l'honneur d'accepter d'être examinateurs de ma thèse.

Ma gratitude va également à ma tante Nadia, pour ses encouragements, ses remarques et précieux conseils qui ont abouti à une amélioration du document de la thèse.

Je tiens à remercier mon frère Sadek, pour son aide précieuse à la maitrise du logiciel Latex.

Je remercie du fond du cœur et avec un grand amour ma mère, mes frères et sœur, pour le soutien et l'amour dont ils ont toujours fait preuve à mon égard. Qu'ils trouvent ici l'expression de ma profonde gratitude. Un grand merci à toute ma famille et à ma belle famille.

Je ne terminerai pas sans exprimer ma profonde reconnaissance à mon mari que je remercie pour sa patience et son aide considérable. Enfin, mon affection va à mes deux adorables enfants Celia et Walid.

RESUME

La théorie de files d'attente vise à fournir la méthodologie d'évaluation de performance quantitative dans le cadre de certaines questions pratiques provenant de systèmes de communication et réseaux (débit, charge, temps de réponse...) et aussi une évaluation qualitative (stabilité, ergodicité, comparabilité...).

Étant donné que les approches classiques en théorie des files d'attente conduisent à des expressions complexes ou ne s'appliquent pas pour des systèmes complexes (multiserveurs), plusieurs méthodes d'évaluation ont été utilisées. Parmi les principales approches introduites ces dernières années, on trouve la méthode des martingales.

Les martingales constituent une classe très importante de processus stochastiques pour laquelle les propriétés sont basées sur celles de l'espérance mathématique conditionnelle. L'interprétation de ce processus stochastique est intéressante. En effet la valeur d'une martingale peut changer ; cependant, ses espérances restent constantes dans le temps. Plus important, l'espérance d'une martingale n'est pas affectée par l'échantillonnage aléatoire (optional sampling). A l'aide des martingales, on peut formuler des énoncés généraux très forts, et souvent intuitivement surprenants. Outre leur intérêt d'un point de vue purement mathématique, elles ont des applications clés en probabilités appliquées, en particulier les résultats de convergence des martingales et le théorème d'arrêt qui peuvent être appliqués une fois une martingale appropriée a été trouvée. L'avantage de cette approche est de permettre de formuler et d'analyser des problèmes plus généraux, en étudiant une extension plus large, que les méthodes traditionnelles.

Dans cette thèse, nous étudions l'application de ces méthodes à quelques modèles de systèmes de files d'attente.

Dans un premier temps, nous présentons une nouvelle approche basée sur la théorie des martingales pour analyser le système M/G/1 avec rappels. En utilisant l'équation récursive du processus induit aux instants de départ de ce système, nous avons construit une martingale arrêtée au premier instant où le système redevient vide. Nous avons obtenu le résultat de stabilité de ce système et le nombre moyen de clients dans le système.

Dans un deuxième temps, nous utilisons la décomposition de Doob-Meyer des semi martingales pour analyser un système multiserveur non-markovien avec pertes. Tout d'abord, nous considérons le problème général où les processus d'arrivées et de départs sont des processus ponctuels. Nous obtenons les équations de la distribution du nombre de clients dans le système. Ensuite nous considérons le cas où le processus ponctuel est un processus de Poisson homogène et non-homogène. Nous complétons notre travail par des exemples numériques illustrant la manière dont des praticiens pourraient exploiter ces résultats du point de vue d'aide à la décision : nombre minimal de serveurs pour garantir une probabilité de perte (refus) inférieure à un seuil α fixé.

Mots-clés : Systèmes multiserveurs, pertes, processus ponctuels, martingales et semimartingales, système avec rappels, chaîne de Markov incluse, théorème d'arrêt, période d'activité.

ABSTRACT

Queueing Theory aims to provide quantitative performance evaluation methodology in connection with some practical questions arising in Communication Systems and Networks (throughput, load, response time ...) and also qualitative evaluation (stability, ergodicity, comparability ...). has this effect, several evaluation techniques were used, Another tool which can be used to study the queueing systems is a martingale methode.

Martingales constitute a very important class of stochastic processes about which very strong, and often intuitively surprising, general statements can be made. In addition to their interest from a purely mathematical point of view, they have key applications in applied probability. Its properties are based on those of the conditional mathematical expectation. The interpretation of this stochastic process interesting in connection with Game Theory. Indeed a martingale's value can change ; however, its expectation remains constant in time. More important, the expectation of a martingale is unaffected by optional sampling. Martingales constitute a very important class of stochastic processes about which very strong, and often intuitively surprising, general statements can be made. In addition to their interest from a purely mathematical point of view, they have key applications in applied probability, in particular the martingale convergence results, the martingale inequalities and the very useful optional stopping theorem can be applied once an appropriate martingale has been found. The advantage of this approach is that, it provides a deepen analysis of the system helping to study a more wide its extension, than the traditional methods.

In this thesis, we study the application of these methods to different models of queueing systems.

As the first step, we present a new approach which uses martingale for analysing the $M/G/1$ retrial queue. Using a technique due to Baccelli and Makowski we define a discrete-time martingale stopped at the first passage time where the system becomes empty with respect to an embedded process and from this, we derive the stability condition and study the busy period of this system.

In the second step, we use the martingale method for analysing a non-

Markovian multiserver queue with n identical servers and losses. Such a queue can be used to model a switching center that allows a maximum of k simultaneous calls. We first state the general problem when the arrival and departure processes are quite general point processes where customers that arrive when all servers are busy are dropped and lost. we construct a martingale representation for the queue-length process. Using the Doob-Meyer semimartingale decomposition, we derives equation for the queue-length distribution and then solve it for particular special case when the arrival and departure processes are Markovian and the case when these processes are nonhomogeneous Poisson processes. The paper also study the problem of optimal number of servers to decrease the loss proportion for a given value.

Key words : Multiserver queues, losses, point processes, martingales and semimartingales, Retrial queues, Embedded Markov Chain, Doob's optional sampling theorem, Busy Period.

Table des matières

RESUME 1

Table des figures **9**

Introduction 10

1 SYSTÈMES DE FILES D'ATTENTE **16**

1.1 Introduction 16
1.2 Caractérisation des modèles de files d'attente 17
1.2.1 Processus d'arrivée 17
1.2.2 Nombre de serveurs 17
1.2.3 Capacité du système 18
1.2.4 Discipline de service 18
1.3 Notation des modèles de files d'attente 18
1.4 Analyse des performances 19
1.4.1 Évaluation de performances en régime stationnaire . . 20
1.4.2 Formule de Little 21
1.5 Exemples 21
1.5.1 Domaine systèmes de production 22
1.5.2 Domaine systèmes informatiques 22
1.6 Les files d'attente markoviennes 23
1.6.1 File d'attente $M/M/1$ 23
1.6.2 File d'attente $M/M/s$ 25
1.6.3 File d'attente $M/M/s/k$ 27
1.6.4 File d'attente $M/M/s/s$ 27
1.7 Modèles non Markoviens 28
1.7.1 File d'attente $M/G/1$ 29
1.7.2 File d'attente $G/M/m$ 30
1.7.3 File d'attente $G/G/1$ 32
1.8 Conclusion 32

2 SYSTÈMES DE FILES D'ATTENTE AVEC RAPPELS 33
2.1 Introduction . . . 33
2.2 Description du modèle . . . 34
2.3 Exemples d'applications de modèle de files d'attente avec rappels 35
2.3.1 Centre d'appel . . . 35
2.3.2 Réseaux à commutation par paquets . . . 36
2.3.3 Réseaux locaux : le protocole CSMA . . . 36
2.4 Modèles markoviens . . . 37
2.4.1 Système $M/M/n$ avec rappels . . . 38
2.5 Modèles semi-markoviens . . . 39
2.5.1 Système $M/G/1$ avec rappels . . . 39
2.6 Période d'activité . . . 42
2.7 Notes bibliographiques . . . 43
2.8 Conclusion . . . 44

3 MARTINGALES A TEMPS DISCRET 45
3.1 Introduction . . . 45
3.2 Définitions - Généralités . . . 45
3.2.1 Filtrations et martingales . . . 45
3.2.2 Exemples . . . 47
3.2.3 Interprétation dans le contexte d'un jeu d'argent . . . 47
3.3 Théorèmes d'arrêt . . . 48
3.3.1 Temps d'arrêt . . . 48
3.3.2 Théorème d'arrêt . . . 49
3.3.3 Propriétés des martingales par rapport aux temps d'arrêts 49
3.3.4 Décomposition . . . 50
3.4 Convergence des martingales . . . 50
3.4.1 Convergence des martingales dans L^2 . . . 50
3.4.2 Convergence des martingales dans L^1 . . . 51
3.5 Conclusion . . . 51

4 ANALYSE DU SYSTÈME M/G/1 AVEC RAPPELS 52
4.1 Introduction . . . 52
4.2 Description du modèle . . . 53
4.2.1 Éléments de probabilité . . . 54
4.2.2 Temps d'arrêt . . . 55
4.3 Martingale . . . 55
4.4 Stabilité du système . . . 56

4.5 Condition d'instabilité 57
4.5.1 Période d'activité 59
4.6 Conclusion 63

5 ANALYSE D'UN SYSTÈME MULTISERVEUR NON-MARKOVIEN AVEC PERTES 64
5.1 Introduction 64
5.2 Description du modèle 65
5.3 Décomposition en semimartingale du processus du nombre de clients 66
5.4 Renormalisation du processus du nombre de clients dans le système 67
5.5 Analyse de la distribution limite du nombre de clients dans le système 68
5.6 Cas particulier 73
5.7 Exemples numériques 75
5.7.1 Exemple 1 76
5.7.2 Exemple 2 76
5.8 Conclusion 77
CONCLUSION 78

Annexe A 79

A Lois de Probabilités et Processus Stochastiques 80
A.1 Lois de probabilités 80
A.1.1 Loi géométrique 80
A.1.2 Loi de Poisson 80
A.1.3 Loi exponentielle 81
A.2 Processus Stochastiques 81
A.2.1 Processus de comptage 82
A.2.2 Processus ponctuels 82
A.2.3 Processus de Poisson 83
A.2.4 Processus de Poisson non homogène 84

Annexe B 85

B Rappels : Espérance conditionnelle et Théorèmes de Convergence de Lebesgue **86**
B.1 Espérance conditionnelle 86
B.1.1 Propriétés de l'espérance conditionnelle analogues à celles de l'espérance 87
B.1.2 Propriétés spécifiques à l'espérance conditionnelle . . . 88
B.2 Théorèmes de Convergence pour l'intégrale de Lebesgue 89

Bibliographie **90**

Table des figures

1.1 Représentation graphique d'un système de file d'attente simple 17
1.2 Ligne de production 22
1.3 Diagramme de transition d'un système de file d'attente $M/M/1$ 23

2.1 Le schéma général d'un système d'attente avec rappels 35

INTRODUCTION

La Théorie des files d'attente (Queueing Theory) est une théorie mathématique relevant du domaine des probabilités, qui permet de modéliser un système admettant un phénomène d'attente, de calculer ses performances pour aider les gestionnaires dans leurs prises de décisions.

Tout a commencé au Danemark, entre 1909 et 1915, avec le développement du téléphone. La compagnie de Copenhague souhaitait à l'époque mettre en place une plateforme permettant aux utilisateurs d'être mis en relation par l'intermédiaire d'opérateurs, mais ne savait pas quelle taille devait avoir une telle structure, ni combien d'appels elle aurait à gérer. Si le centre était trop gros, l'entreprise risquait la faillite. Si elle voyait trop petit, les utilisateurs, faute d'être connectés, auraient manifesté leur mécontentement. La compagnie a donc demandé à l'un de ses ingénieurs, Agner Krarup Erlang, de travailler à une conceptualisation mathématique des files d'attente. C'est donc en 1909 que les bases de ce formalisme sont jetées, grâce à l'article du mathématicien danois A.K. Erlang "The theory of probabilities and telephone conversations". Les premiers résultats sont variés : Erlang observe le caractère poissonnier des arrivées des appels à un central téléphonique, et le caractère exponentiel des durées des appels ; il réussit à calculer de manière relativement simple la probabilité d'avoir un appel rejeté. La notion d'équilibre stationnaire d'un système d'attente est introduite.

À partir des années 30, les travaux de plusieurs mathématiciens à l'image de Molina, Fry, Pollaczek aux États-Unis, Kolmogorov et Khintchine en Russie, Palm en Suède, ou Crommelin en France permettent à la théorie des files d'attente de se développer lentement.
Ce sont ensuite les années 50 qui verront l'essor important de la théorie. Les applications de ces travaux sont alors très pratiques, et concernent les disciplines de recherche opérationnelle et génie industriel. On peut citer les flux de trafic (véhicule, avion, personnes, télécommunications), l'ordonnancement, c'est-à-dire la planification : les patients dans les hôpitaux, les programmes d'un ordinateur, etc . . . ou encore le dimensionnement (banque, poste, réseaux, téléphone, ordinateur).

Dans les années 80, cette discipline devient beaucoup plus mathématique, et la littérature regorge d'articles décrivant des techniques mathématiques

permettant de trouver des solutions exactes aux modèles.
Au cours de la décennie suivante, les chercheurs s'intéressent d'avantage à la création de modèles, et au calcul scientifique associé pour résoudre ces modèles. En effet, le développement de la puissance des ordinateurs permet maintenant d'obtenir des solutions approches des modèles suffisamment fiables pour être utilisées. Actuellement ce sont les applications dans le domaine de l'analyse de performance des réseaux (téléphone mobile, Internet, multimédia,...) qui suscitent le plus de travaux. De ce fait la théorie des files d'attente est aujourd'hui largement utilisée et ses applications sont multiples.

L'évolution rapide des systèmes informatiques et de réseaux de télécommunication ont montré les limites de la théorie des files d'attente dites classiques qui ne permettent pas d'expliquer le comportement stochastique de certains systèmes complexes comme les systèmes téléphoniques où les abonnés répétaient leurs appels en recomposant le numéro plusieurs fois jusqu'à l'obtention de la communication. Ce qui a conduit certains chercheurs à développer d'autres modèles plus élaborés qu'on appelle "files d'attente avec rappels", en anglais "Retrial Queue" (Cohen, 1957). Cependant, l'influence de ce phénomène a été longtemps négligée durant les décennies suivantes. Ce n'est que vers les années 1970-1980 qu'on a vu un net regain d'intérêt pour cette catégorie de modèles, avec l'avènement de nouvelles technologies, notamment dans les systèmes de télécommunication : réseaux ATM.

Les systèmes de files d'attente avec rappels peuvent être appliqués pour résoudre les problèmes pratiques, tels que l'analyse du comportement des abonnés dans les réseaux téléphoniques, l'analyse du temps d'attente pour accéder à la mémoire sur les disques magnétiques, ... Ce type de modèles se rencontre également dans la modélisation de protocoles spécifiques de communication, tels que CSMA (Carrier Sense Multiple Access) ou encore les disciplines Auto-Repeat, Ring-Back-When-Free, Repeat-LastNumber,... (Khomichkov, 1995).
Par conséquent, de nombreux travaux ont été publiés dans des journaux spécialisés en probabilités appliquées et modèles stochastiques, statistiques et recherche opérationnelle, télécommunication et ingénierie industrielle, et informatique. Le grand intérêt de ce domaine est confirmé par l'organisation d'une série de workshops sur les systèmes de files d'attente avec rappels : Madrid(1998), Minsk (1999, 2011), Amsterdam (2000), Cochin (2002), Seoul (2004), Miraflores de la Sierra (2006), Athens (2008) et Beijing (2010).

A cet effet, quelques revues de renommées internationales ont dédié des numéros spéciaux ; c'est le cas du journal Annals of Operation Research [13], European Journal of Operation Research [18], Mathematical and Computer Modelling [12] , Queueing Systems [93] et Top [10].

Parmi les premières contributions sérieuses sur les modèles d'attente avec rappels, on trouve celles de Cohen [[34], 1957], de Eldin [[40], 1967], de Hashida et Kawashima [[53], 1979] et de Lubacz et Roberts [[74], 1984]. Les progrès récents sont résumés dans les articles de synthèse de A. Aïssani [[4], 1994], Kulkani et Liang [[65], 1997), Templeton [[94],1999], dans les monographies de Falin et Templeton [[46], 1997], Artalejo et Gòmez-Corral [[19],2008], Gòmez-Corral et Ramalhoto [[50], 2000] (74), Rodrigo [[84], 2006] et dans les travaux bibliographiques de Artalejo (1999 et 2010) [11, 14].

Pour évaluer la performance d'un système, on utilise soit les méthodes analytiques, telles que les réseaux de files d'attentes, soit la simulation. Chacune de ces méthodes comporte ses avantages et ses inconvénients. Les solutions analytiques bénéficient de temps de résolution très rapides. Les résultats peuvent être immédiats, car ils sont déterminés à partir d'équations mathématiques issues du formalisme emprunté [49].

La théorie des martingales constitue sans doute la technique mathématique de base des probabilités modernes. Une martingale est un processus aléatoire qui ne possède pas de partie prévisible relativement à l'information dont on dispose. Cette théorie a eu de grandes répercussions dans de nombreux champs d'application, en probabilité bien sûr, mais aussi pour la résolution numérique des équations aux dérivées partielles (voir l'UP "théorie du potentiel", EDP), en assurance (théorie de la ruine) et en finance.

Pour les probabilistes, les martingales sont avant tout, des processus intégrables vérifiant une propriété précise d'espérance conditionnelle.

Outre l'usage financier mentionné précédemment, elles sont appliquées à divers problèmes stochastiques ou analytiques et représentent, avec les processus de Markov, l'une des catégories de processus dépendant du passé les plus importantes. On pourra se référer avec profit à [Williams, 1991], la notion semble provenir assez directement de l'idée de stratégie pour un jeu de

hasard. Bien que l'on ait eu très tôt l'intuition qu'une stratégie toujours gagnante pour un jeu défavorable n'existait pas, il faut attendre le début du vingtième siècle pour obtenir une formalisation des notions et du problème (en partie suite au débat sur les axiomes des probabilités proposés par R. von Mises).

Les pionniers du concept de martingale sont alors S. Bernstein, P. Lévy, J. Ville, E. Borel et J. Doob. Cependant, on peut trouver à posteriori des premiers exemples de martingales dans des travaux plus anciens dont par exemple ceux de Pascal sur le problème des partis comme l'explique Y. Derriennic [2003].
En ce qui concerne l'origine du mot (et non du concept), la première citation se trouve dans la thèse de J. Ville introduit au chapitre IV § 3 dans l'expression "système de jeu ou martingale". Mais à partir du chapitre suivant, J. Ville abandonne définitivement l'appellation "système de jeu" et ne garde que "martingale". Ce dernier précise par ailleurs, [Ville, 1985], que la dénomination est directement empruntée au vocabulaire des joueurs. le lecteur curieux de l'histoire de la théorie des martingales pourra consulter Lévy [[70], 1937], le très intéressant petit livre de ville [[96], 1939] et le premier article de Doob [[38], 1940] où figurent le "lemme maximal" et les théorèmes de convergence les plus importants, l'article de Snell[90], qui introduit la notion de sousmartingale et le célèbre chapitre 7 du livre de Doob [39]. Le lecteur pourra aussi consulter les ouvrages de Williams [97], de Neveu [75], et Rogers et Williams [85]. Si l'on peut attribuer – sans trop de risque d'erreur – la découverte de la notion de martingale à Jean Ville (1910 - 1988) exposée dans sa thèse : Étude critique de la notion de collectif, Paris (1939), ce sont les travaux de J. Doob qui développent la théorie des martingales, en établissant les théorèmes de convergence, et de nombreuses utilisations importantes des martingales. Nous ne mentionnerons en particulier que deux résultats fondamentaux :
– le premier est le théorème d'arrêt qui exprime que la propriété de constance en espérance d'une martingale, espérance prise en tout temps t déterministe, s'étend lorsque l'on remplace t par n'importe quel temps d'arrêt borné T.
- Un second résultat fondamental concerne la majoration en norme $L^p(p > 1)$ d'une sous-martingale positive $\{X_u, u \leq t\}$, par un multiple de la norme L^p de X_t.

L'objectif de notre travail est d'élargir le champs d'application de la théorie des martingales aux systèmes de files d'attente. Nous avons choisi deux

types de modèles.

Le premier est le modèle $M/G/1$ avec rappels qui est l'un des modèles fondamentaux de la théorie des files d'attente avec rappels, le plus étudié par les spécialistes Falin [43], Artalejo [9], Kulkarni [71],... et qui a été largement utilisé pour modéliser beaucoup de situations pratiques dans les systèmes de communications téléphoniques et réseaux de télécommunication. Il a été étudié par différentes approches mathématiques : méthodes de la variable supplémentaire et semi-Markovienne, la méthode de la chaine de Markov induite, approche régénérative de Markov...

Le second est un modèle multiserveur non-markovien avec pertes où à l'arrivée d'un client, si l'un des serveurs est libre, le client sera pris en charge immédiatement, dans le cas contraire, le client est perdu. Ce modèle peut être utilisé pour modéliser un centre de commutation permettant un maximum de k appels simultanés.

Notre thèse est composée de cinq chapitres organisés comme suit :

Le premier chapitre débute par une présentation détaillée des composantes d'un système de files d'attente, suivie d'une étude de quelques modèles markoviens et non markoviens.

Dans le chapitre 2, nous rappelons quelques notions sur les systèmes de files d'attente avec rappels.

Le chapitre 3 est consacré aux résultats fondamentaux de la théorie des martingales discrètes, essentiellement les théorèmes de convergence et d'arrêt.

Dans le chapitre 4, nous nous sommes intéressés à l'analyse mathématique du système d'attente $M/G/1$ avec rappels. En utilisant l'équation récursive du processus induit aux instants de départ de ce système, nous avons construit une martingale arrêtée au premier instant où le système redevient vide. Nous avons obtenu la condition de stabilité de ce système et le nombre moyen de clients dans le système.

Enfin dans le chapitre 5, nous utilisons la décomposition de Doob-Meyer des semi martingales pour analyser un système de files d'attente à plusieurs

serveurs et avec perte. Tout d'abord, nous considérons le problème général lorsque les processus d'arrivées et de départs sont des processus ponctuels, ensuite nous considérons le cas où le processus ponctuel est un processus de Poisson homogène et non-homogène. Des exemples numériques sont donnés où nous cherchons à optimiser le nombre de serveurs pour garantir une probabilité de perte inférieure à un seuil fixé.

les résultats issus des deux derniers chapitres ont fait l'objet d'une publication dans le journal "Advanced Studies in Contemporary Mathematics" [80], et de plusieurs communications et publications dans des conférences et proceedings internationaux [[76], [79], [78], [77]].

Une conclusion finale termine ce travail, donnant quelques suggestions et perspectives pour des recherches futures.

En Annexe A, nous rappelons quelques lois de probabilités et de processus stochastiques qui sont nécessaires dans la modélisation des files d'attente.

L' Annexe B contient un rappel sur les règles de calculs des espérances conditionnelles.

Chapitre 1

SYSTÈMES DE FILES D'ATTENTE

1.1 Introduction

Un modèle de file d'attente est une description abstraite d'un système réel de file d'attente. Le modèle classique de la file d'attente consiste en un système dans lequel des serveurs sont soumis à un flux de requêtes qu'ils doivent traiter. Il a un grand nombre d'applications dans les réseaux de télécommunication, dans les réseaux informatiques, les analyses de trafic ou même dans de "vraies" files d'attente, au magasin, au cinéma,... Il permet de répondre à des questions de temps de traitement, de structuration de réseaux ou de dimensionnement.

La configuration basique d'un système de file d'attente peut être décrite de la manière suivante (Figure 1.1) : des "clients" arrivent à un certain endroit et réclament un certain service. Les instants d'arrivée et les durées de service sont généralement des quantités aléatoires. Si un poste de service est libre, le client qui arrive se dirige immédiatement vers ce poste où il est servi, sinon il prend sa place dans la file d'attente dans laquelle les clients se rangent suivant leur ordre d'arrivée. Une fois le serveur libéré, le client entre en service et occupe le serveur pendant tout son temps de service. Puis le client libère le serveur et quitte le système.

Un système d'attente comprend donc un espace de service avec une ou

plusieurs stations de service montées en parallèle, et un espace d'attente dans lequel se forme une éventuelle file d'attente.

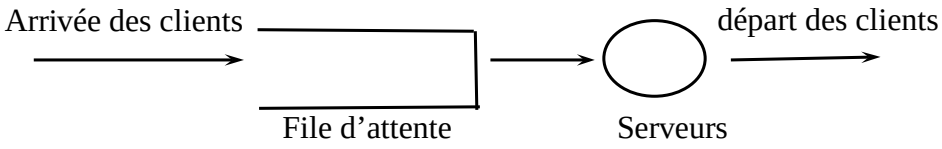

FIGURE 1.1 – Représentation graphique d'un système de file d'attente simple

1.2 Caractérisation des modèles de files d'attente

La définition d'un modèle de file d'attente nécessite principalement la caractérisation du processus d'arrivée, de la distribution du temps de service, du nombre de serveurs, de la capacité du système et de la discipline de service.

1.2.1 Processus d'arrivée

L'arrivée des clients à la station sera décrite à l'aide d'un processus stochastique de comptage $\{N(t), t \geq 0\}$ où $N(t)$ la variable aléatoire indiquant le nombre d'arrivées dans un intervalle de temps $(0, t]$.
Soit $T_n = A_n - A_{n-1}$ le temps séparant l'arrivée du $(n-1)^{ème}$ client et celle du $n^{ème}$ client.
Un processus de comptage $\{N(t), t \geq 0\}$ est un processus de renouvellement si et seulement si les variables aléatoires $\{T_n\}_{n=1,2,...}$ sont des variables indépendantes et identiquement distribuées (i.i.d). La loi T décrivant le temps d'inter-arrivée suffit alors à caractériser le processus de renouvellement.

La plupart du temps, l'arrivée des clients à une file simple est supposée décrite par un processus de renouvellement. Le processus d'arrivée le plus simple à étudier (et donc le plus couramment employé) est le processus de Poisson.

1.2.2 Nombre de serveurs

Le nombre de serveurs indique le nombre maximal d'exécutions en parallèle du même service. Dans un système de file d'attente multiserveur, les

clients qui arrivent se placent dans une seule file d'attente. Chaque fois qu'un serveur est libéré, un client en attente dans la file entre en service. Les temps de service des serveurs sont généralement indépendants et identiquement distribués.

1.2.3 Capacité du système

Dans certains systèmes de files d'attente, des contraintes physiques ou organisationnelles peuvent exister et limitent la longueur maximale de la file. Dans ces types de cas, la capacité du système indique le nombre maximal de clients qui peuvent se retrouver dans le système (en attente de service et en service). Cette capacité peut être finie ou infinie. Par conséquent, si un client arrive et qu'il y a déjà C clients dans le système, le client peut être accepté ou rejeté suivant la politique de débordement de la station.
Dans un système de production, cette capacité peut être liée à une limite de l'espace de stockage.

1.2.4 Discipline de service

La discipline de service détermine l'ordre dans lequel les clients sont sélectionnés pour le service. Les disciplines les plus utilisées sont : premier arrivé premier servi (First In First Out (FIFO)), First Come First Served (FCFS), dernier arrivé premier servi (Last In First Out (LIFO)), Last Come First Served (LCFS), sélection aléatoire, temps de service le plus court d'abord, règles de priorité préemptives (le service en cours d'exécution peut être interrompu) ou non-préemptives.

1.3 Notation des modèles de files d'attente

Dans la théorie des files d'attente, la notation de Kendall (premièrement proposée par D.G. Kendall en 1953) est un système standard pour décrire les caractéristiques essentielles d'un modèle de file d'attente. La notation de Kendall a la forme $A/B/C/K/N/D$ où :

A : distribution des temps d'inter-arrivées
B : distribution du temps de service
C : nombre de serveurs

K : capacité du système
N : nombre de clients existant dans l'univers considéré
D : discipline de service

Les différents symboles utilisés pour la caractérisation de la distribution des temps d'inter-arrivées et du temps de service sont :

M : loi exponentielle (Markovienne)
G : loi générale
GI : lois générales indépendantes
D : loi constante (déterministe)
E_k : loi d'Erlang-k
H_k : loi hyperexponentielle-k
C_k : loi de Cox-k
PH : loi de type phase

Remarque 1.1. *Lorsque un système de file d'attente est caractérisé en utilisant seulement trois champs $A/B/C$, on sous-entend que le système est à capacité infinie, que la population est infinie et que la discipline de service est FIFO.*

1.4 Analyse des performances

Les réseaux des files d'attente servent à analyser les performances du système modélisé. Cette analyse peut être menée en étudiant le comportement du système selon deux axes :

Étude en régime transitoire : L'étude du régime transitoire permet de répondre à des questions de performance qui sont liées à des instants donnés ou sur des périodes de court terme. Par exemple, « combien de clients demandant un service X vont être servis durant la prochaine heure ? ».

Étude en régime stationnaire ou permanent : dit aussi à l'équilibre consiste à vérifier si le système tend vers un équilibre (en terme de probabilité) lorsque le temps croît (à long terme). Cette analyse permet de répondre aux questions telles que : durant une longue période, quel est le

taux moyen d'occupation du serveur ?

Pour étudier cela, des méthodes stochastiques sont utilisées. Elles consistent à estimer la distribution du processus stochastique engendré par le modèle analysé, soit à un instant donné (analyse transitoire), ou bien à long terme (à l'équilibre). Elles permettent de calculer les probabilités pour que le système se trouve dans chacun des états du processus. Ces probabilités sont utilisées pour le calcul des paramètres de performance.

Les modèles de files d'attente les plus simples à analyser sont les modèles markoviens (distributions exponentielles des inter-arrivées et service). Ceux-ci engendre une chaîne de Markov (d'où le nom markovien).

1.4.1 Évaluation de performances en régime stationnaire

Soit $X(t)$ le nombre de clients dans un système de file d'attente (le nombre de clients en attente de service plus le nombre de clients en service) à l'instant $t,\ t \geq 0$. Sous certaines conditions, la distribution de $X(t)$ a une limite pour $t \to \infty : P_n = \lim_{t\to\infty} P\{X(t) = n\}$.

L'existence de cette limite montre que, à long terme, le système atteint un régime permanent indépendant de son état initial. La limite P_n est interprétée comme la probabilité d'avoir exactement n clients dans le système en régime permanent. Quand les probabilités existent, on dit que le processus stochastique $\{X(t), t \geq 0\}$ est ergodique. Pour la plupart des systèmes de files d'attente, la condition générale pour l'existence des probabilités, est la stabilité du système.

Un système de file d'attente est dit stable si le nombre de clients dans le système ne peut pas augmenter jusqu'à l'infini.

Les autres quantités fondamentales utilisées afin d'analyser les performances d'un système sont :
X_f : le nombre de clients en attente dans la file en régime permanent
W : le temps de séjour des clients dans le système en régime permanent (le temps d'attente plus le temps de service)
W_f : le temps de séjour des clients dans la file d'attente en régime permanent

Dans la théorie des files d'attente, on s'intéresse plutôt aux mesures de performances espérées :

$\bar{n} = E[X]$: le nombre moyen de clients dans le système

$\bar{n_f} = E[X_f]$: le nombre moyen de clients en attente

$\bar{w} = E[W]$: le temps moyen de séjour des clients dans le système

$\bar{w_f} = E[W_f]$: le temps moyen de séjour des clients dans la file d'attente

1.4.2 Formule de Little

La formule de Little [Kleinrock [60], 1975] est une loi générale (désigne des relations entre les mesures de performances) qui s'énonce comme suit : « le nombre moyen des clients dans un système est égal au produit du débit du système par le temps moyen passé dans le système par chaque client ».
Soit $X(t)$ le nombre de clients arrivés dans un intervalle de temps $(0, t]$.
Le taux moyen d'arrivée $\lambda_a = \lim_{t\to\infty} \frac{X(t)}{t}$ exprime le nombre moyen de clients arrivés dans le système par unité de temps.
La loi de Little indique que

$$E[X] = \lambda_a E[W] \tag{1.1}$$

Si on applique la loi de Little seulement à la file d'attente, on obtient

$$E[X_f] = \lambda_a E[W_f] \tag{1.2}$$

La loi de Little est valide pour presque tous les systèmes de files d'attente indépendamment du processus d'arrivée, du nombre de serveurs, ou de la discipline de service.

1.5 Exemples

Les exemples que nous allons présentés sont tirés de Bruno Baynat [27].

1.5.1 Domaine systèmes de production

Une ligne de production est la configuration la plus simple d'un système de production. Toutes les pièces doivent, en effet, passer par toutes les machines de l'atelier et dans le même ordre. Entre chaque machine existe un stock de capacité finie. Le nombre total de pièces en attente de la machine i ou en usinage sur cette machine est donc limité à une certaine valeur (soit K_i).

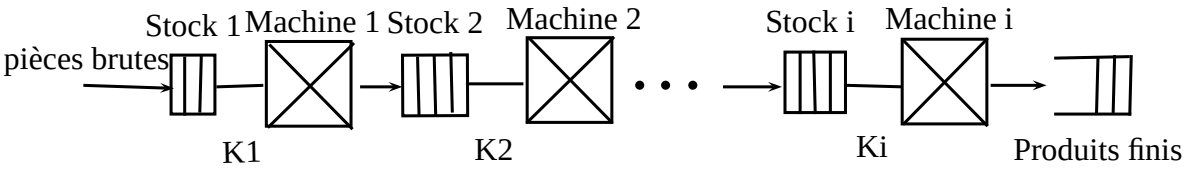

FIGURE 1.2 – Ligne de production

Ce type de système se modélise naturellement par un réseau de files d'attente "en tandem" à capacités limitées avec blocage après service. Les clients du modèle sont alors les pièces du système de production. La file d'attente de la station i modélise l'attente des pièces devant la machine i et le serveur modélise le temps d'usinage de la machine i.

1.5.2 Domaine systèmes informatiques

Le système considéré comporte une unité centrale qui exécute des processus et un ensemble d'unités d'entrée-sortie (disque, disquette, bande magnétique, etc.). Ces différentes composantes constituent les ressources du système. Les entités qui circulent dans le système sont des processus. Ces processus peuvent être dans les différents états suivant :
1. Prêt : en attente de libération de l'unité centrale,
2. élu : en exécution sur l'unité central,
3. en attente : en attente ou en exécution d'une entrée/sotie sur une des unités de stockage.
On modélise ce système par un réseau de files d'attente ouvert. Les clients du modèle sont les processus du système.

1.6 Les files d'attente markoviennes

Les modèles markoviens caractérisent les systèmes dans lesquels les deux quantités stochastiques principales, qui sont le temps inter-arrivées et la durée de service, sont des variables aléatoires indépendantes et exponentiellement distribuées. La propriété d'absence de mémoire de loi exponentielle facilite l'étude de ces modèles. L'étude mathématique de tels systèmes se fait par l'introduction d'un processus stochastique approprié. Ce processus est souvent le processus $\{X(t),\ t \geq 0\}$ défini comme étant le nombre de clients dans le système à l'instant t. L'évolution temporelle du processus markovien est complètement définie grâce à la propriété d'absence de mémoire.

1.6.1 File d'attente $M/M/1$

La file $M/M/1$ est la file la plus simple et la plus utilisée pour modéliser les systèmes informatiques. L'utilisation de cette file est motivée par l'ensemble de ses résultats permettant de déterminer les paramètres de performances moyens. Elle est définie par le processus stochastique suivant :
- le processus d'arrivée des clients est distribué selon un processus de Poisson de paramètre λ.
- le processus de temps de service est indépendant du processus d'arrivée et suit la loi exponentielle de paramètre μ.
Soit $X(t)$ le nombre de clients dans le système à l'instant $t \geq 0$. Le processus stochastique $\{X(t),\ t \geq 0\}$ est une chaîne de Markov à temps continu ayant le diagramme de transition de la Figure 1.3

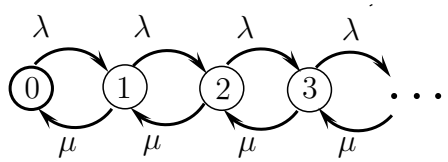

FIGURE 1.3 – Diagramme de transition d'un système de file d'attente $M/M/1$

Régime transitoire

Grâce aux propriétés fondamentales du processus de Poisson et de la loi exponentielle, nous avons pour un petit intervalle de temps Δt les probabilités suivantes :

$$P[\textit{exactement 1 arrivée pendant } \Delta t] = \lambda\Delta t + O(\Delta t)$$

$$P[\textit{aucune arrivée pendant } \Delta t] = 1 - \lambda\Delta t + O(\Delta t)$$

$$P[\textit{2 arrivées ou plus pendant } \Delta t] = O(\Delta t)$$

$$P[\textit{exactement 1 départ pendant } \Delta t / X(t) \geq 1] = \mu\Delta t + O(\Delta t)$$

$$P[\textit{aucun départ pendant } \Delta t / X(t) \geq 1] = 1 - \mu\Delta t + O(\Delta t)$$

$$P[\textit{2 départs ou plus pendant } \Delta t] = O(\Delta t).$$

Ces probabilités ne dépendent ni du temps t ni de l'état $X(t)$ dans lequel se trouve le système.
Les probabilités $P_n(t) = P(X(t) = n), n \geq 0$, sont les solutions du système :

$$P_n^{'}(t) = -(\lambda + \mu)P_n(t) + \lambda P_{n-1}(t) + \mu P_{n+1}(t)$$

$$P_0^{'}(t) = -\lambda P_0(t) + \mu P_1(t).$$

Ces équations sont connues sous le nom d'équations différentielles de Kolmogorov.

Régime stationnaire

Lorsque $t \to \infty$, on a $lim_{t\to\infty} P_n(t) = P_n$ existent et indépendantes de l'état initial du processus et $lim_{t\to\infty} P_n^{'}(t) = 0$.
On obtient un système d'équations linéaires et homogènes :

$$\lambda P_0 = \mu P_1$$

$$\lambda P_{n-1} + \mu P_{n+1} = (\lambda + \mu)P_n, \ n \geq 1$$

$$\sum_{n=0}^{\infty} P_n = 1$$

La solution de ce système nous donne

$$P_n = P\{X = n\} = \big(1 - \frac{\lambda}{\mu}\big)\big(\frac{\lambda}{\mu}\big)^n, \ n \geq 1 \tag{1.3}$$

avec

$$P_0 = P\{X = 0\} = 1 - \frac{\lambda}{\mu} \tag{1.4}$$

La variable aléatoire X suit alors une distribution de probabilité géométrique. Notons qu'une condition nécessaire pour l'existence de la probabilité P_n pour $n = 0, \ldots, \infty$ est $\lambda < \mu$. Autrement dit, le taux d'utilisation $\rho = \frac{\lambda}{\mu}$ qui exprime la proportion du temps pendant lequel le serveur est occupé doit satisfaire la condition $\rho < 1$. C'est la condition de stabilité du système. Quand $\lambda > \mu$, le nombre de clients dans le système augmente sans limite et donc les probabilités $P_n,\ n \geq 0$, n'existent pas.

Caractéristiques du système

Les mesures de performances du système sont obtenus en utilisant les expressions (1.3) et (1.4) et les relations (1.1) et (1.2) où $\lambda_a = \lambda$:

- Le nombre moyen de clients dans le système :

$$\bar{n} = \frac{\lambda}{\mu - \lambda}$$

- Le temps moyen de séjour dans le système :

$$\bar{w} = \frac{1}{\mu - \lambda}$$

- Le nombre moyen de clients dans la file :

$$\bar{n_f} = \frac{\lambda^2}{\mu(\mu - \lambda)}$$

- Le temps moyen d'attente dans la file :

$$\bar{w_f} = \frac{1}{\mu(\mu - \lambda)}$$

1.6.2 File d'attente $M/M/s$

On considère à présent la file d'attente $M/M/s$ où les instants d'arrivée sont toujours poissoniens, les temps de services exponentiels mais il y a s guichets ou "serveurs" disponibles. Si lorsque un client arrive il y a au moins un serveur disponible, le client entre en service immédiatement. Dans le cas contraire, il est placé dans la file d'attente. Les temps de services aux guichets sont bien sûr mutuellement indépendants.

Soit à nouveau $X(t)$ le nombre de clients dans le système à l'instant t. $\{X(t),\ t \geq 0\}$ est un processus de naissance et de mort dont les taux de transitions sont $\lambda_n = \lambda,\ \forall n \geq 0$ et

$$\mu_n = \begin{cases} n\mu & 0 \leq n \leq s \\ s\mu & n \geq s \end{cases} \tag{1.5}$$

On appelle $s\mu$ le taux de service global du système.
La distribution stationnaire du système :

$$P_n = \begin{cases} \frac{1}{n!}(\frac{\lambda}{\mu})^n P_0 & si\ n \leq s \\ \frac{1}{s!s^{n-s}}(\frac{\lambda}{\mu})^n P_0 & si\ n \geq s \end{cases} \tag{1.6}$$

où

$$P_0 = \big[\sum_{n=0}^{s-1} \frac{1}{n!}(\frac{\lambda}{\mu})^n + \frac{1}{s!(1-\frac{\lambda}{s\mu})}(\frac{\lambda}{\mu})^s\big]^{-1}. \tag{1.7}$$

Ces relations sont valables si $\frac{\lambda}{s\mu} = \rho_s < 1$.
La probabilité qu'un client soit placé dans la file d'attente en arrivant est,

$$P(attente) = P(X \geq s) = \frac{P_s}{1-\rho} \tag{1.8}$$

Le cas $s = \infty$

Dans ce cas, il y a un nombre infini de serveurs. Un client entre toujours immédiatement en service à son arrivée. La résolution de processus de naissances et de morts associé donne :

$$P_n = P_0 \frac{\rho^n}{n!} \tag{1.9}$$

avec $\sum_{n=0}^{\infty} P_n = 1$, on obtient $P_0 = e^{-\rho}$.
Par conséquent la distribution stationnaire de la file $M/M/\infty$ est

$$P_n = \frac{\rho^n}{n!} e^{-\rho},\ n \geq 0 \tag{1.10}$$

Cette distribution est identique à celle de Poisson de paramètre ρ. Notons que cette distribution qui est la distribution limite du résultat du paragraphe précédent, existe quelles que soient les valeurs de λ et de μ. En ce qui concerne les caractéristiques du système, on a :

$\bar{n} = E(X) = \frac{\lambda}{\mu}$ et $\bar{w} = \frac{1}{\mu}$, tandis que $\bar{n_f} = \bar{w_f} = 0$.

1.6.3 File d'attente $M/M/s/k$

On considère un système de file d'attente à s serveurs et un nombre limité de clients ($s < k$). Les arrivées sont poissoniens de taux λ. Les durées de service sont exponentielles de taux μ. Si un client trouve à son arrivée le système complet, il s'en va.
Puisque les clients sont en nombre fini k dans le système, on a alors :

$$\lambda_n = \begin{cases} \lambda & 0 \leq n \leq k \\ \\ 0 & n > k \end{cases} \tag{1.11}$$

$$\mu_n = \begin{cases} n\mu & 0 < n < s \\ s\mu & s \leq n \leq k \end{cases} \tag{1.12}$$

On a :

$$P_n = P_0 \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i}. \tag{1.13}$$

1. Si $n \geq k+1$, $P_n = 0$.
2. Si $n \leq k$,
a. $0 < n < s$, $P_n = P_0 \frac{\rho^n}{n!}$,

b. $s \leq n \leq k$, $P_n = P_0 \frac{\lambda^n}{s! \mu^n s^{n-s}}$,

avec

$$P_0 = \Big[\sum_{n=0}^{s-1} \frac{1}{n!} \big(\frac{\lambda}{\mu}\big)^n + \sum_{n=s}^{k} \frac{1}{s! s^{n-s}} \big(\frac{\lambda}{\mu}\big)^n \Big]^{-1}.$$

1.6.4 File d'attente $M/M/s/s$

Dans ce modèle il y a s serveurs, mais pas de file d'attente. Lorsqu'un client arrive, s'il y a au moins un serveur disponible, le client entre directement en service. Sinon le client est rejeté. C'est le modèle de fonctionnement d'un central téléphonique. On établit immédiatement que

$$\lambda_n = \begin{cases} \lambda & 0 \leq n \leq s \\ \\ 0 & n > s \end{cases} \tag{1.14}$$

d'où

$$P_n = \frac{1}{n!}(\frac{\lambda}{\mu})^n P_0,\ n \leq s \tag{1.15}$$

avec

$$P_0 = [\sum_{n=0}^{s} \frac{1}{n!}(\frac{\lambda}{\mu})^n]^{-1} \tag{1.16}$$

Cette formule est connue sous le nom de la formule d' Erlang-B.
Notons qu'aucune restriction ne doit être imposée à λ et μ pour assurer l'existence d'une distribution stationnaire.
La probabilité de pertes du système, qui est la probabilité qu'un client qui arrive ne puisse entrer, est égale à la probabilité pour le système de se trouver dans l'état s :

$$P_s = \frac{1}{s!}(\frac{\lambda}{\mu})^s P_0 \tag{1.17}$$

On trouve également

$\bar{w} = \frac{1}{\mu}$ et $\bar{n} = \frac{\lambda}{\mu}(1 - P_s)$, tandis que $\bar{n_f} = \bar{w_f} = 0$.

1.7 Modèles non Markoviens

Nous appellerons systèmes de files d'attente non markoviens ceux pour lesquels la distribution des intervalles du flux d'arrivées et/ou la distribution des temps de service est (sont) différente(s) de la distribution exponentielle. La technique de base utilisée pour l'étude de tels systèmes consiste à construire un certain processus de Markov judicieusement choisi à l'aide de l'une des méthodes d'analyse suivantes :

Méthode de la chaîne de Markov induite : elle consiste à choisir une suite d'instants déterministes ou aléatoires tels que la chaîne induite soit markovienne et homogène.

Méthode des variables supplémentaires : Elle consiste à compléter l'information sur le processus de telle manière à lui donner le caractère markovien.

Simulation : C'est un procédé d'imitation artificielle d'un processus réel effectué sur ordinateur. Elle nous permet d'étudier les systèmes les plus complexes, de prévoir leurs comportements et de calculer leurs caractéristiques. Les résultats obtenus ne sont qu'approximatifs, mais peuvent être utilisés avec une bonne précision.

1.7.1 File d'attente $M/G/1$

La file M/G/1 est une file à capacité illimitée ayant un seul serveur. Le processus d'arrivée est toujours un processus de Poisson de taux λ tandis que les durées de service suivent une loi générale G de moyenne $\frac{1}{\mu}$ de transformée de Laplace-Stieltjes $B^*(s),\ Re(z) > 0$. Ici, le processus $X(t)$ n'est plus un processus de Markov, et les techniques précédentes ne s'appliquent plus. En fait, la probabilité d'avoir une transition de l'état $\{X = n\}$ vers l'état $\{X = n-1\}$ dépend maintenant de la quantité de service que le client en service a déjà reçu. Dans ce cas particulier, on peut déterminer tous les paramètres de performances, même si son service ne vérifie pas la propriété sans mémoire. Ces paramètres sont déterminés grâce à la méthode de la chaîne de Markov induite. Cette méthode consiste à ramener l'étude à une chaîne de Markov à temps discret en considérant des instants d'observation particuliers (instants de début de service ou instants de fin de service).

Chaîne de Markov induite

Pour tout $n \geq 0$, on note X_n le nombre de clients dans le système juste après le $n^{ème}$ départ, qui satisfait l'équation

$$X_{n+1} = X_n + A_{n+1} - 1_{\{X_n>0\}},\ n \geq 0 \tag{1.18}$$

où A_n est le nombre de clients qui arrivent pendant le service du $n^{ème}$ client. $\{A_n, n \geq 0\}$ est une suite de variables aléatoires indépendantes et identiquement distribuées de loi donnée par

$$P(A_n = k) = \int_0^\infty \frac{(\lambda t)^k}{k!} e^{-\lambda t} dB(t),\ k = 0, 1, 2, ...;\ n \geq 1 \tag{1.19}$$

En particulier,

$$E[A_n] = \rho = \frac{\lambda}{\mu} \tag{1.20}$$

$$a(z) = \sum_{k=0}^{\infty} z^k P(A_n = k) = B^*(\lambda - \lambda z),\ 0 \leq z \leq 1. \tag{1.21}$$

la fonction génératrice de la distribution stationnaire est donnée par la formule suivante appelée formule de Pollaczek-Khintchine. :

$$\pi(z) = (1-\rho)\frac{B^*(\lambda - \lambda z)(1-z)}{B^*(\lambda - \lambda z) - z},\ pour z < 1. \tag{1.22}$$

La distribution stationnaire existe si $\rho < 1$ et de plus

$$\pi_k = \lim_{k\to\infty} P(X(t) = k) = \lim_{k\to\infty} P(X_n = k) \tag{1.23}$$

En d'autre termes les probabilités stationnaires π_n de la chaîne de Markov induite $\{X_n;\ n = 1, 2, ...\}$ sont identiques aux probabilités stationnaires P_n du processus à temps continu $\{X(t);\ t \geq 0\}$.
Notons cependant que ce résultat, valable pour le système $M/G/1$, ne s'étend généralement pas à d'autres processus non markoviens.

Paramètres de performances

Tous les paramètres de performances sont calculés dans le cas où la file est stable ($\lambda < \mu$) et pour le régime stationnaire de la file.
La formule de Pollaczek-Khintchine nous donne l'expression du nombre moyen de clients dans la file en fonction des deux premiers moments de la loi G :

$$\rho + \frac{\rho^2(1 + \lambda^2 Var(Y))}{2(1-\rho)} \tag{1.24}$$

1.7.2 File d'attente $G/M/m$

Le processus $\{X(t);\ t \geq 0\}$ où $X(t)$ est le nombre de clients dans le système à la date t, n'est pas markoviens. Utilisons la même méthode que précédemment en prenant pour suite $\{t_n\}$ les instants d'arrivées des clients dans le systèmes. Le processus $X_n = X(t_n - 0)$ (nombre de clients dans le système à l'instant qui précède immédiatement l'arrivée du $n^{ème}$ client) est une chaine de Markov. On a :

$$X_{n+1} = X_n - D_{n+1} + 1,\ n \geq 0 \tag{1.25}$$

où D_n est le nombre de clients servis entre les arrivées des clients C_{n-1} et C_n. Soit $P_{ij} = P(X_{n+1} = j/X_n = i)$, la probabilité pour que dans l'intervalle $[t_n,\ t_{n+1}]$ il y ait exactement $i + 1 - j$ départs.

1. Si $j > i + 1$ on a : $P_{ij} = 0$

2. Si $j \leq i + 1 \leq m$, tous les clients sont en cours de service et

$$P_{ij} = \int_0^\infty C_{i+1}^{i+1-j}(1 - e^{-\mu x})^{i+1-j} e^{-\mu x j} dA(x) \tag{1.26}$$

3. Si $m \leq j \leq i + 1,\ i \geq m$

$$P_{ij} = \int_0^\infty \frac{(m\mu x)^{i+1-j}}{(i+1-j)!} e^{-\mu x m} dA(x) \tag{1.27}$$

4. Si $j < m < i + 1$

$$P_{ij} = \int_0^\infty C_m^j e^{-j\mu x} \Big[\int_0^\infty \frac{(m\mu y)^{i-m}}{(i-m)!} (e^{-\mu y} - e^{-\mu t})^{m-j} m y dy \Big] dA(x). \tag{1.28}$$

Connaissons la matrice de transitions $P = \|P_{ij}\|$, on peut obtenir la distribution stationnaire $\{\pi_k\}$ de la chaine $\{X_n\}$, qui est solution du système d'équations algébriques : $\Pi = \Pi P$ où $\Pi = (\pi_1, \pi_2, ...)$.
Si $\lambda < m\mu$, la distribution stationnaire de la chaine de Markov $\{X_n\}$ est de la forme :

$$\pi_k = \begin{cases} K\sigma^k & k \geq m - 1 \\ \\ KR_k\sigma^{m-1} & m \leq m - 1 \end{cases} \tag{1.29}$$

où σ est l'unique solution de l'équation fonctionnelle $\sigma = \hat{A}(m\mu - m\mu\sigma)$ dans le domaine $0 < \sigma < 1$.
Les équations R_k peuvent être déterminées récursivement

$$\begin{cases} R_{m-1} = 1 \\ \\ R_{k-1} = \frac{R_k - \sum_{i=k}^{m-1} - R_i P_{ik} - \sum_{i=m-1}^{\infty} \sigma^{i+1-m} P_{ik}}{P_{k-1,k}} & k = 1, m^- - 1 \end{cases} \tag{1.30}$$

et la constante $K = \left[\frac{\sigma^{m-1}}{1-\sigma} + \sigma^{m-1} \sum_{k=0}^{m-2} R_k\right]^{-1}$.
La distribution stationnaire de la chaine de Markov incluse $\{X_n\}$ pour le système $G/M/1$ est géométrique $\pi_k = (1 - \sigma)\sigma^k,\ k = 0,\ 1, ...$ où σ est la solution unique de l'équation $\sigma = \hat{A}(\mu - \mu\sigma)$ dans le domaine $0 < \sigma < 1$.

1.7.3 File d'attente $G/G/1$

La file $G/G/1$ généralise les modèles de files d'attente à un seul serveur (les inter-arrivées et les services sont arbitraires). Cette généralisation rend impossible la détermination des résultats exacts des paramètres de performance pour ce modèle. En effet, un modèle contenant deux lois générales (non Markoviennes) ne peut ni être résolu par la détermination des probabilités transitoires, ni par l'intermédiaire d'une chaîne de Markov induite [27]. Les principaux résultats de cette file sont des bornes inférieures et supérieures qui encadrent le temps moyen d'attente, noté $\bar{w}$, dans la file [60] :

$$\frac{\lambda\sigma_X^2 - X(2-\rho)}{2(1-\rho)} \geq \bar{w} \geq \frac{\lambda(\sigma_T^2 + \sigma_X^2)}{2(1-\rho)} \tag{1.31}$$

avec X : le temps de service moyen.
σ_X : l'écart type de la variable aléatoire qui décrit le temps de service.
σ_T : l'écart type de la variable aléatoire qui décrit les inter-arrivées.

1.8 Conclusion

Dans ce chapitre, nous avons rappelé et présenté les concepts et techniques de base de la théorie de files d'attente classiques. Plus précisément, nous avons exposé quelques modèles d'attente particuliers et nous avons déterminé leurs principales caractéristiques, tout en abordant les files markoviennes et les files non markoviennes.

L'évolution rapide des systèmes informatiques et de réseaux de télécommunication ont montré les limites de la théorie des files d'attente dites classiques qui ne permettent pas d'expliquer le comportement stochastique de certains systèmes complexes comme les systèmes téléphoniques. Ce qui a conduit certains chercheurs à développer d'autres modèles plus élaborés qu'on appelle "files d'attente avec rappels". Ces systèmes d'attente avec rappels feront l'objet du chapitre suivant.

Chapitre 2

SYSTÈMES DE FILES D'ATTENTE AVEC RAPPELS

2.1 Introduction

Dans la théorie des files d'attente classique, il est supposé qu'un client qui ne peut pas obtenir son service immédiatement dès son arrivée, rejoint la file d'attente ou quitte le système définitivement. Une situation intermédiaire envisage la possibilité pour un client qui trouve le(s) serveur(s) occupé(s) de rappeler ultérieurement à des intervalles aléatoires et autant de fois que nécessaire, jusqu'à ce qu'il trouve le serveur libre, on parle alors de système de files d'attente "avec rappels" ou encore "avec répétitions d'appels". Dans la terminologie Anglo-Saxonne, on utilise depuis des années le terme "Retrial Queueing Systems".

Les files d'attente avec rappels ont été largement utilisées pour modéliser de nombreux problèmes dans les systèmes de communications téléphoniques, informatiques, des réseaux locaux et des situations de la vie quotidienne. Les progrès récents dans ce domaine sont résumés dans les articles de synthèse de Aïssani [4], Kulkani et Liang [65], Templeton [[94],1999] et dans les monographies de Falin et Templeton [46], Artalejo et Gòmez-Corral[[19],2008], Gòmez-Corral et Ramalhoto [50], Rodrigo [[84], 2006] et Kim [57].
Pour identifier un système de files d'attente avec rappels, il faut ajouter une nouvelle spécification concernant le processus de répétition de demandes de service.

2.2 Description du modèle

Les systèmes de files d'attente avec rappels sont caractérisés par le fait qu'un client qui arrive et trouve le serveur et la salle d'attente occupés quitte le système définitivement, ou rappelle ultérieurement à des instants aléatoires. Un client qui attend pour rappeler est dit en "orbite" (devient source d'appels secondaires) et refait sa tentative d'avoir un service ultérieurement selon une politique de rappels spécifiée. Si le client qui arrive trouve le serveur libre il prend son service et quitte le système.

Un système de files d'attente avec rappels contient un espace de service composé de s ($s \geq 1$) dispositifs de service et d'un espace d'attente (buffer) ayant $m - s$ ($m \geq s$) positions d'attente et d'une orbite de capacité finie ou infinie. A l'arrivée d'un client primaire, s'il y a un ou plusieurs serveurs libres, et en bon état, le client sera servi immédiatement et quittera le système à la fin de son service. Sinon, s'il y a une position d'attente libre, le client rejoint la file d'attente. Lorsque tous les serveurs et positions d'attente sont occupés, le client quittera le système définitivement avec une probabilité $1 - H_0$ ou entre en orbite avec la probabilité H_0 et rappelle ultérieurement, après un temps aléatoire suivant une loi de probabilité et une intensité de rappels bien définie (rappels constants, rappels classiques, ou bien rappels linéaires, ...). Chacun de ces clients secondaires est traité comme un client primaire.
Le schéma général d'un système avec rappels est représenté par la Figure 2.1.

La classification des modèles avec rappels reposera sur les notations de Yang et Templeton(1987) : $A/B/s/M/O/H$.

Comme pour les notations de Kendall, A désigne le type de la loi des inter-arrivées primaires, B celui de la loi de service, s est le nombre de serveurs, M est la capacité du système (attente plus service), et O est la capacité de l'orbite (qui peut être supprimé lorsque sa capacité est infinie). La séquence $H = \{H_i,\ i \geq 0\}$ est la fonction de persévérance, où H_i est la probabilité pour qu'un abonné fasse une $(i+1)^{ème}$ tentative de rappel, après une $i^{ème}$ tentative avortée. Lorsque tous les clients sont persévérants, $H_i = 1$ pour tout i, le symbole H pourra être également supprimé.

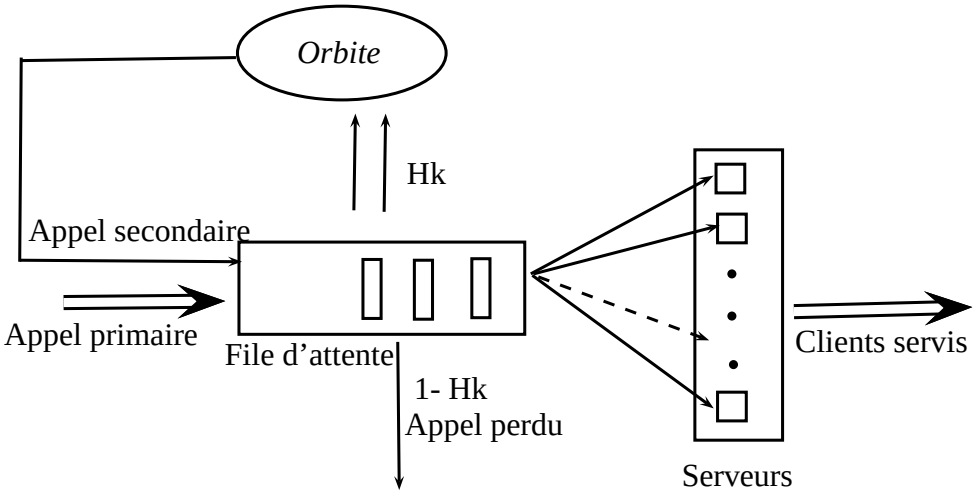

FIGURE 2.1 – Le schéma général d'un système d'attente avec rappels

2.3 Exemples d'applications de modèle de files d'attente a

Dans cette section, nous présentons quelques exemples de problèmes dus aux applications du monde réel pouvant être modéliser par des files d'attente avec rappels.

2.3.1 Centre d'appel

Un centre d'appels ou call-center est important pour une entreprise car il fournit un canal pour les clients pour pouvoir contacter l'entreprise. Dans un centre d'appels, les agents sont les personnes qui répondent à la demande des clients. Lorsqu'un client effectue un appel téléphonique, s'il y a un agent d'appel inactif, il répond immédiatement au client. Si tous les agents sont occupés, le client peut entendre certains messages par exemple "le système est occupé pour le moment, veuillez patienter un instant". A ce moment, le client décide de raccrocher le téléphone immédiatement et peut rappeler de nouveau après un certain temps aléatoire, ou bien de continuer à entendre le message. L'une des mesures de performance la plus importante pour un centre d'appels est la probabilité de blocage.

Un modèle de files d'attente avec rappels est le modèle mathématique le plus approprié pour la conception de centre d'appels .

2.3.2 Réseaux à commutation par paquets

Considérons un réseau de communications d'ordinateurs dans lequel on trouve un ensemble d'interfaces IMP (Interface Message Processors) reliées entre elles par des câbles. Un ordinateur principal est connecté à l'une de ces interfaces. Si l'ordinateur veut envoyer un message à un autre ordinateur principal, il doit en premier lieu envoyer le message avec l'adresse de destination à l'interface à laquelle il est connecté. L'interface à son tour envoie le message à l'ordinateur destinataire directement si elle y est connectée, ou indirectement via d'autres interfaces. Considérons une interface à laquelle un ordinateur principal est connecté. Les messages arrivent de l'extérieur selon un processus aléatoire. Après la réception du message, l'ordinateur l'envoie immédiatement à son interface. S'il y a un tampon libre, le message est accepté. Dans le cas contraire, le message est rejeté et l'ordinateur doit réessayer une autre fois après une période de temps. S'il existe des tampons libres, le message rejeté sera stocké dans un tampon de l'ordinateur principal. Dans le cas contraire, le message est considéré comme perdu. On peut se poser les questions suivantes :
o Quelles sont les probabilités pour qu'un message soit rejeté par l'interface et par l'ordinateur principal ?
o Quel est le nombre moyen de messages dans le tampon de IMP ?
o Quel est le nombre moyen de messages dans le tampon de l'ordinateur principal ?
o Quel est le temps d'attente d'un message dans le tampon de l'ordinateur principal ?

Le problème présenté peut être modélisé comme un système avec rappels à serveur unique (interface IMP) possédant des tampons (positions d'attente). Le nombre de tampons de l'ordinateur principal constitue la capacité de l'orbite.

2.3.3 Réseaux locaux : le protocole CSMA

Dans un réseau local (LAN), plusieurs nœuds partagent un lien physique (bus) afin de transmettre leurs données (paquets). En supposant que plusieurs nœuds envoient leurs paquets en même temps, une collision peut se produire, et tous les paquets seront détruits. Or il est possible d'éviter certaines collisions si l'on fait en sorte que chaque nœud écoute ce qui se

passe sur le réseau avant d'émettre et évite d'émettre lorsque le réseau est occupé compte tenu de cette propriété, CSMA "Carrier Sense Multiple Access" (Écoute d'un Support à Accès Multiple) protocoles ont été développés. Il s'agit d'un ensemble de protocoles d'accès à un média. Ceux-ci vérifient que le support est disponible avant de commencer l'envoi d'une trame. Ils permettent également de détecter ou bien éviter les collisions de messages dans les transmissions.

Quand un nœud veut envoyer des données à un autre nœud, sa carte écoute le bus pour s'assurer qu'aucun signal n'est en cours de transmission ; si le réseau est silencieux, elle émet sa trame sur le bus et les autres seront stockés dans le buffer. Chaque nœud examine l'adresse du destinataire. Si la trame ne lui est pas destinée, il l'ignore. Si la trame lui est destiné, il lit les données, vérifie qu'il n'y a pas eu d'erreur, et envoie un accusé de réception à l'émetteur qui peut alors envoyer la trame suivante. Si deux nœuds émettent un message simultanément, la collision entre les trames provoque un signal d'interférence qui se propage sur le bus et qui est reconnu par les émetteurs. Le premier émetteur détectant la collision émet un signal indiquant celle-ci aux autres nœuds. Les transmissions sont alors arrêtées ; les nœuds qui veulent émettre doivent attendre une durée aléatoire avant de chercher à émettre de nouveau. Le processus se répète jusqu'à ce qu'un nœud puisse émettre sa trame sans qu'il y ait collision.

Parce que les collisions dans les protocoles CSMA ne peuvent être totalement évitées, le phénomène avec rappels se produit dans les réseaux locaux. Les modèles de files d'attente avec rappels sont donc jugés plus appropriés que les modèles de files d'attente classiques dans la modélisation et l'analyse de performances de ces protocoles.

Ce problème peut-être modélisé par un système de files d'attente avec rappels à un seul serveur, qui est le bus, et les buffers des stations représentent l'orbite.

2.4 Modèles markoviens

Les modèles Markoviens sont des systèmes où les inter-arrivées primaires, les durées de service et les temps inter-rappels sont des variables aléatoires

indépendantes et exponentiellement distribuées.

2.4.1 Système $M/M/n$ avec rappels

On considère un système de files d'attente avec rappels sans positions d'attente. Le service est assuré par n serveurs ($n \geq 1$). Les clients primaires arrivent selon un processus de Poisson de taux λ. Les durées de service suivent une loi exponentielle de taux μ. Les temps entre deux rappels consécutifs sont également exponentiels de paramètre ν.
L'état du système peut être décrit par le processus markovien $X(t) = \{C(t),\ R(t)\}$, d'espace d'états $S = \{0, 1, ..., n\} \times N$.
Où $C(t)$ est le nombre de clients en cours de service à la date t et $R(t)$ est le nombre de clients en orbite à l'instant t.

Les conditions d'existence d'un régime stationnaire ont été établies par Falin [43]. Dans ce cas, les probabilités stationnaires

$$P_{ij} = P_{ij}(t) = P(C(t) = i, R(t) = j),\ i = 0, ..., n,\ j \geq 0.$$

Les probabilités de transitions à l'état stationnaire sont données par :

Pour $0 \leq i \leq n-1$

$$P_{ij}(kl) = \begin{cases} \lambda & si(k,l) = (i+1, j), \\ i\mu & si(k,l) = (i-1, j), \\ j\nu & si(k,l) = (i+1, j-1), \\ -(\lambda + i\mu + j\nu) & si(k,l) = (i,j), \\ 0 & sinon \end{cases} \tag{2.1}$$

Pour $i = n$

$$P_{nj}(kl) = \begin{cases} \lambda & si(k,l) = (n, j+1), \\ n\mu & si(k,l) = (n-1, j), \\ -(\lambda + n\mu) & si(n,l) = (n,j), \\ 0 & sinon \end{cases} \tag{2.2}$$

Dans le cas où Le service est assuré par un seul serveur c.à.d $n = 1$, sous la condition $\rho < 1$, les probabilités stationnaires existent et sont données par

[43].

$$P_{0j} = \frac{\rho^j}{j!\nu^j} \prod_{i=0}^{j-1} (\lambda + i\nu)(1-\rho)^{1+\frac{\lambda}{\nu}}, \tag{2.3}$$

$$P_{1j} = \frac{\rho^{j+1}}{j!\nu^j} \prod_{i=0}^{j} (\lambda + i\nu)(1-\rho)^{1+\frac{\lambda}{\nu}}, \tag{2.4}$$

de fonctions génératrices

$$P_0(z) = \sum_{n=0}^{\infty} z^n P_{0n} = (1-\rho)\big(\frac{1-\rho}{1-\rho z}\big)^{\frac{\lambda}{\nu}}, \tag{2.5}$$

$$P_1(z) = \sum_{n=0}^{\infty} z^n P_{1n} = \rho\big(\frac{1-\rho}{1-\rho z}\big)^{\frac{\lambda}{\nu}}. \tag{2.6}$$

Toutes les mesures de performance s'obtiennent en utilisant les fonctions génératrices (voir par exemple [43].

2.5 Modèles semi-markoviens

2.5.1 Système $M/G/1$ avec rappels

Le modèle $M/G/1$ avec rappels est le modèle le plus étudié par les spécialistes. Il existe une littérature abondante sur ses diverses propriétés [43, 51, 64, 101]...
Les clients arrivent dans le système selon un processus de Poisson de taux $\lambda > 0$. Le service des clients est assuré par un seul serveur. La durée de service est de loi générale, de distribution $B(x)$, de transformée de Laplace-Stieltjes $B^*(s)$, $Re(s) > 0$ et de moments $\beta_k = (-1)^k B^*(0)$. La durée entre deux rappels successifs d'une même source secondaire est exponentiellement distribuée de paramètre ν.

Le premier résultat sur les systèmes $M/G/1$ avec rappels a été obtenu par Keilson, Cozzolono et Young [56], en utilisant la méthode de la variable auxiliaire. Ils ont obtenu les probabilités d'états et les fonctions génératrices du nombre de clients dans le système.

L'état du système peut-être décrit par le processus

$$X(t) = \begin{cases} R(t) & si\ S(t) = 0, \\ \{S(t),\ R(t),\ \xi(t)\} & si\ S(t) = 1. \end{cases} \tag{2.7}$$

Où $R(t)$ est le nombre de clients en orbite à la date t et $\xi(t)$ est une variable aléatoire à valeurs dans R^+ et désignant :
- La durée de service écoulé à la date t.
- La durée de service résiduelle à la date t.

Chaîne de Markov induite

Ce paragraphe introduit une technique qui permet d'étudier des processus qui ne sont pas forcément markoviens. Cette technique a été utilisée pour la première fois par Choo et Conolly [31].
Pour tout $n \geq 0$, soit X_i la chaîne de Markov induite aux instants de départs, où $X_i = X_{t_i}$ est le nombre de clients dans le système juste après le $i^{ème}$ départ. Il est clair que :

$$X_{i+1} = X_i + A_{i+1} - \delta_{X_i}, \tag{2.8}$$

où A_i est le nombre de clients qui arrivent pendant le service du $i^{ème}$ client. $\{A_n, n \geq 0\}$ est une suite de variables aléatoires indépendantes et identiquement distribuées de loi donnée par

$$P(A_i = k) = \int_0^\infty \frac{(\lambda t)^k}{k!} e^{-\lambda t} dB(t),\ k = 0, 1, 2, ...;\ n \geq 1, \tag{2.9}$$

de fonction génératrice

$$a(z) = \sum_{k=0}^{\infty} z^k P(A_n = k) = B^*(\lambda - \lambda z),\ 0 \leq z \leq 1, \tag{2.10}$$

et de moyenne $E[A_i] = \rho = \lambda\beta_1$.
δ_{X_i} est la variable de Bernoulli définie par :

$$\delta_{X_i} = \begin{cases} 1 & si\ le\ (i+1)^{ème}\ client\ servi\ provient\ de\ l'orbite \\ 0 & sinon. \end{cases} \tag{2.11}$$

La distribution conditionnelle de δ_{X_i} est donnée par :

$$P\{\delta_{X_n} = 0 / X_n = k\} = \frac{\lambda}{\lambda + k\nu}, \tag{2.12}$$

$$P\{\delta_{X_n} = 1 / X_n = k\} = \frac{k\nu}{\lambda + k\nu}. \tag{2.13}$$

Caractéristiques moyennes

Nombre moyen de clients dans le système :

$$\bar{n} = \rho + \frac{\lambda^2 \beta_2}{2(1-\rho} + \frac{\lambda\rho}{\mu(1-\rho)}. \tag{2.14}$$

Nombre moyen de clients en orbite :

$$\bar{n_0} = \bar{n} - \rho = \frac{\lambda^2 \beta_2}{2(1-\rho} + \frac{\lambda\rho}{\mu(1-\rho)}. \tag{2.15}$$

Temps moyen d'attente d'un client : Pour trouver le temps moyen d'attente $\bar{w}$, on utilise la formule de Little $\bar{n} = \bar{w}\lambda$. On aura :

$$\bar{w} = \frac{\lambda^2 \beta_2}{2(1-\rho} + \frac{\rho}{\mu(1-\rho)}. \tag{2.16}$$

Régime stationnaire

La fonction génératrice du nombre de clients dans le système en régime stationnaire ($\rho < 1$) est :

$$\phi(z) = (1-\rho)\frac{B^*(\lambda - \lambda z)(1-z)}{B^*(\lambda - \lambda z) - z} exp\left(\frac{-\lambda}{\nu}\int_1^z \frac{1 - B^*(\lambda - \lambda x)}{B^*(\lambda - \lambda x) - x} dx\right),\ pour z < 1. \tag{2.17}$$

En posant

$$\psi(z) = exp\left(\frac{-\lambda}{\nu}\int_0^z \frac{1 - B^*(\lambda - \lambda x)}{B^*(\lambda - \lambda x) - x} dx\right), \tag{2.18}$$

on obtient

$$\phi(z) = \pi(z)\frac{\psi(z)}{\psi(1)} \tag{2.19}$$

Cette propriété de décomposition signifie que le nombre de clients dans un système $M/G/1$ avec rappels est égal au nombre de clients dans un système $M/G/1$ classique plus une variable aléatoire positive de fonction génératrice $\frac{\psi(z)}{\psi(1)}$.

2.6 Période d'activité

La période d'activité d'un système de files d'attente avec rappels est définie comme étant l'intervalle de temps qui débute à l'instant t_0 d'arrivée d'un premier client dans un système vide jusqu'à l'instant t_1, où le système redevient vide pour la première fois :

$t_1 = inf\{t : t > 0, S(t) = 0, R(t) = o\}$.

On notera $L = t_1 - t_0$, la durée de la période d'activité du système, cette période est constituée d'une alternance de périodes d'activité et d'inactivité du serveur.

Falin [42] procède à une étude détaillée de la période d'activité en utilisant la méthode des catastrophes qui permet de donner des résultats plus explicites dans le cas $M/G/1$. La transformée de Laplace de la distribution conjointe de la période L et du nombre de clients I servis au cours de cette période :

$$E[e^{-sL}z^I] = \frac{s+\lambda}{\lambda} \tag{2.20}$$

$$-\frac{\nu}{\lambda}\int_0^{\pi_\infty(s,z)} exp\{-\frac{\lambda}{\nu}\int_0^x \frac{s+\lambda-\lambda zB^*(s+\lambda-\lambda y)}{zB^*(s+\lambda-\lambda y)-y}dy\}\frac{dx}{zB^*(s+\lambda-\lambda x)-x}$$

où $\pi_\infty(s,y) = E[e^{-sL_\infty}y^{I_\infty}]$ est l'équivalent de $\pi(s,y)$ pour le système classique $M/G/1$ avec attente. Cette fonction est l'unique solution de l'équation :

$$yB^*(s+\lambda-\lambda\pi_\infty) - \pi_\infty = 0.$$

En outre,

a. Si $\rho > 1$, alors $P(L = \infty) = P(I = \infty) > 0$.

b. Si $\rho = 1$, alors $E[I] = E[L] = \infty$.

c. Si $\rho < 1$, alors

$$E[L] = -\frac{1}{\lambda}\frac{1}{\lambda(1-\rho)}exp\left(\frac{\lambda}{\nu}\int_0^1 \frac{1-B^*(\lambda-\lambda x)}{B^*(\lambda-\lambda x)-x}dx\right), \tag{2.21}$$

$$E[I] = \frac{1}{1-\rho} exp\left(\frac{\lambda}{\nu}\int_0^1 \frac{1-B^*(\lambda-\lambda x)}{B^*(\lambda-\lambda x)-x}dx\right). \tag{2.22}$$

Les probabilités de transition à une étape de la chaîne sont :

$$P_{ij} = \frac{\lambda}{\lambda+i\mu}P_{j-1} + \frac{i\mu}{\lambda+i\mu}P_{j-i+1}. \tag{2.23}$$

la fonction génératrice de la distribution stationnaire est donnée par la formule suivante appelée formule de Pollaczek-Khintchine. :

$$\pi(z) = (1-\rho)\frac{B^*(\lambda-\lambda z)(1-z)}{B^*(\lambda-\lambda z)-z},\ pour\ z<1. \tag{2.24}$$

2.7 Notes bibliographiques

Les files d'attente avec rappels à un seul serveur avec appels prioritaires ont été étudiées par plusieurs auteurs [45, 68]. Différents problèmes théoriques pour les files multi-classes ont été exposés et résolus par Kulkarni [63]. Langaris et Moutzoukis [69] ont étudié le système $M/G/1$ avec rappels et arrivées par groupes, deux types de clients et avec vacation du serveur, ils ont obtenu la distribution stationnaire du système. Des contributions récentes sur certaines situations de files d'attente avec rappels et arrivées par groupes incluent les travaux de Aissani [5], Krishna Kumar et Pavai Madheswari [66], Artalejo et Atencia [16] et aussi Atencia et autres [23]. Une variété importante de systèmes avec rappels non fiable existe, nous pouvons citer les travaux de Aissani [3, 6], Kulkarni et choi [64], Artalejo [8], Oukid [81, 82], Djellab [37] et Almasi Roszik et Sztrik [7] . Récemment, Jain et autres [55] en 2007 ont étudié le système d'attente M/G/1 avec rappels, pannes et lancement. D'autre part, Falin [44] en 2010 a étudié un modèle $M/G/1$ avec rappels et pannes lorsque le temps de service et les temps de réparation ont des distributions générales, En 2003, Wu, et autres [99] sont les premiers à considérer deux orbites dans le système $M/G/1$ avec rappels. La première orbite (I) est dans le sens traditionnel avec la discipline FCFS. La deuxième orbite (II) est réservée spécifiquement pour clients interrompus par une panne de serveur. Temps de réparation et de rappels de l'orbite (I) sont généralement distribués tandis que les rappels de l'orbite (II) sont distribués de façon exponentielle. Chakravarthy et Dudin [29] ont étudié un modèle de files d'attente avec rappels et avec deux types de clients dans lesquels les arrivées suivent des processus

markovien. Pour illustrer le rôle actif des files d'attente avec rappels au cours de ces dernières années, citons quelques articles récents publiés dans la revue "Applied Mathamatical Modelling" [20, 32, 33, 67, 88, 95, 98]. Pour les systèmes avec temps de rappels général, Yang et autres [100], en appliquant la propriété de la décomposition stochastique, proposent une méthode d'approximation efficace pour le calcul des probabilités stationnaires et les mesures de performance du système. Récemment, Artalejo et Phung-Duc [[21], 2011] ont réalisé une étude détaillée de la file d'attente $M/M/1$ avec rappels à communication bidirectionnelle. Les résultats obtenus sont des expressions explicites pour la distribution stationnaire conjointe de l'état du serveur et du nombre de clients en orbite, ainsi que les moments factoriels partiels. La plupart des formules explicites dans [21] sont exprimées en termes de séries hypergéométriques, en accord avec le rôle particulier joué par ces fonctions spéciales dans le calcul de solutions analytiques pour beaucoup d'autres files d'attente avec rappels [17, 30, 52, 58, 62, 83]. En 2013 J. R. Artalejo et T. Phung-Duc [22] ont étudié le comportement d'un état d'équilibre d'une file $M/G/1$ avec rappels dans laquelle il y'a deux flux d'arrivées entrants à savoir les appels effectués par des clients réguliers et les appels sortants effectués par le serveur lorsqu'il est inactif. Ils ont effectué une analyse stationnaire du système, y compris la condition de stabilité, la chaîne de Markov induite, la distribution stationnaire de l'état du serveur, le nombre de clients en orbite et le calcul des premiers moments. Ils ont aussi obtenu les résultats asymptotiques pour le nombre de clients en orbite.

2.8 Conclusion

Dans ce chapitre nous avons présenté les résultats existants ainsi que les différentes méthodes utilisées pour analyser les modèles de files d'attente avec rappels.

Nous avons vu que la résolution de problèmes mathématiques de ces modèles sous des hypothèses différentes de celles des modèles classiques est assez difficile. Compte tenu de ces difficultés, plusieurs auteurs ont tenté de développer des méthodes approximatives d'analyse de ce type de systèmes.

Chapitre 3

MARTINGALES A TEMPS DISCRET

3.1 Introduction

La théorie des martingales est l'un des outils les plus puissants de la théorie des probabilités. Nous présentons ici un bref aperçu des définitions et des résultats élémentaires concernant cette théorie. Pour plus de détails on pourra consulter les ouvrages de williams [97], de Neveu [75] ou C. Dellacherie [36] et Rogers et Williams [85, ?].

La théorie des martingales a son origine dans l'étude des jeux : elle modélise d'une part le caractère aléatoire d'un phénomène mais aussi son évolution dans le temps. On étudie ici le temps discret, c'est à dire lorsque le paramètre de temps est un entier.

3.2 Définitions - Généralités

Dans tout ce qui suit, le triplet $(\Omega, \Im, P)$ désigne un espace de probabilité.

3.2.1 Filtrations et martingales

Définition 3.1. On appelle filtration sur $(\Omega, \Im, P)$ toute suite croissante $(\Im_n)_{n\geq 0}$ de sous-tribus de $\Im$: $\Im_0 \subset \Im_1 \subset ... \subset \Im_n$

Interprétation

La filtration est un formalisme probabiliste pour décrire l'information dont on dispose. La dernière propriété traduit simplement que l'information augmente au cours du temps.
Pour un processus aléatoire $X = (X_t)_{t\geq 0}$, il y a une filtration naturelle, constituée des sous-tribus engendrées par les variables aléatoires $X_s,\ s \leq t$: $\Im_t = \sigma(X_s, s \leq t)$.

Définition 3.2. Soit $(X_n)_{n\geq 0}$ une suite de variables aléatoires et $(\Im_n)_{n\geq 0}$ une filtration. On dit que la suite $(X_n)_{n\geq 0}$ est $(\Im_n)_{n\geq 0}$ adaptée si pour tout n, X_n est $\Im_n$-mesurable.

Définition 3.3. Soit $(\Im_n)_{n\geq 0}$ une filtration et $(X_n)_{n\geq 0}$ une suite de variables aléatoires. On dit que la suite $(X_n)_{n\geq 0}$ est une martingale adaptée à la filtration $(\Im_n)_{n\geq 0}$ si

1. la suite est $(\Im_n)_{n\geq 0}$ adaptée.

2. Pour tout n, X_n est intégrable.

3. Pour tout n, $X_n = E(X_{n+1}/\Im_n)$, presque sûrement.

On dit que X est une sous-martingale si pour tout entier n, $X_n \geq E(X_{n+1}/\Im_n)$, presque sûrement.

On dit que X est une sur-martingale si pour tout entier n, $X_n \leq E(X_{n+1}/\Im_n)$, presque sûrement.

Remarque

La plupart du temps, la filtration est en fait définie à partir de la suite $(X_n)_{n\geq 0}$ par $\Im_n = \sigma(X_0, ..., X_n)$ (tribu du passé avant l'instant n). On dit que c'est la filtration associée au processus $(X_n)_{n\geq 0}$, ou simplement la filtration naturelle.
Si on ne précise pas la filtration quand on parle d'une martingale, c'est que l'on considère implicitement celle-ci.

3.2.2 Exemples

• Si $(\Im_n)_{n\in N}$ est une filtration et si X est une variable aléatoire intégrable, alors $X_n = E(X|\Im_n)$ définit une martingale. C'est la martingale de Doob.

• Si $(X_n)_{n\geq 0}$ est un processus adapté intégrable, alors $S_n = X_0 + ... + X_n$ définit une martingale si et seulement si $E(X_{n+1}|\Im_n) = 0$. En particulier si $(X_n)_{n\geq 0}$ est une suite de variables aléatoires indépendantes centrées telles que $X_0 = 0$ alors $S_n = X_0 + ... + X_n$ est une martingale par rapport à $\Im_n = \sigma(X_0, ..., X_n)$

Remarque

Une martingale est toujours adaptée à sa filtration naturelle.

3.2.3 Interprétation dans le contexte d'un jeu d'argent

Les martingales sont utilisées pour modéliser les jeux de hasard équitables. Supposons en effet qu'à un jeu de hasard, la v.a. X_n représente la fortune d'un joueur à la $n^{ème}$ partie. La propriété de martingale stipule que, si le joueur a la connaissance de l'évolution passée de sa fortune (c'est à dire jusqu'au temps $n-1$) alors l'espérance de la valeur de sa fortune après la partie suivante (la $n^{ème}$) est égale à celle actuelle (i.e. à $n-1$). En moyenne ses gains restent inchangés. Donc :

• Une martingale est un jeu équilibré : on ne peut espérer ni perdre ni gagner.

• Une sous-martingale (resp. sur-martingale) est un jeu avantageux (resp. désavantageux).

Proposition 3.4. *1. Une martingale est constante en moyenne,* $E[X_n] = E[X_0]$ *pour tout* $\forall; \in N$.

2. $E(X_{n+m}/\Im_n) = X_n$, *p.s.,* $\forall n,\ m \in N$.

3.3 Théorèmes d'arrêt

Il y a en fait un seul véritable théorème d'arrêt, connu sous le nom d'Optional Sampling Theorem, mais plusieurs résultats liés à l'échantillonnage aléatoire des (sous-, sur-) martingales. Dans sa version originelle, le théorème d'arrêt traduit le fait que si un jeu est équilibré ou au contraire favorable à l'une des parties, cette propriété reste valable si l'on échantillonne (sampling) le processus à des instants aléatoires, pour peu que ces instants soient choisis de façon non anticipative (optional time).
Dans un premier temps, précisons la notion d'échantillonnage aléatoire non anticipatif, plus communément appelé temps d'arrêt (optional time ou stopping time).

3.3.1 Temps d'arrêt

Lorsqu'on considère un processus aléatoire, on s'intéresse souvent à des instants particuliers tels que celui pour lequel un certain seuil est atteint. Bien sûr, un tel instant dépend de chaque trajectoire du processus et est aléatoire. On définit ainsi la notion de temps d'arrêt :

Définition 3.5. Soit $(\Im_n)_{n\geq 0}$ une filtration sur $(\Omega, \Im, P)$ et T une variable aléatoire de Ω dans $\{0, 1, 2\ldots; +\infty\}$.
On dit que T est un temps d'arrêt si pour tout entier n, $\{T \leq n\} \in \Im_n$.
Dans ce cas, l'ensemble $\Im_T = \{A \in \Im,\ A \cap \{T \leq n\} \in \Im_n,\ \forall n \geq 0\}$ est une sous-tribu de $\Im$.

Interprétation

La définition d'un temps d'arrêt traduit que le choix de l'instant aléatoire $T(\omega)$ dépend seulement du passé (au sens large : incluant le présent).
Les temps d'arrêts vérifient les propriétés élémentaires suivantes : si S et T sont des temps d'arrêts, $\sup(S,T)$, $\inf(S,T)$ sont des temps d'arrêts, $S+T$ est un temps d'arrêt. Si S_n est une suite de temps d'arrêt, alors $\limsup_n S_n$ et $\liminf_n S_n$ sont des temps d'arrêts.

Notation

Dans toute la suite, nous désignerons le sup par le symbole $\vee$ et l'inf par le symbole $\wedge$. Ainsi $\sup(S,T) = S \vee T$, $\inf(S,T) = S \wedge T$.

3.3.2 Théorème d'arrêt

Théorème 3.6. *Soit $(X_n, n \in N)$ une martingale et soient S et T deux temps d'arrêts bornés avec $S \leq T$.*
Alors X_S et X_T sont intégrables et on a :

$$X_S = E(X_T/\Im_S) \tag{3.1}$$

Démonstration. Voir Neveu [75] □

3.3.3 Propriétés des martingales par rapport aux temps d'arrêts

Soit $(X_n, n \in N)$ un processus adapté à une filtration $(\Im_n)_{n\geq 0}$ et T un temps d'arrêt adapté à la même filtration. On définit un nouveau processus, appelé processus arrêté et noté $X_{T\wedge n},\ n \geq 0$, en posant

$$X_{T\wedge n}(w) = \begin{cases} X_n(w) & si\ n < T(w) \\ X_{T(w)}(w) & si\ n \geq T(w) \end{cases}$$

Par conséquent on a

$$X_{T\wedge n}(w) = X_n(w)1_{\{n<T(w)\}} + X_{T(w)}(w)_{\{T(w)\leq n\}}$$

ce qui montre bien que ce processus est encore adapté à la filtration $(\Im_n)_{n\geq 0}$. Une autre manière d'exprimer ce processus est la suivante

$$X_{T\wedge n} = X_0 1_{\{T=0\}} + X_1 1_{\{T=1\}} + ... + X_n 1_{\{T=n\}} + X_n 1_{\{T>n\}}$$

$$= X_0 + \sum_{k=0}^{n-1} X_{k+1} - X_k 1_{\{T>k\}}.$$

Théorème 3.7. *Soit $(\Im_n)_{n\geq 0}$ une filtration et $(X_n)_{n\geq 0}$ une martingale adaptée à la filtration $(\Im_n)_{n\geq 0}$. Soit T un temps d'arrêt adapté à la filtration $(\Im_n)_{n\geq 0}$. Alors la suite $(X_{T\wedge n})_{n\geq 0}$ est une martingale adaptée à la filtration $(\Im_n)_{n\geq 0}$.*

En particulier $E(X_0) = E(X_{T\wedge n}), \forall n$.

Le résultat est valable en remplaçant partout martingale par sous-martingale (resp. surmartingale) et l'égalité par $\leq$ (resp. $\leq$).

3.3.4 Décomposition

La première proposition décompose une sous-martingale en la somme d'une martingale et d'une suite croissante de variables aléatoires :

Théorème 3.8. *soit X une sous-martingale ; il existe une martingale M et un processus croissant prévisible A, nul en 0, uniques, tels que pour tout entier n,*

$$X_n = M_n + A_n.$$

Le processus A est appelé "compensateur" de X.

L'unicité de la décomposition s'écrit de la même façon, et remarquant que, si une telle décomposition a lieu, on doit avoir

$$E(X_{n+1} - X_n/\Im_n) = A_{n+1} - X_n,$$

ce qui caractérise A_n si l'on sait que $A_0 = 0$.

3.4 Convergence des martingales

Les sous-martingales et les surmartingales sont les généralisations aux processus des suites monotones. Si l'on impose que ces processus soient bornés, il est alors tout à fait naturel qu'ils convergent. L'énoncé du théorème est le suivant :

Théorème 3.9. *Soit X_n une martingale bornée dans L^1, i.e.* $\sup_{n\geq 0} E(|X_n|) < \infty$. *Alors $(X_n)_{n\geq 0}$ converge presque sûrement vers une variable aléatoire X_∞ intégrable.*

3.4.1 Convergence des martingales dans L^2

Théorème 3.10. *Soit $(X_n)_{n\geq 0}$ une martingale bornée dans L^2,*
i.e. $\sup_{n\geq 0} E(|X_n^2|) < \infty$. *Alors $(X_n)_{n\geq 0}$ converge dans L^2 et presque sûrement vers une variable aléatoire X_∞ telle que $X_n = E[X_\infty|\Im_n]$.*

En particulier $E[X_\infty] = E[X_0]$

3.4.2 Convergence des martingales dans L^1

Théorème 3.11. *Soit $(X_n)_{n\geq 0}$ une martingale. Les deux conditions suivantes sont équivalentes.*
(i) La suite X_n converge vers X_∞ p.s. et dans L^1.
(ii) Il existe une variable aléatoire Y intégrable telle que $X_n = E[Y|\Im_n]$ pour tout $n \in N$.
De plus, si ces conditions sont satisfaites, on peut prendre $Y = X_\infty$ dans (ii). On dit alors que la martingale est fermée.

Liant la convergence en probabilité et la convergence L^1.

Théorème 3.12. *Soit X_n une martingale bornée dans L^1, et soit X_∞ la limite de X_n lorsque $n \to \infty$. Les propositions suivantes sont équivalentes*
1. X_n converge dans L^1 vers X_∞.
2. X_n est uniformément intégrable.
3. $X_n = E[X_\infty|\Im_n]$.
4. Il existe une variable intégrable X telle que $X_n = E[X|\Im_n]$. De plus, dans ce cas, $X_\infty = E[X|\Im_\infty]$.

3.5 Conclusion

Dans ce chapitre, nous avons présenté un bref aperçu des définitions et des résultats élémentaires concernant la théorie des martingales ainsi que les théorèmes fondamentaux de cette théorie, c'est-à-dire essentiellement les théorèmes de convergence et d'arrêt.

Chapitre 4

ANALYSE DU SYSTÈME M/G/1 AVEC RAPPELS

4.1 Introduction

Dans ce chapitre, nous utilisons une approche différente des approches traditionnelles, celle des martingales pour analyser un système de files d'attente avec rappels. Plus précisément, nous montrons comment la technique proposée par F. Baccelli et A. M. Makowski [24] peut être étendue à la file d'attente $M/G/1$ avec rappels. En utilisant l'équation récursive du processus induit aux instants de départ de ce système, nous avons construit une martingale arrêtée au premier instant où le système redevient vide. Nous avons obtenu le résultat de stabilité de ce système et le nombre moyen de clients dans le système. Nous avons aussi démontré l'instabilité de ce système au sens de divergence de la chaîne de Markov induite.

Cette approche a été introduite dans la littérature de files d'attente par Baccelli et Makowski (1985) [24], qui ont étudié la stabilité et la période d'activité du système M/G/1 classique. À l'aide des martingales, Rosenkrantz [86] dérive une formule explicite pour la transformée de Laplace de la période d'occupation d'une file d'attente $M/G/1$. Kinateder et Lee [59] proposent une nouvelle approche pour le calcul de la transformée de Laplace de la longueur de la période d'occupation d'une file d'attente M/M/1 avec une charge de travail délimitée. Cependant, ils adoptent non seulement la méthode des martingales, mais aussi la technique par laquelle Feller [47] calcule la trans-

formée de Laplace pour un mouvement brownien standard avec l'absorption, et pour cette raison, la dérivation Kinateder et Lee est un peu longue. Jongho Bae [26] obtient en utilisant uniquement l'argument de martingale la formule explicite de la transformée de Laplace de la période d'occupation d'une file d'attente $M/M/1$ avec une charge de travail délimitée (barrage fini), une dérivation directe et beaucoup plus simple est fournie, en faisant usage du «Optional Stopping Theorem ». De même M.Roughan et C. Pearce(2002) ont développé une approche alternative, en utilisant la technique de Baccelli et Makowski, ils obtiennent la fonction génératrice de la distribution stationnaire du nombre de clients dans le système $M/G/1$ à plusieurs phases de service.

4.2 Description du modèle

Considérons un système de files d'attente $M/G/1$ avec rappels, dans lequel les clients arrivent suivant un processus de Poisson de paramètre λ, le temps de service d'un client est distribué selon une variable aléatoire générale de fonction de répartition $B(x)$ et de transformée de Laplace -Stieljes $\beta(\theta) = \int_0^\infty e^{-\theta t} dB(t)$. Soient les moments $\beta_k = (-1)^k \beta^{(k)}(0)$, l'intensité du trafic $\rho = \lambda\beta_1$. La durée entre deux rappels successifs d'une même source secondaire est exponentielle de taux ν. Nous supposons également que le temps des inter arrivées, de rappels et la durée de service sont mutuellement indépendants.
Considérons le processus $\{X_t,\ t \geq 0\}$, où $X(t)$ est le nombre de clients dans le système à l'instant t. Évidemment, ce processus n'est pas celui de Markov, mais il possède une chaîne de Markov induite.

A cet effet, nous considérons le processus $\{X_n,\ n \geq 1\}$: nombre de clients juste après le départ du $n^{ème}$ client.
Le processus $\{X_n = X(t_n);\ n = 1, 2, ...\}$ est un processus stochastique à temps discret (on ne s'intéresse qu'aux instants de fins de service), à espace d'état discret et sans mémoire. C'est donc une chaîne de Markov incluse pouvant être facilement étudiée, dont l'équation fondamental est :

$$X_{n+1} = X_n + A_{n+1} - \delta_{X_n},\ n \geq 0, \tag{4.1}$$

où A_{n+1} est le nombre de clients primaires arrivant dans le système pendant le service du $(n+1)^{ème}$ client. Elle ne dépend pas des événements qui se

sont produit avant l'instant t_{n+1} du début de service du $(n+1)^{ème}$ client (la numérotation se fait dans l'ordre de service).
La distribution de A_n est

$$P(A_n = k) = \int_0^\infty \frac{(\lambda t)^k}{k!} e^{-\lambda t} dB(t),\ k = 0, 1, 2, ...;\ n \geq 1, \tag{4.2}$$

de fonction génératrice

$$a(z) = \sum_{k=0}^{\infty} z^k P(A_n = k) = \beta(\lambda - \lambda z),\ 0 \leq z \leq 1. \tag{4.3}$$

La variable aléatoire δ_{X_n} est une variable de Bernoulli définie par :

$$\delta_{X_i} = \begin{cases} 1 & \textit{si le } (i+1)^{ème} \textit{ client servi provient de l'orbite} \\ 0 & \textit{sinon} \end{cases} \tag{4.4}$$

4.2.1 Éléments de probabilité

Toutes les variables aléatoires (v.a) et éléments stochastiques présentés dans ce chapitre sont définies sur le même espace probabilisé $(\Omega, \Im, P)$.
Les v.a $\{\delta_{X_n},\ n = 0, 1, ...\}$ peuvent être représentées comme suit

$$\delta_{X_n} = 1\left[U_{n+1} \leq \frac{X_n \nu}{\lambda + X_n \nu}\right],\ n = 0, 1, ... \tag{4.5}$$

où $\{U_{n+1},\ n = 0, 1, ...\}$ est une suite de variables aléatoires indépendantes uniformément distribuées sur $(0, 1)$.
Les filtrations $\{\Im_n,\ n = 0, 1, ...\}$ considérées dans ce chapitre sont définies comme étant engendrées par les variables aléatoires A_i et U_i est notée $\Im_n = \sigma(A_0, A_1, ..., A_n;\ U_1, ..., U_{n+1})$. Par conséquent les v.a X_n sont $\Im_n$-mesurables et les v.a $\{A_{n+1},\ n = 0, 1, ...\}$ et $\{U_{n+1},\ n = 0, 1, ...\}$ sont mutuellement indépendantes. Nous noterons $\Im = \bigcup_1^\infty \Im_n$.

Avec les notations ci-dessus et en utilisant les propriétés de l'espérance conditionnelle, nous obtenons

$$E(z^{X_{n+1}}/\Im_n) = z^{X_n - \delta_{X_n}} a(z) \qquad p.s. \tag{4.6}$$

4.2.2 Temps d'arrêt

Nous considérons σ un temps d'arrêt arbitraire pour $\Im_n$, et nous définissons la variable aléatoire $\nu(\sigma)$ comme étant le premier instant après le temps σ, où le système redevient vide. C'est à dire :

$$\nu(\sigma) = \begin{cases} inf\{n \geq 1 : X_{\sigma+n} = 0\} & si\ \sigma < \infty \\ \infty & sinon, \end{cases} \tag{4.7}$$

avec la convention $inf\{\emptyset\} = +\infty$.

4.3 Martingale

Nous pouvons maintenant définir une martingale $M_n(z)$ avec la filtration $(\Im_n)$ qui nous aidera à obtenir la majorité des résultats.

Théorème 4.1. *H. Oukid [80]*
Pour tout $0 < z \leq 1$, *on définit*

$$M_n(z) = \begin{cases} z^{X_0}, & si\ n = 0, \\ z^{X_n} z^{\frac{\sum_{k=0}^{n-1} \delta_{X_k}}{a(z)^n}}, & pour\ n = 1, 2, ... \end{cases} \tag{4.8}$$

Le processus $\{M_n(z), n \in N\}$ *adapté à la filtration* $\{\Im_n,\ n \geq 0\}$, *est une martingale positive intégrable.*

Démonstration. Pour démontrer que la suite $\{M_n(z)\}$ est une martingale, nous utiliserons (4.1) et (4.6). Nous obtenons :

$$\begin{aligned} E(M_{n+1}(z)/\Im_n) &= E\left(z^{X_{n+1}} \frac{z^{\sum_{k=0}^{n} \delta_{X_k}}}{a(z)^{n+1}} / \Im_n \right) \\ &= \frac{z^{\sum_{k=0}^{n} \delta_{X_k}}}{a(z)^{n+1}} E\big(z^{X_{n+1}}/\Im_n\big) = M_n(z)\ a.s. \end{aligned} \tag{4.9}$$

Il est facile de vérifier que $M_n(z)$ est intégrable, en effet

$$E\big(|M_{n+1}(z)|\big) = E\big(M_n(z)\big) < \infty. \tag{4.10}$$

□

4.4 Stabilité du système

Dans cette section, nous obtenons la condition de stabilité du système $M/G/1$ avec rappels par le biais de la martingale introduite à la section 4.3. La clé de ce résultat est l'utilisation du théorème d'arrêt de Doob "Optional Stopping Theorem "[75].

Théorème 4.2. *H. Oukid [80]*
Si $\rho \leq 1$, alors pour $0 < z \leq 1$, on a :

$$E\left(1_{[\sigma < \infty,\, \nu(\sigma) < \infty]} \frac{z^{\sum_{k=\sigma}^{\tau(\sigma)-1} \delta_{X_k}}}{a(z)^{\nu(\sigma)}} / \Im_\sigma\right) = 1_{[\sigma < \infty]} z^{X_\sigma} \qquad p.s. \tag{4.11}$$

Démonstration. Soit σ un temps d'arrêt relatif à la suite croissante de sous-tribus $\{\Im_n,\ n \in N\}$ et $\tau(\sigma) = \sigma + \nu(\sigma)$. la variable aléatoire $\tau(\sigma)$ est aussi un temps d'arrêt pour $\Im_n$.

Pour tout $n \geq 0$, $\tau(\sigma) \wedge n$ et $\sigma \wedge n$ sont encore des temps d'arrêt.
Il est clair que $\sigma \wedge n \leq \tau(\sigma) \wedge n$. Comme $\{M_n,\ n \in N\}$ est une martingale intégrable, alors d'après le théorème d'arrêt IV-2-6 [75], on a :

$$E\big[M_{\tau(\sigma)\wedge n}(z) / \Im_{\sigma \wedge n}\big] = M_{\sigma \wedge n}(z) \qquad p.s. \tag{4.12}$$

Qui peut s'écrire sous la forme suivante

$$E\left[z^{X_{\tau(\sigma)\wedge n}} \frac{z^{\sum_{k=0}^{\tau(\sigma)\wedge n-1} \delta_{X_k}}}{a(z)^{\tau(\sigma)\wedge n}} / \Im_{\sigma \wedge n}\right] = z^{X_{\sigma \wedge n}} \frac{z^{\sum_{k=0}^{\sigma \wedge n-1} \delta_{X_k}}}{a(z)^{\sigma \wedge n}} \qquad p.s. \tag{4.13}$$

Ce qui conduit à :

$$E\big[M_{\tau(\sigma)\wedge n}(z) - M_{\sigma \wedge n}(z) / \Im_{\sigma \wedge n}\big] = 0. \tag{4.14}$$

Par conséquent,

$$E\big[1_{[\sigma < n,\, \nu(\sigma) < n]} M_{\tau(\sigma)}(z) / \Im_\sigma\big] = 1_{[\sigma < n]} M_\sigma(z). \tag{4.15}$$

Ensuite, par convergence monotone, en faisant tendre n vers l'infini, nous obtenons :

$$\lim_{n \to \infty} E\left[1_{[\sigma < n,\, \nu(\sigma) < n]} z^{X_{\tau(\sigma)}} \frac{z^{\sum_{k=0}^{\tau(\sigma)-1} \delta_{X_k}}}{a(z)^{\tau(\sigma)}} / \Im_\sigma\right]$$

$$= E\left[1_{[\sigma<\infty,\,\nu(\sigma)<\infty]} z^{X_{\tau(\sigma)}} \frac{z^{\sum_{k=0}^{\tau(\sigma)-1}\delta_{X_k}}}{a(z)^{\tau(\sigma)}} / \Im_\sigma\right]$$

$$= 1_{[\sigma<\infty]} z^{X_\sigma} \frac{z^{\sum_{k=0}^{\sigma-1}\delta_{X_k}}}{a(z)^\sigma} \qquad p.s. \tag{4.16}$$

Or sur l'événement $[\sigma<\infty, \nu(\sigma)<\infty]$, nous avons $[\tau(\sigma)<\infty]$ et donc $X_{\tau(\sigma)}=0$.
Finalement,

$$E\left[1_{[\sigma<\infty,\,\nu(\sigma)<\infty]} \frac{z^{\sum_{k=0}^{\tau(\sigma)-1}\delta_{X_k}}}{a(z)^{\tau(\sigma)}} / \Im_\sigma\right] = 1_{[\sigma<\infty]} z^{X_\sigma} \frac{z^{\sum_{k=0}^{\sigma-1}\delta_{X_k}}}{a(z)^\sigma} \qquad p.s. \tag{4.17}$$

En tenant compte de la $\Im_\sigma$-mesurabilité des temps d'arrêt σ et $\sum_{k=0}^{\sigma-1}\delta_{X_k}$, nous obtenons (4.11), qui prouve le théorème (4.2). □

Corollaire 4.3. *Sous la supposition $\rho \le 1$ et pour $0<z<1$ on a*

$$P\left[\sigma<\infty, \nu(\sigma)<\infty / \Im_\sigma\right] = 1_{[\sigma<\infty]} \qquad p.s. \tag{4.18}$$

En particulier, si $\sigma<\infty$ p.s., alors $\nu(\sigma)<\infty$ p.s.

Démonstration. En faisant tendre $z \to 1$ dans (4.11) le corollaire (4.3) est une conséquence immédiate du théorème de Convergence dominée.

□

4.5 Condition d'instabilité

Nous étudions l'instabilité de ce système au sens de divergence de la chaîne de Markov induite.

Théorème 4.4. *H. Oukid [76]*
Si $\rho>1$, alors

$$\lim_{n\to\infty} X_n = \infty \; p.s. \tag{4.19}$$

Démonstration. Pour tout $0 < z < 1$ et tout $n \in N$, la relation (4.6) implique que

$$E\big[z^{X_{n+1}}/\Im_n\big] = z^{X_n - \delta_{X_n}} a(z) \leq z^{X_n}\big(\frac{a(z)}{z}\big) \qquad p.s. \qquad (4.20)$$

Supposons que $\rho > 1$ et comme la fonction $a(.)$ est convexe, alors d'après le lemme de Takács [92], il existe z_0 dans l'intervalle $(0,1)$ tel que $a(z_0) < z_0$. Soit c_0 la constante définie par $c_0 = \frac{a(z_0)}{z_0} < 1$.

Par conséquent,

$$E(z_0^{X_{n+1}}/\Im_n) \leq c_0 z_0^{X_n} \leq z_0^{X_n} \qquad p.s. \qquad (4.21)$$

Ce qui prouve que la suite $\{z_0^{X_n},\ n \in N\}$ est $\Im_n-$surmartingale positive majorée qui convege p.s.

De plus, par récurrence sur n, de l'équation (4.21), on en déduit que :

$$E(z_0^{X_n}) \leq c_0^n E(z_0^{X_0}) \leq c_0^n. \qquad (4.22)$$

En utilisant le théorème de la convergence dominée, nous obtenons :

$$\lim_n E(z_0^{X_n}) = E\big(\lim_n z_0^{X_n}\big).$$

En passant à la limite quand n tend vers l'infini, nous avons :

$$\lim_n E(z_0^{X_n}) = E(\lim_n z_0^{X_n}) = 0. \qquad (4.23)$$

Par conséquent, $lim_n z_0^{X_n} = 0$ p.s. pour $0 < z_0 < 1$, et le résultat (4.19) suit immédiatement.

□

Dans la partie suivante, nous nous intéressons au cas où σ est un temps d'arrêt pour $\Im_n$ tel que $X_\sigma = 0$ sur l'événement $[\sigma < \infty]$. Dans ce cas la variable aléatoire $\nu(\sigma)$ représente le nombre de clients servis au cours d'une période d'activité.

4.5.1 Période d'activité

Théorème 4.5. *H. Oukid [76]*

Nous considérons σ un temps d'arrêt pour $\Im_n$ tel que $X_\sigma = 0$ sur $[\sigma < \infty]$. Sous la supposition $\rho \leq 1$, le nombre moyen de clients servis au cours d'une période d'activité est donné par

$$E[\nu(\sigma)] = \begin{cases} \frac{1}{1-\rho}\psi(1) & si\ \rho < 1 \\ \infty & si\ \rho = 1 \end{cases} \tag{4.24}$$

où

$$\Psi(1) = exp\left(\frac{\lambda}{\nu}\int_0^1 \frac{1-a(y)}{a(y)-y}dy\right) \tag{4.25}$$

Démonstration. Si $\rho < 1$, pour tout z dans l'intervalle $(0,1)$, nous avons $z < a(z) \leq 1$, choisissons un temps d'arrêt σ fini pour $\Im_n$, il est clair que

$$E\left[1_{[\nu(\sigma)<\infty]}\frac{z^{\sum_{k=\sigma}^{\tau(\sigma)-1}\delta_{X_k}}}{a(z)^{\nu(\sigma)}}\right] - E\left[1_{[\nu(\sigma)<\infty]}z^{\sum_{k=\sigma}^{\tau(\sigma)-1}\delta_{X_k}}\right]$$

$$= E\left[1_{[\nu(\sigma)<\infty]}z^{\sum_{k=\sigma}^{\tau(\sigma)-1}\delta_{X_k}}\left(\frac{1-a(z)^{\nu(\sigma)}}{a(z)^{\nu(\sigma)}}\right)\right]$$

$$= \big(1-a(z)\big)E\left[1_{[\nu(\sigma)<\infty]}\frac{z^{\sum_{k=\sigma}^{\tau(\sigma)-1}\delta_{X_k}}}{a(z)^{\nu(\sigma)}}\left(\frac{1-a(z)^{\nu(\sigma)}}{1-a(z)}\right)\right]$$

$$= \big(1-a(z)\big)E\left[1_{[\nu(\sigma)<\infty]}\frac{z^{\sum_{k=\sigma}^{\tau(\sigma)-1}\delta_{X_k}}}{a(z)^{\nu(\sigma)}}\sum_{k=0}^{\nu(\sigma)-1}a(z)^k\right]. \tag{4.26}$$

En plus,

$$\lim_{z\to 1}\frac{z^{\sum_{k=\sigma}^{\tau(\sigma)-1}\delta_{X_k}}}{a(z)^{\nu(\sigma)}} = 1.$$

En utilisant le théorème de la Convergence Monotone, lorsque $z \to 1$ dans (4.26), nous obtenons

$$\lim_{z\to 1} E\left[1_{[\nu(\sigma)<\infty]} \frac{z^{\sum_{k=\sigma}^{\tau(\sigma)-1}\delta_{X_k}}}{a(z)^{\nu(\sigma)}} \sum_{k=0}^{\nu(\sigma)-1} a(z)^k\right] = E\big[1_{[\nu(\sigma)<\infty]}\nu(\sigma)\big],$$

il découle que

$$\lim_{z\to 1}\big(1-a(z)\big)^{-1}\left(E\left[1_{[\nu(\sigma)<\infty]}\frac{z^{\sum_{k=\sigma}^{\tau(\sigma)-1}\delta_{X_k}}}{a(z)^{\nu(\sigma)}}\right] - E\left[1_{[\nu(\sigma)<\infty]}z^{\sum_{k=\sigma}^{\tau(\sigma)-1}\delta_{X_k}}\right]\right)$$
$$= E\big[1_{[\nu(\sigma)<\infty]}\nu(\sigma)\big]. \tag{4.27}$$

On utilise le théorème (4.2) pour obtenir

$$E\left[1_{[\nu(\sigma)<\infty]}\frac{z^{\sum_{k=\sigma}^{\tau(\sigma)-1}\delta_{X_k}}}{a(z)^{\nu(\sigma)}}\right] = E\big[z^{X_\sigma}\big], \tag{4.28}$$

maintenant, (4.27) peut être réécrite sous la forme

$$E\big[1_{[\nu(\sigma)<\infty]}\nu(\sigma)\big]$$
$$= \lim_{z\to 1}\big(1-a(z)\big)^{-1}\left(E\big[z^{X_\sigma}\big] - E\left[1_{[\nu(\sigma)<\infty]}z^{\sum_{k=\sigma}^{\tau(\sigma)-1}\delta_{X_k}}\right]\right). \tag{4.29}$$

Appliquons l'égalité suivante

$$E\big[1_{[\nu(\sigma)<\infty]}z^{\sum_{k=\sigma}^{\tau(\sigma)-1}\delta_{X_k}}\big] = E\big[1_{[\nu(\sigma)<\infty]}z^{\sum_{i=1}^{\nu(\sigma)}\delta_{X_i}}\big] = \varphi_{\nu(\sigma)}[\varphi_{\delta_{X_1}}(z)], \tag{4.30}$$

nous obtenons en vertu de (4.29) et (4.30)

$$E\big[1_{[\nu(\sigma)<\infty]}\nu(\sigma)\big] = \lim_{z\to 1}(a(z)-1)^{-1}\big[\varphi_{\nu(\sigma)}(\varphi_{\delta_{X_1}}(z)) - E[z^{X_\sigma}]\big], \tag{4.31}$$

où $\varphi_{\nu(\sigma)}$ est la fonction génératrice du nombre de clients servis au cours d'une période d'activité (voir [43]), donnée par

$$\varphi_{\nu(\sigma)}(z) = \pi(0,z)$$
$$= 1 - \frac{\nu}{\lambda}\int_0^{\pi_\infty(0,z)} exp\{-\frac{\lambda}{\nu}\int_0^x \frac{1-za(y)}{za(y)-y}dy\}\frac{dx}{za(x)-x}. \tag{4.32}$$

Sur l'événement $[\sigma < \infty]$, $X_\sigma = 0$ et en vertu du corollaire (4.3), on a $[\nu(\sigma) < \infty]$.
Nous avons

$$E[\nu(\sigma)] = \lim_{z\to 1}\frac{\varphi_{\nu(\sigma)}(\varphi_{\delta_{X_1}}(z)) - 1}{a(z)-1}. \tag{4.33}$$

Appliquons la règle d'hôpital, quand $z \to 1$, de (4.33), nous obtenons

$$E[\nu(\sigma)] = \lim_{z\to 1}\frac{[\varphi_{\nu(\sigma)}(\varphi_{\delta_{X_1}}(z))]'}{a'(z)} = \lim_{z\to 1}\frac{\varphi'_{\delta_{X_1}}(z)\varphi'_{\nu(\sigma)}\big(\varphi_{\delta_{X_1}}(z)\big)}{a'(z)}. \tag{4.34}$$

Afin de prouver cela, calculons la première dérivée de $\varphi_{\delta_{X_1}}$ au point z. Utilisons l'équation

$$\varphi_{\delta_{X_1}}(z) = \sum_{k=0}^{1} z^k P(\delta_{X_1} = k), \tag{4.35}$$

nous obtenons

$$\varphi'_{\delta_{X_1}}(z) = \sum_{k=0}^{1} k z^{k-1} P(\delta_{X_1} = k), \tag{4.36}$$

$$\lim_{z\to 1}\varphi'_{\delta_{X_1}}(z) = E(\delta_{X_1}) = \rho, \tag{4.37}$$

et

$$\varphi'_{\nu(\sigma)}\big(\varphi_{\delta_{X_1}}(z)\big)$$
$$= \pi'_\infty\big(0,\varphi_{\delta_{X_1}}(z)\big)exp\left\{\frac{\lambda}{\nu}\int_0^{\pi_\infty\left(0,\varphi_{\delta_{X_1}}(z)\right)}\frac{1-\pi_\infty\big(0,\varphi_{\delta_{X_1}}(z)\big)a(y)}{\pi_\infty\big(0,\varphi_{\delta_{X_1}}(z)\big)a(y)-y}dy\right\}. \tag{4.38}$$

Notons que

$$\pi'_\infty\big(0,\varphi_{\delta_{X_1}}(z)\big) = E'(\varphi_{\delta_{X_1}}(z)^{I_\infty}) = \left[\sum_{k=0}^{\infty}\varphi^k_{\delta_{X_1}}(z)P(I_\infty = k)\right]' \tag{4.39}$$

$$= \sum_{k=0}^{\infty} k\varphi_{\varphi_{\delta_{X_1}}}^{k-1}(z)P(I_\infty = k), \tag{4.40}$$

où I_∞ est le nombre de clients servis au cours d'une période d'activité dans un système $M/G/1$ classique.

Substituons (4.36) et (4.38) dans (4.34), nous obtenons pour $0 < z < 1$

$$\lim_{z\to 1} \frac{\varphi'_{\delta_{X_1}}(z)}{a'(z)} \pi'_\infty\big(0, \varphi_{\delta_{X_1}}(z)\big) exp\Big\{ \frac{\lambda}{\nu} \int_0^{\pi_\infty\left(0,\varphi_{\delta_{X_1}}(z)\right)} \frac{1 - \pi_\infty\big(0, \varphi_{\delta_{X_1}}(z)\big)a(y)}{\pi_\infty\big(0, \varphi_{\delta_{X_1}}(z)\big)a(y) - y} dy \Big\}$$

$$= E[\nu(\sigma)]. \tag{4.41}$$

En tenant compte que $\lim_{z\to 1} a'(z) = \rho$, l'équation (4.41) donne :

$$E[\nu(\sigma)] = E(I_\infty) exp\{\frac{\lambda}{\nu} \int_0^1 \frac{1 - a(y)}{a(y) - y} dy\}. \tag{4.42}$$

□

Enfin, le théorème suivant donne une image globale de l'évolution du système M/G/1 avec rappels sous l'hypothèse que $\rho < 1$.

Théorème 4.6. *Supposons que $\rho = -\lambda\beta'(0) \leq 1$, et la séquence de temps de service forme une séquence de renouvellement, alors il existe une suite $\{\tau_n,\ n = 1, 2, ...\}$ finie p.s. de temps d'arrêt pour $\Im_n$, définis par $\tau_{n+1} = \tau_n + \nu(\tau_n)$ pour tout $n > 0$ tel que $X_{\tau_n} = 0$ sur $\{\tau_n < \infty\}$ et $\tau_n + 1 \leq \tau_{n+1},\ \forall n \in N^*$. Dans ce cas les v.a $\{\nu_n\}_2^\infty$ forment une suite i.i.d indépendante de τ_1 et*

$$E(\nu_{n+2}) = \begin{cases} \frac{1}{1-\rho}\Psi(1) & si\ \rho < 1 \\ \infty & si\ \rho = 1 \end{cases} \tag{4.43}$$

où

$$\Psi(1) = exp\left(\frac{\lambda}{\nu} \int_0^1 \frac{1 - a(y)}{a(y) - y} dy\right) \tag{4.44}$$

Démonstration. On définit $\tau_{n+1} = \tau_n + \nu(\tau_n),\ n = 0,\ 1,\ ...$, avec $\nu(\tau_n) = \nu_{n+1}$ et $\tau_0 = 0$.

La preuve de ce théorème 4.6 suit la méthodologie de Baccelli et Makowski [24], en utilisant le théorème 4.2, le corollaire 4.3 et le fait que pour la file d'attente M/G/1 avec rappels, la période d'activité satisfait l'équation

$$E\big[\nu(\sigma)\big] = E(I_\infty) exp\left(\frac{\lambda}{\nu}\int_0^1 \frac{1-a(y)}{a(y)-y}dy\right) \tag{4.45}$$

□

4.6 Conclusion

Dans ce chapitre, nous avons développé une nouvelle méthode basée sur la théorie des Martingales pour analyser le système $M/G/1$ avec rappels. Nous avons construit une martingale à temps discret arrêtée au premier instant où le système redevient vide. Sous la condition de stabilité et en utilisant cette martingale, le calcul des caractéristiques du système concerné est obtenue. Nous avons aussi démontré l'instabilité de ce système au sens de divergence de la chaîne de Markov incluse.

Chapitre 5

ANALYSE D'UN SYSTÈME MULTISERVEUR NON-MARKOVIEN AVEC PERTES

5.1 Introduction

L'utilisation de plusieurs serveurs (tels que des ordinateurs, des opérateurs et machines) est omniprésente. Dans notre vie quotidienne, nous voyons souvent plus d'un caissier dans une banque, plus d'un vérificateur dans un supermarché et plus d'un caissier dans un restaurant de Fast-Food, dans les situations où un guichet unique, vérificateur ou caissier est insuffisante pour traiter le volume de clients. De même, l'utilisation de plusieurs ordinateurs est fréquente dans les systèmes informatiques tels que des batteries de serveurs web et centres de calcul intensif, parce que l'utilisation de plusieurs ordinateurs est une solution rentable et évolutive pour atteindre la haute performance et la fiabilité. Lorsque nous concevons un système multiserveur (ou même un système de serveur unique), une étape importante est d'analyser et de comprendre leur performance. L'étude et l'analyse de performance des systèmes remonte à l'ouvrage de A. K. Erlang en 1917 [50], où il a fourni des formules qui peuvent servir à évaluer le rendement à un central téléphonique. Les formules d'Erlang étaient utilisées par les compagnies de téléphone pour fournir les ressources nécessaires et suffisantes, qui ont

conduit à la croissance rapide et réussi des réseaux téléphoniques. Le travail d'Erlang a été suivi par de nombreux chercheurs et est devenu une théorie connue aujourd'hui Sous « la théorie des files d'attente ». Aujourd'hui, les systèmes multiserveurs peuvent employer des configurations plus complexes et/ou allocation de ressources plus complexe vers des opérations plus efficaces et plus rentables. Étant donné que les approches classiques en théorie des files d'attente conduisent à des expressions complexes ou ne s'appliquent pas pour des systèmes complexes (multiserveurs), en raison de la complexité des résultats connus. En effet, dans la majorité des cas, on se retrouve confronté à des systèmes d'équations dont la résolution est complexe ou possédant des solutions qui ne sont pas facilement interprétables afin que le praticien puisse en bénéficier. Un autre outil qui peut être utilisé pour étudier les systèmes des files d'attente est la méthode des martingales. Habituellement, elle est probabiliste et évite les calculs fastidieux. L'approche Martingale est une technique analytique élégante qui donne souvent des résultats d'intérêt indépendant. La technique est également d'intérêt indépendant du point de vue de modélisation et elle peut être utilisée dans d'autres domaines d'application que la théorie des files d'attente.

Dans ce chapitre, nous montrons comment cette approche peut être adaptée à un système multiserveur non-markovien avec pertes. Une telle file d'attente peut être utilisée pour modéliser un centre de commutation permettant un maximum de k appels simultanés.

5.2 Description du modèle

On considère un système de files d'attente à plusieurs serveurs ayant la structure suivante :

- Soit $A(t)$ la variable aléatoire indiquant le nombre d'arrivées dans un intervalle de temps $[0, t]$ et $D(t)$ le nombre de clients qui quittent le système au temps t.
- $A(t)$ et $D(t)$ sont des processus ponctuels à accroissements stationnaires et ergodiques.
- On a n serveurs.
- A l'arrivée d'un client, si l'un des serveurs est libre, le client sera pris en charge immédiatement, dans le cas contraire, le client est perdu.
- Soit $Q(t)$ le nombre de clients dans le système à l'instant t, ce qui

coïncide avec le nombre de serveurs occupés à l'instant t et on suppose que le système est initialement vide, c'est-à-dire, $Q(0) = 0$. De plus, on suppose que les temps de service sont des variables aléatoires indépendantes, indépendantes de processus des arrivées
- Soit $L(t)$ le nombre de clients perdus jusqu'à l'instant t.
- Tous les processus ponctuels considérés dans le présent document sont continus à droite ayant de limite à gauche.

5.3 Décomposition en semimartingale du processus du nombre de clients

Nous pouvons obtenir une représentation sous la forme d'une martingale pour les processus stochastiques $Q(t) + L(t)$ à l'aide de la décomposition de Doob-Meyer des semimartingales et la méthode de Abramov [1]. Nous avons la représentation fondamentale suivante :

$$Q(t) + L(t) = A(t) - D(t),\ t \geq 0, \tag{5.1}$$

où le processus de départ $D(t)$ est défini à l'aide du processus ponctuel $D_i(t),\ i = 1, ..., n$ comme suit

$$D(t) = \int_0^t \sum_{i=1}^{n} I\{Q(s^-) \geq i\} dD_i(s),\ t \geq 0, \tag{5.2}$$

avec $I\{A\}$ est la fonction indicatrice de l'événement A.
En tenant compte du fait que $A(t)$ et $D_i(t),\ i = 1, 2, ..., n$ sont des semimartingales adaptées à la filtration $\Im_n$ donnée sur l'espace de probabilité $(\Omega, \Im, P)$, alors les processus $A(t)$ et $D(t)$ peuvent être réécrits en utilisant la décomposition de Doob-Meyer [e.g. Liptser et Shiryayev, [72]] comme suit :

$$A(t) = \hat{A}(t) + M_A(t), \tag{5.3}$$

et

$$D(t) = \hat{D}(t) + M_D(t), \tag{5.4}$$

où $M_A(t)$ et $M_D(t)$ sont des martingales locales de carré intégrables, $\hat{A}(t)$ et $\hat{D}(t)$ les compensateurs (processus croissant prévisible) des processus $A(t)$ et $D(t)$ respectivement, admettant la représentation suivante :

$$\hat{A}(t) = \int_0^t X(s) ds,\ t \geq 0, \tag{5.5}$$

et

$$\hat{D}(t) = \int_0^t Q(s)Y(s)ds,\ t \geq 0, \tag{5.6}$$

où $X = \{X(t) : t \geq 0\}$ et $Y = \{Y(t) : t \geq 0\}$ sont les intensités stochastiques des processus ponctuels A et D respectivement adaptés à la filtration $\Im$.
En vertu de l'équation (5.3), (5.4) et l'équation (5.1), le processus $Q(t)+L(t)$ peut être réécrit, en utilisant la décomposition de Doob-Meyer des semimartingales comme suit :

$$Q(t) + L(t) = \hat{A}(t) - \hat{D}(t) + M_A(t) - M_B(t). \tag{5.7}$$

5.4 Renormalisation du processus du nombre de clients dans le système

Dans cette section, nous étudions le processus de nombre de clients dans le système sous la propriété de renormalisation.
Rappelons d'abord que pour un processus $X(t); t > 0$ quelconque sa renormalisation est notée par la lettre minuscule, $x(t) = \frac{1}{t}X(t)$. Ainsi $q(t) = \frac{1}{t}Q(t)$, $l(t) = \frac{1}{t}L(t)$, $m_A(t) = \frac{1}{t}M_A(t)$ et $m_B(t) = \frac{1}{t}M_B(t)$.
L'équation (5.7) s'écrira alors sous la forme

$$q(t) + l(t) = \frac{1}{t}\hat{A}(t) - \frac{1}{t}\hat{D}(t) + m_A(t) - m_B(t), \tag{5.8}$$

ou encore

$$q(t) + l(t) = \frac{1}{t}\int_0^t X(s)ds - \frac{1}{t}\int_0^t Q(s)Y(s)ds + m_A(t) - m_B(t). \tag{5.9}$$

Dans le présent document, le nombre de clients dans le système est toujours limité, et la condition de convergence suivante est utilisée $P\lim_{t\to\infty} a(t) = \lambda$, où $P\lim$ est la limite en probabilité. Si nous supposons de plus que t tend vers ∞, le terme $q(t)$ converge vers 0.
Prouvons maintenant que $P\lim_{t\to\infty} m_A(t) = 0$. En appliquant l'inégalité de Lenglart-Rebolledo [[72], pp66], pour tout δ positif, nous obtenons

$$P\{|M_A(t)| > \delta\} \leq P\Big\{\sup_{0\leq s\leq t} |m_A(s)| > \delta\Big\}$$

$$= P\Big\{ \sup_{0\le s\le t} |M_A(s)| > \delta t\Big\}$$

$$= P\Big\{ \sup_{0\le s\le t} |A(s) - \hat{A}(s)| > \delta t\Big\}$$

$$\le \frac{\epsilon}{\delta^2} + P\{\frac{A(t)}{t} > \epsilon t\}. \tag{5.10}$$

Puisque ϵ est arbitraire, les deux termes de la partie droite de l'équation (5.10) converge vers 0 quand $t \to \infty$, et de plus $P \lim_{t\to\infty} |m_A(t)| = 0$.
Par conséquent le terme

$$\frac{1}{t}\big[\hat{A}_i(t) - A_i(t)\big] \tag{5.11}$$

converge en probabilité vers 0 lorsque $t \to \infty$. Cela signifie que les deux termes $\frac{1}{t}\hat{A}(t)$ et $\frac{1}{t}A(t)$ ont la même limite en probabilité lorsque $t \to \infty$.
Notons, que par analogie à (5.10) nous avons le même résultat que pour le processus $D_i(t)$:

$$\frac{1}{t}[\hat{D}_i(t) - D_i(t)] \tag{5.12}$$

converge en probabilité vers 0 quand $t \to \infty$, donc $P \lim_{t\to\infty} |m_D(t)| = 0$.
En suite, en appliquant le théorème de la convergence dominée, nous obtenons

$$P \lim_{t\to\infty} l(t) = \lambda - P \lim_{t\to\infty} \frac{1}{t}\int_0^t Q(s)Y(s)ds$$

$$= \lambda - \lim_{t\to\infty} \frac{1}{t} E \int_0^t Q(s)Y(s)ds. \tag{5.13}$$

5.5 Analyse de la distribution limite du nombre de clients dans le système

Dans cette section nous présentons les équations relatives aux limites :

$$\lim_{t\to\infty} \frac{1}{t} E \int_0^t I\{Q(s) = i\} dA(s),\ i = 1, 2, ..., n, \tag{5.14}$$

basé sur la décomposition de Doob-Meyer de la fonction indicatrice du nombre de clients dans le système.

Dans ce qui suit, on introduit le processus

$$I_i(t) = I\{Q(t) = i\};\ i = 0, 1, ..., n;\ \forall t \geq 0. \tag{5.15}$$

En prenant en considération que $I_{-1}(t) = 0$. Notons le saut du processus $I_i(t)$ par $\triangle I_i(t)$ ainsi que les sauts des processus $A(t)$, $D(t)$ et $Q(t)$ par $\triangle A(t)$, $\triangle D(t)$ et $\triangle Q(t)$ respectivement. Ainsi, nous avons le théorème suivant.

Théorème 5.1. *H. Oukid [77]*

Soit $A(t)$ et $D_i(t)$ deux processus ponctuels à accroissements strictement stationnaires et ergodiques, nous avons le système d'équations suivant.

Pour i=0,...,n-1,

$$\lim_{t\to\infty} \frac{1}{t} E \int_0^t [I\{Q(s) = i-1\} - I\{Q(s) = i\}] dA(s)$$

$$= i \lim_{t\to\infty} \frac{1}{t} E \int_0^t I\{Q(s) = i\} dD_i(s)$$

$$-(i+1) \lim_{t\to\infty} \frac{1}{t} \int_0^t I\{Q(s) = i+1\} dD_{i+1}(s). \tag{5.16}$$

Pour i=n,

$$\lim_{t\to\infty} \frac{1}{t} E \int_0^t [I\{Q(s^-) = n-1\} - I\{Q(s^-) = n\}] dA(s)$$

$$= n \lim_{t\to\infty} \frac{1}{t} \int_0^t I\{Q(s) = n\} dD_n(s). \tag{5.17}$$

Démonstration. Nous avons les équations suivantes

Pour $i = 0, 1, ..., n-1$,

$$I\{Q(t^-) + \triangle Q(t) = i\} = I_{i-1}(t^-)\triangle A(t)$$

$$+I_{i+1}(t^-)\triangle D_{i+1}(t) + I_i(t^-)[1 - \triangle A(t)][1 - \triangle D_i(t)]. \tag{5.18}$$

Pour $i = n$,

$$I\{Q(t^-) + \triangle Q(t) = n\} = I_{n-1}(t^-)\triangle A(t)$$

$$+I_n(t^-)[1-\triangle A(t)][1-\triangle D_n(t)]. \tag{5.19}$$

Aussi,

$$\triangle I_i(t) = I\{Q(t^-)+\triangle Q(t)=i\} - I\{Q(t^-)=i\},\ i=0,1,...,n. \tag{5.20}$$

Alors,

$$\sum_{s\le t} \triangle I_i(s) = I_i(t) - I_i(0), \tag{5.21}$$

les sauts de $A(t)$ et $D(t)$ sont disjoints. De (5.18)-(5.21), le processus $I_i(t)$ peut être représenté comme suit :
Pour $i=0,1,...,n-1$,

$$I_i(t) = I_i(0) + \int_0^t I_{i-1}(s^-)dA(s)$$

$$+\int_0^t I_{i+1}(s^-)dD_{i+1}(s) - \int_0^t I_i(s^-)dA(s)$$

$$-\int_0^t I_i(s^-)dD_i(s). \tag{5.22}$$

Pour $i=n$,

$$I_n(t) = I_n(0) + \int_0^t I_{n-1}(s^-)dA(s)$$

$$-\int_0^t I_n(s^-)dA(s) - \int_0^t I_n(s^-)dD_n(s). \tag{5.23}$$

En appliquant la décomposition des semimartingales de Doob-Meyer, de (5.22) et (5.23), nous obtenons :

Pour $i=0,1,...,n-1$,

$$I_i(t) = I_i(0) + \int_0^t [I_{i-1}(s^-) - I_i(s^-)]d\hat{A}(s)$$

$$+(i+1)\int_0^t I_{i+1}(s)d\hat{D_{i+1}}(s) - i\int_0^t I_i(s)d\hat{D}_i(s) + M_i(t), \tag{5.24}$$

avec la martingale locale de carré intégrable

$$M_i(t) = \int_0^t [I_{i-1}(s^-) - I_i(s^-)]dM_A(s) - \int_0^t I_i(s^-)dM_{D_i}(s)$$

$$+\int_0^t I_{i+1}(s^-)dM_{D_{i+1}}(s),$$

ou encore

$$M_i(t)=\int_0^t [I_{i-1}(s^-)-I_i(s^-)]d\big[A(s)-\hat{A}(s)\big]-\int_0^t I_i(s^-)d\big[D_i(s)-\hat{D}_i(s)\big]$$

$$+\int_0^t I_{i+1}(s^-)d\big[D_{i+1}(s)-\hat{D_{i+1}}(s)\big],$$

et
pour $i=n$,

$$I_n(t)=I_n(0)+\int_0^t [I_{n-1}(s^-)-I_n(s^-)]d\hat{A}(s)$$

$$-n\int_0^t I_n(s)d\hat{D}_n(s)+M_n(t), \tag{5.25}$$

avec,

$$M_n(t)=\int_0^t [I_{n-1}(s^-)-I_n(s^-)]dM_A(s)-\int_0^t I_n(s^-)dM_{D_n}(s).$$

Ensuite en passant à la limite quand t tend vers l'infini , de (5.24) et (5.25), il vient
Pour $i=0,1,...,n-1$,

$$P\lim_{t\to\infty}\frac{1}{t}(I_i(t)-I_i(0))=P\lim_{t\to\infty}\frac{1}{t}\int_0^t [I_{i-1}(s^-)-I_i(s^-)]d\hat{A}(s)$$

$$+P\lim_{t\to\infty}\frac{(i+1)}{t}\int_0^t I_{i+1}(s)d\hat{D_{i+1}}(s)-P\lim_{t\to\infty}\frac{i}{t}\int_0^t I_i(s)d\hat{D}_i(s)$$

$$+P\lim_{t\to\infty}\frac{1}{t}M_i(t). \tag{5.26}$$

Pour $i=n$,

$$P\lim_{t\to\infty}\frac{1}{t}(I_n(t)-I_n(0))=P\lim_{t\to\infty}\frac{1}{t}\int_0^t [I_{n-1}(s^-)-I_n(s^-)]d\hat{A}(s)$$

$$-P\lim_{t\to\infty}\frac{n}{t}\int_0^t I_n(s)d\hat{D}_n(s)+P\lim_{t\to\infty}\frac{1}{t}M_n(t). \tag{5.27}$$

En utilisant le théorème de Lebesgue sur la convergence dominée, nous obtenons

$$P\lim_{t\to\infty}\frac{1}{t}\int_0^t[I_{i-1}(s^-)-I_i(s^-)]d\hat{A}(s)$$

$$=\lim_{t\to\infty}\frac{1}{t}E\int_0^t[I_{i-1}(s^-)-I_i(s^-)]d\hat{A}(s). \tag{5.28}$$

Laissez nous réécrire la partie droite du (5.28) comme suit :

$$\lim_{t\to\infty}\frac{1}{t}E\int_0^t[I_{i-1}(s^-)-I_i(s^-)]d\hat{A}(s)$$

$$=\lim_{t\to\infty}\frac{1}{t}E\int_0^t[I_{i-1}(s^-)-I_i(s^-)]dA(s)$$

$$-\lim_{t\to\infty}\frac{1}{t}E\int_0^t[I_{i-1}(s^-)-I_i(s^-)]d[A(s)-\hat{A}(s)]. \tag{5.29}$$

En tenant compte du fait que $\frac{1}{t}A$ et $\frac{1}{t}\hat{A}$ ont la même limite en probabilité, nous concluons que le terme

$$P\lim_{t\to\infty}\frac{1}{t}\int_0^t[I_{i-1}(s^-)-I_i(s^-)]d[A(s)-\hat{A}(s)]$$

$$=\lim_{t\to\infty}\frac{1}{t}E\int_0^t[I\{Q(s)=i-1\}-I\{Q(s)=i\}]d[A(s)-\hat{A}(s)], \tag{5.30}$$

tend vers 0 quand $t\to\infty$.
De la combinaison (5.28), (5.29) et (5.30), nous obtenons

$$P\lim_{t\to\infty}\frac{1}{t}\int_0^t[I_{i-1}(s^-)-I_i(s^-)]d\hat{A}(s)$$

$$=P\lim_{t\to\infty}\frac{1}{t}\int_0^t[I_{i-1}(s^-)-I_i(s^-)]dA(s)$$

$$=\lim_{t\to\infty}\frac{1}{t}E\int_0^t[I\{Q(s)=i-1\}-I\{Q(s)=i\}]dA(s). \tag{5.31}$$

Notons que, par analogie, nous avons

$$P\lim_{t\to\infty}\frac{1}{t}\int_0^t I_i(s)dD_i(s) = Plim_{t\to\infty}\frac{1}{t}\int_0^t I_i(s)d\hat{D}_i(s)$$

$$+P\lim_{t\to\infty}\frac{1}{t}\int_0^t I_i(s)d[D_i(s)-\hat{D}_i(s)]. \tag{5.32}$$

Gardant à l'esprit que $\frac{1}{t}D_i(s)$ et $\frac{1}{t}\hat{D}_i(s)$ ont la même limite en probabilité, nous obtenons

$$P\lim_{t\to\infty}\frac{1}{t}\int_0^t I_i(s)dD_i(s) = P\lim_{t\to\infty}\frac{1}{t}\int_0^t I_i(s)d\hat{D}_i(s)$$

$$=\lim_{t\to\infty}\frac{1}{t}E\int_0^t I\{Q(s)=i\}dD_i(s). \tag{5.33}$$

Le côté gauche de (5.26) et de (5.27) est égal à zéro.
Finalement, en substituant (5.31) et (5.33) dans (5.26) et (5.27), Théorème (5.1) est prouvé. □

5.6 Cas particulier

Dans ce qui suit, nous allons considérer deux cas particuliers de processus ponctuels. Le premier est le cas où $A(t)$ est un processus de Poisson de paramètre λ, et le temps de service est exponentiel de paramètre μ. Le second est le cas où $A(t)$ et $D(t)$ sont des processus de Poisson non homogènes.

Notons

$$P_i(t) = P\{Q(t)=i\},\ i=0,1,...,n, \tag{5.34}$$

$$P_i = \lim_{t\to\infty}P\{Q(t)=i\},\ i=0,1,...,n. \tag{5.35}$$

Corollaire 5.2. *H. Oukid [78]*

Supposons que les processus $A(t)$ et $D_i(t)$ sont des processus de Poisson de paramètres λ et μ respectivement, alors nous avons le système d'équations linéaires et homogènes suivant.

Pour $i = 0$,

$$\lambda P_0 = \mu P_1. \tag{5.36}$$

Pour $i = 1, 2, ..., n-1$,

$$(\lambda + i\mu)P_i = \lambda P_{i-1} + (i+1)\mu P_{i+1}. \tag{5.37}$$

Pour $i = n$,

$$(\lambda + n\mu)P_n = \lambda P_{n-1}. \tag{5.38}$$

Démonstration. Nous considérons le régime stationnaire du processus $Q(t)$:

$$P_i = \lim_{t\to\infty} P\{Q(t) = i\}$$

$$= \lim_{t\to\infty} \frac{1}{t}\int_0^t P[Q(s) = i]ds,\ i = 0, 1, ..., n. \tag{5.39}$$

Dans le cas où $A(t)$ est un processus de Poisson de paramètre λ, le compensateur $\hat{A}(t) = \lambda t$. La décomposition de semimartingale du processus de Poisson $A(t)$, nous donne, $A(t) = \lambda t + M_A(t)$.
Le théorème de convergence dominée de Lebesgue donne facilement.

Pour $i = 0, 1, ..., n$,

$$\lim_{t\to\infty} \frac{1}{t}E\int_0^t I[Q(s^-) = i]dA(s) = \lim_{t\to\infty} \frac{\lambda}{t}\int_0^t P[Q(s^-) = i]ds$$

$$+ \lim_{t\to\infty} \frac{1}{t}E\int_0^t I[Q(s^-) = i]dM_A(s). \tag{5.40}$$

Le premier terme de la partie droite de (5.40) est égal à λP_i, et le second terme s'annule. Et par conséquent,

$$\lim_{t\to\infty} \frac{1}{t}E\int_0^t I[Q(s^-) = i]dA(s) = \lambda P_i. \tag{5.41}$$

Ensuite, notons que par analogie à (5.40), si $D_i(t)$ sont des processus de Poisson de paramètre μ, alors

$$\lim_{t\to\infty} \frac{1}{t}E\int_0^t I\{Q(s) = i\}dD_i(s) = \lim_{t\to\infty} \frac{\mu}{t}\int_0^t P\{Q(s) = i\}ds$$

$$= \mu P_i. \tag{5.42}$$

En Substituant (5.41) et (5.42) dans (5.16) et (5.17), le corollaire est prouvé. □

Corollaire 5.3. *H. Oukid [78]*

Supposons que les processus $A(t)$ et $D_i(t)$ sont des processus de Poisson non homogènes, donc nous avons le système d'équations suivant.

Pour $i = 0, 1, ..., n-1$,

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t [P_i(s) - P_{i-1}(s)]X(s)ds = i \lim_{t\to\infty} \frac{1}{t} \int_0^t P_i(s)Y(s)ds$$

$$-(i+1) \lim_{t\to\infty} \frac{1}{t} \int_0^t P_{i+1}(s)Y(s)ds. \tag{5.43}$$

Pour $i = n$,

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t [P_n(s) - P_{n-1}(s)]X(s)ds = n \lim_{t\to\infty} \frac{1}{t} \int_0^t P_n(s)Y(s)ds. \tag{5.44}$$

Démonstration. Il découle des équations (5.5), (5.6), (5.31) et (5.33) que pour $i = 0, 1, ..., n$,

$$\lim_{t\to\infty} \frac{1}{t} E \int_0^t I[Q(s^-) = i]dA(s) = \lim_{t\to\infty} \frac{1}{t} E \int_0^t I[Q(s^-) = i]X(s)ds$$

$$= \lim_{t\to\infty} \frac{1}{t} \int_0^t P[Q(s) = i]X(s)ds, \tag{5.45}$$

et

$$\lim_{t\to\infty} \frac{1}{t} E \int_0^t I\{Q(s) = i\}dD_i(s) = \lim_{t\to\infty} \frac{1}{t} E \int_0^t I[Q(s^-) = i]Y(s)ds$$

$$= \lim_{t\to\infty} \frac{1}{t} \int_0^t P\{Q(s) = i\}Y(s)ds. \tag{5.46}$$

Le corollaire est prouvé, en substituant (5.45) et (5.46) dans (5.16) et (5.17).

□

5.7 Exemples numériques

Dans cette section, nous présentons des solutions numériques pour deux systèmes de files d'attente. Les paramètres λ et μ sont censés être donnés. Donc, nous cherchons à connaitre le nombre minimal de serveurs n nécessaires pour garantir une probabilité de perte inférieure à un seuil α fixé.

5.7.1 Exemple 1

On considère une file $M/M/n/n$ de paramètres $\lambda = 10$ et $\mu = 1$.
La probabilité de pertes du système, qui est la probabilité pour le système de se trouver dans l'état n est :

$$P_n = \frac{\rho^n}{n!} P_0, \tag{5.47}$$

avec

$$P_0 = [\sum_{i=0}^{n} \frac{\rho^i}{i!}]^{-1}.$$

Pour $\alpha = 0.001$, le nombre minimum de serveurs est $n = 21$.
Pour $\alpha = 0.0001$, on trouve $n = 24$.
Pour $\alpha = 0.00001$, on a $n = 27$.

5.7.2 Exemple 2

On considère une file $D/M/n/n$ où le processus des arrivées est déterministe de longueur $\frac{1}{10}$ et de paramètre $\mu = 1$.
La probabilité de pertes du système est donnée par [2] :

$$P_n = [\sum_{i=0}^{n} \binom{n}{i} \prod_{j=1}^{i} \frac{1 - r_j}{r_j}]^{-1}, \tag{5.48}$$

avec

$$r_j = \int_0^\infty e^{-j\mu x} dA(x),$$

et $A(x)$ la distribution des temps des inter-arrivées.

$$r_j = \exp \frac{-j}{10}.$$

Pour $\alpha = 0.001$, le nombre de serveurs, rendant la probabilité de perte inférieure à 0.001 est $n = 17$
Pour $\alpha = 0.0001$, on trouve $n = 19$.
Pour $\alpha = 0.00001$, on a $n = 21$.

Nous remarquons que le nombre de serveurs nécessaires pour garantir une probabilité de perte inférieure à un seuil fixé dans un système markovien est relativement plus élevé que dans un système où le processus des arrivées est déterministe.

5.8 Conclusion

Dans ce chapitre, une analyse du système de files d'attente multiserveur non-markovien avec perte est fournie à l'aide de la théorie des martingales. Le système d'équations de ce système est obtenu, puis le système d'équation est réduit à d'autres systèmes d'équations, semblables à celle du modèle multiserveur markovien. Les résultats, obtenus dans l'étude, nous permettent d'étudier des modèles non standards des systèmes de télécommunications complexes qui en découlent dans la vie réelle.

CONCLUSION

Cette étude met en évidence l'intérêt et les applications de la méthode des martingales pour l'analyse des systèmes de files d'attente. Dans ce travail, nous nous sommes intéressés au système d'attente $M/G/1$ avec rappels et au système multiserveur non-markovien avec pertes.

Dans un premier temps, nous avons rappelé des résultats connus sur les systèmes d'attente classiques. Ces derniers ne prennent pas en considération le phénomène de répétition de demandes de service : le phénomène en question est étudié par la théorie des files d'attente avec rappels. nous avons ensuite présenté un bref aperçu des définitions et des résultats élémentaires concernant la théorie des martingales.

Nous nous sommes ensuite intéressés à l'étude de l'application de la méthode des martingales aux systèmes de files d'attente.

Tout d'abord, nous avons montré comment la technique proposée par Baccelli et Makowski peut être étendue au système d'attente $M/G/1$ avec rappels. En utilisant l'équation récursive du processus induit aux instants de départ de ce système, nous avons construit une martingale à temps discret arrêtée au premier instant où le système redevient vide. Par ailleurs, sous la condition de stabilité et en utilisant cette martingale, nous obtenons le nombre moyen de clients servis au cours d'une période d'activité. nous avons aussi démontré l'instabilité de ce système au sens de divergence de la chaîne de Markov incluse.

Enfin, nous avons montré comment cette approche peut être adaptée à un système multiserveur non-markovien avec pertes où à l'arrivée d'un client, si l'un des serveurs est libre, le client sera pris en charge immédiatement, dans le cas contraire, le client est perdu. nous avons utilisé la décomposition de Doob-Meyer des semi martingales pour obtenir une représentation sous la forme d'une martingale du processus stochastiques et les équations de la distribution du nombre de clients dans le système. D'abord, nous avons considéré le problème général où les processus d'arrivées et de départs sont des processus ponctuels, ensuite nous nous sommes intéressés aux cas où le processus ponctuel est un processus de Poisson homogène et non-homogène. Des exemples numériques sont donnés où nous cherchons le nombre minimal

de serveurs pour garantir une probabilité de perte (refus) inférieure à un seuil α fixé.

Le domaine d'application de la méthode des martingales aux systèmes de files d'attente mérite d'être élargi surtout aux systèmes avec rappels. les résultats obtenus dans cette thèse, permettent d'envisager de nouvelles perspectives de recherche à savoir :

- Une analyse par simulation statistique des trajectoires.

- Une Analyse d'un système multiserveur avec rappels et avec perte où la capacité de l'orbite est finie ou infinie.

- Une Analyse d'un système multiserveur avec rappels où des temps des rappels de distribution générale au lieu de distribution exponentielle généralement utilisée dans la littérature .

Annexe A

Lois de Probabilités et Processus Stochastiques

Dans cette section, nous présentons les définitions des lois de probabilités et des processus stochastiques utilisés lors de cette thèse.

A.1 Lois de probabilités

A.1.1 Loi géométrique

Soit une expérience aléatoire à deux issues : A (succès) et B (échec) avec $P(A) = p$ et $P(B) = 1 - p = q$. On répète cette expérience, une infinité de fois, de manière indépendante et on note X le nombre d'épreuves nécessaires pour obtenir le premier succès. La variable aléatoire X ainsi obtenue suit une loi géométrique de paramètre p que l'on note $\mathcal{G}(p)$:

$$P\{X = n\} = (1-p)^{n-1}p, \ n = 0, 1, ... \tag{A.1}$$

La moyenne $E(X) = \frac{1}{p}$ et la variance $Var(X) = \frac{1-p}{p^2}$.

A.1.2 Loi de Poisson

Une variable aléatoire X prenant des valeurs discrètes et non-négatives suit une loi de Poisson de paramètre $\lambda > 0$ si sa fonction de distribution de

probabilité est donnée par :

$$P\{X = n\} = e^{-\lambda}\frac{\lambda^n}{n!},\ \lambda > 0,\ et\ n \in N \tag{A.2}$$

La moyenne $E(X) = \lambda$ et la variance $Var(X) = \lambda$.

A.1.3 Loi exponentielle

Une variable aléatoire X prenant des valeurs continues et non-négatives suit une loi exponentielle de paramètre $\lambda > 0$ que l'on note $\mathcal{E}(\lambda)$ si sa fonction de distribution de probabilité est donnée par :

$$f_X(x) = \lambda e^{-\lambda x} 1_{x \geq 0}. \tag{A.3}$$

Moments et propriétés

a. La fonction de répartition est donnée par :

$$F_X(x) = P(X \leq x) = 1 - e^{-\lambda x} 1_{x \geq 0}. \tag{A.4}$$

b. La moyenne $E(X) = \frac{1}{\lambda}$ et la variance $Var(X) = \frac{1}{\lambda^2}$.

En pratique, une v.a. de loi exponentielle représente une durée, typiquement le temps d'attente d'un événement ou une durée de vie. La propriété importante des lois exponentielles est d'être "sans mémoire".
Dans le cas particulier d'un composant électronique dont la durée de vie serait modélisée par une loi exponentielle, cela signifie que la probabilité pour que le composant vive un temps t est la même, qu'il soit neuf ou qu'il ait déjà vécu un temps s. Cette absence de mémoire est caractéristique des lois exponentielles.

Théorème A.1. *Soit X une variable aléatoire exponentielle de paramètre $\lambda > 0$. La variable aléatoire X n'a pas de mémoire :*

$$P\{X \geq s + t / X > s\} = P\{X \geq t\} \tag{A.5}$$

A.2 Processus Stochastiques

Un processus stochastique est une famille de variables aléatoires $\{X(t)\}_{t \in T}$. l'ensemble des temps T peut être discret ou continu. $X(t)$ définit l'état du processus à un instant donné t.

A.2.1 Processus de comptage

Un processus stochastique $N(t);\ t \in R^+$ est un processus de comptage si $N(t)$ représente le nombre total d'événements qui se sont produits entre 0 et t, il doit donc satisfaire

– $N(t) \geq 0$,

– $N(t)$ a des valeurs entières uniquement,

– pour $s < t,\ N(t) - N(s)$ est le nombre d'événements qui ont eu lieu entre s et t.

Un processus de comptage est un processus discret à temps continu.

A.2.2 Processus ponctuels

Un processus ponctuel sur R^+ est décrit par une suite croissante de variables aléatoires $T_0 < T_1 < T_2 < ...T_n < T_{n+1}...$, qui vérifient en outre $T_n \to \infty$ lorsque $n \to \infty$. En posant $S_n = T_n - T_{n-1}$, on peut interpréter :

- T_n comme l'instant où se produit le $n^{ième}$ évènement,

- S_n comme le temps d'attente entre le $(n-1)^{ième}$ et le $n^{ième}$ évènement.

Définition A.2. Soit $\{T_n, n \in N\}$ un processus ponctuel. On appelle fonction aléatoire de comptage (notée f.a. de comptage) le processus $\{N_t, t \geq 0\}$, défini par :

$$N_t = \sup_{n \in N}\{T_n \leq t\} = \sum_{j \in N^*} 1_{\{T_j \leq t\}} \tag{A.6}$$

Le processus N_t représente le nombre d'évènements qui se sont produits jusqu'à l'instant t. On a clairement $N_0 = 0$ et $\forall t \in R^+,\ N_t < \infty$ p.s. puisque $T_n \to \infty$ p.s. lorsque $n \to \infty$.
Une trajectoire type d'une f.a. de comptage est donnée par une fonction en escalier. De plus, par définition, cette trajectoire est dite c.a.d.l.à.g. (continue à droite, limite à gauche).
Notons enfin que la donnée de $\{N_t,\ t \in R^+\}$ est équivalente à celle de la

suite $\{T_n,\ n \in N^*\}$.
De plus, on a le lemme suivant :

Lemme A.3. *Soit* $\{T_n, n \in N\}$ *un processus ponctuel de f.a. de comptage* $\{N_t,\ t \geq 0\}$. *Alors on a :*
$\{N_t \geq n\} = \{T_n \leq t\}$; $\{N_t = n\} = \{T_n \leq t < T_{n+1}\}$; $\{N_s < n \leq N_t\} = \{s < T_n \leq t\}$

A.2.3 Processus de Poisson

On appelle processus de Poisson un processus de comptage vérifiant les trois conditions suivantes :
• Le processus $N(t)$ est homogène dans le temps. Ceci signifie que la probabilité d'avoir n événement dans un intervalle de longueur donnée τ ne dépend que de τ et non pas de la position de l'intervalle dans l'axe temporel :

$$P\{N(t+\tau) - N(t) = n\} = P_n(t),\ \textit{pour tous}\ t, \tau > 0\ \textit{et}\ n = 0, 1, ... \quad \text{(A.7)}$$

• Le processus $N(t)$ est accroissements indépendants et stationnaires. Ceci signifie que, pour tout système d'intervalles disjoints, les nombres d'événements s'y produisant sont des variables aléatoires indépendantes.
• La probabilité que deux événements ou plus se produisent dans un intervalle infiniment petit $\Delta\tau$, est négligeable par rapport à la probabilité qu'il n'y ait qu'un seul événement. d'une manière plus précise, on écrit :

$$P_n(\Delta\tau) = O(\Delta\tau),\ n \geq 2$$
$$P_1(\Delta\tau) = \lambda\Delta\tau + O(\Delta\tau)$$
$$P_0(\Delta\tau) = 1 - \lambda\Delta\tau + O(\Delta\tau)$$

Le coefficient λ est appelé intensité du processus de Poisson.
$m(t) = E[N(t)] = \lambda t$. D'où $\lambda = \frac{m(t)}{t}$. Ceci montre que le paramètre λ désigne le nombre moyen d'événements par unité de temps.
$Var(t) = \lambda t$.

Lien avec la loi de Poisson

Proposition A.4. *Pour tout* $t > 0$ *la variable* $N(t)$ *suit une loi de Poisson de paramètre* λt *:*

$$P\{N(t) = n\} = P_n(t) = e^{-\lambda t}\frac{(\lambda t)^n}{n!},\ \lambda,\ t > 0,\ \textit{et}\ n \in N$$

L'apparition de la loi exponentielle

Lorsque l'on observe un processus de Poisson, il est naturel de s'intéresser au temps d'attente entre les sauts ; on a alors le résultat fondamental suivant :

Proposition A.5. *Si T_n désigne l'instant du $n^{ième}$ saut, alors les variables $T_n - T_{n-1}$ sont i.i.d de loi exponentielle de paramètre $\lambda > 0$.*
Ce résultat justifie la place particulière de la loi exponentielle dans l'étude des modèles de durée. On en déduit que la loi de T_n est la loi d'Erlang de paramètres (n, λ).

Remarque A.6. *Le processus de Poisson est le processus stochastique le plus utilisé dans la théorie des files d'attente. Il modélisera généralement le processus d'arrivée des clients dans un système. On parlera alors "d'arrivées poissoniennes".*

Martingales associées au processus de Poisson

Proposition A.7. *les processus $M_t = N_t - \lambda t$ et $M_t^2 - \lambda t$ sont des martingales relativement à la filtration $\sigma(N_s,\ s \leq t)$.*
Pour $\theta > 0$, le processus $\exp\big(-\theta N_t + \lambda t(1 - e^{-\theta})\big)$ est également une martingale relativement à la filtration $\sigma(N_s,\ s \leq t)$.

Démonstration. Pour le processusM_t, on écrit que $E_s[(N_t - N_s)] = \lambda(t - s)$ par la propriété d'accroissements indépendants ; mais on a aussi
$E_s[(N_t - N_s)] = E_s[N_t] - N_s$; en égalant les 2 formes de l'expression on obtient, $E_s[N_t] - \lambda t = N_s - \lambda s$, ce qui prouve le résultat. Le raisonnement est analogue pour $M_t^2 - \lambda t$. □

Il découle aisément de cette proposition que $\lim_{t\to\infty} \frac{N_t}{t} = \lambda$, dans L^2. En fait la convergence est même presque sur.

A.2.4 Processus de Poisson non homogène

La différence entre un processus de Poisson de base et un processus de Poisson non homogène réside dans le fait que les accroissements ne sont plus stationnaires. Cela est plus réaliste car, en général, le taux d'occurrence d'événements dépend du temps. Par exemple, pour un commerce, le taux d'entrée des clients varie au cours du temps. On imagine bien que pour un restaurant, ce taux sera élevé entre 11h et 13h puis faible dans l'après-midi

(voir nul si le restaurant ferme durant cette période) puis à nouveau élevé entre 18h et 20h, après quoi il diminuera graduellement pour retomber à zéro.

Définition A.8. On définit alors un processus de Poisson non stationnaire par le fait que l'intensité du processus est une fonction du temps, $\lambda(t)$. On obtient un processus qui n'est plus stationnaire, la loi de l'accroissement $N(t+h) - N(t)$ est alors une loi de Poisson de paramètre $\int_t^{t+h} \lambda(u)du$. Les deux autres hypothèses (accroissements indépendants, 1 seule arrivée à la fois) sont conservées.

Annexe B

Rappels : Espérance conditionnelle et Théorèmes de Convergence de Lebesgue

Nous donnons dans cette section quelques définitions et propriétés qui nous seront utiles dans cette thèse.

B.1 Espérance conditionnelle

Définition B.1. Pour une variable aléatoire X de $(\Omega, \Im, P)$ et une sous-tribu de $\Im$ notée β, l'espérance conditionnelle de X sachant β, notée $E[X \mid \beta]$, représente l'unique variable aléatoire $\beta-$mesurable telle que

$$\int_B E[X \mid \beta] dP = \int_B X dP \tag{B.1}$$

pour tout élément B de β. L'espérance conditionnelle est également caractérisée par le fait que pour toute variable aléatoire Y bornée et $\beta-$mesurable, $E[XY] = E[E[X \mid \beta]Y]$.

Les propositions suivantes résument les propriétés élémentaires de l'espérance conditionnelle.

B.1.1 Propriétés de l'espérance conditionnelle analogues à celles de l'espérance

Proposition B.2. *Soit X une variable aléatoire dans $L^1(\Omega, \Im, P)$. On note β une sous-tribu de $\Im$.*
a) Pour tous réels a et b et toute variable aléatoire réelle X intégrable,

$$E[aX + b|\beta] = aE[X|\beta] + b;$$

et pour toutes variables aléatoires réelles X_1, X_2 intégrables

$$E[X_1 + X_2|\beta] = E[X_1|\beta] + E[X_2|\beta].$$

b) Si $X_1 \leq X_2$ p.s. alors $E[X_1|\beta] \leq E[X_2|\beta]$.
c) Si X et X_n sont des variables aléatoires réelles dans $L^1(\Omega, \Im, P)$ alors

$$X_n \to X \Longrightarrow E[X_n|\beta] \to E[X|\beta].$$

d) Si X_n sont des variables aléatoires positives, alors

$$E[\liminf_n X_n|\beta] \leq \liminf_n E[X_n|\beta].$$

e) Si $X_n \to X$ p.s. avec pour tout n, $X_n \leq Z \in L^1(\Omega, \Im, P)$, alors

$$\lim_n E[X_n|\beta] = E[X|\beta].$$

f) Soit f une fonction continue et convexe et X une variable aléatoire réelle telle que X et $f(X)$ sont intégrables, alors

$$f\big(E[X|\beta]\big) \leq E[f(X)|\beta];$$

en particulier

$$|E[X|\beta]| \leq E[|X||\beta],$$

et par conséquent

$$E\big[|E[X|\beta]|\big] \leq E[|X|].$$

B.1.2 Propriétés spécifiques à l'espérance conditionnelle

Proposition B.3. *a) Si X est intégrable alors $E[X|\beta]$ l'est aussi et $E[E[X|\beta]] = E[X]$.*
b) Si X est une variable aléatoire réelle β-mesurable alors $E[X|\beta] = X$ p.s. en particulier $E[1|\beta] = 1$.
Donc si ψ est une fonction mesurable telle que $\psi(Y)$ est intégrable,

$$E[\psi(Y)|\beta] = \psi(Y).$$

c) Soient X et Z deux variables aléatoires réelles intégrables telles que XZ soit aussi intégrable. Supposons Z β-mesurable alors

$$E[XZ|\beta] = ZE[X|\beta] \text{ p.s.};$$

en particulier si ψ est une fonction mesurable telle que $\psi(Y)$ et $X\psi(Y)$ soient intégrables,

$$E[X\psi(Y)|Y] = \psi(Y)E[X|Y].$$

d) Si β_1 et β_2 sont deux sous-tribus de $\Im$ telles que $\beta_1 \subset \beta_2$, alors

$$E[E[X|\beta_1]|\beta_2] = E[X|\beta_1]$$

e) Soit β une sous-tribus de $\Im$. Si X est une variable aléatoire dans $L^1(\Omega, \Im, P)$, telle que $\sigma(X)$ et β sont indépendantes, alors

$$E[X|\beta] = E[X];$$

en particulier si X et Y sont deux variables aléatoires réelles indépendantes telles que $X \in L^1(\Omega, \Im, P)$ alors

$$E[X|Y] = E[X];$$

la réciproque de ce dernier point étant fausse.

Nota Bene Il faut noter qu'en général l'espérance conditionnelle $E[Y \mid X]$ est une variable aléatoire et non un nombre. On peut l'interpréter comme la valeur moyenne prise par Y lorsque l'on connaît X. Elle pourra donc s'écrire comme une fonction de X.

B.2 Théorèmes de Convergence pour l'intégrale de Lebesgue

Théorème B.4. (de convergence monotone).

Soit $(f_n)_{n\in N}$ une suite croissante de fonctions mesurables positives sur $(\Omega, \Im, P)$, convergeant ponctuellement vers f. Alors f est mesurable et

$$\lim_{n\to\infty} \int f_n d\mu = \int f d\mu.$$

Théorème B.5. (de convergence dominée de Lebesgue).

Soit $(f_n)_{n\in N}$ une suite de fonctions telles que $|f_n| \leq g$ où g est intégrable et f_n converge simplement vers f. Alors f est intégrable et

$$\lim_{n\to\infty} \int f_n d\mu = \int f d\mu.$$

Lemme B.6. (Inégalité de Lenglart-Rebolledo)).

Soient X et Y deux processus F-adapté avec trajectoire dans D (espace des fonctions càadlàg).. Supposons que Y est croissante et que pour tout T temps d'arrêt fini on a $E[X(T)] \leq E[Y(T)]$. Si Y est prévisible, alors Pour tout T temps d'arrêt fini et pour tout $\epsilon, \eta > 0$,

$$P(\sup_{s\geq T} \mid X_s \mid \geq \epsilon) \leq \frac{\eta}{\epsilon} + P(Y(T) \geq \eta). \tag{B.2}$$

Démonstration. Voir [54]. □

Bibliographie

[1] V. M. Abramov. Analysis of multiserver retrial queueing system : A martingale approach and an algorithm of solution. *Annals of Operations Research*, 141 :19–52, 2006.

[2] V. M. Abramov. Asymptotic analysis of loss probabilities in GI/M/m/n queueing systems as n increases to infinity. *Quality Technology & Quantitative Management*, 4(3) :379–393, 2007.

[3] A. Aissani. Unreliable queuing with repeated orders. *Microelectronics Reliability*, 33(14) :2093–2106, 1993.

[4] A. Aissani. A survey of retrial queues. *Actes des Journées de Statistiques Appliquées*, pages 1–11, 1994.

[5] A. Aissani. An MX/G/1 retrial queue with unreliable server and vacations. In *Proceedings of the 10th International Conference on Analytical and Stochastic Modelling Techniques and Applications, ASMTA*, volume 3, pages 175–180, 2003.

[6] A. Aissani and J. R. Artalejo. On the single server retrial queue subject to breakdowns. *Queueing systems*, 30(3-4) :309–321, 1998.

[7] B. Almási, J. Roszik, and J. Sztrik. Homogeneous finite-source retrial queues with server subject to breakdowns and repairs. *Mathematical and Computer Modelling*, 42(5-6) :673–682, 2005.

[8] J. R. Artalejo. New results in retrial queueing systems with breakdown of the servers. *Statistica Neerlandica*, 48(1) :23–36, 1994.

[9] J. R. Artalejo. Analysis of an M/G/1 queue with constant repeated attempts and server vacations. *Computers & Operations Research*, 24(6) :493–504, 1997.

[10] J. R. Artalejo. 1^{st} international workshop on retrial queues. *Top*, 7(2) :169–353, 1999.

[11] J. R. Artalejo. Accessible bibliography on retrial queues. *Mathematical and computer modelling*, 30(3-4) :1–6, 1999.

[12] J. R. Artalejo. Retrial queueing systems. *Mathematical and Computer Modelling*, 30(3-4) :1–228, 1999.

[13] J. R. Artalejo. Algorithmic methods in retrial queues. *Annals of Operation Research*, 141 :1–301, 2006.

[14] J. R. Artalejo. Accessible bibliography on retrial queues : Progress in 2000–2009. *Mathematical and computer modelling*, 51(9) :1071–1081, 2010.

[15] J. R. Artalejo. *Retrial Queues*. Wiley Encyclopedia of Operation Ressearch and Management Science, 2011.

[16] J. R. Artalejo and I. Atencia. On the single server retrial queue with batch arrivals. *Sankhyā : The Indian Journal of Statistics*, pages 140–158, 2004.

[17] J. R. Artalejo and A. Gomez-Corral. Steady state solution of a single-server queue with linear repeated requests. *Journal of Applied Probability*, 34(1) :223–233, 1997.

[18] J. R. Artalejo and A. Gómez-Corral. Advances in retrial queues. *European Journal of Operation Research*, 2008.

[19] J. R. Artalejo and A. Gomez-Corral. *Retrial Queueing Systems - A Computational Approach*. Berlin : Springer, 2008.

[20] J. R. Artalejo and M. J. Lopez-Herrero. On the distribution of the number of retrials. *Applied mathematical modelling*, 31(3) :478–489, 2007.

[21] J. R. Artalejo and T. Phung-Duc. Markovian single server retrial queues with two way communication. In *Proceedings of the 6th International Conference on Queueing Theory and Network Applications*, pages 1–7. ACM, 2011.

[22] J. R. Artalejo and T. Phung-Duc. Single server retrial queues with two way communication. *Applied Mathematical Modelling*, 37(4) :1811–1822, 2013.

[23] I. Atencia, G.l Bouza, and P. Moreno. An M [x]/G/1 retrial queue with server breakdowns and constant rate of repeated attempts. *Annals of Operations Research*, 157(1) :225–243, 2008.

[24] F. Baccelli and A. M. Makowski. Direct martingale arguments for stability : the M/G/1 case. *Systems & control letters*, 6(3) :181–186, 1985.

[25] F. Baccelli and A. M. Makowski. Martingale relations for the M/GI/1 queue with markov modulated poisson input. *Stochastic processes and their applications*, 38(1) :99–133, 1991.

[26] J. Bae. The derivation of the laplace transform of a wet period in a finite dam via martingales. *Applied mathematics letters*, 19(2) :186–190, 2006.

[27] B. Baynat. *La théorie des files d'attente : des chaînes de Markov aux réseaux à forme produit*. Hermès, 2000.

[28] P. Brémaud. *Point processes and queues : martingale dynamics*. Springer-Verlag, 1981.

[29] SR. Chakravarthy and A. Dudin. Analysis of a retrial queuing model with map arrivals and two types of customers. *Mathematical and Computer Modelling*, 37(3-4) :343–363, 2003.

[30] B. D. Choi, Y. C. Kim, and Y. W. Lee. The M/M/c retrial queue with geometric loss and feedback. *Computers & Mathematics with Applications*, 36(6) :41–52, 1998.

[31] Q . H. Choo and B. Conolly. New results in the theory of repeated orders queueing systems. *Journal of applied Probability*, 16(3) :631–640, 1979.

[32] G. Choudhury. Steady state analysis of an M/G/1 queue with linear retrial policy and two phase service under bernoulli vacation schedule. *Applied Mathematical Modelling*, 32(12) :2480–2489, 2008.

[33] G. Choudhury and J-Ch. Ke. A batch arrival retrial queue with general retrial times under bernoulli vacation schedule for unreliable server and delaying repair. *Applied Mathematical Modelling*, 36(1) :255–269, 2012.

[34] J. W. Cohen. Basic problems of telephone traffic theory and the influence of repeated calls. *Philips Telecommunication Review*, 18(2) :49–100, 1957.

[35] C. Dellacherie. *Capacités et processus stochastiques*, volume 357. Springer-Verlag Berlin, 1972.

[36] C. Dellacherie and P. Meyer. Probabilités et potentiel II. *Hermann, Paris*, 1 :980, 1980.

[37] N. V. Djellab. On the M/G/1 retrial queue subjected to breakdowns. *RAIRO-Operations Research*, 36(4) :299–310, 2002.

[38] J. L. Doob. Regularity properties of certain families of chance variables. *Transactions of the American Mathematical Society*, 47(3) :455–486, 1940.

[39] J. L. Doob. *Stochastic processes*, volume 7. Wiley New York, 1953.

[40] A. Elldin. Approach to the theoretical description of repeated call attempts. *Ericsson Technics*, 23(3) :346–407, 1967.

[41] A. K. Erlang. Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges. *Elektrotkeknikeren*, 13 :5–13, 1917.

[42] G. I. Falin. A single-line system with secondary orders. *Engineering Cybernetics*, 17(2) :76–83, 1979.

[43] G. I. Falin. A survey of retrial queues. *Queueing systems*, 7(2) :127–167, 1990.

[44] G. I. Falin. An M/G/1 retrial queue with an unreliable server and general repair times. *Performance Evaluation*, 67(7) :569–582, 2010.

[45] G. I. Falin, J. R. Artalejo, and M. Martin. On the single server retrial queue with priority customers. *Queueing systems*, 14(3-4) :439–455, 1993.

[46] G. I. Falin and J. G. C. Templeton. *Retrial queues*, volume 75. CRC Press, 1997.

[47] W. Feller. *An introduction to probability theory and its applications, volume II*, volume 2. Wiley, New York, 1971.

[48] E. Gelenbe and G. Pujolle. Introduction aux réseaux de files d'attente. 1982.

[49] E. Gelenbe and Guy. Pujolle. *Introduction aux réseaux de files d'attente*. Eyrolles, 1982.

[50] A. Gomez-Corral and M. F. Ramalhoto. On the waiting time distribution and the busy period of a retrial queue with constant retrial rate. *Stochastic Modelling and Applications*, 3(2) :37–47, 2000.

[51] B. S. Greenberg. M/G/1 queueing systems with returning customers. *Journal of Applied Probability*, 26(1) :152–163, 1989.

[52] T. Hanschke. Explicit formulas for the characteristics of the M/M/2/2 queue with repeated attempts. *Journal of applied probability*, 24(2) :486–494, 1987.

[53] O. Hashida and Kawashima. Buffer behavior with repeated calls. *Electronics and Communication in Japan*, 26-B(3) :222–228, 1979.

[54] J. Jacod and A. Shiryaev. *Limit theorems for stochastic processes*, volume 288. Springer Science & Business Media, 2013.

[55] Sharma G. C. Jain, M. and R. S. Pundhir. The M/G/1 retrial queueing system with set up, server breakdown and repair. *Ganita*, 58(2) :137–155, 2007.

[56] J. Keilson, J. Cozzolino, and H. Young. A service system with unfilled requests repeated. *Operations Research*, 16(6) :1126–1137, 1968.

[57] J. Kim and B. Kim. A survey of retrial queueing systems. *Annals of Operations Research*, 247(1) :3–36, 2016.

[58] J-S. Kim. Retrial queueing system with collision and impatience. *Communications of the Korean Mathematical Society*, 25(4) :647–653, 2010.

[59] K. KJ Kinateder and E. Y. Lee. A new approach to the busy period of the M/M/1 queue. *Queueing systems*, 35(1) :105–115, 2000.

[60] L. Kleinrock. *Queueing systems, volume 1.* Wiley interscience, New York, 1975.

[61] L. Kleinrock. *Queueing systems, volume 2 : Computer applications*, volume 66. wiley New York, 1976.

[62] A. Krishnamoorthy, T. G. Deepak, and V. C. Joshua. An M| G| 1 retrial queue with nonpersistent customers and orbital search. *Stochastic Analysis and Applications*, 23(5) :975–997, 2005.

[63] V. G. Kulkarni. Expected waiting times in a multiclass batch arrival retrial queue. *Journal of Applied Probability*, 23(1) :144–154, 1986.

[64] V. G. Kulkarni and B. D. Choi. Retrial queues with server subject to breakdowns and repairs. *Queueing systems*, 7(2) :191–208, 1990.

[65] V. G. Kulkarni and H. M. Liang. Retrial queues revisited. *Frontiers in queueing : Models and applications in science and engineering*, 7 :19, 1997.

[66] B. K. Kumar and S. P. Madheswari. M$\hat{\text{X}}$/G/1 retrial queue with multiple vacations and starting failures. *Opsearch-New Delhi*, 40(2) :115–137, 2003.

[67] B. K. Kumar, R. Rukmani, and V. Thangaraj. On multiserver feedback retrial queue with finite buffer. *Applied Mathematical Modelling*, 33(4) :2062–2083, 2009.

[68] Ch. Langaris and E. Moutzoukis. A retrial queue with structured batch arrivals, priorities and server vacations. *Queueing Systems*, 20(3-4) :341–368, 1995.

[69] Ch. Langaris and E. Moutzoukis. A retrial queue with structured batch arrivals, priorities and server vacations. *Queueing systems*, 20(3) :341–368, 1995.

[70] P. Levy. *Théorie de l'addition des variables aléatoires*. Gauthier-Villars Paris, 1954.

[71] H-M. Liang and V. G. Kulkarni. Stability condition for a single-server retrial queue. *Advances in Applied Probability*, 25(3) :690–701, 1993.

[72] R. Liptser and A. N. Shiryayev. Theory of martingales. *Mathematics and its Applications. Kluwer, Dordrecht*, 1989.

[73] R. Liptser and A. N. Shiryayev. *Theory of martingales*, volume 49. Springer Science & Business Media, 2012.

[74] J. Lubacz and J. Roberts. A new approach to the single server repeat attempt system with balking. In *Proc. 3rd Intern. Seminar Teletraffic Theory*, pages 290–293, 1984.

[75] J. Neveu. *Martingales à temps discret*. Masson, Paris, 1972.

[76] H. Oukid and A. Aissani. Some martingale relations for M/G/1 retrial queue. *International Conference "Modern Probabilistic Methods for Analysis and Optimisation of Information and Telecommunication Networks" (21st Belarusian Winter Workshop in Queueing Theory-BWWQT*, 2011.

[77] H. Oukid and A. Aissani. Analysis of multiserver queues with losses via martingale. *Marrakesh International Conference on Probability and Statistics "MICPS"*, 2013.

[78] H. Oukid and A. Aissani. Analyse de systèmes d'attente à plusieurs serveurs via les martingales. *Laboratoire de Recherche en Informatique Intelligente, Mathématiques et Applications (RIIMA)*, 2015.

[79] H. Oukid and A. Aissani. A new look at the M/G/1 retrial queue. *16éme Congrés Annuel de la Société Française de Recherche Opérationnelle et d'Aide á la Décision*, 2015.

[80] H. Oukid and A. Aissani. A new look at the M/G/1 retrial queue : A martingale approach. *Advanced Studies in Contemporary Mathematics*, 28(3) :413–422, 2018.

[81] N. Oukid. *Comparaisons stochastiques de files d'attente. Thèse de Magister*. Université de Blida, 1995.

[82] N. Oukid and A. Aissani. Bounds on busy period for queues with breakdowns. *Advances and Applications in Statistics*, 11(2) :137–156, 2009.

[83] T. Phung-Duc, Kasahara Sh. Masuyama, H., and Y. Takahashi. State-dependent M/M/c/c+ r retrial queues with bernoulli abandonment. *Journal of Industrial and Management Optimization*, 6(3) :517–540, 2010.

[84] A. Rodrigo. Estimators of the retrial rate in M/G/1 retrial queues. *Asia-Pacific Journal of Operational Research*, 23(02) :193–213, 2006.

[85] L. C. G. Rogers and D. Williams. *Diffusions, Markov processes and martingales : Volume 2, Itô calculus*, volume 2. Cambridge university press, 2000.

[86] W. A. Rosenkrantz et al. Calculation of the laplace transform of the length of the busy period for the M/G/1 queue via martingales. *The Annals of Probability*, 11(3) :817–818, 1983.

[87] M. Roughan and C. E. M. Pearce. Martingale methods for analysing single-server queues. *Queueing Systems*, 41(3) :205–239, 2002.

[88] Y. W. Shin and T. S. Choo. M/M/s queue with impatient customers and retrials. *Applied Mathematical Modelling*, 33(6) :2596–2606, 2009.

[89] A. N. Shiryaev. *Problems in Probability*. New York : Springer, 2012.

[90] J. L. Snell. Applications of martingale system theorems. *Transactions of the American Mathematical Society*, 73(2) :293–312, 1952.

[91] Sh. Stidham Jr. Analysis, design, and control of queueing systems. *Operations Research*, 50(1) :197–216, 2002.

[92] L. Takács. *Introduction to the Theory of Queues*. London : Oxford University Press, 1962.

[93] J. G. C. Templeton. Retrial queues. *Queueing systems*, 7(2) :125–227, 1990.

[94] J. G. C. Templeton. Retrial queues. *Top*, 7(2) :351–353, 1999.

[95] T. Van Do. An efficient computation algorithm for a multiserver feedback retrial queue with a large queueing capacity. *Applied Mathematical Modelling*, 34(8) :2272–2278, 2010.

[96] J. Ville. *Etude critique de la notion de collectif*. Gauthier-Villars, Paris, 1939.

[97] D. Williams. *Probability with martingales*. Cambridge university press, 1991.

[98] J. Wu, Z. Liu, and G. Yang. Analysis of the finite source MAP/PH/N retrial g-queue operating in a random environment. *Applied Mathematical Modelling*, 35(3) :1184–1193, 2011.

[99] X. Wu, P. Brill, M. Hlynka, and J. Wang. An M/G/1 retrial queue with balking and retrials during service. *International Journal of Operational Research*, 1(1-2) :30–51, 2005.

[100] T. Yang, M. J. M. Posner, J. G. C. Templeton, and H. Li. An approximation method for the M/G/1 retrial queue with general retrial times. *European Journal of Operational Research*, 76(3) :552–562, 1994.

[101] T. Yang and J. G. C. Templeton. A survey on retrial queues. *Queueing systems*, 2(3) :201–233, 1987.