Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université Abderrahmane Mira de Bejaia
Faculté des Sciences Exactes
Département de Mathématiques

Mémoire de Master Académique en Mathématiques

Spécialité : Probabilités Statistique et Applications

Thème

Analyse stochastique des systèmes d'attente avec rappels et impatience

Présenté par :

M. BAHIOU Ouali

M. OUAKLI Sami

Devant le jury

M. OUAZINE Sofiane	Président	M.C.B	Université de Bejaia
Mme. BOURAINE Louiza	Promotrice	M.C.A	Université de Bejaia
M. BOUALAM Mohamed	Examinateur	Professeur	Université de Bejaia
M*lle*. **LAKAOUR** Lamia	Examinatrice	Docteur	Université de Béjaia

Année universitaire : 2018/2019

Remerciements

Louange à Dieu, le miséricordieux, sans lui rien de tout cela n'aurait pu être.

Nous tenons à exprimer toutes nos profondes gratitudes envers notre encadreur Mme. BOURAINE-BERDJOUDJ Louiza, pour la manière dont elle nous a encadré, ses conseils avisés, sa disponibilité ses remarques constructives nous ont beaucoup apporté tout au long de ce travail.

On remercie M^r OUAZINE Sofiane d'avoir accepter de présider le jury. Ainsi que les membres du jury M^r BOUALAM Mohamed et M^{lle} LAKAOUR Lamia qui nous ont honoré de leurs présence et d'avoir accepté d'évaluer notre travail à sa juste valeur.

Que toute personne qui a contribué, de près ou de loin, à l'élaboration de ce mémoire, veuillez bien trouver ici l'expression de nos sincères remerciements.

Pour terminer, nos derniers mots de remerciements vont tout naturellement à nos chers parents, frères et soeurs, pour leur soutien et leur confiance en nous.

J'espère que ce mémoire saura combler vos attentes.

Dédicaces

Du profond de mon cœur, je dédie ce travail . . .

A la mémoire de ma grand-mère le destin ne nous a pas laissé le temps pour jouir ce bonheur ensemble et de t'exprimer tout mon respect. Puisse le Dieu le tout puissant vous accueillira dans son saint paradis.

A celle qui m'a transmis la vie, l'amour, le courage, à toi ma très chère et douce mère.

Au meilleur des pères : Aucune dédicace ne saurait exprimer l'amour, l'estime et le respect que j'ai toujours eu pour vous.

A mon cher frère Zahir et son épouse, ainsi que leurs enfants.

A mes chers frères Nassim et Bouzid, les mots ne suffisent guère pour exprimer l'affection que je porte pour vous.

A mes chères soeurs, leurs époux et leurs enfants.

A mes amis, Fares, Farouk, Kahina, Cylia, Sara, Kiki.

O. BAHIOU

TABLE DES MATIÈRES

TABLE DES FIGURES

INTRODUCTION GÉNÉRALE

La théorie des files d'attente, ou queues, est l'un des outils les plus puissants pour la modélisation des systèmes de logistique et de communication. Cette théorie tire son origine des recherches de l'ingénieur Danois Agner Krarup Erlang entre (1909 - 1920), elle a été inspirée au Danemark avec le développement de la téléphonie. 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 banqueroute. 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 meilleurs ingénieurs, A. K. Erlang, de travailler à une conceptualisation.
Le sujet a inspiré et continue à inspirer de nombreux chercheurs comme en témoignent les nombreuses publications parues à ce jour dans le domaine. C'est grace aux apports des mathématiciens Kendall, Pollaczek et Kolmogorov que la théorie s'est vraiment développée.

Le modèle classique de files 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, le trafic routier 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. Le développement du temps-réel est aujourd'hui une préoccupation majeure. Dans le contexte multimédia par exemple, on cherche à transmettre les données

en un temps très bref. Des flux de données de types parfois très différents doivent cohabiter, l'intégrité de ces flux doit être respectée et les réseaux qui les transmettent doivent être à la fois flexibles et rapides. Toute donnée doit alors avoir une "durée de vie" très limitée dans le système, puisque son traitement doit être instantané.

Pour prendre compte la contrainte du temps-réel dans les réseaux, on doit enrichir le modèle classique de la file d'attente par un nouveau paramètre, le délai des requêtes qui entrent dans la file. On considère qu'une requête est perdue dès qu'elle dépasse ce délai sans avoir commencé son traitement. En termes idéalistes, on parle de files d'attente avec clients impatients. Les clients ont une patience pour entrer en service, au delà de laquelle ils choisissent de quitter la file. Ce modèle a initialement été construit pour décrire les réseaux téléphoniques où les clients mis en attente raccrochaient au bout d'un certain temps si leur appel n'était pas pris en compte. L'impatience dans les systèmes d'attente est due à plusieurs facteurs : nombre de serveurs insuffisant, mauvaise gestion du système ..., et le phénomène d'impatience a un impact négatif sur l'économie des entreprises. Le 1^{er} article publié sur les systèmes avec impatience remonte aux travaux de Barrer (1957) [12] et une bibliographie générale sur les systèmes d'attente avec clients impatients est donné par Wang et al. (2010) [31].
Généralement, il existe trois formes d'impatience. La première est la réticence d'un client, à se joindre à une file d'attente à son arrivée (balking), la seconde est la réticence de rester dans la file après avoir attendu (reneging) et la troisième est le jockey entre les files d'attente qui sont en parallèle (jockeying). Pour plus de détails, le lecteur peut se référer à [1, 20, 22, 25].

D'autres systèmes réels de plus en plus complexes apparaissent, tel que les systèmes téléphoniques, où les abonnés répétaient leurs appels en recomposant le numéro plusieurs fois jusqu'à l'obtention de la communication. "Attendre", constitue la tâche la plus désagréable de la vie moderne. Plusieurs situations d'attente ont donc la caractéristique que les clients doivent rappeler, pour être servis. Entre les rappels succéssifs, le client en question se trouve en orbite. Depuis, ce phénomène de répétition de demandes du service a poussé plusieurs chercheurs à étendre le modèle d'attente classique à celui dit avec rappels.
Parmi les premières contributions sérieuses sur les modèles d'attente avec rappels, on trouve celle de Cohen (1957)[15]. Une description complète de situations où les systèmes de files d'attente avec rappels se présentent peut être trouvée dans les articles de synthèse

de Aissani (1994)[2] et dans les monographies de Falin et Templeton (1997)[16], Artalejo et Gomez-Corral (2008) [6] et J. Kim et B. Kim (2016) [21]. Une classifiction bibliographique récente est donnée dans les articles de Artalejo (1999) et (2010) [4, 5] et dans l'article de Shekkar et al.(2016)[28].

Le phénomène d'impatience est présent aussi dans l'orbite, en effet un client secondaire peut quitter le système après plusieurs tentatives échouées de rappels (voir Falin [16]). Suganthi et al. (2015)[30] ont donné une analyse stationnaire du système $M/M/1$ avec rappels et impatience. Lubacz et Roberts [24] ont présenté une nouvelle approche aux systèmes avec rappels et balking. Hamache (2018) [19] a étudié le système $M/M/1/N$ avec découragement. Azhagappan et al. (2018) [11] ont donné une solution transitoire du système $M/M/1$ avec rappels reneging. Gao et al. (2017) [17] ont analysé le système M/M/1 avec rappels constants et impatience.

L'objectif de notre travail concerne l'analyse stochastique des systèmes de files d'attente $M/M/1$ et $M/G/1$ avec rappels tout en considérant l'impatience des clients en orbite d'une part. D'autre part, il s'agit d'effectuer une étude numérique pour montrer l'influence des taux de rappels et de l'impatience sur les mesures de performance et de comparer les résultats en utilisant différentes distributions du temps de service dans le cas du modèle $M/G/1$ avec rappels et impatience.

Ce mémoire est constitué de trois chapitres, d'une conclusion générale et une bibliographie.

- ✓ Le premier chapitre présente brièvement le formalisme des systèmes de files d'attente classiques avec impatience puis avec rappels. Nous nous focalisons surtout sur les systèmes M/M/1 avec rappels et M/G/1 avec rappels.
- ✓ Le deuxième chapitre est consacré à l'analyse stochastique et numérique du système M/M/1 avec rappels et impatience dans l'orbite.
- ✓ Le troisième chapitre concerne la généralisation de la distribution de service du systeme $M/M/1$ au système $M/G/1$ avec rappels et impatience en orbite. L'analyse sera faite en utilisant la chaîne de Markov induite et les fonctions génératrices.
- ✓ Le travail s'achève par une conclusion en mettant l'accent sur les perspectives de recherche qui ont découlent.

CHAPITRE 1

LA THÉORIE DES FILES D'ATTENTE

Introduction

Les files d'attente peuvent être considérées comme un phénomène caractéristique de la vie contemporaire. On les rencontre dans divers domaines d'activité : guichet, traitement des instructions par un processeur, gestion de communications téléphoniques, trafic routier, etc.
On parle de file d'attente chaque fois que les clients se présentent d'une manière aléatoire à des stations pour réclamer un service dont la durée est généralemeent aléatoire. On s'intéresse essentiellement à deux grandeurs : Le nombre de clients dans le système, et le temps passé par un client dans le système. Ce dernier se décompose en un temps d'attente et un temps de service. La théorie des files d'attente est un des outils analytiques les plus puissants pour la modélisation des systèmes dynamiques.
Dans ce chapitre, on présente le formalisme de la théorie de files d'attente en se basant sur quelques modèles à savoir le modèle M/M/1 et M/G/1 classiques, et avec rappels

1.1 Les systèmes de files d'attente classiques

Pour décrire une file d'attente, on doit préciser les éléments suivants :
– La nature du processus des inter-arrivées qui est définie par la distribution des intervalles séparant deux arrivées consécutives.

– La distribution du temps aléatoire de service.
– Le nombre s des stations de service montées en parallèle.
– La capacité N du système : Si $N > 1$, la file ne peut dépasser une longueur de $(N - s)$ unités. Dans ce cas, certains clients qui arrivent vers le système n'ont pas la possibilité d'y entrer.

1.1.1 Notions élémentaires sur les files d'attente

a- Une file simple :

Une file d'attente simple est un système constitué d'un ou plusieurs serveurs et d'un espace d'attente. Les clients arrivent de l'extérieur, patientent éventuellement dans la file d'attente, reçoivent un service, puis quittent la station. Afin de spécifier complètement une file d'attente simple, on doit caractériser le processus des arrivées des clients, le temps de service ainsi que la structure et la discipline de service de la file d'attente.

b- Processus d'inter-arrivées :

L'arrivée des clients à la station sera décrite à l'aide d'un processus stochastique de comptage $(N_t)_{t\geq 0}$. Si A_n désigne la variable aléatoire mesurant l'instant d'arrivée du $n^{ème}$ client dans le système, on aura ainsi : $A_0 = 0$ et $A_n = inf\{t; N_t = n\}$. Si T_n désigne la variable aléatoire mesurant le temps séparant l'arrivée du $(n-1)^{ème}$ client et du $n^{ème}$ client, on a alors

$$T_n = A_n - A_{n-1}.$$

c- Temps de service :

Considérons tout d'abord une file à serveur unique. On note D_n la variable aléatoire mesurant l'instant de départ du $n^{ème}$ client du système et Y_n la variable aléatoire mesurant le temps de service du $n^{ème}$ client (le temps séparant le début et la fin du service). L'instant de départ correspond toujours à une fin de service, mais ne correspond pas forcément à un début de service. Il se peut en effet qu'un client qui quitte la station laisse celle-ci vide. Le serveur est alors inoccupé jusqu'à l'arrivée du prochain client.

d- Notations de Kendall :

Un système de files d'attente est généralement représenté suivant la notation Kendall définie comme suit : $T/Y/s(/N/K/D_s)$ tels que

T : Indique le processus d'arrivées des clients.

Y : Décrit la distribution des temps de service d'un client.

s : Nombre de serveurs.

N : Capacité du système. C'est le nombre maximal de clients dans le système y compris ceux en service.

K : Population des usagers.

D_s : Discipline de service, c'est la façon dont les clients sont ordonnés pour être servi. Les disciplines utilisées sont les suivantes :

- FIFO (First-In-First-Out) ou FCFS (First-Come-First-Served) : C'est la file standard dans laquelle les clients sont servis dans leur ordre d'arrivée. Notons que les disciplines FIFO et FCFS ne sont pas équivalentes lorsque la file contient plusieurs serveurs. Dans la première, le premier client arrivé sera le premier à quitter la file alors que la deuxième, il sera le premier à commencer son service. Rien n'empêche alors qu'un client qui commence son service après lui, dans un autre serveur, termine avant lui.
- LIFO (Last-In-First-Out) ou LCFS (Last-Come-First-Served). Cela correspond à une pile, dans laquelle le dernier client arrivé sera le premier traité (retiré de la pile). A nouveau, les disciplines LIFO et LCFS ne sont équivalentes que pour une file monoserveur.
- SIRO (Served In Random Order), les clients sont servis aléatoirement.
- PS (Processor Sharing), les clients sont servis de manière égale. La capacité du système est partagée entre les clients.
  Lorsque les trois derniers éléments de la notation de Kendall ne sont pas précisés, ils sont pris par défaut comme suit : $N = \infty, K = \infty, D_s = FIFO$.

Exemples de distributions des inter-arrivées et de service

1. **Distribution exponentielle :**
   La distribution exponentielle de paramètre μ modèlise la durée de vie d'un phénomène sans mémoire ou sans usure. Une variable aléatoire X est de loi exponentielle si sa

densité de probabilité est donnée par :

$$f(x) = \mu e^{-\mu x}, \quad x \geq 0.$$

et sa fonction de répartition :

$$\text{F(x)}=\begin{cases} 1 - e^{-\mu x}, & x \geq 0\,; \\ 0, & \text{Sinon.} \end{cases}$$

D'espérance : $\text{E(X)} = \frac{1}{\mu}$.

2. **Distribution Gamma ($\gamma(k, \mu)$) :**
   La distribution est définie comme somme de k variables aléatoires independantes de loi $\exp(\mu)$.
   La densité de probabilité d'une variable aléatoire X de loi gamma est donnée par :

$$f(x; k, \mu) = x^{k-1} \frac{\mu^k e^{-\mu x}}{\Gamma(k)} \quad , x \geq 0.$$

   avec $\Gamma(k) = (k-1)!$ est la fonction Gamma.
   D'espérance :

$$\mathbb{E}(X) = \frac{k}{\mu}.$$

   De transforme de Laplace :

$$\mathbb{E}(e^{-sX}) = B^*(s) = (\frac{s}{s+\mu}).$$

3. **Distribution déterministe (D) :**
   Les temps inter-arrivées des clients ou les temps de service sont constants et toujours les mêmes avec une moyenne de m. Sa transforme de Laplace est :

$$B^*(s) = \exp(-sm).$$

4. **Distribution Hyper-Exponentielle :**
   La distribution hyper-exponentielle de paramètres $(\mu_1, \mu_2, \mu_3, \ldots, \mu_k)$ est la distribution d'une combinaison en parallele de k variables aléatoires indépendantes suivant chacune une loi exponentielle de paramètre μ_i. Les paramètres de mélange sont

notés $(q_i, 1 \leq i \leq k)$ et verifient $\sum_{i=1}^{k} q_i = 1$. Sa densité de probabilité est :

$$f(x) = \sum_{i=1}^{k} q_i \mu_i e^{-\mu_i x}, \quad x \geq 0.$$

D'espérance :

$$\mathbb{E}(X) = \sum_{i=1}^{k} \frac{q_i}{\mu_i}.$$

Sa fonction de répartition est donnée par :

$$F(x) = 1 - \sum_{i=1}^{k} q_i e^{-\mu_i x}, \quad x \geq 0.$$

Sa transformée de Laplace est :

$$B^*(s) = \sum_{i=1}^{k} \frac{q_i \mu_i}{s + \mu_i}.$$

Analyse mathématique d'un système de files d'attente

L'étude mathématique d'un système de files d'attente (S.F.A) se fait généralement par l'introduction d'un processus stochastique défini de façon appropriée. On s'intéresse principalement au nombre de clients $X(t)$ se trouvant dans le système à l'instant t $(t \geq 0)$. En fonction des quantités qui définissent le système, on cherche à déterminer :

Le régime transitoire :

Les probablités d'etats $P_n(t) = P(X(t) = n)$, qui définissent le régime transitoire du processus $\{X(t),\ t \geq 0\}$. Il est évident que les fontions $P_n(t)$ dépendent de l'etat initial ou de la distribution initiale du processus.

Le régime stationnaire :

$$P_n = \lim_{t \to +\infty} P_n(t) = P(X = n); \quad n = 0, 1, \ldots$$

$(P_n)_{n \geq 0}$: Distribution stationnaire du processus $\{X(t),\ t \geq 0\}$

Formule de Little

La formule de Little est une relation très générale qui s'applique à une grande classe de systèmes à condition que ce dernier soit stable. A partir de la distribution stationnaire du processus $\{X(t), \quad t \geq 0\}$, on peut calculer les caractéristiques suivantes :

1- L : Nombre moyen de clients dans le système.

2- L_q :Nombre moyen de clients dans la file d'attente.

3- W : Temps moyen de séjour d'un client dans le système.

4- W_q : Temps moyen d'attente d'un client dans la file.

Ces caractéristiques sont liées par les relations suivantes :

a- $L = \lambda_e W$,

b- $L_q = \lambda_e W_q$,

c- $L = Lq + \frac{\lambda_e}{\mu}$,

d- $W = W_q + \frac{1}{\mu}$,

Les deux deux premières formule sont appelées "Formules de Little", où $\lambda_e < \lambda$ est le taux d'entrée dans le système.

Remarque 1.1. *Le calcul explicite du régime transitoire s'avère généralement pénible,voir impossible, pour la plus part des modèles donnés. On se contente donc de déterminer le régime stationnaire .*
Si la capacité du système et la population des usagers sont illimitées alors $\lambda_e = \lambda$.
Si la capacité du système et la population des usagers sont limitées alors $\lambda_e < \lambda$.

1.1.2 Les systèmes de files d'attente markoviens

1.1.2.1 Le système M/M/1

Le systeme d'attente M/M/1 est un système formé d'une file à capacité infinie, d'un unique serveur et la discipline de service est FIFO. Supposons que les instants d'arrivées des clients sont distribués selon un processus de Poisson de taux λ et que les temps de service sont indépendants suivant la loi exponentielle de taux μ. Le processus $(X(t))_{t\geq 0}$

décrivant le nombre de clients à l'instant t est un processus de naissance et de mort de taux de transitions :

$$\lambda_n = \lambda, \quad \forall \quad n \geq 0 \text{ et } \quad \mu_n = \mu, \quad \forall \quad n \geq 1.$$

Graphe de transition

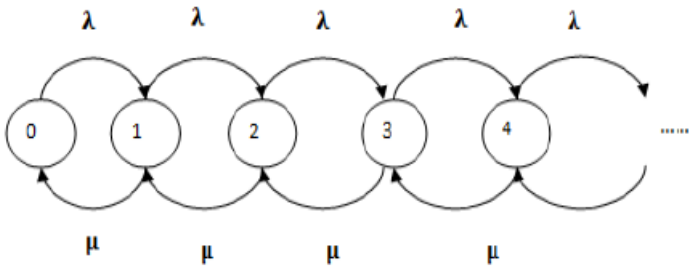

FIGURE 1.1: Graphe de transitions du modèle M/M/1

Régime stationnaire du système

On dit que le système est stable si et seulement si $\rho = \frac{\lambda}{\mu} < 1$. Supposons que $\rho < 1$, les équations de balance sont données par :

$$\begin{cases} \lambda p_0 = \mu p_1, \\ p_1 + \mu p_1 = \lambda p_0 + \mu p_2, \\ . \\ . \\ . \\ \lambda p_n + \mu p_n = \lambda p_{n-1} + \mu p_{n+1}, \qquad \forall\ n \geq 1. \end{cases}$$

D'où

$$\begin{cases} p_1 = \frac{\lambda}{\mu} p_0, \\ p_2 = \frac{\lambda}{\mu} p_1 = (\frac{\lambda}{\mu})^2 p_0, \\ p_3 = \frac{\lambda}{\mu} p_2 = (\frac{\lambda}{\mu})^3 p_0, \\ . \\ . \\ p_n = \frac{\lambda}{\mu} p_{n-1} = (\frac{\lambda}{\mu})^n p_0. \end{cases}$$

comme

$$\sum_{n=0}^{\infty} p_n = 1 \Rightarrow \sum_{n=0}^{\infty} (\frac{\lambda}{\mu})^n p_0 = 1$$

$$\Rightarrow p_0 \sum_{n=0}^{\infty} (\frac{\lambda}{\mu})^n = 1$$

$$\Rightarrow p_0 = \frac{1}{\sum_{n=0}^{\infty} (\frac{\lambda}{\mu})^n}$$

$$\Rightarrow p_0 = 1 - \frac{\lambda}{\mu}.$$

D'où :

$$p_n = (1 - \frac{\lambda}{\mu})(\frac{\lambda}{\mu})^n = (1 - \rho)\rho^n.$$

Mesures de performance

1. Le nombre moyen de clients dans le système :

$$L = \frac{\rho}{1 - \rho}.$$

2. Le nombre moyen de client dans la file d'attente :

$$L_q = L - \frac{\lambda}{\mu} = \frac{\rho^2}{1 - \rho}.$$

3. Le temps moyen de séjour d'un client dans le système :

$$W = \frac{1}{\mu - \lambda}.$$

1.1.2.2 Système M/M/S

Au lieu d'un seul serveur, ce système est comme le système M/M/1 mais avec S serveurs. Les arrivées et les départs sont modélisés par un processus de naissance et de mort

où :

$$\lambda_n = \lambda, \quad \forall n \geq 0, \quad \text{et} \quad \mu_n = \begin{cases} n\mu, & \text{si } n \leq s, \\ s\mu, & \text{sinon.} \end{cases}$$

Graphe de transitions

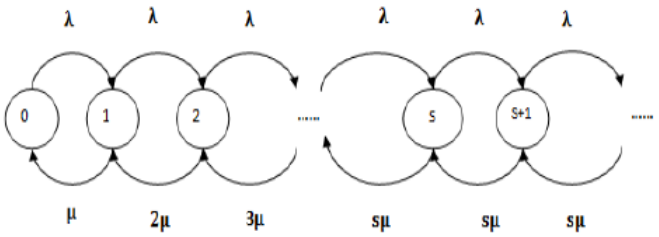

FIGURE 1.2: Graphe de transitions du modèle M/M/s

Distribution stationnaire du système

Supposos que le système est stable ($\rho = \frac{\lambda}{s\mu} < 1$).
Les equations de balance sont :

Cas où $n \leq s$

$$\begin{cases} \lambda p_0 = \mu p_1, \\ \lambda p_1 + \mu p_1 = \lambda p_0 + \mu p_2, \\ . \quad . \\ . \quad . \\ . \quad . \\ \lambda p_{s-1} + (s-1)\mu p_{s-1} = \lambda p_{s-2} + s\mu p_s. \end{cases}$$

D'où

$$\begin{cases} p_1 = \frac{\lambda}{\mu} p_0, \\ p_2 = \frac{\lambda}{2\mu} p_1 = \frac{1}{2}(\frac{\lambda}{\mu})^2 p_0, \\ p_3 = \frac{\lambda}{3\mu} p_2 = \frac{1}{2.3}(\frac{\lambda}{\mu})^3 p_0, \\ \quad . \quad . \\ \quad . \quad . \\ p_s = \frac{\lambda}{s\mu} p_{s-1} = \frac{1}{s!}(\frac{\lambda}{\mu})^s p_0. \end{cases}$$

Cas où $n > s$:

$$\begin{cases} \lambda p_s + s\mu p_{(s-1)} &= \lambda p_{(s-1)} + s\mu p_{(s+1)}, \\ \lambda p_{s+1} + s\mu p_{s+1} &= \lambda p_s + s\mu p_{s+2}, \\ . & . \\ . & . \\ . & . \end{cases}$$

$$\begin{cases} p_s &= \frac{1}{s!}(\frac{\lambda}{\mu})^s p_0, \\ p_{(s+1)} &= \frac{1}{s!s}(\frac{\lambda}{\mu})^{s+1} p_0, \\ p_{(s+2)} &= \frac{1}{s!s^2}(\frac{\lambda}{\mu})^{s+2} p_0, \\ & . \\ & . \\ p_n &= \frac{\lambda}{s!s^{n-s}\mu}(\frac{\lambda}{\mu})^n p_0. \end{cases}$$

Donc,

$$p_n = \begin{cases} \frac{1}{n!}(\frac{\lambda}{\mu})^n p_0 & si \quad n \leq s, \\ \frac{1}{s!s^{n-s}}(\frac{\lambda}{\mu})^n p_0 & si \quad n > s. \end{cases}$$

Avec,

$$p_0 = \left[\sum_{n=0}^{s} \frac{(\frac{\lambda}{\mu})^n}{n!} + \frac{(\frac{\lambda}{s\mu})^{s+1}}{s!(s - \frac{\lambda}{\mu})} \right]^{-1}.$$

Caractéristiques usuelles du système M/M/s

1- $L_q = E(X_q) = \sum_{n=s+1}^{\infty} (n-s)p_n$

$= \sum_{n=s+1}^{\infty} (n-s)\rho^{n-s} P_s = P_s \sum_{n=s+1}^{\infty} (n-s)\rho^{n-s}$

On pose $n - s = j \Rightarrow L_q = P_s \sum_{j=1}^{\infty} j\rho^j = P_s \rho \sum_{j=1}^{\infty} j\rho^{j-1}$

$L_q = P_s \rho \frac{1}{(1-\rho)^2}$.

2- $L = L_q + \frac{\lambda}{\mu} = s\rho + \frac{\rho P_s}{(1-\rho)^2}$.

A l'aide des formules de Little, on trouve :

$$W_q = \frac{P_s}{s\mu(1-\rho)^2} \qquad et \qquad W = \frac{P_s}{s\mu(1-\rho)^2} + \frac{1}{\mu}.$$

1.1.3 Le modèle d'attente semi Markovien M/G/1

Le modèle M/G/1 est un système à un seul serveur où les arrivées sont poissoniennes de taux $\lambda > 0$, c'est-à-dire, le temps entre deux inter-arrivées successives suit une loi exponentielle de moyenne $\frac{1}{\lambda}$. Si le serveur est libre, le client sera pris en charge immédiatement. Dans le cas contraire, il rejoint la file d'attente (de capacité illimitée et discipline FIFO), les durées de service Y sont des variables aléatoires indépendantes et identiquement distribuées de loi générale dont la fonction de répartition $F(y)$, d'espérance mathématique $\mathbb{E}(Y) = \frac{1}{\mu}$.
Le processus $\{X(t), t \geq 0\}$ décrivant l'état du système n'est pas markovien. Pour l'analyse de ce système on procède par la méthode de la chaîne de Markov induite.

1.1.3.1 Chaîne de Markov induite

Considérons le processus $(X(t))_{t>0}$ aux instants t_1, $t_2 \ldots$ où les clients quittent le système aprés être servi. Soit $\{X_n = X(t_n), n \in \mathbb{N}\}$ le processus stochastique à temps discret où t_n est l'instant de départ du $n^{ème}$ client. Cette suite forme une chaîne de Markov. Soit la suite de variables aléatoires (A_n) indépendantes et identiquement distribuées telle que A_n est le nombre de clients arrivant pendant $n^{ème}$ service.

$$P(A_n = k) = a_k = \int_0^\infty \frac{exp(-\lambda t)(\lambda t)^k}{k!} F(t)dt,$$

avec $a_k > 0 \quad (k = 0, 1, 2, \ldots)$,
alors

$$X_{n+1} = \begin{cases} X_n - 1 + A_{n+1}, & si \quad X_n > 0, \\ A_{n+1}, & si \quad X_n = 0. \end{cases}$$

X_{n+1} peut s'écrire de la manière suivante :

$$X_{n+1} = X_n - \delta_n + A_{n+1},$$

avec :

$$\delta_n = \begin{cases} 1, \text{ si } X_n > 0, \\ 0, \quad si \quad X_n = 0 \ . \end{cases}$$

X_{n+1} ne dépend que de X_n et de A_{n+1} et non pas des valeurs prises par X_{n-1}, $X_{n-2} \ldots$
La suite de variables aléatoires $\{X_n, n \geq 1\}$ est une chaîne de Markov induite du processus $\{X_t, t \geq 0\}$. Ses probabilités de transition se calculent par :

$$\begin{cases} p_{0j} = a_j, \quad si \quad j \geq 0, \\ p_{ij} = a_{j-i+1}, \quad si \quad 1 \leq i \leq j+1, \\ p_{ij} = 0, \qquad sinon. \end{cases}$$

Ainsi la matrice de transition P est donnée par :

$$P = \begin{pmatrix} a_0 & a_1 & a_2 & a_3 & ... & ... \\ a_0 & a_1 & a_2 & a_3 & ... & ... \\ 0 & a_1 & a_2 & a_3 & ... & ... \\ 0 & 0 & a_0 & a_1 & a_2 & ... \\ ... & ... & ... & ... & ... & ... \end{pmatrix}$$

La chaîne est une chaîne de Markov irréductible car on peut passer de chaque état vers n'importe quel autre état.
La fonction génératrice de distribution stationnaire $\pi(z)$ de la chaîne existe si :

$$\rho = \frac{\lambda}{\mu} < 1,$$

et donné par :

$$\pi(z) = \sum_{n=0}^{\infty} z^n \pi_n = \frac{(1-\rho)B^*(\lambda - \lambda z)(1-z)}{B^*(\lambda - \lambda z) - z}.$$

Avec $B^*(z) = \mathbb{E}(e^{-zY})$ est la transformée de laplace de la durée de service.

Mesures de performence

- Nombre moyen de clients dans le système :

$$L = \rho + \frac{\rho^2 + \lambda^2 Var(Y)}{2(1-\rho)}.$$

- Nombre moyen de clients dans la file d'attente L_q :

$$L_q = L - \rho \rightarrow L_q = \frac{\rho^2 + \lambda^2 V(Y)}{2(1-\rho)}.$$

- Temps moyen de séjour d'un client dans le système :

$$W = \frac{L}{\lambda} = \frac{1}{\mu} + \left(\frac{\lambda(V(Y) + \frac{1}{\mu^2})}{2(1-\rho)} \right).$$

- Temps moyen d'attente d'un client :

$$W_q = \frac{L_q}{\lambda} = \frac{\lambda(V(Y) + \frac{1}{\mu^2})}{2(1-\rho)}.$$

1.2 Files d'attente avec impatience

Pour modèliser un système de files d'attente avec impatience, on doit ajouter une contrainte au système en spécifiant que les clients sont perdus si le temps qu'ils passent dans le système est plus grand qu'un délai qui leur est alloué. Le modèle que nous considèrons est donc une file d'attente où les clients ont un dèlai de patience pour être servis, au delà du quel ils sortent du système sans être servi.

Considérons le $n^{ème}$ client C_n entrant dans la file d'attente qui est affecté du délai D_n, qui est une variable aléatoire strictement positive. S'il n'a pas pu atteindre le serveur à la fin de sa patience (i.e à l'instant $T_n + D_n$), il est éliminé du système. Les délais seront donc considérés éliminatoires et ce jusqu'au début de service.
Par ailleurs, la suite D_n , $n \in N^*$, sera toujours supposée stationnaire de loi générique D. On supposera de plus que :

$$E(D) < 1.$$

Et, on notera que

$$\alpha = (E(D))^{-1}.$$

1.2.1 Mesures de performance

Définissons à présent les différentes variables aléatoire caractérisant notre problème et permettant d'évaluer ses performances :

1. Le nombre de clients moyen L dans le système (file+serveur) avec $L = E(N)$
2. le temps d'attente W_n proposé au client C_n est le temps qu'aura à attendre ce client avant la fin de son service.
3. Le temps de séjour τ_n est le temps que passe effectivement le client C_n dans le système (file+serveur) :

$$\tau_n = (W_n + s_n)\mathbf{1}_{\{\mathbf{W_n < D_n}\}} + \mathbf{D_n}\mathbf{1}_{\{\mathbf{W_n \geq D_n}\}}.$$

4. Le temps d'attente τ_n^q est le temps que passe effectivement le client C_n dans la file :

$$\tau_n^q = (W_n))\mathbf{1}_{\{\mathbf{W_n < D_n}\}} + \mathbf{D_n})\mathbf{1}_{\{\mathbf{W_n \geq D_n}\}}.$$

Pour toute file d'attente à perte, notons π_n la probabilité de perte du client C_n. Ainsi, π_n est la probabilité qu'un client C_n se voit proposé un temps d'attente supérieur à son délai de patience. Autrement dit :

$$\pi_n = P(W_n > D_n)$$

1.2.2 Modèle M/M/1

1.2.2.1 Cas où l'impatience est markovienne (M/M/1+M)

Considérons la file M/M/1+M où les arrivées sont markoviens de taux λ, le temps de service est exponentiel de taux μ et le délai de patience est exponentiel de taux α. Le processus $(X(t))_{t\geq 0}$ est un processus de naissance et de mort, dont les taux de transitions sont donnés dans le graphe ci-dessous :

$$\lambda_n = \lambda, \quad \forall n \geq 0. \quad et \quad \mu_n = \mu + n\alpha, \quad \forall n \geq 1.$$

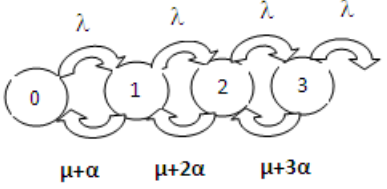

FIGURE 1.3: Graphe de transition du modèle M/M/1 avec impatience

Posons P_n la probabilité associée à l'état n. Ainsi, et par conservation de flux nous obtenons la formule de récurrence suivante :

$$\lambda P_n = [\mu + (n+1)\alpha]P_{n+1}.$$

Et, on montrera aisément par récurrence que,

$$P_n = \left(\prod_{k=1}^{n} \frac{\lambda}{\mu + k\alpha}\right) P_0.$$

Par souci de simplification, posons,

$$\rho_k = \frac{\lambda}{\mu + k\alpha}, \quad \forall k \geq 1,$$

ainsi,

$$P_n = (\prod_{k=1}^{n} \rho_k) P_0.$$

D'autre part, et comme les probabilités somment à 1, alors, on déduit que,

$$P_0 = \frac{1}{1 + \sum\limits_{n=1}^{\infty} \prod\limits_{k=1}^{n} \rho_k}.$$

P_0 existe car $\sum\limits_{n=1}^{\infty} \prod\limits_{k=1}^{n} (\frac{\lambda}{\mu+k\alpha})$ converge.
En effet,

$$\begin{aligned} \mu + k\alpha > k\alpha \quad &\Rightarrow \quad \frac{1}{\mu + k\alpha} \leq \frac{1}{k\alpha} \\ &\Rightarrow \quad \prod_{k=1}^{n} \frac{\lambda}{\mu + k\alpha} \leq (\frac{\lambda}{\alpha})^n \prod_{k=1}^{n} \frac{1}{k} = (\frac{\lambda}{\alpha})^n \frac{1}{n!} \\ &\Rightarrow \quad \sum_{k=1}^{\infty} \prod_{k=1}^{n} \frac{\lambda}{\mu + k\alpha} \leq \sum_{k=1}^{\infty} (\frac{\lambda}{\alpha})^n \frac{1}{n!} = e^{\frac{\lambda}{\alpha}} < \infty. \end{aligned}$$

Enfin et par souci de simplification nous garderons toujours la constatnte P_0 dans l'expression de P_n.

1.2.2.2 Calcul des mesures de performance

En gardant les simplifications déjà faites, nous trouverons les résultats suivants :

$$L = \sum_{n=1}^{\infty} nP_n = n \sum_{n=1}^{\infty} P_0 \prod_{k=1}^{n} \rho_k.$$

En appliquant la loi de Little $L = \lambda W$, on trouve :

$$\begin{aligned} W &= \frac{L}{\lambda} \\ &= \frac{P_0}{\lambda} \sum_{n=1}^{\infty} n \prod_{k=1}^{n} \rho_k. \end{aligned}$$

Probabilité de perte d'un client à l'etat stationnaire

$$\Pi = P(W > D) = \frac{\alpha}{\lambda} \sum_{n=1}^{\infty} P_0 \prod_{k=1}^{n} \rho_k = \frac{\alpha}{\lambda}(1 - P_0).$$

1.3 Les systèmes de files d'attente avec rappels

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. Les systèmes de file d'attente développés ces dernières années tentent de prendre en considération des phénomènes de répétition de demandes de service, et ceci aprés une durée du temps aléatoire. Un tel système est connu comme « système de files d'attente avec rappels ou répétition d'appels»(Retrial Queues dans la terminologie anglo-saxonne). Les systèmes de files d'attentes avec rappels sont caractérisés par la propriété qu'un client qui trouve à son arrivée tous les serveurs occupés quitte l'espace de service et rappelle ultérieurement à des instants aléatoires. Entre deux rappels successifs, le client est dit "en orbite". Ces systèmes de files d'attente sont largement utilisés dans la modélisation des systèmes informatiques et des réseaux de télécommunications ([8],[10],[22]), par exemple supposons qu'un client essaie d'appeler un magasin local. Si le

client reçoit un signal occupé, il pose le téléphone et réessaie plus tard, tout en espérant que le serveur deviendera libre avant le prochain appel. Cependant, les autres clients peuvent également appeler pendant cette période, de sorte que le premier client peut toujours recevoir un signal occupé au deuxième essai. Cet exemple motive le modèle conceptuel de base d'une file d'attente avec rappels, étudiés dans [7, 9, 22].

1.3.1 Modèle général d'un système de file d'attente avec rappels

Afin d'identifier un système de files d'attente avec rappels, généralement on a besoin de la nature stochastique du processus des arrivées, la distribution du temps de service, le nombre de serveurs qui composent l'espace de service, la capacité et la discipline d'attente ainsi que la spécification concernant le processus de répétition d'appels. Ce système de files d'attente avec la notation de Kendall $T/Y/s/N/O/s$ est composé de $s \geq 1$ serveurs et $(N-s)$ places d'attente. Les arrivées dans le système suivent un processus aléatoire avec une loi de probabilité donnée. A l'arrivée d'un client, s'il y a une position d'attente libre, le client rejoint la file d'attente. Dans le cas contraire, il quitte l'espace de service temporairement pour tenter sa chance après une durée de temps aléatoire avec une probabilté H_k, ou il quitte définitivement le système avec une probabilité $1 - H_k$. Les arrivées de l'éxtérieur dans le système forment un flux d'appels primaires contrairement aux clients en orbite qui présentent des appels secondaires. La capacité O de l'orbite peut être finie ou infinie. Dans le cas où O est finie et si l'orbite est pleine, le client quitte le système pour toujours. Le schéma général d'un système avec rappels est donné dans la Figure 1.4.

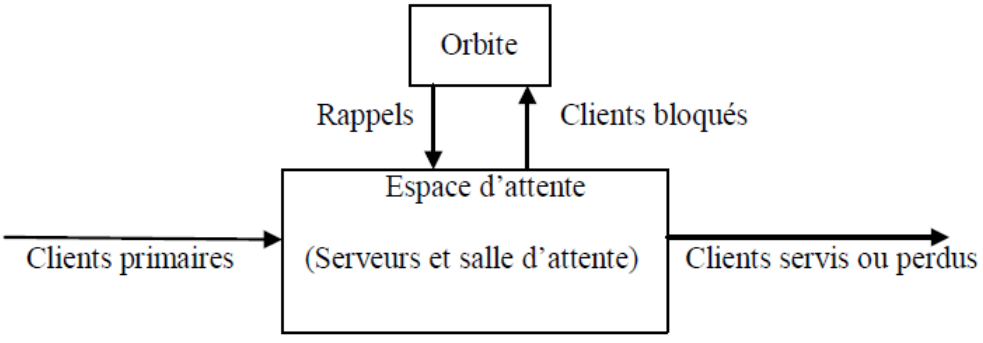

FIGURE 1.4: Schéma général d'un systeme avec rappels

1.3.2 Le modèle M/M/1 avec rappels

1.3.2.1 Description du modèle M/M/1 avec rappels

Le système M/M/1 avec rappels est un système de files d'attentes avec un seul serveur. Les clients primaires arrivent selon un processus de poisson de taux $\lambda > 0$. Les durées de service suivent une loi exponentielle de fonction de répartition
$F(x) = 1 - e^{-\mu x},\ x \geq 0$ de moyenne finie $\frac{1}{\mu}$ et les temps entre deux rappels consécutifs sont également exponentiels de paramètre $\theta > 0$ de fonction de répartition
$B(x) = 1 - e^{-\theta x},\ x \geq 0$. Nous admettons que les durées de service, les durées entre deux rappels consécutifs ainsi que entre deux arrivées primaires succéssives sont mutuellement indépendantes. L'état du système peut être décrit par le processus

$$X(t) = \{C(t), N(t), t \geq 0\}.$$

Où $C(t)$ est égale à 0 ou 1 selon que le serveur est libre ou non, $N(t)$ est le nombre de clients dans l'orbite à l'instant t. Supposons que le régime stationnaire existe ($\rho = \frac{\lambda}{\mu} < 1$) le processus $X(t) = \{(C(t), N(t)), t \geq 0\}$ est de Markov d'éspace d'états $S = \{0, 1\} \times \mathbb{N}$.

1.3.2.2 Graphe de transitions

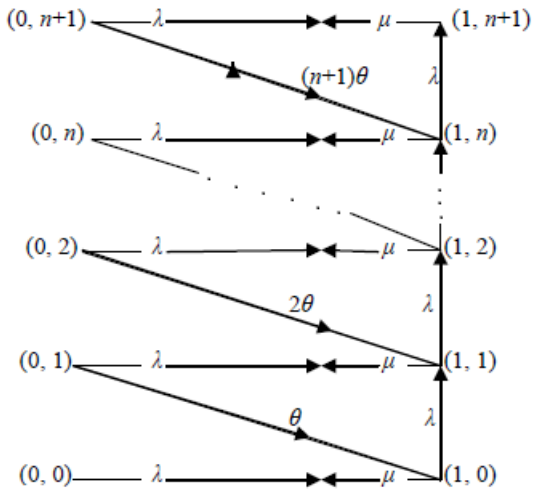

FIGURE 1.5: Graphe de transition du modèle M/M/1 avec rappels

Les équations d'équilibre sont :

$$(\lambda + n\theta)p_{0n} = \mu p_{1n}, \qquad n = 0. \tag{1.3.1}$$

$$(\lambda+\mu)p_{1n} = \lambda p_{0n} + (n+1)\theta p_{0,n+1} + \lambda p_{1,n-1},\ \ \forall n \geq 1. \quad (1.3.2)$$

Avec, $P_{in} = \lim_{t\to\infty} \mathbb{P}((C(t)=i, N(t)=n))$, $i=0,1$ et $n \geq 0$, représentent la distribution stationnaire conjointe de l'état du serveur et du nombre de clients en orbite. Introduisons les fonctions génératrices suivantes :

$$P_0(z) = \sum_{n=0}^{\infty} P_{0n} z^n,$$

$$P_1(z) = \sum_{n=0}^{\infty} P_{1n} z^n.$$

A l'aide de ces fonctions et à partir des équations (1.3.1) et (1.3.2), on obtient(voir [3]) :

$$P_0(z) \ = \ (1-\rho)\left(\frac{1-\rho}{1-z\rho}\right)^{\frac{\lambda}{\theta}}, \quad (1.3.3)$$

$$P_1(z) \ = \ \rho(\frac{1-\rho}{1-z\rho})^{\frac{\lambda}{\theta}+1}. \quad (1.3.4)$$

Les transformées inverses des (1.3.3) et (1.3.4) nous donnent les formules analytiques explicites

$$p_{0n} = \frac{\rho}{n!\theta^n} \prod_{k=0}^{n-1} (1+k\theta)(1-\rho)^{\frac{\lambda}{\theta}+1},$$

$$p_{1n} = \frac{\rho^{n+1}}{n!\theta^n} \prod_{k=1}^{n} (\lambda+k\theta)(1-\rho)^{\frac{\lambda}{\theta}+1}.$$

1.3.3 Le Modèle M/G/1 Avec rappels

Description du modèle

Les clients arrivent dans le système selon un processus de Poisson de taux $\lambda > 0$: $P(\tau_n^e \leq x) = 1 - e^{-\lambda x}$. Le service des clients est assuré par un seul serveur. La durée de service est de loi générale $P(\tau_n^s \leq x) = F(x)$ et de transformée de laplace $B^*(s)$. Soient les moments $\beta_k = (-1)^k B^{*(k)}(0)$, l'intensité du trafic $\rho = \lambda\beta_1$ et $\mu = \frac{1}{\beta_1}$. La

durée entre deux rappels successifs d'une même source secondaire est exponentiellement distribuée de paramètre $\theta > 0$: $P(\tau_n^r \leq x) = 1 - e^{-\theta x}$.
Le système évolue de la manière suivante : On suppose que le $(n-1)^{ème}$ client termine son service à l'instant ξ_{n-1} (les clients sont numérotés dans l'ordre de service) et le serveur devient libre ; même s'il y a des clients dans le système, ils ne peuvent pas occuper le serveur immédiatement à cause de leur ignorance de l'état de ce dernier. Donc il existe un intervalle de temps R_n durant lequel le serveur reste libre avant que le $n^{ème}$ client n'entre en service. A l'instant $\tau_n = \xi_n + R_n$ le $n^{ème}$ client débute son service durant un temps τ_n^s . Les rappels qui arrivent durant ce temps de service n'influent pas sur ce processus. A l'instant $\tau_n = \xi_n + \tau_n^s$ le $n^{ème}$ client achève son service, le serveur devient libre et ainsi de suite.

Distribution stationnaire de l'état du système

Le premier résultat sur le système M/G/1 avec rappels a été obtenu par Keilson et al. (1968) [20], en utilisant la méthode de la variable supplémentaire. Ils ont obtenu les probabilités d'états et les fonctions génératrices de nombre de clients dans le système. L'état du système peut-être décrit par le processus :

$$X(t)=\begin{cases} N_0(t), & si \quad C(t)=0, \\ (C(t); N_0(t); \xi(t)) & si \quad C(t)=1. \end{cases}$$

où $\xi(t)$ est une variable aléatoire supplémentaire à valeurs dans $\mathbb{R}^+$, et désignant la durée de service écoulé à la date t, si $C(t) = 1$ et $N_0(t)$ représente le nombre de clients dans l'orbite.
Notons

$$P_{0j}(t) = P(C(t)) = 0, N_0(t) = j),$$
$$\mathbb{P}_{1j}(t, x) = \mathbb{P}(C(t) = 1, N_0(t) = j, x < \xi(t) < x + dx), j > 0,$$

Si $\frac{\lambda}{\mu} < 1$, le système est stable. La fonction génératrice du nombre de clients dans le système est donnée par :

$$Q(z) = \frac{(1-\rho)(1-z)A(z)}{A(z)-z}\frac{\phi(z)}{\phi(1)},$$

où

$$\phi(z) = exp\{\frac{\lambda}{\theta}\int_1^z \frac{1-\beta(\lambda-\lambda x)}{\beta(\lambda-\lambda x)-x}d\}.$$

On aura alors,

$$Q(z) = \frac{(1-\rho)(1-z)\beta(\lambda-\lambda z)}{\beta(\lambda-\lambda z)-z} exp\{\frac{\lambda}{\theta}\int_1^z \frac{1-\beta(\lambda-\lambda x)}{\beta(\lambda-\lambda x)-x)}dx\}.$$

Cette formule est appelée "Decomposition Stochastique", signifie que le nombre de clients dans un système M/G/1 avec rappel s'écrit comme somme de la variable qui représente le nombre de clients dans le système M/G/1 classique et l'autre variable aléatoire positive de fonction génératrice $\frac{\phi(z)}{\phi(1)}$, qui représente le nombre de clients dans l'orbite lorsque le serveur est libre.

Graphe de transitions

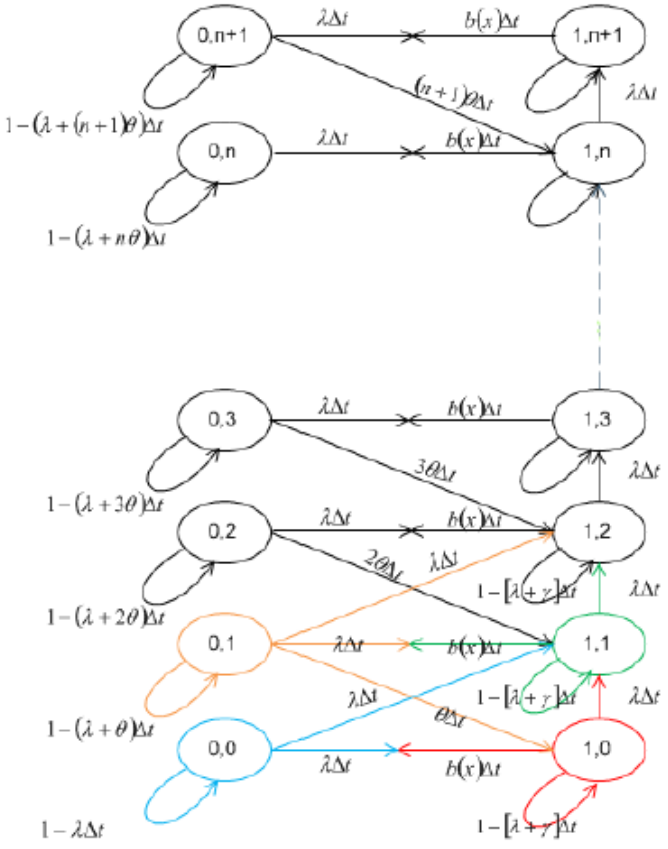

FIGURE 1.6: Graphe des transitions du modèle M/G/1 avec rappels

Chaîne de Markov induite du système

[13] Considérons le processus $X(t) = \{(C(t); N_0(t)); t \geq 0\}$ non markovien
où :
$C(t)$: la variable aléatoire indiquant l'état du serveur à l'instant t

$$C(t) = \begin{cases} 0, & \text{si le serveur est libre ;} \\ 1, & \text{si le serveur est occupé.} \end{cases}$$

$N_0(t)$: le nombre de clients en orbite à l'instant t. $\xi(t)$ représente le temps de service écoulé du client en service si $C(t) = 1$.
Ce processus possède une chaîne de Markov induite ($q_n = N_0(\xi_n)$) représentant le nombre de clients en orbite aprés le $n^{ème}$ départ. Cette chaîne a été décrite pour la première fois par Choo et Conolly (1979).

Soit V_n, le nombre de clients qui arrivent dans le système durant le service du $n^{ème}$ client dont la distribution est donnée par :

$$\mathbb{P}(V_n = i) = a_i = \int_0^\infty exp(\lambda x)\frac{(\lambda x)^i}{i!}dB(x),$$

avec $a_i > 0$, $i \geq 0$.

Remarque :
Si $\lim\limits_{n\to\infty} V_n = V$ et $\mathbb{E}(V) = \rho$,
alors :

$$A(z) = \beta(\lambda - \lambda z) = \int_0^\infty e^{-x(\lambda-\lambda z)}dB(x) = \sum_{i\geq 0} a_i z^i.$$

Soit δ_{q_n} une variable aléatoire de Bernoulli dépendante de q_n définie par :

$$\delta_{Xn} = \begin{cases} 1, & \text{si n(ième) client provient de l'orbite ;} \\ 0, & \text{sinon.} \end{cases}$$

De distribution conditionnelle

$$P(\delta_{X_n} = 1|X_n = i)=\frac{i\theta}{\lambda+i\theta}.$$
$$P(\delta_{X_n} = 0|X_n = i)=\frac{\lambda}{\lambda+i\theta}.$$

L'équation fondamentale de la chaîne de Markov induite est :

$$X_{n+1}=X_n - \delta_{X_n} + V_{n+1}.$$

Les probabiités de transition de la chaîne en une étape sont données par :

$$P_{ij} = \mathbb{P}(X_{n+1} = j|X_n = i)=\frac{i\theta}{\lambda+i\theta}a_{j-i+1} + \frac{\lambda}{\lambda+i\theta}a_{j-i},$$

en effet,

$$\begin{aligned}
P_{ij} &= \mathbb{P}(X_{n+1} = j|X_n = i) = \mathbb{P}(X_n - \delta_{X_n} + V_{n+1} = j|X_n = i) \\
&= \mathbb{P}(V_{n+1} = j - i + \delta_{X_n}|X_n = i) \\
&= \mathbb{P}(V_{n+1} = j - i|X_n = i\delta_{X_n} = 1)P(\delta_{X_n} = 1|X_n = i) \\
&+ \mathbb{P}(V_{n+1} = j - i|X_n = i, \delta_{X_n} = 0)\mathbb{P}(\delta_{X_n} = 0|X_n = i) \\
&= \mathbb{P}(V_{n+1} = j - i)(\mathbb{P}(\delta_{X_n} = 0|X_n = i) + P(V_{n+1} = j - i + 1)(P(\delta_{X_n} = 1|X_n = i) \\
&= \frac{i\theta}{\lambda + i\theta}a_{j-i+1} + \frac{\lambda}{\lambda + i\theta}a_{j-i}.
\end{aligned}$$

Distribution stationnaire de la chaîne de Markov induite

Si $\rho < 1$ la chaîne de Markov induite est stationnaire et possède une fonction génératrice notée $Q(z)$:

$$Q(z) = \sum_{n=0}^{\infty} \pi_n z^n = \frac{(1-\rho)(1-z)\beta(\lambda-\lambda z)}{\beta(\lambda-\lambda z)-z} exp\{\frac{\lambda}{\theta}\int_1^z \frac{1-\beta(\lambda-\lambda x)}{\beta(\lambda-\lambda x)-x}dx\}.$$

Mesures de performence

Les caractéristiques du Modèle M/G/1 avec rappels sont :

a) Le nombre moyen des clients dans le système,

$$L = Q^{'}(1) = \rho + \frac{\lambda^2\beta_2}{2(1-\rho)} + \frac{\lambda\rho}{\theta(1-\rho)}.$$

b) Nombre moyen des clients en orbite

$$L_0 = L - \rho = \frac{\lambda^2 \beta_2}{2(1-\rho)} + \frac{\lambda \rho}{\theta(1-\rho)}.$$

c) Temps moyen d'attente d'un client :

$$W = \frac{L_0}{\lambda} = \frac{\lambda \beta_2}{2(1-\rho)} + \frac{\rho}{\theta(1-\rho)}.$$

d) Nombre moyen de rappels par client (d'après la formule de Little) :

$$R = \theta W = \frac{\lambda \beta_2}{2(1-\rho)} + \frac{\rho}{(1-\rho)}.$$

Conclusion

Nous avons rappelé et présenté les notions et techniques de base sur les systèmes de files d'attente classiques et avec rappels. Dans ces derniers, les clients peuvent quitter le système sans être servis, après plusieurs tentatives échouées. Il s'agit des clients impatients dans l'orbite. Cette notion d'impatience sera traitée dans le chapitre suivant pour le cas de $M/M/1$ avec rappels et clients impatients (voir Chapitre 2).

CHAPITRE 2

LE MODÈLE M/M/1 AVEC RAPPELS ET IMPATIENCE EN ORBITE

Introduction

L'impatience est un concept très naturel et important dans les modèles de files d'attente. Il existe plusieurs situations (centres d'appels) dans lesquelles les clients peuvent devenir impatients s'ils ne reçoivent pas le service assez rapidement. Les clients en attente dans l'orbite perdent patience à cause du long délai d'attente pour le service. Si le client secondaire trouve le serveur libre avant l'expiration du délai, il a la possibilité d'être servi et quitte le système une fois le service terminé. Si la minuterie expire avant qu'il ne puisse être servi, il quitte le système sans être servi.
Dans la monographie de Falin [16], l'impatience a été modélisée en utilisant la fonction de persévérance H. Le but de ce chapitre est de décortiquer l'article de Suganthi et al. [30] qui consiste à analyser le système M/M/1 avec rappels et impatience en considérant le délai d'impatience dans l'orbite est exponentiellement distribué.

2.1 Description mathématique du modèle

Considérons un système de files d'attente M/M/1 avec rappels et des clients impatient dans l'orbite, ces derniers arrivent de l'extérieur (clients primaires) selon un processus de Poisson de taux $\lambda > 0$. Un client primaire reçoit le service immédiatement si

le serveur est inoccupé, sinon il rejoint l'orbite. Un client en orbite essaie de manières répétées d'entrer dans le service avec un temps de rappels distribué de manière exponentielle avec un taux $\theta > 0$. Le temps de service suit une distribution exponentielle de moyenne $\frac{1}{\mu}$. En rejoignant l'orbite les clients secondaires, active un délai d'attente distribué de manière exponentielle avec un taux α. Si le serveur n'est pas disponible pour le client avant l'expiration du délai, le client quitte l'orbite, c'est-à-dire ne retourne jamais dans le système.

Soit le processus $\{(C(t), N(t)), \quad t \geq 0\}$ décrivant l'état du système à l'instant t, où $C(t) = 1$ ou 0 selon que le serveur est actif ou inactif et $N(t)$ est le nombre de clients dans l'orbite à l'instant t.

Notons par :

$p_{0n}(t)$= P [le serveur est inactif et il y a n clients dans l'orbite à l'instant t]

$p_{1n}(t)$= P [le serveur est occupé et il y a n clients dans l'orbite à l'instant t].

2.1.1 Distribution stationnaire

Graphe de transition

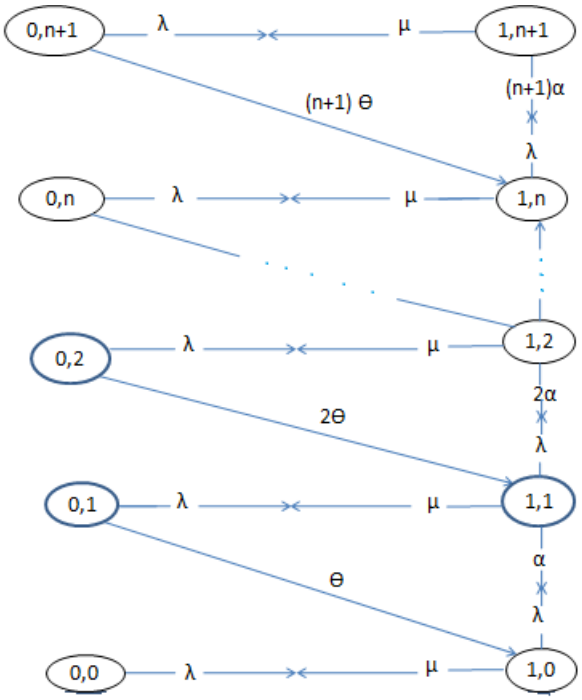

FIGURE 2.1: Graphe de transition du modèle M/M/1 avec rappels et impatience

Les équations de Chapmann Kolmogorov de ce modèle sont données par :

$$p'_{0n}(t) = -(\lambda + n\theta)p_{0n}(t) + \mu p_{1n}(t), \quad n \geq 0, \tag{2.1.1}$$

$$\begin{aligned} p'_{1n}(t) &= -(\lambda + \mu + n\alpha)p_{1n}(t) + \lambda p_{1n-1}(t) + \lambda p_{0n}(t) \\ &\quad +(n+1)\theta p_{0n+1}(t) + (n+1)\alpha p_{1n+1}(t), \quad n \geq 1, \end{aligned} \tag{2.1.2}$$

$$p'_{10}(t) = -(\lambda + \mu)p_{10}(t) + \lambda p_{00}(t) + \theta p_{01}(t) + \alpha p_{11}(t), \quad n = 0. \tag{2.1.3}$$

A l'état d'équilibre (stationnaire), on a : $\lim_{t\to\infty} P_{0n}(t) = P_{0n}$ et $\lim_{t\to\infty} P_{1n}(t) = P_{1n}$.
D'où les équations de balance

$$(\lambda + n\theta)p_{0n} = \mu p_{1n}, \quad n \geq 0, \tag{2.1.4}$$

$$\begin{aligned} (\lambda + \mu + n\alpha)p_{1n} &= \lambda p_{1n-1} + \lambda p_{0n} + (n+1)\theta p_{0n+1} \\ &\quad +(n+1)\alpha p_{1n+1}, \quad n \geq 1, \end{aligned} \tag{2.1.5}$$

$$(\lambda+\mu)p_{10} = \lambda p_{00} + \theta p_{01} + \alpha p_{11}, \quad n=0. \tag{2.1.6}$$

Pour résoudre le système des équations, nous suivons l'approche donnée dans Falin et Templeton (1997) [16]. À partir de (2.1.4)

$$p_{1n} = \frac{[\lambda+n\theta]}{\mu} p_{0n},$$

$$p_{1n+1} = \frac{[\lambda+(n+1)\theta]}{\mu} p_{0n+1},$$

$$p_{1n-1} = \frac{[\lambda+(n-1)\theta]}{\mu} p_{0n-1}.$$

En remplaçant dans (2.1.5), nous obtenons :

$$\begin{aligned}(n+1)[\mu\theta+\alpha[\lambda+(n+1)\theta]]p_{0n+1} - \lambda(\lambda+n\theta)p_{0n} &= n[\mu\theta+\alpha(\lambda+n\theta)]p_{0n} \\ &- \lambda(\lambda+(n-1)\theta)p_{0n-1} \end{aligned} \tag{2.1.7}$$

En posant :

$$x_n = n[\mu\theta+\alpha(\lambda+n\theta)] \quad \text{et} \quad y_n = \lambda(\lambda+n\theta).$$

On aura :

$$x_{n+1}p_{0n+1} - y_n p_{0n} = x_n p_{0n} - y_{n-1}p_{0n-1}.$$

Soit la constante C telle que :

$$C = x_{n+1}p_{0n+1} - y_n p_{0n}, \quad \text{pour} \;\; n \geq 0. \tag{2.1.8}$$

La constante C peut être déterminée à partir de (2.1.6) comme suit :
Résoudre les equations de p_{10} et p_{11} dans (2.1.4) puis substituer dans (2.1.6), nous obtenons :

$$[\mu\theta+\alpha(\lambda+\theta)]p_{01} - \lambda^2 p_{00} = 0,$$

qui peut être écrite comme :

$$x_1 p_{01} - y_0 p_{00} = 0,$$

qui est identique à l'équation (2.1.8) en remplaçant $n = 0$. Donc $C = 0$.
En général, de l'équation (2.1.8), on aura :

$$p_{0n+1} = \frac{y_n}{x_{n+1}} p_{0n}, \quad \text{pour} \quad n \geq 0,$$

$$\begin{aligned}
\text{Pour } n \geq 1 \quad , p_{0n} &= \prod_{i=0}^{n-1} \frac{y_i}{x_{i+1}} p_{00} \\
&= \prod_{i=0}^{n-1} \frac{\lambda(\lambda + i\theta)}{(i+1)[\mu\theta + \alpha(\lambda + (i+1)\theta)]} p_{00} \\
&= p_{00} \frac{\lambda^n}{n!} \prod_{i=0}^{n-1} \frac{\frac{\lambda}{\theta} + i}{[\mu + \alpha(\frac{\lambda}{\theta} + i + 1)]} \\
&= p_{00} \frac{\lambda^n}{n!\alpha^n} \prod_{i=0}^{n-1} \frac{a+i}{[\frac{\mu}{\alpha} + (a+1) + i]}.
\end{aligned}$$

Donc :

$$p_{0n} = \frac{\lambda^n}{n!\alpha^n} \frac{(a)_n}{(b)_n} p_{00}. \tag{2.1.9}$$

Où $a = \frac{\lambda}{\theta}$ $,b = \frac{\mu}{\alpha} + a + 1$ et $(a)_n = a(a+1)(a+2)(a+3)...(a+n-1)$ est la fonction factorielle croissante(Indice de Pochammar).
Pour obtenir la probabilité de n clients lorsque le système est occupé

$$\begin{aligned}
p_{1n} &= \frac{[\lambda + n\theta]}{\mu} p_{0n} \\
&= \frac{\theta}{\mu}[a+n] p_{0n} \\
&= \frac{\theta}{\mu}[a+n] \frac{\lambda^n}{n!\alpha^n} \frac{(a)_n}{(b)_n} p_{00} \\
&= \frac{\theta}{\mu}[a+n] \frac{\lambda^n}{n!\alpha^n} \frac{a(a+1)(a+2)(a+3)...(a+n-1)}{(b)_n} p_{00}.
\end{aligned}$$

Donc :

$$p_{1n} = \frac{\lambda}{\mu} \frac{\lambda^n}{n!\alpha^n} \frac{(a+1)_n}{(b)_n} p_{00}. \tag{2.1.10}$$

On définit les fonctions génératrices de probabilité comme suit :

$$P_0(z) = \sum_{n=0}^{\infty} p_{0n} z^n,$$

$$P_1(z) = \sum_{n=0}^{\infty} p_{1n} z^n.$$

En multipliant les équations (2.1.9) et (2.1.10) par z^n, en faisant la somme sur n de 0 à ∞ et en appliquant la fonction confluente Kummer définie par $\phi(a,b;z) = \sum_{n=0}^{\infty} \frac{(a)_n z^n}{(b)_n n!}$ nous obtenons,

$$\begin{aligned} P_0(z) &= p_{00} \sum_{n=0}^{\infty} \frac{\lambda^n}{n!\alpha^n} \frac{(a)_n}{(b)_n} z^n \\ &= p_{00} \sum_{n=0}^{\infty} \frac{(a)_n}{n!(b)_n} (\frac{\lambda z}{\alpha})^n. \\ P_1(z) &= p_{00} \sum_{n=0}^{\infty} \frac{\lambda \lambda^n}{\mu n! \alpha^n} \frac{(a+1)_n}{(b)_n} \\ &= p_{00} \frac{\lambda}{\mu} \sum_{n=0}^{\infty} \frac{(a+1)_n}{n!(b)_n} (\frac{\lambda z}{\alpha})^n. \end{aligned}$$

Cela peut être écrit comme :

$$P_0(z) = \phi(a,b;\frac{\lambda z}{\alpha}) p_{00}, \tag{2.1.11}$$

$$P_1(z) = \rho \phi(a+1,b;\frac{\lambda z}{\alpha}) p_{00} \tag{2.1.12}$$

Ici p_{00} peut être déterminée en utilisant la condition de normalisation $P_0(1) + P_1(1) = 1$, ce qui implique :

$$p_{00} = \frac{1}{\phi(a,b;\frac{\lambda}{\alpha}) + \rho \phi(a+1,b;\frac{\lambda}{\alpha})} \tag{2.1.13}$$

2.1.2 Mesures de performance

Les diverses mesures de performance peuvent être dérivées des fonctions génératrices de probabilité.

a. La probabilité d'occupation du serveur est donnée par :

$$p_1 = \sum_{n=0}^{\infty} p_{1n} = P_1(1) = \frac{\rho \phi(a+1,b;\frac{\lambda}{\alpha})}{\phi(a,b;\frac{\lambda}{\alpha}) + \rho \phi(a+1,b;\frac{\lambda}{\alpha})}$$

$$où \qquad \rho^* = \frac{\rho\phi(a+1,b;\frac{\lambda}{\alpha})}{\phi(a,b;\frac{\lambda}{\alpha})}$$

On aura :

$$p_1 = \frac{\rho^*}{1+\rho^*} \tag{2.1.14}$$

b. Le nombre moyen de clients dans l'orbite L_0

$$L_0 = P_0'(1) + P_1'(1) \tag{2.1.15}$$

On a la dérivée de : $\Phi(a,b,z) = \sum_{n=0} \frac{(a)_n}{(b)_n}.\frac{z^n}{n!}$ est donnée par

$$\begin{aligned}\Phi'(a,b,z) &= \sum_{n=1}^{\infty} n\frac{(a)_n}{(b)_n}.\frac{z^{n-1}}{n!} \\ &= \sum_{n=1}^{\infty} \frac{(a)_n}{(b)_n}\frac{z^{n-1}}{(n-1)!} \\ &= \sum_{n=0}^{\infty} \frac{(a)_{n+1}}{(b)_{n+1}}\frac{z^n}{n!}.\end{aligned}$$

comme

$$(a)_{n+1} = a(a+1)(a+2)\dots(a+n),$$

$$(a+1)_n = (a+1)(a+2)\dots(a+n),$$

$$(a)_{n+1} = a(a+1)_n,$$

$$(b)_{n+1} = b(b+1)_n,$$

on aura :

$$\begin{aligned}\Phi'(a,b,z) &= \sum_{n=0}^{\infty} \frac{a(a+1)_n}{b(b+1)_n}\frac{z^n}{n!} \\ &= \frac{a}{b}\Phi(a+1,b+1;z).\end{aligned}$$

Par la suite, on aura :

$$P_0^{'}(z) = \frac{\lambda}{\alpha}\Phi^{'}(a,b,\frac{\lambda}{\alpha}z)p_{00}, \quad (2.1.16)$$

$$P_1^{'}(z) = \rho\frac{\lambda}{\alpha}\Phi^{'}(a+1,b,\frac{\lambda}{\alpha}z)p_{00}$$

$$= \rho\frac{\lambda}{\alpha}\frac{a+1}{b}\Phi^{(}a+2,b+1,\frac{\lambda}{\alpha}z)p_{00}, \quad (2.1.17)$$

pour $z=1$, on aura :

$$P_0^{'}(1) = \frac{\lambda}{\alpha}\frac{a}{b}\Phi^{'}(a+1,b+1,\frac{\lambda}{\alpha})p_{00}, \quad (2.1.18)$$

$$P_1^{'}(1) = \frac{a+1}{b}\rho\frac{\lambda}{\alpha}\Phi(a+2,b+1,\frac{\lambda}{\alpha})p_{00}, \quad (2.1.19)$$

par conséquant en combinant (2.1.18) (2.1.19) nous obtenons

$$L_0 = \frac{p_{00}}{\theta\mu+\alpha\lambda+\alpha\theta}[\lambda^2\phi(a+1,b+1,\frac{\lambda}{\alpha}) + \rho(\lambda^2+\lambda\theta)\phi(a+2,b+1,\frac{\lambda}{\alpha})] \quad (2.1.20)$$

Ainsi, on peut déterminer d'autres mesures de performance en utilisant les formules de Little :

c. Le nombre moyen de clients dans le système est

$$L = L_0 + p_1.$$

d. Le temps d'attente moyen des clients dans l'orbite W_0

$$W_0 = \frac{L_0}{\lambda}.$$

e. Le temps moyen d'attente des clients dans le système W

$$W = \frac{L}{\lambda}.$$

Remarque 2.1. *Si le taux d'impatience $\alpha = 0$, notre modèle de file d'attente devient $M/M/1$ avec rappels étudié par Donald Gross et al. [18].*

2.2 Illustrations numériques

Dans cette section, quelques résultats numériques sont présentés pour illustrer l'effet de divers paramètres sur les mesures de performance du système. Tous les calculs ont été réalisés avec le logiciel Matlab.

Partie 1

En générant deux suites deux variables aléatoires et en calculant leurs moyennes $\lambda = 2$, $\mu = 5$, pour $\theta = 1, 2, 3, ...16$ et $\alpha = 2, 3$ et 4. Les résultats obtenus sont exhibés dans les figures suivantes :

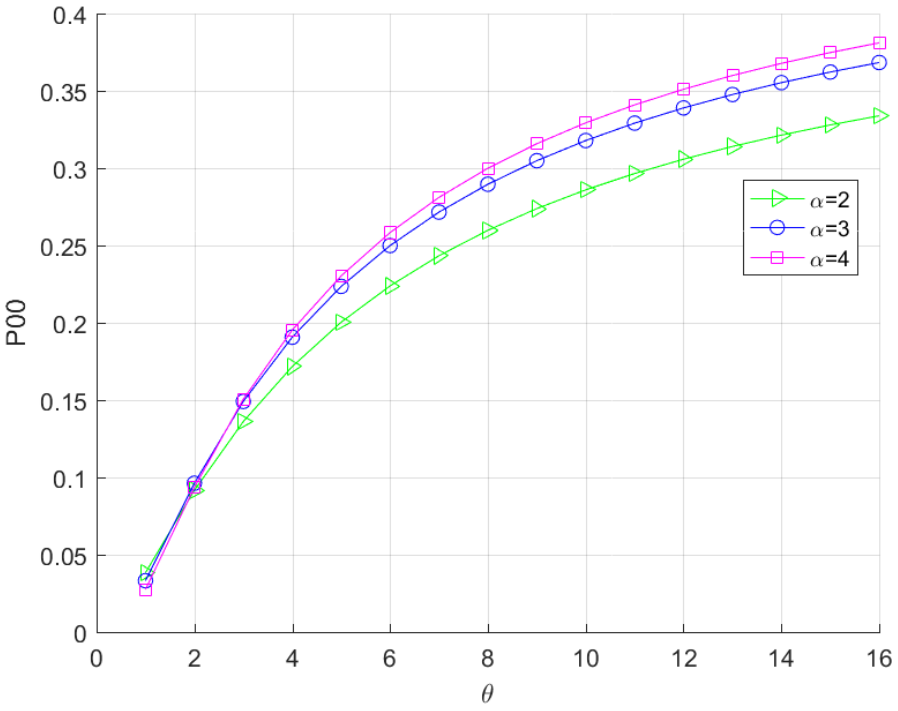

FIGURE 2.2: La probabilité que le système est vide en foction de $\theta et \alpha$.

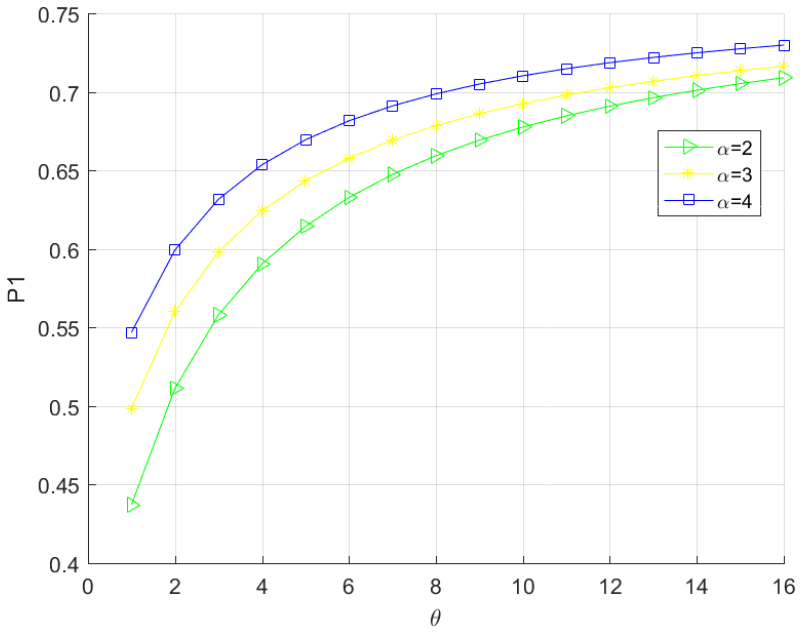

FIGURE 2.3: La probabilité d'occupation du serveur

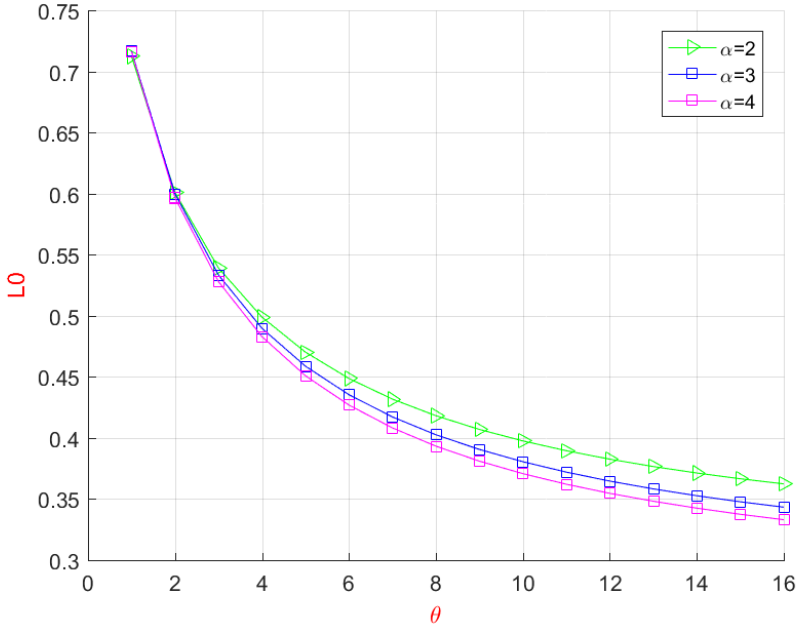

FIGURE 2.4: Nombre moyen de clients en orbite

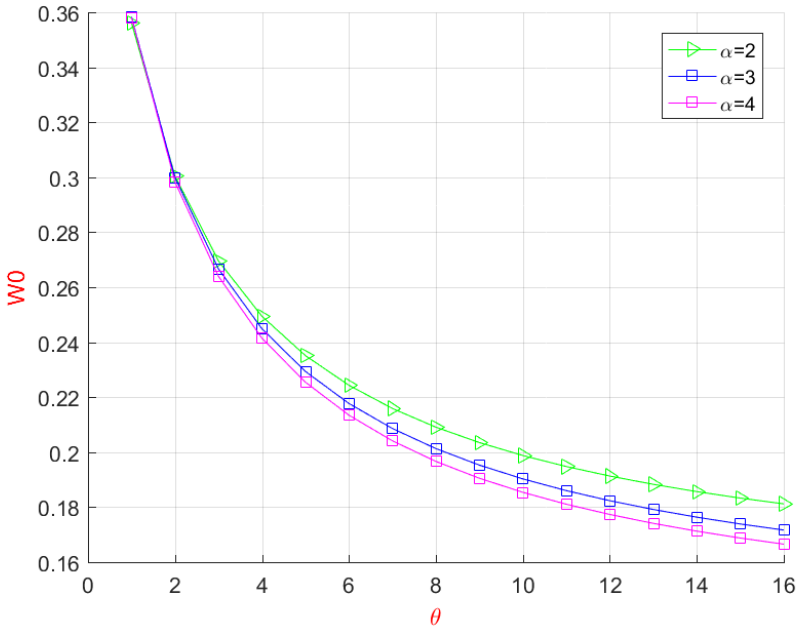

FIGURE 2.5: Temps moyen d'attente en orbite

Interprétations :

- Lorsque le taux de rappels croit, la probabilité p_{00} que le serveur est libre et l'orbite est vide croit. Cette croissance est plus importante lorsque le taux d'impatience est en croissance. Ce qui est logique comme le montre la Figure 2.2.
- D'après la Figure 2.3, on remarque que lorsque le taux de rappels croit, la probabilité d'occupation de serveur croit, cette croissance est moins importante avec la décroissance de taux d'impatience.
- D'après les Figures 2.4 et 2.5, on constate que le nombre moyen de clients et le temps moyen de séjours dans l'orbite décroit avec la croissance des taux de rappels et de l'impatience.

Partie 2

En générant deux suites deux variables aléatoires et en calculant leurs moyennes pour les valeurs $\lambda = 2$, $\mu = 16$, pour $\alpha = 1, 2, 3, ...16$ et $\theta = 2, 3, 4$.

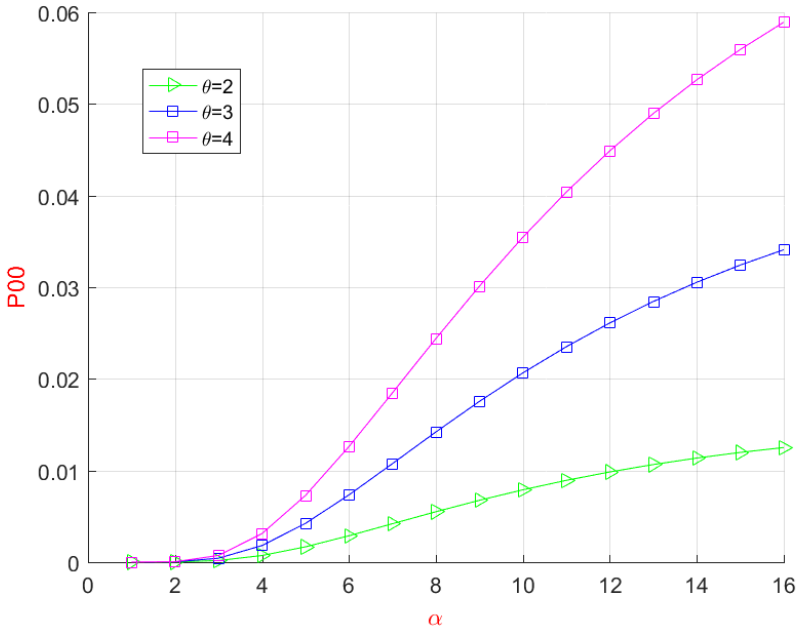

FIGURE 2.6: La probabilité que le système soit vide

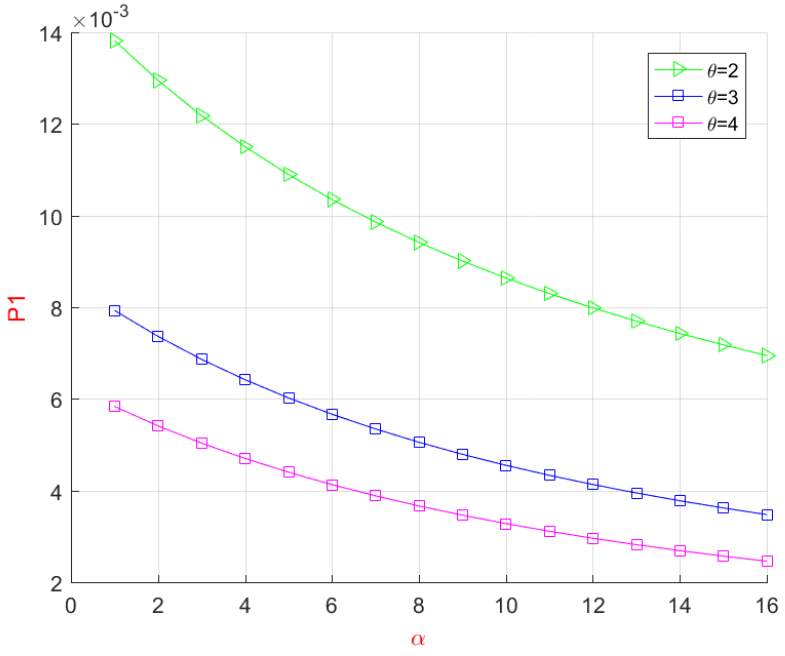

FIGURE 2.7: La probabilité d'occupation du serveur

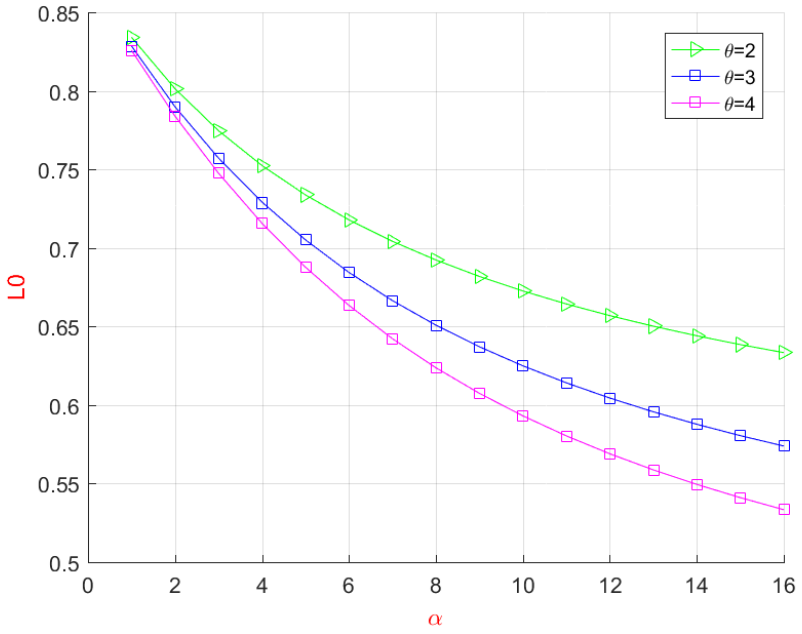

FIGURE 2.8: Nombre moyen de clients en orbite

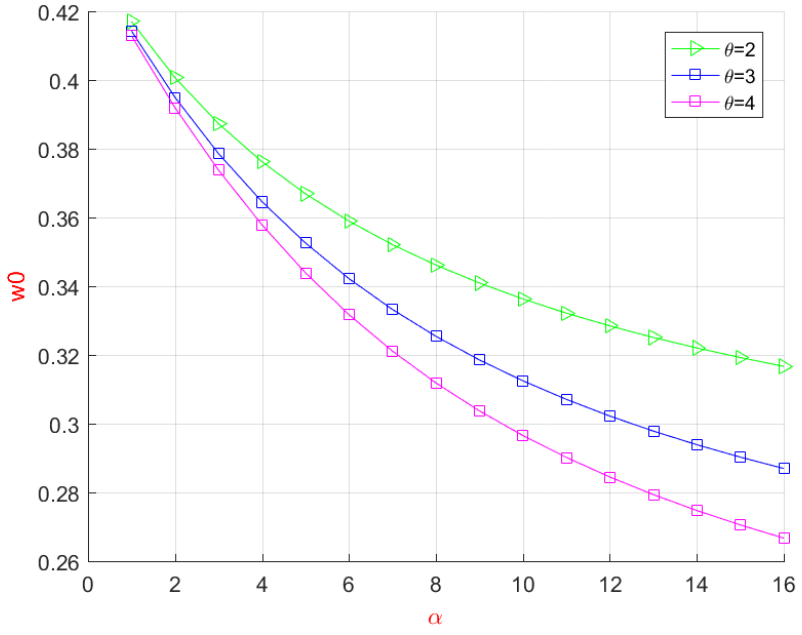

FIGURE 2.9: Temps moyen d'attente en orbite

Interprétation :

- D'après la Figure 2.6 la probabilité que le système est vide croit avec la croissance de taux d'impatience et de taux de rappels (ce qui est logique).

- La probabilité d'occupation du serveur décroit avec la croissance de taux d'impatience et de taux de rappels (ce qui est logique) comme le montre la Figure 2.7.
- Le nombre moyen de clients dans l'orbite décroit avec la croissance de taux d'impatience et de taux de rappels comme l'indique la Figure 2.8.
- Le temps moyen d'attente dans l'orbite décroit avec la croissance de taux d'impatience et de taux de rappels (ce qui est logique) comme l'indique la Figure 2.9.

Conclusion

Dans le cas des systèmes markoviens avec rappels, la modélisation de l'impatience est souvent simple à faire. Dans la pratique, la distribution de temps de service n'est pas en général une loi exponentielle, et l'ajout de toute autre élément décrivant le système, complique l'étude. C'est le cas de système $M/G/1/1$ avec rappels et impatience dans l'orbite qui fera l'objet de notre étude dans le chapitre suivant.

CHAPITRE 3

LE MODÈLE M/G/1 AVEC RAPPELS ET IMPATIENCE EN ORBITE

Introduction

Dans ce chapitre, nous allons généraliser la distribution de service du modèle markovien $M/M/1$ avec rappels et impatience au modèle non markovien $M/G/1$ avec rappels et impatience.

Exemple illustratif

Dans les réseaux locaux (LAN), l'un des protocoles de communication les plus utilisés est CSMA (Carrier-Sence Multiple Acces) non-persistant. Supposons qu'un réseau local est composé de n stations connectées par un seul bus. La communication entre les stations est réalisée au moyen de ce bus. Les messages arrivent aux stations du monde extérieur. En recevant le message, la station consulte le bus pour voir s'il est occupé. Si le bus est libre, le message est transmis via ce bus à la station de destination. Autrement, le message est stocké dans le tampon et la station peut consulter le bus après une certaine durée aléatoire. Les questions concernant ce problème sont : Quel est le temps moyen d'attente d'un message ? Quel est le nombre moyen de messages (paquets) dans le tampon d'une station ? Si les messages arrivent selon un processus de Poisson et la durée de service suit une loi général alors le système peut être modélisé comme un système M/G/1 avec rappels et impatience. Le serveur est le bus et les tampons des stations représentent l'orbite. Si la capacité des tampons est très grande, on a un système de files d'attente avec

rappels et une capacité de l'orbite infinie.
Pour repondre à toute ces questions, on procède par l'étude de la solution stationnaire de ce modèle en utilisant la méthode de la chaîne de Markov induite.

3.1 Description mathématique du modèle

Considérons le système $M/G/1/1$ avec rappels où les clients primaires arrivent selon un processus de Poisson de paramètre $\lambda > 0$. Il y a un seul serveur et il n'y a pas de places d'attente. A l'arrivée d'un client primaire, si le serveur est occupé, le client entre en orbite et devient un client secondaire. Chaque client de l'orbite forme un processus de Poisson d'intensité $\theta > 0$. Si le serveur est libre à l'arrivée d'un client primaire ou secondaire, le client commence son service et quitte le système après sa complétion. La distribution de temps de service est $B(t)$ (non exponentielle) de transformée de laplace $B^*(s) = \int_0^\infty e^{-st} dB(t)$. Soit $\beta_k = (-1)^k B^{(k)}(0)$ le $k^{ième}$ moment du service. Les clients secondaires activent délai d'impatience indépendament l'un de l'autre qui suit une distribution exponentielle de paramètre $\alpha > 0$. Si le minuteur expire, avant le service, le client abandonne l'orbite sans être servi. De plus, on suppose que les durées inter-arrivées, les durées inter-rappels, les durées de service et les durées d'impatience sont mutuellement indépendantes.
Soit $N(t)$ le nombre de clients en orbite à l'instant t et $C(t)$ l'état du serveur ($C(t) = 1$ ou 0 selon que le serveur est occupé ou libre).
Le processus $\{(N(t),C(t)),\ t \geq 0\}$ n'est pas markovien.
Pour l'analyse de ce système, on procède par la méthode de la chaîne de Markov induite.

3.2 La chaîne de Markov induite

Soit $N_i = N(t_i)$ le nombre de clients en orbite à l'instant du départ t_i (en raison de la fin du service ou de l'impatience). Nous avons

$$N_i = N_{i-1} - B_i - L_i + A_i, \tag{3.2.1}$$

où $\mathrm{B}_i = \begin{cases} 1, & \text{si le client provient de l'orbite ;} \\ 0, & \text{si le client provient de l'extérieur.} \end{cases}$

Sa distribution conditionnelle est donnée par :

$$\mathbb{P}(B_i = 1|N_{i-1} = n) = \frac{n\theta}{\lambda + n\theta} \quad et \quad \mathbb{P}(B_i = 0|N_{i-1} = n) = \frac{\lambda}{\lambda + n\theta}. \tag{3.2.2}$$

L_i est le nombre de clients en orbite perdus au cours du $i^{ème}$ service et A_i est le nombre de clients primaires qui arrivent durant le $i^{ème}$ service.

La distribution conditionnelle du vecteur aléatoire (L_i, A_i) est donnée par :

$$\begin{aligned}\mathbb{P}(L_i = k, A_i = n|N_{i-1} - B_i = m) &= \int_0^\infty C_m^k e^{-\alpha(m-k)t}(1 - e^{-\alpha t})^k \times \frac{\left[\frac{\lambda}{\alpha}(1 - e^{\alpha t})\right]^n}{n!} \\ &\quad \times \exp\{-\frac{\lambda}{\alpha}(1 - e^{-\alpha t})\}dB(t).\end{aligned}$$

Avec la fonction génératrice

$$\mathbb{E}(x^{L_i}y^{A_i}|N_{i-1} - B_i = m) = \int_0^\infty [x + (1 - x)e^{-\alpha t}]^m \times \exp\{\frac{\lambda}{\alpha}(1 - e^{-\alpha t})(y - 1)\}dB(t).$$

Cela donne

$$\mathbb{E}(L_i|N_{i-1} - B_i = m) = \int_0^\infty m(1 - e^{-\alpha t})dB(t) = m[1 - B^*(\alpha)],$$

la variable aléatoire A_i ne depend pas de N_{i-1} et de B_i,
et

$$\mathbb{E}(A_i) = \int_0^\infty \frac{\lambda}{\alpha}(1 - e^{-\alpha t})dB(t) = \frac{\lambda}{\alpha}(1 - B^*(\alpha)). \tag{3.2.3}$$

En raison de la structure récursive de l'équation (3.2.1), pour étudier l'ergodicité de la chaîne considérée, nous utiliserons le critère basé sur l'accroissement moyen (critère de Foster) [22]. Nous avons :

$$\begin{aligned}\triangle_n &= \mathbb{E}(N_i - N_{i-1}|N_{i-1} = n) \\ &= \mathbb{E}(-B_i - L_i + A_i|N_{i-1} = n) \\ &= -\mathbb{E}(B_i|N_{i-1} = n) - \mathbb{E}(L_i|N_{i-1} = n) + \mathbb{E}(A_i|N_{i-1}) \\ &= -\frac{n\theta}{\lambda + n\theta} + \frac{\lambda}{\alpha}(1 - B^*(\alpha)) - \mathbb{E}(L_i|N_{i-1} = n).\end{aligned}$$

Mais

$$\begin{aligned}\mathbb{E}(L_i|N_{i-1}=n) &= \mathbb{E}(L_i|N_{i-1}-B_i=n)\mathbb{P}(B_i=0|N_{i-1}=n)\\ &\quad +\mathbb{E}(L_i|N_{i-1}-B_i=n-1)\mathbb{P}(B_i=1|N_{i-1}=n)\\ &= n(1-B^*)(\alpha)\frac{\lambda}{\lambda+n\theta}+(n-1)(1-B^*(\alpha))\frac{n\theta}{\lambda+n\theta}\\ &= (n-\frac{n\theta}{\lambda+n\theta})(1-B^*(\alpha)). \qquad (3.2.4)\end{aligned}$$

D'où

$$\triangle_n = -[n-\frac{\lambda}{\alpha}][1-B^*(\alpha)]-\frac{n\theta}{\lambda+n\theta}B^*(\alpha).$$

$\lim_{n\to\infty} \triangle_n = -\infty$. D'où la chaîne de Markov $\{N_i\}$ est toujours ergodique.

3.3 Mesures de performance

Pour déterminer des mesures de performance du modèle, on prend les espérances mathématiques des deux membres de l'equation (3.2.1) :

$$0 = -\mathbb{E}(B_i)-\mathbb{E}(L_i)+\mathbb{E}(A_i).$$

En utilisant les équations (3.2.2), (3.2.3) et (3.2.4), nous avons :

$$\frac{B^*(\alpha)}{1-B^*(\alpha)}\mathbb{E}(\frac{\theta N_i}{\lambda+\theta N_i}) = \frac{\lambda}{\alpha}-\mathbb{E}(N_i). \qquad (3.3.1)$$

De l'inégalité de Jensen et comme la fonction $\varphi(x)=\frac{x}{\lambda+x}$ est concave, nous avons :

$$\mathbb{E}[\frac{\theta N_i}{\lambda+\theta N_i}] \le \frac{\theta\mathbb{E}(N_i)}{\lambda+\theta\mathbb{E}(N_i)}.$$

D'où :

$$N^2-(\frac{\lambda}{\alpha}-\frac{\lambda}{\theta}-\frac{B^*(\alpha)}{1-B^*(\alpha)})N-\frac{\lambda^2}{\theta\alpha}\ge 0,$$

où $N=\mathbb{E}(N_i)=L_0$ est le nombre moyen de clients en orbite.
Soit N_+, la racine positive de cette équation. Puisque, pour $N=0$, la partie gauche de cette équation est négative, cela implique que $N \ge N_+$. Comme le service consiste en une periode d'activité et d'inactivité avec β_1 et la durée moyenne d'activité, la longueur moyenne de la période d'inactivité est de :

$$\sum_{n=0}^{\infty} \pi_n \frac{1}{\lambda + n\theta} = \mathbb{E}(\frac{1}{\lambda + \theta N_i}),$$

où π_n est la distribution stationnaire de la chaîne induite, ainsi la probabilité du temps pour que le serveur soit occupé est alors donnée par :

$$p_1 = \frac{\beta_1}{\beta_1 + \mathbb{E}(\frac{1}{\lambda+\theta N_i})} = \frac{\rho}{\rho + \mathbb{E}(\frac{\lambda}{\lambda+\theta N_i})}, \quad \text{avec} \quad \rho = \frac{\lambda}{\mu} = \lambda\beta_1.$$

En utilisant (3.3.1),
On a :

$$\begin{aligned} p_1 &= \frac{\rho}{\rho + \mathbb{E}[1 - \frac{\theta N_i}{\lambda+\theta N_i}]} \\ &= \frac{\rho}{1 + \rho - \mathbb{E}[\frac{\theta N_i}{\lambda+\theta N_i}]} \\ &= \frac{\rho}{1 + \rho + (N - \frac{\lambda}{\alpha})(\frac{1-B^*(\alpha)}{B^*(\alpha)})} \\ &\leq \frac{\rho}{1 + \rho + (N_+ - \frac{\lambda}{\alpha})(\frac{1-B^*(\alpha)}{B^*(\alpha)})} \\ &= \frac{\rho + \theta\beta_1 N_+}{1 + \rho + \theta\beta_1 N_+}. \end{aligned}$$

Remarque :
D'après les lois de Little on peut déduire que :

- Le Nombre moyen des clients dans le système est : $L = L_0 + p_1$.
- Le temps moyen d'attente en orbite est : $W_0 = \frac{L_0}{\lambda}$.
- Le temps de sejours moyen dans le système est : $W = \frac{L}{\lambda}$.

3.4 Illustration numérique

L'objectif de cette section est d'étudier l'influence des taux des rappels et d'impatience sur les diverses mesures de performance du modèle $M/G/1$ avec rappels et impatience. Nous avons élaboré un simulateur sous Matlab pour générer le taux des inter-arrivées et le taux de service de ce modèle pour quelques distributions du temps de service.

► **La durée de service est déterministe :**

Pour les exemples qui suivent, nous avons généré le taux des inter-arrivées $\lambda = 3$ et le taux de la durée de service $\mu = 7$.

a- L'influence sur la probabilité que le système soit occupé ($\mathbf{p}_1$) :

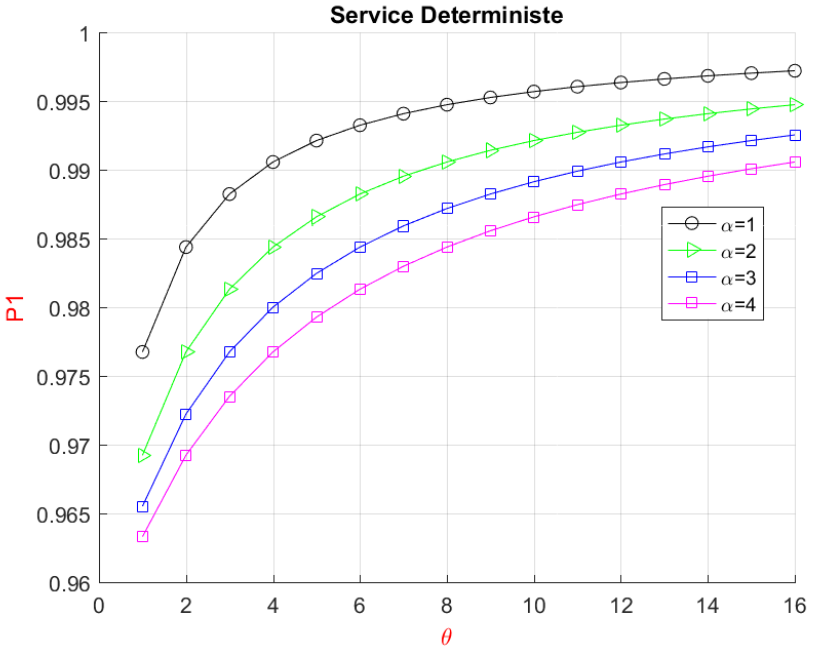

FIGURE 3.1: La probabilité que le serveur soit occupé

Interprétation :

D'après la Figure 3.1, la probabilité d'occupation du serveur croit avec la croissance de taux de rappels, cette croissance est moins importante lorsque le taux d'impatience α croit.

b- Nombre moyen de clients dans l'orbite L_0 et dans le système L :

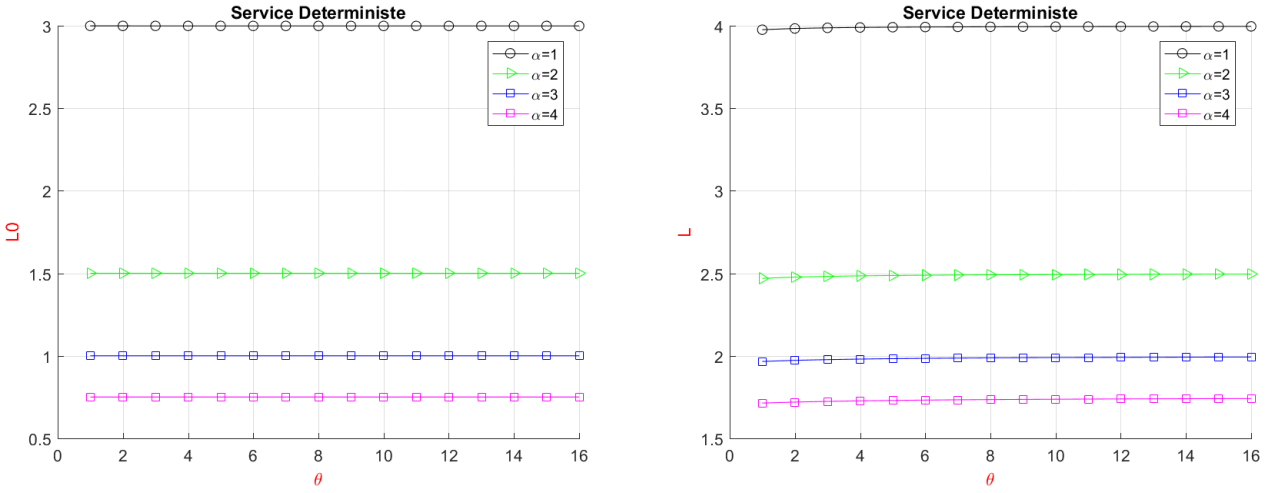

FIGURE 3.2: Nombre moyen de clients dans le système et dans l'orbite

c- Le temps d'attente moyen dans l'orbite W_0 et dans le système W :

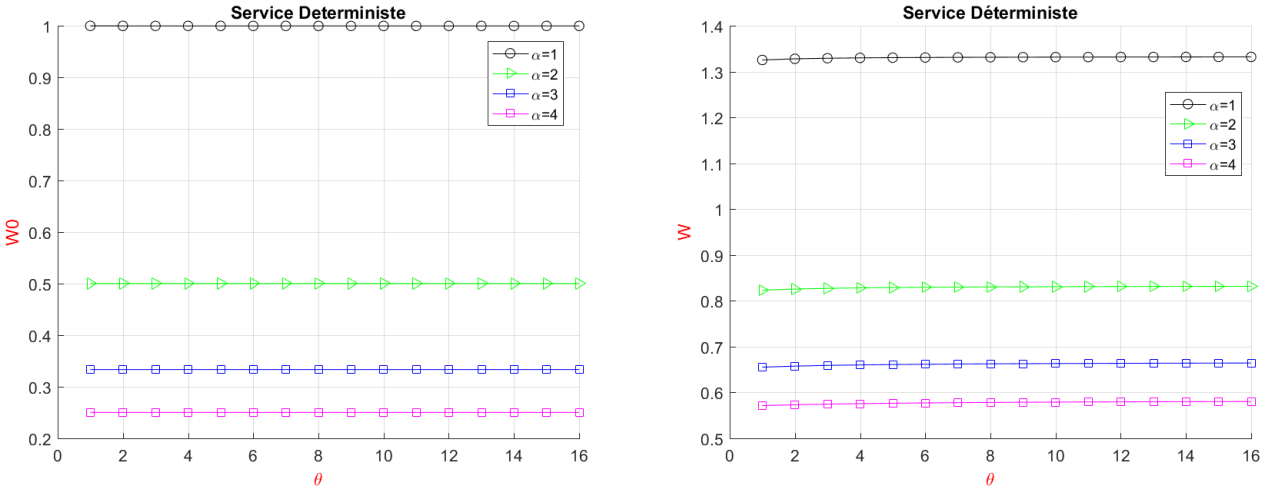

FIGURE 3.3: Le temps d'attente moyen dans l'orbite W_0 et dans le système W

Interprétation :

D'après la Figure 3.2, on constate que le nombre moyen de clients dans l'orbite L_0 décroit légèrement avec la croissance de taux de rappels θ, mais il est moins important avec la croissance de taux d'impatience α. Par contre le nombre moyen de clients dans le système L croit légèrement avec la croissance de taux de rappels tout en décroissant avec la croissance de taux d'impatience α. On en déduit que l'influence de taux d'impatience sur L et L_0 est très considérable par

rappot à θ.

Le temps moyen d'attente dans l'orbite décroit légèrement avec la croissance de taux de rappels θ, mais il est plus important avec la décroissance de taux d'impatience α. Par contre le temps moyen d'attent dans le système L croit légèrement avec la croissance de taux de rappels tout en décroissant avec la croissance de taux d'impatience α. On en déduit que l'influence de taux d'impatience α sur W et W_0 est très important par rappot à θ.

▶ **La durée de service suit la loi Gamma d'ordre 2 :**

Pour les exemples qui suivent, nous avons repris les mêmes valeurs que pour le service déterministe $\lambda = 3$ et le taux de la durée de service $\mu = 7$.

a- L'influence sur la probabilité que le système soit occupé (p_1) :

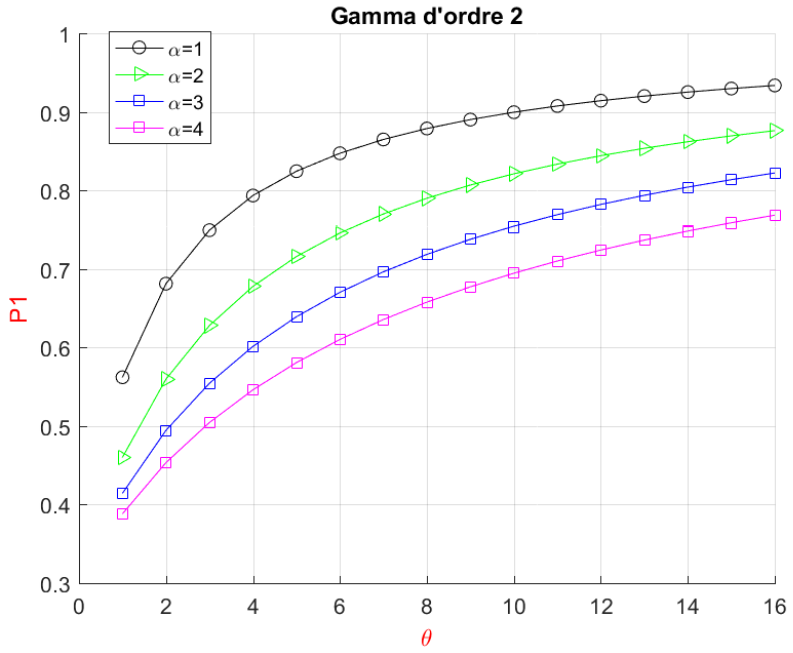

FIGURE 3.4: La probabilité que le serveur soit occupé

Interprétation :

La probabilité d'occupation du système p_1 croit avec la croissance de taux de rappels θ et cette croissance est moins importante avec la croissance de taux d'impatience α comme l'indique la Figure 3.4.

b- Nombre moyen de clients en orbite (L_0) et dans le système (L) :

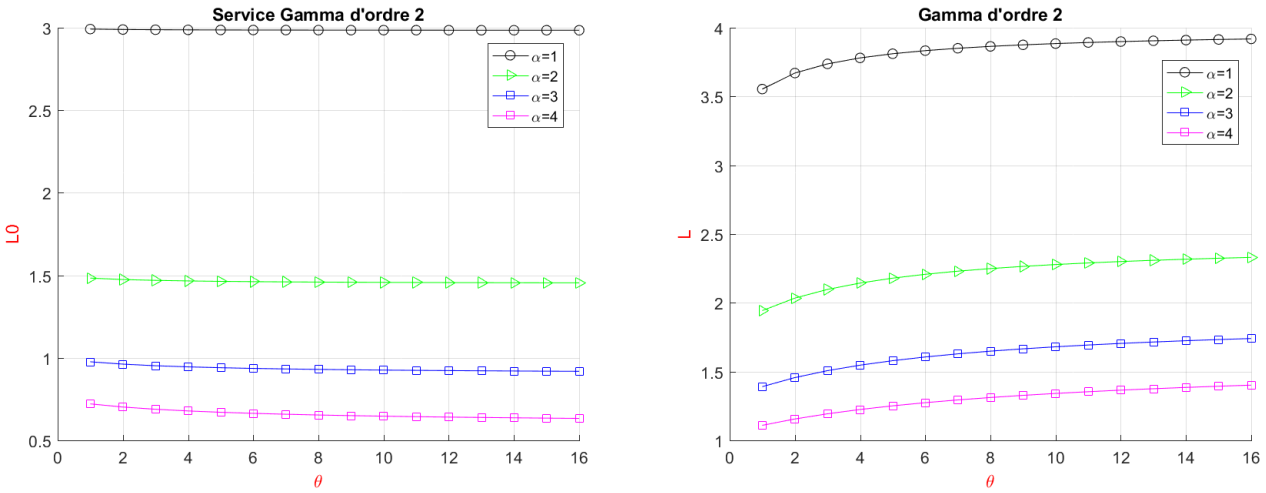

FIGURE 3.5: Nombre moyen de clients dans le système et dans l'orbite

c- Le temps moyen d'attente en orbite (W_0) et dans le système (W) :

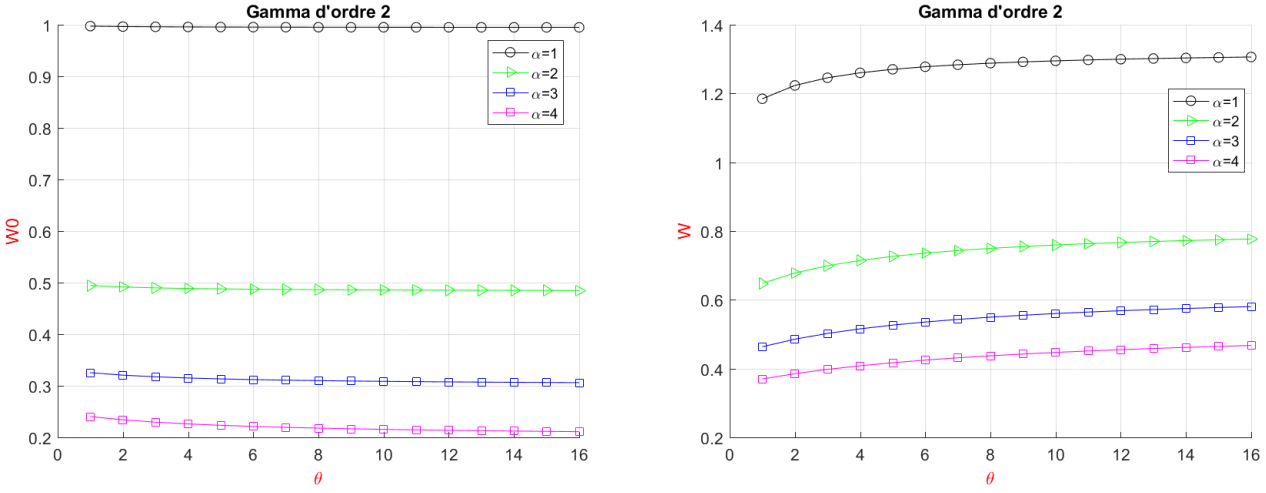

FIGURE 3.6: Le temps moyen d'attente dans l'orbite W0 et dans le système W

Interprétations :

- D'après la Figure 3.5, on constate que le nombre moyen de clients en orbite L_0 décroit légèrement avec la croissance de taux de rappels θ, cette décroissance est beaucoup plus importante avec la croissance de taux d'impatience α. Par contre, le nombre moyen de clients dans le système L croit légèrement avec la croissance de taux de rappels tout en décroissant avec la croissance de taux

d'impatience α. On en déduit que l'influence de taux d'impatience sur L et L_0 est très considérable par rappot à θ.

- On remarque que le temps moyen d'attente des clients en orbite L_0 décroit légèrement avec la croissance de taux de rappels θ, cette décroissance est beaucoup plus importante avec la croissance de taux d'impatience α. Par contre, le temps moyen de séjours des clients dans le système L croit légèrement avec la croissance de taux de rappels tout en décroissant avec la croissance de taux d'impatience α. On en déduit que l'influence de taux d'impatience sur L et L_0 est très considérable par rappot à θ comme le montre la Figure 3.6.

▶ **La durée de service suit la loi hyper-exponentielle d'ordre** 2

On considère la file d'attente $M/H_2/1$ avec rappels et impatience tel que le processus d'arrivés des clients est exponentiel de taux $\lambda = 3$ et de durée du service hyper-exponentielle d'ordre 2 de taux $\mu_1 = 5$ et $\mu_2 = 9$ pour les paramètres $\alpha = 2, 3$ et 4, $\theta = 1, 2, 3, \ldots, 16$ (choisis arbitrairement).

a- La probabilité que le système soit occupé p_1 :

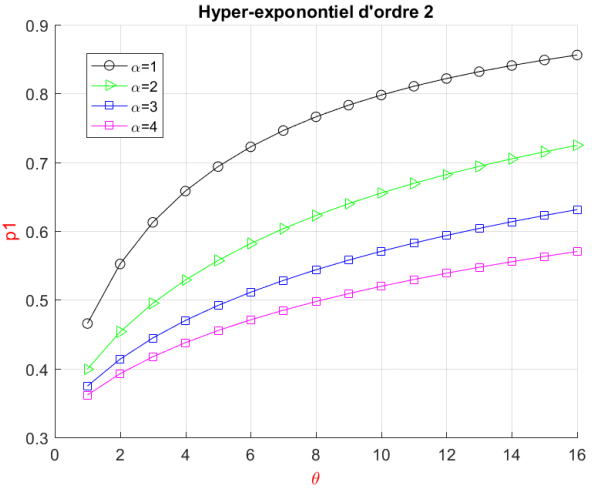

FIGURE 3.7: La probabilité que le système soit occupé

Interprétation :

D'après la Figure 3.7, on remarque que la probabilité d'occupation du serveur croit avec la croissance de taux de rappels θ et cette croissance est moins importante avec la croissance de taux d'impatience α.

b- Le nombre moyen de clients dans l'orbite $L0$ et dans le système L :

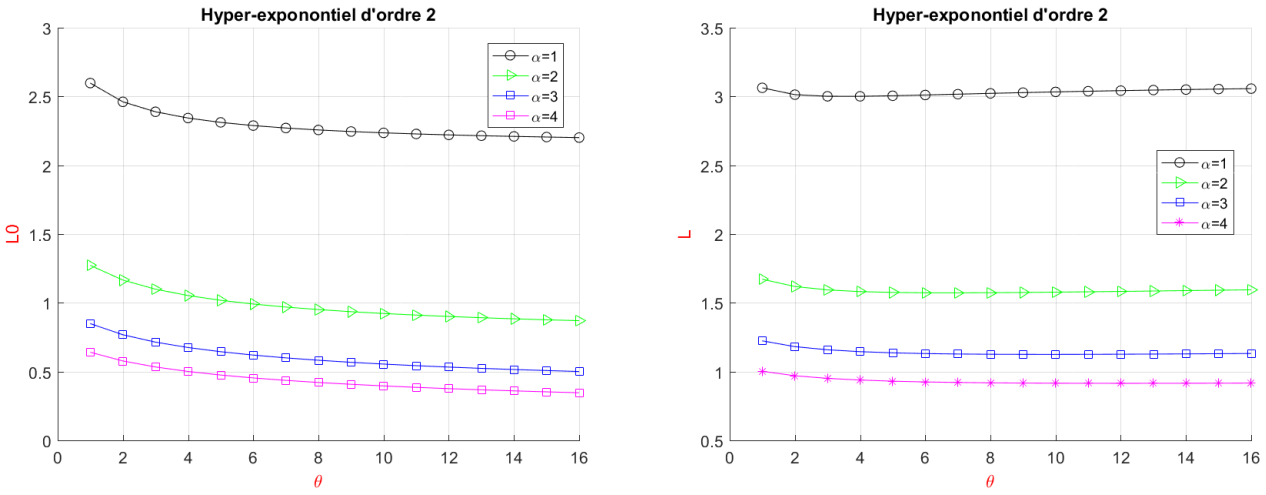

FIGURE 3.8: Nombre moyen de clients dans le système et dans l'orbite

c- Le temps moyen d'attente en orbite W_0 et dans le système W :

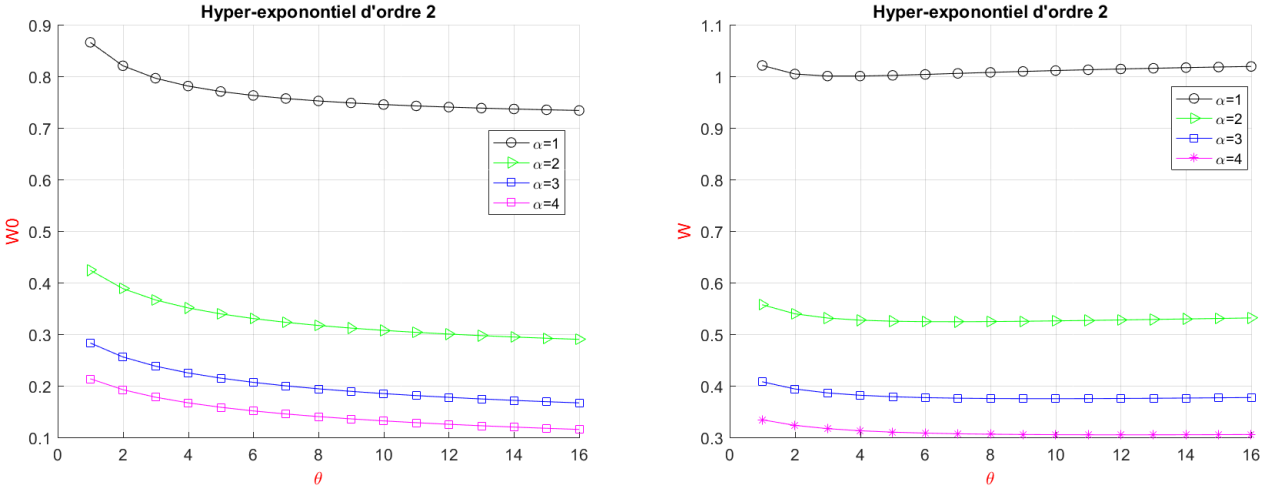

FIGURE 3.9: Le temps moyen d'attente dans l'orbite W_0 et dans le système W

Interprétations :

D'après les Figures 3.8 et 3.9 on remarque que :

- Le nombre moyen de clients et le temps moyen d'attente dans l'orbite décroit avec la croissance de taux de rappels θ et de taux d'impatience α (ce qui est logique).
- Le nombre moyen de clients et le temps moyen de séjour dans le système décroit avec la croissance de θ et cette décroissance est plus importante losque le taux d'impatience croit. On souligne que pour chaque valeur de

taux d'impatience α, une légère croissance de L et de W en fonction des dernières valeurs de θ ceci est dû à la croissance de p_1.

Conclusion

Dans ce chapitre nous avons éffectué une analyse stochastique et numérique du modèle non markovien $M/G/1$ avec rappels et impatience dans l'orbite, en utilisant différentes distributions pour la durée de service, on en déduit que le taux d'impatience a un effet plus considérable que le taux des rappels sur les mesures de performance du modèle.

CONCLUSION GÉNÉRALE ET PERSPECTIVES

Dans ce travail, nous nous sommes intéressés à l'étude des systèmes de files d'attente avec rappels et impatience dans l'orbite. En particulier, à l'analyse du modèle $M/M/1$ et $M/G/1$ afin de voir l'influence des paramètres sur les mesures de performance et aussi comparer les résultats numériques obtenus lors de notre simulation en utilisant les différentes distributions du temps de service dans le cas $M/G/1$.

Dans la première application, nous avons présenté quelques résultats numériques réalisés avec le logiciel Matlab pour illustrer l'effet de divers paramètres sur les mesures de performance du système $M/M/1$ avec rappels et impatience. On a constaté que l'impatience et les rappels influent sur les caractéristiques du modèle telle que le temps moyen d'attente en orbite et dans le système W_0, W et le nombre de clients dans le système et dans l'orbite L, L_0.

Dans la seconde application, nous avons utilisé le logiciel Matlab pour étudier le modèle M/G/1 avec rappels et impatience en utilisant les différentes distributions du temps de service dans le but de voir l'effet de divers paramètres sur les mesures de performance.

Dans le cas des systèmes markoviens avec rappels, malgrès que la modélisation de l'impatience est souvent simple à faire mais les résultats obtenus ne sont pas exploitables en pratique. Par contre, dans le cas non markovien, la distribution du temps de service n'est pas souvent exponentielle et le meilleur choix pour cette distribution dépends des taux de rappels et de l'impatience. D'où la nécessité d'approximation dans de tels modèles complexes.

Perspectives :

- ✓ On peut envisager d'élargir l'étude aux systèmes avec rappels et impatience en considérant d'aures disciplines de rappels à savoir : rappels constants et linéaires.
- ✓ Appliquer la théorie des jeux pour voir la meilleure stratégie à choisir pour minimiser l'impatience des clients.
- ✓ Elaboration d'un simulateur du modèle $M/G/1$ avec rappels et impatience dans l'orbite et comparaison des résultats de simulations avec ceux obtenus dans ce travail.

BIBLIOGRAPHIE

[1] M.O.Abou El Aba and A.M.A. Harris. The M/M/C/N queue with balking and reneging. Computer and Operations Research. 19 (13), pp.713-716 (1992).

[2] A. Aïssani. A Survey on Retrial Queueing Models. Actes des Journées Statistiques Appliquées, U.S.T.H.B., Alger, pp.1-11 (1994).

[3] N.K. Arrar. Problèmes de convergence, optimisation d'algorithmes et analyse stochastique de systèmes de files d'attente avec rappels. Thèse de Doctorat en Mathématiques Appliquées. Université de Annaba (2012).

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

[5] J. R. Artalejo. Accessible bibliography on retrial queues : Progress in 2000-2009. Mathematical and computer modelling 51, pp.1071-1081 (2010).

[6] J.R. Artalejo and A. Gomez-Corral. Retrial Queueing Systems. A Computational Approach. Springer (2008).

[7] J. R. Artalejo and M. J. Lòpez-Herrero. On the distribution of the number of retrials. Applied Mathematical Modelling, 31, pp.478-489 (2007).

[8] J. R. Artalejo and V. Pla. On the impact of customer balking, impatience and retrials in telecommunication systems. Computers Mathematics with Applications, 57, pp.217-229 (2009).

[9] J. R. Artalejo, V. Rajagopalan and R. Sivasamy. On finite Markovian queues with repeated attempts. Investigacion Operativa, 9, pp.83-94 (2000).

[10] K. Avrachenkov and U. Yechiali. Retrial networks with finite buffers and their application to internet data traffic. Probability in the Engineering and Informational Sciences, 22, pp.519-536 (2008).

[11] A. Azhagappan, E. Veermani, W. Monica and K. Sonabharathi. Transient Solution of an M/M/1 retrial queue with Reneging from orbit (13) (2), pp. 628-638 (2018).

[12] D. Y. Barrer Queueing with impatient Customers and Ordered service, Operations Research, 5, pp. 650-656 (1957).

[13] L. Berdjoudj. Stabilité forte dans les systèmes de files d'attente avec rappels. Thèse magister en Mathmatiques Appliquées. Université de Bejaia (2000).

[14] N. K. Boots, H. Tijms. A Multiserver queueing System with impatient Customers. Management Sciences. 45 (3), pp. 444-448 (1999).

[15] J. W. Cohen. Basic Problems of Telephone Traffic Theory and Influance of Repeated Calls. Philips Telecom. Review, 18(2), pp.49-100 (1957).

[16] G.I. Falin and J.G.C. Templeton. Retrial Queues. Chapman and Hall (1997).

[17] S. Gao, X. Niu and T. Li. Analysis of a constant retrial queue with joining strategy and impatient retrial customers. Mathematical Problems in Engineering. Volume, Article 9618215, 8 pages. (2017).

[18] D. Gross, J. F. Shortele, J.M. Thomposon, and C.M. Harris. Fundamental of Queueing Theory. John Wiley qnd Sons, London (2008).

[19] S. Hammache. Systèmes de file d'attente avec découragement. Mémoire de Master en Mathématiques, Analyse Stochastique, Statistique des Processus et Applications (ASSPA). Université de Saida (2018).

[20] J. Keilson, V.A. Cozzolino, and H. Young. A service system with unfilled requests repeated. Operations Research, 16, pp.1126–1137, (1968).

[21] J. Kim and B. Kim. A servey of retrial queuing systems. Annals of operations Research, pp.1-34 (2016)

[22] L. Kleinrock. Queueing Systems (Theory) Volume 1. John Wiley and Sons Edition, (1975).

[23] R. Kummer, and S.K. Sherma. M/M/1/N queueing with retention of reneged customers. Pakistan Journal of statistics and operation research, 8, pp.853-866 (2012).

[24] J. Lubacz and J. Roberts. A new approach to the single-server repeated attempt system with balking. Proceedings of the Third Int. Seminar on Teletraffic Theory. Moscow, pp.290-293 (1984).

[25] F. Machihara and M. Saitoh. Mobile customer model with retrials. European Journal of Operational Research, 189, pp.1073-1087, (2008).

[26] T. Meziani. Sur les files d'attente avec impatience. Mémoire de Master en Mathématiques option Statistique et Analyse Decisionnelle. Université de Bejaia (2016).

[27] T. Saaty. Elements of Queueing Theory and Applications. Mc Graw-Hill Book Company, N.Y, (1961).

[28] A. Shekhar, A. Raina and A. Kumar. A brief review on retrial queue. International Journal of Applied Sciences and Engineering Research, 5 (4), (2016).

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

[30] P. Suganthi, S. Pavai Madheswari. Retrial Queueing System with customer Impatience. Global Journal of Pure and Applied Mathematics. 11 (5), pp.3177–3188, (2015).

[31] K. Wang, N. Li and Z. Jiang. Queueing System with Impatient Customers. A Review. IEEE International Conference on Logistics and Informatics. pp. 82-87. (2010).

Résumé

Dans ce travail, nous avons abordé l'analyse stochastique et numérique des systèmes de files d'attente avec rappels et impatience des clients dans l'orbite.

Dans un premier lieu, en utilisant les fonctions génératrices nous avons obtenu la distribution stationnaire du système M/M/1 avec rappels et impatience ainsi que les mesures de performance.

Dans un second lieu, on a généralisé les résultats obtenus au système M/G/1 avec rappels et impatience en utilisant la chaîne de Markov induite. Enfin, une illustration numérique a été faite pour montrer l'effet des taux de rappels et de l'impatience sur les mesures de performances des deux modèles considérés.

Mots clés : Systèmes d'attente avec rappels - orbite- impatience- fonction génératrice- chaîne de Markov induite- mesures de performance.

Abstract

In this work, we invistigate the stochastic and numerical analysis of retrial queueing systems and impatience in the orbit.

Firstly, we used the generating functions to obtain the stationnary distribution of the M/M/1 retrial queue with impatience. Also, we derived the performance measures. Secondly, we generalized the obtained results to the M/G/1/ retrial queue with impatience by using the imbedded Markov chain. Finally, a numerical illustration was made to illustrate the effect of the retrial and impatience rates on the performance measures of the two considered models .

Key words : Retrial queueing systems-orbit- impatience - generating function - Embedded Markov chain - performance measures.