الجمهورية الجزائرية الديمقراطية الشعبية
République algérienne démocratique et populaire
وزارة التعليم العالي والبحث العلمي
Ministère de l'enseignement supérieur et de la recherche scientifique
جامعة عين تموشنت بلحاج بوشعيب
Université –Ain Temouchent- Belhadj Bouchaib
Faculté des Sciences et de Technologie
Département de Mathématiques et Informatique

Projet de Fin d'Etudes
Pour l'obtention du diplôme de Master en : Mathématiques

Domaine : Mathématiques et Informatique
Filière : Mathématiques
Spécialité : Probabilités et Statistique Appliquées

Thème

Analyse d'un réseau de file d'attente sous des disciplines de service avec priorité

Présenté Par :
Mlle Yekhlef Hadjer

Devant le jury composé de :

Dr. MECENE Rahmouna	MCB	UAT.B.B (Ain Temouchent)	Présidente
Dr. BALASKA Lamia	MAA	UAT.B.B (Ain Temouchent)	Examinatrice
Dr. SAKHI Hanane	MCB	UAT.B.B (Ain Temouchent)	Encadrante

Année Universitaire 2024/2025

Résumé

Dans ce travail nous étudions la stabilisation d'un modèle de réseaux de files d'attente sous des disciplines de service avec priorité ; un système de files d'attente fluide multi classe avec priorité composé de N stations, N ≥ 3 et 2N classes (chaque station admet deux classes). Nous nous basons sur le modèle fluide pour l'étude de la stabilité.

Abstract

In this work, we study the stabilization of a queueing network model under service disciplines with priority; a multi-class fluid queueing system with priority consisting of N stations (N ≥ 3) and 2N classes (each station accommodates two classes). We base our stability analysis on the fluid model.

ملخص

في هذا العمل، ندرس استقرار نموذج لشبكات صفوف الانتظار تحت سياسات خدمة ذات الأولوية؛ نظام صفوف انتظار انسيابي متعدد المحطات، بأولوية تتكون من N محطة، $N \geq 3$، و2N فئة (كل محطة تقبل فئتين). نعتمد على نموذج انسيابي لدراسة الاستقرار في هذه الحالة .

Dédicaces

♡ Je dédie ce modeste travail ♡
à ma chèr mère et mon chèr père
qui dieu les protége et leur procure
bonne santé et longue vie.
A toute ma famille, ma soeur
et mon frére, qui m'ont soutenue
tout au long de ce projet,
je vous aimes trés fort
et merçi,

❤

♡

Remerciements

Je remercie tout d'abord Allah de nous avoir donné le courage et la volonté pour mener à bien ce mémoire.

Je tiens à remercier sincèrement mon directrice de recherche, Mme Sakhi Hanane pour sa patience et son soutien qui m'ont été précieux afin de mener ce travail à bon port.

Je remercie également les membres de jury, pour avoir accepté d'évaluer ce travail.

Mes remerciements les plus sincères à mes enseignants, Mathématiques et Informatique.

HADJER

Table des matières

Table des figures

Index des notations

Notation	Description
CMTD	chaine de Markov à temps discret
CMTC	chaine de Markov à temps continu
FIFO	first in, first out
i.e.	c'est à dire
P.G.C.D	Plus Grand Commun Diviseur
M/M/1	Système de files d'attente M/M/1
M/M/c	Système de files d'attente M/M/c
BCMP	Réseau à forme produit

Introduction

Le **formalisme de files d'attente** est la technique la plus largement utilisée pour l'évaluation de performances des systèmes. Ceci s'explique par le fait qu'elle permet d'abstraire le comportement de ces systèmes de façon assez réaliste.

Dans les systèmes, de nombreuses entités partagent les ressources communes. Par exemple, les messages partagent les bus de communication. En général, la ressource utilisée ayant une capacité limitée, toutes les entités ne peuvent donc pas utiliser la ressource en même temps. Ainsi, lorsqu'une première entité accède à la ressource, toutes les autres doivent attendre leur tour en file d'attente, ou alors être rejetées suivant la politique de gestion choisie.

La théorie des files d'attente, permet de représenter les ressources et les mécanismes de gestion assez fidèlement, mais permet également d'obtenir un certain nombre de résultats assez intéressants concernant les performances du système étudié.

L'étude d'un système par la théorie des files d'attente fait appel à la notion de serveur, file d'attente et de clients. Cette terminologie s'adapte quel que soit le domaine concerné.

La théorie des files d'attente est une technique de la recherche opérationnelle qui permet de modéliser un système admettant un phénomène d'attente, de calculer ses performances et de déterminer ses caractéristiques, pour aider les décideurs dans leurs prises de décisions.

Les files d'attente peuvent être considérées comme un phénomène caractéristique de la vie contemporaine. On les rencontre dans les domaines d'activité les plus divers. L'étude mathématique des phénomènes d'attente, constitue un champ d'application important des processus stochastiques.

On parle de phénomène d'attente à chaque fois que certaines unités, appelées (clients), se présentent d'une manière aléatoire à des (stations) afin de recevoir un service dont la durée est généralement aléatoire.

La théorie des files d'attente est un formalisme mathématique qui permet de mener des analyses quantitatives à partir de la donnée des caractéristiques de flux d'arrivées et des temps de service.

la théorie traditionnelle se fond sur les Chaînes de Markov, c'est-à-dire que tous les événements tels que les arrivées, les fins de service, les changements de files d'attente, etc, dépendent uniquement de l'état actuel du système et non de son comportement

antérieur. Cela simplifie non seulement le traitement mathématique, mais aussi la collecte de données puisque seules les moyennes sont requises, par exemple une moyenne des délais de service, une moyenne des arrivées par unité de temps et ainsi de suite.

Les réseaux de files d'attente ont été intégrés dans la modélisation de divers domaines d'activité dans les années 1960. C'est l'échange de données entre ordinateurs qui a permis une évolution rapide de la théorie des réseaux de file d'attente, notamment Les problèmes concernent les temps de traitements de requêtes sur un système centralisé ou encore les délais, le taux d'occupation des différents nœuds du réseau. L'apparition de l'internet et son dévloppement entre les années $70 - 90$ ont poussé les mathématiciens à essayer de modéliser ces réseau.

L'objectif de notre travail est de comprendre, modéliser et optimiser le fonctionnement du système, en tenant compte des différences de traitement entre les classes de clients. Ce mémoire se compose de trois chapitres organisés de la manière suivante :

Dans le chapitre 1, nous introduisons la terminologie de la théorie des files d'attente et certaines définitions et notations qui sont nécessaires dans l'étude des systèmes de files d'attente. Nous donnons d'abord une introduction aux concepts de base de la théorie des chaines de Markov en temps discret et en temps continu. Nous présentons également les propriétés fondamentales des chaines de Markov. Nous introduisons la terminologie de la théorie des files d'attente et de certaines définitions et notations qui sont nécessaires dans l'étude des systèmes de files d'attente. En suite nous étudions quelque modèles de files d'attente markoviennes (M/M/1, M/M/c).

Dans le chapitre 2, qui a été inspiré par le document de Philippe Robert, nous introduisons un ensemble de travaux relatifs à l'étude des réseaux stochastiques. Les réseaux de files d'attente à forme produit seront données (le réseaux de Jackson, le Réseaux multi classe). Enfin les limites fluides notamment seront données, qui sont les limites des processus renormalisées.

Dans le chapitre 3, nous concentrons sur stabilisation d'un autre modèle de réseau de file d'attente sous des disciplines de service avec priorité ; un système de files d'attente fluide multi classe avec priorité composé de N stations, $N \geq 3$ et $2N$ classes (chaque station admet deux classes). Nous nous basons sur le modèle fluide pour l'étude de la stabilité.

Chapitre 1

Chaîne de Markov et file d'attente.

Dans ce chapitre, nous considérons deux exemples de processus stochastiques à espace d'état discret, connus sous le nom de chaines de Markov, le premier à temps discrets, le deuxième à temps continu. Ces deux modéles fouiront des outils simples de modélisation et d'analyse d'une classe particulière de systèmes à événements discrets. Par ailleurs, L'analyse des Chaînes de Markov est un préliminaire nécessaire à l'étude des systèmes de Files d'attente.

Les notions présentées dans ce chapitre sont tirées des références : [1] [3] [12] [15] [18] [24] [26] [27].

1.1 Chaîne de Markov

Les chaînes de Markov fournissent des moyens très puissants et efficaces pour la description et l'analyse des propriétés des systèmes dynamiques (files d'attente, réseaux informatiques et téléphoniques, physique, biologiques ou économiques etc). Leurs utilisations pour la modélisation des phénomènes aléatoires évoluant dans le temps, donnent aux chaînes de Markov une place importante. La structure simple des chaînes de Markov permet de dire beaucoup de choses sur le comportement et l'évolution de ses phénomènes. En effet, l'ensemble des études mathématiques des processus aléatoires peut être considéré comme une généralisation d'une manière ou d'une autre de la théorie des chaînes de Markov.

Définition 1.1.1 *Soit* $(X_n)n \in \mathbf{N}$ *une suite de variables aléatoires prenant leurs valeurs dans un espace dénombrable* $\mathbf{E}$ *(dans notre cas, nous prendrons N ou un sous-ensemble de* $\mathbf{N}$*) .* $(X_n)n \in \mathbf{N}$ *est une chaîne de Markov si et seulement si*

$$P(X_{n+1} = j \mid X_0 = i_0, \ldots, X_{n-1} = i_{n-1}, X_n = i) = P(X_{n+1} = j \mid X_n = i).$$

Ainsi, pour une chaîne de Markov, la probabilité d'un comportement futur particulier du processus, lorsque son présent actuel est bien connu, n'est pas modifié par toute connaissance supplémentaire du passé. Le processus est dit sans mémoire ou sans usure.
Soit la notation :

$$p_{i,j} = P(X_n = j \mid X_{n-1} = i).$$

p_{ij} *est appelé la probabilité de transition de l'état i vers l'état j. Lorsque cette probabilité ne dépend pas de n, nous disons que la chaîne de Markov est homogène*

1.1.1 Les Chaînes de Markov à temps discret (CMTD)

On considère un processus stochastique $\{X_n\}_{n\in\mathbf{N}}$ à espace d'état discret et à temps discret. E est l'espace d'état. Il peut être de dimension finie ou infinie (mais dénombrable cas discret).

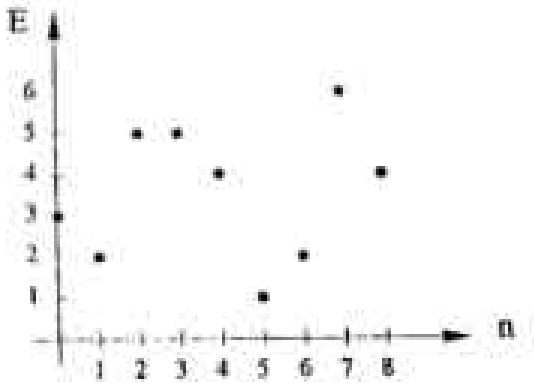

FIGURE 1.1 – Processus stochastique à espace d'état discret et à temps discret.

Définition 1.1.2 $\{X_n\}_{n\in\mathbf{N}}$ *est une chaine de Markov à temps discret ssi*

$$P\left[Xn = j | X_{n-1} = i_{n-1}, X_{n-2} = i_{n-2}....., X_0 = i_0\right] = P\left[X_n = j | X_{n-1} = i_{n-1}\right].$$

la probabilité pour que la chaine soit dans un certain état à la $n^{ième} \ll$ *étape* $\gg$ *du processus ne dépend donc que de l'état du processus à l'étape précédente (la* $n-1^{ième}$ *étape) et pas des états dans lesquels il se trouvait aux étapes antérieures (les étape j . Pour tout* $j = 0, ..., n-2$*).*

1.1.2 États périodiques

Définition 1.1.3 *La périodicité. La période d(i) d'un état i est définie par :*

$$d(i) = P.G.C.D.\{n \in \mathbf{N}^*; p_{i,i}(n) > 0\}.$$

(Avec $d(i) = 0$ si pour tout $n \in \mathbf{N}^*, p_{i,i}(n) = 0$).
Si $d(i) = 1$, i est dit apériodique.

Théorème 1.1.1 *La période est une propriété de classe. Si les états i et j communiquent entre eux alors ils ont la même période.*

Démonstration

Comme i et j communiquent, alors il existe des entiers N et M tel que $p_{i,j}(M) > 0$ et $p_{j,i}(N) > 0$. Pour tout $k > 1$.

$$p_{i,i}(M + nk + N) \geq p_{i,j}(M)(p_{j,j}(k))^n p_{j,i}(N).$$

(En effet, le chemin $X_0 = i; \quad X_M = j; \quad X_{M+k} = j; \quad X_{M+nk} = j; \quad X_{M+nk+N} = i$ est une possibilité d'aller de i à i dans $M + nk + N$ étapes).

Par conséquent, pour tout $k > 1$ tel que $p_{j,j}(k) > 0$; on a $(M + nk + N) > 0$ pour tous $n > 1$.

Et par conséquent, d_i divise $M + nk + N$ pour tout $n > 1$, et en particulier, d_i divise k. Nous avons donc démontré que d_i divise tous les k de telle sorte que $p_{j,j}(k) > 0$, et en particulier , d_i divise d_j. Par symétrie, d_j divise d_i , et donc, finalement, $d_i = d_j$.

Théorème 1.1.2 [23] *Soit P une matrice stochastique irréductible avec la période d, alors pour tous les états i, j il existent $m > 0$ et $n_0 > 0$ (m et n peuvent être dépendant de i, j) de telle sorte que*

$$p_{i,j}(m + nd) > 0 \quad \forall n \geq 0.$$

Démonstration

Il suffit de démontrer pour $i = j$. En effet, il existe m tel que $p_{i,j}(m) > 0$, parce que j est accessible à partir de i, la chaîne étant irréductible, et donc, si pour une $n_0 > 0$ nous avons $p_{j,j}(nd) > 0$ pour tout $n > n_o$, alors $p_{i,j}(m + nd) > p_{i,j}(m)p_{j,j}(nd) > 0$ pour tout $n > n_0$. Le PGCD de l'ensemble $A = k \geq 1; p_{j,j}(k) > 0$ est d et A est fermé pour addition. L'ensemble A contient donc tout sauf un nombre fini des multiples positifs de d. En d'autres termes, il existe n_0 tel que $n > n_0$ implique $p_{jj}(nd) > 0$.

Propriété 1.1.3 *Tous les états d'une même classe de communication ont la même période.*

Démonstration

Si x et x' communiquent, il existe k et l tels que $p_{xx'}(k) > 0$ et $p_{x'x}(l) > 0$. Alors $p_{xx}(k + l) > 0$. Soit m tel que $p_{x'x'}(m) > 0$. Alors $p_{xx}(k + m + l) > 0$. Donc $d(x)$ divise $k+l$ et $d(x)$ divise $k+m+l$; donc $d(x)$ divise m et ceci, pour tout m tel que $p_{x'x'}(m) > 0$, donc $d(x)$ divise $d(x)$. Comme x et x' jouent le même rôle, alors $d(x) = d(x')$.

Remarque 1.1.4 *Une classe dont un élément admet une boucle, (c'est-à-dire $p_{x,x} > 0$), est obligatoirement apériodique (c'est-à-dire de période 1), mais ce n'est pas une condition nécessaire.*

Théorème 1.1.5 *Si x et y sont dans une même classe de communication de période d, si $p_{x,y}(n) > 0$ et si $p_{x,y}(m) > 0$, alors d divise $m - n$.*

Démonstration

Comme y mène à x, il existe k tel que $p_{y,x}(k) > 0$. On a donc $p_{x,x}(m+k) > 0$ et $p_{x,x}(n+k) > 0$. Donc d divise $m+k$ et $n+k$, il divise donc la différence. Ainsi, on peut partitionner les classes périodiques de la façon suivante.

Théorème 1.1.6 *Soit X une chaîne de Markov homogène dans un espace d'états $\mathbf{E}$, et soit $\mathbf{E}' \subseteq \mathbf{E}$ une classe irréductible non vide . Alors pour tout $i, j \in \mathbf{E}'$, les périodes de i et j sont les même, i.e., $d(i) = d(j)$.*

1.1.3 Etats récurrents et états transients

Définition 1.1.4 *Si au temps 0 , la chaîne est en l'état e_i , la variable aléatoire $T_i = \min\{n \geq 1 \ ; \ X_n = e_i\}$ définit le temps de premier retour en e_i .*

Définition 1.1.5 *Un état e_i est récurrent si partant de e_i on y revient sûrement , ce qui s'exprime par :*

$$\mathbf{P}\left((\quad \textit{il existe} \quad n \geq 1 \quad \textit{tel que} \quad X_n = e_i) | X_0 = e_i\right) = 1,$$

soit encore :

$$\mathbf{P}\left(T_i < +\infty \,|\, X_0 = e_i\right) = 1.$$

Un état récurrent est donc visité un nombre infini de fois.

Définition 1.1.6 *Un état e_i est dit transient s'il n'est pas récurrent :*

$$\mathbf{P}\left(T_i < +\infty \,|\, X_0 = e_i\right) < 1.$$

Un état est transient s'il n'est visité qu'un nombre fini de fois.
Par définition, une chaîne irréductible ne contient aucun état transient ou absorbant.

Définition 1.1.7 *Une chaîne dont tous les états sont récurrents est dite récurrente, une chaîne dont tous les états sont transients est dite transiente.*

Définition 1.1.8 *La probabilité de première transition en n unités de temps, de l'état e_i à l'état e_j , notée $f_{i,j}^{(n)}$ est définie par :*

$$f_{i,j}^{(n)} = \mathbb{P}\left(T_j = n \,|\, X_0 = e_i\right) = \mathbb{P}\left(X_n = e_j, X_{n-1} \neq e_j, \ldots, X_1 \neq e_j \,|\, X_0 = e_i\right).$$

$f_{i,j}^{(n)}$ est la probabilité de premier retour à l'état e_i en n unités de temps .

Théorème 1.1.7 *Pour e_i un état d'une chaîne de Markov, on a :*

$$e_i \text{ est récurrent } \Leftrightarrow \sum_{n=1}^{+\infty} f_{i,i}^{(n)} = 1,$$

$$e_i \text{ est transitoire } \Leftrightarrow \sum_{n=1}^{\infty} f_{ii}^{(n)} < 1.$$

Définition 1.1.9 *On appelle temps moyen de retour en e_j :*

$$\mu_j = \mathbb{E}[T_j = n \mid X_0 = e_j] = \sum_n n\, P_{j,j}^{(n)}.$$

Il existe deux sortes d'états récurrents
- *Les états récurrents positifs.*
- *Les états récurrents nuls.*

Définition 1.1.10 *Un état e_j est récurrent positif (ou non nul) s'il est récurrent et si le temps moyen μ_j de retour est fini . Dans le cas où le temps moyen de retour est infini, l'état est dit récurrent nul.*

Théorème 1.1.8 *(Caractérisation des états)*
Pour e_i un état d'une chaîne de Markov, on a :

$$(e_i \text{ transient}) \Leftrightarrow \sum_{n\geq 1} P_{i,i}^{(n)} < +\infty,$$

$$(e_i \text{ récurrent nul}) \Leftrightarrow \sum_{n\geq 1} P_{i,i}^{(n)} = +\infty \quad et \quad \lim_{n\to+\infty} P_{i,i}^{(n)} = 0,$$

$$(e_i \text{ récurrent positif}) \Leftrightarrow \sum_{n\geq 1} P_{i,i}^{(n)} = +\infty \quad et \quad \lim_{n\to+\infty} P_{i,i}^{(n)} > 0.$$

Théorème 1.1.9 *Soit X_n un chaîne de Markov à espace d'états fini.*
1. *Aucun état n'est récurrent nul.*
2. *Aucune chaîne finie n'est transiente ; en revanche, certains de leurs états peuvent être transients.*
3. *Il existe au moins un état récurrent positif.*
4. *Si en outre la chaîne est irréductible, elle est récurrente positive.*

Démonstration

* (1) Supposons qu'il existe un état récurrent nul e_j , et C_j sa classe de communication, alors

$$\forall n \in \mathbf{N}; \sum_{k \in C_j} P_{i,i}^{(k)} = 1. \tag{1.1}$$

  Mais $\lim_{n \to +\infty} p_{j,k}^{(n)} = 0$, qui est en contradiction avec (1.1) .
* (2) Supposons tous les états transients, on montre alors que $\forall e_i, e_j$ transients , la série $\sum p_{j,k}^{(n)}$ converge, donc

$$\lim_{n \to +\infty} p_{j,k}^{(n)} = 0. \tag{1.2}$$

  Or,

$$\sum_{j=1}^{card(E)} p_{i,j}^{(n)} = 1 (\forall n \in \mathbf{N}).$$

  Donc par passage à la limite quand $n \to +\infty$, on arrive à une contradiction de (1.2).
* (3)Est conséquence de (1) et (2).
* (4) La chaîne ayant au moins un état récurrent d'après (c) et une seule classe est récurrente positive.

Définition 1.1.11 *Un état récurrent positif et apériodique est dit ergodique. Une chaîne irréductible, apériodique et récurrente positive est dite ergodique.*

Théorème 1.1.10 *Les états d'une classe de communication qui contient au moins un état ergodique sont ergodiques.*

Preuve

On sait que les propriétés de récurrence et d'apériodicité sont des propriétés de classe : l'ergodicité l'est donc aussi. Il suffit donc qu'un état de la classe soit ergodique pour que la classe qui le contient le soit.

Théorème 1.1.11 *(théorème ergodique)*
Soit $(X_n) n \in \mathbf{N}$ une chaîne de Markov irréductible récurrente positive , de probabilité invariante π, soit $f : \mathbf{E} \to \mathbf{R}$ une fonction bornée . Alors

$$\frac{1}{n} \sum_{k=1}^{n} f(X_k) \xrightarrow[n \to +\infty]{p.s.} \sum_{k=1}^{n} \pi_k f(x). \tag{1.3}$$

Une chaîne de Markov à temps continu peut souvent être décrite par une chaîne à temps discret sur les sauts entre états, avec des temps d'attente exponentiels entre ces sauts.

1.1.4 Les Chaînes de Markov à temps continu (CMTC)

Dans cette section, nous introduisons la définition et les principaux Théorèmes pour les processus continus. Un processus de Markov continu est un processus aléatoire $(X_t)_{t \in \mathbf{T}}$ indexé par un espace continu. Dans nôtre cas , il sera indexé par $\mathbf{R}^+$.

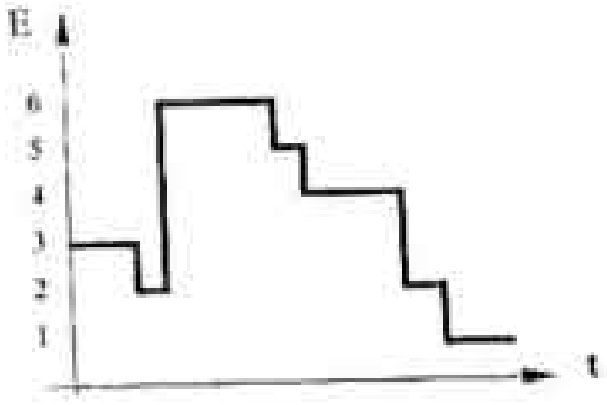

FIGURE 1.2 – Processus stochastique à espace d'état discret et à temps continu.

Définition 1.1.12 *$\{X(t)\}_{t\geqslant 0}$ est une chaine de Markov à temps continu ssi :*

$$P[X(t_n) = j | X(t_{n-1}) = i_{n-1}, X(t_{n-2}) = i_{n-2}, ..., X(t_0) = i_0]$$

$$= P[X(t_n) = j | X(t_{n-1}) = i_{n-1}].$$

$\forall n$ et $\forall t_0 < t_1 < .... < t_n$.

Définition 1.1.13 ***Matrice de transition***
La matrice de transition $P(t)$ dont le terme générale $p_{i,j}(t) = \mathbb{P}(X_t = j \mid X_0 = i)$ est une matrice stochastique qui vérifie :

$$\forall i, j \in E \quad p_{i,j}(t) \geq 0,$$
$$et \quad \forall i \in E \quad \sum_{j\in E} p_{i,j}(t) = 1.$$

Théorème 1.1.12 *Les probabilités de transition d'une chaîne de Markov à temps continu satisfont les équations suivantes dites de* ***Chapman-Kolmogorov*** *:*

$$\forall i, j \in E, \ \forall s, t \geq 0, \quad on\ a \quad p_{i,j}(t+s) = \sum_{k\in E} p_{i,k}(t) p_{k,j}(s).$$

Proposition 1.1.13 *Nous associons à la chaîne de Markov à temps continu (CMTC) la matrice Q, dite* ***générateur infinitésimal****, définie par :*

$$\begin{cases} \forall i \neq j, \quad q_{i,j} = \mu_{ij} = \mu_i P_{i,j}, \\ \forall i, \quad q_{i,i} = -\sum_{j\neq i} \mu_{ij} = -\sum_{j\neq i} \mu_i P_{i,j} = -\mu_i. \end{cases}$$

Ainsi,

$$Q = \begin{pmatrix} -\sum_{j\neq 1} \mu_{1j} & \mu_{12} & \mu_{13} & \cdots & \mu_{1n} \\ \mu_{21} & -\sum_{j\neq 2} \mu_{2j} & \mu_{23} & \cdots & \mu_{2n} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mu_{n1} & \mu_{n2} & \mu_{n3} & \cdots & -\sum_{j\neq n} \mu_{nj} \end{pmatrix}$$

avec $\sum_j q_{i,j} = 0$.

Corollaire 1.1.14 *La matrice de transition $P(t)$ d'une CMTC admet la représentation suivante :*

$$P(t) = e^{tQ} = \sum_{n=0}^{+\infty} \frac{t^n}{n!} Q^n.$$

Q Générateur infinitésimal associé à CMTC.

Propriété 1.1.15 *Le temps passé dans un état d'une CMTC a une distribution exponentielle.*

Preuve

On s'intéresse à deux instants d'observation très proches, t et t+dt. On peut affimer, d'après la définition 1.2.1, que la probabilité pour que le processus ait quitté l'état i au temps t+dt, sachant qu'il y était au temps t, soit $P[X(t+dt) \neq i | X(t) = i]$, est indépendante de tout ce qui a pu se produire avant l'instant t. Il est clair, dans ces conditions, que le temps passé dans l'état i est ≪ sans mémoire ≫ (car la probabilité pour le quitter en dt ne dépend pas du temps qu'on y a déjà passé), donc exponentiel.

Les chaînes de Markov offrent un cadre mathématique puissant pour analyser les files d'attente et peuvent fournir des prédictions précises sur le comportement du système sous différentes conditions de charge et de configuration.

1.2 File d'attente

La théorie de file d'attente fournit un outil très puissant et efficace pour la modélisation des systèmes admettant un phénomène d'attente. Cette théorie datent du début du XXème siècle par les travaux de l'ingénieur danois Agner Krarup Erlang (1878, 1929).

Les modèles Markoviens sont des systèmes où les temps entre deux arrivées successives et les durées de service sont des variables aléatoires indépendantes et exponentiellement distribuées. On s'intéresse au nombre $N(t)$ de clients se trouvant dans le système à l'instant t. On introduit donc le processus stochastique

$$\{N(t), t \geq 0\}. \tag{1.4}$$

Les modèles Markoviens ont en commun que le processus (1.4) constitue une chaîne de Markov à temps continu homogène, ce qui signifie que le comportement futur ne dépend que de l'état présent et non pas de l'évolution dans le passé :

$$\Pr(N(t) = j / N(t_n) = i_n, N(t_{n-1}) = i_{n-1}, \ldots, N(t_1) = i_1) = \Pr(N(t) = j / N(t_n) = i_n),$$

où $t_1 < t_2 < .... < t_n < t$. Supposons que la discipline d'attente est FIFO (voir le tableau de l'index de notation), alors (1.4) est un processus de naissance et de mort, qui est une chaîne de Markov à temps continu et à espace état discret.

Définition 1.2.1 *Une file d'attente est composé d'un certain nombre de places d'attente, d'un ou plusieurs serveurs et de clients qui arrivent, attendent, se font servir selon des règles de priorité, puis quittent le système.*

Description d'une file simple

Une file d'attente est un système caractérisé par un espace d'attente qui contient une ou plusieurs places, et une espace de service composé d'un ou plusieurs serveurs. Les clients arrivent de l'extérieur à des instants aléatoires, ils attendent que l'un des serveurs soit libre pour pouvoir être servi puis quittent le système. Les clients peuvent être des individus, des appels téléphoniques, des signaux électriques, des véhicules, des accidents, . . . etc.

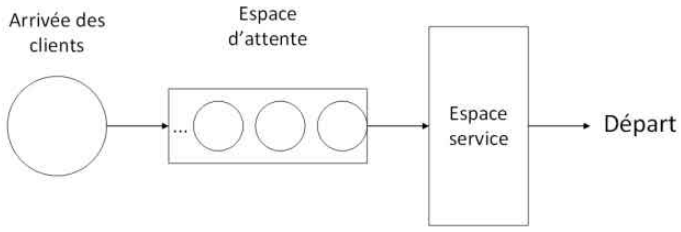

FIGURE 1.3 – Système de file d'attente simple.

Une file d'attente réelle peut être modélisée par un système M/M/1 si ses caractéristiques correspondent aux hypothèses du modèle.

1.2.1 Système de files d'attente M/M/1

1) Description du modèle :

La file d'attente M/M/1 est un modèle caractérisé par des arrivées suivants un processus de Poisson de taux λ, temps de service exponentiel de paramètre μ et un seul serveur. Les clients arrivent à la station selon un processus de Poisson de taux λ, si le serveur est vide le client est pris en charge immédiatement sinon il rejoint la file d'attente (de capacité illimitée et discipline FIFO), les temps des interarrivées sont indépendants.

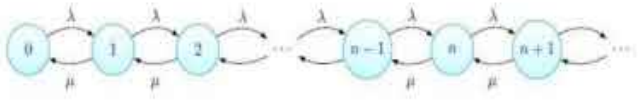

FIGURE 1.4 – Graphe de transitions du modèle M/M/1.

2) Analye du modèle :

L'état du système à la date t peut être décrit par le processus stochastique (1.4).Grâce aux propriété fondamentales du processus de Poisson et de la loi exponentielle, on a pour

un petit intervalle du temps Δt les probabilités suivantes :

$$\begin{cases} \Pr(\text{exactement une arrivée pendant } \Delta t) = \lambda\Delta t + o(\Delta t) \\ \Pr(\text{ aucune arrivée pendant } \Delta t) = 1\text{-}\lambda\Delta t + o(\Delta t) \\ \Pr(\text{ deux arrivées ou plus pendant } \Delta t) = o(\Delta t) \\ \Pr(\text{exatement un départ pendant } \Delta t / N(t) > 0) = \mu\Delta t + o(\Delta t) \\ \Pr(\text{aucun départ pendant } \Delta t / N(t) > 0 = 1\text{-}\mu\Delta t + o(\Delta t) \\ \Pr(\text{deux départs ou plus pendant } \Delta t) = o(\Delta t) \end{cases}$$

Ces probabilités ne dépendent ni de temps ni de l'état t dans lequel le système se trouve.

Soient $p_{i,j}(\Delta t) = Pr(N(t+\Delta t) = j / N(t) = i), i = 0, 1, 2, ...$ Ces probabilités de transition ne dépendent pas de l'instant t.

On suppose que les arrivées et les départ sont mutuellement indépendants.

Régime transitoire :

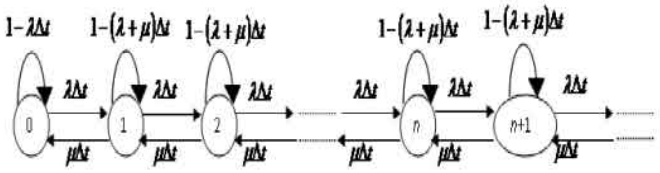

FIGURE 1.5 – Graphe de transitions de la file M/M/1.

Soit $p_n(t) = Pr(N(t) = n), n = 0, 1, 2, ...$ Le graphe des transitions est la figure (1.5) A partir du graphe des transitions, on obtient

$$\begin{cases} p_0(t+\Delta t) = \mu\Delta t p_1(t) + [1 - \lambda\Delta t]p_0(t), \\ p_n(t+\Delta t) = \mu\Delta t p_{n+1}(t) + \lambda\Delta t p_{n-1}(t) + [1 - (\lambda+\mu)\Delta t]p_n(t), n \geqslant 1. \end{cases}$$

Puis, les équations de Kolmogorov :

$$\begin{cases} p_0'(t) = -\lambda p_0(t) + \mu p_1(t), \\ p_n'(t) = -(\lambda+\mu)p_n(t) + \lambda p_{n-1}(t) + \mu p_{n+1}(t), n \geqslant 1. \end{cases} \tag{1.5}$$

Ces équations permettent, en principe, de calculer les probabilités d'état $p_n(t)$, si l'on connaît en plus les conditions initiales du processus, c'est-à-dire la distribution de $N(0)$.

Régime stationnaire :

Il est démontré que $\lim_{t\to\infty} p_n(t) = p_n, n \geqslant 0$, existent et sont indépendantes de l'état initial du processus (1.4) ; et $\lim_{t\to\infty} p_n'(t) = p_n, n \geqslant 0$.

De (1.5), on obtient le système d'équations de balance suivant :

$$\begin{cases} \mu p_1 = \lambda p_0 \\ \lambda p_{n-1} + \mu p_{n+1} = (\lambda+\mu)p_n, n \geqslant 1 \end{cases} \tag{1.6}$$

La résolution du système (1.6) (la résolution du modèle) s'effectue de la manière suivante :

$$\begin{cases} p_1 = \dfrac{\lambda}{\mu} p_0 \\ \text{pour } n = 1, \quad \lambda p_0 + \mu p_2 = (\lambda + \mu) p_1, \quad p_2 = \left(\dfrac{\lambda}{\mu}\right)^2 p_0 \\ \text{pour } n > 1, \quad p_n = \left(\dfrac{\lambda}{\mu}\right)^n p_0 \end{cases}$$

Pour trouver la probabilité p_0, on utilise l'équation de normalisation. En effet,

$$p_0 + \frac{\lambda}{\mu} p_0 + \left(\frac{\lambda}{\mu}\right)^2 p_0 + ... = 1.$$

$$p_0 = \frac{1}{1 + \frac{\lambda}{\mu} + \left(\frac{\lambda}{\mu}\right)^2 + ...}.$$

où $1 + \frac{\lambda}{\mu} + \left(\frac{\lambda}{\mu}\right)^2 + ...$ est une progression géométrique de raison $\frac{\lambda}{\mu}$. Elle converge si $\frac{\lambda}{\mu} < 1$,et elle est égale à $\frac{1}{1-\frac{\lambda}{\mu}}$.

Alors, $p_0 = 1 - \frac{\lambda}{\mu}$. D'où

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

$\rho = \frac{\lambda}{\mu}$ est l'intensité du trafic. $\rho = \frac{\lambda}{\mu} < 1$ est la condition d'existence du régime stationnaire.

Encore, $p_n = (1-\rho)\rho^n, n \geqslant 0$, est la distribution stationnaire du nombre de clients dans le système M/M/1.

Remarque 1.2.1

1. *Si $\lambda \geqslant \mu$, on a $\lim_{t\to\infty} p_n(t) = 0, n \geqslant 0$. Ceci signifie que la longueur de la file d'attente dépasse toute limite.*

2. *En ce qui concerne le processus de sortie du système M/M/1, il est (en régime stationnaire) évident que le taux de sortie est égal au taux d'arrivée λ. Il est démontré que le processus de sortie d'un système M/M/1 est à nouveau de type poissonien.*

3) Caractéristiques du système M/M/1 (Mesures de performance)

Soit $N = \lim_{t\to\infty} N(t)$.

a) Le nombre moyen de clients dans le système :

$$\overline{n} = E[N] = \sum_{n=0}^{\infty} n p_n = (1-\rho)\sum_{n=0}^{\infty} n\rho^n = (1-\rho)\rho[1 + 2\rho + 3\rho^2 + ...] = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu - \lambda}.$$

b) Le nombre moyen de clients dans la file d'attente :

Soit $N_f = \lim_{t\to\infty} N_f(t)$ où $N_f(t)$ est le nombre de clients dans la file d'attente à la date t. La variable N_f est définie de la manière suivante :

$$N_f = \begin{cases} 0 & si \quad N = 0 \\ N-1 & si \quad N \geq 1 \end{cases}$$

et

$$\overline{n}_f = \sum_{n=1}^{\infty}(n-1)p_n = \frac{\lambda^2}{\mu(\mu-\lambda)}.$$

ou bien

$$\overline{n}_f = \overline{n} - \rho.$$

c) Le temps moyen d'attente et de séjour d'un client dans le système :

Le temps moyen d'attente $\overline{W}$ et le temps moyen de séjour $\overline{W}_s$ peuvent être calculés à l'aide de formule de Little, en effet,

$$\overline{W} = \frac{\overline{n}_f}{\lambda} = \frac{\lambda}{\mu(\mu-\lambda)}.$$

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

Le M/M/1 est un cas particulier du M/M/c avec $c = 1$. Beaucoup de formules du M/M/c deviennent celles du M/M/1 lorsque $c = 1$.

1.2.2 Système de files d'attente M/M/c

1) Description du modèle :

Les clients arrivent vers le système selon un processus de Poisson de taux $\lambda > 0$. Le service est assuré par $c \geq 1$ serveurs montés en parallèle. A l'arrivée d'un client, si l'un des serveurs est libre, le client commence immédiatement son service. Dans le cas contraire (tous les serveurs sont occupés par le service), le client prend place dans la file d'attente, commune pour tous les serveurs. La capacité d'attente est illimitée (le nombre de positions d'attente est infini). Lorsqu'un serveur se libère, le client en tête de la file d'attente occupe le serveur libéré. Par conséquent, la discipline d'attente est FIFO. Les temps de service sont exponentiellement distribués de moyenne finie $\frac{1}{\mu}$. Les durées entre deux arrivées consécutives et les durées de service sont mutuellement indépendantes.

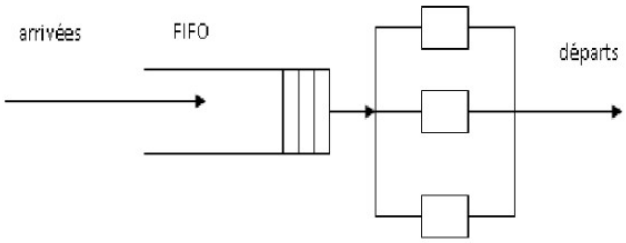

FIGURE 1.6 – Modèle d'attente M/M/c.

2) Analyse du modèle :

L'état du système à la date t peut être décrit à l'aide du processus (1.4), dont l'espace des états est $S = \{0, 1, 2, ...\}$. Ce dernier est un processus de naissance et de mort dont les taux de transition sont :

$$\lambda_n = \lambda, \quad n \geq 0$$
$$\mu_n = \min\{n, c\}\mu, \quad n \geq 1$$

Régime transitoire :

Le graphe des transitions est donné dans la figure (1.5).

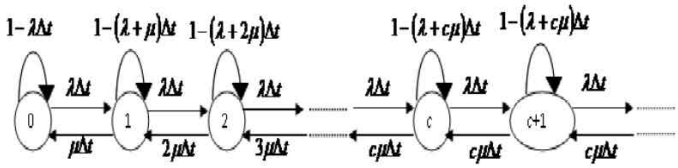

FIGURE 1.7 – Graphe de transitions de la file M/M/c.

Le système d'équations pour les probabilités d'état $p_n(t) = Pr(N(t) = n), n \geq 0$ est

$$\begin{cases} p_0(t + \Delta t) = (1 - \lambda\Delta t)p_0(t) + \mu\Delta t p_1(t). \\ p_n(t + \Delta t) = \lambda\Delta t p_{n-1}(t) + [1 - (\lambda + n\mu)\Delta t]p_n(t) + (n + 1)\mu\Delta t p_{n+1}(n), 1 \leq n < c. \\ p_n(t + \Delta t) = \lambda\Delta t p_{n-1}(t) + [1 - (\lambda + c\mu)\Delta t]p_n(t) + c\mu\Delta t p_{n+1}(n), n \geq c. \end{cases}$$

Le système d'équations de Kolmogorov se présente de la manière suivante :

$$\begin{cases} p_0'(t) = -\lambda p_0(t) + \mu p_1(t). \\ p_n'(t) = \lambda p_{n-1}(t) - (\lambda + n\mu)p_n(t) + (n + 1)\mu p_{n+1}(n), 1 \leq n < c. \\ p_n'(t) = \lambda p_{n-1}(t) - (\lambda + c\mu)p_n(t) + c\mu p_{n+1}(n), n \geq c. \end{cases}$$

Régime stationnaire :

Soit $p_n = \lim_{t\to\infty} p_n(t), n \geq 0$. Cette distribution stationnaire satisfait les équations de balance

$$\begin{cases} 0 = -\lambda p_0 + \mu p_1. \\ 0 = \lambda p_{n-1} - (\lambda + n\mu)p_n + (n + 1)\mu p_{n+1}, 1 \leq n < c. \\ 0 = \lambda p_{n-1} - (\lambda + c\mu)p_n + c\mu p_{n+1}, n \geq c. \end{cases}$$

La résolution du système d'équations ci-dessus nous donne

$$\begin{cases} p_n = \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n p_0, 1 \leq n \leq c. \\ p_n = \frac{1}{c!}\frac{1}{c^{n-c}}\left(\frac{\lambda}{\mu}\right)^n p_0, n \geq c. \end{cases}$$

On remarque que pour $n = c$, les deux formules donnent la même valeur. Pour calculer la probabilité pour que le système est vide p_0, on applique l'équation de normalisation $\sum_{n=0}^{\infty} p_n = 1$. En effet,

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

La deuxième somme peut être réécrite de la manière suivante

$$\frac{1}{c!}\left(\frac{\lambda}{\mu}\right)^c \left[1 + \frac{\lambda}{c\mu} + \left(\frac{\lambda}{cu}\right)^2 + \left(\frac{\lambda}{c\mu}\right)^3 + ...\right].$$

La somme [...] possède une limite égale à $\frac{1}{1-\frac{\lambda}{c\mu}}$ si $\frac{\lambda}{c\mu} < 1$. Par conséquent, le système considéré est en régime stationnaire si , $\rho = \frac{\lambda}{c\mu} < 1$, ρ est l'intensité globale du trafic. On obtient ainsi

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

Encore,

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

et

$$p_n = \frac{1}{c!}\left(\frac{\lambda}{\mu}\right)^c \left(\frac{\lambda}{c\mu}\right)^{n-c} p_0 = \rho^{n-c} p_c.$$

3) Mesures de performance :

a)Le nombre moyen de clients dans le système :

$$\overline{n} = \sum_{n=0}^{\infty} n p_n;$$

$$\overline{n} = \sum_{n=1}^{c-1} \frac{n\left(\frac{\lambda}{\mu}\right)^n}{n!} p_0 + \sum_{n=c}^{\infty} \frac{n\left(\frac{\lambda}{\mu}\right)^n}{c!c^{n-c}} p_0;$$

$$\overline{n} = \frac{\lambda}{\mu} + \frac{\left(\frac{\lambda}{\mu}\right)^{c+1}}{cc!\left(1-\frac{\lambda}{c\mu}\right)^2} p_0.$$

b) Le nombre moyen de clients dans la file d'attente :

$$\overline{n}_f = \sum_{k=0}^{\infty} k p_{c+k};$$

$$\overline{n}_f = \frac{\left(\frac{\lambda}{\mu}\right)^c}{c!} \sum_{k=0}^{\infty} k \left(\frac{\lambda}{c\mu}\right)^k p_0;$$

$$\overline{n}_f = \frac{\left(\frac{\lambda}{\mu}\right)^{c+1}}{cc! \left(1 - \frac{\lambda}{c\mu}\right)^2} p_0.$$

c) Le temps moyen de séjour d'un client dans le système :

$$\overline{W}_s = \frac{\overline{n}}{\lambda}.$$

$$\overline{W}_s = \frac{1}{\mu} + \frac{\left(\frac{\lambda}{\mu}\right)^c}{c\mu c! \left(1 - \frac{\lambda}{c\mu}\right)^2} p_0.$$

d) Le temps moyen d'attente d'un client :

$$\overline{W} = \frac{\overline{n}_f}{\lambda}.$$

$$\overline{W} = \frac{c\mu \left(\frac{\lambda}{\mu}\right)^c}{c!(c\mu - \lambda)^2} p_0.$$

On définit ainsi, dans ce qui suit markoviens M/M/1 avec rappels et files d'attente avec feedback

1.2.3 Modèle M/M/1 avec rappels

Un système de files d'attente où un client arrivant dans le système et trouvant tous les serveurs et, éventuellement, positions d'attente occupés tente de nouveau son service après une durée de temps, est appelé système de files d'attente avec rappels. Son étude est motivée par diverses applications pratiques dans le domaine des télécommunications.

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 du système
- La discipline de service,
- La spécification concernant le processus de répétition d'appels.

Le modèle général :

Le système est composé de $c \geq 1$ dispositifs de service et de $(m - c)$ positions d'attente. Les clients arrivent dans le système selon un processus aléatoire avec une loi de probabilité donnée, et forment un flux de clients primaires.

A l'arrivée d'un client primaire, s'il y a un ou plusieurs serveurs libres, le client sera immédiatement pris en charge. Sinon, 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 une source de clients secondaires ou de clients répétés.

La capacité de l'orbite "O" 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 (secondaire) 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-ème tentative échouée).

La notation de Kendall est : A/B/c/m/O/H, où

- A et B décrivent respectivement la distribution du temps entre deux arrivées consécutives et la distribution du temps de service,
- c est le nombre de serveurs identiques et indépendants,
- $(m - c)$ est la capacité d'attente,
- O est la capacité de l'orbite,
- H est la fonction de persistance :$H = \{H_k, k \geq 0\}$.

Si m, O, H sont absents dans la notation de Kendall, alors $m = c, O = \infty, H_k = 1$ pour tout $k \geq 0$.

La distribution du temps inter-rappels(du temps entre deux tentatives consécutives d'un client secondaire d'accédes au serveur) n'est pas indiquée.

Description du modèle

Les clients primaires arrivent dans le système selon un processus homogène de Poisson de taux λ. La durée de temps entre deux arrivées primaires consécutives suit une loi exponentielle de fonction de répartition $A(t) = 1 - \exp\{-\lambda t\}, t \geq 0$.

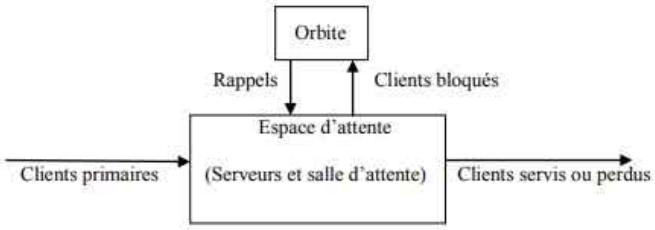

FIGURE 1.8 – Schéma général d'un système avec rappels de file d'attente.

Le service des clients est assuré par un seul serveur.

A l'arrivée d'un client primaire, si le serveur est libre, il est immédiatement pris en charge. Dans le cas contraire, le client en question entre en orbite et devient une source de tentatives répétés (devient source de clients secondaires).

Les durées de service suivent une loi générale commune de fonction de répartition $B(t) = 1 - \exp\{-\gamma t\}, t \geq 0$, et de moyenne finie $\frac{1}{\gamma}$.

La durée de temps entre deux tentatives consécutives (rappels) d'un même client secondaire est distribuée selon une loi de probabilité de fonction de répartition

$T(t) = 1 - \exp\{-\theta t\}, t \geq 0$, de moyenne finie $\frac{1}{\delta}$.

Les trois variables aléatoires introduites sont supposées mutuellement indépendantes. L'état du système à la date t peut être décrit par le processus stochastique suivant :

$$\{C(t), N_0(t),\, t \geq 0\}; \tag{1.7}$$

où $C(t)$ est 1 ou 0 selon le fait que le serveur est occupé ou non, $N_0(t)$ est le nombre de clients en orbite à la date t.

Il s'agit d'un processus de Markov. Supposons que le régime stationnaire existe, c'est-à-dire $\rho = \frac{\lambda}{\gamma} < 1$.

Théorème 1.2.2 *Pour un système de files d'attente M/M/1 avec rappels, la distribution stationnaire conjointe de l'état du serveur et du nombre de clients en orbite* $p_{i,n} = \lim_{t\to\infty} \Pr\{C(t) = i,\, N_0(t) = n\}, i = 0, 1$ *et* $n \leq 0$*, est donnée par*

$$p_{0,n} = \frac{\rho^n}{n!\theta^n} \prod_{k=0}^{n-1} (\lambda + k\theta)(1-\rho)^{1+\frac{\lambda}{\theta}}; \tag{1.8}$$

$$p_{1,n} = \frac{\rho^{n+1}}{n!\theta^n} \prod_{k=1}^{n} (\lambda + k\theta)(1-\rho)^{1+\frac{\lambda}{\theta}}. \tag{1.9}$$

Les fonctions génératrices partielles correspondantes sont données par

$$P_0(z) = \sum_{n=0}^{\infty} z^n p_{0,n} = (1-\rho)\left(\frac{1-\rho}{1-\rho z}\right)^{\frac{\lambda}{\theta}}; \tag{1.10}$$

$$P_1(z) = \sum_{n=0}^{\infty} z^n p_{1,n} = \rho\left(\frac{1-\rho}{1-\rho z}\right)^{\frac{\lambda}{\theta}+1}. \tag{1.11}$$

Preuve

Le processus (1.7) a pour espace d'états $S = \{0, 1\} \times \mathbb{N}$. Les transitions possibles sont :

- de l'état (0,n) vers l'état (1,n) avec un taux λ, ainsi que vers l'état $(1, n-1)$ avec un taux $n\theta$;
- vers l'état (0,n) à partir de l'état (1,n) avec un taux γ ;
- de l'état (1,n) vers l'état (1,n + 1) avec un taux λ, ainsi que vers l'état (0,n) avec un taux γ ;
- vers l'état (1,n) à partir de l'état (0,n) avec un taux λ, de l'état (0,n+1) avec un taux $(n+1)\theta$, ainsi que de l'état $(1, n-1)$ avec un taux λ.

Les équations d'équilibre statistique (de balances) sont

$$(\lambda + n\theta)p_{0,n} = \gamma p_{1,n}; \tag{1.12}$$

$$(\lambda + \gamma)p_{1,n} = \lambda p_{0,n} + (n+1)\theta p_{0n+1} + \lambda p_{1n-1}. \tag{1.13}$$

A l'aide de fonctions génératrices, telles que

$$P_0(z) = \sum_{n=0}^{\infty} z^n p_{0,n} \quad \text{et} \quad P_1(z) = \sum_{n=0}^{\infty} z^n p_{1,n}.$$

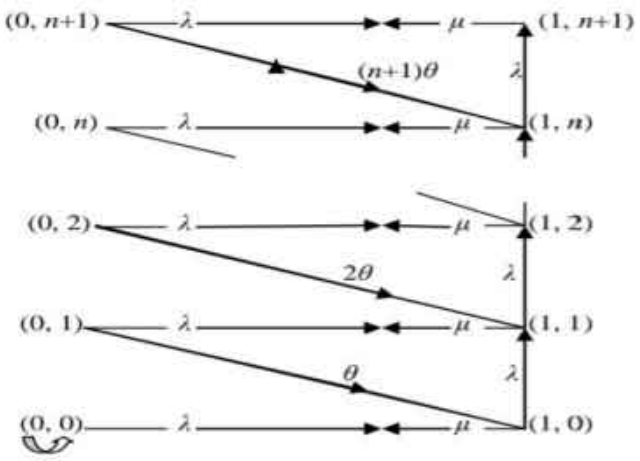

FIGURE 1.9 – Graphe de transitions de la file M/M/1 avec rappels.

les équations (1.12)-(1.13) deviennent

$$\lambda P_0(z) + \theta z P_0^{'}(z) = \gamma P_1(z); \tag{1.14}$$

$$(\lambda + \gamma - \lambda z) P_1(z) = \lambda P_0(z) + \theta P_0^{'}(z).$$

D'où

$$P_0^{'}(z) = \frac{\lambda\rho}{\theta(1-\rho z)} P_0(z).$$

La solution de cette dernière équation est

$$P_0(z) = k^{'}(1-\rho z)^{-\frac{\lambda}{\theta}}.$$

(1.15)

Avec $k^{'} \in \mathbf{R}$.

Des équations (1.14),

$$P_1(z) = \rho P_0(z) + \frac{\theta z}{\gamma} P_0^{'}(z) = P_0(z)\frac{\rho}{1-\rho z} = \frac{\rho\, k^{'}}{(1-\rho z)^{\frac{\lambda}{\theta}+1}}. \tag{1.16}$$

Vu que $\sum_{n=0}^{\infty}(p_{0,n} + p_{1,n}) = P_0(1) + P_1(1) = 1$, on obtient :

$$k^{'} = (1-\rho)^{\frac{\lambda}{\theta}+1}. \tag{1.17}$$

A partir des équations (1.15)-(1.17), on déduit les équations (1.10) et (1.11).

A présent, à l'aide de l'équation (1.12), on élimine $p_{1,n}$ de l'équation (1.13). De cette manière, on trouve

$$(n+1)\theta\,\gamma\, p_{0,n+1} - \lambda(\lambda + n\theta)\, p_{0,n} = n\theta\,\gamma\, p_{0,n} - \lambda(\lambda + (n-1)\theta)\, p_{0,n-1}.$$

Ceci implique que

$$n\theta\gamma p_{0,n} - \lambda\,(\lambda + (n-1)\theta)\, p_{0,n-1} = 0.$$

D'où

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

De l'équation (1.12), on a

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

La probabilité $p_{0,0}$ sera trouvée à l'aide de l'équation de normalisation $\sum_{n=0}^{\infty} p_{0,n} + \sum_{n=0}^{\infty} p_{1,n} = 1$

$$p_{0,0} = \left[\sum_{n=0}^{\infty} \frac{\rho^n}{n!\,\theta^n} \prod_{k=0}^{n-1} (\lambda + k\theta) + \sum_{n=0}^{\infty} \frac{\rho^{n+1}}{n!\,\theta^n} \prod_{k=1}^{n} (\lambda + k\theta) \right]^{-1}.$$

A l'aide de la formule binomiale

$$(1+x)^m = \sum_{n=0}^{\infty} \frac{x^n}{n!} \prod_{i=0}^{n-1} (m-i),$$

on obtient

$$p_{0,0} = (1-\rho)^{\frac{\lambda}{\theta}} + \rho(1-\rho)^{\frac{\lambda}{\theta}+1} = (1-\rho)^{\frac{\lambda}{\theta}+1}.$$

En fin, on peut former les équations (1.8) et (1.9).

Conséquences

1. La distribution stationnaire de processus (1.7) existe si $\rho = \frac{\lambda}{\gamma} < 1$.
2. La fonction génératrice de la distribution stationnaire marginale du nombre de clients en orbite $N_0 = \lim_{t\to\infty} N_0(t)$ est définie par

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

3. La fonction génératrice de la distribution stationnaire du nombre de clients dans le système $N = \lim_{t\to\infty} (N(t) = N_0(t) + C(t))$ est définie

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

.

4. La distribution stationnaire marginale du nombre de serveurs occupés est

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

$$P_1 = \lim_{t\to\infty} \Pr(C(t) = 1) = P_1(1) = \rho.$$

Mesures de performance

e)Nombre moyen de clients dans le système :

$$\overline{n} = E[N] = Q^{'}(1) = \rho + \frac{\lambda^2 \beta_2}{2(1-\rho)} + \frac{\lambda\rho}{\theta(1-\rho)}.$$

f)] Nombre moyen de clients en orbite :

$$\overline{n}_0 = E[N_0] = \overline{n} - \rho = P^{'}(1) = \frac{\lambda^2\beta_2}{2(1-\rho)} + \frac{\lambda\rho}{\theta(1-\rho)}.$$

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

$$\overline{W} = \frac{\overline{n}_0}{\lambda} = \frac{\lambda\beta_2}{2(1-\rho)} + \frac{\lambda\beta_1}{\theta(1-\rho)}.$$

h) Nombre moyen de rappels par client :

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

Ici, $\beta_1 = \frac{1}{\gamma}$ et $\beta_2 = \frac{2}{\gamma}$

1.2.4 Files d'attente avec feedback

Le mot feedback est un mot qui caractérise le client qui quitte la file du à plusieurs facteurs, soit par l'insuffisance du nombre de serveurs, soit par une qualité de service médiocre ou bien le système est mal géré. Dans ce cas, le client retourne à la file pour demander son service.

Modèle d'attente M/M/1 avec Bernoulli feedback

Considérons un système de file d'attente M/M/1 avec Bernoulli feedback. Cette dernière peut modéliser un guichet unique où chaque client reçoit un service dont la durée est une variable exponentielle de paramètre μ et le processus d'arrivée des clients dans la file est un processus de Poisson de taux λ, $N(t)$ est le nombre de clients arrivant pendant un intervalle de temps $[0, t]$ suit une distribution de Poisson. Après avoir obtenu un service avec une probabilité β, le client peut rejoindre le système en tant que client Bernoulli feedback pour recevoir un autre service supplémentaire avec une probabilité $1 - \beta$. Sinon, il quitte définitivement le système, avec une probabilité β.

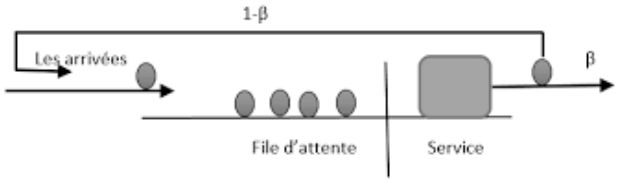

FIGURE 1.10 – Representation d'une file d'attente M/M/1 avec Bernoulli feeedback.

Une file d'attente est un composant élémentaire, tandis qu'un réseau de files d'attente est une structure composée de plusieurs files interconnectées, permettant de modéliser des systèmes plus complexes.

Chapitre 2

Réseaux de file d'attente

Un réseau de file d'attente se compose des entrées, une file d'attente et des serveurs comme des centres de services. En général, il se compose d'un ou plusieurs serveurs pour servir les clients qui arrivent d'une manière quelconque comportant des exigences de service. Les clients (les flux d'entités) représentent les utilisateurs, les emplois, les opérations ou programmes. Ils arrivent au centre de service, attendent pour qu'ils soient servis s'il y a une salle d'attente, et ils quittent le système à la fin du service. Parfois, les clients sont perdus. Donc, les systèmes de files d'attente sont décrits par la distribution des temps interarrivées, la distribution des temps de service, le nombre de serveurs, la discipline de service et la capacité maximale. Notons aussi qu'un réseau de files d'attente est un ensemble de files d'attente interconnectées classé en deux catégories :

* Les réseaux de files d'attente monoclasses, dans lesquels circule une seule classe de clients.

* Les réseaux de files d'attente multiclasses, dans lesquels circulent plusieurs classes de clients.

Les notions présentées dans ce chapitre sont tirées des références : [1] [6] [12] [20].

2.1 Quelques modèles de réseaux de file d'attente

2.1.1 Réseau de file d'attente ouvert et Réseau de file d'attente fermé :

Un réseau de file d'attente est un ensemble de files simples (stations) interconnectées. Soit M le nombre de stations du réseau.

Définition 2.1.1 *Dans un réseau de file d'attente ouvert, les clients arrivent de l'extérieur, circulent dans le réseau à travers les différentes stations, puis quittent le réseau. Le nombre de clients pouvant se trouver à un instant donné dans un réseau ouvert n'est donc pas limité. Afin de spécifier complètement un réseau ouvert, il faut bien sûr caractériser chaque station, mais également le processus d'arrivée des clients et le routage (cheminement) des clients dans le réseau. La figure* 2.1 *donne un exemple d'un réseau de file d'attente ouvert comportant* 4 *stations.*

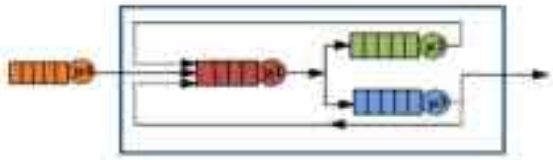

FIGURE 2.1 – Réseau de file d'attente ouvert.

Définition 2.1.2 *Dans un réseau de file d'attente fermé les clients sont en nombre constant. Soit N le nombre total de clients du système. Il n'y a donc pas d'arrivée ni de départ de client. La spécification d'un réseau fermé se réduit donc à celle des différentes stations et à celle du routage des clients. La figure* 2.2 *donne un exemple de réseau de file d'attente fermé comportant* 3 *stations et dans lequel circulent N clients.*

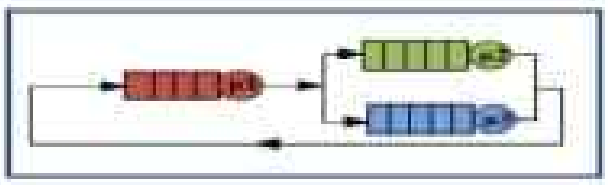

FIGURE 2.2 – Réseau de file d'attente fermé.

2.1.2 Le réseau de Jockson

Un réseau de Jackson est caractérisé par un ensemble de N files d'attente avec la displine FIFO, i.e. avec la discipline de service premier arrivé premier servi. Dans la file d'attente d'ordre i avec $1 < i < N$, le service est exponentiel de paramètre μ_i et l'arrivée des clients dans le réseau à la file "i" est un processus de Poisson de paramètre λ. Une fois le service du client par la file i est terminé, le client rejoint la file j avec probabilité $P_{i,j}$ (avec $P_{i,i} = 0$ i.e. étant donné dans la file i le client ne peut jamais revenir à la file "i") où quitte définitivement le réseau avec la probabilité résiduelle voir figure (2.3).

La matrice $R = (R_{ij}, i, j = 0, ...N)$ est définie par, si $i \neq 0$ et $j \neq 0$,

$r_{ij} = P_{i,j}$;

$r_{i0} = 1 - \sum_{j=1}^{N} P_{i,j}$;

$r_{00} = 1$.

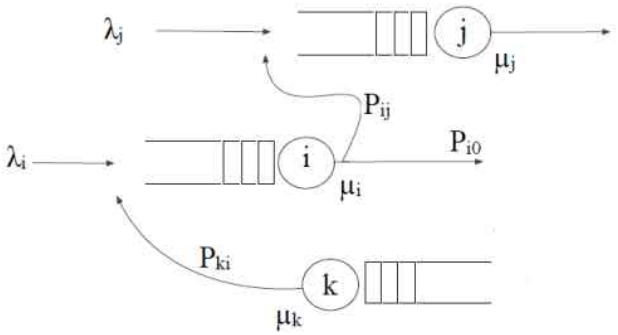

FIGURE 2.3 – Réseau de Jackson.

de telle sorte que "**R**" est une matrice markovienne. On suppose que toutes les variables aléatoires utilisées sont indépendantes. Le processus de Markov associé à ce réseau de file d'attente est à valeurs dans $S = \mathbf{N}^N$. En notant pour $1 < i < N, \quad e_i = (l_{j=i}; 1 \leq j \leq N)$ le $i^{éme}$ vecteur unité, la matrice Q de ce processus de Markov est donnée par

$$q_{n,n+e_i-e_j} = \mu_j \quad n_j > 0, i, j \leq N,$$

$$q_{n,n-e_j} = \mu_j \quad n_j > 0, i, j \geq N,$$

$$q_{n,n+e_i} = \lambda_i \quad i \leq N.$$

La matrice markovienne $R = (R_{ij}; i, J = 0, ..., N)$ est supposée avoir 0 comme unique point absorbant, si (Y_n) est une chaîne de Markov de matrice de transition **R**, presque sûrement (Y_n) est constante égale à 0 à partir d'un certain rang. On fait en outre l'hypothèse que $r_{ii} = 0$ pour tout $1 \leq i \leq N$ (un client ne revient pas en fin de file d'attente après son service). Si cette condition n'est pas remplie, l'expression du générateur montre qu'il suffit de remplacer μ_i par $\mu_i/(1 - r_{ii})$ et les r_{ij} par $r_{ij}/(1 - r_{ii}), \quad j \neq i$, pour se ramener à cette situation.

Lemme

Il existe une suite positive $(\overline{\lambda}_i; 1 \leq i \leq N)$ vérifiant les équations de trafic

$$\overline{\lambda}_i = \lambda_i + \sum_{j=1}^{N} \overline{\lambda}_j p_{j,i}. \tag{2.1}$$

Pour $1 \leq i \leq N$, on posera $\rho_i = \overline{\lambda}_i/\mu_i$.

Exemple 2.1.1

Dans un réseau de Jackson, qui est composé de N files d'attente, dans la file d'attente d'ordre i avec $1 < i < N$*, le service est exponentiel de paramètre* μ_i *et l'arrivée des clients dans le réseau à la file i est un processus de Poisson de paramètre* λ*. Une fois le service du*

client par la file i est terminé, le client rejoint la file j avec probabilité $P_{i,j}$ *(avec* $P_{i,i} = 0$ *ou quitte définitivement le réseau avec la probabilité résiduelle.*

Montrer qu'il existe une suite positive $\left(\overline{\lambda}_i\,;\, 1 \leq i \leq N\right)$ *et vérifiant les équations de trafic*

$$\overline{\lambda}_i = \lambda_i + \sum_{j=1}^{N} \overline{\lambda}_j p_{j,i}.$$

.

Pour $1 \leq i \leq N$*, on posera* $\rho_i = \dfrac{\overline{\lambda}_i}{\mu}$*. On pose* $\lambda = \sum_{j=1}^{N} \lambda_i$ *et* $\alpha = (\lambda_i/\lambda, 1 \leq i \leq N)$*. Si* α *est la distribution initiale de la chaîne de Markov associée à la matrice R, comme cette chaîne est transiente sur* $1, ..., N$ *et l'état* 0 *est absorbant, le nombre de visites* N_i *à l'état* $i = 1, ..., N$ *est intégrable et*

$$E(N_i) = P(Y_0 = i) + \sum_{j=1}^{N} \sum_{n=0}^{\infty} P(Y_n = j, Y_{n+1} = i).$$

.

La propriété de Markov donne l'identité

$$\lambda E(N_i) = \lambda_i + \sum_{j=1}^{N} \sum_{n=0}^{\infty} \lambda P(Y_n = j) r_{ji} = \lambda_i + \sum_{j=1}^{N} \lambda E(N_j) r_{ji}.$$

le vecteur $(\lambda E(N_i),\ 1 \leq i \leq N)$ *est donc la solution de* $\frac{\lambda_1}{\mu_{12}} + \frac{\lambda_2}{\mu_{22}} < 1$*.*

2.1.3 Réseaux de files d'attente multiclasses :

Les réseaux de files d'attente multiclasses, dans lesquels circulent plusieurs classes de clients. ces différentes classes pouvant se distinguer par un schéma de routage spécifique et par des comportements différents au niveau de chaque station, tant au niveau du service que de l'ordonnancement de l'attente. Dans le cas de réseaux multiclasses, si toutes les classes de clients sont des classes ouvertes, on parlera de "réseaux pûrement ouverts" et si toutes les classes de clients sont des classes fermées, on parlera de "réseaux purement fermés". Un réseau parcouru à la fois par des classes ouvertes et des classes fermées sera qualifié comme " réseau mixte".

Exemple 2.1.2 *La figure* (2.4) *donne un exemple de réseau de file d'attente ouvert parcouru par deux classes de clients*

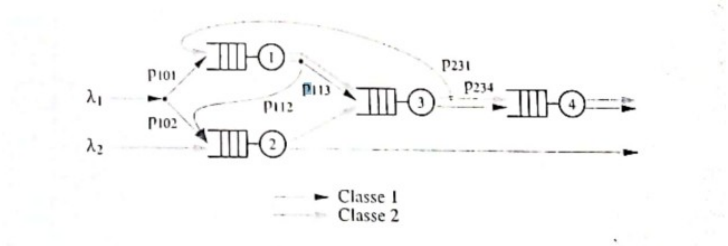

FIGURE 2.4 – Réseau de file d'attente multiclasse.

On est alors amené à caractériser pour chaque classe r :

- ***Pour un réseau ouvert***, *le processus d'arrivée (pour une classe donnée) est un processus d'arrivée poissonien. Il suffit alors de classer les taux d'arrivée λ_r des clients de classe r ;*
- ***Pour un réseau fermé***, *le nombre total N de clients est fixé ;*
- ***Le routage des clients*** *: si on se limite aux routages probabilistes, on définit p_{ij}^r comme la probabilité qu'un client de classe r, quittant la station i, se rende à la station j. (Si i ou j est égal à 0, cela fait référence à l'extérieur : un réseau ouvert.)*

2.1.4 Réseaux à forme produit

Définition 2.1.3 *On considère un réseau de file d'attente parcouru par différentes classes de clients. On suppose dans un premier temps que les clients ne changent pas de classe lors de leur cheminement dans le réseau. Ce réseau possède les caractéristiques suivantes :*

- *un seul serveur à chaque station*
- *une capacité de stockage illimitée à toutes les stations*
- *des routages probabilistes pour chaque classe de clients.*

On note M le nombre de stations du réseau et R le nombre de classes qui le parcourent. Les clients d'une classe donnée ne pouvant changer de classe. Chaque classe est donc soit une classe ouverte, soit une classe fermée. On note O l'ensemble des classes ouvertes du réseau et F l'ensemble des classes fermés : $O \cap F = \varnothing$ et $O \cup F = (1, ..., \mathbf{R})$. Les clients d'une classe ouverte arrivent dans le système, accomplissent un certain nombre d'opérations, puis quittent le système. Les clients d'une classe fermée sont, quant à eux, en nombre constant, et ne peuvent ni arriver de l'extérieur ni quitter le système.

Définition 2.1.4 *Pour une classe ouverte donnée $r \in O$, on impose de plus : un parcourent d'arrivée des clients de classe r dans le système poissonien de taux λ_r.*
On note alors p_{0i}^r la probabilité qu'un client de classe r qui arrive dans le système se rende à la station i, $p_{i,j}^r$ la probabilité qu'un client de classe r qui termine son service à la station i se rende à la station j et p_{i0}^r la probabilité qu'un client de classe r qui termine son service à la station i quitte le systéme. Ces probabilités satisfont la relation :

$$\sum_{j=0}^{M} p_{i,j}^r = 1 \quad pour \quad i = 0, \ldots, M. \tag{2.2}$$

avec la convention $p_{0,0}^r = 0$.
Pour une classe fermée donnée $r \in F$, soit N_r le nombre de clients de classe r. On note comme précédemment p_{ij}^r la probabilité qu'un client de classe r qui termine son service à la station i se rende à la station j. Ces probabilités satisfont la relation :

$$\sum_{j=1}^{M} p_{i,j}^r = 1 \quad pour \quad i = 1, \ldots, M. \tag{2.3}$$

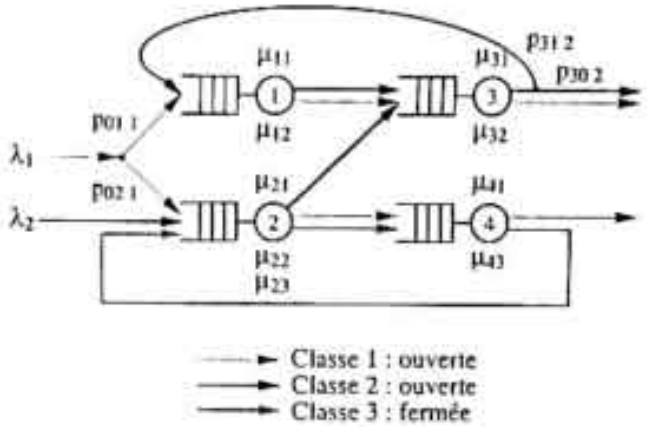

FIGURE 2.5 – Réseau multiclasse à forme produit.

Finalement, chaque station peut utiliser l'une des disciplines suivantes : FIFO, PS, IS ou LCFS-PR (voir le tableau 2.1). À chaque type correspond une discipline de service particulière à laquelle est associée une loi de service.

On constate que seule la discipline FIFO (premier arrivé, premier servi) impose une restriction quant à la loi de service autorisée, puisque celle-ci doit être exponentielle et, de plus, indépendante de la classe du client en service. Cela signifie que tous les clients, quelles que soient leurs classes, auront un temps de service distribué selon une loi exponentielle de taux μ.

Les trois autres disciplines de service, PS (temps partagé), IS (nombre infini de serveurs) et LCFS-PR (dernier arrivé, premier servi, avec préemption du service), n'imposent comme seule contrainte sur la loi de service, que la transformée de Laplace associée à cette dernière puisse s'exprimer sous forme d'une fraction rationnelle.

Exemple 2.1.3 *le service des clients de classe 1 pourra être distribué selon une loi exponentielle de taux $\mu_{i,1}$ tandis que le service des clients de classe 2 sera distribué selon une loi de Erlang-2 de taux $\mu_{i,2}$.*

Type	**Discipline de service**	**Lois de service**
1	FIFO (premier arrivé, premier servi)	Exponentielles indépendantes de la classe du client en service : taux de service μ_i
2	PS (temps partagé)	Générales différentes pour chaque classe (à transformée de Laplace rationnelle) : $1/\mu_{ir}$temps de service moyen des clients de classe r
3	IS (nombre de serveurs infinis)	
4	LCFS-PR (dernier arrivé, premier servi, avec préemption du service)	

TABLE 2.1 – Réseaux BCMP sans changement de classe (et à taux indépendant de l'état)

Un grand intérêt a été attribué à la compréhension de la dynamique des réseaux de files d'attente multiclasse, et en particulier leurs propriétés de stabilité. De nombreuses techniques ont été développées pour l'analyse de la stabilité ou l'ergodicité en utilisant une variété de méthodes.
∗ L'analyse de la stabilité et de l'évaluation de la performance des réseaux de files d'attente ont reçu beaucoup d'attention. Ceci est dû à plusieurs exemples qui démontrent que les conditions de stabilité habituelles (intensité du trafic inférieure à 1 à chaque station) ne sont pas suffisantes pour la stabilité, même dans le cadre de la discipline de service FIFO voir par exemple, Rybko et Stolyar, Bramson (1993).

Les méthodes utilisées pour établir la stabilité des réseaux de files d'attente ont été développées par plusieurs chercheurs, en utilisant des limites fluides, voir Dai, Chen (1995), et les fonctions de Lyapunov par exemple Kumar et Meyn (1995), (1996).

2.2 Les méthodes de stabilité :

2.2.1 La méthode de limite fluide

L'étude du processus $(X(t))$ à l'état sationnaire est en général difficile, dans la plupart des modèles ou, le comportement à l'équilibre est explicité, la loi stationnaire est représentée par la fameuse forme produit. Pour l'étude du comportement asymptotique du réseau, on utilise la méthode des limites fluides. Cette méthode consiste à modifier le processus en accélérant le temps et en renormalisant en espace par un paramétre et d'éudier le comportement du processus modifié quand ce paramétre tend vers l'infini. Ce procédé présente l'avantage de gommer toutes les fluctuations indésirables qui n'ont aucun impact sur le comportement principal du réseau.

Définition 2.2.1 *Une limite fluide associée au processus de Markov (X(t)) et à une fonction positive f est un point d'accumulation des lois de probabilité des processus :*

$$\left\{\left(\frac{f(X(x,f(x)t))}{f(x)}\right);x\in S\right\}=\{(||X||_f(x,t)),x\in S\}$$

.
sur l'espace des fonctions càdlàg muni de la topologie de Skorokhod.
Si l'espace d'états S est inclus dans un espace vectoriel de dimension finie, un point d'accumu- lation des lois de probabilité de l'ensemble de processus

$$\left\{\left(\frac{1}{f(x)}X(x,f(x)t)\right);x\in S\right\}=\{(\overline{X}_f(x,t)),x\in S\}$$

.
est aussi appelé par abus de langage une limite fluide. Une limite fluide est donc une loi de probabilité $\mathbf{Q}$ *d'un processus càdlàg. Quitte à agrandir l'espace de probabilité initial, il est possible de représenter une limite fluide comme un processus (W(t)) de loi* $\mathbf{Q}$ *défini sur l'espace de probabilité de base.*
Les limites fluides donnent une expression asymptotique des trajectoires du processus de Markov, et donc le comportement qualitatif de celui-ci pour des grandes valeurs initiales.

Exemple 2.2.1 *Dans une file d'attente M/M/1, les arrivées des clients suivent un processus de Poisson de taux λ, le taux de service est μ, et $L(t)$ est le nombre de clients dans la file à l'instant $t > 0$.*

Par la méthode des limites fluides, on calculer $L(t)$.

$f(z) = z$ si $z \in \mathbb{N}$,le processus renormalisé est défini par

$$\overline{L}(N,t) = \frac{1}{N} L(Nt).$$

si $L(0) = N \in \mathbb{N}$, alors

$$\overline{L}(N,t) \longrightarrow ((1 + (\lambda - \mu)t)^+).$$

sur l'espace $D(\mathbb{R}_+, \mathbb{R})$ muni de la convergence uniforme sur les compacts. La convergence a donc lieu aussi pour la topologie de Skorokhod. La fonction $(1 + (\lambda - \mu)t)^+$ est donc l'unique limite fluide de ce processus de Markov.

2.2.2 La méthode de fonction de Lyapunov

Cette méthode a été adaptée pour l'étude de la stabilité de modèles stochastiques. La méthode des fonctions de Lyapunov s'inspirent de la méthode classique de Lyapunov pour la stabilité des équations différentielles ordinaires, où la stabilité d'une solution est prouvée sans la connaissance préalable de cette solution. Ceci se fait via l'utilisation d'une fonction vérifiant certaines propriétés de régularité. (voir[6])

Chapitre 3

Analyse d'un réseau de file d'attente avec N stations et 2N classes de clients

Les réseaux de traitement stochastiques apparaissent comme des modèles dans les domaines de la fabrication, des télécommunications, des systèmes informatiques et des services. Ces réseaux partagent des caractéristiques communes : ils comportent des entités, telles que des tâches, des clients ou des paquets, qui se déplacent le long de routes, attendent dans des files d'attente, reçoivent un traitement de diverses ressources, et sont soumis à des effets de variabilité stochastique, tels que les temps d'arrivée, les temps de traitement et les protocoles de routage. Les réseaux issus des applications modernes sont souvent très complexes et hétérogènes. Leur analyse et leur contrôle posent généralement des problèmes mathématiques difficiles. Une approche pour relever ces défis consiste à considérer des modèles approximatifs.

Au cours des 15 dernières années, des progrès significatifs ont été réalisés dans l'utilisation de modèles approximatifs pour comprendre la stabilité et la performance d'une classe de réseaux de traitement stochastiques appelés réseaux de files d'attente multiclasses ouverts HL. HL désigne une discipline de service sans attente inutile, dite Head-of-the-Line, c'est-à-dire que les tâches sont extraites d'un file d'attente dans l'ordre de leur arrivée. Des exemples de telles disciplines sont FIFO (premier arrivé, premier servi) et les priorités statiques. Des approximations de premier ordre (lois des grands nombres fonctionnelles), appelées modèles fluides, ont été utilisées pour étudier la stabilité de ces réseaux. Des approximations de second ordre (théorèmes de la limite centrale fonctionnels), appelées modèles de diffusion, ont été utilisées pour analyser la performance des réseaux fortement congestionnés.

Le développement de l'approche fluide a été inspiré par l'étude de certains contre-exemples dans les travaux de Kumar et Seidman, Rybko et Stolyar, et Bramson, entre autres, où les réseaux de files d'attente multiclasses ne sont pas stables, même lorsque l'intensité du trafic à chaque station du réseau est inférieure à un.

Un résultat élégant de l'approche par modèle fluide a été proposé d'abord par Rybko et Stolyar, puis généralisé et affiné par Dai, Chen, Dai et Meyn , Stolyar et Bramson. Ce résultat affirme qu'un réseau de files d'attente soit stable si le modèle de réseau fluide correspondant est stable. Une réciproque partielle de ce résultat est également donnée par Meyn, Dai et Puhalskii et Rybko . Heng Quing Ye a utilisé le réseau de Kumar-Rybko-Seidman-Stolyar pour établir la stabilité du réseau de files d'attente fluide.

Dans ce chapitre, nous nous concentrons sur la capacité de certaines grandes classes de réseaux de files d'attente fluides multiclasses sous une discipline de service par priorité. Plus précisément, nous établissons une condition de stabilité pour certains réseaux fluides avec priorité comportant N stations et $2N$ classes de clients, où dans le système, chaque station peut servir plusieurs classes de clients avec des priorités de service différenciées.

Pour stabiliser nos réseaux, un certain nombre de stations devrait être ajouté, ceux-ci agissent plus tard en tant que régulateurs pour les systèmes, l'ajout de ces stations n'est pas aléatoire, il dépend essentiellement de priorité supérieure et inférieure de classes de clients et sur nombre de stations du réseau. L'approche du modèle fluide est utilisé pour prouver la stabilité.

Les notions présentées dans ce chapitre sont tirées des références : [2] [4] [5] [6] [7] [8] [9] [10] [11] [13] [14] [16] [19] [21] [23] [25].

3.1 Les modèles de réseaux de files d'attente multi-classes fluides sous des disciplines de service avec priorité

Nous décrivons les modèles de réseau de files d'attente fluides multiclasses sous des disiplines se service avec priorité $(J, K, \lambda, m, C, P, \pi)$. Le réseau fluide se conspose de J stations (buffers) $(J = N)$ indexé par $j \in J = \overline{1, N}$, sert K, $K = 2N$ classes de clients indexé par $k \in K = \overline{1, 2N}$.

Une classe fluide est servie exclusivement à une station, mais cette dernière peut servir plus d'une classe fluide . $\sigma(k)$ indique la station qu'elle sert le fluide de classe k.

Une classe k peut arriver exogènement dans le réseau à des taux λ_1 et λ_{N+1} (avec ≥ 0), puis elle est servie dans la station $\sigma(k)$, avec un temps de service moyen $m_k = 1/\mu_k$, pour $k = \overline{1, 2N}$.

Après service, une fraction p_{kl} de fluide transforme le fluide de classe k à un fluide de classe l, et la fraction restante, $1 - \sum_{l=1}^{K} p_{kl}$, c'est le fluide qui quitte le réseau.

Soit $C(j)$ l'ensemble des classes résidant dans la station j. Alternativement, on définit une matrice $J \times K$ notée $C = (c_{ij})_{J \times K}$, appelée matrice constitutive, où $c_{jk} = 1$ si $\sigma(k) = j$, et $c_{jk} = 0$ sinon. Soit $Q_k(t)$ le nombre de clients de classe k dans le réseau au temps t, (avec $Q(0) = Q_k(0)$), et $\lambda = (\lambda_k)$ deux vecteurs K-dimensionnels à composantes non négatives. $P = (p_{kl})_{K \times K}$ est une matrice stochastique dont le rayon spectral est strictement inférieur à un, et $\mu = (\mu_k)$ un vecteur K-dimensionnel à composantes strictement positives.

Le vecteur $Q(0)$ est appelé vecteur de niveau de fluide iniale, λ représente le vecteur des taux d'arrivée exogènes, μ le vecteur des taux de service, et la matrice P est appelée matrice de transfert de flux.

Lorsque la station $\sigma(k)$ consacre toute sa capacité à servir le fluide de classe k , elle génère un taux $\mu_k > 0$, pour tout $k \in K$.

Entre les classes, le fluide suit une discipline de priorité, qui est décrite par une application bijective π de l'ensemble $\{1, \ldots, K\}$ sur lui-même. Plus précisément, une classe k a une priorité plus élevée par rapport la classe l si $\pi(k) < \pi(l)$ et $\sigma(l) = \sigma(k)$.

On suppose que les clients de classe k ne peuvent être servis à la station $\sigma(k)$ que s'il n'y a aucun client de classe l ayant une priorité supérieure ou égale. Notre réseau de fluides multiclasse consiste en N stations et $2N$ classes de clients. On suppose que le

processus d'arrivée des clients de classe k, pour $k = \overline{1, 2N}$, suit un processus de Poisson avec des taux d'arrivée $\lambda_1 \geq 0$ et $\lambda_{N+1} \geq 0$. Le temps de service pour chaque client de classe k est distribué exponentiellement avec une durée moyenne $m_k > 0$. On suppose également que tous les temps inter-arrivés et les temps de service sont indépendants.

Pour décrire la dynamique du réseau fluide, on introduit le processus de niveaux de fluide K-dimensionnel $\overline{Q} = \{\overline{Q}(t), t \geq 0\}$, dont la $k^{ème}$ composante $\overline{Q}_k(t)$ désigne le niveau de fluide de la classe k à l'instant t ; le processus d'allocation de temps $\overline{T} = \{\overline{T}(t), t \geq 0\}$, dont la $k^{ème}$ composante $\overline{T}_k(t)$ représente le temps total que la station $\sigma(k)$ a consacré pour servir le fluide de classe k durant l'intervalle $[0, t\,]$; et le processus de capacité cumulative $\overline{Y} = \{\overline{Y}(t), t \geq 0\}$, dont la $k^{ème}$ composante $\overline{Y}_k(t)$ désigne la capacité cumulative de la station $\sigma(k)$ pendant l'intervalle $[0, t\,]$, après avoir servi toutes les classes à la station $\sigma(k)$ ayant une priorité au moins égale à celle de la classe k.

On note par D la matrice diagonale de dimension k avec la $k^{ème}$ composante est μ_k, et e un vecteur de dimension k dons toutes les composantes sont égales à un. Soit

$$H_k = \{l : \sigma(l) = \sigma(k), \pi(l) \leq \pi(k)\}$$

est ensemble des indices pour toutes les classes qui sont servies à la même station que la classe k et ont une priorité supérieur à celle de la classe k. On pase $k \in H_k$ par définition.

Alors, la dynamique du modèle de réseau fluide est décrite par :

$$\overline{Q}(t) = \overline{Q}(0) + \lambda t - (I - P^{'})D\overline{T}(t) \geq 0, \tag{3.1}$$

$$\overline{T}(\cdot) \text{ est croissante avec } \overline{T}(0) = 0, \tag{3.2}$$

$$\overline{Y}_k(t) = t - \sum_{l \in H_k} \overline{T}_l(t) \text{ est croissante}, \quad k \in K, \tag{3.3}$$

$$\int_0^\infty \overline{Q}_k(t)\, d\overline{Y}_k(t) = 0, \quad k \in K. \tag{3.4}$$

L'équation suivante est la $k^{ème}$ coordonée de équation de flux (3.1) :

$$Q_k(t) = Q_k(0) + \lambda_k t + \sum_{l=1}^{K} p_{lk}\mu_l T_l(t) - \mu_k T_k(t) \geq 0, \quad k = 1, \ldots, K. \tag{3.5}$$

L'équation (3.1) établit la relation entre le processus d'allocation de temps $T(\cdot)$ et le processus de capacité cumulative $Y(\cdot)$. La relation (3.4) signifie à chaque temps t il existe une capacité positive dans la station $\sigma(k)$ pour servir les classes des clients qui ont une priorité strictement inférieure à k seulement lorsque les niveaux de fluide de toutes les classes dans H_k sont égaux à zéro.

On dit que le couple $\left(\overline{Q}, \overline{T}\right)$ est une solution fluide s'il satisfait l'ensemble des équations (3.1) à (3.4). Par commodité, on appelle également $\overline{Q}$ une solution fluide s'il existe un $\overline{T}$ tel que le couple $\left(\overline{Q}, \overline{T}\right)$ soit une solution fluide.

Le réseau fluide $(J, K, \lambda, m, C, P, \pi)$ est dit stable s'il existe un temps $\tau \geq 0$ tel que $Q(\tau + \cdot) \equiv 0$ pour toute solution fluide $\overline{Q}$ avec $\|Q(0)\| = 1$; et il est dit faiblement stable si $\overline{Q}(\cdot) = 0$ pour toute solution fluide $\overline{Q}$ avec $\overline{Q}(0) = 0$.

Les processus $\overline{Q}$, $\overline{Y}$ et $\overline{T}$ sont lipschitziens, et donc dérivables presque partout sur $[0, \infty)$, cette propriété bien connue sera utilisée dans ce chapitre.

Il est bien connu que le processus de la longueur de file d'attente $Q(t)$ est une chaîne de Markov en temps continu sous les hypothèses d'arrivée de type Poisson et de service exponentiel.

On dit que le réseau fluide $(J, K, \lambda, m, C, P, \pi)$ est stable si la chaîne de Markov $Q(t)$ est positivement récurrente. Il est également bien connu que la chaîne de Markov $Q(t)$ est positivement récurrente seulement si l'intensité de trafic pour chaque station est inférieure à un, c'est-à-dire $\rho_j < 1$ (où ρ_j est la j-ième composante de ρ, l'intensité de trafic pour la station j) pour tout $j \in J$, ou en résumé, $\rho < e$ où e est un vecteur J-dimensionnel avec toutes les composantes égales à un.

L'espérance stationnaire de la longueur totale de la file $\overline{Q}$ est définie par

$$\overline{Q} = \lim_{t\to\infty} \mathbb{E}\left[\sum_{k\in K} Q_k(t)\right].$$

La longueur de la file d'attente $\overline{Q}(t)$ est finie si et seulement si le processus Q est positivement récurrent.

Chen et Zhang ont donné un résultat très important concernant la stabilité des systèmes de fluide avec priorité. Les auteurs ont établi une condition suffisante de stabilité basée sur l'existence d'une fonction de Lyapunov linéaire, cette dernière fournit aussi une condition nécessaire et suffisante pour la stabilité.

Leur résultat est présenté dans le théorème 3.1.1. Afin de l'énoncer, nous avons besoin de quelques hypothèses additionnelles :

Posons

$$h(k) = \begin{cases} \arg\max\{\pi(l) : l \in H_k^+\} & \text{si } H_k^+ \neq \emptyset, \\ 0 & \text{sinon,} \end{cases} \tag{3.6}$$

avec $H_k^+ = H_k \setminus \{k\}$, en d'autres termes : si k n'est pas la classe prioritaire à la station $\sigma(k)$, alors $h(k)$ est l'indice de la classe ayant la priorité immédiatement supérieure à celle de la classe k à la station $\sigma(k)$, sinon $h(k) = 0$.

$$\theta = \lambda - (I - P')\mu_H^0, \tag{3.7}$$

où $\mu_H^0 = De_H^0$ ($e_H^0 = (e_1^0, \ldots, e_K^0)'$) est un vecteur K-dimensionnel avec $e_k^0 = 1$ si $H_k^+ = \emptyset$ et $e_k^0 = 0$ sinon.

$$R = (I - P')D(I - B), \tag{3.8}$$

où $B = (b_{lk})$ est une matrice $K \times K$ avec $b_{lk} = 1$ si $k = h(l)$, et $b_{lk} = 0$ sinon (pour $l, k = 1, \ldots, K$).

Enfin, l'intensité de trafic de notre réseau de file d'attente est donnée par

$$\rho = CD^{-1}(I - P')^{-1}\lambda. \tag{3.9}$$

Théorème 3.1.1 *On considérons un réseau de fluide (λ, μ, P, C) sous une discipline de service par priorité π. Soit le vecteur θ et la matrice R définis respectivement par les équations (3.7) et (3.8). Supposons que $\rho < e$. Alors, le réseau de fluides est stable s'il existe un vecteur h de dimension K avec $h \geq 0$ tel que, pour toute partition a et b de K satisfaisant si la classe $l \in a$, alors toute classe k avec*

$$\textit{Si } \sigma(k) = \sigma(l) \textit{ et } \pi(k) > \pi(l), \textit{ est aussi dans } a. \tag{3.10}$$

nous avons :

$$h_a^{'}\left(\theta_a + R_{ab}x_b\right) < 0 \tag{3.11}$$

Pour tout $x_b \in S_b := \{u \geq 0 : \theta_b + R_b u = 0$ *et* $u \leq e\}$ *lorsque* $b \neq \emptyset$*, et* $x_b = 0$ *lorsque* $b = \emptyset$*. L'inégalité (3.11) est supposée satisfaite par défaut lorsque* $S_b = \emptyset$*.*

L'ensemble a regroupe toutes les classes ayant un taux de capacité cumulative nul, tandis que l'ensemble b regroupe toutes les classes ayant un taux de capacité cumulative positif à l'instant t.

3.2 La stabilisation de quelques modèles de réseau de file d'attente

Dans cette partie, nous présentons deux théorèmes. Nous fournissons la démonstration du premier théorème, tandis que la démonstration du second est omise similaire à la première.

3.2.1 Stabilisation du réseau de file d'attente de N stations sous des disciplines de serivices avec priorités avec quelques stations additionnelles :

Notre réseau de files d'attente multiclasses est constitué de N stations et $2N$ classes de clients. Supposons que le processus d'arrivée de la classe k, $k = \overline{1, 2N}$, suit un processus de Poisson avec des taux d'arrivée λ_1 et λ_{N+1} (≥ 0) vers respectivement les classes et $N + 1$. Le temps de service pour chaque client de classe k suit une loi exponentielle avec un temps de service moyen $m_k > 0$. Nous supposons également que tous les temps inter-arrivées et les temps de service sont indépendants.

Supposons que chaque classe paire à la station $\overline{1, N}$ est prioritaire.

Nous modifions notre réseau de manière que, s'il est composé d'un nombre pair de stations, nous ajoutons N stations additionnelles ; sinon, nous ajoutons $(N - 1)$. L'explication de ce choix sera donnée dans la suite de la section. Le réseau modifié est illustré dans les Figures 3.1 et 3.2 ; les stations additionnelles sont nommées station $N + 1, \ldots,$ station $2N$ (si N est pair), ou station $N + 1, \ldots,$ station $2N - 1$ (si N est impair).

Théorème 3.2.1 *On suppose que même si* $\rho < e$ *l'équation (3.11) n'est pas satisfaite. Si*

$$\lambda_{k1} > \frac{\left(1 - \frac{m_{k_1^{'}}}{m_{k_1^{''}}}\right)}{m_{k_1}}. \tag{3.12}$$

Le processus de la longueur de la file d'attente $Q(\cdot)$ *est* ***réccurent positif.***

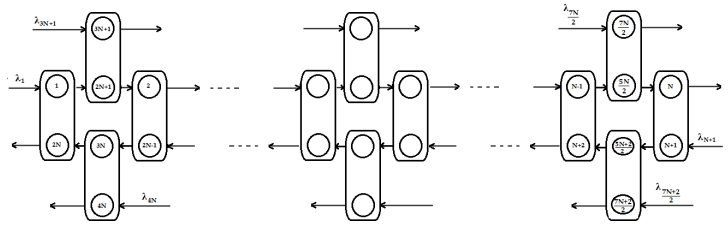

FIGURE 3.1 – Réseau de file d'attente fluide avec 2N stations sous des disciplines de services avec priorité.

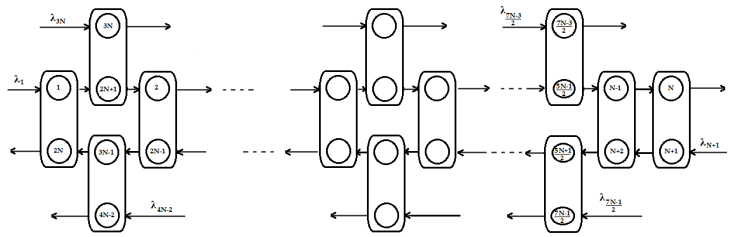

FIGURE 3.2 – Réseau de file d'attente fluide avec 2N-1 stations sous des disciplines de services avec priorité.

Avec λ_{k_1} (resp. m_{k_1}) est le taux d'arrivée exogène (resp. le temps de service moyen) de la classe de fluide de priorité supérieure des stations additionnelles $i = \overline{N+1, 2N}$ (N pair) (resp. $i = \overline{N+1, 2N-1}$ (N impair)), tel que :

$k_1 = \overline{3N+1, 4N}, (N \text{ pair})$ $k_1 = \overline{3N, 4N-2}, (N \text{ impair})$

$m_{k_1'}$ est le temps de service moyen de la classe de fluide de priorité inférieure des stations additionnelles$[i = \overline{N+1, 2N}, (N : \text{pair}](resp.i = \overline{N+1, 2N-1}, (N = impair))$.
$m_{k_1''}$ est celui de la classe de priorité supérieure du réseau original.

On a :

$$m'_{k_1} = \begin{cases} m_{k_1-N}, & k_1 = \overline{3N+1, 4N}, \quad (N \text{ pair}) \\ m_{k_1-(N-1)}, & k_1 = \overline{3N, 4N-2}, \quad (N \text{ impair}) \end{cases}$$

$$m_{k_1''} = \begin{cases} m_{k_1'-(2N-j_1)}, & k_1' = \overline{2N+1, \frac{5N}{2}}, \ j_1 = \overline{1, \frac{N}{2}} \\ & \qquad\qquad\qquad\qquad\qquad\qquad (\text{N : pair}), \\ m_{(k_1-k_1')+j_1}, & k_1' = \overline{\frac{5N+2}{2}, k_1 - N}, \ j_1 = \overline{2, N} \\ m_{k_1'-(2N-j_1)}, & k_1' = \overline{2N+1, \frac{5N-1}{2}}, \ j_1 = \overline{1, \frac{N-1}{2}} \\ & \qquad\qquad\qquad\qquad\qquad\qquad (\text{N : impair}). \\ m_{(k_1-k_1')+j_1}, & k_1' = \overline{\frac{5N+1}{2}, k_1 - (N-1)}, \ j_1 = \overline{4, N+1} \end{cases}$$

Où pour chaque k_1' correspond k_1 et j_1, (j_1 est un nombre pair).

où pour chaque k_1' correspond un k_1 et un j_1 (avec j_1 nombre pair).

Dans Chen et Yao et Dai, il a été démontré que pour prouver la stabilité d'un modèle de files d'attente, il suffit d'étudier la stabilité de son modèle de file d'attente fluide correspondant. Notre démonstration est basée sur ce résultat. Pour mieux comprendre ce phénomène, examinons la dynamique du réseau original travail initial. Lorsque les classes

de priorité supérieure sont en cours de service, les classes de priorité inférieure restent en attente (un client de classe 1 ne peut pas passer en classe 2, ni un client de classe 2 en classe 3, etc., pour un service ultérieur, et vice-versa). Ainsi, ces classes ne seront jamais servies en même temps et forment en effet des stations virtuelles (Dai et Vande Vate). Par conséquent, l'intensité nominale totale du trafic pour ces classes réunies, c'est-à-dire pour les stations virtuelles, ne doit pas dépasser un pour que le réseau soit stable. Un argument similaire montre que le réseau devient instable lorsque l'intensité nominale du trafic pour les stations virtuelles dépasse un, c'est-à-dire lorsque la condition (3.11) n'est pas satisfaite.

Considérons maintenant le réseau modifié. Les classes additionnelles agissent comme des régulateurs qui régulent les trafics du réseau afin de le stabiliser. Lorsque les charges de travail des classes k_1 (définies dans le théorème) sont légères, une grande capacité de service des stations additionnelles est disponible pour les classes k_1' (également définies dans le théorème), et ces dernières ne bloquent donc pas le trafic, évitant ainsi l'accumulation de files d'attente dans les classes prioritaires du réseau original. Ainsi, l'effet de station virtuelle prédomine et le réseau reste instable si la condition (3.11) n'est pas satisfaite.

Cependant, lorsque les charges de travail des classes k_1 sont suffisamment lourdes pour que la condition (3.12) soit satisfaite, le service pour les classes de priorité inférieure dans les stations additionnelles est effectivement ralenti, et le trafic dans le réseau original est retenu (ces classes ne se bloquent pas mutuellement leurs services). Finalement, l'effet de station virtuelle est évité et le réseau modifié est stabilisé.

La dynamique de notre modèle de réseau fluide modifié peut être décrite comme suit :

$$\overline{Q}_{k_1}(t) = \overline{Q}_{k_1}(0) + \lambda_{k_1} t - \mu_{k_1} \overline{T}_{k_1}(t) \geq 0, \tag{3.13}$$

$k_1 = 1, N{+}1, \overline{3N+1, 4N}$ (N pair), (resp. $k_1 = 1, N{+}1, \overline{3N, 4N-2}$ (N impair)),

$$\overline{Q}_k(t) = \overline{Q}_k(0) + \mu_l \overline{T}_l(t) - \mu_k T^k(t) \geq 0, \tag{3.14}$$

(k, l) = deux classes de clients successives, où les clients de la classe k arrivent à la classe l,

$$\overline{T}_k(\cdot) \text{ est croissante avec } \overline{T}_k(0) = 0, \tag{3.15}$$

$k = \overline{1, 4N}$ (N pair) (resp. $k = \overline{1, 4N-2}$ (N impair),

$$\begin{cases} \overline{Y}_{k_1}(t) = t - \overline{T}_{k_1}(t), \\ \quad \text{est une fonction croissante,} \\ \overline{Y}_{k_1''}(t) = t - \overline{T}_{k_1''}(t), \end{cases} \tag{3.16}$$

$$\overline{Y}_k(t) = t - \overline{T}_l(t) - T_k(t) \quad \text{est croissante,} \tag{3.17}$$

(k,l) = (classe de client de priorité inférieure, classe de client de priorité supérieure) à la station i, i=$\overline{1, 2N}$ (N pair), (resp. $i = \overline{1, 2N-2}$ (N impair)),
$\int_0^\infty \overline{Q}_k(t) d\overline{Y}_k(t) = 0$, $k = \overline{1, 4N}$ (N pair) (resp. $k = \overline{1, 4N-2}$ (N impair)). (3.18)

L'étude de stabilité du réseau fluide modifié se fera en trois étapes.

Première étape. Nous prouvons qu'il existe un temps $\tau_1 \geq 0$ tel que

$$\overline{Q}_{k_1}(t) = 0, \quad \text{pour tout } t \geq \tau_1, \tag{3.19}$$

avec $k_1 = \overline{3N+1, 4N}$ (pour N pair), (respectivement $k_1 = \overline{3N, 4N-2}$ pour N impair).

Si $\dot{\overline{Q}}_{k_1}(t) > 0$, alors d'après l'équation (3.18),

$$\dot{\overline{Y}}_{k_1}(t) = 0, \tag{3.20}$$

Puis, en utilisant les conditions (3.16) et (3.20),

$$\dot{\overline{T}}_{k_1}(t) = 1, \tag{3.21}$$

et en combinant (3.13) et (3.21), on obtient

$$\dot{\overline{Q}}_{k_1}(t) = \lambda_{k_1} - \mu_{k_1}. \tag{3.22}$$

Notons que la condition $\rho < e$ implique $\lambda_{k_1} < \mu_{k_1}$.

Posons

$$\tau_1^{(l)} = \frac{\dot{\overline{Q}}_{k_1}(0)}{\mu_{k_1} - \lambda_{k_1}}, \quad l = \overline{1, \frac{N}{2}} \quad (N \text{ pair}) \quad (\text{respectivement } l = \overline{1, \frac{N-1}{2}} \quad (N \text{ impair})).$$

Alors, nous avons

$$\overline{Q}_{k_1}(t) = 0 \quad \text{pour tout } t \geq \tau_1^{(l)}. \tag{3.23}$$

En posant

$$\tau_1 = \max\left(\frac{1}{\mu_{k_1} - \lambda_{k_1}}\right),$$

nous avons que $\tau_1 \geq \max(\tau_1^{(l)})$ (chaque l correspondant à un k_1) sous l'hypothèse $\|Q(0)\| = 1$.

La conclusion (3.23) conduit donc à l'assertion (3.19).

Deuxième étape. Nous prouvons qu'il existe un temps $\tau_2 \geq \tau_1$ tel que

$$\overline{Q}_{k_1''}(t) = 0, \quad \text{pour tout } t \geq \tau_2, \tag{3.24}$$

où k_1'' désigne la classe de client de priorité supérieure à la station i, $i = \overline{1, N}$.

La définition de k_1'' est la suivante :

$$k_1'' = \begin{cases} k_1' - (2N - j_1), \quad k_1' = \overline{2N+1, \frac{5N}{2}}, \quad j_1 = \overline{1, \frac{N}{2}} \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (\text{N pair}), \\ (k_1 - k_1') + j_1, \quad k_1' = \overline{\frac{5N+2}{2}, \quad k_1 - N}, \quad j_1 = \overline{2, N}, \\ k_1' - (2N - j_1), \quad k_1' = \overline{2N+1, \frac{5N-1}{2}}, \quad j_1 = \overline{1, \frac{N-1}{2}} \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (\text{N impair}), \\ (k_1 - k_1') + j_1, \quad k_1' = \overline{\frac{5N+1}{2}, \quad k_1 - (N-1)}, \quad j_1 = \overline{4, N+1}. \end{cases}$$

Sous la condition (3.19), on a $\dot{\overline{Q}}_{k_1}(t) = 0$, et donc

$$\dot{\overline{T}}_{k_1}(t) = \lambda_{k_1} m_{k_1}, \quad \text{pour } k_1 = \overline{3N+1, 4N} \quad (N \text{ pair}),$$

(resp. $k_1 = \overline{3N, 4N-2}$ pour N impair), pour tout temps $t \geq \tau_1$.

Combiné avec (3.17), ceci donne

$$\dot{\overline{Y}}_{k_1'}(t) = 1 - \dot{\overline{T}}_{k_1'}(t) - \dot{\overline{T}}_{k_1}(t) \geq 0.$$

k_1' sont des classes de priorité inférieure dans les stations additionnelles,

$$k_1' = \begin{cases} k_1 - N, & k_1 = \overline{3N+1, 4N}, \quad (N \text{ pair}), \\ k_1 - (N-1), & k_1 = \overline{3N, 4N-2}, \quad (N \text{ impair}). \end{cases}$$

et

$$\dot{\overline{T}}_{k_1'}(t) \leq 1 - \dot{\overline{T}}_{k_1}(t) = 1 - \lambda_{k_1} m_{k_1}, \quad \text{pour tout } t \geq \tau_1 \tag{3.25}$$

Alors,
$\dot{\overline{Q}}_{k_1''}(t) = \mu_{k_1'} \dot{\overline{T}}_{k_1'}(t) - \mu_{k_1''} \dot{\overline{T}}_{k_1''}(t) \leq \mu_{k_1'}(1 - \lambda_{k_1} m_{k_1}) - \mu_{k_{1''}} < 0$, où pour chaque k_1' correspond un k_1'', pour tout $t \geq \tau_2$, et où la dernière inégalité est déduite de l'hypothèse que

$$\lambda_{k_1} > \frac{1 - m_{k_1'}/m_{k_1''}}{m_{k_1}}.$$

Posons $\tau_2^{(l)} = \frac{\dot{\overline{Q}}_{k_1''}(\tau_1)}{\mu_{k_1''} - \mu_{k_1'}(1-\lambda_{k_1} m_{k_1})}, \quad l = \overline{1, N} \quad (N \text{ pair}), \quad (\text{resp. } l = \overline{1, N-1} \text{ si } N \text{ impair}).$
Alors, on a

$$\overline{Q}_{k_1''}(t) = 0 \quad \text{pour tout } t \geq \tau_2^{(l)}. \tag{3.26}$$

Posons

$$\tau_2 = \max\left(\frac{1 + \Theta\tau_1}{\mu_{k_1''} - \mu_{k_1'}(1 - \lambda_{k_1} m_{k_1})}\right)$$

où Θ est la constante de Lipschitz du processus de niveau de fluide $\overline{Q}(t)$. Alors, on a $\tau_2 \geq \max(\tau_2^{(l)})$.

Maintenant, la conclusion (3.26) implique l'assertion (3.24).

Avant de passer à l'étape suivante, nous prouvons séparément que $\overline{Q}_{N+1}(t) = 0$ pour tout $t \geq \tau_2$, dans le cas d'un réseau avec un nombre pair de stations.

Si $\dot{\overline{Q}}_{N+1}(t) = 0$, cela implique $\dot{\overline{Y}}_{N+1}(t) = 0$, ce qui implique à son tour que $\dot{\overline{T}}_{N+1}(t) = 1$. Ainsi, on a $\dot{\overline{Q}}_{N+1}(t) = \lambda_{N+1} - \mu_{N+1}$ avec $\lambda_{N+1} < \mu_{N+1}$ (car $\rho < e$). Donc, il existe $\tau_2' = \frac{\dot{\overline{Q}}_{N+1}(0)}{\mu_{N+1} - \lambda_{N+1}}$, tel que $\dot{\overline{Q}}_{N+1}(t) = 0$ pour tout $t \geq \tau_2$.

Troisième étape. Nous prouvons qu'il existe un temps $\tau \geq \tau_2$ (avec $\tau \geq 0$) tel que

$$\overline{Q}_l(t) = 0, \quad \text{pour } t \geq \tau, \tag{3.27}$$

où l représente les classes de client de priorité inférieure à la station $i = \overline{1, 2N}$ (si N est pair) (resp. $i = \overline{1, 2N-1}$ si N est impair), ce qui, combiné avec les équations (3.19) et (3.24), implique

$$\overline{Q}(t) = 0 \quad \text{pour tout} \quad t \geq \tau.$$

Posons

$$\overline{W}_i(t) = (\lambda_1 m_{l_1} + \lambda_{N+1} m_{l_2})t - \sum_{k:\sigma(k)=i} \overline{T}_k(t), \quad i = \overline{1, N}.$$

Avec $l_1 = \overline{1, N}$ et $l_2 = \overline{N = 1, 2N}$ classes de clients dans la même stations dans le réseau initial.

$W_{i'}(t) =$

$$\begin{cases} \lambda_1 m_{k_1'} t - \overline{T}_{k_1'}(t), \quad k_1' = \overline{2N+1, \frac{5N}{2}}, \quad N: \text{ pair}(resp. k_1' = \overline{2N+1, \frac{5N-1}{2}}, N: impair \\ \lambda_{N+1} m_{k_1'} t - \overline{T}_{k_1'}(t), \quad k_1' = \overline{\frac{5N+2}{2}, 3N}, \quad N: \text{ pair}(resp. k_1' = \overline{\frac{5N+1}{2}, 3N-1}, N: impair \end{cases}$$

pour $\tau \geq \tau_2$. Ici, $\overline{W}(t)$ peut être interprété comme la charge de travail immédiate dans le système à l'instant t. Définissons

$$f_i(t) = k_1' \overline{W}_i(t),$$

avec k_1' une classe de client de priorité inférieure dans les stations additionnelles.

et

$$f_{i'}(t) = k_1'' \overline{W}_{i'}(t),$$

avec k_1'' une classe de client de priorité supérieure dans le réseau initial.

Pour chaque i (resp. i') correspond un k_1' (resp. k_1'').

Alors, il est direct de vérifier que, pour $t \geq \tau_2$:

$$\dot{f}_i(t) < 0 \quad \text{si} \quad \dot{Q}_i(t) > 0, \quad \text{pour} \quad \begin{cases} i = \overline{1, 4N}, & N \text{ pair} \\ i = \overline{1, 4N-2}, & N \text{ impair} \end{cases}$$

Et :

$$f_1(t) \leq f_N(t) \quad \text{si} \quad \overline{Q}_1(t) = 0,$$

$$f_i(t) \leq f_{i-1}(t) \quad \text{si} \quad \overline{Q}_i(t) = 0, \quad i = \overline{2, 3N}, \quad N \text{ pair} \quad (\text{resp. } i = \overline{2, 3N-1}, \quad N \text{ impair}),$$

$$f_j(t) \leq f_i(t) \quad \text{si} \quad \sum_{j \neq i} \overline{Q}_j(t) = 0, \quad j = \overline{1, 3N}, \quad N \text{ pair} \quad (\text{resp. } j = \overline{1, 3N-1}, \quad N \text{ impair}),$$

$$f_N(t) \leq f_N(t) \quad \text{si} \quad \overline{Q}_{3N}(t) = 0, \quad (N \text{ pair}) \quad (\text{resp. } \overline{Q}_{3N-1}(t) = 0, \quad N \text{ impair})$$

En appliquant maintenant l'approche de fonction de Lyapunov linéaire par morceaux pour le modèle de réseau de fluide multiclasses décrit dans le Théorème 3.1 de Chen et Ye, nous obtenons la conclusion (3.27).

3.2.2 Stabilisation de réseaux de files d'attente fluides avec priorité compse de N stations avec N stations additionnelles

Notre réseau multiclasses à N stations est le même que précédemment. Supposons cette fois que la priorité supérieure est attribuée aux classes $N, \overline{N+2, 2N}$.

Nous modifions notre réseau en ajoutant N stations additionnelles (voir Figure 3.3).

- les classe de client $3N+1$ sont prioritaires dans $N+1$,
- les classes $\overline{3N+2, 4N}$ sont prioritaire dans les stations $\overline{N+2, 2N}$.

Nous présentons maintenant le deuxième résultat principal.

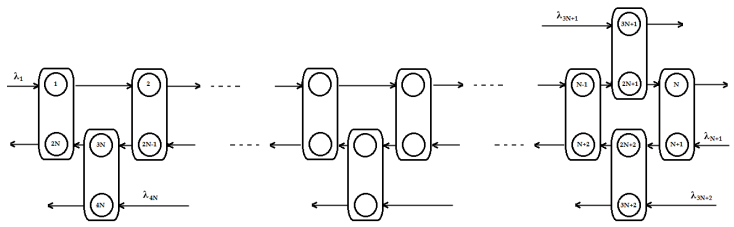

FIGURE 3.3 – Réseau de file d'attente fluide avec 2N stations sous des disciplines de services avec priorité

Théorème 3.3.2 Supposons que $\rho < e$ et que l'équation (3.11) n'est pas satisfaite. Si

$$\lambda_{3N+1} > \frac{(1 - m_{2N+1}/m_N)}{m_{3N+1}}, \quad \lambda_{k_2} > \frac{\left(1 - m_{k_2'}/m_{k_2''}\right)}{m_{k_2}} \tag{3.28}$$

$k_2'' = \overline{N, 2N}, \quad k_2' = \overline{2N+1, 3N}, \quad k_2 = \overline{3N+1, 4N}$ où à chaque k_2 correspondent k_2' et k_2''.

Alors le processus de longueur de file d'attente $Q(\cdot)$ est **récurrent positif**.

Dans ce cas, lorsque les classes de priorité supérieure sont en cours de service, les classes de priorité inférieure ne peuvent pas être servies (la classe 1 ne peut pas passer à la classe 2, la classe 2 ne peut pas passer à la classe 3, etc., pour un service ultérieur, et vice versa). Ces classes forment donc des stations virtuelles. Par conséquent, ces dernières ne doivent pas dépasser une unité pour que le réseau soit stable.

Considérons maintenant le réseau modifié. Les classes additionnelles $2N+1$, et $\overline{3N+1, 4N}$ agissent comme des régulateurs qui régulent les trafics.

Lorsque les charges de travail des classes $3N+1$, et $\overline{3N+2, 4N}$ sont légères, une grande capacité de service des stations $N+1, \ldots, 2N$ est laissée aux classes $\overline{2N+1, 3N}$, respectivement. Par conséquent, ces classes ne retardent pas le trafic, ce qui évite l'accumulation de files d'attente pour les classes de haute priorité du réseau original. Ainsi, l'effet des stations virtuelles prévaut, et le réseau reste instable.

Cependant, lorsque les charges de travail des classes $\overline{3N+1, 4N}$ sont suffisamment lourdes pour que la condition (3.28) soit satisfaite, le service pour les classes de priorité inférieure $\overline{2N+1, 3N}$ est effectivement ralenti, et le trafic vers les classes de haute priorité N et $\overline{N+2, 2N}$ est retenu.

Finalement, l'effet des stations virtuelles est évité et le réseau modifié est ainsi stabilisé.

En suivant les mêmes étapes que dans le théorème 3.3.1, il n'est pas difficile de prouver qu'il existe un temps $\tau_1 \geq 0$ tel que

$$\overline{Q}_{k_2}(t) = 0, \quad k_2 = \overline{3N+1, 4N}, \quad \text{pour tout } t \geq \tau_1. \tag{3.29}$$

Ensuite, on prouve qu'il existe un temps $\tau_2 \geq \tau_1$ tel que

$$\overline{Q}_N(t) = \overline{Q}_{k_2''}(t) = 0, \quad k_2'' = \overline{N+2, 2N}, \quad \text{pour tout} \quad t \geq \tau_2. \tag{3.30}$$

Enfin, on prouve qu'il existe un temps $\tau \geq \tau_2 (\geq 0)$ tel que

$$\overline{Q}_{k'_4}(t) = 0, \quad k'_4 = \text{classe de client de priorité inférieure aux stations } i = \overline{1, 2N}, \quad \text{pour tout } t \geq \tau. \tag{3.31}$$

Conclusion

Dans ce mémoire nous nous somme intéréssés aux systèmes de file d'attente fluides muticlasses.

Le chapitre 1 est une introduction sur les Chaîne de markov et quelques modèles de systèmes de file d'attente.

En raison des difficultés d'analyse des systèmes de files d'attente avec rappels et la complexité des résultats analytiques obtenus, plusieurs chercheurs ont tenté de développer des méthodes approximatives d'analyse des phénomènes ou d'appliquer la propriété de décomposition.

Dans le chapitre 2, on a donné quelques modèles de réseaux de file d'attente, ces derniers sont des outils efficaces pour modéliser de nombreux environnements industriels. Un environnement pour lequel le modèle est particulièrement attractif est le flux de production au sein des installations de fabrication de semi conducteurs. la fin de ce chapitre est consacré à l'une des méthodes de stabilité appelée la méthode de limite fluide.

Dans le chapitre 3 , nous avons étudié la stabilisation de réseaux de files d'attente à N stations en utilisant leur réseau fluide correspondant. Le modèle résultant, réseaux de files d'attente fluides avec stations additionnelles dépendant de la priorité de service et du nombre de stations dans le réseau, est présenté formellement dans ce chapitre. La principale préoccupation de notre travail est la stabilisation des réseaux fluides sous des disciplines de service prioritaire avec des stations additionnelles. La stabilité du réseau fluide modifié implique la stabilité du réseau original.

Annexe

Processus Stochastiques

Définition 3.2.1 *Soient $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité, $T \subset \mathbb{R}^+$ un espace de temps (ou de paramètres) et $E \subset \mathbb{R}$ un espace des états. Un* ***processus stochastique*** *est une application mesurable*

$$X : \Omega \times T \to E$$

$$(\omega, t) \mapsto X(t, \omega) = X_t(\omega)$$

où à t_0 fixé, $\omega \mapsto X_{t_0}(\omega)$ est une variable aléatoire, et à ω_0 fixé, $t \mapsto X(t, \omega_0)$ est une application mesurable (trajectoire).

Processus de naissance et de mort

Définition 3.2.2 [27] *Soient $(\lambda_n)_{n\in\mathbb{N}}$ et $(\mu_n)_{n\in\mathbb{N}^*}$ deux suites de réels strictement positifs. On appelle processus de naissance et de mort de taux de naissance (λ_n) et taux de mort (μ_n) respectivement toute chaîne de Markov en temps continu à valeurs dans $\mathbb{N}$.*
Dans de tels processus les seules transitions possibles à partir de n sont, soient vers $n+1$ en cas de naissance ou vers $n-1$ en cas de décès.

Matrice stochastique

Définition 3.2.3 [15] *Une matrice stochastique est une matrice (a_{ij}) avec $i, j \in \{1, 2, \ldots, r\}$ telle que :*

$$a_{ij} \geq 0$$

$$\sum_{j=1}^{r} a_{ij} = 1 \quad \forall i \in \{1, 2, \ldots, r\}$$

Irréductible

Définition 3.2.4 *Une chaîne de Markov X est dite **irréductible** si tous ses états communiquent, c'est-à-dire si elle possède une seule classe d'équivalence ou si son graphe admet une seule composante fortement connectée :*

$$i \leftrightarrow j, \quad \forall i, j \in S$$

Théorème 3.2.2

[3] *Soit $(X_n)_{n\in\mathbb{N}}$ une chaîne de Markov. Nous supposons que X_n est irréductible, apériodique et homogène. La chaîne de Markov peut posséder une distribution d'équilibre, également appelée distribution stationnaire, soit un vecteur de distribution π (avec* $\pi = (\pi_0, \pi_1, \ldots)$ *) satisfaisant :*

$$\pi P = \pi, \tag{3.32}$$

et

$$\sum_{k\in E} \pi_k = 1. \tag{3.33}$$

Si nous pouvons trouver un vecteur π satisfaisant les équations (3.32) et (3.33), la distribution est unique et

$$\lim_{n\to+\infty} \mathbb{P}(X_n = i \mid X_0 = j) = \pi_i$$

de sorte que π est la distribution limite de la chaîne de Markov.

Formule de Little

Théorème 3.2.3 *La loi de Little établit une connexion entre trois concepts principaux dans la théorie des files d'attente (voir [1è]) : le nombre moyen de clients dans un système $\overline{L}$, le temps moyen passé dans le système $\overline{W}$, et le taux d'entrée dans le système λ. Le théorème de Little est souvent exprimé par l'équation :*

$$\overline{L} = \lambda \overline{W}$$

En se concentrant sur l'attente dans la file (sans considérer le service), la loi de Little permet de relier le nombre moyen de clients en attente $\overline{L_q}$ au temps moyen d'attente d'un client $\overline{W_q}$ avant le service :

$$\overline{L_q} = \lambda \overline{W_q}$$

En ne tenant compte que des serveurs, la loi de Little relie le nombre moyen de clients en service $\overline{L_s}$ au temps moyen de séjour d'un client $\overline{W_s}$ dans le service par la relation :

$$\overline{L_s} = \lambda \overline{W_s}$$

En d'autres termes, à partir de ces trois relations, on peut déduire :

$$\overline{L} = \overline{L_s} + \overline{L_q} \quad et \quad \overline{W} = \overline{W_s} + \overline{W_q}$$

Définition 3.2.5 *La suite $(x_n)_{n\in\mathbb{N}}$ dans $D(\mathbb{R}_+,\mathbb{R}^N)$ converge uniformément sur des ensembles compacts (c.u.c) vers $x \in D(\mathbb{R}_+,\mathbb{R}^N)$ si pour chaque $T>0$, on a :*

$$\lim_{n\to\infty} \sup_{t\in[0,T]} \|x_n(t) - x_n(0)\| = 0.$$

Définition 3.2.6 [21] *Une fonction f d'un intervalle J dans $\mathbb{R}$ est de Lipschitz d'ordre $\alpha > 0$ s'il existe une constante $C > 0$ telle que pour tout $x, y \in J$,*

$$|f(x) - f(y)| \leq C|x - y|^\alpha.$$

Proposition 3.2.4 *Pour $T > 0$ il existe une constante K_T telle que, si Y et Y' sont des fonctions càdlàg telles que $Y(0) \geq 0$ et $Y'(0) \geq 0$, alors :*

$$\|X_Y - X_{Y'}\|_{\infty,T} \leq \|Y - Y'\|_{\infty,T}, \tag{3.34}$$

$$\|R_Y - R_{Y'}\|_{\infty,T} \leq \|Y - Y'\|_{\infty,T}, \tag{3.35}$$

où l'on note

$$\|Z\|_{\infty,T} = \sup_{0\leq s\leq T} \max_{1\leq i\leq d} |Z_i(s)|, \quad \text{si } Z(t) = (Z_i(t))_{1\leq i\leq d}.$$

Si les coordonnées d'une fonction càdlàg Y sont de Lipschitz d'ordre α sur l'intervalle $[0,T]$, il en va de même pour R_Y et X_Y.

Théorème 3.2.5 ***(Théorème de représentation de Skorokhod).***
Si une suite de probabilités $(\mathbb{P}_n)$ sur $D([0,T],\mathbb{R}^d)$ converge vers la probabilité $\mathbb{P}$, il existe un espace de probabilité $(\Omega,\mathcal{F},\mathbb{Q})$ sur lequel sont définis des processus càdlàg $(Y_n(t))_{t\in[0,T]}$, $n \geq 1$, et $(Y(t))$ tels que, Pour tout $n \geq 1$, la loi de $(Y_n(t))$ soit $\mathbb{P}_n$, P La loi de $(Y(t))$ et $\mathbb{Q}$-presque sûrement, $(Y_n(t))$ converge, pour la topologie de Skorokhod, vers $(Y(t))$ quand n tend vers l'infini.

Définition 3.2.7 ***Loi d'Erlang***

** La loi d'Erlang B est donnée par la formule suivante :*

$$P_b = \frac{A^{N_c}/N_c!}{\sum_{i=0}^{N_c} A^i/i!}$$

avec P_b : probabilité de blocage.
A :trafic en Erlang.
N_c :normbre de canaux disponibles.

Bibliographie

[1] Baynat.B. La théoire des files d'attente : des chaînes de Markov aux réseaux à forme produit. *Hermés,*(2000).

[2] Bramson, M. Stability of two famillies of queueing networks and a discussion of fluid limits, *Queueing Systems Theory and Applications*, 23 (1998), 7–31.

[3] Claudie Chabriac. *Processus stochastiques et modélisation.* (2012 − 2013).

[4] Chen, H. Fluid approximations and stability of multiclass queuing networks : Work-conserving discipline, *Annals of Applied Probability*, 5 (1995), 637–655.

[5] Chen, H. and Yao, D.D. Yao, *Fundamentals of Queuing Networks : Performance, Asymptotics and Optimization, Springer-Verlag New York, Inc.* 2001.

[6] Chen, H. and Ye, H.Q. Piecewise liner Lyapunov function for the stability of priority multiclass queuing networks, *IEEE Transactions on Automatic Control*, 47(4) (2002), 564–575.

[7] Chen, H. et Zhang, H. Stability of multiclass queueing networks under priority, *Operations Research*, 48 (2000), 26–37.

[8] Dai, J.G, On positive Harris recurrence of multiclass queueing networks : a une unified approach via fluid models, *Annals of Applied Probability,* 5 (1995), 49–77.

[9] Dai, J.G. A fluid-limit model criterion for l'instability of multiclass queueing networks, *Annals of Applied Probability.,* 6 (1996), 751–757.

[10] Dai, J.G. et Meyn, S.P. Stability and Convergence of moments for multiclass queueing networks via fluid models, *IEEE Transactions on Automatic Control*, 40 (1995), 1899–1904.

[11] Dai, J.G. and J. H. Vande Vate, Global Stability of Two-Station Queueing Networks. *Proceedings of Workshop on Stochastic Networks : Stability and Rare Events, Editors : Paul Glasserman, Karl Sigman and David Yao, Springer-Verlag, Columbia University, New York.* (1996), 1–26.

[12] Dr. Mokhtor Kadi. Réseaux de files d'attente stochastiques. Kadi1969@yahoo.fr.

[13] H. SAKHI et al. *Analyse d'un réseau fluide muti-classe.* PhD thesis, 2017.

[14] Heng-Qing Ye. A paradox for admission control of multiclass queueing network with différentiated service, *J. AppL Probab,* 44(2) (2007), 321–331.

[15] Jean Louis Poss, Probabilité et statistique version 2.1.p74. Mai (2003).

[16] Kumar, P.R. and T.I. Seidman, Dynamic intabilities and stabilization methods in distributed realtime scheduling of manufacturing systems, *IEEE Transactions on Automatic Control*, 35 (1990), 289–298.

[17] Little .John.D.C. A proof of the queueing formula $L = \lambda W$, *Oper. Res, 9(3), 383-387,* (1961).

[18] Lakàtós. Làszló Szeidl. Miklós Telek. Introduction to Queueing Systems with Telecommunication Applications. *Springer, p* 88.(2010).

[19] Meyn, S, Transience of des multiclass queueing networks via fluid limit models, *Annals of Applied Probability*, 5 (1995), 946–957.

[20] Philippe Robert Réseaux et files d'attente : méthodes probabilistes, (2000).

[21] Puhalskii, A. and Rybko, A.N. Non-ergodicity of queueing networks under nonstability of their fluid models, *Problems of information transmission,* 36(1) (2000), 26–48.

[22] P. Robert. Réseaux et files d'attente : méthodes probabilistes, volume 35. *Springer Science Business Media, 2000.*

[23] Rybko, A.N. and Stolyar, A.L. Ergodicity of stochastic processed describing the operations of open queueing networks. *Problemy Peredachi Informatsii*, 28 (1992), 2–26.

[24] Samuel Karlin. Howard, M. Taylor. A first course in stochastic processes, second edition. *Academic press New york San fransisco.* (1975).

[25] Stolyar, A.L. On the stability of multiclass queueing network : a relaxed sufficient condition via limiting fluid processes, *Markov Processes and Related Fields,* 1(4) (1995), 491–512.

[26] Sébastien Loustau. Chaînes de Markov et Processus markoviens de sauts. Applications aux files d'attente. Ecole Centrale de Marseille, Année (2008 − 2009).

[27] Yves Caumel. Probabilité et Processus Stochastiques. *Springer-Verlag FRance,* (2011).