Résumé

Les systèmes avec rappel apparaissent dans beaucoup de domaines, tels que les réseaux téléphoniques, les réseaux informatiques et la télécommunication. Ces systèmes sont caractérisés par le fait que les clients qui trouvent tous les serveurs occupés ou non disponibles, rappellent ultérieurement pour le service, à des intervalles de temps aléatoires. Cependant, la prise en considération du phénomène d'appels répétés a rendu les résultats de la théorie des files d'attente standards inadéquats, et a introduit de grandes difficultés analytiques. En fait, des résultats explicites détaillés existent pour certaines files d'attente avec rappel particulières, avec des hypothèses contraignantes sur certains paramètres, tel que le nombre de serveurs, la taille de la population, la fiabilité des serveurs, l'homogénéité des clients et des serveurs, etc., alors que pour beaucoup d'autres systèmes, les résultats obtenus restent très limités. L'intérêt de la démarche présentée dans cette thèse, réside essentiellement dans la proposition d'un modèle d'analyse des systèmes avec rappel complexes, autre que le modèle conventionnel des files d'attente avec rappel, et donc d'adapter ses méthodes d'analyse, dans le but de dériver des résultats exacts des indices de performance. Vu l'importance pratique des systèmes avec rappel, à source finie de clients et serveurs non fiables d'une part, et d'autre part le besoin de considérer des systèmes avec plusieurs classes de clients et classes de serveurs à la fois.

REMERCIEMENTS

Tout d'abord, Je tiens à exprimer mes sincères remerciement .
– A mon Professeur pour m'avoir confié ce memoire pour m'avoir encadré avec beaucoup de patience, de disponibilité et de professionnalisme
– Je voudrais également associer mes meilleurs remerciements à ma femme pour son soutien moral et pour ses remarques qui ont eu une contribution importante pour ma réussite et à mes parents.
– Et pour finir, merci à toutes les personnes que j'ai oubliées de citer et qui m'ont permis de mener à bien cette memoire

Dédicaces

Je dédie ce modeste travail à mes parents qui m'ont aidé pour arriver à ce que je suis aujourd'hui :

- A ma femme Salima pour tout son aide, son soutien.
- A mon enfant Djawad.
- A mes frères et soeurs à qui je souhaite une réussite dans la vie.
- A mes amis et tout les collegues.

Table des matières

Introduction générale

La théorie des files d'attente est une théorie mathématique relevant du domaine des probabilités, né en 1917, des travaux de l'ingénieur danois Erlang sur la gestion des réseaux téléphoniques de Copenhague. Entre 1909 et 1920, elle a étudié notamment les systèmes d'arrivée dans une queue, les différentes priorités de chaque nouvel arrivant, ainsi que la modélisation statistique des temps d'exécution. C'est grâce aux apports des mathématiciens Khintchine, Palm, Kendall, Pollaczek et Kolmogorov que la théorie s'est vraiment développée. La théorie de files d'attente est aujourd'hui largement utilisée et ses applications sont multiples.

Vue l'apparition d'autre systèmes réels de plus en plus complexes, tel que les systèmes téléphoniques où les abonnés répétaient leurs appels en recomposant le numéro plusieurs fois jusqu'à l'obtention de la communication, des chercheurs tels que Kosten et Wilkinson ont mis en évidence les limites de la théorie classique des files d'attente qui ne permettait pas d'expliquer le comportement stochastique de ce type de système.

Ce phénomène de répétition de demandes du service a poussé certains chercheurs à étendre le modèle d'attente classique à celui dit avec rappels. Ce type de systèmes de files d'attente avec rappels peut être appliqué pour résoudre les problèmes pratiques, tels que l'analyse du comportement des abonnés dans les réseaux téléphoniques, l'analyse du temps d'attente pour accéder à la mémoire sur les disques magnétiques. Ce type de modèles se rencontre également dans la modélisation de protocoles spécifiques de communication, tels que CSMA (Carrier Sens Multiple Access) ou encore les disciplines Auto-Repeat, Ring-Back-When-Free, Repeat-LastNumber. Les progrès récents dans ce domaine sont résumés dans les articles de synthèse de Falin (1990) [10], Aïssani (1994) [1], Kulkarni et Liang (1997) [15], Templeton (1999) [18] et dans les monographies de Falin et Templeton (1997) [14] et Artalejo et Gòmez (2008) [3].

Les files d'attente avec rappels ont été largement utilisés pour modéliser de nombreux problèmes dans les systèmes de communications téléphoniques, informatiques, des réseaux locaux et des situations de la vie quotidienne. Dans la plupart des publications sur les files d'attente avec rappels, le serveur ne fournit que le service aux arrivées entrantes effectuées par les clients réguliers. Cependant, il existe des situations réelles par exemple : les centres d'appels où un opérateur non seulement sert les appels entrants, mais il effectue aussi des appels sortants vers l'extérieur lorsque le serveur est libre. Cette fonction est connue sous le nom de files d'attente avec rappels à communication bidirectionnelle [14].

L'étude des modèles de files d'attente avec rappels à communication bidirectionnelle remonte aux travaux de Falin (1979) [12]. Ce qui a ouvert les portes à diverses publications de plusieurs auteurs, à savoir Artalejo (2010) [2], Artalejo et Phung-Duc (2011) [5], Artalejo et Phung-Duc (2013) [6].

La théorie analytique des modèles d'attente avec rappels s'avère d'une portée limitée en raison de la complexité des résultats connus. En effet, dans la majorité des cas, on se retrouve confronté à des systèmes d'équations dont la résolution est complexe ou possédant des solutions qui ne sont pas facilement interpretable afin que le praticien puisse en bénéficier. Par ailleurs, on peut citer le degré de difficulté pour l'obtention de certaines caractéristiques dans quelques modèles tels que les modèles de files d'attente avec rappels et vacances, avec rappels et priorité, avec rappels à communication bidirectionnelle, avec rappels de distribution générale ayant deux types de clients. Cette difficulté réside essentiellement dans l'utilisation des inverses des transformées de Laplace-Stieljes et des distributions marginales. Pour pallier à toutes ces difficultés, les chercheurs ont recouru aux méthodes d'approximation telle que les méthodes de comparaison stochastique , qui permettent d'avoir des estimations qualitatives pour certaines mesures de performance. L'idée générale de cette méthode est de borner un système complexe par un autre système, plus simple à résoudre et fournissant des bornes qualitatives pour ces mesures de performance.

Le but de notre travail est d'appliquer les méthodes de comparaison stochastique, pour étudier les propriétés de monotonie du modèle de files d'attente avec rappels à communication bidirectionnelle relativementà l'ordre stochastique, l'ordre convexe et l'ordre en transformée de Laplace afin d'obtenir des bornes simples pour la distribution stationnaire de la chaîne de Markov induite liée à ce modèle .

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

Dans le premier chapitre, on cible deux importants Processus stochastiques dans la modélisation des files d'attente. Le premier est le processus de Poisson, qui est un outil extrêmêment utilisé dans les phénomènes de comptage. Il apparaît en effet naturellement dans ces situations, le deuxième est celui de naissances et morts qui sert à modéliser l'évolution d'une population au cours du temps.

Ensuite dans le deuxième chapitre, nous introduisons la terminologie de la théorie des files d'attente et de certaines définitions et notations qui sont nécessaires dans l'étude des systèmes de files d'attente (la notation de KANDELL,la formule de LITTLE $\cdots$).En suite nous étudions quelque modèles de files d'attente markoviennes ($M/M/1$, $M/M/1/K$, $M/M/c$, $M/M/\infty$,) et l'évaluation des paramètres de performance.

Enfin Le troisième chapitre est consacré à des files d'attente avec clients impatients. On traite les cas des files d'attente $M/M/1$ et $M/M/2$ - $M/G/1$ avec rappels. Nous donnons aussi quelques exemples d'application de notre modèle.
.

Chapitre 1

Processus stochastique

1.1 Le processus de poisson

1.1.1 Le processus de comptage

Définition 1.1.1. *Un processus $(N_t)_{t\in\mathbb{R}_+}$ est appelé processus de comptage si c'est un processus croissant, c'est-à-dire si pour tout $s \leq t$, $N_s \leq N_t$. La variable aléatoire $N_t - N_s$ est alors appelée accroissement du processus sur $]s, t]$.*

par exemple :
- $N(t)$ = nombre de poissons capturés dans l'intervalle de temps $[0, t]$,
- $N(t)$ = taille d'une population à la date t.

Définition 1.1.2. *Un processus de comptage $(N_t)_{t\in\mathbb{R}_+}$ est appelé processus à accroissements indépendants si pour tout $n \in \mathbb{N}^*$ et pour tous $t_1, ..., t_n$ tels que $t_1 < t_2 ... < t_n$, les accroissements $N_{t_1} - N_0$, $N_{t_2} - N_{t_1}, ..., N_{t_n} - N_{t_{n-1}}$ sont des variables aléatoires indépendantes.*

Définition 1.1.3. *Le processus est dit stationnaire (ou homogène dans le temps), si pour tout s et pour tout t, l'accroissement $N_{t+s} - N_s$ a même loi que N_t.*

Définition 1.1.4. *Un processus à accroissements indépendants stationnaire $(N_t)_{t\in\mathbb{R}_+}$ est dit à événements rares si*

$$\lim_{h\to 0_+} \mathbb{P}([N_h > 0]) = 0$$

et si

$$\lim_{h\to 0_+} \frac{\mathbb{P}([N_h > 1])}{\mathbb{P}([N_h = 1])} = 0.$$

1.1.2 processus de poisson

Définition 1.1.5. *Un processus de comptage* $(N_t)_{t\in\mathbb{R}_+}$ *tel que* $N_0 = 0$ *est un processus de Poisson si*

- *C1 :* $(N_t)_{t\in\mathbb{R}_+}$ *est stationnaire,*
- *C2 :* $(N_t)_{t\in\mathbb{R}_+}$ *est un processus à accroissements indépendants,*
- *C3 :* $(N_t)_{t\in\mathbb{R}_+}$ *est un processus à événements rares.*

On s'intéresse ici au comptage du nombre d'occurrences d'un événement, par exemple la naissance d'un individu. On note $N(t)$ le nombre d'événements survenus dans l'intervalle $[0, t]$. Un tel processus a une trajectoire en escalier (voir figure 1.1). L'événement d'intérêt survient aux dates t_1, t_2, ..., à chacune de ces dates, le comptage $N(t)$ augmente de 1 :

$$\begin{aligned} N(t) &= 0 \quad si\ t < t_1, \\ &= 1 \quad si\ t_1 \leq t \leq t_2, \\ &\vdots \\ &= k \quad si\ t_k \leq t < t_{k+1}, \\ &etc. \end{aligned}$$

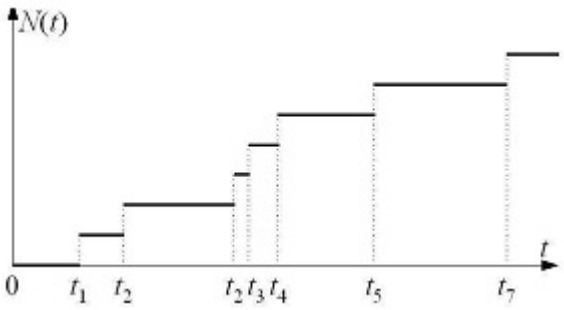

FIGURE 1.1 – Exemple de trajectoire d'un processus de comptage.

Exemples. Les exemples de ce genre de ce processus ne se limitent évidemment pas à la biologie :

- Appels téléphoniques à un standard,
- Prise d'un poisson par un pêcheur,
- Arrivée d'un client à un guichet,
- Passage d'un autobus.

Définition. On dit qu'un tel processus est poissonnien s'il vérifie les hypothèses suivantes :

A *Le processus est sans mémoire :* l'occurrence d'événements avant la date t n'influe en rien sur l'occurrence d'événements après t :

$$N(t+h) - N(t) \perp N(t) - N(t-k).$$

B *Le processus est homogène dans le temps :* la loi de l'accroissement $[N(t+h) - N(t)]$ du processus ne dépend que de h et pas de t (et est donc la même que celle de $N(h)$) [1] :

$$N(t+h) - N(t) \overset{\mathcal{L}}{=} N(h) - N(0).$$

Remarques.
- L'hypothèse **A** induit que le processus de comptage de événements vérifie l'*hypothèse de Markov* : toute l'information issue du passé du processus qui conditionne la loi de $N(t)$ est résumée par $n(t^-)$.
- Pour l'hypothèse **B**, on parle parfois d'hypothèse d'*homogénéité temporelle* ou de stationnarité.

Le nom donné au processus de Poisson s'explique par ce qui suit :

Propriété 1.1.1. *Un processus de comptage $(N_t)_{t\in\mathbb{R}_+}$ tel que $N_0 = 0$ est un processus de Poisson si et seulement si :*
- *C1 : $(N_t)_{t\in\mathbb{R}_+}$ est stationnaire,*
- *C2 : $(N_t)_{t\in\mathbb{R}_+}$ est un processus à accroissements indépendants,*
- *C3 : il existe $\lambda > 0$ tel que, pour tout $t \geq 0$, la variable aléatoire N_t suive la loi de Poisson de paramètre λt.*

Proposition 1.1.1. *Si $(N_t)_{t\in\mathbb{R}_+}$ est un processus de Poisson de paramètre λ, le temps aléatoire U qui sépare un instant θ du prochain événement et le temps aléatoire V qui sépare θ du dernier événement suivent la loi exponentielle $\mathcal{E}(\lambda)$.*

1.1.3 Caractérisation d'un processus de Poisson par ses temps d'arrivée

Soit A_n l'instant de la n^{ime} arrivée : $A_n = \inf\{t \geq 0; N_t = n\}$ et T_n le n^{ime} temps d'attente pour $n \in \mathbb{N}^*$: $T_n = A_n - A_{n-1}$ (en convenant $A_0 = 0$). On a $A_n = \sum_{i=1}^{n} T_i$ et $N_t = \max\{n \geq 0; A_n = t\}$.

Théorème 1.1.1. *$(N_t)_{t\in\mathbb{R}_+}$ est un processus de Poisson de paramètre λ si et seulement si les variables aléatoires T_n sont indépendantes de même loi exponentielle $\mathcal{E}(\lambda)$ (de densité $f_{T_n}(t) = \lambda e^{-\lambda t}\mathbb{1}_{]0,+\infty[}(t)$).*

1. La notation $X \overset{\mathcal{L}}{=} Y$ signifie que les variables aléatoires X et Y ont la même loi de probabilité, pas la même valeur.

Loi de probabilité. En terme de probabilités, si on considère la probabilité qu'un événement survienne dans un intervalle d'amplitude Δt

$$Pr\{N(t+\Delta t) - N(t) = 1\}$$

en opérant un développement limité au premier ordre, et en remarquant que $Pr\{N(t) - N(t) = 1\} = 0$, on obtient

$$Pr\{N(t+\Delta t) - N(t) = 1\} = 0 + \lambda\Delta t + o(\Delta t)$$

et les hypothèses **A** et **B** impliquent que λ ne dépend pas de t. λ est appelé *intensité du processus.*

Système différentiel. Pour Δt suffisamment petit, le résultat précédent nous permet d'écrire le système

$$\begin{cases} Pr\{N(t+\Delta t) - N(t) \geq 2\} = o(\Delta t), \\ Pr\{N(t+\Delta t) - N(t) = 1\} = \lambda\Delta t + o(\Delta t), \\ Pr\{N(t+\Delta t) - N(t) = 0\} = 1 - \lambda\Delta t + o(\Delta t). \end{cases}$$

1.1.4 Loi du nombre d'événements et interprétation

Loi de $N(t)$

D'après le système différentiel, en notant

$$p_n(t) = Pr\{N(t) = n\},$$

on a

$$\begin{aligned} p_n(t+\Delta t) &= Pr\{N(t) = n\}Pr\{N(t+\Delta t) - N(t) = 0\} \\ &\quad + Pr\{N(t) = n-1\}Pr\{N(t+\Delta t) - N(t) = 1\} + o(\Delta t) \\ &= p_n(t) \times (1 - \lambda\Delta t) + p_{n-1}(t) \times \lambda\Delta t + o(\Delta t) \\ &= p_n(t) + \lambda\Delta t[p_{n-1}(t) - p_n(t)] + o(\Delta t) \end{aligned}$$

d'où

$$\frac{p_n(t+\Delta t) - p_n(t)}{\Delta t} = \lambda[p_{n-1}(t) - p_n(t)] + \frac{o(\Delta t)}{\Delta t},$$

or

$$p_n'(t) = \lim_{\Delta t \to 0} \frac{p_n(t+\Delta t) - p_n(t)}{\Delta t}$$

et donc, en passant à la limite pour $\Delta t \to 0$, il vient

$$p_n'(t) = \lambda[p_{n-1}(t) - p_n(t)].$$

Il faut cependant isoler le cas particulier $n = 0$:

$$\begin{aligned} p_0(t+\Delta t) &= Pr\{N(t)=0\}Pr\{N(t+\Delta t)-N(t)=0\} \\ &= p_0(t)\times[1-\lambda\Delta t+o(\Delta t)] \end{aligned}$$

qui donne

$$p_0'(t) = -\lambda p_0(t).$$

Les fonctions $p_n(t)$ vérifient donc le système différentiel

$$\left\{ \begin{array}{ll} p_0'(t) & = -\lambda p_0(t), \\ p_n'(t) & = \lambda[p_{n-1}(t)-p_n(t)] \quad pour\ n>0. \end{array} \right.$$

Résolution du système.

1. On a $p_0'(t) = -\lambda p_0(t) \quad \Leftrightarrow \quad p_0(t) = C_0 e^{-\lambda t}$ or $p_0(0) = 1$ donc

$$p_0(t) = e^{-\lambda t}.$$

2. On a $p_1'(t) = \lambda p_0(t) - \lambda p_1(t) = \lambda e^{-\lambda t} - \lambda p_1(t)$.
On résout tout d'abord

$$p_1'(t) = -\lambda p_1(t) \quad \Leftrightarrow \quad p_1(t) = C_1 e^{-\lambda t}$$

puis on fait varier la constante $C_1 = C_1(t)$ ce qui donne

$$p_1'(t) = e^{-\lambda t}[C_1'(t) - \lambda C_1(t)]$$

et en reportant dans l'équation de départ, on a

$$\begin{aligned} e^{-\lambda t}[C_1'(t)-\lambda C_1(t)] = e^{-\lambda t}[\lambda - C_1(t)] \quad &\Rightarrow C_1'(t) = \lambda \\ &\Rightarrow C_1(t) = \lambda t + c_1 \end{aligned}$$

d'où

$$p_1(t) = (\lambda t + c_1)e^{-\lambda t} \quad or \quad p_1(0) = 0$$

donc

$$p_1(t) = \lambda t e^{-\lambda t}.$$

Interprétation des résultats

Espérance et variance. On en déduit immédiatement l'espérance et la variance de $N(t)$:

$$\mathbb{E}[N(t)] = \lambda t, \qquad \mathbb{V}[N(t)] = \lambda t.$$

On a donc

$$N(1) \sim \mathcal{P}(\lambda) \qquad \Leftrightarrow \qquad \mathbb{E}[N(1)] = \lambda$$

ce qui signifie que le nombre moyen d'événements survenant en une unité de temps est égal à λ.

On peut aussi donner un intervalle de prédiction pour niveau $1-\alpha = 0.95$ donné, comme présenté en figure (2.2) : on fait l'approximation de la loi de Poisson par une loi normale :

$$\mathcal{P}(\lambda t) \approx \mathcal{N}(\lambda t, \lambda t)$$

Ainsi, à t fixé, un intervalle de prédiction pour $N(t)$ est

$$[\lambda t \pm 1.96\sqrt{\lambda t}]$$

Interprétation binomiale. On peut retrouver ce résultat par une autre approche : on découpe l'intervalle $[0, t]$ en m intervalles de taille $\frac{t}{m}$ (voir figure 1.3) suffisamment petits pour que chacun puisse contenir au plus un événement et ce, avec probabilité $\frac{\lambda t}{m}$:

$$\text{soit } I_k = \left[k\frac{t}{m}, (k+1)\frac{t}{m}\right], \quad Pr\{A \in I_k\} = \frac{\lambda t}{m}.$$

Pour tout k, la variable qui vaut 1 si l'événement se produit dans l'intervalle I_k (et 0 sinon) suit une loi de Bernoulli de paramètre $\frac{\lambda t}{m}$.
$N(t)$ est la somme de toutes ces variables, donc

$$N(t) \sim \mathcal{B}\left(m, \frac{\lambda t}{m}\right)$$

et on sait que quand n tend vers l'infini et $n\pi$ tend vers une constante, la loi binomiale $\mathcal{B}(n, \pi)$ tend vers la loi de Poisson $\mathcal{P}(n\pi)$, or ici m peut être pris arbitrairement grand et $m \times \frac{\lambda t}{m} = \lambda t$, donc

$$N(t) \sim \mathcal{P}(\lambda t).$$

Loi des dates d'arrivée des événements

Un processus $\{N(t), t \geq 0\}$ ne peut se décrire uniquement à partir du nombre d'occurences de l'événement à l'instant t, $N(t)$. Il faut savoir à quelles dates se sont passés les événements. On appelle T_k la variable aléatoire représentant la date à laquelle se produit le $k$ème événement. Conditionnellement

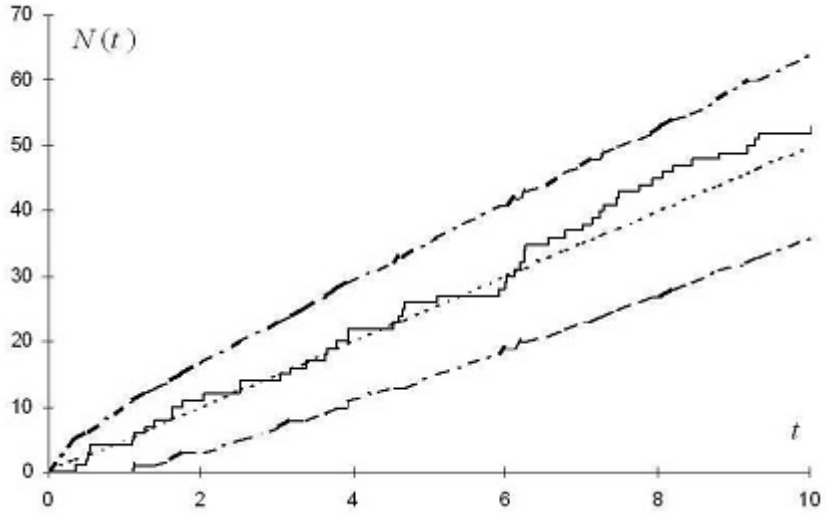

FIGURE 1.2 – Trajectoire d'un processus de Poisson ($\lambda = 5$) + Espérance + Intervalle à 95%.

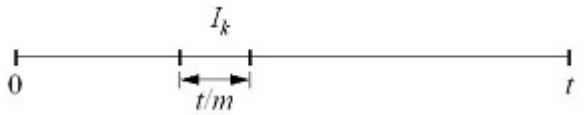

FIGURE 1.3 – Approche binomiale.

à $N(t) = n$, on connaît la loi de ces temps : si on note $T_{(1)} < T_{(2)} < \ldots < T_{(n)}$ la statistique d'ordre de $(T_1, T_2, \ldots, T_n)$, on a que

$$T_{(1)} < T_{(2)} < \ldots < T_{(n)} | N(t) = n \sim \mathcal{U}_{[0,t]^n}$$

Cela signifie que si on connaît le nombre d'événements qui sont survenus sur l'intervalle $[0, t]$, les dates (ordonnées) auxquelles les événements se produisent sont uniformément répartis sur l'intervalle sur $[0, t]$.

1.1.5 Comparaison avec un modèle déterministe

On note

$$n(t) = \text{ le nombre d'événements observés dans l'intervalle } [0, t],$$

L'équation différentielle correspondant aux hypothèses (**A** = absence de mémoire du processus) et (**B** = homogénéité temporelle) s'écrit

$$n'(t) = \lambda$$

avec la condition initiale naturelle $n(0) = 0$. On obtient ainsi l'équation

$$n(t) = \lambda t$$

qui correspond au comportement "moyen" (*i.e.* en espérance) du processus de Poisson.

1.1.6 Temps séparant deux événements successifs

Loi de la durée séparant deux événements

On s'intéresse maintenant à la durée (aléatoire) séparant deux occurrences de l'événement. On se place à une date t_0 et on s'intéresse à la variable T :

T = temps d'attente jusqu'à l'occurrence du prochain événement.

On a

$$\begin{aligned} Pr\{T > t\} &= Pr\{N(t_0 + t) - N(t_0) = 0\} \\ &= Pr\{N(t) = 0\} \quad \left(\begin{array}{l} \text{grâce à l'hypothèse} \\ \text{d'indépendance temporelle} \end{array} \right) \\ &= p_0(t) = e^{-\lambda t} \end{aligned}$$

La loi de T est donc indépendante de t_0, et on a

$$\begin{array}{rcrl} & Pr\{T > t\} & = & e^{-\lambda t} \\ \Leftrightarrow & Pr\{T \leq t\} & = & 1 - e^{-\lambda t} \\ \Leftrightarrow & T & \sim & \mathcal{E}(\lambda) \end{array}$$

Il est important de remarquer qu'on ne se préoccupe pas de savoir si t_0 est elle-même *une date d'occurrence ou pas* : cela ne change pas la loi de T à cause de l'hypothèse d'indépendance temporelle.
Interprétation. T suit une loi exponentielle de paramètre λ, on a donc

$$\mathbb{E}(T) = \frac{1}{\lambda}$$

ce qui signifie que la durée moyenne séparant deux événements est égale à $\frac{1}{\lambda}$.
Remarque. On démontre au passage que la loi de Poisson de paramètre λ est la loi du nombre d'événements survenant dans une unité de temps quand ces événements sont séparés par des durées exponentielles indépendantes. Cette propriété fournit un algorithme de simulation d'une variable aléatoire poissonnienne à partir de variables aléatoires exponentielles :

1. on simule des X_i .i.i.d, $X_i \sim \mathcal{E}(\lambda)$,
2. on calcule $S_i = \sum_{j \leq i} X_j$,
3. on prend $N = n$ tel que $S_n \leq 1 < S_{n+1}$.

N est distribuée selon une loi de Poisson $\mathcal{P}(\lambda)$.

Date du n-ème événement

On rappelle que T_n représente la date (aléatoire) à laquelle survient le n-ème événement. On vient de voir que

$$T_1 \sim \mathcal{E}(\lambda)$$

et, de façon générale, que

$$T_n - T_{n-1} \sim \mathcal{E}(\lambda) \text{ pour } n > 0 \text{ et avec } T_0 = 0,$$

donc, T_n est la *somme de n variables exponentielles de paramètre* λ , sa loi est appelée loi gamma et notée

$$T_n \sim \gamma(n, \lambda)$$

1.2 Processus de naissances et de morts

Ce processus est la fusion des deux processus précédents : le processus de naissance et le processus de mort.

Un processus de naissance et de mort est un processus qui consiste à faire évoluer un système entre une infinité dénombrable ou non dénombrable (processus continu) d'états, le système étant à chaque instant dans un état est un seul. A titre d'exemple : une file d'attente devant un carrefour à feux, les états du systémeétant le nombre de voitures dans le lieu de service. L'arrivée des voitures peut être considérée comme une naissance et le départ des voitures peut être considéré comme une mort.

Définition 1.2.1. *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 variables en fonction de l'état du système.

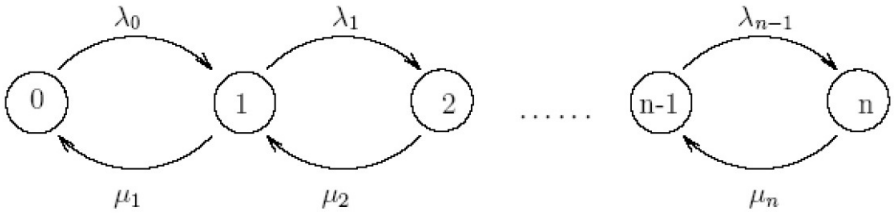

FIGURE 1.4 – Graphe de transition d'un processus de naissance et de mort.

1.2.1 Processus de naissances pur

Nous allons maintenant procéder à une présentation plus systématique de quelques processus de naissance et de mort particuliers.

Définition 1.2.2. *Dans un processus de naissance pur,* $\lambda_n = n$ *et* $\mu_n = 0$ *pour* $n = 0, 1, ...$ *Donc, les arrivées ont lieu à taux constant et il n'y a pas de départs. Pour un tel processus, le nombre de clients dans le système est évidemment égal au nombre d'arrivées enregistrées pour un processus de Poisson classique, si bien que*

$$\begin{aligned} \pi_n(t) &= \textit{probabilité que l'état du système à l'époque } t \textit{ soit égal à } n \\ &= e^{-\lambda t}\frac{(\lambda t)^n}{n!} \quad (n = 0, 1, ...) \end{aligned}$$

Définition 1.2.3. *Dans un processus de mort pur, l'ensemble des états possibles du système est* $\{0, 1, ..., N\}$ *et*

$$\lambda_n = 0 \text{ pour } n = 0, 1, ..., N$$
$$\mu_n = \begin{cases} 0 & sin = 0 \\ \mu & sin = 1, 2, ..., N. \end{cases}$$

Les files d'attente de type Markovien (M/M) sont des cas particuliers très importants de processus de naissance et de mort. Leur étude complète sera effectuée dans le chapitre suivant.

Loi de la durée entre deux événements successifs

Comme pour le processus de Poisson, on s'intéresse à la loi de la durée séparant deux événements successifs. On note X_k la variable aléatoire représentant la durée entre le $k$ème et le $k+1$ème événement. On a que

$$X_k = T_{k+1} - T_k,$$

si T_k représente le temps où le $k$ème événement se produit.

Prenons tout d'abord $k = 0$. X_0 représente le temps d'attente jusqu'à la division d'une cellule sachant qu'il y a n_0 cellules. Si on note X_0^i le temps d'attente jusqu'à que la cellule i se divise, le temps d'attente jusqu'au premier événement pour toute la population correspond au premier temps d'attente des n_0 cellules, donc X_0 est

$$X_0 = \inf\{X_0^i, i = 1, ..., n_0\}$$

Par analogie avec le processus de Poisson, on a que X_0^i suit une loi exponentielle de paramètre λ. De plus, on a supposé que les n_0 cellules se comporter indépendamment les unes des autres, ce qui implique que les variables X_0^i sont indépendantes. En utilisant la loi de l'inf d'exponentielle, on montre que X_0 suit une loi exponentielle de paramètre la somme des paramètres des n_0 exponentielles, donc

$$X_0 \sim \mathbb{E}(\lambda n_0)$$

De la même façon, on peut montrer que

$$X_k \sim \mathbb{E}(\lambda n_k)$$

où n_k est le nombre de cellules à l'instant T_k, i.e. soit $n_0 + k$.

Chapitre 2

Les files d'attente

Introduction

La théorie des files d'attente, ou queues, et des réseaux de files d'attente sont des outils analytiques les plus puissants pour la modélisation de systèmes logistiques et de communication. En quelques mots, cette théorie a pour objet l'étude des systèmes où des entités, appelées clients, cherchant à accéder à des ressources, généralement limitées, afin d'obtenir un service. La demande concurrente d'une même ressource par plusieurs clients engendre des délais dans la réalisation des services et la formation de files de clients désirant accéder à une ressource indisponible. L'analyse théorique de tels systèmes permet d'établir à l'avance les performances de l'ensemble, d'identifier les éléments critiques ou, encore, d'appréhender les effets d'une modification des conditions de fonctionnement.

2.1 Description du modèle d'attente classique

Une file d'attente ou queue est un système stochastique composé d'un certain nombre (fini ou non) de places d'attente d'un ou plusieurs serveurs et bien sûr de clients qui arrivent, attendent, se font servir selon des règles de priorité données et quittent le système. La description précédente d'une file d'attente, dont une représentation schématique est donnée en figure (2.1), ne saurait capturer toutes les caractéristiques des différents modèles que comptent la littérature, mais elle identifie leséléments principaux permettant la classification de la grande majorité des files d'attente simples.

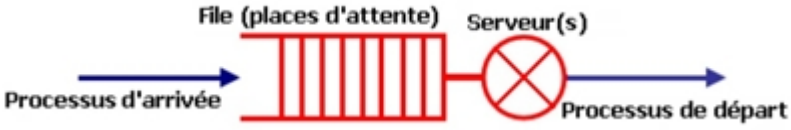

FIGURE 2.1 – Le système de files d'attente

2.2 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 $(t \geq 0)$.

En fonction des quantités qui définissent le système, on cherche à déterminer :

- Les probabilités d'état $\pi_n(t) = \mathbb{P}(N(t) = n)$, qui définissent le régime transitoire du processus stochastique $\{N(t), t \geq 0\}$. Il est évident que les fonctions $\pi_n(t)$ dépendent de l'état initial ou de la distribution initiale du processus.
- Le régime stationnaire du processus stochastique est défini par :

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

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

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

2.3 Classification des systèmes d'attente

Afin de spécifier un système de file d'attente , on se base sur trois éléments :

- le processus d'arrivée :Pour le processus d'arrivée ,on s'intéresse aux instants d'arrivés des clients dans la file.Ils sont en général aléatoires. Certaines hypothèses sont faites sur leurs lois. Tout d'abord, il n'arrive qu'un client à la fois. La deuxième hypothèse est l'homogénéité dans le temps. Cela se traduit par le fait que les temps d'interarrivée des clients sont des v.a. de même loi.Ils sont également supposés indépendants.Enfin, la loi des temps d'interarrivée est supposée connue. Le cas le plus courant est celui où cette loi est exponentielle. Dans ce cas, le modèle des temps d'arrivée des clients est un processus de Poisson. Évidemment d'autres cas peuvent se présenter : temps d'interarrivés constants, de loi uniforme, loi normale ou encore gamma.
- le processus de service :qui compris (le nombre de serveurs et la loi probabiliste décrivant la durée des services)
- Structure et discipline de la file :La discipline de service détermine l'ordre dans lequel les clients sont rangés dans la file et y sont retirés pour recevoir un service.

2.4 Notation de Kendall (1953)

La notation suivante , appelée la notation de Kendall , est largement utilisé pour classer les différents systèmes de files d'attente :

$$T/Y/C/K/m/Z$$

avec

1. T :indique le processus d'arrivée des clients . Les codes utilisés sont :
   - M (Markov) : Interarrivées des clients sont indépendamment , identiquement distribuées selon une loi exponentielle . Il correspond à un processus de Poisson ponctuel(propriété sans mémoire).

- D (Répartition déterministe) :les temps Interarrivées des clients ou les temps de service sont constants et toujours les même.
- GI (général indépendant) : Interarrivées des clients ont une distribution générale (il n'y a aucune hypothèse sur la distribution mais les interarrivées sont indépendentes et identiquement distribuées) .
- G (général) : Interarrivées de clients ont une distribution générale et peuvent être dépendantes .
- E_k : Ce symbole désigne un processus où les intervalles de temps entre deux arrivées successives sont des variables aléatoires indépendantes et identiquement distribuées suivant une loi d'Erlang d'ordre k.

2. Y : décrit la distribution des temps de service d'un client . les codes sont les mêmes que T.
3. C : nombre de serveurs
4. K : capacité de la file (c'est le nombre de places dans le système en d'autre tèrme c'est le nombre maximal de clients permis dans le système y compris ceux en service.
5. m : population des usagers
6. Z : discipline de service (c'est la façon dont les clients sont ordonnés pour être servi). Les codes utilisés sont les suivants :
   - FIFO (first in, first out) ou FCFS (first come first served) : c'est la file standand dans laquelle les clients sont servis dans leur ordre d'arrivée. Notons que les disciplines FIFO et FCFS ne sont pas équivalentes lorsque la file contient plusieurs serveurs. Dans la première, le premier client arrivé sera le premier à quitter la file alors que dans la deuxième, il sera le premier à commencer son service. Rien n'empêche alors qu'un client qui commence son service après lui, dans un autre serveur, termine avant lui.
   - LIFO (last in, first out) ou LCFS (last come, first served) . Cela correspond à une pile, dans laquelle le dernier client arrivé (donc posé sur la pile) sera le premier traité (retiré de la pile). À nouveau, les disciplines LIFO et LCFS ne sont équivalentes que pour une file monoserveur.
   - SIRO (Served In Random Order) , les clients sont servis au aléatoirement.
   - PNPN (Priority service) , les clients sont servis à leur priorité . Tous les clients de la plus haute priorité sont servis premier , puis les clients de moindre priorité sont servis , et ainsi de suite.

- PS (Processor Sharing) , les clients sont servis de manière égale. La capacité du système est partagé entre les clients.

Remarque : Dans sa version courte, seuls les trois premiers symboles $T/Y/C$ sont utilisés. Dans un tel cas, on suppose que la file est régie par une discipline FIFO et que le nombre de places d'attente ainsi que celui des clients susceptibles d'accéder au système sont illimités.

2.5 La loi de Little

La loi de Little est une relation trés générale qui s'applique à une grande classe de systémes. Elle ne concerne que le régime permanent du systéme. Aucune hypothése sur les variables aléatoires qui caractérisent le systéme (temps d'interarrivées, temps de service,etc) . La seule condition d'application de la loi de Little est que le systéme soit stable. Le debit du systéme est alors indifféremment soit le débit d'entrée, soit le débit de sortie La loi de Little s'exprime telle que dans la propriété suivante :

Théorème 2.5.1. *(Formule de Little) :Le nombre moyen de clients, le temps moyen passé dans le systéme et le débit moyen d'un systéme stable en régime permanent se relient de la façon suivante :*

$$\overline{N} = \lambda_e \overline{T}$$

où λ_e est le taux d'entrée dans le système ($\lambda_e = \lambda$ pour une file$(M/M/1)$

On a vu que la loi de Little nous dit qu'il existe une relation entre le nombre moyen de clients dans la file (en attente ou en service) et le temps moyen total de séjour d'un client dans la file(temps d'attente + temps de service) :

$$\overline{N} = \lambda_e \overline{T}$$

La loi de Little peut aussi s'appliquer en considérant uniquement l'attente dans la queue(sans le service).Elle permet alors de relier le nombre moyen de clients en attente($\overline{N}_Q$)au temps moyen d'attente d'un client avant service($\overline{T}_Q$)

par la relation :$\overline{N}_Q = \lambda_e \overline{T}_Q$

Enfin, on peut appliquer la loi de Little en ne considérant que le serveur. Dans ce cas, elle relie le nombre moyen de clients en service ($\overline{N}_S$)), au temps moyen de séjour d'un client dans le serveur qui n'est rien d'autre que le temps moyen de service($\overline{T}_S$)par la relation : $\overline{N}_S = \lambda_e \overline{T}_S$

On a obtenu trois relations en appliquant la loi de Little successivement au systéme entier à la file d'ttente seule et enfin, au serveur seul. Ces trois

relations ne sont bien sûr pas indépendantes. On peut en effet déduire l'une d'entre elles à partir des deux autres en remarquant que $\overline{N} = \overline{N}_Q + \overline{N}_S$ et $\overline{T} = \overline{T}_Q + \overline{T}_S$

Remarque 2.5.1. *La loi de Little s'applique à tous les modèles de file d'attente rencontrés en pratique (pas seulement à la file $M/M/1$).*

2.6 Mesures de performance d'une file d'attente

L'étude d'une file d'attente ou d'un réseau de files d'attente a pour but de calculer ou d'estimer les performances d'un système dans des conditions de fonctionnement données, et les mesures les plus fréquemment utilisées sont :

- $\overline{N} = \mathbb{E}(N)$: nombre moyen de clients dans le système,
- $\overline{N}_S$: nombre moyen de clients en train d'être servis,
- $\overline{N}_Q$: nombre moyen de clients dans la file d'attente.

N_Q, N_S et N sont les v.a. correspondantes.

- $\overline{T}$: temps moyen qu'un client passe dans le système,
- $\overline{T}_S$: temps moyen de service,
- $\overline{T}_Q$ temps moyen d'attente d'un client dans la file.

T_Q, T_S et T sont les v.a. correspondantes.

De 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 c, une file est stable si et seulement si

$$\lambda < m\mu \Leftrightarrow \rho = \lambda/c\mu < 1,$$

où, ρ est appelé l'intensité du trafic.

2.7 Arrivée avant un départ et départ avant une arrivée

- Temps pour qu'une nouvelle arrivée se produise :

$$A \sim \mathcal{E}(\lambda).$$

- Temps pour qu'un nouveau départ se produise :

$$D \sim \mathcal{E}(\mu).$$

(A et D sont indépendantes).

– Probabilité qu'une arrivée se produise avant un départ :

$$\mathbb{P}(A < D) = \frac{\lambda}{\lambda + \mu}.$$

– Probabilité qu'un départ se produise avant une arrivée :

$$\mathbb{P}(D < A) = \frac{\mu}{\lambda + \mu}.$$

2.8 Analyse en régime stationnaire

Il est difficile d'étudier la variable aléatoire $N(t)$ représentant le nombre de clients au temps t dans le système. On s'intéresse plutôt à $N = \lim_{t\to\infty} N(t)$. On parle alors d'analyse en régime stationnaire (ou analyse à l'équilibre). Pour qu'une file $M/M/1$ puisse atteindre l'équilibre, il faut que $\lambda < \mu$ (sinon la taille de la file augmentera à l'infini). À l'équilibre, on peut montrer que

$$\mathbb{P}(N = n) = \frac{\lambda}{\lambda + \mu}\mathbb{P}(N = n - 1) + \frac{\mu}{\lambda + \mu}\mathbb{P}(N = n + 1).$$

Il s'agit de la règle des probabilités totales. Le terme $\frac{\lambda}{\lambda+\mu}$ représente la probabilité qu'un nouveau client arrive avant que le client en service quitte le système, et $\frac{\mu}{\lambda+\mu}\mathbb{P}(N = n+1)$ est la probabilité que le client en service quitte avant qu'un nouveau client n'arrive.

2.9 La file $M/M/1$

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, la discipline d'attente est FIFO, la file d'attente est de capacité infinie.

FIGURE 2.2 – La file $M/M/1$.

La file peut être considérée comme un processus de naissance et de mort, pour lequel :

$$\lambda_n = \lambda \quad \forall n \geq 0$$
$$\mu_n = \begin{cases} \mu & \forall n \geq 1 \\ 0 & \text{si } n = 0 \end{cases}$$

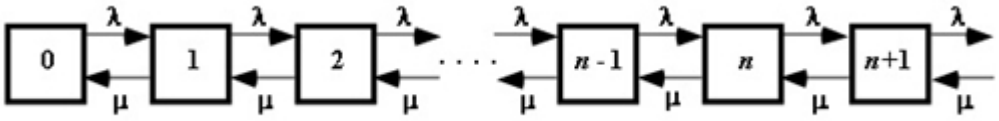

FIGURE 2.3 – Evaluation de l'état dans la file d'attente $M/M/1$.

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 $p_n(t) = \mathbb{P}[N(t) = n]$ peuvent être calculées par les équations différentielles de Kolmogorov ci-dessous, connaissant les conditions initiales du processus.

$$\begin{cases} p'_n(t) = -(\lambda + \mu)p_n(t) + \lambda p_{n-1}(t) + \mu p_{n+1}(t), \\ \text{et} \\ p_0(t) = \lambda p_0(t) + \mu p_1(t). \end{cases}$$

Régime stationnaire

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

$$\pi_n = \lim_{t\to\infty} p_n(t) = (1-\rho)\rho^n, \quad \forall n \in \mathbb{N}. \tag{2.1}$$

$\pi = \{\pi_n\}_{n\geq 0}$ est appelé distribution stationnaire, elle suit une loi géométrique.

Caractéristiques du système

- Le nombre moyen de clients dans le système est :
  $\overline{N} = \mathbb{E}(N) = \sum_{n\geq 0} n\pi_n = (1-\rho)\sum_{n\geq 0} n\rho^n.$
  D'où :
  $$\overline{N} = \frac{\rho}{1-\rho} \tag{2.2}$$
- Nombre moyen de clients en train d'être servis :
  $$\overline{N}_S = 1 - \pi_0 = \rho. \tag{2.3}$$
- Le nombre moyen de clients dans la file :
  $$\overline{N}_Q = \sum_{n\geq 1}(n-1)\pi_n = \frac{\rho^2}{1-\rho}. \tag{2.4}$$

Le temps moyen qu'un client passe dans le système $\overline{T}$, le temps moyen de service $\overline{T}_S$ et le temps moyen d'attente dans la file $\overline{T}_Q$ sont obtenus à partir des formules de Little, ou des distributions du système :

- Temps moyen qu'un client passe dans le système :
  $$\overline{T} = \overline{N}/\lambda = \frac{\rho}{\lambda(1-\rho)} = \frac{1}{\mu-\lambda}. \tag{2.5}$$
- Temps moyen de service :
  $$\overline{T}_S = 1/\mu. \tag{2.6}$$
- Temps moyen d'attente :
  $$\overline{T}_Q = \overline{T} - \overline{T}_S = \frac{\lambda}{\mu(\mu-\lambda)}. \tag{2.7}$$

2.10 La file $M/M/1/K$

On considère un système à serveur simple identique à la file $M/M/1$ excepté que la capacité de la file d'attente est finie. On a donc toujours les hypothèses suivantes : le processus d'arrivée des clients dans la file est un processus de Poisson de taux λ et le temps de service d'un client est une variable aléatoire exponentielle de taux μ. Soit K la capacité de la file d'attente : c'est le nombre maximal de clients qui peuvent être présents dans le système, soit en attente, soit en service. Quand un client arrive alors qu'il y a déjà K clients présents dans le système, il est perdu. Ce système est connu sous le nom de file $M/M/1/K$. L'espace d'états E est maintenant fini : $E = \{0, 1, 2, ..., K\}$. La capacité de la file étant limitée, même si les clients arrivent en moyenne beaucoup plus vite que ce que le serveur de la file est capable de traiter, dès que celle-ci est pleine, les clients qui se présentent sont rejetés. Le nombre de clients dans la file ne peut donc jamais "partir" à l'infini. De plus, dès qu'un client est autorisé à entrer, il sortira un jour et son temps de séjour dans la file est fini, puisqu'il correspond au temps de service de tous les clients devant lui et que ce nombre est limité par K. Sur un temps très long, le débit de sortie sera donc bien égal au débit d'entrée, ce qui correspond bien à la stabilité inconditionnelle du système.

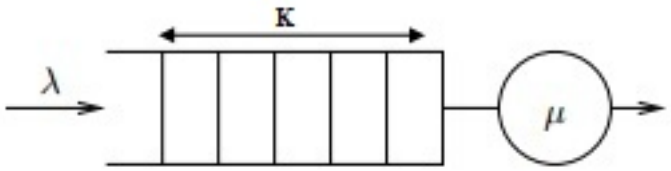

FIGURE 2.4 – La file $M/M/1/K$.

Le processus de naissance et de mort modélisant ce type de file d'attente est alors défini de la façon suivante :

$$\lambda_n = \begin{cases} \lambda & \text{si } n < K \\ 0 & \text{si } n \geq K \end{cases}$$
$$\mu_n = \mu \text{ si } 1 \leq n \leq K$$

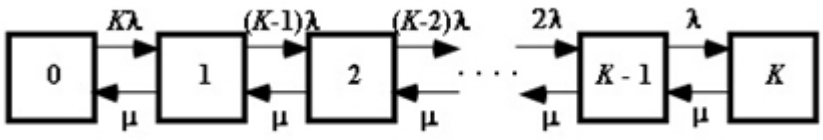

FIGURE 2.5 – Evaluation de l'état dans la file d'attente $M/M/1/K$.

L'intégration de l'équation récurrente permettant de calculer π_n se fait alors comme suit :

$$\pi_n = \pi_0 \rho^n \text{ pour } n \leq K$$

$$\pi_n = 0 \text{ pour } n > K$$

$$\pi_0 = \frac{1}{\sum_{n=0}^{K} \rho^n} = \frac{1-\rho}{1-\rho^{K+1}} \text{ si } \lambda \neq \mu(\text{ et } \frac{1}{K+1} \text{ si } \lambda = \mu).$$

Caractéristiques du système

- Le nombre moyen de clients dans le système est :

$$\overline{N} = \sum_{n=0}^{K} n\pi_n = \frac{\rho}{1-\rho} \frac{1-(K+1)\rho^K + K\rho^{K+1}}{1-\rho^{K+1}} \tag{2.8}$$

  À nouveau, lorsque K tend vers l'infini et $\rho < 1$, on retrouve les résultats de la file $M/M/1$:

$$\overline{N} = \frac{\rho}{1-\rho}$$

- Le nombre moyen de clients dans la file :

$$\overline{N}_Q = \sum_{n=1}^{\infty} (n-1)\pi_n = \overline{N} - (1-\pi_0) \tag{2.9}$$

Le temps moyen qu'un client passe dans le système $\overline{T}$ et le temps moyen d'attente dans la file $\overline{T}_Q$ sont obtenus à partir la loi de Little :

- Temps moyen qu'un client passe dans le système :

$$\overline{T} = \frac{\overline{N}}{\lambda} \tag{2.10}$$

- Temps moyen d'attente :

$$\overline{T}_Q = \frac{\overline{N}_Q}{\lambda} \tag{2.11}$$

2.11 La file $M/M/C$

On considère un système identique à la file $M/M/1$ excepté qu'il comporte C serveurs identiques et indépendants les uns des autres. On conserve les hypothèses : processus d'arrivée des clients poissonien de taux λ et temps de service exponentiel de taux μ (pour chacun des serveurs). Ce système est connu sous le nom de file $M/M/C$. L'espace d'états E est, comme pour la $M/M/1$ infini : $E = \{0, 1, 2, ...\}$. La file d'attente est de capacité infini. Si l'un des serveurs est libre, le client qui arrive se dirige immédiatement vers ce serveur. Dans le cas contraire, le client prend sa place dans une file d'attente commune pour tous les serveurs. Lorsqu'un serveur se libère, le client en tête de la file occupe ce serveur. Par conséquent, la discipline d'attente est FIFO.

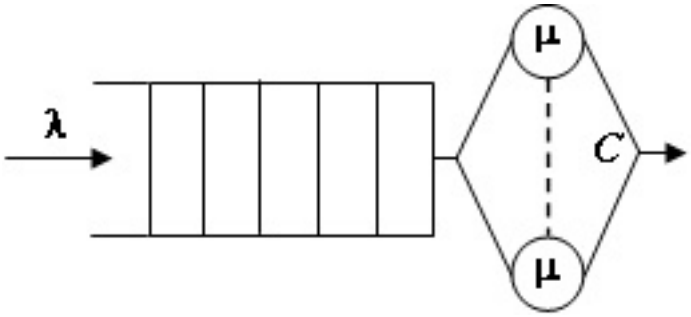

FIGURE 2.6 – La file $M/M/C$.

Le processus de naissance et de mort modélisant ce type de file d'attente est alors défini de la façon suivante :

$$\lambda_n = \lambda \quad \forall n \geq 0$$
$$\mu_n = \begin{cases} 0 & \text{si } n = 0 \\ n\mu & \forall n = 1, ..., C \\ C\mu & \forall n \geq C. \end{cases}$$

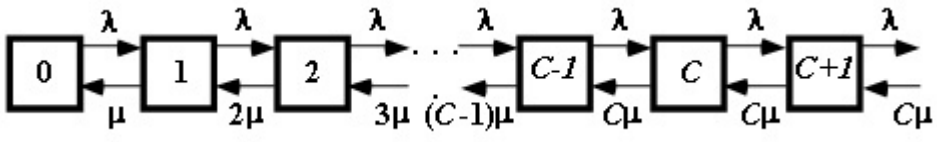

FIGURE 2.7 – Evaluation de l'état dans la file d'attente $M/M/C$.

Du diagramme, on déduit les résultats qui suivent. L'analyse du système en régime stationnaire, à l'aide de la procédure des équations de Chapman-Kolmogorov aboutit aux équations suivantes :

$$\begin{aligned} \lambda\pi_0 &= \mu\pi_1, \\ (\lambda + n\mu)\pi_n &= \lambda\pi_{n-1} + (n+1)\mu\pi_{n+1} \qquad 1 \leq n < c, \\ (\lambda + c\mu)\pi_n &= \lambda\pi_{n-1} + c\mu\pi_{n+1} \qquad n \geq c, \end{aligned}$$

avec

$$\sum_{n=0}^{\infty} \pi_n = 1.$$

La résolution du système ci-dessus présente la distribution stationnaire suivante :

$$\pi_n = \frac{\rho^n}{n!}\pi_0, \qquad 0 \leq n \leq C, \tag{2.12}$$

$$\pi_n = \frac{\rho^C}{C!}(A)^{n-C}\pi_0, \qquad n \geq C, \tag{2.13}$$

où

$$\pi_0 = \left[\sum_{n=0}^{C-1} \frac{\rho^n}{n!} + \frac{\rho^C}{C!}\sum_{n=C}^{\infty} \rho^{n-C}\right]^{-1}, \quad \rho = \frac{\lambda}{\mu} \;\; et \;\; A = \frac{\lambda}{C\mu}.$$

Cette dernière existe si $\lambda < C\mu$.

Caractéristiques du système

A partir de la distribution stationnaire du processus $\{N(t), t \geq 0\}$, on peut calculer les caractéristiques du système. En effet,

- Le nombre moyen de clients dans le système :

$$\overline{N} = \rho + \frac{\rho^{C+1}}{C.C!(1-A)^2}\rho_0, \tag{2.14}$$

- Le nombre moyen de clients dans la file :

$$\overline{N}_Q = \frac{\rho^{C+1}}{C.C!(1-A)^2}\rho_0, \tag{2.15}$$

- Temps moyen qu'un client passe dans le système :

$$\overline{T} = \frac{C\mu\rho^C}{C!(C\mu-\lambda)^2}\rho_0, \tag{2.16}$$

- Temps moyen d'attente :

$$\overline{T}_Q = \frac{1}{\mu} + \frac{\rho^C}{\mu C.C!(1-A)^2}\rho_0. \tag{2.17}$$

2.12 La file $M/M/\infty$

On considère un système composé d'un nombre illimité de serveurs identiques et indépendants les uns des autres. Dès qu'un client arrive, il rentre donc instantanément en service. Danc cette file particulière, il n'y a donc pas d'attente. On suppose toujours que le processus d'arrivée des clients est poissonien de taux λ et que les temps de service sont exponentiels de taux μ (pour tous les serveurs). Ce système est connu sous le nom de file $M/M/\infty$.

Comme cela a été fait pour la file $M/M/C$, on peut facilement démontrer que le taux de transition d'un état n quelconque vers l'état $n-1$ est égal à $n\mu$ et correspond au taux de sortie d'un des n clients en service. De même, le taux de transition d'un état n vers l'état $n+1$ est égal à λ et correspond au taux d'arrivée d'un client.

De façon intuitive, la capacité de traitement de la file est infinie puisque tout nouveau client se présentant à l'entrée de la file est instantanément traité. La condition de stabilité exprimant que "le nombre moyen de client arrivant à la file par unité de temps doit être inférieure à la capacité de traitement de la file" est donc toujours satisfaite.

Soit π_n la probabilité stationnaire d'être dans l'état n. Les équations d'équilibre nous donnent

$$\pi_{n-1}\lambda = \pi_n n\mu \text{ pour } n = 1, 2, ...$$

soit $\pi_n = \frac{\rho}{n}\pi_{n-1}$ pour $n = 1, 2, ...$, où $\rho = \frac{\lambda}{\mu}$.
On peut alors exprimer toutes les probabilités en fonction de π_n :

$$\pi_n = \frac{\rho^n}{n!}\pi_0 \text{ pour } n = 1, 2, ...$$

La condition de normalisation nous donne alors immédiatement π_n :

$$\pi_n = \frac{1}{\sum_{n=0}^{+\infty} \frac{\rho^n}{n!}} = e^{-\rho}.$$

Notons que la série $\sum_{n=0}^{+\infty} \frac{\rho^n}{n!}$ converge pour toutes valeurs de ρ (donc de λ et de μ), ce qui est cohérent avec la stabilité inconditionnelle de la file. On obtient finalement :

$$\pi_n = \frac{\rho^n}{n!}e^{-\rho} \text{ pour } n = 1, 2, ...$$

Caractéristiques du système

- Nombre moyen de clients $\overline{N}$

$$\overline{N} = \sum_{n=1}^{+\infty} n\pi_n = e^{-\rho} \sum_{n=1}^{+\infty} \frac{\rho^n}{(n-1)!} = e^{-\rho}\rho e^{\rho} = \rho$$

- Temps moyen de séjour $\overline{T}$

  Intuitivement, le temps moyen passédanslesystème est réduit au temps moyen de service, soit $\frac{1}{\mu}$. On peut redémontrer ce résultat en utilisant la loi de Little :

$$\overline{T} = \frac{\overline{N}}{\lambda} = \frac{1}{\mu}$$

Chapitre 3

Les files d'attente avec rappels

Introduction

Dans la théorie des files d'attente classique, il est supposé qu'un client qui ne peut pas obtenir son service immédiatement dès son arrivée, rejoint la file d'attente ou quitte le système définitivement. Les systèmes de 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».

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.

3.1 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, d'un buffer de capacité $K - c$ ($K \geq c$) et d'une orbite de capacité N. 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 le buffer, le client la rejoindra. Par ailleurs, si un client arrive et trouve tous les serveurs et toutes les positions d'attente du buffer 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 du buffer 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 le buffer, 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^{me} 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 figure (3.1).

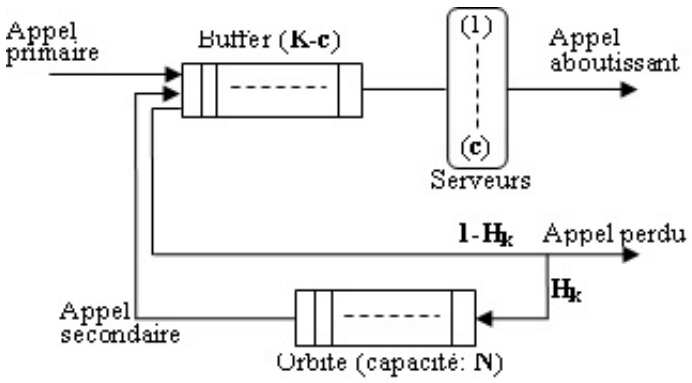

FIGURE 3.1 – Le schéma général d'un système d'attente avec rappels.

Remarque 3.1.1. *1. Le modèle d'attente avec rappels décrit ci-dessus est un modèle général. Plusieurs systèmes de files d'attente avec rappels peuvent être considérés comme des cas particuliers tels que : les systèmes sans buffer, 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.*

Notation :
Une file d'attente avec rappels est entièrement caractérisée par le quintuple

$$A/S/s/s + B - X$$

où :
- A : représente la loi de la suite des interarrivées $(U_n, n \in \mathbb{N}^*)$,
- S : représente la loi de la suite $(\sigma_n, n \in \mathbb{N}^*)$ des temps de services demandés par les clients,
- s : représente le nombre (éventuellement infini) de serveurs,
- B : est la taille (éventuellement infini) du buffer (nombre de clients maximal qu'il peut contenir),
- X : représente la discipline de service.

En particulier, les deux premiers paramètres A et S caractérisant les lois des suites des interarrivées et des temps de services prennent essentiellement les valeurs suivantes :
- G : (général) si la suite est stationnaire ergodique,
- GI : (général indépendant) si elle est i.i.d,
- M : (memoryless) si la suite est i.i.d. de loi exponentielle (si $A = M$, le processus des entrées est donc un processus de Poisson),
- D : (deterministic) si la suite est déterministe et constante.

En général, si aucune ambiguïté n'est possible, on notera $././s$, une file avec s serveurs, de buffer de capacité infinie, régie par la discipline FIFO. Par exemple, la file $M/M/1$ est la file d'attente où les clients entrent suivant un processus de Poisson, en demandant un service de loi exponentielle, avec un serveur, un buffer de taille infinie et traitée en FIFO.

3.2 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)[10] et concerne l'aspect modélisation du système sous étude. 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 α. Donc, il y a une probabilité $n\alpha dt + o(dt)$ d'un nouveau rappel dans le prochain intervalle $(t, t + dt)$ sachant que n 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.

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 $\nu dt + o(dt)$. 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) [7], Shikata (1999) [17]). Artalejo et 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 n client sont en orbite à l'instant t est $(\nu(1 - \delta_{0n}) + n\alpha)dt + o(dt)$.
On mentionne aussi l'existence d'une autre politique dite politique de rappels quadratiques .

3.3 Modèles markoviens

3.3.1 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 $\frac{1}{\mu}$. 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

$$\{C(t), N_o(t), t \geq 0\}, \tag{3.1}$$

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 ($\rho = \dfrac{\lambda}{\mu} < 1$). Le processus (3.1) est de Markov d'espace d'états $S = \{0, 1\} \times \mathbb{N}$.

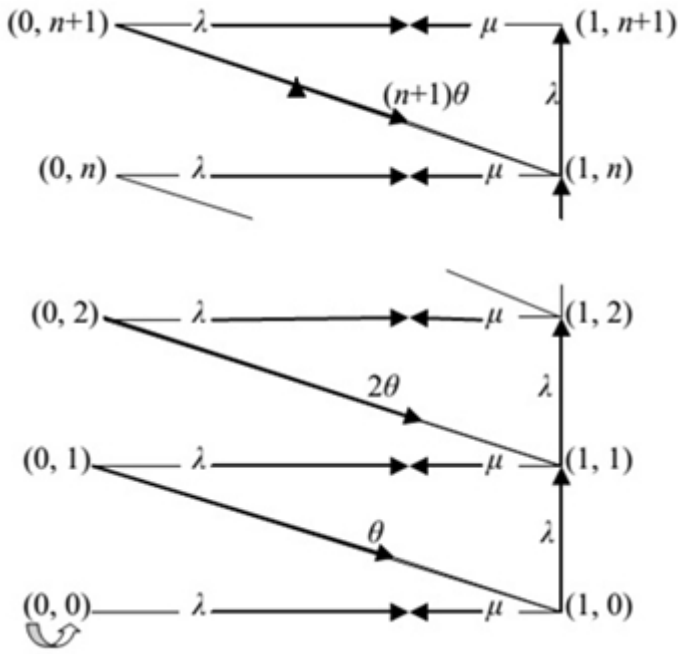

FIGURE 3.2 – La structure générale du modèle $M/M/1$ avec rappels.

Les équations d'équilibre statistique sont :

$$(\lambda + n\theta)\pi_{0n} = \mu\pi_{1n}, \tag{3.2}$$

$$(\lambda + \mu)\pi_{1n} = \lambda\pi_{0n} + (n+1)\theta\pi_{0,n+1} + \lambda\pi_{1,n-1}. \tag{3.3}$$

.

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

$$\pi_0(z) = \sum_{n=0}^{\infty} Z^n \pi_{0n},$$

$$\pi_1(z) = \sum_{n=0}^{\infty} Z^n \pi_{1n}.$$

A l'aide de ses fonctions et à partir des équations (3.2) et (3.3), on obtient :

$$\pi_0(z) = (1-\rho)\left(\frac{1-\rho}{1-z\rho}\right)^{\frac{\lambda}{\theta}}, \tag{3.4}$$

$$\pi_1(z) = \rho \left(\frac{1-\rho}{1-z\rho} \right)^{\frac{\lambda}{\theta}+1}. \tag{3.5}$$

Les transformées inverses des (3.4) et (3.5) nous donnent les formules analytiques explicites [25] :

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

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

3.3.2 Modèle $M/M/2$ avec rappels

Nous considérons un système de files d'attente avec rappels où l'espace de service comprend $c = 2$ serveurs. Les clients primaires arrivent dans le système selon un processus de Poisson de taux $\lambda > 0$. Si un client primaire trouve au moins un serveur libre, il commence son service. Sinon, il entre en orbite. Les durées de service et les durées entre deux rappels consécutives sont exponentiellement distribuées de moyennes finies, respectivement $\frac{1}{\mu}$ et $\frac{1}{\theta}$ Nous supposons que toutes les variables aléatoires introduites sont mutuellement indépendantes.

L'état du système à la date t peut décrit par le processus (3.1), dont l'espace d'états est $S = \{0,1,2\} \times \mathbb{N}$. Les probabilités d'état sont

$$\pi_{in} = \mathbb{P}(C(t) = i, N_o(t) = n), \quad (i,n) \in S.$$

Les transitions possibles sont données dans la figure (3.3).
La condition d'existence d'un régime stationnaire est $\lambda < 2\mu$.

Supposons que $\rho = \frac{\lambda}{2\mu} < 1$ et $\mu = 1$. A partir du graphe des transitions, il est possible d'obtenir les équations d'équilibre statistique, telles que :

$$(\lambda + n\theta)\pi_{0n} = \pi_{1n}. \tag{3.6}$$

$$(\lambda + 1 + n\theta)\pi_{1n} = \lambda\pi_{0n} + (n+1)\theta\pi_{0,n+1} + 2\pi_{2n}, \tag{3.7}$$

$$(\lambda + 2)\pi_{2n} = \lambda\pi_{1n} + (n+1)\theta\pi_{1,n+1} + \lambda\pi_{2,n-1}. \tag{3.8}$$

Ainsi que l'équation de normalisation :

$$\sum_{n=0}^{\infty} \pi_{.0n} + \sum_{n=0}^{\infty} \pi_{.1n} + \sum_{n=0}^{\infty} \pi_{.2n} = 1$$

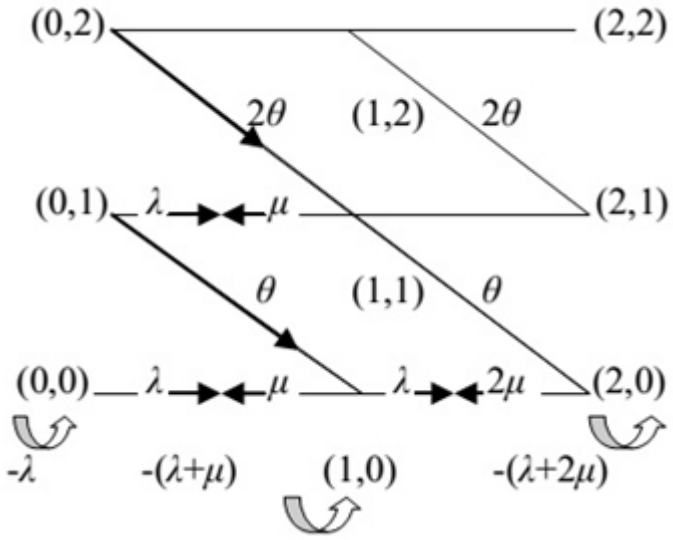

FIGURE 3.3 – La structure générale du modèle $M/M/2$ avec rappels.

pour $\pi_{in} = \lim_{t\to\infty} \pi_{.in}(t)$.
La résolution des équations (3.6)-(3.7)-(3.8) nous donne

$$\begin{aligned}
\pi_{0n} &= \frac{\lambda^n}{n!\theta^n} \prod_{k=0}^{n-1} \frac{(\lambda + k\theta)^2 + k\theta}{2 + 3\lambda + 2\theta + 2k\theta} \pi_{00}, \\
\pi_{2n} &= (\lambda + n\theta) \frac{\lambda^n}{n!\theta^n} \prod_{k=0}^{n-1} \frac{(\lambda + k\theta)^2 + k\theta}{2 + 3\lambda + 2\theta + 2k\theta} \pi_{00}, \\
\pi_{2n} &= [1 + \lambda + (n+1)\theta] \frac{\lambda^n}{n!\theta^n} \prod_{k=0}^{n} \frac{(\lambda + k\theta)^2 + k\theta}{2 + 3\lambda + 2\theta + 2k\theta} \pi_{00}.
\end{aligned}$$

3.4 Modèle d'attente $M/G/1$ avec rappels

Le modèle $M/G/1$ avec rappels est le modèle le plus étudié par les spécialistes. Il existe une littérature abondante sur ses diverses propriétés.

3.4.1 Description du modèle

Les clients arrivent dans le système selon un processus de Poisson de taux $\lambda > 0 : \mathbb{P}(\tau_n^e \leq x) = 1 - e^{\lambda x}$. .
Le service des clients est assuré par un seul serveur.
La durée de service τ est de loi générale $\mathbb{P}(\tau_n^s \leq x) = B(x)$ et de transformée de Laplace-Stieltjes $\widetilde{B}(s)$, $Re(s) > 0$. Soient les moments $\beta_k = (-1)^k \widetilde{B}(k)(0)$, l'intensité du trafic $\rho = \lambda\beta_l$ et $\gamma = \frac{1}{\beta_1}$. La durée entre deux rappels successifs d'une même source secondaire est exponentiellement distribuée de paramètre $\theta > 0 : T(x) = \mathbb{P}(\tau_n^r \leq x) = 1 - e^{-\theta x}$.

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

3.4.2 Chaîne de Markov induite

Considérons le processus $\{C(t), N_o(t), t \geq 0\}$, 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)[8]

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^{me} départ, dont

l'équation fondamentale est :

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

où ν_{n+1} est le nombre des clients primaires arrivant dans le système durant le service du $(n+1)^{me}$ 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 $(n+1)^{ime}$ client. La distribution de ν_{n+1} est la suivante :

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

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

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

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

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

Elle dépend de q_n et sa distribution est

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

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

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

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

La condition d'existence du régime stationnaire peut être obtenue comme suit : L'accroissement moyen de la chaîne vaut

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

Si $\rho < 1$, alors $\lim_{i\to\infty} \mathbb{E}[q_{n+1} - q_n/q_n = i] = \rho - 1 < 0$ et la chaîne est donc ergodique. Par contre, si $\rho \geq 1$, alors $\lim_{i\to\infty} \mathbb{E}[q_{n+1} - q_n/q_n = i] = \rho - \frac{i\theta}{\lambda+i\theta} \geq 1 - \frac{i\theta}{\lambda+i\theta} = \frac{\lambda}{\lambda+i\theta} > 0$. 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 $\pi_n = \lim \mathbb{P}(N_o(\xi_i) = n)$. Les équations de Kolmogorov se présentent de la manière suivante :

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

Vu la présence de convolution, cette équation peut être transformée, à l'aide des fonctions génératrices $\phi(z) = \sum_{n=0}^{\infty} z^n \pi_n$ et $\psi(z) = \sum_{n=0}^{\infty} z^n \frac{\pi_n}{\lambda + n\theta}$,

$$\phi(z) = A(z)(\lambda\phi(z) + \theta\psi'(z)).$$

D'un autre côté,

$$\begin{array}{rcl} \phi(z) & = & \sum_{n=0}^{\infty} z^n \pi_n = \sum_{n=0}^{\infty} z^n \pi_n \frac{\lambda + n\theta}{\lambda + n\theta} \\ & = & \lambda \sum_{n=0}^{\infty} z^n \pi_n \frac{\pi_n}{\lambda + n\theta} + \theta \sum_{n=0}^{\infty} n z^n \frac{\pi_n}{\lambda + n\theta} \\ & = & \lambda\psi(z) + \theta\psi'(z). \end{array}$$

Par conséquent

$$\begin{array}{rcl} \lambda\psi(z) + \theta\psi'(z) & = & A(z)(\lambda\psi(z) + \theta\psi'(z)), \\ \theta\psi'(z)[A(z) - z] & = & \lambda\psi(z)[1 - A(z)]. \end{array} \tag{3.9}$$

Lemme 3.4.1. *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

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

Théorème 3.4.1. *Soit $\rho < 1$. La distribution stationnaire de la chaîne de Markov induite possède la fonction génératrice suivante (58)[17]*

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

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

3.4.3 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)[9] . Elle consiste à décrire le processus des arrivées comme processus de Markov avec dépendance de l'état de paramètre λ_{in} quand $\{C(t), N_o(t)\}$ est dans l'état (i, n) et à appliquer les schémas récursifs. L'état du système peut être décrit par le processus

$$X(t) = \begin{cases} N_o(t) \text{ si } C(t) = 0 \\ \{C(t); N_o(t); \xi(t)g \text{ si } C(t) = 1, \end{cases},$$

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

$$p_{0n} = \lim_{t\to\infty} P(C(t) = 0, N_o(t) = n),$$

$$p_{1n}(x) = \lim_{t\to\infty} \frac{d}{dx} P(C(t) = 1, \xi(t) \le x, N_o(t) = n).$$

A partir du graphes **??**, les probabilités p_{0n} et $p_{1n}(x)$ vérifient le système d'équations de balance :

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

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

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

D'où

$$\begin{cases} \lambda P_0(z) + \theta z \lambda P_0'(z) = \int_0^\infty P_1(z,x) b(x) dx, \\ P_1'(z,x) = (\lambda z - \lambda - b(x)) P_1(z,x), \\ P_1(z,0) = \lambda P_0(z). \end{cases} \tag{3.10}$$

De la deuxième équation de (3.10), on a

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

Donc, la première équation de (3.10) devient

$$\begin{array}{rcl} \lambda P_0(z) + \theta z P_0'(z) & = & \int_0^\infty P_1(z,0)[1 - B(x)] \exp(-(\lambda - \lambda z)x) b(x) dx \\ & = & P_1(z,0) \widetilde{B}(\lambda - \lambda z) = P_1(z,0) A(z). \end{array} \tag{3.11}$$

A partir des équations (3.10) et (3.11), on a

$$P_1(z,0) f(z) = \lambda P_0(z) + \theta z (\frac{P_1(z,0)}{\theta} - \frac{\lambda}{\theta} P_0(z)), \tag{3.12}$$

$$P_1(z,0) = \frac{\lambda - \lambda z}{A(z) - z} P_0(z)[1 - B(x)] \exp(-\lambda(\lambda - \lambda z)x). \tag{3.13}$$

En intégrant cette équation, et en utilisant la formule $\int_0^\infty \exp(-sx)[1 - B(x)]dx = (1 - \widetilde{B}(s))/s$,on obtient

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

De (3.10) et (3.11), on peut obtenir $P_0(z)$

$$\lambda P_0(z) + \theta z P_0'(z) = A(z)[\lambda P_0(z) + \theta P_0'(z)], \tag{3.14}$$

$$\theta[A(z) - z] P'(z) = \lambda[1 - A(z)] P_0(z). \tag{3.15}$$

Considérons $f(z) = A(z) - z$. Du lemme 3.1, $f(z)$ 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 3.2 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

$$\begin{array}{rcl} P_0(z) & = & \sum_{n=0}^\infty z^n p_{0n} \quad = \quad (1 - \rho) \exp\left[\frac{\lambda}{\theta} \int_1^z \frac{1 - A(u)}{A(u) - u} du\right] \\ P_1(z) & = & \sum_{n=0}^\infty z^n p_{1n} \quad = \quad \frac{1 - A(z)}{A(z) - z} P_0(z). \end{array}$$

3.4.4 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

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

- Temps moyen d'attente d'un client

$$\overline{T} = \frac{\overline{N}_o}{\lambda} = \frac{\lambda \beta_2}{2(1-\rho)} + \frac{\rho}{\theta(1-\rho)},$$

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

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

conclusion générale

Les files d'attente avec clients impatients sont devenues un axe de recherche très important à cause de leurs impacts négatifs (des pertes colosales) sur l'économie. La recherche d'une gestion des phénomènes d'attentes avec rappel, rappel et priorité est une des vastes domaines de la recherche opérationnelle.
Les phénomènes d'attente dans les sociétés dites modernes sont nombreux et plus visibles, par exemple dans le domaine des transports (terrestre, aérien,$\cdots$) ou des services (banque, poste, $\cdots$).
Le type des phénomènes d'attentes avec rappel et priorité, dits discrets sont retrouvés dans certains systèmes tels les réseaux téléphoniques, les systèmes informatiques, ou les réseaux informatiques.
Dans ce travail nous nous sommes intéréssés aux systèmes de fils 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 améliorie et mieux evaluer 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.

Bibliographie

[1] . A. Aïssani. A Survey on Retrial Queueing Models. Actes des Journées Statistiques Appliquées, U.S.T.H.B., Alger, 1-11.1994

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

[3] .J. R. Artalejo and A. Gòmez-Corral, Advances in Retrial Queues, European Journal of Operation Research...2008

[4] .J. R. Artalejo and A. Gòmez-Corral. Analysis of an M/G/1 queue with constant repeated attempts and server vacations. Computers and Operations Research, 24(6) : 493-504...2008

[5] .J.R. Artalejo. T. Phung-Duc.Markovian single server retrial queues with two way communication, ,in : Proceedings of the 6th International Conference on Queueing Theory and Network Applications, Seoul, pp. 1...7...2011

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

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

[8] .Q. H. Choo and B. Conolly. New results in the theory of repeated orders queueing systems. Journal of applied Probability, 16 :631-640..1979

[9] A. G. De Kok. Algorithmic methods for single server systems with repeated attempts. Statistica Neerlandica, 38 :23-32..1984

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

[11] .G.I. Falin. A single line system with secondary orders. Engineering Cybernitics Review, 17(2) : 76-93...1979

[12] .G.I. Falin. A Single-line System with Secondary Orders. Engineering Cybernetics Review, 17(2), 76-83...1979

[13] .G. I. Falin.A Single-line repeated orders queueing systems. Engineering Cybernetics Review, 21 (6), 21-25,

[14] .G.I. Falin.A Single server retrial queues with two way communication,Applied Mathematical Modelling. 37, 1811-1822

[15] .G.I. Falin and J.G.C. Templeton. Retrial Queues. Chapman and Hall.....1997

[16] ..V. G. Kulkarni and H. M. Liang. Retrial Queues Revisited. Frontiers in Queueing (J.H. Dshalov, ed.) CRC Press Boca Raton, pp 19-34....1997

[17] .shikata.Optimizing the menezes-vanstone algorithm or Non Super singular ellliptic currves...1999

[18] . .J. G. C. Templeton, Retrial Queues, Top, 7 : 351-353....1999