RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE
Ministère de l'Enseignement Supérieure et de la Recherche Scientifique
Université Saad Dahleb de Blida 1
Faculté des Sciences
Département des Mathématiques

Mémoire de fin de cycle

en vue de l'obtention du diplôme de Master en Mathématiques
Option : Modélisation Stochastique et Statistique

SUR LE MODÈLE D'ATTENTE *M/M/1* AVEC RAPPELS ET RECHERCHE DES CLIENTS EN ORBITE.

Présenté par:

- M_r BAGHDADI Nouh.
- M_r SAAD Mohamed.

Encadré par:

- *Madame*: BOUSSAHA Zina "MAA"
Académie Militaire de Cherchell D.P.H.B

Devant le jury composé de :

Présidente: *Madame*: BOUTAOUS Fatiha "MCB" USDB 1.
Examinatrice: *Madame*: OUKID Nadia "MCA" USDB 1.

Année universitaire : 2018/2019.

Table des matières

Remerciements

Avant d'entamer ce rapport du projet de fin d'étude, nous tenons à exprimer notre sincère gratitude envers tous ceux qui nous ont aidé ou participé au bon déroulement de ce projet.

Je remercie vivement mon encadreur Madame Z.BOUSSAHA, pour m'avoir proposé le sujet de ce mémoire, pour sa patience et ses conseils qui m'ont été d'un grand apport, et encouragement rendant toute la durée de l'élaboration de ce travail.

Je tiens également à remercier :

Le président de jury Madame F.BOUTAOUS, docteur de l'université de blida, d'avoir eu l'amabilité d'accepter la présidence du jury de soutenance.

Mes remerciements s'adressent également à Madame N.OUKID, (M.C.A à l'université de blida) de m'avoir honoré en acceptant d'examiner ce travail.

Je remercie tous ceux qui ont participé de prés ou de loin à l'élaboration de ce travail.

Je tiens également à remercier ma famille, tous mes professeurs, et tous mes amis(es).

Résumé :

Dans ce mémoire, nous avons utilisé la méthode de la fonction génératrice pour l´étude d'un système de file d'attente M/M/1 avec rappels et recherche de clients en orbite. Nous avons obtenu la fonction génératrice du nombre moyen de clients dans le système et des mesures de performances. En outre, nous avons effectué l'analyse de sensibilité des performances du modèle étudié par rapport à la variation des valeurs des paramètres. Plusieurs exemples numériques ont été réalises.

Mots-clés : Files d'attente avec rappels ; Chaîne de Markov induite ; Fonction génératrice ;la recherche des clients.

Abstract :

In this Master thesis, we have used the generating function method for analyzing a retrial queueing model with orbital search of customer. We have obtained the generating function of the length system and some performances measures. Furthermore,we have provided a sensitivity analysis of measures performances with respect to the variability values of the parameters models.

Keywords : Retrials queueing models ; Embedded Markov chain ; Generating function.customer search

0.1 Introduction générale

La théorie des files d'attente, est l'un des outils analytiques les plus puissants pour la modélisation des systèmes dynamiques.

Cette théorie a été initie au Danemark, entre 1909 et 1915 avec le développement du 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'operateurs, 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 à défaut 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 pour déterminer le nombre de circuits nécessaires afin de fournir un service téléphonique acceptable. Par la suite, les files d'attente ont été intègre dans la modélisation de divers domaines d'activité. On a assisté alors à une évolution rapide de la théorie des files d'attente qu'on appliquera à l'évaluation des performances des systèmes informatiques et aux réseaux de communication. Les chercheurs ouvrant dans cette branche d'activité ont élaboré plusieurs nouvelles méthodes qui ont été ensuite appliquées avec succès dans d'autre domaines, notamment dans le secteur de fabrication. On a aussi constaté une résurgence des applications pratiques de la théorie des files d'attente dans des secteurs plus traditionnels de la recherche opérationnelle, un mouvement mené par Peter Kolesar (1979) et Richard Larson(1987). La théorie mathématique des files d'attente s'est développé par la suite, grâce aux travaux de Palm, de Kolmogov et de Khintchin. Grâce à tous ces développements, cette théorie est aujourd'hui largement utilisée et ses applications sont multiples.

Dès la fin des années 1940 ; des chercheurs ont mis en évidence les limites de la théorie classique des files d'attente qui ne permettait pas d'expliquer le comportement stochastique des systèmes réels de plus en plus complexes, tels que les systèmes téléphoniques où les abonnés répétaient leurs appels en recomposant le numéro plusieurs fois jusqu' à l'obtention

de la communication. Ce phénomène de répétition de demandes du service a poussé certains chercheurs à étendre le modèle d'attente classique à celui dit avec rappels. 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. Les progrès récents dans ce domaine sont résumés dans les articles de synthèse de Yang et Templeton [24], Falin (1990) [17], A. Aïssani (1994) [3] et dans la monographie d'Artalejo et Gòmez (2008) [6].

La plupart des modèles avec rappels étudié dans la littérature supposent qu'après chaque service, le serveur restera inactif dans le système jusqu'à l'arrivée du prochain client primaire ou un client de l'orbite. Même s'il y a des clients dans le système qui veulent obtenir un service, ils ne peuvent pas occuper le serveur immédiatement, en raison de leur ignorance de l'état du serveur. Par conséquent, après l'achèvement de chaque service, le prochain client entre en service uniquement après un certain intervalle de temps pendant lequel le serveur est libre alors qu'il peut y avoir des clients en attente dans l'orbite. Mais dans la vrais vie, nous voulons toujours minimiser le temps d'inactivité du serveur et minimiser les coûts de conservation. Donc , il est nécessaire d'étudier les systèmes de file d'attente avec recherche des clients . La recherche de clients immédiatement à la fin d'un service était d'abord discuté dans le cas des files d'attente classiques par Neuts et Ramalhoto (1984) [19]. Artalejo et al (2002) [5] ont envisagé un système de files d'attente avec rappels dans lequel, immédiatement après l'achèvement d'un service, le serveur recherche un client en orbite ou reste inactif. Si une recherche est faite un service est suivi d'un autre service, sinon, il sera suivi d'un temps d'inactivité. Cependant, dans Artalejo et al (2002) [5] il n'y a pas de temps libre ni de temps de recherche (le temps de recherche est égal à zéro). Dudin et al [15] considèrent le même modèle que dans Artalejo et al. (2002) [5] avec une entrée BMAP et rechercher des clients. Cependant, le mécanisme de recherche est lancé juste après l'achèvement du service. Quelques autres extensions se trouvent dans, Artalejo et Phung-Duc (2012) [9] envisagent un modèle avec une communication bidirectionnelle

où, après le temps d'inactivité du serveur initiant un appel sortant dont la durée est exponentielle distribué. Ceci peut être considéré comme le temps de recherche dans notre modèle. cependant, après un appel sortant, le serveur reste inactif, c'est-à-dire qu'aucun client de l'orbite est ramassé. Dans tous les travaux ci-dessus, le temps d'inactivité et le temps de recherche sont considérés séparément. L'article Tuan Phung [23] est le premier qui propose un mécanisme de recherche qui est lancé après un temps d'inactivité du serveur. D'autres travaux sur les modèles avec recherches des clients pour les lecteurs ; Artalejo, Joshua, et Krishnamurthy (2002)[8], Dudin, Krishnamoorthy, Joshua, et Tsarenkov (2004) [15], Krishnamoorthy, Deepak, et Joshua (2005) [18], Chakravarthy, Krishnamoorthy, et Joshua (2006) [11], Wang (2006), Zhang et Wang(2012), Deepak, Dudin, Joshua, et Krishnamoorthy (2013) [18].

L'apparition de la virtualisation, de l'infogérance, de l'externalisation et de la démocratisation de l'informatique à la fin de XXe siècle, a permis d'assister ces dernières décennies à l'explosion du Cloud Computing (l'informatique en nuage). Ce concept constitue une tendance, mondiale en matière d'acquisition de services technologiques et une véritable révolution dans l'utilisation de l'informatique en amenant de nouvelles possibilités de mutualisation de services et d'économies pour les entreprises, notamment par le fait de diminuer les coûts d'exploitation des infrastructures technologiques et des applications. Par conséquent, développer des méthodes efficaces et précises pour évaluer les performances de ces services est devenu un problème de recherche très important qui a suscité une attention considérable de la part des milieux universitaires et industriels. Parmi les approches fondamentales permettant d'évaluer les performances dans le Cloud Computing, on trouve celles basées sur la modélisation stochastique.

Dans ce mémoire, nous considérons une étude analytique du modèle de file d'attente M/M/1 avec rappels et recherche des clients en orbite, Le modèle est motivé par les systèmes de cloud computing où l'unité du traitement et l'unité de stockage sont séparées . Nous obtenons la fonction génératrice du nombre moyen de clients dans le système, et celle du nombre moyen de clients dans l'orbite, ainsi plusieurs mesures de performance du modèle

étudie. En plus, nous intéressons à l'analyse de sensibilité de certaines performances du modèle étudie par rapport à la variabilité de leurs paramètres.

Ce mémoire comprend une introduction générale et quatre chapitres.

- Dans le premier chapitre nous présentons les notions et techniques de base sur les systèmes de files d'attente classiques et nous réalisons une étude des modèles markoviens et non-markoviens.

- Le deuxième chapitre comprend une étude des systèmes de files d'attente avec rappels. Une attention particulière est consacrée au modèles M/M/1 et M/G/1 avec rappels.

- Le troisième chapitre est pour l'étude du modèle de file d'attente M/M/1 avec rappels linéaire et recherche des clients en orbite par le serveur (temps d'inactivité et temps de recherche son négligeable).

-Le quatrième chapitre présente une application sur un modèle de file d'attente M/M/1 avec rappels et recherche des clients en orbite (temps d'inactivité et temps de recherche non négligeable), des résultats numérique pour étudier l'effet de certains paramètres clés sur les caractéristiques du modèle son présenté.

Finalement, nous terminerons par une conclusion générale et une bibliographie.

Chapitre 1

Modèles de files d'attente classiques

1.1 Introduction :

La théorie des files d'attente, et des réseaux de file d'attente sont des outils analytiques les plus puissants pour la modélisation des systèmes logistiques et de communication . En quelques mots, cette théorie à pour objet l'étude des systémes où des entités, appelées clients, cherchant à accéder à des ressources, généralement limitées, a fin d'obtenir un service . La demande concurrente d'une même ressouce par plusieurs clients engendre des délais dans la réalisation des services et la formation de file des clients désirant accéder à une ressource indisponible . L'analyse théorique de tels systèmes permet d'établir à l'avance les performance de l'ensemble, d'identifier les éléments critiques ou, encore, d'appréhen les effets d'une modification des condition de fonctionnement . Dans ce chapitre on va présenter quelques notions fondamentals à l'étude des files d'attente, puis on va présenté une étude approfondi des systèmes de files d'attente $M/M/1$ et $M/G/1$.

1.2 Description du modèle d'attente classique :

Une file d'attente est un système dans lequel arrivent des clients auquel des serveurs fournissent un service. Ce formalisme peut être utilisé dans des situations diverses : guichet, traitement des instructions par un processeur, gestion de communications téléphoniques, etc. Le modèle général d'un phénomène d'attente peut être résumé comme suit : des clients arrivent suivant un processus quelconque à un intervalle du temps aléatoire pour acquérir un service auprès d'un serveur. A l'arrivée d'un client, si un dispositif de service (serveur)

est libre, il se dirige immédiatement vers ce dispositif où il est servi. Dans le cas contraire, le client prend place dans une file d'attente, sinon il quitte le système.La durée du service auprès de chaque serveur est aussi aléatoire.

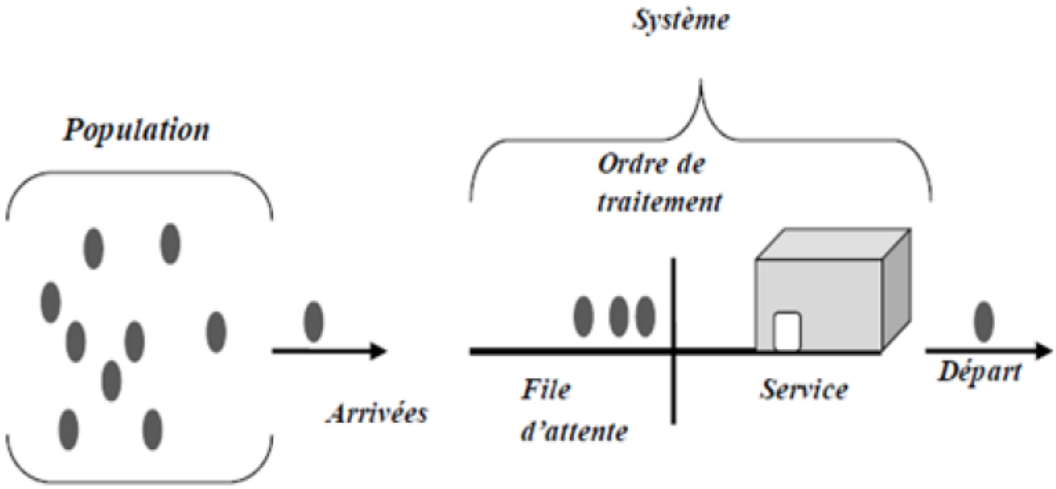

Fig1.1 : Modèle de file d'attente.

-Population : La population constitue la source de clients potentiels.elle est caractérisée par son nombre d'élément (fini ou infini).

-File d'attente : La file d'attente est caractérisée par le nombre maximum permis de clients en attente (fini ou infini).

-Clients : Les clients (issus de la population) se joignent au système avec un taux moyen d'arrivée.

-Service : Le service peut être assuré par un ou plusieurs serveurs. le temps qui s'écoule entre le début et la fin de service d'un client est dénoté,

le temps de service suivant une loi de probabilité.Le taux de service est une autre caractéristique du système.

-Stratégie de service : La stratégie de service réfère à l'ordre selon laquelle les clients sont servis : premier arrivée premier servi, au hasard,selon des priorités.

1.3 Analyse mathématique d'un système de files d'attente

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

-Les probabilités d'état $\pi_n(t) = p(N(t) = n)$, qui définissent le régime transitoire du processus stochastique $\{N(t), t \geq 0\}$. Il est évident que les fonction $\pi_n(t)$ dépendent de l'état initial ou de la distribution intial du processus.

-Le régime stationnaire du processus stochastique est définir par :

$$\pi_n = \lim_{n\to\infty} \pi_n(t) = p(N(+\infty) = n) = p(N = n \ , \ n = 0, 1, 2, 3, ..),$$

ou $\{\pi_n\}_{n\geq 0}$ est appelée distribution stationnaire du processus stochastique $\{N(t), t \geq 0\}$.

Le calcul explicite du régime transitoire s'avère généralement pénible , voire impossible, pour la plupart des modèles donnèes. On se contente donc de déterminer le régime stationnaire.

1.3.1 Processus stochastique

Le processus d'état stochastique $\{N(t) : t \geq 0\}$ est un processus de naissance et de mort si pour chaque $n = 0, 1, 2..$il existe des paramètres tels que,

Lorsque le système est dans l'état n, le processus d'arrivée est poissonnien de taux λ et le processus de sortie est poissonnien de taux μ.

Un processus stochastique $\{N(t) \ , \ t \in T\}$ est une fonction du temps dont la valeur à chaque instant dépend de l'issue d'une expérience aléatoire.

A chaque instant $t \in T$, $N(t)$ est donc une variable aléatoire.

Un processus stochastique peut être considéré comme une famille de variables généralement non indépendantes. L'ensemble des temps T peut être discret ou

continu. $N(t)$ définit l'état du processus à un instant donné t ,on classifie les processus

stochastiques de la façon suivante :

-Processus à temps discret et à espace d'état discret.

-Processus à temps continu et à espace d'état discret.

-Processus à temps discret et à espace d'état continu.

-Processus à temps continu et à espace d'état continu.

1.3.2 Processus de naissance et de mort

Rappelons que l'état d'un système de file d'attente à l'instant t, noté $N(t)$ est simplement le nombre de clients présents dans le système à l'instant t. L'ensemble des variables aléatoires d'état décrit un processus stochastique $\{N(t), t \geq 0\}$. Considérons à présent un système très général, dans lequel nous ferons abstraction (du moins, à première vue) des caractéristiques telles que nombre de serveurs,capacité,... etc. On peut utilement visualiser un tel système comme une "boîte noire", simplement caractérisée par un processus d'arrivée, un processus de sortie et un processus d'état résultant de la combinaison des arrivées et des départs :à chaque instant $t + \Delta t$, l'état $N(t + \Delta t)$ du système résulte des arrivées et sorties enregistrées entre t et $t + \Delta t$. Sans être parfaitement rigoureuse, la définition suivante permet d'introduire les caractéristiques d'un tel système auxquelles nous allons nous intéresser.

définition :

Le processus d'état stochastique $\{N(t) : t \geq 0\}$ est un processus de naissance et de mort si, pour chaque $n = 0, 1, 2, ...$, il existe des paramètres λ_n et μ_n(avec $\mu_0 = 0$) tels que, lorsque le système est dans l'état n, le processus d'arrivée est poissonnien de taux λ_n et le processus de sortie est poissonnien de taux μ_n. Dans un processus de naissance et de mort, les taux d'arrivée et de service sont donc des variables en fonction de l'état du système.

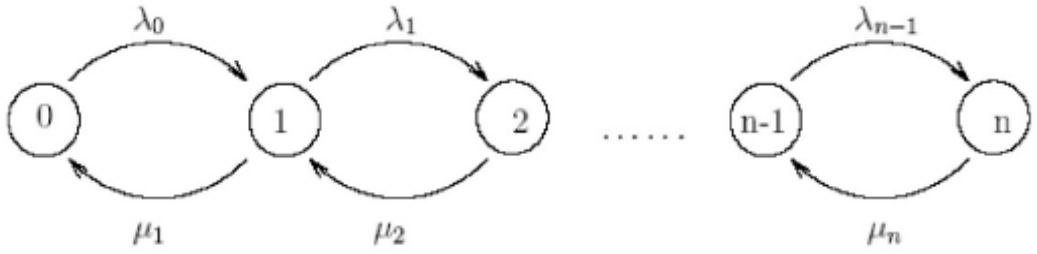

Fig1.2 : Graphe de transition d'un processus de naissance et de mort.

Ces processus permettent de façon générale de décrire l'évolution temporelle de la taille d'une population d'un type donné. Dans le cas d'un système d'attente,on considère par exemple des populations comprenant tous les clients qui sont dans le système à l'instant t. Les processus de naissance et de mort sont des processus stochastiques à temps continu et à espace d'états discret $n = 0, 1, 2, ..$ Ils sont caractérisés par deux conditions importantes : ils sont sans mémoire,et à partir d'un état donné n des transitions ne sont possibles que vers l'un ou l'autre des états voisins $(n+1)$ et $(n-1)$ pour $n \geq 1$. Alors, soit $\{N(t), t \geq 0\}$ un processus de naissance et de mort à états discrets et homogène dans le temps, c'est-à-dire : $p(N(t+s) = j/N(s) = i)$ ne dépende pas de s . Ce processus est de naissance et de mort si :

$$p_{i,i+1}(\Delta t) = \lambda_i(\Delta t) + o(\Delta t) \qquad i \geq 0, \tag{1.1}$$

$$p_{i,i-1}(\Delta t) = \mu_i(\Delta t) + o(\Delta t) \qquad i \geq 1, \tag{1.2}$$

$$p_{i,i}(\Delta t) = 1 - (\lambda_i + \mu_i)\Delta t + o(\Delta t) \quad i \geq 0, \tag{1.3}$$

$$p_{0,0}(\Delta t) = 1 - \lambda_i(\Delta t) + o(\Delta t), \tag{1.4}$$

$$p_{i,j}(\Delta t) = o(\Delta t) \quad |i - j| \geq 2, \tag{1.5}$$

$$p_{i,j}(0) = \delta_{ij} = \begin{cases} 1 & si \quad i = j \\ 0 & si \quad i \neq j \end{cases}, \tag{1.6}$$

λ_i : taux de naissance,

μ_i :taux de mort.

Processus des arrivées (processus de comptage "poisson") :

Un processus de comptage est une fonction $N(t)$ telle que :

1. $N(t) \geq 0$, pour t $\in \mathbb{R}$,
2. $N(t)$ est entier pour t $\in \mathbb{R}$,
3. Si $s \leq t$ alors $N(s) \leq N(t)$ peu importe {s, t $\in \mathbb{R}$}.

On dit d'un processus de comptage qu'il a des incréments indépendants si le nombre d'événements de toute paire d'intervalles de temps disjoints sont statistiquement indépendants, on dit d'un processus de comptage qu'il a des incréments stationnaires si le nombre d'événements dans un intervalle de temps ne dépend que de la longueur de l'intervalle, pour tout nombre Δt suffisamment petit :

$$\begin{aligned} p(N(t+\Delta t) - N(t) &= 0) = 1 - \lambda \Delta t, \\ p(N(t+\Delta t) - N(t) &= 1) = \lambda \Delta t, \\ p(N(t+\Delta t) - N(t) &> 1) = 0. \end{aligned}$$

Le paramètre λ indique le nombre moyen de clients qui arrivent par unité de temps.

Nous notons $N(t)$ le nombre d'arrivées de clients survenues dans l'intervalle de temps $[0, t]$, pour $t \geq 0$. La quantité $N(a+t) - N(a)$ représente alors le nombre d'arrivées enregistrées entre les instants a et $a+t$, pour tout $a \geq 0$ et $t \geq 0$. En règle générale, pour chaque $t \geq 0$, $N(t)$ est une variable aléatoire. L'ensemble de ces variables aléatoires fournit une représentation mathématique, c'est-à-dire un modèle, des arrivées de clients dans le système. processus (stochastique) d'arrivée cet ensemble $\{N(t) : t \geq 0\}$.

Le processus d'arrivée peut bien sûr présenter des caractéristiques variées en fonction de la situation modélisée. Mais il est fréquent dans la pratique ,que ce processus soit un processus de Poisson, ce qui signifie qu'il existe un paramètre $\lambda \geq 0$ (appelé taux du processus) tel que :

1. Le nombre d'arrivées dans tout intervalle $[a, a+t]$ de longueur t suit une loi de Poisson de moyenne λt , c'est-à-dire : pour $a, t \geq 0$ et $n = 0, 1, 2, ...$

$$p_n(t) = \frac{(\lambda t)^n}{n!} e^{-\lambda t}. \tag{1.7}$$

2. Si $[a, b]$ et $[c, d]$ sont des intervalles de temps disjoints, alors le nombre d'arrivées dans $[a, b]$ est indépendant du nombre d'arrivées dans $[c, d]$.

3. $N(0) = 0$.

Processus de service (exponentielle)

Rappelons que la distribution exponentielle est une loi continue dont les fonctions de densité et de répartition sont :

$$f(x) = \lambda e^{-\lambda x} \quad x \geq 0, \tag{1.8}$$

$$F(x) = 1 - e^{-\lambda x} \quad x \geq 0. \tag{1.9}$$

En particulier, si x suit une loi exponentielle et que $N(t)$ est un processus de Poisson, alors :

$$\begin{aligned} p(x \geq t) &= 1 - F(t) \\ &= e^{-\lambda t} = p(N(t) = 0). \end{aligned} \tag{1.10}$$

On en conclut donc que les temps entre les arrivées suivent des lois exponentielles indépendantes de moyenne $1/\lambda$.

1.4 Classification des systémes d'attente

Pour identifier un systéme d'attente, on a besoin des spécifications suivantes :

– La nature stochastique du processus des arrivées, qui est défini par la distribution des intervalles séparant deux arrivées consécutives .

– La distribution du temps aléatoire de service .

– Le nombre m de serveurs (stations de service). On admet généralement que les temps de service correspondants suivent la même distribution et que les clients qui arrivent forment une seule file d'attente.

– La capacité N du systéme. Si $N < \infty$, la file d'attente ne peut dépasser une longueur de $N-m$ unités. Dans ce cas, certains clients arrivant vers le systéme n'ont pas la possibilité d'y entrer.

1.4.1 Notation de Kendall (1953)

Un modéle des files d'attente est totalement décrit selon la notation de Kendall. Dans sa version étendue, un modéle est spécifié par une suite de six symboles :

$$A/B/C/D/E/F.$$

La signification de chacun de ces symboles est :

A : nature du processus des arrivées.

B : nature du processus de service.

C : nombre de serveurs.

D : capacité du systéme.

E : taille de la population (la source).

F : discipline de la file.

Dans la description des processus d'arrivée et de service, les symboles les plus courants sont :

M : loi Exponentielle (memoryless).

E : loi d'Erlang.

Γ : loi Gamma .

D : loi Déterministe (temps d'inter-arrivées ou de service constant).

G : loi Générale (quelconque).

La forme abrégé : $A/B/C$ signifie que D et E sont infinies.

1.4.2 Les diférentes disciplines de service

La discipline de service décrit l'ordre avec lequel les arrivées dans le système vont accéder au service. Ces disciplines sont :

$FIFO$ (First In First Out) : Le premier arrivée est le premier servi.

$LIFO$ (Last In First Out) : Le dernier arrivé sera le premier servi.

Random (aléatoire) : Les clients accèdent au serveur de manière aléatoire, indépendamment de l'ordre des arrivées.

Exemples des files d'attente classiques :

1. Modèle $M/M/1$: Les clients se présentent suivant un processus de Poisson.Le temps de service suit une loi exponentielle de taux μ,indépendamment d'un client à un autre. La file d'attente peut s'étendre à l'infini. Le cas d'un guichet de poste avec un seul serveur.
2. Modèle $M/M/\infty$: Les arrivées suivent une loi de Poisson Les temps de service suivent une loi Exponentielle. Le nombre de serveurs est infini : Les clients sont traités simultanément et indépendamment. On les retrouve généralement dans le cas des réseaux téléphoniques.
3. Modèle $M/G/1$: Les arrivées suivent une loi de Poisson. Les temps de service sont arbitrairement distribuées suivant une loi Générale .

1.5 Mesures de performance d'une file d'attente

L'étude d'une file d'attente a pour but de calculer ou d'estimer les performances d'un système dans des conditions de fonctionnement données. Ce calcul se fait le plus souvent pour le régime stationnaire uniquement et les mesures les plus fréquemment utilisées sont :

$L = E(X)$: nombre moyen de clients dans le système,

L_q : nombre moyen de clients dans la file d'attente,

W : temps moyen de séjour d'un client dans le système,

W_q : temps moyen d'attente d'un client dans la file.

Ces valeurs ne sont pas indépendantes les unes des autres, mais sont liées par les relations suivante :(formule de little)

$$\begin{aligned} L &= \lambda W, && \text{où } \lambda \text{ représente le taux d'arrivées.} \\ L_q &= \lambda w_q, \\ w &= w_q + \frac{1}{\mu}, && \text{où } \mu \text{ représente le taux de service.} \\ L &= L_q + \frac{\lambda}{\mu}. \end{aligned}$$

D'une manière générale, une file est stable si et seulement si le nombre moyen d'arrivées de clients par unité de temps, noté λ, est inférieur au nombre moyen de clients pouvant être servis par unité de temps. Si chaque serveur peut traiter μ clients par unité de temps et si le nombre de serveurs est s,une file est stable si et seulement si :

$$\lambda < s\mu \Leftrightarrow \rho = \frac{\lambda}{s\mu} < 1$$

où, ρ est appelé l'intensité du trafic.

1.6 Les files d'attente markoviennes

Ils caractérisent les systèmes dans lesquels les deux quantités stochastiques principales qui sont le temps des inter-arrivées et la durée de service sont des variables aléatoires indépendantes exponentiellement distribuées (modèle $M/M/1$). La propriété d'absence de mémoire de la 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$\{N(t), t \geq 0\}$ défini comme étant le nombre de clients dans le

système à l'instant t. L'évolution temporelle du processus markovien $\{N(t), t \geq 0\}$ est complètement définie grâce à la propriété d'absence de mémoire.

1.6.1 Modèle d'attente $M/M/1$

Le système de files d'attente $M/M/1$ est le système le plus élémentaire de la théorie des files d'attente. Le flot des arrivées est poissonnien de paramètre λ et la durée de service est exponentielle de paramètre μ.

Régime transitoire

Soit $N(t)$ le nombre de clients présents dans le système à l'instant t ($t \geq 0$). Grâce aux propriétés fondamentales du processus de Poisson et de la loi exponentielle $N(t)$ est un processus markovien homogène. Les probabilités d'état $P_n(t) = P[X(t) = n]$ peuvent être calculées par les équations différentielles de Kolmogorov ci-dessous, connaissant les conditions initiales du processus. D'aprés la formule des probabilités totales :

$$\begin{aligned}
P_n(t+\Delta t) &= P[N(t+\Delta t) = n], \\
&= \sum_{i=0}^{\infty} P_i(t) P_{i\ n}(\Delta t), \\
&= P_{n-1}(t) P_{n-1\ n}(\Delta t) + P_n(t) P_{n\ n}(\Delta t) + P_{n+1}(t) P_{n+1\ n}(\Delta t) + o(\Delta t), \\
&= P_{n-1}(t)\lambda_{n-1}(\Delta t) + P_n(t)(1 - (\lambda_n - \mu_n)\Delta t) + P_{n+1}(t)\mu_{n+1}(\Delta t) + o(\Delta t), \\
&= P_{n-1}(t)\lambda_{n-1}(\Delta t) + P_n(t) - P_n(t)(\lambda_n - \mu_n)(\Delta t) + P_{n+1}(t)\mu_{n+1}(\Delta t) + o(\Delta t), \\
P_n(t+\Delta t) - P_n(t) &= P_{n-1}(t)\lambda_{n-1}(\Delta t) - P_n(t)(\lambda_n - \mu_n)(\Delta t) + P_{n+1}(t)\mu_{n+1}(\Delta t) + o(\Delta t), \\
\frac{P_n(t+\Delta t) - P_n(t)}{\Delta t} &= P_{n-1}(t)\lambda_{n-1} - P_n(t)(\lambda_n - \mu_n) + P_{n+1}(t)\mu_{n+1} + \frac{o(\Delta t)}{\Delta t},
\end{aligned}$$

Lorsque $\Delta t \longrightarrow 0$ alors :

$$P_n^{'}(t) = \lambda_{n-1} P_{n-1}(t) - (\lambda_n - \mu_n) P_n(t) + \mu_{n+1} P_{n+1}(t), \forall n \geq 1.$$

De la mème maniere, on obtient :

$$P_0^{'}(t) = -\lambda_0 P_0(t) + \mu_1 P_1(t).$$

D'ou le système d'équation de Chapmain-kolmogorov et :

$$\begin{cases} P_n'(t) = -(\lambda_n + \mu_n)P_n(t) + \lambda_{n-1}P_{n-1}(t) + \mu_{n+1}P_{n+1}(t) \\ P_0'(t) = -\lambda_0 P_0(t) + \mu_1 P_1(t) \end{cases}, \tag{1.11}$$

Sous la condition d'ergodicité du système $\rho = \frac{\lambda}{\mu} < 1$, pour laquelle le régime stationnaire existe, il est aisé d'obtenir les probabilités stationnaires

$$P_n = \lim_{t\to\infty} p_n(t) = (1-\rho)\rho^n \ , \ \forall n \in N \tag{1.12}$$

$P = \{P_n\}\ n \geq 0$ est appelé distribution stationnaire, elle suit une loi géométrique.

Caractéristiques du système -Le nombre moyen de clients dans le système est :

$$L = E(x) = \sum_{n\geq 0} nP_n = (1-\rho)\sum_{n\geq 0} n\rho^n,$$

D'où :

$$L = \frac{\rho}{1-\rho}. \tag{1.13}$$

-Le nombre moyen de clients dans la file :

$$L_q = \sum_{n\geq 1}(n-1)P_n = \frac{\rho^2}{1-\rho}. \tag{1.14}$$

Le temps moyen de séjour dans le système W et le temps moyen d'attente dans la file W_q sont obtenus à partir des formules de Little :

-Le temps moyen de séjour dans le système :

$$W = \frac{\rho}{\lambda(1-\rho)}. \tag{1.15}$$

-Le temps moyen d'attente dans la file :

$$W_q = \frac{\rho^2}{\lambda(1-\rho)}. \tag{1.16}$$

1.7 Les files d'attente non-markoviennes

En l'absence de l'exponentialité ou plutôt lorsque l'on s'écarte de l'hypothèse d'exponentialité de l'une des deux quantités stochastiques : le temps des inter-arrivées et la durée de service, ou en prenant en compte certaines spécificités des problèmes par introduction de paramètres supplémentaires,on aboutit à un modèle non markovien. La combinaison de tous ces facteurs rend l'étude mathématique du modèle très délicate. On essaye alors de se ramener à un processus de Markov judicieusement choisi à l'aide de l'une des méthodes d'analyse suivantes :

Méthode des étapes d'Erlang : Son principe est d'approximer toute loi de probabilité ayant une transformée de Laplace rationnelle par une loi de Cox (mélange de lois exponentielles),cette derniére posséde la propriété d'absence de mémoire par étapes.

Méthode de la chaîne de Markov induite : Cette méthode élaborée par Kendall , est souvent utilisée. Elle consiste à choisir une séquence d'instants $1, 2, 3, ...n$ (déterministes ou aléatoires) telle que la chaîne induite $\{N_n \ ; n \geq 0\}$, ou $N_n = N(n)$ soit markovienne et homogéne.

Méthode des variables auxiliaires : Elle consiste à compléter l'information sur le processus $\{N(t) \ ; t \geq 0\}$ de telle maniére à lui donner le caractére markovien. Ainsi, on se raméne à l'étude du processus :

$\{N(t), A(t_1), A(t_2), ...A(t_n)\}$ Les variables $A(t_k)$, $k \in \{1, 2, 3...n\}$ sont dites auxiliaires.

Méthode des événements fictifs : Le principe de cette méthode est d'introduire des événements fictifs qui permettent de donner une interprétation probabiliste aux transformées de Laplace et aux variables aléatoires décrivant le systéme étudié.

Simulation : C'est un procédé d'imitation artificielle d'un processus réel donné sur ordinateur.Elle nous permet d'étudier les systémes les plus complexes,de prévoir leurs com-

portements et de calculer leurs caractéristiques. Les résultats obtenus ne sont qu'approximatifs, mais peuvent être utilisés avec une bonne précision.Cette technique se base sur la génération de variables aléatoires suivant les lois gouvernant le systéme.

1.7.1 Loi d'erlang :

Considerant une variable aléatoire X de densité de probabilité :

$$\begin{aligned} f(X) &= \frac{1}{\Gamma(\alpha)\ \beta^{\alpha}}\ X^{(\alpha-1)}\ e^{-\frac{X}{\beta}}. \\ \Gamma(\alpha) &= \int_0^{\infty} X^{(\alpha-1)}\ e^{-\alpha}\ dX. \end{aligned} \tag{1.17}$$

α, β sont les paramétres de la distribution de probabilité on a

$$\begin{aligned} E(X) &= \alpha\beta, \\ var\ X &= \alpha\beta^2. \end{aligned}$$

Exemple 1 Considerant un cas particulier de cette distribution ou :

$$\begin{cases} \alpha = k & tq \quad k \in N^* \\ \beta = \frac{1}{k\mu} & \mu = cste > 0 \end{cases},$$

on obtient la famille de distribution de probabilité d'erlang :

$$f(X) = \frac{(k\mu)^k}{(k-1)!}\ X^{(k-1)}\ e^{-k\ \mu\ X}\ , \qquad 0 < X < \infty \tag{1.18}$$

k, μ sont les paramétres de la distribution d'erlang ;

$$\begin{aligned} E(X) &= \alpha\beta = k\frac{1}{k\mu} = \frac{1}{\mu}, \\ var\ X &= \alpha\beta^2 = k\frac{1}{k^2\mu^2} = \frac{1}{k\mu^2}. \end{aligned}$$

La distribution d'erlang qui correspande a une valeur particuliaire E_K est appelé erlang de type (ordre) k.

Remarques : la somme de k variable aléatoire exponentiel indépandant identiquement distribué avec une moyenne de $\frac{1}{k\mu}$ est une distribution d'erlang de type k avec une moyenne $\frac{1}{\mu}$ cette relation nous permet de d'ecrire les systéme d'attente ou le service peut étre divisé ou plusieur étapes identique :

$\left[\begin{matrix}\text{1ere étape}\\ \frac{1}{k\mu}\end{matrix}\right]$.....$\left[\begin{matrix}\text{2eme étape}\\ \frac{1}{k\mu}\end{matrix}\right]$.....$\left[\begin{matrix}\text{k eme étape}\\ \frac{1}{k\mu}\end{matrix}\right]$

est le service globale est une d'erlang de type k.

Il faut noté que :

1. Tout les étape (phase) de service sont indépandants et identiques.
2. Chaque fois on a un seul client dans le service c-a-d c'est on a un client qui rentre dans la premier phase de service une fois qu'il termine, il se dirige vers une autre phase est un autre client qui est en attente ne peut pas axider a la premier phase jusqu'a se client qui est en service quite complètement le systéme.

1.7.2 Modèle d'attente $M/G/1$

Le flux des arrivées dans le systéme $M/G/1$ est poissonnien de paramétre λ et la durée de service est distribuée selon une loi générale G de moyenne $1/\mu$. La particularité de ce systéme est que, contrairement au cas $M/M/1$, le processus $N(t)$ n'est pas markovien.

Chaîne de Markov induite et probabilités de transition : Soit N_n le nombre de clients dans le système M/G/1 à la fin de service du n^{ieme} client.Notons par G(s) la

distribution de la durée de service et par le paramètre μ de la distribution exponentielle régissant la duréee entre deux arrivées consécutives.Le processus $\{N_n, n \geq 0\}$ est une chaîne de Markov, d'opérateur de transition $P = [p_{ij}]$ $i \geq 0, j \geq 1$où :

$$p_{ij} = \begin{cases} p_j & si \quad i = 0 \\ p_{j-i+1} & si \quad j+1 \geq i \geq 1 \\ 0 & sinon \end{cases},$$

avec

$$p_k = \int_0^\infty \frac{e^{-\lambda s}(\lambda s)^k}{k!} dG(s) \ , \quad k = 1, 2, 3, ... \tag{1.19}$$

En effet, si A_n est le nombre de clients qui entrent dans le systéme pendant le $n^{iéme}$ service, on a :

$$N_{n+1} = N_n - \delta_n + A_{n+1} \qquad avec \qquad \delta_n = \begin{cases} 1 & \text{si} \quad N_n > 0, \\ 0 & \text{si} \quad N_n = 0. \end{cases}$$

Ceci montre que N_{n+1} ne d´epend que de N_n et de A_{n+1} et non pas de N_{n-1}, N_{n-2}.... Ce qui signifie que la suite $\{N(t) \ , \ t \geq 0\}$est markovienne, où $N(t)$ est le nombre de clients dans le systéme à l'instant t. Car le nombre de clients A_n qui entrent dans le systéme, est distribué suivant une loi de Poisson de paramétre λt. Et d'aprés le théoréme des probabilités totales :

$$p(A_n = k) = p_k = \int_0^\infty \frac{e^{-\lambda t}(\lambda t)^k}{k!} dG(t) \ , \qquad \text{oÙ} \qquad 1 > \ p_k > 0 \qquad (k = 1, 2, ...)$$

La suite $\{N(t) \ , \ t \geq 0\}$ est une chaîne de Markov induite du processus. Ces probabilités de transition ;

$$p_{ij} = p(N_{n+1} = j / N_n = i),$$

se calcule par :

$$\begin{cases} p_{0j} = p_j & si \quad j \geq 0 \\ p_{ij} = p_{j-i+1} & si \quad 1 \leq \ i \leq j+1 \\ p_{ij} = 0 & \text{sinon} \end{cases},$$

la matrice des transitions et

$$p = \begin{bmatrix} a_0 & a_1 & a_2 & a_3 & .. & .. \\ a_0 & a_1 & a_2 & a_3 & .. & .. \\ 0 & a_0 & a_1 & a_2 & .. & .. \\ 0 & 0 & a_0 & a_1 & .. & .. \\ .. & .. & .. & .. & .. & .. \end{bmatrix}.$$

Régime stationnaire Le régime stationnaire du systéme existe et est identique à l'état stationnaire de la chaÎne de Markov induite N_n, si $\rho = \frac{\lambda}{\mu} < 1$. Il ne sera généralement pas possible de trouver la distribution stationnaire $P = (p_0, p_1, ...)$. Cependant, nous pouvons calculer la fonction génératrice correspondante $P(z)$:

$$P(z) = G^*(\lambda - \lambda z)\frac{(1-\rho)(1-z)}{G^*(\lambda - \lambda z) - z'}. \tag{1.20}$$

où G^* représente la transformée de Laplace de la densité de probabilité du temps de service, et z est un nombre complexe vérifiant $|z| \leq 1$. La formule est appelée formule de Pollaczek-Khintchine.

Caractéristiques du systéme On note λ le taux d'arrivée des clients. Cela signifie que l'espérance de la durée séparant deux arrivées successives est $E(X) = 1/\lambda$. On note μ le taux de service des clients. Cela signifie que l'espérance de la durée de service est $E(Y) = 1/\mu$.

L'intensité du trafic s'exprime de la maniére suivante :

$$\rho = \frac{\lambda}{\mu} = \frac{E(Y)}{E(X)}.$$

ou X est la loi des inter-arrivées et Y est la loi de service.

-Le nombre moyen de clients dans le systéme : Cette quantité peut être déterminée, en régime stationnaire, en utilisant la relation :

$$E(X) = \lim_{z \to 1} P'(z)$$

Néanmoins, ce calcul s'avére compliqué. Par contre, elle peut être obtenue aisément en utilisant la relation

$$l = E(X_n) = \rho + \frac{\rho^2 + \lambda^2 v(y)}{2(1-\rho)}, \tag{1.21}$$

où $V(Y)$ est la variance de la variable aléatoire Y.

-Le nombre moyen de clients dans la file est :

$$l_q = \frac{\rho^2 + \lambda^2 v(y)}{2(1-\rho)}, \tag{1.22}$$

en utilisant la formule de Little, on obtient :

-Le temps moyen de séjour dans le systéme :

$$w = \frac{1}{\mu} + \lambda \left(\frac{v(y) + \frac{1}{\mu^2}}{2(1-\rho)} \right), \tag{1.23}$$

-Le temps moyen d'attente dans la file :

$$w_q = w - \frac{1}{\mu} = \lambda \left(\frac{v(y) + \frac{1}{\mu^2}}{2(1-\rho)} \right). \tag{1.24}$$

1.8 Conclusion

Dans ce chapitre nous nous sommes intéressés à quelques notions de base sur la théorie de file d'attente. Ce chapitre est consacré à la présentation des systémes d'attente classiques répartis en deux sections différentes à savoir ; les systémes markoviens et les systémes non markoviens ainsi que leurs caractéristiques.

Les modéles d'attente développées ces derniéres décennies tentant de prendre en considération des phénoménes de répétition de demande de service comme dans les résaux téléphonique. Ces phénoménes affectent les caractéristique de performance des systémes réel. Il s'agit donc des systémes de files d'attente avec rappels.

Chapitre 2

Modèles 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. Les systèmes de files d'attente développés tentent de prendre en considération des phénomènes de répétition de demandes de service, et ceci après une durée du temps aléatoire. Un tel système est connu comme «système de files d'attente avec rappels». Son étude est motivée par diverses applications pratique dans le domaine de télécommunication voir : falin et templeton(1997)[16], Pour identifier un système de files d'attente avec rappels, on a besoin des spécifications suivantes : 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 discipline d'attente ainsi que la spécification concernant le processus de répétition d'appels. Ce chapitre est consacré à l'étude des files d'attente $M/M/1$ et $M/G/1$ avec rappels.

2.2 Description du modèle d'attente avec rappels

Un système d'attente avec rappels (Retrial Queue) est un système composé de c, $(c \geq 1)$ serveurs identiques et indépendants, et d'un buffer (file) de capacité $K - c$, $(K \geq c)$ et d'une orbite de capacité j. À l'arrivée d'un client, 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 dans la file, le client la rejoindra. Par

ailleurs, si un client arrive et trouve tous les serveurs et toutes les positions d'attente de la file occupés, il quittera le système définitivement avec la probabilité $1 - H_0$, ou bien entre en orbite avec la probabilité H_0 et devient une source d'appels répétés et tentera sa chance après une durée de temps aléatoire. Les clients qui reviendront et rappelleront pour le service sont dits en "orbite". Cette dernière peut être finie ou infinie. Dans le cas d'une orbite à capacité finie, si elle est pleine, un client qui trouve tous les serveurs et les positions d'attente de la file occupés, sera obligé de quitter le système définitivement sans être servi.chaque client en orbite appelé aussi «client secondaire», est supposé rappeler pour le service à des intervalles de temps 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 c'est-à-dire un nouveau client qui arrive de l'extérieur du système. s'il trouve un serveur libre,il sera servi immédiatement puis quittera le système.Sinon, s'il y a des positions d'attente disponibles dans la file, il le rejoindra.Par contre, si tous les serveurs et les positions d'attente sont encore occupés,le client quittera le système pour toujours avec la probabilité $1 - H_k$ (si c'est le $k^{éme}$ rappel sans succès) ou bien entre en orbite avec la probabilité H_k si l'orbite n'est pas pleine. Le schéma général d'un système d'attente avec rappels est donné par la Fig(2.1) :

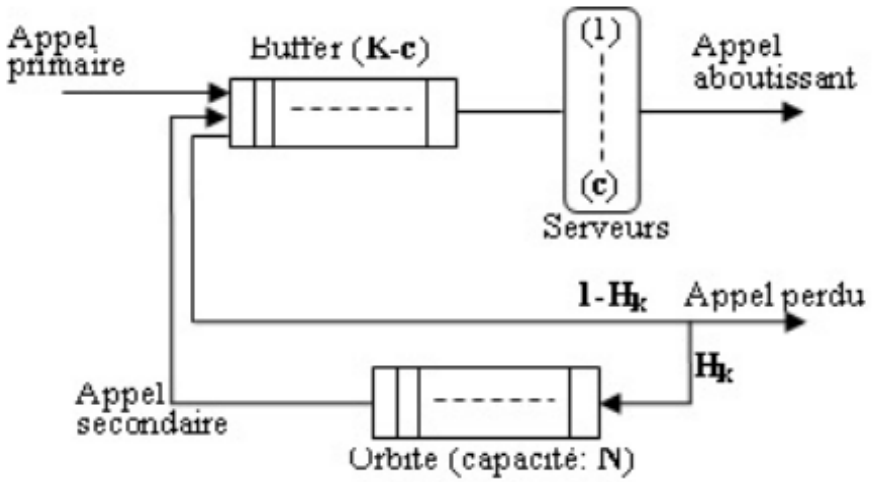

Fig2.1 : Schéma illustratif d'un système d'attente avec rappels.

Remarques :

1. Le modèle d'attente avec rappels décrit ci-dessus est un modèle général. Plusieurs systèmes de files d'attente avec rappel peuvent être considérés comme des cas particuliers tels que : les systèmes sans files, les systèmes à un seul serveur, ...
2. La description d'un système de files d'attente ordinaire (classique) se fait avec ses éléments principaux : le processus d'arrivées, le mécanisme de service (disponibilité et nombre de serveurs), et la discipline d'attente. Pour un système avec rappels, on doit ajouter un élément décrivant la loi des répétitions d'appels. En fonction du modèle considéré,on pourra introduire d'autres éléments décrivant la fiabilité du serveur, les types de priorité, ...
3. Les clients primaires ou secondaires qui arrivent durant un temps de service, entrent en orbite sans aucune influence sur le processus de service.

2.3 Politiques d'accès au serveur à partir de l'orbite

La définition du protocole de rappels est en effet un sujet de controverse (voir : Falin 1990)[17] et concerne l'aspect modélisation du système sous étude. Il existe dans la littérature des modéles avec rappels essentiellement trois politique de rappels : La politique classique, la politique constante et la politique linèaire (versatile).

2.3.1 Politique de rappels classique

Le protocole le plus décrit dans la théorie classique des files d'attente avec rappels est la politique de rappels classiques dans laquelle chaque source dans l'orbite rappelle après un temps exponentiellement distribué avec un paramètre θ.

Donc, il y a une probabilité $j\theta dt + o(dt)$ d'un nouveau rappel dans le prochain intervalle $(t, t+dt)$ sachant que j clients sont en orbite à l'instant t. Une telle politique a été motivée par des applications dans la modélisation du comportement des abonnés dans les réseaux téléphoniques depuis les années 1940.

2.3.2 Politique de rappels constante

Dans les années précédentes, la technologie a considérablement évoluée. La littérature de files d'attente avec rappels décrit différents protocoles de rappels spécifiques à certains réseaux informatiques, et de communication modernes dans lesquels le temps inter-rappels est contrôlé par un dispositif électronique et par conséquent, est indépendant du nombre d'unités demandant le service. Dans ce cas, la probabilité d'un rappel durant $(t, t+dt)$,sachant que l'orbite est non vide, est $\alpha dt + o(dt)$. Ce type de discipline de rappels est appelé politique de rappels constants. Le premier travail dans cette direction est celui de Fayolle qui considère une file d'attente $M/M/1$, où uniquement le client en tête de la file en orbite peut demander un service après un temps de rappels exponentiellement distribué avec un taux constant.

Cette sorte de politique de contrôle de rappels est bien connue pour le protocole ALOHA

dans les systèmes de communication. Certains autres travaux décrivent des applications aux réseaux locaux, protocole de communication,systèmes mobiles et autres (Choi 1992)[12], (Shikata 1999). Artalejo et al (2002)[5].

2.3.3 Politique de rappels linèaire (versatile)

Gómez-Corral (1997)[4] traitent les deux cas d'une manière unifiée en définissant une politique de rappels linéaires pour laquelle la probabilité d'un rappel durant $(t, t + dt)$ sachant que j client sont en orbite à l'instant t est $(\alpha(1-\delta_{0j})+j\theta)dt+o(dt)$. On mentionne aussi l'existence d'une autre politique dite politique de rappels quadratiques .

2.4 Modèle $M/M/1$ avec rappels

On considère un système de files d'attente sans positions d'attente. Le service est assuré par 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 ;

$$B(x) = 1 - e^{-\mu x}, x \geq 0.$$

et de moyenne finie $1/\mu$. Les temps entre deux rappels consécutifs sont également exponentiels de paramètre $\theta > 0$, la fonction de répartition ;

$$T(x) = 1 - e^{-\theta x}, x \geq 0.$$

Nous admettons que les durées de service, les durées entre deux rappels consécutifs ainsi que entre deux arrivées primaires successives sont mutuellement indépendantes. L'état du système peut être décrit par le processus

$$\{C(t), N_0(t), t \geq 0\},$$

où $C(t)$ est égale à 0 ou 1 selon le fait que le serveur est libre ou non, $N_o(t)$ est le

nombre de clients en orbite l'instant t.

Supposons que le régime stationnaire existe ($\rho = \frac{\lambda}{\mu} < 1$) Le processus est de Markov d'espace d'états $S = \{0, 1\} \times \mathbb{N}$.

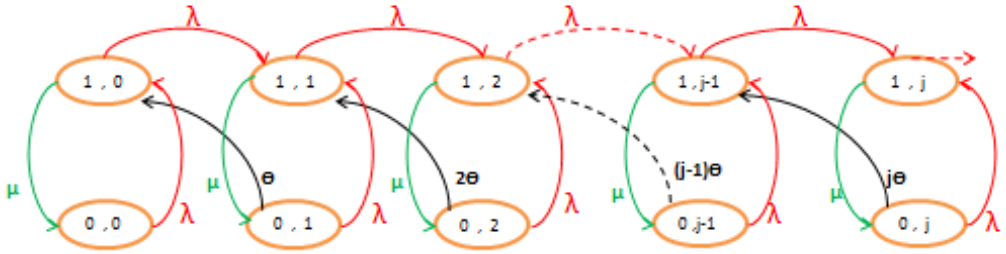

Fig2.2 : Graphe des transitions du modèle M/M/1 avec rappels.

Les équations d'équilibre statistique sont :

$$(\lambda + j\theta)p_{0\,j} = \mu p_{1\,j}, \tag{2.1}$$

$$(\lambda + \mu)p_{1\,j} = \lambda p_{0\,j} + (j+1)\theta p_{0\,j+1} + \lambda p_{1\,j-1}, \tag{2.2}$$

ici :

$$p_{ij} = \lim_{t\to\infty} P(C(t) = i, N_0(t) = j), \quad i = 0, 1 \; et \; j \geq 0,$$

les probabilités de transition en régime stationnaire sont définies par :

$$q_{i,j}(k,l) = \begin{cases} \lambda & si \;\; (k,l) = (i+1, j), \\ \mu & si \;\; (k,l) = (i-1, j), \\ j\theta & si \;\; (k,l) = (i+1, j-1), \\ -(\lambda + \mu + j\theta) & si \;\; (k,l) = (i,j) \\ 0 & \text{sinon} \end{cases} \tag{2.3}$$

représentent la distribution stationnaire conjointe de l'état du serveur et du nombre de clients en orbite.

Introduisons les fonctions génératrices suivantes :

$$P_0(z) = \sum_{n=0}^{\infty} z^j p_{0j}, \tag{2.4}$$

$$P_1(z) = \sum_{n=0}^{\infty} z^j p_{1j}. \tag{2.5}$$

A l'aide de ses fonctions et à partir des équations, on obtient les fonction génératrices partielles :

$$P_0(z) = (1-\rho)(\frac{1-\rho}{1-z\rho})^{\frac{\lambda}{\theta}}, \tag{2.6}$$

$$P_1(z) = \rho(\frac{1-\rho}{1-z\rho})^{\frac{\lambda}{\theta}+1}. \tag{2.7}$$

Les transformées inverses nous donnent les formules analytiques explicites :

$$p_{0j} = \frac{\rho}{j!\theta^j}\prod_{k=0}^{j-1}(1-k\theta)(1-\rho)^{\frac{\lambda}{\theta}+1}, \tag{2.8}$$

$$p_{1j} = \frac{\rho^{j+1}}{j!\theta^j}\prod_{k=1}^{n}(\lambda+k\theta)(1-\rho)^{\frac{\lambda}{\theta}+1}. \tag{2.9}$$

Par conséquent,la fonction génératrice de la distribution de la taille de l'orbite sera définie par :

$$P(z) = P_0(z) + P_1(z) = (1+\rho-\rho z)(\frac{1-\rho}{1-z\rho})^{\frac{\lambda}{\theta}+1}. \tag{2.10}$$

Le nombre moyenne de client en orbite est, la dérivé de $P(z)$ quand z tend vers 1 :

$$E(N_0(t)) = P'(1), \tag{2.11}$$

$$E(N_0(t)) = \frac{\rho(\lambda+\rho\mu)}{\mu(1-\rho)}. \tag{2.12}$$

Et celle de la distribution stationnaire du nombre de client dans le systéme s'obtiandra de la maniére suivante :

$$Q(z) = P_0(z) + zP_1(z) = (\frac{1-\rho}{1-z\rho})^{\frac{\lambda}{\theta}+1}. \tag{2.13}$$

On peut également trouver la distribution du nombre de serveur occupé :

$$p_0 = \lim_{t\to\infty} P(C(t)=0) = P_0(1) = 1-\rho, \tag{2.14}$$

$$p_1 = \lim_{t\to\infty} P(C(t)=1) = P_1(1) = \rho. \tag{2.15}$$

2.5 Modèle $M/G/1$ avec rappels

Le modèle M/G/1 avec rappels est le modèle le plus étudié par les spécialistes.

2.5.1 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 τ est de loi générale $P(\tau_n^s \leq x) = B(x)$ et de transformée de Laplace-Stieltjes e $\widetilde{B}(s)$, $Re(s) > 0$. Soient les moments $\beta_k = (-1)^k \widetilde{B}(k)(0)$,l'intensité du trafic $\rho = \lambda\beta_\iota$ et $\gamma = \frac{1}{\beta_1}$. La durée entre deux rappels successifs d'une même source secondaire est exponentiellement distribuée de paramètre $\theta > 0 : T(x) = P(\tau_n^r \leq x) = 1 - e^{\theta x}$.

Le système évolue de la manière suivante : On suppose que le $(n-1)^{ieme}$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^{ieme} client n'entre en service. A l'instant $\xi_n = \eta_n + R_n$ le n^{ieme} 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 $\xi_n = \eta_n + \tau_n^s$ le n^{ieme} client achève son service, le serveur devient libre et ainsi de suite.

2.5.2 Chaîne de Markov induite

Considérons le processus $\{C(t), N_0(t), t \geq 0\}$, où $C(t)$ représente l'état du serveur

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

et $N_o(t)$ est le nombre de clients en orbite à la date t. En général, ce processus n'est pas un processus de Markov, mais il possède une chaîne de Markov induite,Cette chaîne a été décrite pour la première fois par Choo et Conolly (1979)[13], Soit (q_n) la chaîne de Markov induite aux instants de départs, où $q_n = N_o(\xi n)$ représente le nombre de clients en orbite après le n^{ieme} départ, dont l'équation fondamentale est :

$$q_{n+1} = q_n - \delta_{q_n} + \nu_{n+1}, \tag{2.16}$$

où ν_{n+1} est le nombre des clients primaires arrivant dans le système durant le service du $(n+1)^{ieme}$ client. Elle ne dépend pas des événements qui se sont produits avant l'instant η_{n+1} (où l'instant 0 en faisant une translation) du début de service du $(n+1)^{ieme}$ client. La distribution de ν_{n+1} est la suivante :

$$P(\nu_n = i) = a_i = \int_0^\infty \frac{(\lambda x)^i}{i!} \exp(-\lambda x) dB(x), \tag{2.17}$$

où $a_i > 0$,$i \geq 0$. On a les résultats suivants si

$$\nu = \lim_{n\to\infty} \nu_n \ , \ E[\nu] = \rho \ ; \quad alors \ A(z) = \sum_0^\infty a_i z^i = \widetilde{B}(\lambda - \lambda z).$$

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

$$\delta_{q_n} = \begin{cases} 1 \quad \text{si} \quad \text{le } (n+1)^{ieme} \text{ client servi provient de l'orbite,} \\ 0 \quad \text{si} \quad \text{le } (n+1)^{ieme} \text{client servi est primaire.} \end{cases}.$$

Elle dépend de q_n et sa distribution est :

$$P(\delta_{q_n} = 1/q_n = i) = \frac{i\theta}{\lambda + i\theta'}, \tag{2.18}$$

$$P(\delta_{q_n} = 0/q_n = i) = \frac{i}{\lambda + i\theta}. \tag{2.19}$$

Les probabilités de transition de l'état i à l'état j , ($\forall j \geq 0$ et $0 \leq i \leq j$) sont :

$$r_{ij} = P(q_{n+1} = j/q_n = i) = a_{j-i}\frac{\lambda}{\lambda + i\theta} + a_{j-i+1}\frac{i\theta}{\lambda + i\theta}. \tag{2.20}$$

La condition d'existence du régime stationnaire peut être obtenue comme suit :

L'accroissement moyen de la chaîne vaut

$$\begin{aligned} E[q_{n+1} - q_n \ / \ q_n = i] &= E[\nu_{n+1}] - E[\delta q_n = 1 \ / \ q_n = i] \\ &= \rho - \frac{i\theta}{\lambda + i\theta}. \end{aligned}$$

Si $\rho < 1$, alors :

$$\lim_{i\to\infty} E[q_{n+1} - q_n/q_n = i] = \rho - 1 < 0$$

et la chaîne est donc ergodique. Par contre, si $\rho \geq 1$, alors :

$$\lim_{i\to\infty} E[q_{n+1} - q_n/q_n = i] = \rho - \frac{i\theta}{\lambda + i\theta} \geq 1 - \frac{i\theta}{\lambda + i\theta} = \frac{\lambda}{\lambda + i\theta} > 0.$$

Puisque la chaîne est bornée inférieurement par la chaîne induite du système $M/G/1$ classique, donc la chaîne n'est pas ergodique (elle est transitoire).

Soit $P_n = \lim P(N_0(\xi_i) = n)$. Les équations de Kolmogorov se présentent de la manière suivante :

$$P_n = \sum_{m=0}^{n} p_m \frac{\lambda}{\lambda + m\theta} a_{n-m} + \sum_{m=1}^{n+1} p_m \frac{m\theta}{\lambda + m\theta} a_{n-m+1}. \ \text{ et } n = 0, 1, 2, ..$$

Vu la présence de convolution, cette équation peut être transformée, à l'aide des fonctions

génératrices $\phi(z) = \sum_{n=0}^{\infty} z^n p_n$ et $\psi(z) = \sum_{n=0}^{\infty} z^n \frac{p_n}{\lambda + n\theta}$,

$$\phi(z) = A(z)(\lambda\phi(z) + \theta\psi'(z)). \tag{2.21}$$

D'un autre côté ;

$$\begin{aligned} \phi(z) &= \sum_{n=0}^{\infty} z^n p_n = \sum_{n=0}^{\infty} z^n p_n \frac{\lambda + n\theta}{\lambda + n\theta}, \\ &= \lambda \sum_{n=0}^{\infty} z^n p_n \frac{p_n}{\lambda + n\theta} + \theta \sum_{n=0}^{\infty} n z^n \frac{p_n}{\lambda + n\theta}, \\ &= \lambda\psi(z) + \theta\psi'(z). \end{aligned} \tag{2.22}$$

Par conséquent ;

$$\begin{aligned} \lambda\psi(z) + \theta\psi'(z) &= A(z)(\lambda\psi(z) + \theta\psi'(z)), \\ \theta\psi'(z)\,[A(z) - z] &= \lambda\psi(z)\,[1 - A(z)]\,. \end{aligned}$$

Lemme :

La fonction analytique $f(z) = A(z) - z$ est positive, croissante et pour $z \in [0, 1]$, $\rho < 1 : z < A(z) < 1$.

Démonstration : Soit

$$f(z) = \widetilde{B}(\lambda - \lambda z) - z \ , \ f(1) = \widetilde{B}(0) - 1 = 0$$

en plus :

$$f\,'(z) = -\lambda \widetilde{B'}(\lambda - \lambda z) - 1 \ \text{ et } \ f\,'(1) = \rho - 1 < 0,$$

alors 1 est le seul zéro de f. En outre ;

$$f\,''(z) = -\lambda \widetilde{B'}(\lambda - \lambda z) + \lambda^2 \widetilde{B''}(\lambda - \lambda z) \geq 0.$$

Alors $f(z)$ est décroissante sur $[0,1]$, positive pour $\rho = \frac{\lambda}{\gamma} < 1$ et pour $z \in [0,1]$:

$$z < f(z) < 1.$$

Notons aussi que :

$$\lim_{z \to 1-} \frac{1 - \widetilde{B}(\lambda - \lambda z)}{\widetilde{B}(\lambda - \lambda z) - z} = \frac{\rho - 1}{1 - \rho} < \infty. \tag{2.23}$$

Théorème :

Soit $\rho < 1$, la distribution stationnaire de la chaîne de Markov induite possède la fonction génératrice suivante :

$$\phi(z) = \sum_{n=0}^{\infty} z^n P_n = \frac{(1-\rho)(1-z)A(z)}{A(z) - z} \exp\left\{ \frac{\lambda}{\theta} \int_1^z \frac{1 - A(u)}{A(u) - u} du \right\}, \tag{2.24}$$

où $A(z) = \widetilde{B}(\lambda - \lambda z)$.

2.5.3 Distribution stationnaire de l'état du système

Le premier résultat sur le système $M/G/1$ avec rappels est basé sur la méthode des variables supplémentaires. Une des approches permettant de trouver la distribution stationnaire jointe de l'état du serveur et de la taille de l'orbite a été introduite par De Kok (1984)[14]. Elle consiste à décrire le processus des arrivées comme processus de Markov avec dépendance de l'état de paramètre λ_{in} quand $\{C(t), N_0(t)\}$ est dans l'état (i,n) et à appliquer les schémas récursifs. L'état du système peut être décrit par le processus

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

où $\xi(t)$ est une variable aléatoire supplémentaire à valeurs dans $\mathbb{R}^+$, et désignant la durée de service écoulé à la date t. Notons par :

$$\begin{aligned} P_{0n} &= \lim_{t\to\infty} P(C(t)=0, N_0(t)=n), \\ P_{1n}(x) &= \lim_{t\to\infty} \frac{d}{dx} P(C(t)=1\ ,\ \xi(t)\le x\ ,\ N_0(t)=n). \end{aligned}$$

Les probabilités P_{0n} et $P_{1n}(x)$ vérifient le système d'équations de balance :

$$\begin{aligned} (\lambda+n\theta)p_{0\ n} &= \int_0^\infty p_{1\ n}(x)b(x)dx, \\ p'_{1\ n}(x) &= -(\lambda+b(x))p_{1\ n}(x)+\lambda p_{1\ n-1}(x), \\ p_{1\ n}(0) &= -\lambda p_{0\ n}+(n+1)\theta p_{0\ n+1}. \end{aligned}$$

où $b(x)=B'(x)/(1-B(x))$ est l'intensité instantanée du service étant donné que la durée écoulée est égale à x.

Soient les fonctions génératrices, telles que $P_0(z)=\sum_{n=0}^\infty z^n p_{0n}$ et $P_1(z,x)=\sum_{n=0}^\infty z^n p_{1n}$. Le système d'équations de balance devient :

$$\begin{aligned} &\lambda\textstyle\sum_{n=0}^\infty z^n p_{0\ n}+\theta\sum_{n=0}^\infty z^n p_{0\ n}=\int_0^\infty\sum_{n=0}^\infty z^n p_{1\ n}(x)dx, \\ &\textstyle\sum_{n=0}^\infty z^n p'_{1\ n}(x)=-(\lambda+b(x))\sum_{n=0}^\infty z^n p_{1\ n}(x)+\lambda\sum_{n=0}^\infty z^n p_{1\ n-1}(x), \\ &\textstyle\sum_{n=0}^\infty z^n p_{1\ n}(0)=\lambda\sum_{n=0}^\infty z^n p_{0\ n}+\theta\sum_{n=0}^\infty z^n(n+1)p_{0\ n+1}. \end{aligned}$$

D'où ;

$$\begin{cases} \lambda P_0(z)+\theta z\lambda P'_0(z)=\int_0^\infty P_1(z,x)b(x)dx, \\ P'_1(z,x)=(\lambda z-\lambda-b(x))P_1(z,x), \\ P_1(z,0)=\lambda P_0(z). \end{cases} \tag{2.25}$$

De la deuxième équation de (2.25), on a :

$$P_1(z,x)=P_1(z,0)\,[1-B(x)]\exp(-(\lambda-\lambda z)x).$$

Donc, la première équation de (2.25), devient ;

$$\begin{aligned}\lambda P_0(z) + \theta z P_0'(z) &= \int_0^\infty P_1(z,0)\left[1 - B(x)\right]\exp(-(\lambda - \lambda z)x)b(x)dx, \\ &= P_1(z,0)\widetilde{B}(\lambda - \lambda z) = P_1(z,0)A(z). \end{aligned} \tag{2.26}$$

A partir des équations (2.25) et (2.26), on a :

$$P_1(z,0)f(z) = \lambda P_0(z) + \theta z\left(\frac{P_1(z,0)}{\theta} - \frac{\lambda}{\theta}P_0(z)\right), \tag{2.27}$$

$$P_1(z,0) = \frac{\lambda - \lambda z}{A(z) - z}P_0(z)\left[1 - B(x)\right]\exp(-\lambda(\lambda - \lambda z)x). \tag{2.28}$$

En intégrant cette équation, et en utilisant la formule

$$\int_0^\infty \exp(-sx)\left[1 - B(x)\right]dx = \frac{(1 - \widetilde{B}(s))}{s},$$

On obtient ;

$$P_1(z) = \int_0^\infty P_1(z,x)dx = P_0(z)\frac{1 - A(z)}{A(z) - z}. \tag{2.29}$$

De (2.26) et (2.27), on peut obtenir $P_0(z)$

$$\lambda P_0(z) + \theta z P_0'(z) = A(z)\left[\lambda P_0(z) + \theta P_0'(z)\right],$$

$$\theta\left[A(z) - z\right]P_0'(z) = \lambda\left[1 - A(z)\right]P_0(z).$$

Considérons $f(z) = A(z) - z$. Du lemme , $f(z)$ est une fonction décroissante sur $[0,1]$, positive et pour $\rho < 1$ et $z \in [0,1]$:

$z < A(z) < 1$. En plus,

$$\lim_{z \to 1-} \frac{1 - A(z)}{A(z) - z} = \frac{A'(1)}{A(1) - 1} = \frac{\rho}{1 - \rho} < \infty.$$

De ce fait, pour $z = 1$, la fonction $\frac{1-A(z)}{A(z)-z} = \frac{\rho}{1-\rho}$.

Théorème :

Si $\rho = \lambda\beta_1 < 1$, le système est en régime stationnaire et les fonctions génératrices de la distribution conjointe de l'état du serveur et de la taille de l'orbite sont données par

$$P_0(z) = \sum_{n=0}^{\infty} z^n p_{0\,n} = (1 - \rho) \exp\left[\frac{\lambda}{\theta} \int_1^z \frac{1 - A(u)}{A(u) - u} du\right]. \tag{2.30}$$

$$P_1(z) = \sum_{n=0}^{\infty} z^n p_{1\,n} = \frac{1 - A(z)}{A(z) - z} P_0(z). \tag{2.31}$$

Mesures de performance

Les caractéristiques du modèle sont :

-Nombre moyen de clients dans le système

$$\overline{n} = Q'(1) = \rho + \frac{\lambda^2 \beta_2}{2(1 - \rho)} + \frac{\lambda\rho}{\theta(1 - \rho)},$$

-Nombre moyen de clients en orbite

$$\overline{n_0} = P'(1) = \overline{n} - \rho = \frac{\lambda^2 \beta_2}{2(1 - \rho)} + \frac{\lambda\rho}{\theta(1 - \rho)},$$

-Temps moyen d'attente d'un client

$$T^* = \frac{\overline{n_0}}{\lambda} = \frac{\lambda\beta_2}{2(1 - \rho)} + \frac{\rho}{\theta(1 - \rho)},$$

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

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

2.6 Conclusion :

Dans ce chapitre nous nous sommes intéressés aux systèmes de files d'attente de type $M/G/1$ et $M/M/1$ avec rappels. Une étude plus poussée des ce genre de systèmes est nécessaire pour amélioré et mieux évaluer les performances des systèmes informatiques, des réseaux de communications, systèmes industriels et systèmes complexes dans nombreux domaines. Cette technique est devenu inconcevable pour construire un système quelconque sans avoir fait une analyse des performances au préalable. Les modèles d'attente développés ces dernières décennies tentent de prendre en considération la recherche des clients en orbite par le serveur. Ces phénomènes affectent les caractéristiques de performance des systèmes réels.

Chapitre 3

Modèles de files d'attente avec rappels et recherche des clients en orbite

3.1 Introduction

Dans ce chapitre, nous étudions un système de files d'attente avec rappels à serveur unique et une politique de rappels linéaire, où le serveur peut effectuer une recherche des clients immédiatement après l'achèvement de chaque service.

3.2 Modèles d'attente avec rappels et recherche des clients en orbite

Les files d'attente avec rappels considérées par les chercheurs ont jusqu'ici la particularité, que le service est précédé et suivi d'une période d'inactivité qui se termine soit par l'arrivée d'un client primaire (premier tentative), ou par un client de l'orbite (client secondaire). la motivation pour envisager un modèle de file d'attente avec recherche (concurrence) provient d'une application dans un environnement d'un centre d'appels. La recherche de clients en orbite a été introduite pour la premiére fois par Neuts et al[19], où les auteurs ont étudié une file d'attente classique avec la recherche de clients immédiatement à la fin de chaque service.La recherche de clients en orbite pour le système de files d'attente $M/G/1$ avec rappels a été introduite par Artalejo et al[5]. Pour une revue détaillée des principaux résultats et de la littérature sur ce sujet, le lecteur est appelé a consulter Krishnamoorthy et al. [18], Chakravarthy et al. [11], Deepak et al. [14], Gao et Wang [25].

Cependant,nous considérons des modèles de files d'attente avec rappel dans lequel,

même sans salle d'attente, chaque époque d'achèvement du service ne doit pas nécessairement être suivi d'un temps d'inactivité. le temps d'inactivité du serveur est réduit par l'introduction de la recherche des clients en orbite immédiatement après l'achèvement du service. Ceci est réalisé comme suit : immédiatement à l'achèvement de chaque service, le serveur sélectionne un client de l'orbite avec une probabilité p_j, ou il y a j clients dans l'orbite (il est supposé que le serveur est conscient de l'état de l'orbite,par exemple, il a un registre des clients en orbite par contre ces derniers ignorent son état). Avec une probabilité $1 - p_j$ le serveur n'effectue aucune recherche des clients à la fin de son service et dans ce cas, comme dans la file d'attente avec rappel classique, une concurrence prend place entre les clients primaire et les clients secondaire pour atteindre le service. Ainsi, si la recherche est faite, le service sera suivi d'un autre service si non, il sera suivi d'un temps d'inactivité.Notre étude a deux objectifs principaux :

1. Le premier consiste à introduire la recherche des clients en orbite dans un modèle de file d'attente avec rappel, ce qui permet de réduire le temps d'inactivité du serveur. Si le tenu des coûts et de mise en œuvre de la recherche de clients sont introduits. Les résultats obtenus peuvent être utilisés pour le réglage optimal des paramètres du mécanisme de la recherche.
2. Le deuxième objectif est de donner un aperçu du lien entre la file d'attente avec rappel correspondante et la file d'attente classique (sans rappel).

Cependant, on observe que si $p_j = 1$, notre modèle sera réduit à celui d'une file d'attente classique (sans rappel) et quand $p_j = 0$, il devient le modèle de file d'attente avec rappel.

3.3 Modèle d'attente avec rappels linaire et recherche des clients

Les temps d'inter-rappels peuvent être modélisés selon différentes disciplines en fonction de chaque application particulière. Dans les systèmes téléphoniques,les tentatives répétées sont effectués individuellement par chaque client bloqué suivant une loi exponentielle de taux θ. C'est la discipline de rappels classique dont le taux est $j\theta$, quand la taille de

l'orbite est $j \geq 1$. En revanche, il existe d'autres types de situations de file d'attente dont les temps d'inter-rappels sont indépendants du nombre de clients en orbite. Cette deuxième possibilité est la discipline de rappels constante,i.e. le taux d'inter-rappels est $\alpha(1-\delta_{j0})$ où δ_{j0} est la Kronecker fonction. Artalejo et Gomez-Corral[6], unissent les deux disciplines en définissant la politique d'inter-rappels linéaire avec le taux $\alpha(1-\delta_{j0})+j\theta$. L'application suivante motive l'analyse du modèle considéré ci-dessous.

Réparation par le service avec la recherche des clients : le serveur tient un registre des clients qui sont obligés de quitter le système car ils ont rencontré un serveur occupé au moment de l'arrivée. À la fin d'un service, le serveur décide de démarrer immédiatement le prochain service en récupérant un client non satisfait (de l'orbite) avec une probabilité p_j.

Le temps de recherche est supposé être négligeable. La probabilité de ne pas effectuée une recherche des clients est $q_j = 1 - p_j$. Si le serveur ne récupère pas le prochain client à servir de l'orbite, alors il va y avoir une concurrence entre les clients primaires et ceux de l'orbite pour attendre le prochain service. Ainsi, le travail actuel inclut comme cas particulier, la file d'attente classique lorsque $p_j = 1$, $j \neq 0$ et la file d'attente avec rappels lorsque $p_j = 0$.

3.3.1 Modèle d'attente $M/M/1$ avec rappels linèaire et recherche des clients en orbite

Nous considérons un système de file d'attente à serveur unique auquel les clients primaire arrivent selon un flux Poissonienne de taux λ. Tout client qui trouve, le serveur occupé quitte immédiatement la zone de service et rejoint l'orbite. L'intervalle entre deux tentatives répétées successives est distribué de manière exponentielle avec un taux $\alpha(1-\delta_{j0})+j\theta$, étant donner que le nombre de clients en orbite est j, $j \geq 1$. La distribution du temps de service est exponentiels de taux μ. Soit ξ_n l'instant de l'achèvement de service du $n^{ème}$ client. Immédiatement après ce moment, le serveur passe à la recherche d'un client en orbite avec une probabilité p_j , $p_0 = 0$ qui dépend du nombre de clients j en orbite. Avec la probabilité $q_j = 1 - p_j$ le serveur reste libre. Dans ce dernier cas,l'événement à suivre

dépend d'une compitition entre les arrivées primaires du taux λ et le flux d'inter-rappels de taux $\alpha(1-\delta_{j0})+j\theta$. Le temps de recherche est supposé être négligeable. Les flux d'arrivées primaires, d'inter-rappels et du temps de service sont supposés être mutuellement indépendants.

Description du modéle :

Soit $N(t)$ le nombre de clients en orbite et $C(t)$ est l'état du serveur à l'instant t.

Nous avons :

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

L'état du système à la date t peut être décrit par le processus stochastique suivant :

$$\{C(t),N(t),\xi(t);\ t\geq 0\},$$

l'espace d'état du processus est $S=\{0,1\}\times\mathbb{N}$. $\rho=\frac{\lambda}{\mu}<1$, est la condition de stabilité du système proposé. Les transitions entre les états sont illustrées dans la Fig(3.1) :

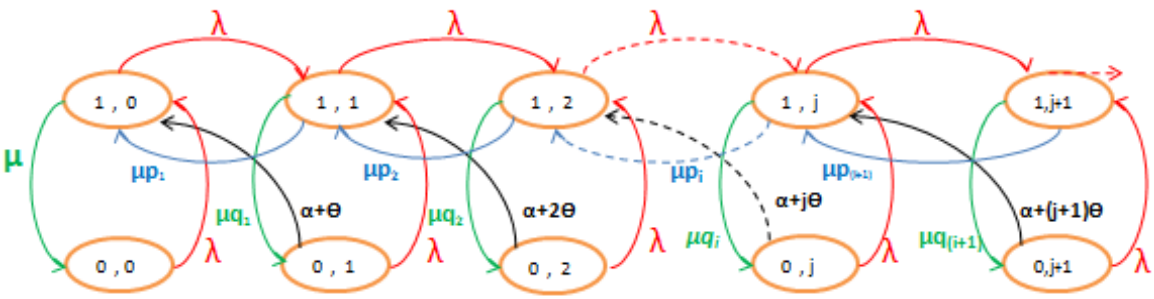

Fig3.1 : Modèle avec rappels linéaire et recherche de clients.

Distribution stationnaire de l'état du système

Théorème :

La distribution de l'état du serveur $C(t)$ et du nombre de clients dans l'orbite $N(t)$ du

système $M/M/1$ au régime stationnaires $p_{ij} = P\left[C(t) = i,\ N(t) = j\right]$ est donnée par :

$$p_{0j} = p_{00} q_j \rho^j \prod_{k=0}^{j-1} \frac{\lambda + \alpha(1 - \delta_{k0}) + k\theta}{p_{k+1}\lambda + \alpha + (k+1)\theta},\ j \geq i, \tag{3.1}$$

$$p_{1j} = p_{00} \rho^{j+1} \prod_{k=1}^{j} \frac{\lambda + \alpha + k\theta}{p_k \lambda + \alpha + k\theta},\ j \geq 0, \tag{3.2}$$

$$p_{00}^{-1} = \sum_{j=0}^{\infty} \rho^{j+1} \left(1 + \frac{q_j \mu}{\lambda + \alpha(1 - \delta_{j0}) + j\theta}\right) \prod_{k=1}^{j} \frac{\lambda + \alpha + k\theta}{P_k \lambda + \alpha + k\theta}. \tag{3.3}$$

Preuve :

L'ensemble des équations d'équilibre statistique pour les probabilités p_{0j} et p_{1j} est :

$$(\lambda + \alpha(1 - \delta_{j0}) + j\theta) p_{0j} = q_j \mu p_{1j},\ j \geq 0, \tag{3.4}$$

$$(\lambda + \mu) p_{1j} = \lambda P_{1,j-1} + \lambda P_{0j} + \left[\alpha + (j+1)\theta\right] P_{0,j+1} + \mu p_{j+1} P_{1,j+1},\ j \geq 0, \tag{3.5}$$

En utilisant l'équation (3.4), éliminer les probabilités p_{1j} de l'équation (3.5). Après quelques calculs algèbres sur l'équation résultante, nous obtenons :

$$\begin{aligned} & \mu q_{j-1} q_j (\alpha + (j+1)\theta + \lambda P_{j+1}) p_{0,j+1} - \lambda q_{j-1} q_{j+1} (\lambda + \alpha + j\theta) p_{0j} \\ = \ & \mu q_{j-1} q_{j+1} (\alpha + j\theta + \lambda p_j) p_{0j} - \lambda q_j q_{j+1} (\lambda + \alpha(1 - \delta_{j-1,0}) + (j-1)\theta) p_{0,j-1}\ , \end{aligned}$$

cela implique que :

$$\mu q_{j-1} q_{j+1} (\alpha + j\theta + \lambda p_j) p_{0j} - \lambda q_j q_{j+1} (\lambda + \alpha(1 - \delta_{j-1,0}) + (j-1)\theta) p_{0,j-1} = 0\ ,$$

donc :

$$p_{0j} = \frac{\lambda q_j (\lambda + \alpha(1 - \delta_{j-1,0}) + (j-1)\theta)}{\mu q_{j-1} (\alpha + j\theta + \lambda p_j)} p_{0,j-1}\ .$$

3.3.2 Modèle d'attente $M/M/1$ avec rappel constant et recherche des clients

Il semble impossible d'exprimer les formule (3.1), (3.3) en fonction d'une fonction connue , même dans le cas de la recherche des clients géométrique (c-à-d $p_j = 1 - p^j$, $p \in [0,1]$, $j \geq 1$). Cependant, nous supposons le cas de la recherche constante $p_j = p$, $p \in [0,1]$, $j \geq 1$, et $\beta(s) = \int_0^\infty e^{-sx} dB(x)$ est le transformer de la Laplace-Stieltjes $B(x)$, $\beta_k = (-1)^k \beta^{(k)}(0)$ est le k^{ieme} moment du temps de service, pour obtenir quelques belles expressions fermées.

Premiérement, nous introduisons quelques notations élémentaire. Soit F la sèrie hypergèomètrique donnée par :

$$F(a, b\ ;\ c\ ;\ z) = \sum_{k=0}^{\infty} \frac{(a)_k (b)_k}{(c)_k} \frac{z^k}{k!},$$

où $(x)_k$ est le symbole Pochhammer défini par :

$$(x)_k = \begin{cases} 1, & \text{si } k = 0, \\ x(x+1)...(x+k-1), & \text{si } k \geq 1. \end{cases}$$

Nous introduisons également les fonctions génératrice partielles :

$$P_i(z) = \sum_{j=0}^{\infty} z^j p_{ij},\ i \in \{0,1\}\,,\ \ |z| \leq 1,$$

et les moments factoriels partiels M_k^i définis par :

$$M_0^i = \sum_{j=0}^{\infty} p_{ij},\quad i \in \{0,1\}\,,$$

$$M_k^i = \sum_{j=k}^{\infty} j(j-1)...(j-k+1) p_{ij},\quad i \in \{0,1\}\,,\ k \geq 1$$

Théorème :

Supposons que $\{N(t); t \geq o\}$ est positif récurrent, alors :

- Les probabilités limites $\{p_{ij}\}_{(i,j)\in s}$ sont données par :

$$p_{0j} = p_{00}\frac{(1-p)\lambda}{\lambda+\alpha}\rho^j\frac{\left(\frac{\lambda+\alpha}{\mu}\right)_j}{\left(\frac{p\lambda+\alpha}{\mu}+1\right)_j}, \quad j\geq 1, \tag{3.6}$$

$$p_{1j} = p_{00}\rho^{j+1}\frac{\left(\frac{\lambda+\alpha}{\mu}+1\right)_j}{\left(\frac{p\lambda+\alpha}{\mu}+1\right)_j}, \quad j\geq 0, \tag{3.7}$$

$$p_{00}^{-1} = F\left(1,\frac{\lambda+\alpha}{\mu}+1\,;\,\frac{p\lambda+\alpha}{\mu}+1\,;\,\rho\right). \tag{3.8}$$

-Les fonctions génératrice partielles $P_i(z), 0\leq i\leq 1$, sont données par :

$$P_0(z) = p_{00}(1-\rho z)F\left(1,\frac{\lambda+\alpha}{\mu}+1\,;\,\frac{p\lambda+\alpha}{\mu}+1\,;\,\rho z\right). \tag{3.9}$$

$$P_1(z) = p_{00}\rho F\left(1,\frac{\lambda+\alpha}{\mu}+1\,;\,\frac{p\lambda+\alpha}{\mu}+1\,;\,\rho z\right) \tag{3.10}$$

-Les moments factoriels partiels M_k^i ,$i\in\{0,1\}$, $k\geq 0$, sont donnés par :

$$\begin{aligned}
M_0^0 &= 1-\rho,\\
M_k^0 &= p_{00}k!\frac{(1-p)\lambda}{\lambda+\alpha}\rho^k\frac{\left(\frac{\lambda+\alpha}{\mu}\right)_k}{\left(\frac{p\lambda+\alpha}{\mu}+1\right)_k}F\left(k+1,\frac{\lambda+\alpha}{\mu}+k\,;\,\frac{p\lambda+\alpha}{\mu}+k+1\,;\,\rho\right), \quad k\geq 1,\\
M_0^1 &= \rho,\\
M_k^1 &= p_{00}k!\rho^{k+1}\frac{\left(\frac{\lambda+\alpha}{\mu}+1\right)_k}{\left(\frac{p\lambda+\alpha}{\mu}+1\right)_k}F\left(k+1,\frac{\lambda+\alpha}{\mu}+k+1\,;\,\frac{p\lambda+\alpha}{\mu}+k+1\,;\,\rho\right), \quad k\geq 1,
\end{aligned}$$

Preuve :

Pour le cas $p_j = p$, $j\geq 1$, formules (3.1),(3.3) réduire à (3.6),(3.8). On utilisent les fonctions génératrice, nous obtenons (3.10) et

$$P_0(z) = \frac{p_{00}}{\lambda+\alpha}\left(p\lambda+\alpha+(1-p)\lambda F\left(1,\frac{\lambda+\alpha}{\mu}\,;\,\frac{p\lambda+\alpha}{\mu}+1\,;\,\rho z\right)\right). \tag{3.11}$$

Après un certain réarrangement (42) donne l'expression alternative (40). La clé pour calculer les moments factoriels partiels est l'identité suivante :

$$p_i(1+z) = \sum_{k=0}^{\infty} M_k^i \frac{z^k}{k!}, \quad i \in \{0,1\}.$$

Après avoir remplacé $(1+z)^j$ par $\sum_{k=0}^{j} \begin{pmatrix} j \\ k \end{pmatrix} z^k$, nous pouvons obtenir M_k^i par une identification directe des coefficients de la série $P_i(1+z)$. Par souci d'exhaustivité, nous donnons ensuite les expressions correspondant à la discipline de rappels classique et constante.

Corollaire 1 (*discipline de rappels classique*) :

Considérons $\alpha = 0$ et $\mu > 0$,alors :

-Les probabilités limites sont données par :

$$\begin{aligned} p_{0j} &= p_{00}(1-p)\rho^j \frac{\left(\frac{\lambda}{\mu}\right)_j}{\left(\frac{\rho\lambda}{\mu}+1\right)_j}, \quad j \geq 1, \\ p_{1j} &= p_{00}\rho^{j+1} \frac{\left(\frac{\lambda}{\mu}+1\right)_j}{\left(\frac{\rho\lambda}{\mu}+1\right)_j}, \quad j \geq 0, \\ p_{00}^{-1} &= F\left(1, \frac{\lambda}{\mu}+1\,;\, \frac{\rho\lambda}{\mu}+1\,;\, \rho\right). \end{aligned}$$

-Les fonctions génératrices partielle sont données par :

$$\begin{aligned} P_0(z) &= p_{00}(1-\rho z)F\left(1, \frac{\lambda}{\mu}+1\,;\, \frac{\rho\lambda}{\mu}+1\,;\, \rho z\right), \\ P_1(z) &= p_{00}\rho F\left(1, \frac{\lambda}{\mu}+1\,;\, \frac{\rho\lambda}{\mu}+1\,;\, \rho z\right). \end{aligned}$$

-Les moments factoriels partiels sont donnés par :

$$\begin{aligned}
M_0^0 &= 1-\rho, \\
M_k^0 &= p_{00}k!(1-p)\rho^k \frac{\left(\frac{\lambda}{\mu}\right)_k}{\left(\frac{\rho\lambda}{\mu}+1\right)_k} F\left(k+1, \frac{\lambda}{\mu}+k\,;\, \frac{\rho\lambda}{\mu}+k+1\,;\,\rho\right), \quad k\geq 1, \\
M_0^1 &= \rho, \\
M_k^1 &= p_{00}k!\rho^{k+1} \frac{\left(\frac{\lambda}{\mu}+1\right)_k}{\left(\frac{\rho\lambda}{\mu}+1\right)_k} F\left(k+1, \frac{\lambda}{\mu}+k+1\,;\, \frac{\rho\lambda}{\mu}+k+1\,;\,\rho\right), \quad k\geq 1.
\end{aligned}$$

Corollaire 2 (*discipline de rappels constante*) :

Considérons $\alpha > 0$ et $\mu = 0$,alors :

-Les probabilités limites sont données par :

$$\begin{aligned}
p_{0j} &= p_{00}\frac{(1-p)\lambda}{\lambda+\alpha}\beta^j, \quad j\geq 1, \\
p_{1j} &= p_{00}\rho\beta^j\ , \ j\geq 0, \\
p_{00} &= 1-\beta.
\end{aligned}$$

-Les fonctions génératrices partielle sont données par :

$$\begin{aligned}
P_0(z) &= \frac{(1-pz)(1-\beta)}{1-\beta z}, \\
P_1(z) &= \frac{\rho(1-\beta)}{1-\beta z}.
\end{aligned}$$

-Les moments factoriels partiels sont donnés par :

$$\begin{aligned}
M_0^0 &= 1-\rho, \\
M_k^0 &= k!\frac{(1-p)\lambda}{\lambda+\alpha}\left(\frac{\beta}{1-\beta}\right)^k, \quad k\geq 1, \\
M_0^1 &= \rho, \\
M_k^1 &= k!\rho\left(\frac{\beta}{1-\beta}\right)^k, \quad k\geq 1.
\end{aligned}$$

Pour les choix $p_j = 1$, $j \geq 1$, et $p_j = 0$, $j \geq 1$, on peut déduire les mésures de performances de la file d'attente $M/M/1$ classique et ceux du $M/M/1$ avec rappels.

3.4 Conclusion :

Dans ce chapitre, on a présenté le système de file d'attente $M/M/1$ avec la discipline de rappels linaire et recherche des clients en orbite par le serveur (temps de recherche est considéré négligeable), quelques résultats comparatifs avec le système standard et celui avec rappels classique à été calculer. Dans le prochain chapitre nous étudions un modèle d'attente $M/M/1$ avec rappels classique et temps de recherche des clients en orbite non négligeable, de plus plusieurs résultats numérique seront présenté.

Chapitre 4

Application : Modèle d'attente $M/M/1$ avec rappels et recherche des clients

4.1 Introduction

Le Cloud Computing (systèmes d'informatique en nuage) est le hype informatique de cette dernière décennie et le fruit des évolutions récentes des technologies de l'information. Il constitue une véritable révolution dans l'utilisation de l'informatique en amenant des nouvelles possibilités de mutualisation de services et d'économies pour les entreprises. Comme tout est fourni aux utilisateurs du Cloud en tant que services, la qualité de service a un impact important sur la croissance et l'acceptabilité du paradigme du Cloud Computing.

Dans ce chapitre, nous allons voir comment les fournisseurs de services Cloud peuvent garantir aux utilisateurs de service Cloud la qualité de service appropriée, en considérant le temps de recherche des clients en orbite non négligeable .Un modèle de files d'attente avec rappels et recherche des clients en orbite est proposé par Tuan Phung [23] est motivé par les systèmes d'informatique en nuage (Cloud Computing) où l'unité de traitement et l'unité de stockage sont séparés. L'unité de traitement à la capacité de servir une seule tache à chaque fois. Nous proposons de notre part, plusieurs résultats numériques pour étudier l'effet de certains paramètres clés sur les caractéristiques du modèle.

4.2 Description du modèle :

Les clients arrivent au serveur selon un processus de Poisson avec taux λ. Le temps de service des clients entrants suit la loi de distribution exponentielle avec une moyenne $1/\mu$. Après l'achèvement du service, le serveur reste inactif pendant une durée de temps qui suit la loi exponentielle avec une moyenne $1/\alpha$. Pendant cette période d'inactivité, un client arrivant (soit primaire ou secondaire) est immédiatement servi. Après le temps de repos, le serveur commence la recherche d'un client en orbite. Le temps de recherche suit la distribution exponentielle avec la moyenne $1/\nu$. Les clients qui arrivent et trouvent le serveur occupé (par un service d'un client ou par la recherche) rejoint l'orbite. Ainsi, les clients de l'orbite tentent d'attendre le serveur après un certain temps exponentiellement distribué avec moyenne $1/\theta$. Ce modèle à été analysé pour la premières fois dans la littérature par Tuan Phung[23] .

4.3 Etude analytique du système :

Soit $C(t)$ l'état du serveur à l'instant t, $t \geq 0$.

$$C(t) = \begin{cases} 0 & \text{le serveur est libre,} \\ 1 & \text{le serveur est occupé,} \\ 2 & \text{le serveur recherche un client.} \end{cases}$$

$N(t)$ est le nombre de clients dans l'orbite à l'instant t, $t \geq 0$. Nous avons $\{X(t) = (C(t), N(t)), t \geq 0\}$ qui forme une chaîne de Markov sur l'espace d'état S=$\{0, 1, 2\} \times \{0, 1, 2, ...\}$. Voir la figure (4.1), pour les transitions entre les états.

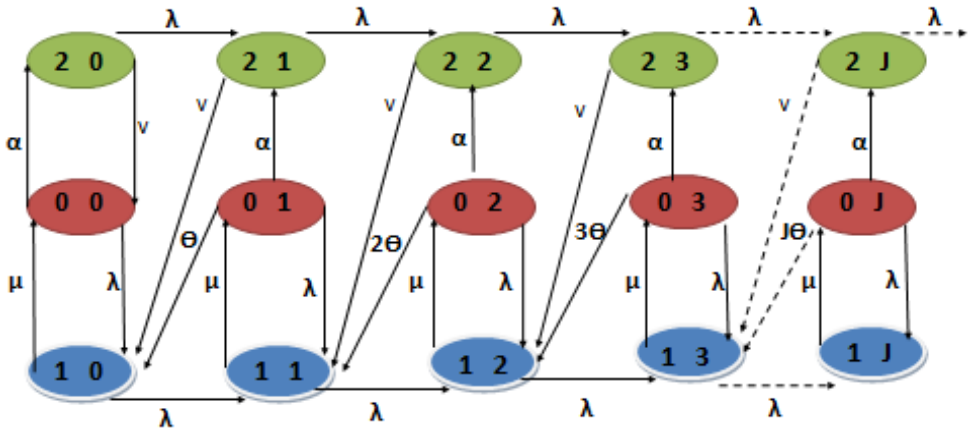

Fig4.1 : Les transitions des états du modèle.

Nous supposons que le système est stable, c'est-à-dire que la distribution stationnaire existe.

La condition nécessaire et suffisante pour que le système soit stable est $\lambda < \mu$ qui sera obtenus plus tard dans l'analyse.

$$p_{ij} = \lim_{t\to\infty} p(C(t) = i, N(t) = j),$$

Les équations d'équilibre de balance pour les états (i, j) sont indiqués comme suit :

$$\begin{aligned}
(\lambda+\alpha)p_{0,0} &= \mu p_{1,0} + \nu p_{2,0} \ , && (4.1)\\
(\lambda+\alpha+j\theta)p_{0,j} &= \mu p_{1,j} && j \geq 1, \ (4.2)\\
(\lambda+\mu)p_{1,j} &= (j+1)\theta p_{0,j+1} + \nu p_{2,j+1} + \lambda p_{1,j-1} + \lambda p_{0,j} && j \geq 0,\\
(\lambda+\nu)p_{2,j} &= \alpha p_{0,j} + \lambda p_{2,j-1} && j \geq 0, \ (4.3)
\end{aligned}$$

où $p_{i,-1} = 0 (i = 1, 2)$. Soient $P_i(z)$ fonctions génératrice de $p_{i,j}$, ie. $P_i(z) = \sum_{j=0}^{\infty} p_{i,j} z^j$ $(i = 0, 1, 2)$. On Transforment les équations d'équilibre ci-dessus en fonctions génératrice

nous obtenons :

$$\begin{aligned}(\lambda+\alpha)P_0(z)+\theta z P_0^{'}(z) &= \mu P_1(z)+\nu p_{2,0}, &(4.4)\\ (\lambda+\mu)P_1(z) &= \theta P_0^{'}(z)+\frac{\nu}{z}(P_2(z)-p_{2,0})+\lambda z P_1(z)+\lambda P_0(z), &(4.5)\\ (\lambda+\nu)P_2(z) &= \alpha P_0(z)+\lambda z P_2(z). &(4.6)\end{aligned}$$

En résument les équations ci-dessus et organisent les rendements de résultats en aura :

$$\lambda(P_1(z)+P_2(z))=\mu P_0^{'}(z)+\frac{\nu(P_2(z)-p_{2,0})}{z}. \tag{4.7}$$

Cette équation représente l'équilibre entre les entrées et sorties de l'orbite. A partir de (4.4) et (4.6), nous obtenons :

$$\begin{aligned}P_1(z) &= \frac{(\lambda+\alpha)P_0(z)+\mu z P_0^{'}(z)-\nu p_{2,0}}{\mu}, &(4.8)\\ P_2(z) &= \frac{\alpha P_0(z)}{\lambda+\nu-\lambda z}. &(4.9)\end{aligned}$$

Substituons ces deux expressions dans l'équation d'équilibre de l'orbite (4.7) et organisons les rendements du résultat en aura :

$$P_0^{'}(z)=A(z)P_0(z)+B(z). \tag{4.10}$$

où

$$A(z)=\frac{\frac{\lambda(\lambda+\alpha)}{\mu}+\frac{\alpha(\lambda-\frac{\nu}{z})}{\lambda+\nu-\lambda z}}{\theta\left(1-\frac{\lambda z}{\theta}\right)}, \qquad B(z)=\frac{p_{2,0}\nu}{\theta z}.$$

Nous décomposons $A(z)$ comme suit :

$$A(z)=\frac{a}{z}+\frac{b}{1-\frac{\lambda z}{\mu}}+\frac{c}{1-\frac{\lambda z}{\lambda+\nu}},$$

où a, b et c sont donnés par :

$$a = -\frac{\alpha\nu}{\theta(\lambda+\nu)}, \qquad b = \frac{\lambda^2(\lambda+\alpha+\nu-\mu)}{\theta\mu(\lambda+\nu-\mu)}, \qquad c = \frac{\lambda^2\alpha\mu}{(\lambda+\nu)^2\theta(\mu-\lambda-\nu)}.$$

nous résolvons d'abord l'équation différentielle non homogène ;

$$P_0^{'}(z) = A(z)P_0(z),$$

qui se transforme en ;

$$\frac{P_0^{'}(z)}{P_0(z)} = \frac{a}{z} + \frac{b}{1-\frac{\lambda z}{\mu}} + \frac{c}{1-\frac{\lambda z}{\lambda+\nu}}.$$

La solution de cette équation différentielle est donnée par :

$$P_0(z) = Cz^a \left(\frac{\mu-\lambda}{\mu-\lambda z}\right)^{\frac{b\mu}{\lambda}} \left(\frac{\nu}{\lambda+\nu-\lambda z}\right)^{\frac{c(\lambda+\nu)}{\lambda}},$$

où C est un nombre constant. Comme d'habitude, nous avons la solution pour notre équation différentielle original (4.10) sous la forme suivante :

$$P_0(z) = C(z)z^a \left(\frac{\mu-\lambda}{\mu-\lambda z}\right)^{\frac{b\mu}{\lambda}} \left(\frac{\nu}{\lambda+\nu-\lambda z}\right)^{\frac{c(\lambda+\nu)}{\lambda}},$$

où $C(z)$ est une fonction inconnue. Substituons dans l'équation différentielle (4.10) les rendements on aura :

$$C^{'}(z)z^a \left(\frac{\mu-\lambda}{\mu-\lambda z}\right)^{\frac{b\mu}{\lambda}} \left(\frac{\nu}{\lambda+\nu-\lambda z}\right)^{\frac{c(\lambda+\nu)}{\lambda}} = \frac{p_{2,0}\nu}{\theta z},$$

ou d'une manière équivalente :

$$C^{'}(z) = \frac{p_{2,0}\nu}{\theta} z^{-(a+1)} \left(\frac{\mu-\lambda}{\mu-\lambda z}\right)^{-\frac{b\mu}{\lambda}} \left(\frac{\nu}{\lambda+\nu-\lambda z}\right)^{-\frac{c(\lambda+\nu)}{\lambda}}.$$

Donc, nous avons

$$C(z) = C_0 - \frac{p_{2,0}\nu}{\theta} \int_z^1 u^{-(a+1)} \left(\frac{\mu-\lambda}{\mu-\lambda u}\right)^{-\frac{b\mu}{\lambda}} \left(\frac{\nu}{\lambda+\nu-\lambda u}\right)^{-\frac{c(\lambda+\nu)}{\lambda}} du,$$

où C_0 est un nombre constant. Puisque $P_0(z)$ est analytique en $z = 0$ et $a < 0$, nous devons avoir $C(0) = 0$ d'où :

$$C_0 = \frac{p_{2,0}\nu}{\theta} \int_0^1 u^{-(a+1)} \left(\frac{\mu-\lambda}{\mu-\lambda u}\right)^{-\frac{b\mu}{\lambda}} \left(\frac{\nu}{\lambda+\nu-\lambda u}\right)^{-\frac{c(\lambda+\nu)}{\lambda}} du.$$

La solution finale de $P_0(z)$ est donnée par :

$$P_0(z) = \frac{p_{2,0}\nu}{\theta} z^a \left(\frac{\mu-\lambda}{\mu-\lambda z}\right)^{\frac{b\mu}{\lambda}} \left(\frac{\nu}{\lambda+\nu-\lambda z}\right)^{\frac{c(\lambda+\nu)}{\lambda}} \times \int_0^z u^{-(a+1)} \left(\frac{\mu-\lambda}{\mu-\lambda u}\right)^{-\frac{b\mu}{\lambda}} \left(\frac{\nu}{\lambda+\nu-\lambda u}\right)^{-\frac{c(\lambda+\nu)}{\lambda}} du. \tag{4.11}$$

De (4.7), (4.9) et (4.10), nous obtenons :

$$P_1(1) + P_2(1) = \left(\frac{\theta}{\lambda} A(1) + \frac{\alpha}{\lambda}\right) P_0(z). \tag{4.12}$$

Nous avons aussi la condition de normalisation :

$$P_0(1) + P_1(1) + P_2(1) = 1. \tag{4.13}$$

De (4.12) et (4.13), nous obtenons :

-La probabilité que le serveur est libre mais le système n'est pas vide $P_0(1)$:

$$P_0(1) = \frac{\nu(1-\frac{\lambda}{\mu})}{\alpha+\nu},$$

où l'expression de $A(1)$ en termes de paramètres donnés est utilisée dans (4.8) et (4.9) et :

-La probabilité que le serveur est occupé $P_1(1)$:

$$P_1(1) = \frac{\lambda}{\nu},$$

-La probabilité que le serveur est en recherche $P_2(1)$:

$$P_2(1) = \frac{\alpha(1-\frac{\lambda}{\mu})}{\alpha+\nu}.$$

Par conséquent, de l'expression $P_0(z)$, nous obtenons l'expression de $p_{2,0}$ comme suit :

$$p_{2,0} = \frac{\theta(1-\frac{\lambda}{\mu})}{(\alpha+\nu)\int_0^1 u^{-(a+1)}\left(\frac{\mu-\lambda}{\mu-\lambda u}\right)^{-\frac{b\mu}{\lambda}}\left(\frac{\nu}{\lambda+\nu-\lambda u}\right)^{-\frac{c(\lambda+\nu)}{\lambda}} du}. \tag{4.14}$$

De cette expression, nous obtenons le fait que la condition de stabilité pour le modèle est $\lambda < \mu$.

4.4 Analyse de sensibilité des mesures du performance du modèle

Notre objectif est d'étudier le comportement de mesures de performance du ce système par rapport à quelque paramètres. Le calcul numérique est réalise à l'aide de logiciel Matlab. Les différents résultats obtenus sont présentés dans des tableaux et des figures.

Dans tout ce qui suits on vas présenté l'influence des paramètres sur :

P_0 : la probabilité que le serveur est libre mais le système n'est pas vide.

P_1 : la probabilité que le serveur est occupé.

P_2 : la probabilité que le serveur est en recherche.

$\bar{n_0}$: le nombre moyen de clients dans l'orbite.

$\bar{n_s}$: le nombre moyen de clients dans le systéme.

W : le temps moyen de séjour des clients dans le système.

W_q : le temps moyen d'attente des clients dans l'orbite.

4.4.1 L'effet du taux d'arrivée λ sur le modèle :

On fait varier le taux d'arrivée λ sur les mesures de performance pour $\mu = 10$, $\nu = 2.5$, $\alpha = 3$, et $\theta = 1.34$.

λ	p0	p1	p2	no	ns	w	wq
1	0,4091	0,1	0,4909	2,9917	3,0917	3,0917	2,9917
1,2	0,4	0,12	0,48	4,5178	4,6178	3,8482	3,7648
1,4	0,3909	0,14	0,4691	6,1354	6,2354	4,4539	4,3824
1,6	0,3818	0,16	0,4582	7,7792	7,8792	4,9245	4,862
1,8	0,3727	0,18	0,4473	9,3653	9,4653	5,2585	5,2029
2	0,3636	0,2	0,4364	10,7964	10,8964	5,4482	5,3982
2,2	0,3545	0,22	0,4255	11,9708	12,0708	5,4867	5,4413
2,4	0,3455	0,24	0,4145	12,9758	12,8958	5,3732	5,3316
2,6	0,3364	0,26	0,4036	13,2037	13,3037	5,1168	5,0783
2,8	0,3273	0,28	0,3927	13,1686	13,2686	4,7388	4,7031
3	0,3182	0,3	0,3818	12,7189	12,8189	4,273	4,2396

Tab4.1 : Variations des mesure de performances dans le systéme en fonction de λ.

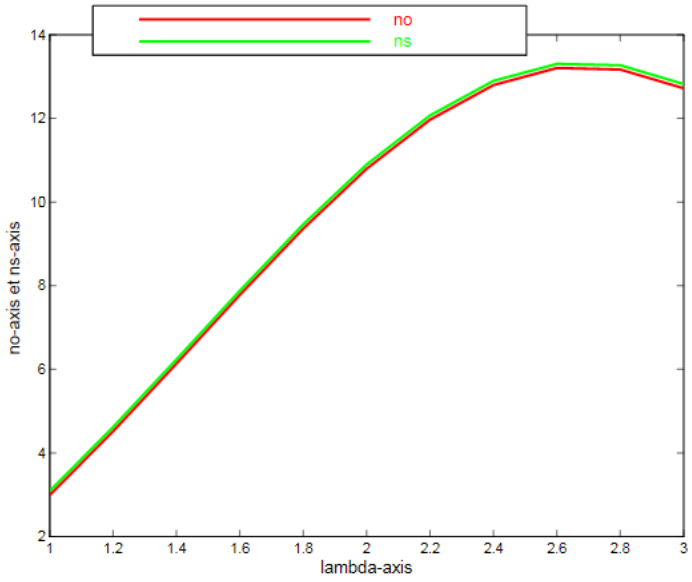

Fig4.2 : Effet de λ sur le nombre moyen de clients dans le système et dans l'orbite.

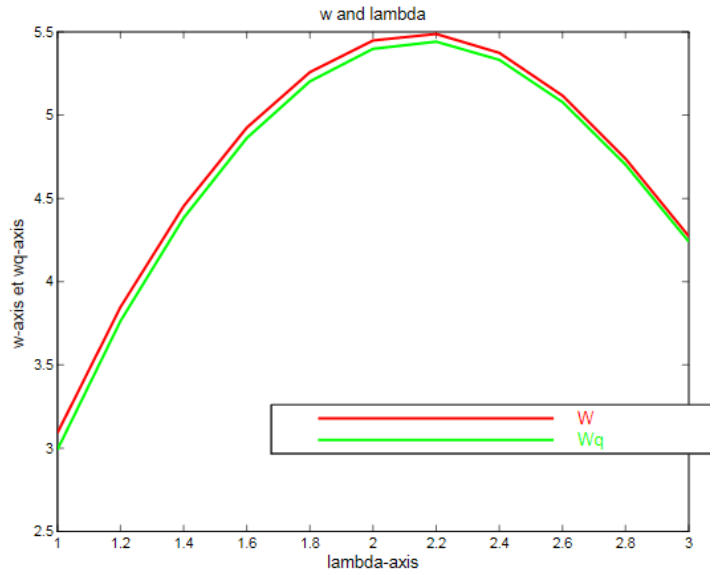

Fig4.3 : Effet de λ sur le temps moyen de séjour des clients dans le système et dans l'orbite.

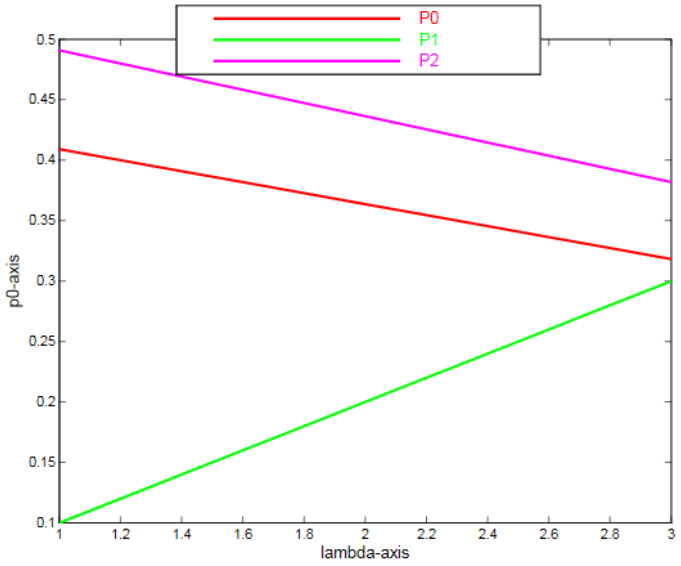

Fig4.4 : Effet de λ sur les probabilités de l'état du serveur.

Commentaire :

À partir du tableau (4.1) et de la figure (4.4), nous pouvons voir que le taux des arrivée influence sur , les probabilités de l'état du serveur P_0, P_1 ou elle décroissent de façon monotone, par contre la probabilité P_2 est croissante. Pour le nombre moyen de clients dans le système et dans l'orbite, ont tendance à augmenter de façon monotone. Ce qui est conforme à notre prévision.

4.4.2 L'effet du taux de recherche ν sur le modèle :

On fait varier le taux de recherche ν sur les mesures de performance pour $\lambda = 2.2$, $\mu = 4$, $\theta = 1.44$, et $\alpha = 1.2$.

v	p0	p1	p2	no	ns	w	wq
1	0,2045	0,55	0,2455	2,5629	3,1129	1,415	1,165
2	0,2812	0,55	0,1687	17,3015	17,8515	8,1143	7,8643
3	0,3214	0,55	0,1286	4,0757	4,6257	2,1026	1,8526
4	0,3462	0,55	0,1038	3,0197	3,5697	1,6226	1,3726
5	0,3629	0,55	0,0871	2,5876	3,1376	1,4262	1,1762
6	0,375	0,55	0,075	2,3456	2,9856	1,3162	1,0662
7	0,3841	0,55	0,0659	2,1893	2,7393	1,2451	0,9951
8	0,3913	0,55	0,0587	2,0795	2,6295	1,1952	0,9452
9	0,3971	0,55	0,0529	1,9978	2,5478	1,1581	0,9081
10	0,4018	0,55	0,0442	1,9347	2,4847	1,1294	0,8791
11	0,4057	0,55	0,0443	1,8843	2,4343	1,1065	0,8565

Tab4.2 : Variations des mesure de performances dans le systéme en fonction de ν.

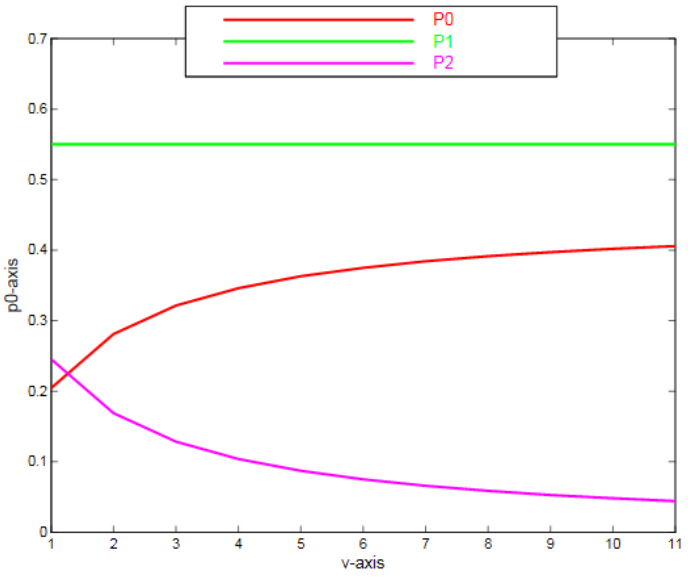

Fig4.5 : Effet de ν sur les probabilités de l'état du serveur.

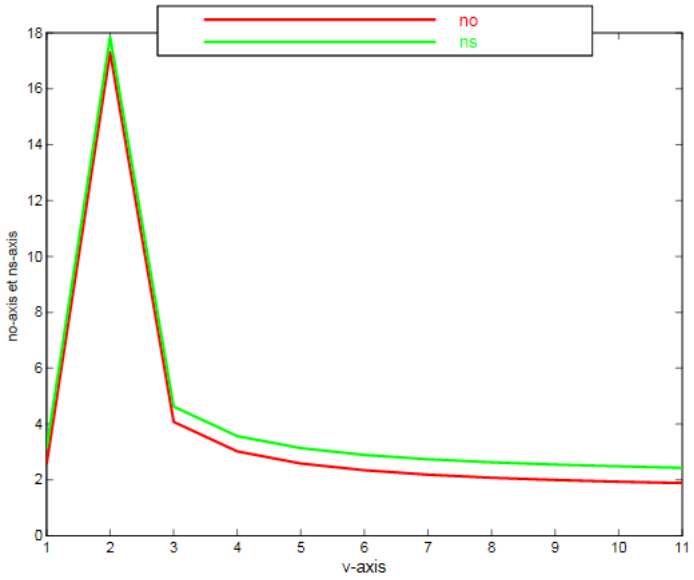

Fig4.6 : Effet de ν sur le nombre moyen de clients dans le système et dans l'orbite.

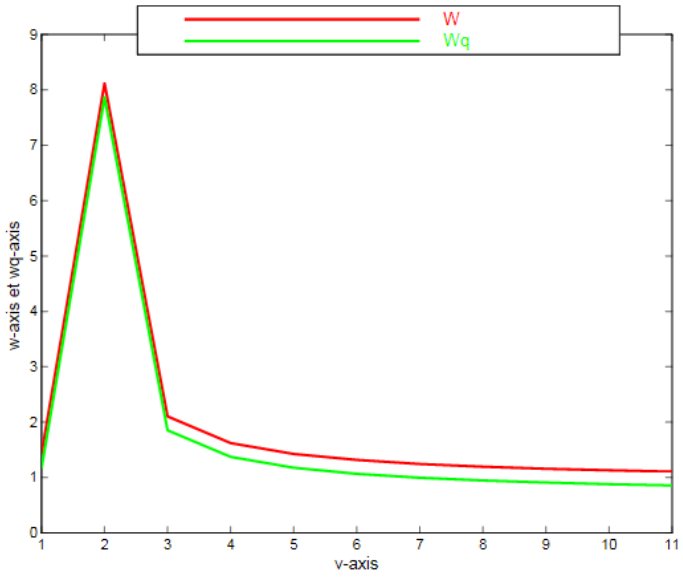

Fig4.7 : Effet de ν sur le temps moyen de séjour des clients dans le système et dans l'orbite.

Commentaire :

Dans le tableau (4.2) , on voit clairement la stabilité de la probabilité P_0, P_2 qui décroit contrairement a P_1 ce qui est montré par la figure (4.5). On remarque que l'augmentation de la valeur du taux ν induit une augmentation du nombre moyen et du temps moyen de séjour des clients dans le système et dans l'orbite, puis elle baisse juste au moment au la valeur de ν se rapproche du taux des arrivée λ.

De plus en plus le serveur est en recherche, le nombre moyen de clients dans le système est plus petit.

4.4.3 L'effet du taux d'inactivité α sur le modèle :

On fait varier le taux d'inactivité α sur les mesures de performance pour $\lambda = 5$, $\mu = 10$, $\theta = 1.33$, et $\nu = 2.5$.

α	p0	p1	p2	no	ns	w	wq
1	0,3571	0,5	0,1429	3,838	4,338	0,8676	0,7676
1,2	0,3378	0,5	0,1622	4,081	4,581	0,9162	0,8162
1,4	0,3205	0,5	0,1795	4,391	4,891	0,9782	0,8782
1,6	0,3049	0,5	0,1951	4,758	5,258	1,0516	0,9516
1,8	0,2907	0,5	0,2093	5,1741	5,6741	1,1348	1,0348
2	0,2778	0,5	0,2222	5,6329	6,1329	1,2266	1,1266
2,2	0,266	0,5	0,234	6,1287	6,6287	1,3257	1,2257
2,4	0,2551	0,5	0,2449	6,6572	7,1572	1,4314	1,3314
2,6	0,2451	0,5	0,2549	7,2144	7,7144	1,5429	1,4429
2,8	0,2358	0,5	0,2642	7,7972	8,2972	1,6594	1,5594
3	0,2273	0,5	0,2727	8,4026	8,9026	1,7805	1,6805

Tab4.3 : Variations des mesure de performances dans le systéme en fonction de α.

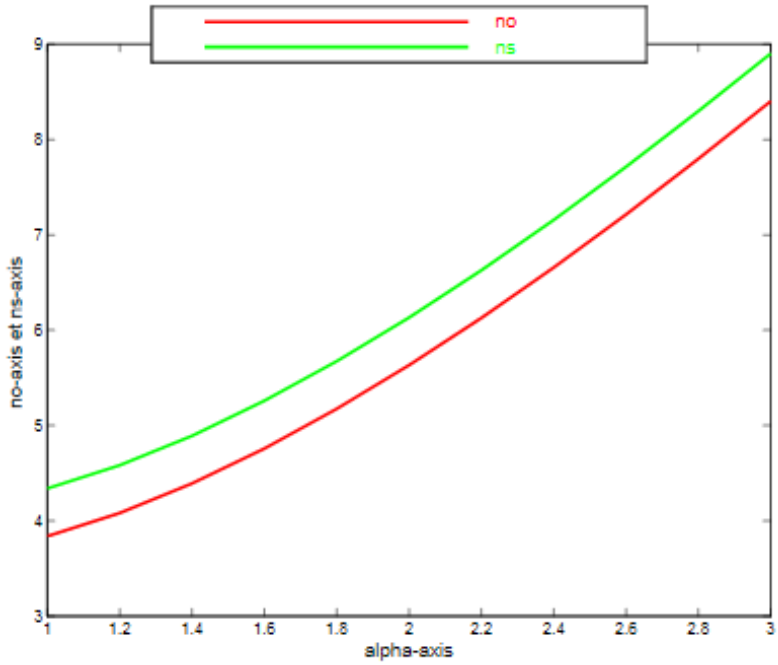

Fig4.8 : Effet de α sur le nombre moyen de clients dans le système et dans l'orbite.

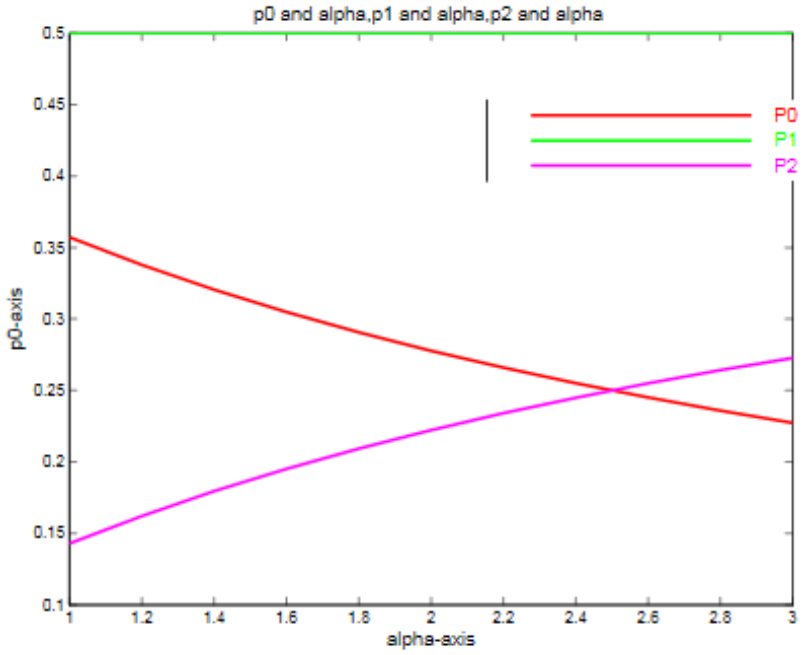

Fig4.9 : Effet de α sur les probabilités de l'état du serveur.

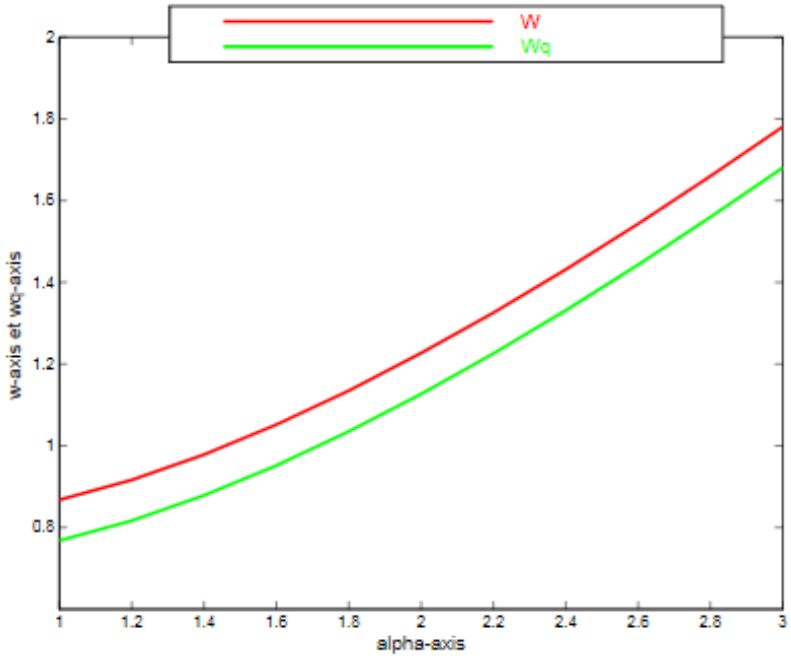

Fig4.10 : Effet de α sur le temps moyen de séjour des clients dans le système et dans l'orbite.

Commentaire :

Dans le tableau (4.3) et d'aprés les figures (4.8),(4.10) en remarque :

L'augmentation de α induit l'augmentation du nombre moyen de clients dans le système et dans l'orbite, ainsi que le temps de séjours des clients dans le système et dans l'orbite.

4.4.4 Effet de variations de ν et α sur le nombre de clients dans l'orbite :

Dans cette partie, on fait varier les deux paramètres ν et α au même temps, afin de voir la variation de nombre moyen de clients dans l'orbite avec $\lambda = 7$, $\mu = 8$ et $\theta = 10$.

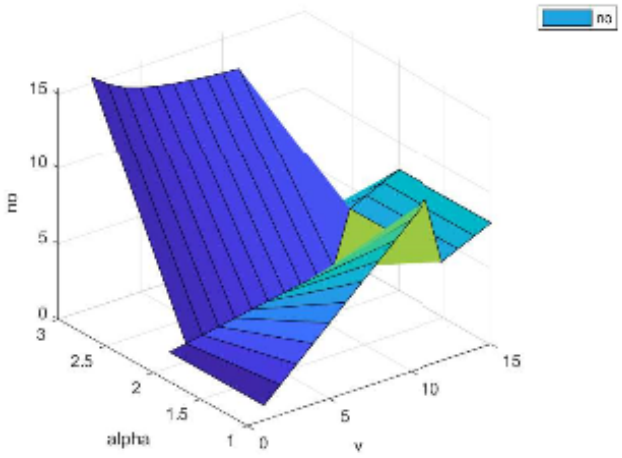

Fig4.11 : Variation de nombre moyen de clients
dans l'orbite en fonction de ν et α.

commentaire :

Le nombre moyen de clients dans l'orbite diminue avec l'augmentation du taux de recherche, et augmente avec l'augmentation du taux d'inactivité.

4.4.5 Effet de variations de ν et α sur le nombre de clients dans le systéme :

On fait varier les deux paramètres ν et α au même temps, afin de voir la variation de nombre moyen de clients dans le système avec $\lambda = 7$, $\mu = 8$ et $\theta = 10$.

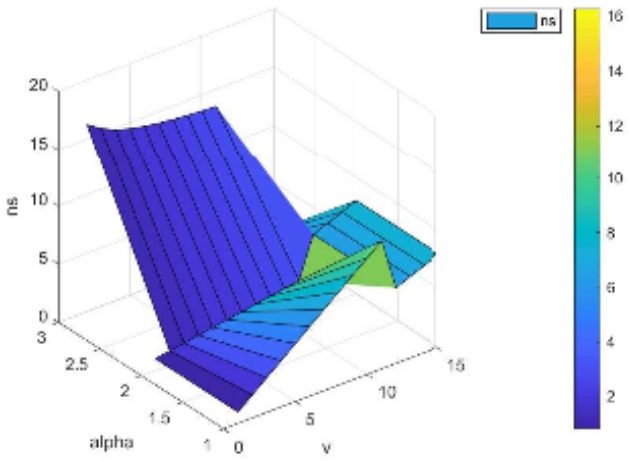

Fig4.12 : Variation de nombre moyen de clients
dans le système en fonction de ν et α.

Commentaire :

Le nombre moyen de clients dans le système augmentent avec l'augmentation du taux de d'inactivité et diminue avec l'augmentation du taux de recherche.

4.5 Conclusion

Dans ce chapitre, nous avons considéré l'analyse de sensibilité des mesures de performance par rapport à leurs paramètres critique, dans le système de file d'attente $M/M/1$ avec rappels et recherche des clients en orbite proposé par Tuan Phung-Duc [23]. Plusieurs exemples numériques ont été illustré.

4.6 Conclusion générale

Les systèmes de files d'attente avec rappels et recherche des clients en orbite par le serveur sont rencontres dans plusieurs situations réelles. L'étude de tels systèmes est certainement très importante pour les applications pratiques, car la recherche des clients par le serveur a une influence importante sur les principaux indices de performance.

Dans ce mémoire, nous avons étudie un modèle de files d'attente avec rappels et recherche des clients en orbite. En utilisant la méthode de la fonction génératrice, nous avons pu obtenir plusieurs résultats analytiques et certaines mesures de performance du modèle en question, telles que le nombre moyen de clients dans le système, et dans l'orbite. De plus, nous avons effectue une analyse de sensibilité du modèle d'attente étudie, tout en montrant l'effet de l'influence de changement des valeurs des paramètres du modèle sur ses caractéristiques stationnaires.

Dans ce travail, nous nous sommes intéresses au modèle $M/M/1$ avec rappels et recherche des clients en orbite. Notre étude a deux objectifs principaux. Le premier consiste à introduire la recherche des clients en orbite au systèmes de files d'attente avec rappels qui permettent de minimiser le temps d'inactivité du serveur. Le deuxième objectif est de fournir un aperçu du lien entre la file d'attente avec rappels correspondante et la file d'attente classique. La contribution principale de ce travail est le développement de modèle analytique pour l'évaluation des performance dans le Cloud Computing en utilisant la théorie des files d'attente. Nous avons d'abord rappelé quelques concepts et techniques de base de la théorie des files d'attente, et nous avons introduit quelques systèmes de file d'attente classiques. Ensuite, nous avons présente les systèmes de files d'attente avec rappels, en particulier les modèles $M/M/1$ et $M/G/1$ avec rappels. Puis une étude approfondis sur le modèle $M/M/1$ avec rappels linéaire et recherche de clients en orbite été présenter. Enfin, nous avons effectue une analyse numérique montrant l'effet de la variation de quelques paramètres sur les mesures de performance (le nombre moyen de client dans le système, dans l'orbite, et les probabilités P_0, P_1 et P_2) du système proposée. En termes de continuité de

ce travail, plusieurs perspectives de recherche peuvent être envisagées, on peut citer :

- Entreprendre la même démarche pour généraliser cette étude aux modèles d'attente avec rappels plus complexes.

- Etude des autres modèles d'attente avec rappels.

References

[1].Alan Ruegg (1989). livre :Processus stochastiques : avec applications aux phénomènes d'attente et de fiabilité.

[2].Alexandre, B. M (2014). les files d'attente .Laboratoire d'informatique formelle Université du Québec Cours 8INF802.

[3].Aïssani, A. (1994). A Survey on retrial queueing models. Actes des journées statistique appliquées. USTHB, alger,1-11.

[4].Artalejo, J R, and a. Gomez-corral (1997). Steady state solution of a single-server queue with linear repeated request. Journal of applied probability,34,223-233.

[5].Artalejo, J R, and all (2002). A second order analysis of the waiting time in the M/G/1 retrial queue. Asia pacific journal of operation research 19(2).

[6].Artalejo, J R, and a. Gomez-corral (2008). Analysis of an M/G/1 queue with constant repeated attempts and server vacations. Computer and operations research, 24(6) :394-504.

[7].Artalejo, J R, and Atencia (2004). On the single server retrial queue with batch arrivals. Sankhya, 66, 1-140-158.

[8].Artalejo, J R, Joshua, v.C. Krishnamoorthy (2002). A, in : j.R Artalejo, a. Krishnamoorthy (Eds.). an M/G/1 retrial queue with orbital search by the server Advances in Stochasic Modelling, Notable Publications Inc,. New Jersey, pp. 41-54 .

[9].Artalejo, J R, and Phung-Duc, T. (2012). Markovian retrial queues with two way communication. Journal of industrial and management optimization, Vol. 8. No. 4, 781-806.

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

[11].Chakravarthy, S. R., Krishnamoorthy, A.. Jeshna, V. C. (2006) Analysis of a multi-server retrial queue with search of custormers from the orbit. performance evaluation, 63(8), 776-798.

[12].Choi.B.D (1992). Retrial queues with collision arising from unslotted CSMA/CD protocol. Queueing Systems 11, 335-356.

[13].choo,Q.H. And Conolly,B (1979). New results in the theory of repeated orders queueing systems, N0.3,631-640.

[14].De kok. A.g (1984). Algorithmic methods for single server systems with repeated attempts. Statistica neerlandica,38 :23-32.

[15].Dudin, A. N., Krishnamoorthy, A., Joshua, V. C., Tsarenkov, G. V. (2004). Analysis of the BMAP/G/1 retrial system with search of customers from the orbit. European Journal of Operational Research, 157(1), 169-179.

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

[17].Falin. G. I (1990). Asurvey of retrial queues. Queueing systems, 7 :127-168.

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

[19].Neuts, et Ramalhoto (1984). A service model in which the server is required to search for customers. Journal of applied probability 21(1).

[20].Phung-Duc. T. Masuyama, H. Kasahara, S. and Takahashi, Y. (2010). A simple algorithm for the rate matrices of level-dependent QBD processes. In Proceedings of the 5th international conference on queueing theory and network applications. ACM. 46-52.

[21].Phung-Duc. T. (2012). An explicit solution for a tandem queue with retrials and losses. Operational Research, 12(2), 189-207.

[22].Phung-Duc. T. Rogiest, W. Takahashi, Y. and Bruneel, H. (2014). Retrial queues with balanced call blending : analysis of single-server and multiserver model, Annals of Operations Research, DOI :10.1007/s10479-014-1598-2.

[23].Phung-Duc. T.(2015). Retrial queue for cloud systems with separated processing and storage units. International conference on queueing theory and network application, 143-151.

[24].T. Yang, et al.(1994). An approximation method for the M/G/1 retrial queue with general retrial times. European Journal of Operation Research, 76 :552-562.

[25].Wang, J., Cao, J., Li, Q (2001).Reliability analysis of the retrial queue with serveur breakdowns and repairs.Que.Syst. 38,363-380.