République Algérienne Démocratique et Populaire
Ministère de l'enseignement supérieur et de la recherche scientifique

N° Attribué par la bibliothèque

Année univ.: 2016/2017

Systèmes de files d'attents $M^X/G/1$ avec rappels et groupes impatients

Mémoire présenté en vue de l'obtention du diplôme de

Master Académique

Université Dr Tahar Moulay - Saïda

Discipline : MATHEMATIQUES

Spécialité : ASSPA

par

Daoudi **A**bdelkarim [1]

Sous la direction de

Mr M. Kadi

Soutenue le 22 Mai 2017 devant le jury composé de

Dr.F. Mokhtari	Université Dr Tahar Moulay - Saïda	Présidente
Mr.M. Kadi	Université Dr Tahar Moulay - Saïda	Encadreur
Mlle.F. Benziadi	Université Dr Tahar Moulay - Saïda	Examinatrice
Dr.S. Rahmani	Université Dr Tahar Moulay - Saïda	Examinatrice

1. e-mail : karimkirak@gmail.com

Remerciements

Par ces quelques lignes, je tiens à remercier toutes les personnes qui ont participé de près ou de loin au bon déroulement de ce mémoire, en espérant n'avoir oublié personne ...

Tout d'abord, Je tiens à exprimer mes sincères remerciements à Mr M. Kadi pour m'avoir encadré durant ces mois avec beaucoup de patience, de disponibilité et de professionnalisme, et pour m'avoir transmis sa passion pour le domaine des Systèmes de files d'attente,.... Ce fut très sincèrement un réel plaisir de travailler à ses côtés pendant ces mois et pour m'avoir ainsi permis de réaliser ce mémoire. Je remercie très chaleureusement le Professeur A. Kandouci pour m'avoir accueilli au sein du laboratoire LMSSA. Je tiens également à le remercier pour sa disponibilité, la confiance qu'il m'a accordé et pour ses conseils et ses commentaires fort utiles qui ont fortement enrichi ma formation.

Je remercie Dr F.Mokhtari d'avoir accepter de présider le jury. J'exprimer ma gratitude aux Mlle F.Benziadi et Dr S.Rahmani qui ont accepté de rapporter ce mémoire et pour l'intérêt qu'elles y ont porté. Je voudrais également associer mes meilleurs remerciements à mes collègues pour leurs aides et leurs soutien moral qui ont eu une contribution importante pour ma réussite dans ce mémoire.

Je remercie tout mes amis pour leur soutien moral : Mokhtar, samir, Hamid et Farouk. Un remerciement très spécial à mon frère Mohammed. Et pour finir, merci à toutes les personnes que j'ai oubliées de citer et qui m'ont permis de mener à bien ce mémoire.

Table des matières

Introduction

La théorie des files d'attente, est un des outils analytiques les plus puissants pour la modélisation des systèmes dynamiques. Cette théorie a commencé en 1909 avec les travaux de recherches de l'ingénieur danois Agner Krarup Erlang (1878,1929) sur le trafic téléphonique de Copenhague pour déterminer le nombre de circuits nécessaires afin de fournir un service téléphonique acceptable. Par la suite, les files d'attente ont été intégrés dans la modélisation de divers domaines d'activité. On assista alors à une évolution rapide de la théorie des files d'attente qu'on appliqua à l'évaluation des performances des systèmes informatiques et aux réseaux de communication. Les chercheurs oeuvrant dans cette branche d'activité ont élaboré plusieurs nouvelles méthodes qui ont été ensuite appliquées avec succés dans d'autres domaines, notamment dans le secteur de la 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 et Richard Larson. Grâce à tous ces développements, la théorie des files d'attente est aujourd'hui largement utilisée et ses applications sont multiples. Dans ce mémoire on s'intéresse aux systèmes de files d'attente avec rappels.
La théorie des files d'attente avec rappel a été surtout utilisée pour modéliser les systèmes téléphoniques, centres d'appels et les réseaux informatiques. Elle sert à résoudre des problèmes pratiques, tels que l'analyse du temps d'attente des abonnés dans les réseaux téléphoniques, l'évitement de collision dans les réseaux locaux, l'analyse du temps d'attente pour accéder à la mémoire sur les disques magnétiques,

Le premier chapitre est un rappel sur les différents processus (le Processus de comptage, processus de Poisson, le processus de renouvellement, le processus de naissance et de mort...etc), qui sont un outil très puissant pour la modèlisation des systèmes dynamiques.

Le deuxième chapitre traitent les modèles de files d'attente $M/G/1$ et $M^X/G/1$ classiques. La distribution de service étant générale les fonctions génératrices de ces deux modèles sont données. Cette méthode permet d'introduire le chapitre suivant.

Dans le chapitre trois, nous présentons une étude de certains modèles d'attente de type M/G/1 avec rappels. Nous commençons par les systèmes d'attente avec rappels et clients persistants, puis ceux avec rappels et clients impatients. Ainsi que le modèle $M^X/G/1$ avec rappels, arrivées par groupes et clients impatients.

Chapitre 1

Les Processus Stochastiques

Introduction

Les processus stochastiques décrivent l'évolution d'une grandeur aléatoire en fonction du temps. Il existe de nombreuses applications des processus aléatoires notamment en physique statistique, en biologie (évolution, génétique et génétique des population, médecine (croissance de tumeurs, épidémie), et bien entendu les sciences de l'ingénieur. Dans ce dernier domaine, les applications principales sont pour l'administration des réseaux, de l'internet, des télécommunications et bien entendu dans les domaines économique et financier.

L'étude des processus stochastiques s'insère dans la théorie des probabilités dont elle constitue l'un des objectifs les plus profonds. Elle soulève des problèmes mathématiques intéressants et souvent très difficiles.

Définitions et propriétés de base :

Définition 1.0.1. *Un processus stochastique est une famille de variables aléatoires X_t, $t \in T$ ou chaque variable aléatoire X_t est indexée par le paramètre $t \in T$, si T est un ensemble de $\mathbb{R}_+$, alors t signifie temps.*
Généralement X_t représente l'état du processus stochastique au temps t.

- *Si T est dénombrable, i.e $T \subseteq \mathbb{N}$, alors nous disons que X_t, $t \in T$ est un processus à temps discret.*
- *Si T est un intervalle de $[0; \infty[$, alors le processus stochastique est dit un processus à temps continu.*

L'ensemble des valeurs de X_t est appelé l'espace d'état, qui peut également être soit discret (fini ou infini dénombrable) ou continu (un sous-ensemble de $\mathbb{R}$ ou

$\mathbb{R}^n$*), donc nous écrivons* $(X_n)_{n\geq 0}$ *pour le processus à temps discret et* $(X_t)_{t\geq 0}$ *pour le processus à temps continu.*

1.1 Le Processus de comptage

Définition 1.1.1. ***(processus de comptage)*** *Un processus stochastique* $[N(t), t \in \mathbb{R}^+]$ *est un processus de comptage si* $N(t)$ *représente le nombre total d'événements qui se sont produits entre 0 et t ,il doit donc satisfaire*

- $N(t) \geq 0$
- $N(t)$ *a des valeurs entières uniquement.*
- *pour* $s < t$,$N(t) - N(s)$ *est le nombre d'événements qui ont eu lieu entre* s*et* t*.*

Un processus de comptage est un processus discret à temps continu. Un second processus peut être associé au processus des temps d'occurrence ; le processus des temps d'inter-arrivées $\{W_n, n \in N_0\}$ *ou* $\forall n \in N_0$ *la variable aleatoire* N_n *est le temps d'attente entre les* $(n-1)^{ieme}$ *et* n^{ieme} *occurrences ,c-à-d :*

$$W_n = T_n - T_{n-1}$$

Proposition 1.1.1. *Les relations suivantes sont triviales tel que* $T_0 = 0$ *à vérifier :*

1. $T_n = W_1 + W_2 + \ldots + W_n \ \forall n \geq 1$;
2. $N(t) = \sup\{n \geq 0 : T_n \leq t\}$;
3. $\mathbb{P}[N(t) = n] = \mathbb{P}[T_n \leq t < T_{n+1}]$;
4. $\mathbb{P}[N(t) \geq n] = \mathbb{P}[T_n \leq t]$;
5. $\mathbb{P}[s < T_n < t] = \mathbb{P}[N(s) < n \leq N(t)]$

Démonstration : on a

$W_n = T_n - T_{n-1}$

$$\begin{aligned} T_n &= W_1 + W_2 + \ldots + W_n \\ &= T_1 - T_0 + T_2 - T_1 + T_3 - T_2 + \ldots + T_{n-1} - T_{n-2} + T_n - T_{n-1} \\ &= T_0 + T_n \\ &= T_n. \quad \text{car} \quad T_0 = 0 \end{aligned}$$

Définition 1.1.2. ***(processus à accroissements indépendants)***
Un processus$\{X_t\}$ *tel que* $X_0 = 0$ *est à accroissements indépendants si pour tout suite finie* $0 < t_1 < t_2 < t_3 \ldots < t_n$ *les variables aléatoires* $X_{t_1}, X_{t_2} - X_{t_1}, \ldots, X_{t_n} - X_{t_{n-1}}$ *sont indépendantes.*

Définition 1.1.3. *(**Un processus à accroissements indépendants est à accroissements stationnaires**)*
si la loi de l'accroissement $(X_{t+s} - X_t)$ *ne dépend pas de* t *pour tout* $t \geq 0$.

Définition 1.1.4. *(**Un processus de comptage**)*
$\{N(t), t \geq 0\}$ *est un processus de poisson d'intensité* $\lambda > 0$ *si :*

- *N(0)=0 ;*
- *le processus est à accroissements stationnaires ;*
- *le processus est à accroissements indépendants ;*
- $\forall 0 \leq s < t$, *la variable aléatoire* $N(t) - N(s)$ *suit une loi de poisson de paramètre* $\lambda(t - s)$.

1.2 Rappels : loi de Poisson et loi exponentielle :

1.2.1 Définitions et généralité :

Définition 1.2.1. *Une variable aléatoire* X *à valeurs entières suit une loi de Poisson de paramètre* $\lambda > 0$ *si :*

$$\forall k \in \mathbb{N}, \mathbb{P}(X = k) = \frac{\lambda^k}{k!} \exp^{-(\lambda k)}$$

Définition 1.2.2. *Une variable aléatoire* Y *à valeurs réelles strictement positives suit une loi exponentielle de paramètre* $\mu > o$ *si :*

$$\forall t > 0, \mathbb{P}(Y = t) = \mu \exp^{-(\mu t)}$$

1.2.2 Distribution de Poisson

Soit n une variable aléatoire discrète avec $n = 0, 1, ...$ qui suit une distribution Poisson. La distribution de probabilité de n est $P_n = \lambda^n \exp^{-\lambda} / n!$. L'espérence et la variance de n sont $E(n) = \lambda$,et $V(n) = \lambda$, respectivement. La distribution de Poisson peut également être définie en unités de temps t. Dans ce cas, la variable discrète n représente le nombre d'occurrences dans le temps t devient,

$P(n, t) = (\lambda t)^n \exp^{-(\lambda t)} / n!$

1.2.3 Distribution exponentielle

Soit t une variable aléatoire avec $t \geq 0$ qui suit une distribution exponentielle. La densité de probabilité de t est $f(t) = \mu \exp^{-(\mu t)}$ et la distribution cumulée correspondante est $F(t) = 1 - \exp^{-(\mu t)}$. L'espérance et la variance de t sont $E(t) = 1/\mu$, et $V(t) = 1/\mu^2$, respectivement.

1.2.4 Relation entre la distribution Exponentielle et la distribution de Poisson :

La densité de probabilité d'une distribution exponentielle $f(t) = \alpha \exp^{-(\alpha t)}$ Supposons τ est exponentielle avec une espérance $1/\alpha$, et n est de Poisson de moyenne α.on a :

$$\begin{aligned} P(\tau > t) &= 1 - F(t) \\ &= \exp^{-(\alpha t)} \\ &= P(n = 0 \quad \text{en} \quad t) \\ &= P(0,t)^{\alpha} \end{aligned}$$

Notons $P(n,t)$ la probabilité d'avoir n unités dans le temps t.

$$\begin{aligned} P(0,t) &= \exp -(\alpha t) \\ P(1,t) &= \int_{\tau=0}^{t} P(0,\tau) f(1-\tau) d\tau &&= \alpha t \exp -(\alpha t) \\ P(2,t) &= \int_{\tau=0}^{t} P(1,\tau) f(1-\tau) d\tau &&= (\alpha t)^2 \exp -(\alpha t)/2! \\ \ldots &= \\ P(n,t) &= \int_{\tau=0}^{t} P(n-1,\tau) f(1-\tau) d\tau &&= (\alpha t)^n \exp -(\alpha t)/n! \end{aligned}$$

Définition 1.2.3. *Une variable aléatoire X est dite **sans mémoire**(ou sans usure)si : $\forall s, t \geq 0$*

$$\mathbb{P}(X > t + s / X > t) = \mathbb{P}(X > s)$$

Si X est la durée de vie d'un matériel quelconque l'équation précédante s'interprété de la manière suivante, sachant le matériel en état de bon fonctionnement au temps t, la loi de probabilité de sa durée de vie future est la même que celle de sa durée de vie initiale. En d'autres termes, le matériel ne s'use pas.

Exemple : Une variable aléatoire de loi exponentielle est sans mémoire.

Remarque 1.1. *L'unique loi de probabilité continue sans mémoire est la loi exponentielle, cette définition est similaire à la version discrète à l'exception des variable s et t sont réelles positives et non entières, plutôt que de compte le nombre d'essais jusqu'au premier succès on peut penser à l'heure d'arrivée du premier **appel téléphonique dans un centre d'appel**.*

1.3 Le Processus de renouvellement

Introduction :

Un processus de renouvellement à pour fonction de dénombrer les occurrences d'un phénomène donné, lorsque les délais entre deux occurrences consécutives sont des variables aléatoires indépendantes et identiquement distribuées.

Exemple

Il peut s'agir de compter le nombre de pannes d'un matériel électronique en théorie de la fiabilité (le matériel est alors renouvelé après chaque panne, d'où la dénomination), de dénombrer les arrivées de clients dans une file d'attente, de recenser les occurrence d'un sinistre pour une compagnie d'assurance...

Définition 1.3.1. ***(processus de renouvellement)***
Un processus de comptage dont la suite des inter-arrivées forme une suite de variables aléatoires indépendantes et identiquement distribuées s'appelle processus de renouvellement.

Définition 1.3.2. ***(processus de renouvellement)***
Soit$(X_n)_{n\geq 0}$ *une suite de variables aléatoire positives on note* S_n *la suite des sommes partielles :* $S_0 = 0$ *et* $S_n = X_n + S_{n-1}$ *pour tout* $n \geq 1$ *on considère alors le processus* R_t *défini comme suit :*

$$R_t = card\{n \geq 1, S_n \leq t\} = \sum_{n\geq 1} \mathbb{1}_{\{S_n \leq t\}}$$

Par exemple, si les X_n *modélisent les durées de vie d'une ampoule* R_t *représente le nombre d'ampoules changées avant l'instant* t*; les* X_n *peuvent également représenter le temps séparant deux ventes successives, ou deux sinistres successifs pour une compagnie d'assurance.* R_t *désignera alors, suivant le cas, le nombre d'articles vendus ou le nombre sinistres survenus au cours de l'intervalle de temps* $[0, t]$*, la suite* (S_n) *est appelée processus de renouvellement associé aux* $(X_n)_{n\geq 0}$ *et le processus* (R_t) *est le processus de comptage.*
Par abus de langage, on appelle également R_t *Processus de renouvellement.*

1.4 Le Processus de Poisson

Introduction :

De nombreux phénomènes aléatoires se manifestent par des "arrivées" survenant une par une à des instants aléatoires successifs.

Exemples :
- arrivées d'appels à un central téléphonique ;
- impacts de micrométéorites sur un satellite ;
- passage de véhicules à un péage d'autoroute ;

De tels phénomènes peuvent se définir par la famille $(A_n)_{n\in\mathbb{N}^*}$ des temps d'arrivées qui sont des variables aléatoires. Mais on peut aussi le faire à partir du processus de comptage $(N_t)_{t\in\mathbb{R}_+}$, ou par la famille $(T_n)_{n\in\mathbb{N}^*}$ des intervalles de temps entre deux arrivées.

- N_t :est le nombre d'événements apparus jusqu'à l'instant t.

$N_{t+u} - N_u$ est le nombre dévènements apparus entre u et u + t.
L'éspace des états du processus $(N_t)_{t\in\mathbb{R}_+}$ est $E = \mathbb{N}$ et l'éspace des temps est $T = \mathbb{R}_+$. Le processus qui modélise convenablement les exemples cités est le processus de Poisson. On conviendra que $N_0 = 0$.

On note :
- A_n l'instant de réalisation du n^{ime} événement ;
- T_n la durée séparant le $(n-1)^{ime}$ événement du n^{ime} événement pour $n \geq 2$ et $T_1 = A_1$.

On a :
- $A_n = T_1 + T_2 + ... + T_n$ pour tout $n \in \mathbb{N}^*$
- $T_1 = A_1$,$T_n = A_n - A_{n-1}$ pour tout $n \geq 2$

Ainsi, la connaissance de la famille $(A_n)_{n\in\mathbb{N}^*}$. équivaut à celle de la famille $(T_n)_{n\in\mathbb{N}*}$. D'autre part,$A_n \leq t$ signifie que le n^{ime} événement a eu lieu à l'instant t ou avant, c'est-àdire qu'à l'instant t, au moins n événements ont eu lieu, c'est-à-dire que $N_t \geq n$ Ainsi

$$F_{A_n}(t)=P([A_n \leq t])=P([N_t \geq n])=1-\sum_{k=0}^{n-1} P([N_t = k]$$

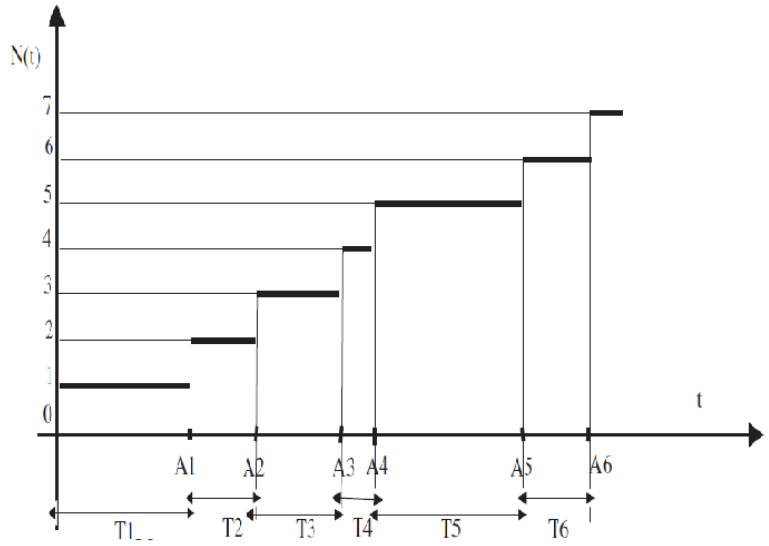


FIGURE 1.1 – Processus de comptage

et

$P([N_t = n])$=$P([N_t \geq n]) - P([N_t \geq n+1])$=$P([A_n \leq t]) - P([A_{n+1} \leq t])$

Par conséquent, la connaissance de $(N_t)_{t\in\mathbb{R}_+}$ équivaut à celle de la famille $(A_n)_{n\in\mathbb{N}^*}$.

Définition 1.4.1. *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.*

Définition 1.4.2. *Un processus de comptage $N_t, t \geq 0$ est un processus de poisson de taux $\lambda > 0$, si*

- *i) le processus est à accroissements indépendants et stationnaires ;*
- *ii) $P(N_h = 1) = \lambda h + o(h)$*
- *iii) $P(N_h \geq 2) = o(h)$*

Théorème 1.4.1. *[3] Les définitions 1.4.1 et 1.4.2 sont équivalentes.*

Preuve :

Nous montrons que définition 1.4.2 ⇒ définition 1.4.1. Posons

$$P_n(t) = P(N_t = n)$$

Montrons dans un premier temps que p_0 est solution d'une certaine équation différentielle

$$\begin{aligned} P_0(t+h) &= P(N_{t+h} = 0) \\ &= P(N_t = 0; N_{t+h} - N_t = 0) \\ &= P(N_t = 0)P(N_{t+h} - N_t = 0) \\ &= P_0(t)[1 - \lambda h + o(h)] \end{aligned}$$

Dans la troisième équation, nous avons utilisé l'assertion i). Nous avons pu poursuivre en combinant ii) et iii) pour aboutir à la dernière équation. On obtient finalement

$$\frac{P_0(t+h) - P_0(t)}{h} = -\lambda P_0(t) + \frac{o(h)}{h}$$

En passant à la limite $h \to 0$, on obtient

$$P_0^{'} = -\lambda P_0$$

L'intégration de cette équation linéaire d'ordre un conduit, en tenant compte de la condition initiale $P_0(0) = 1$ à

$$P_0(t) = e^{-\lambda t}$$

En procédant de la même manière, pour tout $n > 0$, il vient

$$\begin{aligned} P_n(t+h) &= P(N_{t+h} = n) \\ &= P(N_t = n; N_{t+h} - N_t = 0) \\ &+ P(N_t = n-1; P(N_{t+h} - N_t = 0) \\ &+ \sum_{k=2}^{n} P(N_t = n-k; N_{t+h} - N_t = k) \end{aligned}$$

Par l'assertion iii), le dernier terme est d'ordre o(h). En utilisant à nouveau l'assertion i), on otient

$$\begin{aligned} P_n(t+h) &= P_n(t)P_0(h) - P_{n-1}(t)P_1(h) + o(h) \\ &= (1 - \lambda h)P_n(t) + \lambda h P_n(t) + o(h) \end{aligned}$$

Soit, en passant à la limite $h \to 0$

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

Posons

$$q_n(t) = e^{\lambda t} P_n(t)$$

Alors, en multipliant les deux termes par $e^{\lambda t}$ comme

$$\frac{d}{dt} q_n(t) = -\lambda q_{n-1}(t)$$

On termine la démonstration en vérifiant par récurrence que

$$q_n(t) = \frac{\lambda^n t^n}{n!}$$

est l'unique solution de ce système sous la condition $q_n(0) = 0$. La réciproque est facile à établir. C'est un exercice.

1.4.1 Caractérisation d'un processus de Poisson par ses temps d'arrivée :

Soit A_n l'instant de la n^{ime} arrivée : $A_n = inf\{t > 0; N_t = n\}$ 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 $N_t = \max\{n > 0; A_n \leq t\}$

Théorème 1.4.2. *[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 $\varepsilon(\lambda)$ (de densité)*

$$f_{T_n}(t) = \lambda e^{-\lambda t} \prod_{]0;+\infty[}$$

Preuve :

$P([T_1 > t]) = P([N_t = 0]) = e^{-\lambda t} = 1 - F_1(t)$

où F_1 est la fonction de répartition de T_1. On a donc bien T_1 qui suit la loi exponentielle $\varepsilon(\lambda)$.

$$\begin{aligned} P^{[T_1=t_1]}([T_2 > t]) &= P([N_{t_1+t} = 1]/[N_{t_1} = 1] \cap [N_s = 0 \text{ pour tout } s < t_1]) \\ &= P([N_{t_1+t} - N_{t_1} = 0]/[N_{t_1} = 1] \cap [N_s = 0 \quad pour \quad tout \quad s < t_1]) \\ &= P([N_{t_1+t} - N_{t_1} = 0]) \end{aligned}$$

d'après l'indépendance des accroissements.

Or $P([N_{t_1+t} - N_{t_1} = 0]) = P([N_{t_1+t-t_1} = 0]) = P([N_t = 0])$ d'après la stationnarité ; et c'est aussi $e^{-\lambda t}$ car N_t suit la loi de Poisson $P(\lambda t)$.

Donc T_2 est bien indépendante de T_1 et de même loi exponentielle $\varepsilon(\lambda)$.

De façon plus générale,

$$\begin{aligned} P([T_k > t]/[T_1 = t_1] \cap ...[T_{k-1} = t_{k-1}]) &= P([N_{t_{k-1}+t} - N_{t_{k-1}} = 0]) \\ &= P([N_t = 0]) \qquad = e^{-\lambda t} \end{aligned}$$

Donc T_k est indépendante de $T_1, ..., T_{k-1}$ et de même loi exponentielle $\varepsilon(\lambda)$. La réciproque sera admise.

Conséquence :
Les variables aléatoires A_n suivent la loi Gamma $\gamma(\lambda, n)$ (ou loi d'Erlang), de densité d'efinie par

$$f_{A_n}(t) = \frac{\lambda^n}{(n-1)!} e^{-\lambda t} t^{n-1} \prod_{]0;+\infty[}(t)$$

Propriété 1.4.1. *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 $\varepsilon(\lambda)$.*

Preuve :

$$P([U > x]) = P([N_{\theta+x} - N_\theta = 0]) = P([N_x = 0]) = e^{-\lambda t}$$

car $[U > x]$ signifie que pendant la durée x qui suit θ, il n'y a aucune arrivée. De même,

$$P([V > x]) = P([N_\theta - N_{\theta-x} = 0]) = P([N_x = 0]) = e^{-\lambda t}$$

car $[V > x]$ signifie que pendant la durée x qui précédait θ, il n'y a eu aucune arrivée.

Remarque 1.2. *[1] On a alors $\mathbb{E}(U + V) = \mathbb{E}(U) + \mathbb{E}(V) = \frac{2}{\lambda}$ alors que $\mathbb{E}(T_n) = \frac{1}{\lambda}$ pour tout $n \in \mathbb{N}^*$. C'est donc que $U + V = T_n$ n'a pas même loi que les T_n alors que sur $[N_\theta = n]$, on a $T_{N_\theta} = T_n$.*

On peut terminer ce paragraphe en remarquant que $\mathbb{E}(N_1) = \lambda$ et $\mathbb{E}(T_n) = \frac{1}{\lambda}$ Ainsi, plus λ est grand, plus le nombre moyen d'arrivées par unité de temps est important, et plus l'intervalle entre 2 arrivées est court, ce qui semblait a priori évident. Pour cette raison, on appelle également le paramètre λ l'intensité du processus.

1.5 Le Processus de naissance et de mort

1.5.1 Généralités :

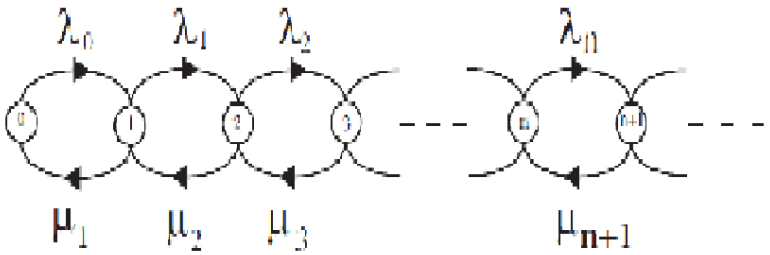
Utilisés plus particulièrement en biologie, démographie, physique, sociologie, pour rendre compte de l'évolution de la taille d'une population, les processus de naissance et de mort sont des processus de Markov continus ($T = \mathbb{R}_+$), à valeurs dans $\mathbb{E} = \mathbb{N}$ tels que les seules transitions non négligeables possibles à partir de k soient vers $k+1$ ou vers $k-1$. Le générateur infinitésimal du processus est donc une matrice dite "tridiagonale" $A = (a_{i,j})_{i,j\in\mathbb{N}}$ vérifiant $a_{i,j} = 0$ si $|i-j| \geq 2$.

On posera $a_{k,k+1} = \lambda_k$ et $a_{k,k-1} = \mu_k$ pour $k \geq 0$ (et $a_{0,1} = \lambda_0$) :λ_k représente le taux de naissance à partir de l'état k et μ_k le taux de mort à partir de l'état n.
Les files d'attente de type Markovien (M/M) sont des cas particuliers trés importants de processus de naissance et de mort.

$$A = \begin{pmatrix} -\lambda_0 & \lambda_0 & \dots & \dots & (0) \\ \mu_1 & -(\lambda_1+\mu_1) & \lambda_1 & \dots & \dots \\ \vdots & \mu_2 & -(\lambda_2+\mu_2) & \lambda_2 & \dots \\ \vdots & \dots & \mu_3 & \ddots & \ddots \\ (0) & \ddots & \ddots & \ddots & \ddots \end{pmatrix}$$

de graphe des taux [4] :

Le graphe des taux de transition est constitué de "boudins", les arcs vers la droite représentant les taux de naissance, et ceux vers la gauche les taux de mort. Ces processus sont l'analogue des "chemins aléatoires" dans le cas où l'échelle des temps est continue.



1.5.2 Régime transitoire :

On rappelle que si $P(t) = (P_{i,j}(t))$, où $P_{i,j}(t) = P([X_{u+t} = j]/[X_u = i])$, alors

$$P(t) = e^{At} = \sum_{k=0}^{+\infty} \frac{t^k}{k} A^k$$

La loi de X_t est alors donnée, en théorie, par $\overrightarrow{\pi}(t) = \overrightarrow{\pi}(0)P(t)$ (On notera ici, par commodité d'écriture, $\overrightarrow{\pi}(t) = (\pi_n(t))_{n\in\mathbb{N}}$, avec $\pi_n(t) = P([X_t = n])$. Malheureusement, le calcul de e^{At} s'avére bien souvent très compliqué (puissances de matrice, puis série à sommer !). On préfèrera, en général, garder l'expression différentielle, même si celle-ci est souvent tout aussi difficile à résoudre.

Équations de Kolmogorov : ona a $\overrightarrow{\pi}(t) = \overrightarrow{\pi}(0)P(t)$ qui redonne, en dérivant

$$\overrightarrow{\pi}(t) = \overrightarrow{\pi}(0)P'(t) = \overrightarrow{\pi}(0)P(t)A = \overrightarrow{\pi}A$$

et donc $\pi_j'(t) = \sum_i \pi_i(t)a_{i,j}$ qui donne ici le système suivant, dit d'équations de Kolmogorov :

$$\begin{cases} \pi_0'(t) = -\lambda_0\pi_0(t) + \mu_1\pi_1(t) \\ \pi_n'(t) = -\lambda_{n-1}\pi_{n-1}(t) - (\lambda_n + \mu_n)\pi_n(t) + \mu_{n+1}\pi_{n+1}(t) \quad pour \quad n > 1 \end{cases}$$

1.5.3 Régime permanent :

Lorsque, quand $t \to +\infty$, les limites $\pi_n = \lim_{t\to+\infty} \pi_n(t)$ existent et sont indépendantes de l'état initial du processus, on a alors $\overrightarrow{\pi}' = \overrightarrow{0}$ et $\overrightarrow{\pi}A = \overrightarrow{0}$ i.e. $\overrightarrow{\pi}$ est distribution stationnaire du processus si le régime permanent existe. Ceci se traduit par les équations dites "de balance"

$$\begin{cases} 0 = -\lambda_0\pi_0 + \mu_1\pi_1 \\ 0 = -\lambda_{n-1}\pi_{n-1} - (\lambda_n + \mu_n)\pi_n + \mu_{n+1}\pi_{n+1} \quad pour \quad n > 1 \end{cases}$$

que l'on retrouve en écrivant en chaque état l'égalité du flux entrant et du flux sortant :

* état 0 : $\mu_1\pi_1 = \lambda_0\pi_0$;

* état 1 : $\mu_2\pi_2 + \lambda_0\pi_0 = \mu_1\pi_1 + \lambda_1\pi_1$

⋮

* état n : $\mu_{n+1}\pi_{n+1} + \lambda_{n-1}\pi_{n-1} = \mu_n\pi_n + \lambda_n\pi_n$

⋮

auxquelles il faut ajouter l'équation $\sum_{n=0}^{+\infty} \pi_n = 1$ pour que $\overrightarrow{\pi}$ définisse bien une probabilité.

Ces équations se simplifient successivement pour donner finalement des égalités "boudins par boudins" :

$$\begin{cases} \mu_1\pi_1 = \lambda_0\pi_0 \\ \mu_2\pi_2 = \lambda_1\pi_1 \\ \vdots \\ \mu_n\pi_n = \lambda_{n-1}\pi_{n-1} \\ \vdots \end{cases}$$

On en déduit alors $\pi_n = \frac{\lambda_0\cdots\lambda_{n-1}}{\mu_1\cdots\mu_n}\pi_0$ pour $n \geq 1$, avec $\pi_0(1 + \sum_{n=1}^{+\infty} \frac{\lambda_0\cdots\lambda_{n-1}}{\mu_1\cdots\mu_n} = 1)$

1.5.4 Étude de quelques cas particuliers :

On étudie ici les cas particuliers des populations où les taux de naissance et de mort sont linéaires :

$$\lambda_n = n\lambda + \alpha$$

- α représente le taux d'immigration : arrivées venant de l'extérieur. Il est supposé constant (et donc indépendant du nombre d'individus déjà présents).
- λ représente le taux de naissance : chaque individu est susceptible de donner naissance à un nouvel individu avec le taux λ et s'il y a déjà n individus, la probabilité qu'il y ait une naissance sur l'intervalle $[t, t+h[$ est alors $n\lambda h + o(h)$.

$$\mu_n = n\mu + \beta \quad \text{si} \quad n \geq 1$$

- β représente le taux d'émigration : départs vers l'extérieur. Il est supposé constant (sauf s'il n'y a personne, auquel cas il est nul).
- μ représente le taux de mort : chaque individu est susceptible de mourir avec le taux μ et s'il y a déjà n individus, la probabilité qu'il y ait une mort sur l'intervalle $[t, t+h[$ est alors $n\mu h + o(h)$.

Croissance pure, par immigration : C'est le cas $\lambda_n = \alpha$ et $\mu_n = 0$ pour tout $n \in \mathbb{N}$

$$\begin{cases} \pi_0'(t) = -\alpha\pi_0(t) \\ \pi_n'(t) = \alpha\pi_{n-1}(t) - \alpha\pi_n(t) \qquad pour \quad n \geq 1 \end{cases}$$

On retrouve les équations vérifiées par le processus de Poisson N_t de paramètr α. En particulier X_t suit la loi de Poisson $P(\alpha t)$: $\pi_n(t) = e^{-\alpha t}\frac{(\alpha t)^n}{n!}$

Croissance pure, par naissance :

C'est le cas $\lambda_n = n\lambda$ et $\mu_n = 0$ pour tout $n \in \mathbb{N}$ Ceci n'a de sens que si $X_0 = 1$

$$\begin{cases} \pi_0'(t) = 0 \\ \pi_n'(t) = (n-1)\lambda\pi_{n-1}(t) - n\lambda\pi_n(t) \qquad pour \quad n \geq 1 \end{cases}$$

Propriétés :

X_t suit la loi géométrique $G(e^{-\lambda t})$: $\pi_n(t) = e^{-\lambda t}(1 - e^{-\lambda t})^{n-1}$

Preuve : On peut trouver $\pi_n(t)$ par récurrence ascendante sur n (matrice A triangulaire supérieure !)

- $\pi_0'(t) = 0$ donc $\pi_0(t) = \pi_0(0) = 0$.
- $\pi_1'(t) = -\lambda\pi_1(t)$ donc $\pi_1(t) = Ce^{-\lambda t}$ avec $C = \pi_1(0) = 1$
- Supposons que $\pi_{n-1}(t) = e^{-\lambda t}(1 - e^{-\lambda t})^{n-2} \quad$ pour $\quad n \geq 2$. on a

$$\pi_n'(t) + n\lambda\pi_n(t) = (n-1)\lambda e^{-\lambda t}(1 - e^{-\lambda t})^{n-2}$$

C'est une équation différentielle linéaire du premier ordre avec second membre donc, pour la résoudre, on applique la méthode de la variation de la constante.

$\pi_n(t) = Ce^{-n\lambda t}$ avec $C'(t)e^{-n\lambda t} = (n-1)\lambda e^{-\lambda t}(1-e^{-\lambda t})^{n-2}$, soit

$$\begin{aligned} C'(t) &= (n-1)\lambda e^{(n-1)\lambda t}(1-e^{-\lambda t})^{n-2} \\ &= (n-1)\lambda e^{\lambda t}(e^{\lambda t}-1)^{n-2} = \tfrac{d}{dt}((e^{\lambda t}-1)^{n-1}) \end{aligned}$$

donc $C(t)-C(0) = (e^{\lambda t}-1)^{n-1}$ et $C(0) = p_n(0) = 0$ pour $n \geq 2$ Finalement,

$$\pi_n(t) = (e^{\lambda t}-1)^{n-1}e^{-n\lambda t} = e^{-\lambda t}(1-e^{-\lambda t})^{n-1}$$

X_t suit donc la loi géométrique de paramètre $e^{-\lambda t}$ et, en particulier $\mathbb{E}(X_t) = e^{\lambda t}$

Décroissance pure, par décès :

C'est le cas $\lambda_n = 0$ et $\mu_n = n\mu$ pour tout $n \in \mathbb{N}$

On suppose $X_0 = N \geq 1$ Ce modèle décrit l'évolution d'un système composé de N dispositifs indépendants non réparables, dont les durées de fonctionnement obéissent à une même loi exponentielle de paramètre μ (durée de vie moyenne $1/\mu$).

Dans ce cas, il est immédiat que $\pi_n(t) == C_N^n e^{-n\mu t}(1-e^{-\mu t})^{N-n}$ pour $n \in \{0, 1, \cdots, N\}$

En effet, la probabilité qu'un dispositif de durée de vie T fonctionne encore à l'instant t est $p([T > t]) = e^{-\mu t}$ (donc la probabilité qu'il ne fonctionne plus est $(1-e^{-\mu t})$). comme $\pi_n(t)$

est la probabilité qu'il y ait exactement n dispositifs qui fonctionnent à l'instant t et qu'il y a exactement C_N^n façons de choisir les n qui fonctionnent, on a bien le résultat.

On peut le retrouver avec les équations de Kolmogorov :

$$\begin{cases} \pi_n'(t) = -n\mu\pi_n(t) + (n+1)\mu\pi_{n+1}(t) & pour \quad n \in \{0, 1, \cdots, N-1\} \\ \pi_N'(t) = -N\mu\pi_N(t) \end{cases}$$

Propriétés : X_t suit la loi binomiale $\mathrm{B}(N, e^{-\mu t})$: $\pi_n(t) = C_N^n e^{-n\mu t}(1-e^{-\mu t})^{N-n}$

Preuve :

On peut trouver $\pi_n(t)$ par récurrence descendante sur n (matrice A triangulaire inférieure !)

- $\pi_N^{'}(t) + N\mu\pi_N(t) = 0$ donc $\pi_N(t) = Ce^{-N\mu t}$ avec $C = \pi_N(0) = 1$
- Supposons maintenant que $\pi_{n+1}(t) = C_N^{n+1}(e^{-\mu t})^{n+1}(1-e^{-\mu t})^{N-n-1}$ pour $n \in \{1, \cdots, N-1\}$ on a

$$\pi_n^{'}(t) + n\mu\pi_n(t) = (n+1)\mu C_N^{n+1}(e^{-\mu t})^{n+1}(1-e^{-\mu t})^{N-n-1}$$

On applique la méthode de la variation de la constante : $\pi_n(t) = C(t)e^{-n\mu t}$ avec

$$\begin{aligned} C'(t) &= e^{-n\mu t} \times (n+1)\mu \frac{N!}{(n+1)!(N-n-1)!} e^{-(n+1)\mu t}(1-e^{-\mu t})^{N-n-1} \\ &= (N-n)\mu C_N^n e^{-\mu t}(1-e^{-\mu t})^{N-n-1} = \frac{d}{dt}(C_N^n(1-e^{-\mu t})^{N-n}) \end{aligned}$$

donc $C(t) - C(0) = C_N^n(1-e^{-\mu t})^{N-n}$ avec $C(0) = \pi_n(0) = 0$ pour $n < N$ et

$$\pi_n(t) = C_N^n e^{-n\mu t}(1-e^{-\mu t})^{N-n}$$

X_t suit donc la loi binomiale $\mathrm{B}(N, e^{-\mu t})$ et, en particulier $\mathbb{E}(X_t) = Ne^{-\mu t}$

Chapitre 2

Les Systèmes de File d'Attente Classiques

Introduction aux files d'attente :
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.
On s'intéresse essentiellement à deux grandeurs : le nombre de clients dans le système, et le temps passé par un client dans le système. Ce dernier se décompose en un temps d'attente et un temps de service.
Classification des files d'attente
Pour décrire une file d'attente, on doit donc se donner les éléments suivants :

- La nature du processus des arrivèes qui est définie par la distribution des intervalles séparant deux arrivées consécutives.
- La distribution du temps aléatoire de service.
- Le nombre s des stations de service montées en parallèle.
- La capacité N du système. Si $N < \infty$, la file ne peut dépasser une longueur de $N - s$ unités. Dans ce cas, certains clients qui arrivent vers le système n'ont pas la possibilité d'y entre.

2.1 File d'Attente Simple

La file simple
Une file d'attente simple est un système constitué d'un ou plusieurs serveurs et

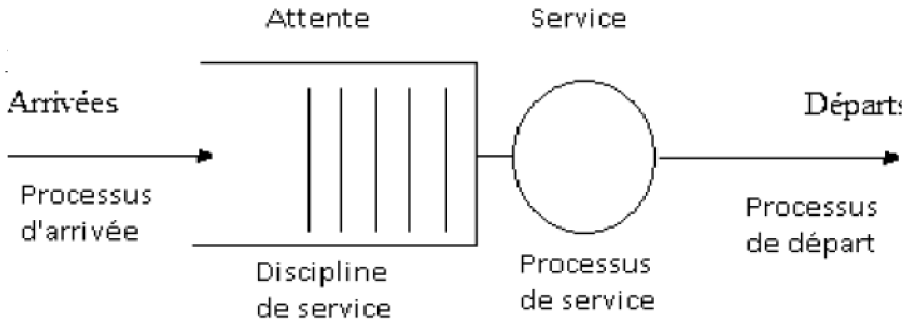


FIGURE 2.1 – Représentation schématique d'une file d'attente simple

d'un espace d'attente. les clients arrivent de l'extérieur, patientent éventuellement dans la file d'attente, reçoivent un service, puis quittent la station. Afin de spécifier complètement une file d'attente simple, on doit caractériser le processus d'arrivée des clients, le temps de service ainsi que la structure et la discipline de service de la file d'attente .

Processus d'arrivée

L'arrivée des clients à la station sera décrite à l'aide d'un processus stochastique de comptage $(N_t)_{t\geq 0}$.

Si A_n désigne la variable aléatoire mesurant l'instant d'arrivée du n^{ime} client dans le système, on aura ainsi : $A_0 = 0$ et $A_n = \inf\{t; N_t = n\}$.

Si T_n désigne la variable aléatoire mesurant le temps séparant l'arrivée du $(n-1)^{ime}$ client et du n^{ime} client, on a alors :

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

Temps de service

Considérons tout d'abord une file à serveur unique.

On note $\mathbf{D}_n$ la variable aléatoire mesurant l'instant de départ du n^{ime} client du système et $\mathbf{Y}_n$ la variable aléatoire mesurant le temps de service du n^{ime} client (le temps séparant le début et la fin du service). Un instant de départ correspond toujours à une fin de service, mais ne correspond pas forcément à un début de service. Il se peut en effet qu'un client qui quitte la station laisse celle-ci vide. le serveur est alors inoccupé jusqu'à l'arrivée du prochain client.

On note μ le taux de service :

$$1/\mu \text{ est la durée moyenne de service.}$$

Structure de la file :

Nombre de serveurs

Une station peut disposer de plusieurs de plusieurs serveurs en parallèle. Soit C le nombre de serveurs. Dès qu'un client arrive à la station, soit il y a un serveur de libre et le client entre instantanément en service, soit tous les serveurs sont occupés et le client se place dans la file en attente de libération d'un des serveurs. Mais en suppose à la plupart du temps que les serveurs sont identiques et indépendants les uns des autres.
Une station particulière est la station IS (infinite servers) dans la quelle le nombre de serveurs est infini. Cette station ne comporte donc pas de file d'attente.

Capacité de la file :
La capacité de la file à accueillir des clients en attente de service peut ètre finie ou infinie. Soit K la capacité de la file, une file à capacité illimitée vérifie $K = +\infty$.

2.2 Notation de Kendall :

La notation suivante, appelée la notation de Kendall, est largement utilisée pour classer les différents Systémes 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 :Interarrivées des clients sont identiquement distribuées selon une loi exponentielle. Il correspond à un processus de Poisson ponctuel (propriété sans mémoire).
   - D :Les temps interarrivées des clients ou les temps de service sont constants et toujours les mêmes.
   - GI :Interarrivées des clients ont une distribution générale (il n'y a aucune hypothèse sur la distribution mais les interarrivées sont indépendantes et identiquement distribuées).
   - G :Interarrivées des clients ont une distribution générale et peuvent être dépendantes.
   - E_k :Ce symbole désinge un processus où les intervalles de temps entre deux arrivées succesives 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.

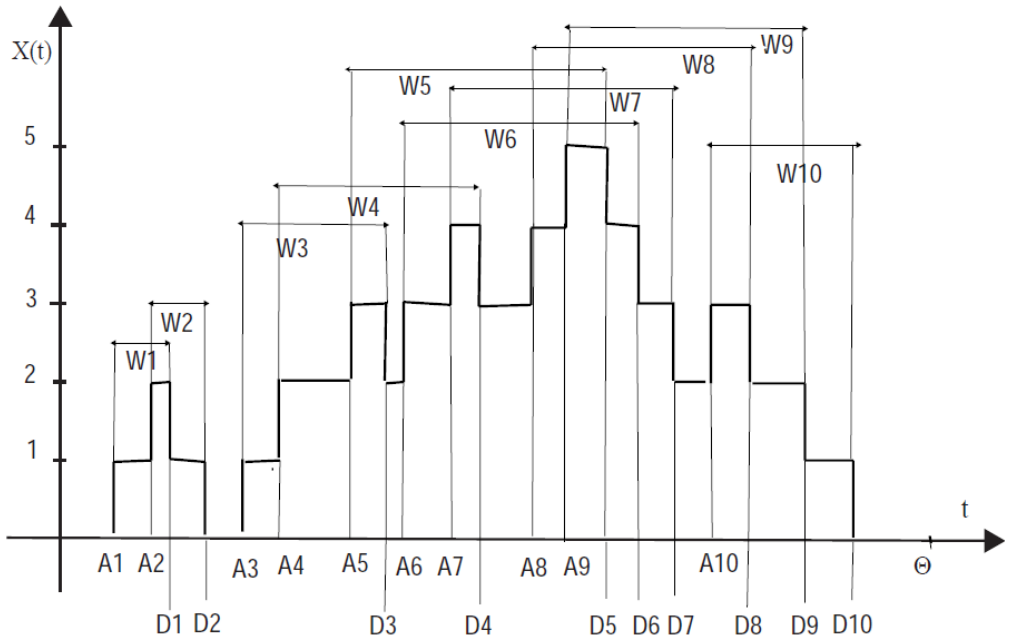
4. K :capacité de la file c'est le nombre de places dans le système en d'autre terme c'est le nombre maximal de clients dans le système y compris ceux en service.

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 standard dans laquelle les clients sont servis dans leur ordre d'arrivée. Notons que les disciplines FIFO et FCFS ne sont pas équivalentes lorsque la file contient plusieurs serveurs. Dans la première, le premier client arrivé sera le premier à quitter la file alors que la deuxième, il sera le premier à commencer son service. Rien n'empêche alors qu'un client qui commence son service après lui, dans un autre serveur, termine avant lui.
- LIFO (last in, first out) ou LCFS (last come, first served). Cela correspond à une pile, dans laquelle le dernier client arrivé (donc posé sur la pile) sera le premier traité (retiré de la pile). A nouveau, les disciplines LIFO et LCFS ne sont équivalentes que pour une file monoserveur.
- SIRO (Served In Random Order), les clients sont servis aléatoirement.
- PNPN (Priority service), les clients sont servis selon leur priorité. Tous les clients de la plus haute priorité sont servis premiers, puis les clients de priorité inférieur sont servis, et ainsi de suite.
- PS (Processor Sharing), les clients sont servis de manière égale. La capacité du système est partagée entre les clients.

Remarque 2.1. *:Dans sa version courte, seuls les tois 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 ilimités.*



2.3 Paramètres de performances opérationnels :

2.3.1 Paramètres de performances en régime transitoire :

Considérons le comportement du système sur une période de temps donnée, par exemple entre $t = 0$ et $t = \theta$. soit X_t le nombre total de clients dans le système à l'instant t. S'intéresser au comportement du système sur l'intervalle de temps $[0, \theta]$ revient à considérer le "régime transitoire" du système.

Définissons les "paramètres opérationnels" suivants :

W_t : temps de séjour du k^{ime} client dans le système : $W_k = D_k - A_k$

θ : temps total de l'observation.

$T(n, \theta)$: temps total pendant lequel le système contient n clients ; on a bien sûr :

$$\sum_{n \geq 0} T(n, \theta) = \theta$$

$P(n, \theta) = \frac{T(n,\theta)}{\theta}$: proportion de temps pendant laquelle le système contient n clients.

$\alpha(\theta)$: nombre de clients arrivant dans le système pendant la période $[0, \theta]$

$\delta(\theta)$: nombre de clients quittant le système pendant la période $[0, \theta]$

À partir de ces quantités, on définit les paramètres de performances en régime transitoire suivants :

Débit moyen d'entrée : Le débit moyen d'entrée est le nombre moyen de clients arrivés dans le système par unité de temps. Sur la période d'observation $[0, \theta]$, c'est donc :

$$d_e(\theta) = \frac{\alpha(\theta)}{\theta}$$

Débit moyen de sortie : Le débit moyen de sortie est le nombre moyen de clients ayant quitté le système par unité de temps. Sur la période d'observation $[0, \theta]$, c'est donc :

$$d_s(\theta) = \frac{\delta(\theta)}{\theta}$$

Nombre moyen de clients : Le nombre moyen de clients présents dans le système est la moyenne temporelle de X_t (ou $X(t)$) sur la période d'observation $[0, \theta]$, c'est donc l'aire sous la courbe de X_t :

$$L(\theta) = \frac{1}{\theta} \sum_{n=0}^{+\infty} nT(n, \theta) = \sum_{n=0}^{+\infty} nP(n, \theta)$$

Temps moyen de séjour : Le temps moyen de séjour d'un client dans le système est, par définition, la moyenne arithmétique des temps de séjour des clients arrivés dans le système pendant l'intervalle de temps $[0, \theta]$:

$$W(\theta) = \frac{1}{\alpha(\theta)} \sum_{k=1}^{\alpha(\theta)} W_k$$

Taux d'utilisateur U : Pour une file simple comportant un unique serveur, en plus des paramètres précédents, débit moyen d'entrée, débit moyen de sortie, nombre moyen de clients, temps moyen de séjour, on peut définir un paramètre de performances supplémentaire : le taux d'utilisation du serveur. Il est défini comme la proportion de temps pendant laquelle le serveur est occupé sur l'intervalle de temps $[0, \theta]$:

$$U(\theta) = \sum_{n=1}^{+\infty} P(n, \theta) = 1 - P(0, \theta)$$

Cas des réseaux de files d'attente :

Pour un réseau de files d'attente, on peut considérer les paramètres de performances du réseau tout entier : débit moyen d'entrée dans le réseau, temps

moyen de séjour dans le réseau. Notons que ces paramètres n'ont d'intérêt que pour un réseau ouvert (ou mixte). En effet, si le réseau est fermé, le débit moyen d'entrée et le débit moyen de sortie sont nuls, le nombre de clients est en permanence égal à la population du réseau (N) et le temps moyen de séjour est égal à la durée d'observation θ (infinie en régime permanent). Pour un réseau multiclasses ouvert (ou mixte), on pourra s'intéresser aux paramètres de performances, par classe ou toutes classes confondues.

On peut également considérer les paramètres de performances pour chacune des stations du réseau : débit moyen d'entrée dans la station i (d_{ei}), débit moyen de sortie de la station i (d_{si}), nombre moyen de clients dans la station i (L_i), temps moyen de séjour dans la station i (W_i). À nouveau, pour un réseau multiclasses, on pourra s'intéresser aux paramètres de performances de chaque classe ou toutes classes confondues.

2.3.2 Paramètres de performances en régime permanent :

Toutes les quantités précédentes définissent les performances du système en régime transitoire (au bout d'un temps θ fini). En régime permanent, on s'intéressera à l'existence et aux valeurs (éventuelles) des limites lorsque θ tend vers l'infini de tous ces paramètres :

$$d_e = \lim_{\theta \to +\infty} d_e(\theta); d_s = \lim_{\theta \to +\infty} d_s(\theta)$$

$$L = \lim_{\theta \to +\infty} L(\theta); W = \lim_{\theta \to +\infty} W(\theta)$$

$$U = \lim_{\theta \to +\infty} U(\theta)$$

2.3.3 Stabilité :

Définition : Un système est stable si et seulement si le débit moyen asymptotique de sortie des clients du système est égal au débit moyen d'entrée des clients dans le système :

$$\lim_{\theta \to +\infty} d_s(\theta) = \lim_{\theta \to +\infty} d_e(\theta) = d$$

D'après les relations précédentes, cela implique que le nombre total de clients arrivés dans le système pendant l'intervalle $[0, \theta]$, $\alpha(\theta)$ ne soit pas croître

plus rapidement que le nombre total de clients ayant quitté le système $\delta(\theta)$, lorsque θ tend vers l'infini :

$$\lim_{\theta \to +\infty} \frac{\delta(\theta)}{\alpha(\theta)} = 1$$

2.3.4 Ergodicité :

L'ergodicité est une notion très importante dans le domaine des processus stochastiques. D'un côté, l'analyse opérationnelle s'intéresse à une évolution particulière d'un système entre deux instants $t = 0$ et $t = \theta$ Nous avons vu que faire tendre θ vers l'infini et considérer les limites de tous les paramètres de performances opérationnels, revient à s'intéresser au régime permanent du système. En fait, cela revient à s'intéresser au régime permanent d'une évolution particulière du système. Il est alors possible d'étudier différentes évolutions du système. La question que l'on se pose immédiatement est bien sûr : toutes ses réalisations ont-elles le même comportement asymptotique ? En d'autres termes, tous les paramètres de performances considérés ont-ils la même limite quelle que soit l'évolution du système considérée ?

D'un autre côté, l'analyse stochastique va associer au système des variables aléatoires et des processus stochastiques :

A_k : variable aléatoire mesurant l'instant d'arrivée du k^{ime} client dans le système ;

D_k : variable aléatoire mesurant l'instant de départ du k^{ime} client du le système ;

W_k : variable aléatoire mesurant le temps de séjour du k^{ime} client dans le système :

$$W_k = D_k - A_k;$$

(α_t) : processus mesurant le nombre de clients arrivés dans le système à l'instant t.

(δ_t) : processus mesurant le nombre de clients ayant quitté le système à l'instant t.

(X_t) : processus stochastique mesurant le nombre de clients dans le système à l'instant t :

$$X_t = \alpha_t - \delta_t$$

$\pi_n(t)$: probabilité pour que le système contienne n clients à l'instant t : $\pi_n(t) = P([X_t = n])$.

On peut alors, comme cela a été fait dans le cadre de l'analyse opérationnelle, calculer tous les paramètres de performance stochastiques, en régime transitoire et en régime permanent. Le nombre moyen de clients présents dans le système à l'instant t se calcule, par exemple, de la façon suivante :

$$L(t) = \sum_{n=0}^{+\infty} n\pi_n(t)$$

La question que l'on se pose est alors : comment se relient les paramètres de performances stochastiques aux paramètres de performances opérationels ? Par exemple, si l'on s'intéresse à caractériser combien de temps un individu passe, en moyenne dans sa vie, à dormir, on s'intéresse bien à la proportion de temps pendant laquelle il dort et non à la probabilité qu'à un instant quelconque, il soit en train de dormir ! Par ailleurs, les techniques de simulation étudient une réalisation particulière du processus et fournissent donc des paramètres de performances opérationnels.

La notion d'ergocité nous permet de définir une classe de Systémes pour laquelle toutes les réalisations particulières de l'évolution du système sont asymptotiquement et statistiquement identiques, c'est-à-dire :

Définition : Un système est ergodique si et seulement si, quelle que soit la réalisation particulière étudiée du processus stochastique :

$$\lim_{\theta \to +\infty} \sum_{n=0}^{+\infty} n^k P(n, \theta) = \lim_{t \to +\infty} \sum_{n=0}^{+\infty} n^k \pi_n(t) \quad \textit{pour tout } k = 1, 2, \cdots$$

Cela implique que tous les paramètres de performances opérationnels en régime permanent (mesurés ou calculés par simulation à partir de n'importe quelle réalisation particulière du processus) sont égaux aux paramètres de performances stochastiques en régime permanent (obtenus à partir d'une étude analytique de performances). Même si cela n'est pas parfaitement rigoureux, on peut donc considérer cette propriété comme définition équivalente à l'ergodicité.

En choisissant comme paramètres de performances les proportions de temps passé par le système dans l'état n et les probabilités associées pour que le système contienne n clients, on obtient que si le système est ergodique :

$$\lim_{\theta \to +\infty} P(n, \theta) = \lim_{t \to +\infty} \pi_n(t)$$

Dans un système ergodique, on pourra donc confondre en régime permanent, proportions de temps passé dans un état et probabilités d'être dans cet état. On parlera indifféremment de "probabilités en régime permanent", de "probabilités à l'équilibre" ou de "probabilités stationnaires". Tous les paramètres de performances fournis par une analyse stochastique de notre système seront égaux aux paramètres de performances opérationnels, c'est-à-dire à ceux que l'on pourrait "observer" sur n'importe quelle réalisation particulière (suffisamment longue) de l'évolution de notre système. On parlera, de la même façon, des "performances stationnaires" du système.

Notons toutefois qu'il existe des Systémes non ergodiques. Par exemple, une chaîne de Markov non irréductible constitue un système non ergodique (certaines réalisations conduisent dans une sous-chaîne absorbante, d'autres dans une autre). Une chaîne périodique est un autre exemple d'un système non ergodique. Les probabilités stationnaires n'existent pas mais on est cependant capable de déterminer les proportions de temps passé dans chaque état de la chaîne. Notons finalement qu'un système instable n'est également pas un système ergodique puisque la limite lorsque θ tend vers l'infini du nombre moyen de clients $L(\theta)$ n'existe pas (est infinie).

2.4 La Loi de Little

La loi de Little est une relation très général qui s'applique à une grande classe de Systémes.
Elle ne concerne que le régime permanent du système. Aucune hypothèse sur les variables aléatoires qui caractérisent le système (temps d'interarrivées, temps de service,...) n'est nécessaire. La seule condition d'application de la loi de Little est que le système soit stable. le débit du système est alors indifféremment, soit le débit de sortie : $d_s = d_e = d$. La loi de Little s'exprime telle que dans la propriété suivante :

Théorème 2.4.1. *[2] :(Formule de Little)Le nombre moyen de clients, le temps moyen passé dans le système et le débit moyen d'un système stable en régime permanent se relient de la façon suivante :*

$$L = W \times d$$

Preuve :

Considérons, dans un premier temps, une durée d'observation θ qui est telle que le système est vide au début et à la fin de l'observation. Dans ces conditions, le nombre de clients qui ont quitté le système pendant $[0, \theta]$ est égal au nombre de clients qui y sont arrivés : $\delta(\theta) = \alpha(\theta)$. Les quantités $L(\theta)$ et $W(\theta)$ peuvent alors être décrites de la façon suivante :

$$L(\theta) = \sum_{n=0}^{+\infty} nP(n,\theta) = \frac{1}{\theta}\sum_{n=0}^{+\infty} nT(n,\theta)$$

$$W(\theta) = \frac{1}{\delta(\theta)}\sum_{k=1}^{\delta(\theta)} W_k$$

La formule de Little repose sur le résultat suivant :

$$\sum_{n=0}^{+\infty} nT(n,\theta) = \sum_{k=1}^{\delta(\theta)} W_k$$

On peut démontrer formellement que les deux sommations intervenant dans cette égalité représentent deux façons de calculer l'aire sous la courbe de X_t

Comme le débit $d(\theta)$ du système est le rapport du nombre de clients sortis $\delta(\theta)$ sur le temps moyen d'observations θ, on déduit immédiatement des trois relations donnant $L(\theta)$, $W(\theta)$ et $d(\theta)$ que : $L(\theta) = W(\theta) \times d(\theta)$. On peut finalement s'affranchir de l'hypothèse que le système est vide au début et à la fin de l'observation, en faisant tendre θ vers l'infini et en se rappelant que si le système est stable,

$$\lim_{\theta \to +\infty} \frac{\delta(\theta)}{\alpha(\theta)} = 1$$

.

La loi de Little a une très grande importance dans l'analyse des Systémes de files d'attente. Elle permet de déduire l'une des trois quantités (L,W, d) en fonction de la connaissance des deux autres. Elle s'applique sous des formes très diverses. Considérons ici le cas d'une file simple comportant un unique serveur et montrons que la loi de Little peut s'appliquer de différentes façons.

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) :

$$L = W \times d$$

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 L_q, au temps moyen d'attente d'un client avant service W_q, par la relation :

$$L_q = W_q \times d$$

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 L_s, au temps moyen de séjour d'un client dans le serveur qui n'est rien d'autre que le temps moyen de service $\frac{1}{\mu}$, par la relation

$$L_s = \frac{1}{\mu} \times d$$

Comme il n'y a jamais plus d'un client en service, L_s s'exprime simplement :

$$L_s(\theta) = 0P(0,\theta) + 1[P(1,\theta) + P(2,\theta) + P(3,\theta) + \cdots] = 1 - P(0,\theta)$$

L_s n'est donc rien d'autre que le taux d'utilisation du serveur, U, défini comme étant la probabilité que le serveur soit occupé (ou la proportion de temps pendant laquelle le serveur est occupé). Ainsi, dans ce cas, la loi de Little s'écrit :

$$U = \frac{1}{\mu} \times d$$

On a obtenu trois relations en appliquant la loi de Little successivement au système entier, à la file d'attente 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 :

$$W = W_q + \frac{1}{\mu} \quad et \quad L = L_q + L_s = L_q + U$$

.

2.5 Modèles de files d'attente :

2.5.1 Modèle d'attente M/G/1 :

Description du modèle

Les clients arrivent dans le système selon un processus de Poisson de taux $\lambda > 0$ De ce fait, le temps entre deux arrivées successives suit une loi exponentielle de moyenne $\frac{1}{\lambda}$. Le service est assuré par un seul serveur. A l'arrivée d'un client, si le serveur est libre, le client sera pris en charge immédiatement. Dans le cas contraire, il rejoint la file d'attente (de capacité illimitée et discipline FIFO) les durées de service (Se) sont des variables aléatoires indépendantes et identiquement distribuées de loi générale dont la fonction de répartition et B(x) la transformée de Laplace-Stieltjes $\widetilde{B}(x)$.
Soient $\mathbb{E}(Se) = \frac{1}{\gamma}$.

Chaîne de Markov induite :

Nous introduisons le processus stochastique $\{N(t), t \geq 0\}$ qui n'est pas un processus de Markov. Pour le rendre markovien, nous utiliserons la méthode des chaînes de Markov induites.

Soit le processus à temps discret $\{N_n = N(\xi_n), n \geq 1\}$ où ξ_n est l'instant où le n^{ime} client a fini son service et quitte le système. Vérifions que cette suite de variables définit bien une chaîne de Markov.

Soient les (A_n) des variables aléatoires indépendantes et identiquement distribuées telles que A_n est le nombre de clients arrivants pendant le n^{ime} service avec la distribution

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

où $a_i > 0 \quad et \quad i > 0$ Déterminons l'équation fondamentale de la chaîne :

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

Soit la variable aléatoire

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

alors l'équation fondamentale de la chaîne devient :

$$N_{n+1} = N_n - \delta_n + A_{n+1}$$

Il est évident que N_{n+1} dépend de $N_n \quad et \quad A_{n+1}$ seulement et non pas de $N_{n-1}, N_{n-2}, \cdots$ D'où la suite $\{N_n, n \geq 1\}$ est une chaîne de Markov induite du processus $\{N_t, t \geq 0\}$

avec les probabilités de transitions $\mathbb{P}(N_{n+1} = j/N_n = i) = p_{ij}$ qui s'expriment de la manière suivante :

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

Par conséquent, la matrice de transition M est donnée par

$$\begin{bmatrix} p_{00} & p_{01} & p_{02} & p_{03} & \cdots \\ p_{10} & p_{11} & p_{12} & p_{13} & \cdots \\ p_{20} & p_{21} & p_{22} & p_{23} & \cdots \\ p_{30} & p_{31} & p_{32} & p_{33} & \cdots \end{bmatrix} = \begin{bmatrix} a_0 & a_1 & a_2 & a_3 & \cdots \\ a_0 & a_1 & a_2 & a_3 & \cdots \\ 0 & a_0 & a_1 & a_2 & \cdots \\ 0 & 0 & p_1 & p_2 & \cdots \end{bmatrix}$$

Puisque nous pouvons passer d'un état à n'importe quel autre, alors la chaîne de Markov est irréductible. En outre la matrice n'est pas décomposable (elle est apériodique), donc la chaîne est ergodique.

La distribution stationnaire existe si $\rho = \frac{\lambda}{\gamma} < 1$ Nous avons la fonction génératrice

$$\begin{aligned} A(z) &= \sum_{i=0}^{\infty} a_i z^i = \sum_{i=0}^{\infty} z^i \int_0^{\infty} \left(\frac{\lambda t}{i!}\right)^k \exp(-\lambda t) dB(t) \\ &= \int_0^{\infty} \exp(-(\lambda - \lambda z)t) dB(t). \end{aligned}$$

posons $\widetilde{B}(s) = \int_0^{\infty} \exp(-st) dB(t)$, alors $A(z) = \widetilde{B}(\lambda - \lambda z)$ converge pour $\mid z \mid \leq 1$:

1. $\mid z \mid < 1 : 0 < a_k < 1 \quad \Rightarrow \mid a_k z \mid < \mid z \mid^k$
2. $\mid z \mid = 1 : \quad A(1) = \widetilde{B}(0) = 1.$

soit $\rho < 1$. La distribution stationnaire de la chaîne de Markov induite $\{N_n, n \geq 1\}$ possède la fonction génératrice suivante [4]

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

Soient les probabilités suivantes :

$$\begin{aligned} p_j &= \lim_{t \to \infty} \mathbb{P}(N(t) = j), j \geq 0 \\ \pi_j &= \lim_{n \to \infty} \mathbb{P}(N(\xi_n) = j), j \geq 0 \\ r_j &= \lim_{n \to \infty} \mathbb{P}(N(\varsigma_n) = j), j \geq 0 \end{aligned}$$

où ς_n est l'instant d'arrivée du n^{ime} client. Comme le processus des arrivées est celui de Poisson de paramètre λ et le nombre de clients dans le système N(t) est discontinu avec un changement de taille ± 1, alors

$$p_j = \pi_j = r_j$$

Par conséquent, le processus $\{N(t), t \geq 0\}$ a une distribution stationnaire identique à celle de la chaîne de Markov induite et la fonction génératrice du nombre de clients dans le système est $Q(z) = \sum_{j=0}^{\infty} p_j z^j = \prod(z)$

Mesures de performance :

– Nombre moyen de clients dans le système

$$\overline{n} = \rho + \frac{\rho^2 + \lambda^2 Var[Se]}{2}(1-\rho)$$

– Temps moyen de séjour d'un client dans le système

$$\overline{W}_s(s) = \frac{s(1-\rho)}{s - \lambda + \lambda\widetilde{B}(s)}\widetilde{B}(s)$$

– Temps moyen d'attente d'un client

$$\overline{W}(s) = \frac{s(1-\rho)}{s - \lambda + \lambda\widetilde{B}(s)}$$

2.5.2 Modèle d'attente $M^X/G/1$:

Description du modèle :

Considérons un système de files d'attente où le service des clients est assuré par un seul serveur. Les clients arrivent par groupes, ces arrivées des groupes de clients primaires suivent une loi de Poisson de paramètre $\lambda > 0$. Le groupe contient K clients ($1 \leq K \leq \infty$) où K est une variable aléatoire discrète qui est égale à k avec la probabilité c_K et dont la fonction génératrice est $C(z) = \sum_{k=1}^{\infty} c_k z^k$. La taille moyenne des groupes est $\mathbb{E}[K] = C'(1) = \overline{c} = \sum_{K=1}^{\infty} k c_k$.

Le temps de service t suit une loi générale de fonction de répartition $B(t)$ et de transformée de Laplace-Stieltjes $\widetilde{B}(s)$, $Re(s) > 0$. Soient les moments $\beta_k = (-1)^k \widetilde{B}^{(k)}(0)$. Les durées entre deux arrivées consécutives des groupes, la taille des groupes ainsi que les durées de service sont supposées mutuellement indépendantes.

Analyse du modèle :

Considérons un intervalle de temps arbitraire de longueur t. Soit $N(t)$ le nombre de clients arrivant dans l'intervalle de temps t et $\nu(t)$ le nombre de groupes arrivant dans cet intervalle de temps. Alors,

$$\mathbb{P}(\nu(t) = n) = \frac{(\lambda t)^n e^{-\lambda t}}{n!}$$

soient $\alpha_1, \alpha_2, \cdots, \alpha_\nu$ les tailles des groupes, où α_i est indépendante de α_j pour $i \neq j$ et les $\alpha_{i1 \leq i \leq \nu}$

sont des variables aléatoires indépendantes et identiquement distribuées. Nous avons alors

$$N(t) = N = \alpha_1 + \alpha_2 + \cdots + \alpha_\nu$$

$$\mathbb{E}[z^N \ \nu] = [\mathbb{E}[z^\alpha]]^\nu = C^\nu$$

$$\mathbb{E}[z^N] = \sum_{\nu=0}^{\infty} C^\nu(z) \frac{(\lambda t)^\nu e^{-\lambda t}}{\nu!} = e^{-\lambda t[1-C(z)]}.$$

$\mathbb{E}[z^N]$ est la fonction génératrice du nombre d'arrivées dans un intervalle de temps arbitraire de taille t.

Chaîne de Markov induite :

A présent, on considére l'intervalle de temps t comme étant le temps de service actuel, alors la fonction génératrice du nombre d'arrivées sera définie par

$$A(z) = \int_{t=0}^{\infty} \exp(-\lambda(1-C(z))t) B(t) dt = \widetilde{B}(\lambda - \lambda C(z)) \tag{2.2}$$

Soit la chaîne de Markov induite aux instants de départs, de la même méthode utilisée dans le cas du modèle M/G/1 ordinaire, on effectue une analyse de cette chaîne.

En effet, en remplaçant la valeur A (z) de l'équation (2.2) dans le résultat du modèle M/G/1 ordinaire, on obtient

$$\prod(z) = \frac{(1-\rho)\widetilde{B}(\lambda - \lambda C(z))(1-z)}{\widetilde{B}(\lambda - \lambda C(z)) - z}$$

avec $\rho = A'(1) = \lambda C'(1)\beta_1 = \lambda \overline{c} \beta_1$.

Approche alternative à l'analyse du modèle $M^X/G/1$:

Dans cette approche, on considère les groupes des arrivées comme étant des requêtes avec leur temps de service égal à la somme des temps de services de tous les clients du groupe. La durée de service d'un groupe τ^* suit une loi

générale de fonction de répartition $B^*(t)$ et de transformée de Laplace-Stieltjes $\widetilde{B^*}(s)$. Alors

$$\widetilde{B^*}(s) = \sum_{k=1}^{\infty} c_k(\widetilde{B}(s))^k = C(\widetilde{B}(s))$$

,

avec

$$\begin{array}{rcccl} \overline{\tau^*} & = & -\frac{d\widetilde{B^*}(s)}{ds}|_{s=0} & = & [-\widetilde{B}'(s)C'(\widetilde{B}(s))]_{s=0} \\ & = & \beta_1 C'(1) & = & \overline{c}\beta_1 \end{array}$$

et

$$\overline{\tau^{*^2}} = \frac{d^2\widetilde{B^*}(s)}{ds^2}|_{s=0} = \sigma_\tau^2\overline{c} + (\overline{c}^2 + \sigma_k^2)\beta_1^2,$$

où σ_τ^2 est la variance de la variable aléatoire τ qui représente la durée de service d'un client et σ_k^2 est la variance de la variable aléatoire K qui représente la taille d'un groupe de clients. Soit
$A^*(z) = \widetilde{B^*}(\lambda - \lambda z) = C(\widetilde{B}(\lambda - \lambda z))$ la fonction génératrice du nombre des arrivées des groupes durant le temps de service d'un groupe. Comme le groupe en entier est considéré comme une requête de service, alors on utilise la chaîne de Markov induite aux instants de départs. Soit $Q^*(z)$ la fonction génératrice du nombre de groupes restant après le départ d'un groupe. On peut maintenant remplacer $A(z)$ par $A^*(z)$ avec $\rho = \lambda\overline{c}\beta_1$. D'où la fonction génératrice du nombre de clients dans le système à l'instant de départ d'un groupe

$$Q^*(z) = \frac{(1-\rho)C(\widetilde{B}(\lambda - \lambda z))(1-z)}{C(\widetilde{B}(\lambda - \lambda z)) - z}$$

Mesures de performance :

Temps moyen d'attente d'un groupe $\overline{W}_{fg}$ avant qu'il ne commence son service

$$\overline{W}_{fg} = \frac{\lambda}{2(1-\rho)}\overline{\tau^{*^2}}$$

on peut écrire $\overline{W}_{fg}$ sous la forme

$$\overline{W}_{fg} = \frac{\frac{\lambda}{2(1-\rho)}\overline{\tau^{*^2}}\rho\overline{c}\beta_1}{2(1-\rho)}[1 + \frac{S_\tau^2}{\overline{c}} + S_k^2],$$

où $S_\tau^2 = \frac{\sigma_\tau^2}{\beta_1^2}$ et $S_k^2 = \frac{\sigma_k^2}{\overline{c}^2}$ sont les coefficient de variation quadratique de τ et k

Temps moyen d'attente $\overline{W}_2$ pour un appel sachant que le service de son groupe a commencé

$$\overline{W}_2 = \widetilde{W_2}(s)|_{s=1} = \frac{\beta_1[C^{"}(1) - \overline{c}]}{2\overline{c}} = \beta_1[\frac{\overline{c}}{2}(1 + S_k^2) - \frac{1}{2}]$$

Temps moyen d'attente global $\overline{W}_f$ pour un appel (dans le groupe) est donné par

$$\begin{aligned} \overline{W}_f &= \overline{W}_{fg} + \overline{W}_2 \\ &= \frac{\lambda}{2(1-\rho)}\overline{\tau^{*2}} + \frac{\beta_1}{2}[\overline{c}(1+S_k^2) - 1] \end{aligned}$$

Chapitre 3

Les Systèmes De File d'Attente Avec Rappels

3.1 Introduction :

Les systèmes de files d'attente avec rappels sont caractérisés par la propriété qu'un client qui trouve à son arrivée tous les serveurs occupés quitte l'espace de service et rappelle ultérieurement à des instants aléatoires. Entre deux rappels successifs, le client est dit "en orbite". Ces systèmes de files d'attente sont largement utilisés dans la modélisation des systèmes informatiques et des réseaux de télécommunications Une description complète de situations où les systèmes de files d'attente avec rappels peut être trouvée dans la monographie de Falin et Templeton (1997) et dans Une classification bibliographique est donnée dans les articles de Artalejo (1999) et (2010). Le modèle M/G/1 avec rappels et clients non-persistants a été considéré par Falin (1990), par Martin et Artalejo (1995) et Martin et Gomez-Corral (1995) . 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 la discipline d'attente ainsi que la spécification concernant le processus de répétition d'appels. Le modèle général d'un système de files d'attente avec répétition d'appels, peut être décrit comme suit : le système est composé de $c \geq 1$ dispositifs de service et de $m - c(m \geq c)$ positions d'attente. Les clients arrivent dans le système selon un processus aléatoire avec une loi de probabilité donnée, et forment un flux d'appels primaires. A l'arrivée d'un client, s'il y a une position d'attente libre, le client rejoint la file d'attente. Dans le cas

contraire, il quitte l'espace de service temporairement avec une probabilité H_0 pour tenter sa chance après une durée de temps aléatoire, ou il quitte le système définitivement avec une probabilité $1 - H_0$ Entre les tentatives, le client est "en orbite" et devient source d'appels répétés ou d'appels secondaires. La capacité O de l'orbite peut être finie ou infinie. Dans le cas où O est finie et si l'orbite est pleine, le client quitte le système pour toujours. Lorsqu'un client est rappelé de l'orbite, il est traité de la même manière qu'un client primaire avec une probabilité H_k (s'il s'agit de la k^{ime} tentative échouée). La notation de Kendall est A/B/c/m/O/H, où A et B décrivent respectivement la distribution du temps inter-arrivées et la distribution du temps de service, c est le nombre de serveurs identiques et indépendants, $m - c$ est la capacité du tampon, O est la capacité de l'orbite, H est la fonction de persistance $H = H_k, k \geq 0$. Si m,O et H sont absents dans la notation de Kendall, alors $m = c$, $O = \infty$ et $H_k = 1$ pour tout $k \geq 1$ La distribution du temps inter-rappels n'est pas indiquée. On décrit l'entrée dans le système par une suite $(\tau_n^e, M_n), n \geq 1$ (60), où τ_n^e est l'intervalle de temps entre les arrivées des n^{ime} et $(n+1)^{ime}$ clients primaires, M_n est une marque associée au n^{ime} client primaire. Cette marque comprend : τ_n^s la durée de service, wn le nombre maximal de rappels autorisés (on suppose $\omega_n \to \infty$), $\tau_n^r = \tau_{n1}^r, \tau_{n2}^r, \cdots, \tau_{n\omega_n}^r$ une suite d'intervalles de temps entre deux rappels successifs. Les variables aléatoires τ_n^e, τ_n^s et τ_n^r sont indépendantes et définies sur l'espace probabilisé $(\Omega, F, \mathbb{P})$. Dans ce qui suit, on suppose que les suites τ_n^e, τ_n^s et τ_n^r sont des suites indépendantes de variables aléatoires indépendantes et identiquement distribuées.

Dans ce chapitre, nous présentons une étude de certains modèles avec rappels. Nous commençons par les systèmes de files d'attente avec rappels de type M/G/1 avec clients persistants et ceux avec rappels et clients impatients. Puis nous réalisons pour la première fois l'analyse stochastique complète du modèle $M^X/G/1$ avec rappels, arrivées par groupes et clients impatients.

3.2 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.

3.2.1 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 $P(\tau_n^e \leq x) = B(x)$ 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 $\rho = \lambda\beta_1$ 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 $(n-1)^{ime}$ 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^{ime} client n'entre en service. A l'instant $\xi_n = \eta_n + R_n$ le n^{ime} 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^{ime} client achève son service, le serveur devient libre et ainsi de suite

3.2.2 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 $N_0(t)$ 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).

soit q_n la chaîne de Markov induite aux instants de départs, où $q_n = N_0(\xi_n)$ représente le nombre de clients en orbite après le n^{ime} départ, dont l'équation fondamentale est :

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

où ν_{n+1} est le nombre d'clients primaires arrivant dans le système durant le service du $(n+1)^{ime}$ 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 $a_i > 0, \quad i > 0$ On a les résultats suivants si

$$\nu = \lim_{n\to\infty} \nu_n, E(\nu) = \rho; \quad alors \quad A(z) = \sum_{i=0}^{\infty} a_i z^i = \widetilde{B}(\lambda - \lambda z)$$

. 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

$$\begin{aligned} \mathbb{P}(\delta_{q_n} = 1/q_n = i) &= \frac{i\theta}{\lambda + i\theta'} \\ \mathbb{P}(\delta_{q_n} = 0/q_n = i) &= \frac{i\theta}{\lambda + i\theta\cdot} \end{aligned}$$

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

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

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

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

si $\rho < 1$ alors $\lim_{n\to\infty} E(q_{n+1} - q_n/q_n = i) = \rho - 1 < 0$ 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 $\pi_n = \lim \mathbb{P}(N_0(\xi_i) = n)$. Les équations de Kolmogorov se présentent de la manière suivante :

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

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

$$\varphi(z) = \sum_{n=0}^{\infty} z^n \pi_n \quad et \quad \psi(z) = \sum_{n=0}^{\infty} z^n \frac{\pi_n}{\lambda + n\theta'}$$

$$\varphi(z) = A(z)(\lambda\psi(z) + \theta\psi^{'}(z))$$

D'un autre côté,

$$\begin{aligned}\varphi(z) &= \sum_{n=0}^{\infty} z^n \pi_n = \sum_{n=0}^{\infty} z^n \pi_n \frac{\lambda + n\theta}{\lambda + n\theta} \\ &= \lambda \sum_{n=0}^{\infty} z^n \frac{\pi_n}{\lambda + n\theta} + \theta \sum_{n=0}^{\infty} n z^n \frac{\pi_n}{\lambda + n\theta} \\ &= \lambda\psi(z) + \theta\psi^{'}(z) \qquad (3.1)\end{aligned}$$

Par conséquent

$$\begin{aligned}\lambda\psi(z) + \theta\psi^{'}(z) &= A(z)(\lambda\psi(z) + \theta\psi^{'}(z)) \\ \theta\psi^{'}(z)[A(z) + z] &= \lambda\psi(z)[1 - A(z)] \qquad (3.2)\end{aligned}$$

Lemme 3.2.1. *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

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

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

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

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

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

Théorème 3.2.1. *soit $\rho > 1$ La distribution stationnaire de la chaîne de Markov induite possède la fonction génératrice suivante*

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

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

Démonstration : Le lemme3.2.1 nous permet de réécrire l'équation (3.2) pour tout $z \in [0,1]$ comme suit

$$\psi^{'}(z) = \frac{\lambda}{\theta}[\frac{1-A(z)}{A(z)-z}]\psi(z)$$

, qui a pour solution

$$\psi(z) = \psi(1)\exp\{\frac{\lambda}{\theta}\int_1^z \frac{1-A(u)}{A(u)-u}du\}$$

si $\rho < 1$

$$\psi(z) = \psi(1)\exp\{\frac{\lambda}{\theta}\int_1^z \frac{1-A(u)}{A(u)-u}du\}$$

De (3.1), On a

$$\begin{aligned}\varphi(z) &= \lambda\psi(z) + \theta\psi^{'}(z)\\ &= \lambda\psi(z) + \theta\frac{\lambda}{\theta}z\frac{1-A(z)}{A(z)-z}\psi(z)\\ &= \lambda\psi(z)A(z)\frac{1-z}{A(z)-z}\end{aligned}$$

Puisque $\varphi(1) = 1$, on a $\psi(1) = \sum_{n=0}^{\infty} z^n \frac{\pi_n}{\lambda+n\theta} = \frac{1-\rho}{\lambda}$ Enfin, on obtient la fonction génératrice

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

3.2.3 Distribution stationnaire de l'état du système :

Le premier résultat sur le système M/G/1 avec rappels a été obtenu par Keilson et al. (1968), 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) . 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

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

A partir du graphes des transitions 3.1, les probabilités p_{0n} et $p_{1n}(x)$ vérifient le système d'équations de balance :

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

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 $p_1(z,x)=\sum_{n=0}^{\infty} z^n p_{1n}(x)$. Le système d'équations de balance devient

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

D'où

$$\begin{cases}\lambda P_0+\theta z P_0'(z)=\int_0^\infty P_1(z,x)b(x)dx;\\ P_1'(z,x)=(\lambda z-\lambda-b(x))P_1(z,x);\\ P_1(z,0)=\lambda P_0(z)+\theta P_0'(z).\end{cases}\qquad(3.3);$$

De la deuxième équation de (3.3), on a

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

.

Donc, la première équation de (3.3) devient

$$\begin{aligned}\lambda P_0+\theta z P_0'(z) &= \int_0^\infty P_1(z,0)[1-B(x)]\exp(-(\lambda-\lambda z)x)b(x)dx\\ &= P_1(z,0)\widetilde{B}(\lambda-\lambda z)=P_1(z,0)A(z)\end{aligned}\qquad(3.4)$$

A partir des équations (3.3) et (3.4), on a

$$\begin{aligned}P_1(z,0)f(z) &= \lambda P_0(z)+\theta z(\tfrac{P_1(z,0)}{\theta}-\tfrac{\lambda}{\theta}P_0(z));\\ P_1(z,0) &= \tfrac{\lambda-\lambda z}{A(z)-z}P_0(z)[1-B(x)]\exp(-(\lambda-\lambda z)x).\end{aligned}$$

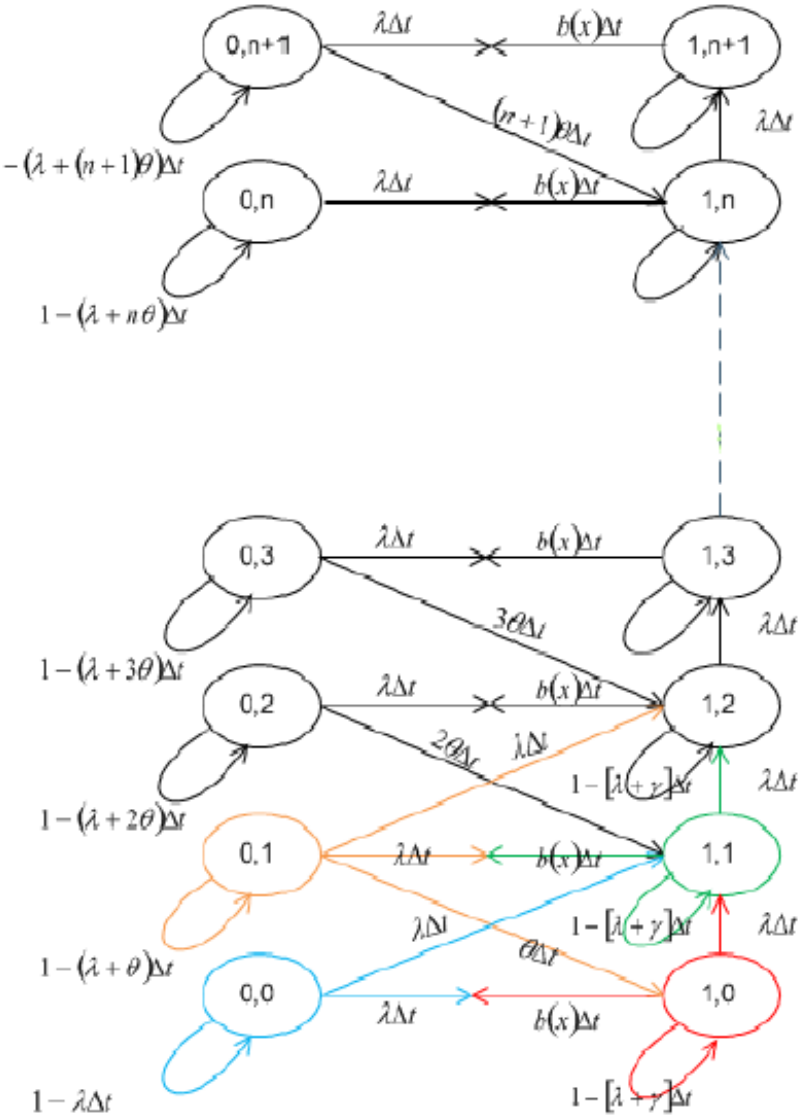


FIGURE 3.1 – Graphe des transitions du modèle M/G/1 avec rappels

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 (3.3) et (3.4), on peut obtenir $p_0(z)$

$$\begin{aligned} \lambda P_0(z) + \theta z P_0'(z) &= A(z)[\lambda P_0(z) + \theta z P_0'(z)]; && (3.5) \\ \theta[A(z) - z]P_0'(z) &= \lambda[1 - A(z)]P_0(z) && (3.6) \end{aligned}$$

Considérons $f(z) = A(z) - z$ Du lemme 3.2.1, $f(z)$ est une fonction décroissante sur $[0,1]$, positive et pour $\rho < 1 \quad et \quad z \in [0,1] : z < A(z) < 1$. En plus $\lim\limits_{z \to 1^-} \dfrac{1 - A(z)}{A(z) - z} = \dfrac{A'(1)}{A(1) - 1} = \dfrac{\rho}{\rho - 1} < \infty$. De ce fait, pour $z = 1$, la fonction $\frac{1-A(z)}{A(z)-z} = \frac{\rho}{\rho-1}$.

Théorème 3.2.2. *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*

$$\begin{aligned} P_0(z) &= \sum_0^\infty z^n P_{0n} = (1-\rho)\exp[\tfrac{\lambda}{\theta}\int_1^z \frac{1 - A(u)}{A(u) - u}du] \\ P_1(z) &= \sum_0^\infty z^n P_{1n} = \tfrac{1-A(z)}{A(z)-z}P_0(z). \end{aligned}$$

Démonstration :

L'équation (3.6) devient

$$P_0'(z) = \frac{\lambda}{\theta}\frac{1 - A(z)}{A(z) - z}P_0(z)$$

La résolution de cette équation nous donne

$$\begin{aligned} P_0(z) &= (1-\rho)\exp[\tfrac{\lambda}{\theta}\int_1^z \frac{1 - A(u)}{A(u) - u}du] \\ P_1(z) &= \tfrac{1-A(z)}{A(z)-z}p_0(z) \end{aligned}$$

De plus, $P_1(1) = \frac{\rho}{1-\rho}P_0(1)$. et, vu que $P_0(1) - P_1(1) = 1$, on obtient $P_1(1) = \rho \quad et \quad P_0(1) = 1 - \rho$.

Par conséquent, la distribution marginale du nombre de serveurs occupés s'exprime de la manière suivante

$$P_0 = \lim_{t\to\infty} P(C(t)=0) = (1-\rho)P_1 = \lim_{t\to\infty} P(C(t)=1) = \rho$$

La fonction génératrice de la distribution marginale de la taille de l'orbite est définie par :

$$\begin{aligned} P(z) &= P_0(z) + P_1(z) = \tfrac{1-z}{A(z)-z} P_0(z) \\ &= \tfrac{(1-\rho)(1-z)}{A(z)-z} \exp[\tfrac{\lambda}{\theta} \int_1^z \frac{1-A(u)}{A(u)-u} du] \end{aligned}$$

et la fonction génératrice de la distribution de l'état stationnaire du nombre de clients dans le système est

$$\begin{aligned} Q(z) &= P_0(z) + P_1(z) \\ &= \tfrac{(1-\rho)(1-z)A(z)}{A(z)-z} \exp[\tfrac{\lambda}{\theta} \int_1^z \frac{1-A(u)}{A(u)-u} du] \end{aligned}$$

3.2.4 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

$$\overline{n}_0 = P'(1) = \overline{n} - \rho = \frac{\lambda^2\beta_2}{2(1-\rho)} + \frac{\lambda\rho}{\theta(1-\rho)}$$

– Temps moyen d'attente d'un client

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

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

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

3.3 Systèmes de files d'attente M^X/G/1 avec rappels et groupes impatients :

La première étude des systèmes de files d'attente avec rappels et arrivées par

groupe était faite par Falin , qui a supposé la règle suivante : " Si le serveur est occupé à une arrivée, alors la totalité du groupe rejoint l'orbite, et si le serveur est libre, alors un des arrivants commence son service et le reste rejoint l'orbite. A l'arrivée d'un groupe, si le serveur est occupé, la totalité du groupe rejoint l'orbite ; dans le cas contraire, l'un des clients arrivant commence son service et le reste rejoint l'orbite.". Ce modèle peut être utilisé pour évaluer la performance des réseaux locaux à bus opérant sous des protocoles comme le CSMA/CD (Carrier Sense Multiple Access with Collision Detection). La plus part des travaux sur les files d'attente avec rappels considéraient le temps d'attente comme une alternative au modèle classique du réseau téléphonique. Dans ce contexte, chaque client bloqué génère des appels répétés indépendamment du reste des clients en orbite. Alors, dans cette situation, les intervalles entre les essais successifs sont exponentiellement distribués de paramètre q, quand le nombre de clients en orbite est j. Ce type de discipline de rappels est connu comme une politique de rappels classique

3.3.1 Description du modèle :

Les clients primaires arrivent dans le système selon un processus de Poisson de taux $\lambda > 1$. Les clients arrivent par groupes de taille K qui est une variable aléatoire, posons $P(K = k) = c_k(t), k \geq 0$, (la probabilité que la taille du groupe soit égale à k à la date t). Le service est assuré par un seul serveur. A l'arrivée d'un groupe primaire, si le serveur est occupé, le groupe entre en orbite avec une probabilité H_1, sinon il quitte le système avec une probabilité $(1 - H_1), (H_2 = 1)$ Par contre si le serveur est libre, l'un des clients sera pris en charge par le serveur et le reste du groupe entre en orbite. Les clients en orbite répètent les appels jusqu'à ce que le serveur soit libre, et ceci avec un taux de rappel $\theta > 0$, qui dépend du nombre de clients en orbite. Les durées inter-rappels sont exponentiellement distribuées ; $T(x) = 1 - e^{-\theta x}, x > 0$.

Les durées de service suivent une loi générale $\mathbb{P}(\tau_n^s \leq x = B(x))$ de transformée de Laplace-Stieltjes $\widetilde{B}(s), Re(s) > 0$. Soient les moments $\beta_k = (-1)^k \widetilde{B}^{(k)}(0)$, le taux de service $\gamma = \frac{1}{\beta_1}$.

Soit la fonction génératrice de la distribution stationnaire de la taille des groupes $C(z) = \sum_{k=1}^{\infty} c_k z^k$ et $\overline{c} = C^{'}(1)$ est la taille moyenne des groupes. L'intensité du trafic est $\rho = \lambda \overline{c} H_1 \beta_1$. Enfin, nous admettons que toutes les variables définies précédemment soient mutuellement indépendantes. L'état du système est décrit par le processus

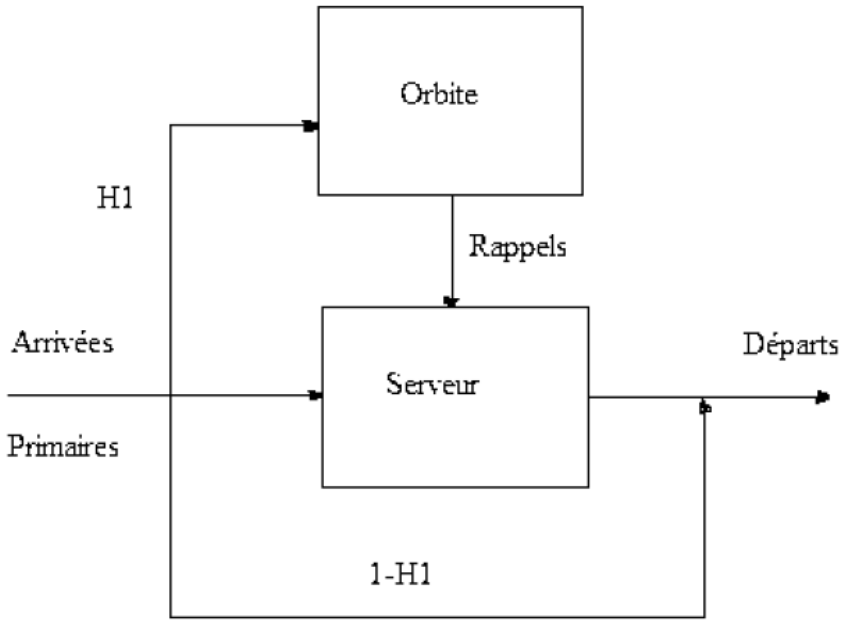


FIGURE 3.2 – Représentation schématique d'un système de files d'attente avec rappels

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

où $N_0(t)$ est le nombre de clients en orbite, $C(t)$ est l'état du serveur à l'instant t. Nous avons $C(t)$ est égale à 0 ou 1 selon le fait que le serveur est libre ou occupé. Si $C(t) = 1, \xi(t)$ représente le temps de service écoulé à la date t.

$$c_k = \lim_{t\to\infty} c_k(t), k \geq 1.$$

Le phénomène de l'impatience est représentée par la fonction de persistance

$$H = H_k, k \geq 1 \quad telle \quad que \quad H_1 < 1 \quad et \quad H_2 = H_3 = \cdots = 1$$

Le processus ci-dessus peut être étudié à l'aide de deux manières, soit en utilisant la méthode de la chaîne de Markov induite, soit avec la méthode des variables supplémetaires.

3.3.2 Chaîne de Markov induite :

Considérons le processus $C(t), N_0(t), t \geq 0$ qui n'est pas en général markovien, mais possède un chaîne de Markov induite. Soit $q_n = N_0(\xi_n)$ le nombre de clients en orbite après le n^{ime} départ. La suite des variables aléatoires $q_n, n \geq 1$ forme une chaîne de Markov induite, dont l'équation fondamentale est

$$q_{n+1} = q_n - \delta_{q_n} + \nu_n. \qquad (3.7)$$

La variable aléatoire ν_{n+1} représente le nombre de clients primaires arrivant dans le système durant le service du $(n+1)^{ime}$ client. Elle ne dépend pas des événements qui se sont produits avant l'instant du début de service du $(n+1)^{ime}$ client.

Théorème 3.3.1. *La distribution du nombre de clients primaires arrivant dans le système durant un service est donnée par*

$$\mathbb{P}(\nu_n = i) = \int_0^\infty \sum_j (\frac{\lambda H_1 x}{j!})^j \exp(-\lambda H_1 x) c_i^{(j)} dB(x)$$

où c_i^j est le j^{ime} produit de convolution de la suite c_i et $a_i > 0, \forall i \geq 0$. Sa fonction génératrice est définie par

$$A(z) = \sum_{n=0}^{\infty} a_i z^n = \widetilde{B}(\lambda H_1 (1 - C(z)))$$

. En outre,

$$\mathbb{E}[\nu_n] = \sum_{i=0}^{\infty} i a_i = \rho$$

.

Lemme 3.3.1. *La fonction $f(z) = A(z) - z$ est décroissante, positive ; et pour $\rho < 1$ et $z \in [0,1]$, on a $z \leq \widetilde{B}(\lambda H_1(1 - C(z))) \leq 1$, où $A(z) = \widetilde{B}(\lambda H_1(1 - C(z)))$.*

3.3.3 Distribution stationnaire de l'état du système :

Nous obtenons la distribution stationnaire du processus $C(t), N_0(t), \xi(t), t \geq 0$ à l'aide de la méthode des variables supplémentaires. Posons $\rho < 1$ et introduisons

$$\begin{aligned} p_{0,n} &= \lim_{t\to\infty} P(C(t) = 0, N_0(t) = n); \\ p_{1,n}(x) &= \lim_{t\to\infty} \frac{d}{dx} P(C(t) = 1, \xi(t) \leq x, N_0(t) = n); \\ p_{1,n} &= \lim_{t\to\infty} P(C(t) = 1, N_0(t) = n) = \int_0^{\infty} p_{1,n}(x)dx. \end{aligned}$$

A partir du graphe de transition (3.3), nous obtenons le système d'équations de balance suivant :

$$(I) \begin{cases} (\lambda + n\theta)p_{0,n} = \displaystyle\int_0^{\infty} p_{1,n}(x)b(x)dx; \\ p'_{1,n}(x) = -(\lambda H_1 + b(x))p_{1,n}(x) + \lambda H_1 \displaystyle\sum_{k=0}^{n} c_k p_{1,n-k}(x) \\ p_{1,n}(0) = \lambda \displaystyle\sum_{k=1}^{n+1} c_k p_{0,n-k+1} + (n+1)\theta p_{0,n+1}. \end{cases}$$

où $b(x) = \frac{B'(x)}{1-B(x)}$ est l'intensité instantanée du service étant donnée que la durée de service écoulé est égale à x. Pour la résolution de ce système on introduit les fonctions génératrices telles que : $P_0(z) = \displaystyle\sum_{n=0}^{\infty} p_{0,n} z^n$ et $P_1(x) = \displaystyle\sum_{n=0}^{\infty} p_{1,n}(x) z^n$. Le système (I) devient

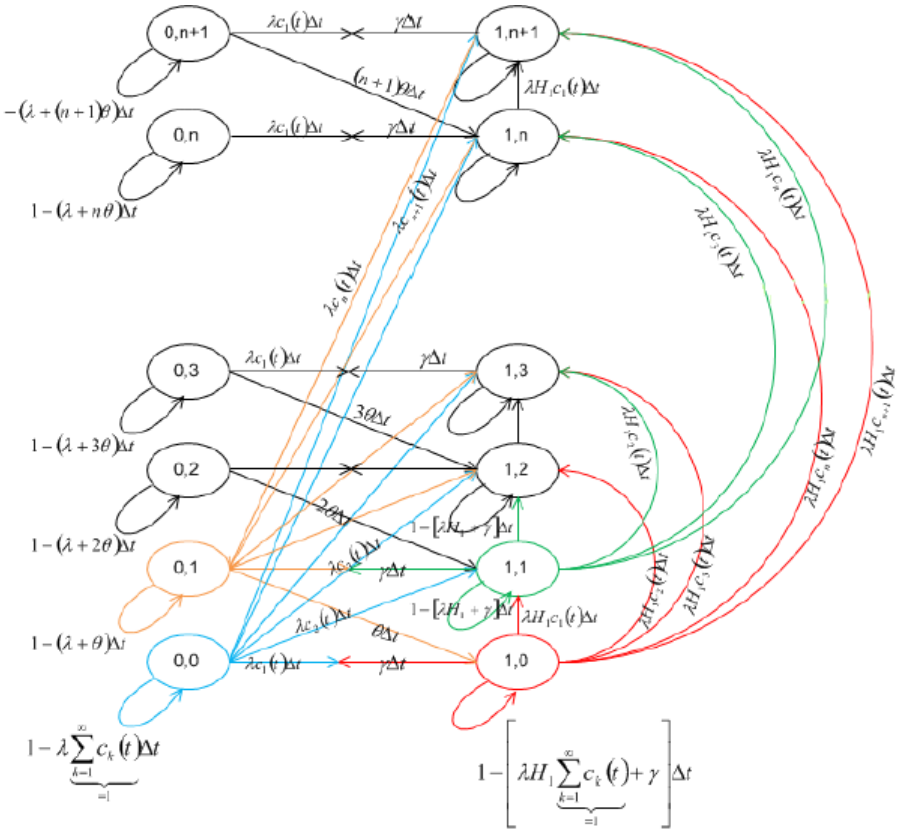


FIGURE 3.3 – Graphe des transitions du modèle $M^X/G/1$ avec rappels et impatience

$$(I)\begin{cases} \lambda\sum_{n=0}^{\infty}p_{0,n}z^n+\theta\sum_{n=0}^{\infty}np_{0,n}z^n=\int_0^{\infty}\sum_{n=0}^{\infty}z^np_{1,n}(x)b(x)dx; \\ \sum_{n=0}^{\infty}z^np'_{1,n}(x)=-(\lambda H_1+b(x))\sum_{n=0}^{\infty}z^np_{1,n}(x)+\lambda H_1\sum_{n=0}^{\infty}z^n\sum_{k=0}^{n}c_kp_{1,n-k}(x) \\ \sum_{n=0}^{\infty}z^np_{1,n}(0)=\lambda\sum_{n=0}^{\infty}z^n\sum_{k=1}^{n+1}c_kp_{0,n-k+1}+\theta\sum_{n=0}^{\infty}z^n(n+1)p_{0,n+1}. \end{cases}$$

$$(I)\begin{cases} \lambda p_0(z)+\theta zp'_0(z)=\int_0^{\infty}p_1(z,x)b(x)dx; \\ p'_1(z,x)=-(\lambda H_1+b(x))p_1(z,x)+\lambda H_1C(x)p_1(z,x) \\ p_1(z,x)=\frac{\lambda}{z}C(z)P_0(z)+\theta P'_0(z) \end{cases} \tag{3.8}$$

De la deuxième équation du système (3.8), on a

$$p_1'(z,x) = -[\lambda H_1(1-C(z)) + b(x)]p_1(z,x)$$

d'où

$$\frac{p_1'(z,x)}{p_1(z,x)} = -\lambda H_1(1-C(z)) - \frac{B'(x)}{1-B(x)}$$

. Alors,

$$p_1(z,x) = p_1(z,0)(1-B(x))\exp[-\lambda H_1(1-C(z))x] \tag{3.9}$$

Donc la première équation du système (3.8) devient

$$\begin{aligned}
\lambda p_0(z) + \theta z p_0'(z) &= \int_0^\infty p_1(z,0)(1-B(x))\exp[-\lambda H_1(1-C(z))x]b(x)dx \\
&= p_1(z,0)\int_0^\infty B'(x)\exp[-\lambda H_1(1-C(z))x]dx \\
&= p_1(z,0)\int_0^\infty \exp[-\lambda H_1(1-C(z))x]dB(x) \\
\lambda p_0(z) + \theta z p_0'(z) &= P_1(z,0)A(z) \qquad (3.10)
\end{aligned}$$

où $A(z) = \widetilde{B}(\lambda H_1(1-C(z)))$. De la troisième équation du système (3.8), l'équation (3.10) donne

$$\begin{aligned}
\lambda p_0(z) + \theta z p_0'(z) &= [\tfrac{\lambda}{z}C(z)P_0(z) + \theta P_0'(z)]A(z) \\
\theta[A(z)-z]P_0'(z) &= \lambda[1 - A(z)\tfrac{C(z)}{z}]P_0(z). \qquad (3.11)
\end{aligned}$$

Théorème 3.3.2. *si $\rho = \lambda\overline{c}H_1\beta_1 < 1$, et le système est en régime stationnaire alors les fonctions génératrices de la distribution jointe de l'état du serveur et de la taille de l'orbite sont données par*

$$\begin{aligned}
P_0(z) &= \sum_{n=0}^{\infty} z^n p_{0n} = \tfrac{H_1(1-\rho)}{\rho + H_1(1-\rho)}\exp[\tfrac{\lambda}{\theta}\int_1^z \frac{1 - A(u)\frac{c(u)}{u}}{A(u)-u}du] \\
P_1(z) &= \sum_{n=0}^{\infty} z^n p_{1n} = \tfrac{1-A(z)}{(A(z)-z)H_1}P_0(z).
\end{aligned}$$

3.3.4 Mesures de performance :

– Nombre moyen de clients dans le système

$$\begin{aligned}
\overline{n} &= \lim_{t\to\infty} E[C(t) + N_0(t)] = Q'(1) \\
&= \rho + \tfrac{\lambda^2 H_1^2 (C'(1))^2 \beta_2 + \rho C"(1)/C'(1)}{2(1-\rho)} + \tfrac{\lambda}{\theta}\tfrac{\rho+\overline{c}-1}{1-\rho};
\end{aligned}$$

– Nombre moyen de clients en orbite

$$\begin{aligned}\overline{n}_0 &= \lim_{t\to\infty} E[N_0(t)] = P'(1) \\ &= \frac{\lambda^2 H_1^2(\overline{c})^2\beta_2+\rho C''(1)/\overline{c}}{2(1-\rho)} + \frac{\lambda}{\theta}\frac{\rho+\overline{c}-1}{1-\rho};\end{aligned}$$

– Temps moyen d'attente d'un client

$$\overline{W} = \frac{\overline{n}_0}{\lambda H_1\overline{c}} = \frac{\lambda H_1\overline{c}\beta_2 + \rho C''(1)/\lambda H_1(\overline{c})^2}{2(1-\rho)} + \frac{1}{\theta H_1\overline{c}}\frac{\rho+\overline{c}-1}{1-\rho}$$

;

– Nombre moyen de rappels par client

$$\overline{R} = \theta\overline{W} = \theta\frac{\lambda H_1\overline{c}\beta_2 + \rho c''(1)/\lambda H_1(\overline{c})^2}{2(1-\rho) + \frac{1}{H_1\overline{c}}\frac{\rho+\overline{c}-1}{1-\rho}}$$

3.3.5 Exemples d'application :

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

Conclusion

Dans ce mémoire, nous nous sommes intéréssés aux Systémes de files d'attente avec rappels. Nous avons traité de types $M/G/1$ et $M^X/G/1$ avec rappels. Pour le premier modèle les arrivées un par un et clients impatients, or le deuxième modèle les arrivées par groupes et clients impatients.

Le premier chapitre, était un rappel sur les processus stochastiques qui sont un outil très puissant pour la modélisation des phénomènes dynamiques.

Le deuxième chapitre, traitent les modèles de files d'attente de type M/G/1, et de type $M^X/G/1$, qui seront utilisées dans la modélisation des Systémes de files d'attente avec des arrivées par groupes.

Dans le chapitre trois, nous présentons une étude le modèle $M^X/G/1$ avec rappels, arrivées par groupes et clients impatients. Nous avons mis en évidence la chaîne de Markov induite associée et donné la distribution du nombre de clients primaires arrivant dans le système durant un service donné. Par la méthode des variables supplémentaires et à l'aide du graphe des transitions nous avons obtenu les fonctions génératrices de la distribution conjointe de l'état du serveur et de la taille de l'orbite, ainsi que la fonction génératrice de la distribution stationnaire du nombre de clients dans le système et la distribution marginale du nombre de serveurs occupés. Enfin, nous avons établi aussi les mesures de performances de ce modèle.

Bibliographie

[1] **Claudie Chabriac**, Processus stochastiques et modélisation, Universite de Toulouse le Mirail, (2012-2013).

[2] **Claudie Hassenforder-Chabriac**, Eléments de théorie des files d'attente, SUPAERO, janvier-(2008).

[3] **Christel Ruwet**, Processus de Poisson, Université de Liège, (2006-2007).

[4] **Nawel Arrar**, Problèmes de convergence, optimisation d'algorithmes et analyse stochastique de Systémes de files d'attente avec rappels, 2.J.(2013).

[5] **Dominique foata**, Processus stochastiques, processus de poisson chaîne de marcove et martingales, DUNOD, (1998).

[6] **E. Lebarbier, S. Robin** , Processus de Poisson Processus de Naissances et Morts, AgroParisTech.

[7] **l'université Rennes**, LE PROCESSUS DE POISSON, Préparation à l'agrégation externe de Mathématiques de l'université Rennes, Année 2008/2009,

[8] **David Coupier**, PROCESSUS STOCHASTIQUES.

[9] **PHAM Cong-Duc**, Cours de Modélisation et d'Evaluation de Performance.