République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

****

Université des Sciences et de la Technologie Houari Boumedienne

Faculté des Sciences Mathématiques Pures et Appliquées

Thèse de Magister

Spécialité : Mathématiques

Option : Méthodes Stochastiques et Recherche Opérationnelle

Présentée par :

HESSAS Fatima

Approximation de systèmes de files d'attente avec rappel

Devant le jury d'examen composé de

M.	M. Bentarzi	Professeur	U.S.T.H.B	Président
M.	A. Aissani	Professeur	U.S.T.H.B	Rapporteur
M.	K. Boukhetala	Maître de conférences	U.S.T.H.B	Examinateur
M.	H. Fellag	Maître de conférences	U.M.M.T.O	Examinateur
M.	D. Hamadouche	Maître de conférences	U.M.M.T.O	Examinateur

soutenue publiquement le 12/11/2002.

Remerciements

Ce travail a été réalisé sous la direction du professeur Amar Aissani auquel je voudrai exprimer toute ma gratitude. Il a suivi et dirigé mes travaux avec patience et compréhension mais aussi avec beaucoup de rigueur et de compétence. Il a su me transmettre son savoir tout au long de cette thèse. Qu'il en soit vivement remercié.

Je souhaite aussi exprimer ma gratitude au professeur Bentarzi pour l'honneur qu'il me fait de présider le jury de ma thèse. Mes remerciements vont aussi à Monsieur Boukhetalla de l'université d'Alger et Messieurs Fellag et Hamadouche de l'université de Tizi-Ouzou pour avoir accepté de juger mon travail.

Enfin, j'aimerai remercier tous mes amis, mes proches et mes parents pour leur soutien moral et matériel et pour les encouragements qu'ils n'ont cessé de m'apporter durant toute cette période.

Je dédie ce modeste travail à :

- Mes chers parents (*Mama et Vava*) sans qui tout cela n'aurait pas été possible.
- Mes frères et soeurs (*Malik, Karim, Nassima, Samia, Assia et Yacine*) que j'aime beaucoup.
- Mon fiancé *Mahdi* que j'aime et que j'aimerai pour toujours.
- *Khalti Hamama* et *Djeddi Mokrane* pour leurs encouragements.
- Mes oncles *Khali Noreddine* et *Khali Rachid* pour leur soutien.
- Ma belle famille et ma belle sœur *Nadia.*

Ce travail est aussi dédié à la mémoire de :

- Mes grand-parents *Djeddi Said*, *Djida Dahbia*, *Mohammed Taïb* et *Fatima Louni*.

– *Khalti Ferroudja* et *Khalti Taous*.

– et toutes les victimes du printemps noir.

TABLE DES MATIERES

INTRODUCTION GENERALE **1**

CHAPITRE I : GENERALITES SUR LES SYSTEMES DE FILES D'ATTENTE.

Introduction **5**

A. Files d'attente classiques **7**

A.1. Description du modèle **7**

A.2. Etude de quelques modèles classiques **9**

A.2.1. Modèles markoviens **9**

A.2.1.1. La file M/M/1 **10**

A.2.1.2. Variantes des files M/M/1 **12**

A.2.1.2.1 Politiques de service autres que FIFO **12**

A.2.1.2.2 Capacités limitées **12**

A.2.2. Modèles non markoviens **13**

A.2.2.1. Système M/G/1 **13**

A.2.2.2. Système G/M/1 **16**

A.2.2.3. Système GI/GI/1 **17**

B. Files d'attente avec rappel **20**

B.1. Description du modèle **20**

B.2. Quelques exemples modélisés par des systèmes de files d'attente avec rappel **22**

B.2.1. Problème de réservation 23
B.2.2. Réseaux locaux CSMA 23
B.2.3. Système informatique à temps réel 24
B.3. Système M/G/1 avec rappel 25
B.4. Autres modèles de systèmes avec rappel 28
B.4.1. Modèle d'attente avec des clients persistants 28
B.4.2. Modèle d'attente avec des clients non persistants 28
B.4.3. Modèle d'attente avec des temps de rappel généraux 30

CHAPITRE II : APPROXIMATIONS DES DEUX MOMENTS DES SYSTEMES CLASSIQUES GI/GI/1.

Introduction 31

A. Méthodes d'approximation des deux moments des files classiques GI/GI/1 32
A.1. Méthodes de diffusion 32
A.1.1.Approximation de Heyman 32
A.1.2. Approximation de Reiser-Kobayashi 32
A.1.3. Approximation de Gelenbe 33
A.1.4. Approximation de Yu 33
A.2. Méthodes heuristiques 33
A.2.1.Approximation de Marchal 33
A.2.2. Approximation de Kramer et Lagenbach-Belz 34
A.2.3. Approximation de Page 35
A.2.4. Approximation de Sakasegawa 36
A.3. Comparaison numérique 37

B. Approximation par interpolation des systèmes GI/GI/s **41**

B.1. Méthodologie de l'interpolation **42**

B.2. Temps moyen d'attente dans la file M/G/s **45**

B.3. Temps moyen d'attente dans la file GI/GI/s **48**

CHAPITRE III : APPROXIMATIONS HEURISTIQUES DU MODELE GI/GI/1 AVEC RAPPEL.

I. Introduction-Position du problème **51**

II. Approximations heuristiques du systeme GI/GI/1 **52**

II.1. Approximations de type linéaire **53**

II.2. Approximations de type harmonique **57**

II.3. Autre approximation heuristique **60**

CHAPITRE IV : SIMULATION ET VALIDATION DES APPROXIMATIONS.

Introduction **62**

A. Simulation des systèmes de files d'attente **62**

A.1. Simulation par événements discrets **62**

A.1.1. Génération de variables aléatoires **63**

A.1.2. Méthodes de génération des nombres aléatoires **63**

A.2. Programme de simulation **64**

A.2.1. Description du modèle **65**

A.2.2. Paramètres d'entrée **67**

A.2.3. Initialisation du système **68**

A.2.4. Evolution du système 68
A.2.5. Calcul des paramètres de performance 70

B. Résultats des simulations 72
B.1. Comparaison numérique 72
B.2. Résultats et discussion 73
B.2.1. Cas $\rho = 0.8$ 73
B.2.2. Cas $\rho = 0.4$ 78
B.2.3. Influence du taux de rappel 83
B.2.4. Cas Ca^2 (Cs^2) > 1 83
B.3. Conclusion 87

Conclusion générale 88

Annexe A : Rappels mathématiques 90
Annexe B : Introduction aux méthodes par diffusion 97

Références 103

INTRODUCTION GENERALE

Ce travail s'inscrit dans la continuité des études intensives menées dans le but de comprendre et de rendre compte des comportements stochastiques des files d'attente avec rappel exponentiel présentant des inter arrivées et des temps de service non exponentiels.

La théorie des files d'attente constitue un outil théorique et pratique pour la modélisation et l'évaluation des performances de systèmes concrets (systèmes de production **[BUZ 1992 ; AIS 1996]**, systèmes informatiques **[SAV 1981; LAZ 1984]**, systèmes de télécommunication **[PEL 1994]**…). Elle fut introduite pour la première fois dans sa version classique pour l'analyse des systèmes téléphoniques **[ERL 1917 ; ENG 1918]**. Cette théorie classique s'est très vite montrée inefficace face à des systèmes réels de plus en plus complexes. Dès la fin des années 1940, des chercheurs tels que Kosten **[KOS 1947]** et Wilkinson **[WIL 1956]** ont mis en évidence les limites de la théorie classique qui ne permettait pas d'expliquer le comportement stochastique des systèmes téléphoniques où les abonnés répétaient leurs appels en recomposant le numéro plusieurs fois jusqu'à l'obtention de la communication.

Ce phénomène de rappel a poussé certains chercheurs à étendre le modèle d'attente classique à celui dit *avec rappel* **[CLO 1948 ; WIL 1956 ; COH 1957]**. Cependant, l'influence de ce phénomène a été longtemps négligée durant les décennies suivantes. Ce n'est que vers les années 1970-1980 qu'on a vu un net regain d'intérêt pour cette catégorie de modèles, avec l'avènement de nouvelles technologies dans les

ateliers de production, les réseaux informatiques, les systèmes de télécommunication,…

Les modèles d'attente avec rappel permettent en effet de mieux modéliser des protocoles de communication spécifiques à certains réseaux de communication (protocoles CSMA, disciplines Auto-repeat, Ring-back-when-free, Repeat-last-number,…**[KEL 1985]**). Les progrès récents dans ce domaine sont résumés dans les articles de synthèse de Yang et Templeton **[YAN 1987]**, Falin **[FAL 1990]** et Aissani **[AIS 1994]**. Pour leurs applications, on pourrait se référer aux articles de Le Gall **[LeG 1970]**, Hashida et Kawashima **[HAS 1975]**, Pourbabai **[POU 1988]**,…

L'évaluation des performances de ce type de systèmes (temps moyen d'attente, nombre moyen de clients, probabilité de blocage,…) se fait suivant différentes techniques. Parmi les mieux adaptées, on distingue trois sortes de méthodes :

- Les techniques dites *analytiques* qui visent à traduire le système à étudier en équations mathématiques, dont la résolution fournira des informations statistiques sur des variables représentant la qualité de service (taux d'occupation des ressources, temps de réponse, ...). La théorie des files d'attente fournit un certain nombre de résultats généraux qui, moyennant le plus souvent quelques hypothèses simplificatrices, permettent de dériver simplement nombre de résultats utiles dans la pratique.

- Les techniques de *simulation* qui permettent d'imiter artificiellement sur ordinateur les systèmes à étudier et de prévoir leurs comportements avec plus ou moins de précision. Celles-ci présentent certains inconvénients comme les temps d'exécution et donc d'implémentation des programmes qui sont assez longs.

- Les techniques d'*approximation des deux moments* qui se satisfont bien de la seule information concernant les deux premiers moments. Celles-ci se basent sur l'interpolation linéaire ou harmonique d'un paramètre de performance à partir de paramètres de performance calculés pour des systèmes connus.

La prise en considération des appels répétés introduit de grandes difficultés analytiques. En effet, des résultats analytiques détaillés n'existent que pour un certain nombre de files d'attente avec rappel particulières, avec des hypothèses contraignantes sur certains paramètres tels que le nombre de serveurs (un seul serveur), les distributions des temps de rappel et d'arrivée (loi exponentielle) et l'état du système (régime stationnaire) par exemple, alors que pour beaucoup d'autres, les résultats obtenus sont extrêmement limités.

Cette contrainte analytique a focalisé l'attention des spécialistes sur le développement d'autres méthodes telles les méthodes numériques **[STE 1983 ; RUS 1984]**, algorithmiques **[WIL 1968 ; HAS 1975 ; FAL 1990]**, de comparaisons stochastiques **[GRE 1987, 1989]** d'approximation heuristique **[ART 1995 ; FAL 1994]** et de simulation.

L'objet de notre travail est de contribuer à cet intérêt particulier pour les files d'attente avec rappel et plus particulièrement aux systèmes GI/GI/1 avec rappel exponentiel. En effet, un examen de la littérature montre que ce modèle est très peu étudié car très mal connu. On propose dans le cadre de la méthode d'approximation à deux moments, différentes approches heuristiques du temps moyen d'attente dans le système par une combinaison linéaire ou harmonique des temps d'attente de systèmes connus (M/M/1, M/D/1) et d'autres moins connus (D/M/1). En l'absence de résultats théoriques exacts, la qualité de telles approximations est testée à l'aide de la simulation statistique.

Ce mémoire est structuré de la manière suivante :

Le premier chapitre comprend une synthèse sur les systèmes de files d'attente classiques et avec rappel. Une description détaillée du modèle classique est présentée avec les résultats connus dans les différentes files de type M/M/1 M/G/1 et G/M/1. Quelques exemples d'application des modèles d'attente avec rappel sont cités et le peu de résultats connus, notamment dans le système M/G/1 avec rappel exponentiel, est présenté.

Dans le deuxième chapitre, on présente, dans un premier lieu quelques résultats de la méthode d'approximation des deux moments dans les files d'attente classiques GI/GI/1 obtenus par les techniques de *diffusion* et *heuristiques* **[SHA 1980]**. La deuxième partie de ce chapitre est consacrée à la méthodologie de l'approximation par interpolation des systèmes (toujours dans le cadre de la méthode des deux moments) GI/GI/s classiques et aux résultats obtenus par Kimura **[KIM 1994]**.

Le troisième chapitre présente l'exploitation de la méthodologie de Kimura pour proposer certaines approximations heuristiques pour les systèmes GI/GI/1 avec rappel. Dans le quatrième chapitre à caractère pratique, après un bref aperçu du programme de simulation utilisé pour la validation des différentes approximations, une comparaison des résultats simulés et interpolés permet de tirer la meilleure approximation et son domaine de validité selon les valeurs des coefficients de variation.

Enfin, une conclusion générale résumera les différents résultats obtenus à travers ce travail ainsi que les perspectives de recherche qui pourraient découler du présent travail.

CHAPITRE I : GENERALITES SUR LES SYSTEMES DE FILES D'ATTENTE.

Introduction

Dans ce chapitre, nous présentons les approches de base de la théorie des files d'attente. La raison principale du succès de cette théorie, introduite pour la première fois dans l'analyse des systèmes téléphoniques **[ERL 1917 ; ENG 1918]**, est la combinaison de la puissance d'expression et de l'efficacité des solutions qu'elle offre.

Plusieurs études ont été menées sur les files d'attente. Une présentation générale et claire est donnée dans le livre de Cox et Smith **[COX 1961]**. Celui de Sakarovitch **[SAK 1978]** propose une approche beaucoup plus mathématique. Une multitude d'autres publications **[KLE 1975, 1976]** a été consacrée à l'application des files d'attente à différents systèmes dans le but de modéliser et résoudre des problèmes tels que : construire des lignes téléphoniques en minimisant les temps d'attente pour obtenir une communication, organiser un système multi-processeurs, un système de temps partagé, un réseau d'ordinateurs,…

La théorie des files d'attente donne deux méthodes principales pour la résolution du conflit qui se produit lorsqu'un client arrive dans le système et trouve le(s) serveur(s) occupé(s) : **i)** Il quitte le système pour toujours sans être servi, ce qui correspond au *"système d'Erlang avec refus"* (Erlang Loss System) appelé aussi *"modèle d'appel perdu"* **[BRO 1948]**. **ii)** Il peut attendre dans une file d'attente pour être servi dès la libération du serveur, ce qui correspond au *"système de files d'attente classique"* (Queueing system) **[KLE 1975 ; LAR 1997]**. Une situation intermédiaire

est envisageable par un rappel ultérieur pour le service, autant de fois qu'il le faut, à des intervalles de temps aléatoires, jusqu'à ce que le client puisse trouver un serveur libre. Ceci correspond à un *"système de files d'attente avec rappels"* (Retrial Queueing System).

L'objet de ce chapitre est de donner une description générale et détaillée des deux systèmes (classique et avec rappel) suivie d'une présentation des principaux modèles étudiés, notamment les systèmes de type M/M/1, M/G/1, G/M/1 et G/G/1.

A. Files d'attente classiques

A.1. Description du modèle

Une file d'attente peut se décrire comme un système où des *clients* (modélisant les activités qui ont besoin d'accéder aux ressources) arrivent à des instants aléatoires vers *une station* (modélisant les ressources) pour recevoir un service. A la lumière des exemples précédents, on voit que les clients peuvent être de toutes sortes (appels téléphoniques, machines,…) de même que la station de service (central téléphonique, processeur,…). La station de service peut comprendre un ou plusieurs serveurs. Quand ceux-ci sont tous occupés, les clients doivent alors patienter dans un espace d'attente (si celui-ci existe) jusqu'à ce qu'un serveur soit disponible. Une représentation graphique d'une file d'attente classique est donnée par la figure I.1.

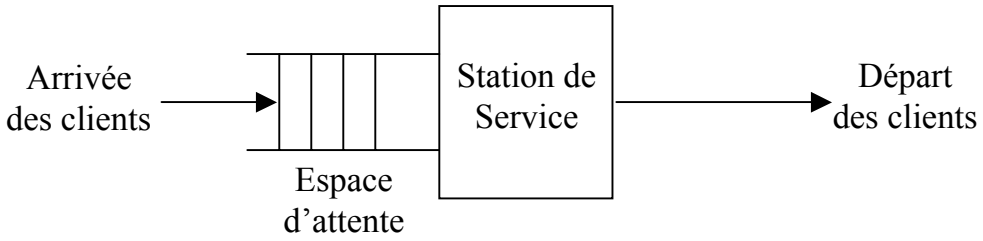


***Figure I.1* : Schéma descriptif d'une file d'attente classique.**

Terminologie et notation

Un système de file d'attente classique est défini par :

a) Un processus générateur d'arrivées, caractérisé par une loi statistique déterminant les intervalles inter arrivées (qui sont des variables aléatoires). Sauf indication contraire (cas d'arrivées groupées), les arrivées se font une à une.

b) La file d'attente proprement dite : elle contient les clients en attente de service. Cependant, quand on parle de temps d'attente pure, on considérera le temps passé dans cette file sans être servi. La file est, de plus, caractérisée par l'ordre dans lequel les clients sont servis.

c) Un ou plusieurs serveurs, caractérisés par une loi statistique déterminant la durée de service (qui est une variable aléatoire).

La plupart des files d'attente qu'on rencontre peuvent être caractérisées par une séquence de six symboles (Conférence Internationale sur la Standardisation des Notations dans la Théorie des Files d'Attente 1971) :

A / B / S / K / m / Z

Avec A : distribution d'inter arrivées.
B : distribution de service.
S : nombre de serveurs.
K : capacité de la file.
m : la borne supérieure du nombre de clients dans le système (peu d'intérêt).
Z : discipline de service.

Où A et B sont donnés par :

M : loi exponentielle (markovienne).
D : loi constante (déterministe).
E_k : loi Erlang d'ordre k.
H_k : loi hyperexponentielle d'ordre k.
G : loi générale
GI : loi générale indépendante.
. . .

La discipline de service détermine l'ordre dans lequel les clients sont rangés dans la file et y sont retirés pour recevoir un service. Les disciplines les plus courantes sont :

a) FIFO (First In First Out) ou FCFS (First Come First Served) : les clients sont servis dans leur ordre d'arrivée. Notons que les discipline FIFO et FCFS ne sont pas équivalentes lorsque la file contient plusieurs serveurs. Dans la première, le premier arrivé sera le premier à quitter la file, alors que dans la deuxième, il sera premier à commencer le service. Rien n'empêche alors qu'un client qui commence son service après lui, dans un autre serveur, termine avant lui.

b) LIFO (Last In First Out) ou LCFS (Last Come First Served) : Cela correspond à une pile, dans laquelle le dernier client arrivé (posé sur la pile) sera le premier traité (retiré de la pile). Comme pour le cas précédent, les disciplines LIFO et LCFS ne sont équivalentes que pour une file mono serveur.

c) RANDOM (Aléatoire) : le prochain client qui sera servi est choisi aléatoirement dans la file d'attente.

Remarque : lorsque les trois derniers symboles de la notation ne sont pas précisés, il est sous-entendu que $K = +\infty$, $m = +\infty$ et Z = FIFO.

Les principaux paramètres de performance des files d'attente qui sont étudiées sont :

- $\overline{N}_s = E(N_s)$: nombre moyen de clients dans le système (file + service).
- $\overline{N}_q = E(N_q)$: nombre moyen de clients dans la file (hors service).
- $\overline{W}_{s(q)} = E(W_{s(q)})$: temps moyen d'attente dans le système (la file).

A.2. Etude de quelques modèles classiques

A.2.1. Modèles markoviens :

Les modèles markoviens de files d'attente sont des systèmes où les temps des inter arrivées et les temps de service sont des variables aléatoires indépendantes

exponentiellement distribuées : la propriété *sans mémoire* (markovienne) de la loi exponentielle rend aisée l'analyse de ce type de modèles.

A.2.1.1. La file M/M/1

C'est l'exemple le plus simple qui modélise le processus d'attente devant un serveur quand les arrivées sont poissoniennes de taux λ (nombre moyen de clients arrivant pendant une unité de temps) et le service est exponentiel de taux μ (nombre de clients servis pendant une unité de temps) et de temps moyen de service $1/\mu$.

Ces deux processus, qui font évoluer le nombre de clients étant markoviens, le processus $(X(t)_{t\geq 0})$ " *nombre de clients dans le système à l'instant t* " est donc markovien.

Dans le régime transitoire, les équations différentielles de Chapman-Kolmogorov sont de la forme :

$$\begin{cases} \dfrac{dp_0(t)}{dt} = \mu\, p_1(t) - (\lambda+\mu) p_0(t) \\ \dfrac{dp_n(t)}{dt} = \lambda\, p_{n-1}(t) + \mu\, p_{n+1}(t) - (\lambda+\mu) p_n(t) \end{cases} \qquad \text{(I.1)}$$

où $p_n(t) = p(X(t) = n)$.

Dans le régime stationnaire, lorsque t tend vers l'infini, le système d'équations (I.1) permet d'obtenir un système d'équations plus simple :

$$\lambda p_n = \mu\, p_{n+1} \qquad \text{(I.2)}$$

Ce qui permet de déduire la solution :

$$p_n = \rho^n\, p_0 \qquad \text{(I.3)}$$

avec

$$p_0 = \frac{1}{\sum_{i=1}^{\infty} \rho^i} = 1 - \rho \qquad \text{(I.4)}$$

où $\rho = \frac{\lambda}{\mu}$ est appelé *"l'intensité de trafic"* (ou taux de charge) de la file.

L'existence de la solution (la condition d'ergodicité) est liée à la convergence de la série figurant au dénominateur de l'expression (I.4). On a donc :

$\lambda < \mu$ ou $\rho < 1$: la file est stable.
$\lambda \geq \mu$ ou $\rho \geq 1$: la file est instable.

Les paramètres de performance de la file M/M/1 sont :

- $\overline{N}_s = \sum_{n=0}^{\infty} n\, p_n = \frac{\rho}{1-\rho}$ (I.5)

- $\overline{W}_s = \frac{1}{\lambda} \cdot \frac{\rho}{1-\rho}$ (I.6)

Ce dernier résultat est obtenu par l'application de la formule de Little (voir annexe A.1).

- $\overline{W}_q = \overline{W}_s - \frac{1}{\mu} = \frac{\rho^2}{\lambda(1-\rho)} = \frac{\rho}{(\mu - \lambda)}$ (I.7)

- $\overline{N}_q = \lambda \overline{W}_q = \frac{\rho^2}{\lambda(1-\rho)} = \frac{\lambda \rho}{\mu - \lambda}$ (I.8)

La file d'attente M/M/1 est largement utilisée dans la modélisation de systèmes pour plusieurs raisons :

- elle est extrêmement simple à traiter avec un ensemble de propriétés facilement exprimables par des formules aisées à manipuler. Elle modélise d'une manière générale tout système de type guichet.

- Les hypothèses qu'elle utilise (processus markoviens en entrée et en sortie) sont classiquement utilisées dans la modélisation de divers systèmes, en l'absence de caractérisation plus précise et justifiée des paramètres.

A.2.1.2. Variantes des files M/M/1

A.2.1.2.1. Politiques de services autres que FIFO

Dans la file M/M/1 classique, la politique de service est supposée, par défaut, premier arrivé- premier sorti (FIFO). En fait, quelque soit la politique de service (FIFO, LIFO,…), le nombre de clients dans la file d'attente (service compris ou non) ne varie pas. Les équations d'équilibre et les statistiques moyennes (nombre moyen, temps moyen d'attente) ne sont pas modifiées. Par contre, la distribution des temps d'attente varie fortement car elle dépend des événements qui se produiront après l'arrivée (au lieu de l'état de la file au moment d'une arrivée dans la politique FIFO) comme pour l'exemple d'une file M/M/1 avec politique de service LIFO avec préemption (le client qui arrive prend le serveur en interrompant le service en cours). Quand un client arrive, son temps d'attente ne dépend pas du nombre de clients présents, et son temps de présence dans la file sera égal à la période d'activité (busy period) du serveur.

A.2.1.2.2. Capacités limitées

Dans la pratique, et pour prendre en compte la capacité limitée des systèmes (mémoire limitée des machines, guichets de service,..), on est souvent conduit à prendre l'hypothèse de file d'attente à capacité limitée du type M/M/1/N.

Dans ce cas, un client qui arrive alors que le système est plein est rejeté. D'une manière équivalente, on peut dire que le taux d'arrivée dépend de l'état de la file : il vaut λ si n est strictement inférieur à N, et 0 sinon.

La théorie générale des processus de naissance et de mort (voir annexe A.2) applicable à ce système permet d'accéder aux paramètres de performance :

$$\overline{N_s} = \frac{\rho\left[1-(N+1)\rho^N + N\rho^{N+1}\right]}{(1-\rho)(1-\rho^{N+1})}$$
$$\overline{W_s} = \frac{\rho\left[1-(N+1)\rho^N + N\rho^{N+1}\right]}{\lambda(1-\rho)(1-\rho^N)} \qquad \text{(I.9)}$$

Quand $N \to \infty$ (capacité illimitée), on aboutit aux mêmes paramètres (I.5) et (I.6) obtenus dans la file M/M/1.

A.2.2. Modèles non markoviens

En s'écartant de l'hypothèse d'exponentialité des deux quantités stochastiques (temps des inter arrivées et durée de service) ou en introduisant des paramètres supplémentaires spécifiques au modèle étudié, on sera alors en présence d'un processus non markovien. Ceci rend donc l'analyse du modèle très délicate, voire impossible. On se ramène alors à un processus markovien judicieusement choisi grâce à différentes méthodes **[KEN 1953 ; GAV 1959 ; DSH 1995 ; BAY 1990 ; KUM 1994]**.

A.2.2.1. Système M/G/1

Si la loi de service n'est plus exponentielle, le système perd ses propriétés markoviennes et ne peut plus être résolu par la théorie des processus de naissance et de mort. Sauf cas particulier où la loi de service a un comportement modélisable par un

processus markovien (par exemple par décomposition en étapes d'Erlang ou des lois hyperexponentielles), la théorie des processus de Markov ne peut plus s'appliquer. On est donc réduit à trouver des méthodes de calcul différentes.

L'une de ces méthodes est celle dite *"des variables auxiliaires"* qui s'applique aux systèmes M/G/1 en complétant l'information sur X(t) par la variable Y(t) qui représente le temps de service déjà écoulé à l'instant t. Le calcul du régime transitoire d'un tel processus fait intervenir des équations aux dérivées partielles.

Pour éviter cela, la méthode dite de *"la chaîne de Markov incluse"* (CMI) ramène l'étude du processus bidimensionnel précédent à celle d'un processus unidimensionnel $X_n = X(t_n)$ vérifiant la propriété de Markov : on considère les instants t_n de départ du $n^{ième}$ client où le processus $\{X(t_n), Y(t_n)\}$ sera identique au processus X_n = *"nombre de client dans le système juste après l'instant t_n"*, car $Y(t_n) = 0$, qui constitue une chaîne de Markov à temps discret. Celle-ci est apériodique et irréductible, les probabilités de transitions étant données par la matrice :

$$P(i,j) = \begin{pmatrix} q_0 & q_1 & q_2 & q_3 & \cdots & \cdots \\ q_0 & q_1 & q_2 & q_3 & \cdots & \cdots \\ 0 & q_0 & q_1 & q_2 & \cdots & \cdots \\ 0 & 0 & q_0 & q_1 & q_2 & \cdots \\ \vdots & \vdots & 0 & q_0 & q_1 & \cdots \\ \vdots & \vdots & \vdots & 0 & q_0 & \cdots \end{pmatrix} \qquad \text{(I.10)}$$

où les q_i représentent la probabilité qu'il y ait eu i arrivées pendant une durée de service. En considérant B(x) la fonction de distribution de la loi de service, q_i s'écrit :

$$q_i = \int_0^{+\infty} \frac{(\lambda x)^i}{i!} \exp(-\lambda x)\, dB(x) \qquad \text{(I.11)}$$

On peut ainsi établir le système des équations reliant les π_i, probabilités d'état de la chaîne incluse :

$$\pi_i = \pi_0 \, q_i + \sum_{j=0}^{i} q_j \, \pi_{i-j+1} \qquad \text{(I.12)}$$

La résolution de ce système se fera par l'introduction de la fonction génératrice $f(z) = \sum_{i=0}^{\infty} \pi_i \, z^i$. On obtient alors l'expression de f(z) donnée par **[BAY 2000]** :

$$f(z) = \pi_0 \, \frac{(1-z)\, B^*(\lambda - \lambda z)}{B^*(\lambda - \lambda z) - z} \qquad \text{(I.13)}$$

où $B^*(s)$ est la transformée de Laplace de la fonction B.

Pour finir, il est utile de rappeler une propriété qui ne sera pas démontrée ici : l'état de la file au moment d'un départ est égal à la probabilité stationnaire. On en déduit la valeur de π_0 égale à 1-ρ. La résolution du système (I.12) est détaillée par B. Benameur dans sa thèse de Magister **[BEN 2002]** ainsi que dans tout ouvrage spécialisé dans ce type de systèmes **[SAK 1978 ; BAY 2000 ; …]**.

La formule (I.13), connue sous le nom de *"la formule des transformées de Pollaczek-Khinchine"*, permet de déduire les probabilités d'état. Par conséquent, l'expression du nombre moyen de clients en fonction des deux premiers moments de la loi G (ou de sa moyenne $m = 1/\mu$ et de son coefficient de variation $Cs^2 = var / m^2$) est donnée par :

$$\overline{N}_s = \rho + \frac{\rho^2 (1 + Cs^2)}{2(1-\rho)} \qquad \text{(I.14)}$$

connue sous le nom de *"formule de Pollaczek-Khinchine"*. A partir de cette formule, et en utilisant la formule de Little, on déduit le temps moyen d'attente :

$$\overline{W}_s = \frac{\rho}{\lambda} + \frac{\rho^2 (1 + Cs^2)}{2\,\lambda(1-\rho)} \qquad \text{(I.15)}$$

A.2.2.2. Système G/M/1

Le système de file d'attente G/M/1 peut, en fait, être vu comme *dual* du système M/G/1 puisque les temps d'inter arrivées et de service sont respectivement généraux et exponentiels de taux μ pour le premier, exponentiels de taux λ et généraux pour le second. De ce fait, le système G/M/1 sera décrit par un processus bidimensionnel $\{X(t), Z(t)\}$, où $Z(t)$ est le temps déjà écoulé depuis l'arrivée du dernier client, qui est sans mémoire. En se basant sur le même principe de la CMI, l'idée est de ne s'intéresser qu'aux instants d'arrivée des clients afin d'éliminer l'information $Z(t)$. Le système est alors décrit par un processus unidimensionnel $X_n = X(t_n)$ = *"nombre de clients dans le système juste avant l'arrivée du* $n^{ième}$ *client*" (Chaîne de Markov à temps discret). Les probabilités de transitions sont données par la matrice :

$$P(i,j) = \begin{pmatrix} \beta_1 & \beta_0 & 0 & 0 & \dots & \dots \\ \beta_2 & \beta_1 & \beta_0 & 0 & 0 & \dots \\ \beta_3 & \beta_2 & \beta_1 & \beta_0 & 0 & \dots \\ \beta_4 & \beta_3 & \beta_2 & \beta_1 & \ddots & \ddots \\ \vdots & \vdots & \beta_3 & \beta_2 & \ddots & \ddots \\ \vdots & \vdots & \vdots & \ddots & \ddots & \ddots \end{pmatrix} \qquad \text{(I.16)}$$

où les β_i représentent la probabilité pour que i clients terminent leur service pendant un temps d'inter arrivée. En considérant $F(x)$ la fonction de distribution de la loi inter arrivée, β_i s'écrit :

$$\beta_i = \int_0^{+\infty} \frac{(\mu x)^i}{i!} \exp(-\mu x)\, dF(x) \qquad \text{(I.17)}$$

Cette chaîne de Markov est clairement irréductible et apériodique. Si, de plus, on suppose que $\rho = \frac{\lambda}{\mu} < 1$ (la file est stable) où $1/\lambda$ est la moyenne d'arrivée des clients

dans la file, alors elle est aussi récurrente positive. En conséquence, le vecteur des probabilités stationnaires existe et est solution du système qui s'écrit sous la forme :

$$\pi_k = \sum_{i=k-1}^{\infty} \pi_i \, \beta_{i-k+1} \qquad k = 0, 1, \ldots \tag{I.18}$$

La solution de ce système est donnée par **[BAY 2000]**:

$$\pi_n = (1-\sigma)\,\sigma^n \tag{I.19}$$

où σ est l'unique solution strictement comprise entre 0 et 1 de l'équation :

$$\sigma = F^*(\mu - \mu\,\sigma) \tag{I.20}$$

avec $F^*(s) = \dfrac{\lambda}{\lambda + s}$ la transformée de Laplace de la distribution inter arrivée.

A partir des probabilités stationnaires, on peut facilement calculer le temps moyen d'attente dans le système :

$$\overline{W}_s = \frac{1}{\mu} + \frac{\sigma}{\mu\,(1-\sigma)} \tag{I.21}$$

et, par la formule de Little, le nombre moyen de clients dans le système :

$$\overline{N}_s = \frac{\lambda}{\mu} + \frac{\lambda\,\sigma}{\mu\,(1-\sigma)} \tag{I.22}$$

A.2.2.3 Système GI/GI/1

Jusque là, on a eu à faire à des systèmes markoviens ou semi-markoviens. Pour prédire l'évolution du système GI/GI/1, on a besoin, cette fois, de connaître, en plus du nombre de clients dans le système X(t), le temps déjà passé dans le serveur par le client

en service Y(t) et le temps déjà écoulé depuis l'arrivée du dernier client Z(t). Le processus {X(t), Y(t), Z(t)} est donc un processus sans mémoire, sauf qu'il s'agit d'un processus dont l'espace d'état est mixte et donc très complexe à analyser. Contrairement aux files précédentes, on n'est pas capable ici d'extraire un processus markovien (sans mémoire) simple en examinant le système à des instants choisis (aucun instant d'observation particulier ne permet d'éliminer simultanément les variables Y(t) et Z(t)). L'une des conséquences est qu'il n'existe pas pour cette file de résultats analytiques exacts tels que le nombre moyen de client ou le temps moyen d'attente en régime stationnaire.

L'attention des spécialistes est actuellement focalisée sur l'étude de méthodes d'approximation parmi lesquelles on peut distinguer :

a) *Les méthodes numériques :* largement étudiées, particulièrement pour les systèmes markoviens. Des approches à temps discret ont été proposées notamment pour l'analyse du régime transitoire **[BEN 1995]**. Ces dernières présentent, cependant, l'inconvénient de temps d'implantation et d'exécution du programme assez longs.

b) *Les méthodes algorithmiques ou itératives :* généralement utilisées pour les systèmes markoviens par Wilkinson **[WIL 1968]**, Hashida et Kawashima **[HAS 1979]**…

c) *Les méthodes d'approximation par les processus de diffusion :* basées sur le principe d'approximation de la densité de probabilité de X(t) par la densité de probabilité *f* d'une loi normale (dont les paramètres seront déterminés par differentes approches) qui représente une solution de l'équation de diffusion de Fokker-Planck . Des barrières dites *réfléchissantes* ou *absorbantes* sont introduites pour prendre en considération le fait que $X(t) \geq 0$ **[KLE 1976]** (voir annexe B).

d) *Les méthodes heuristiques :* qui, bien que sujettes à controverse car dépendant plus de l'intuition et de la créativité, sont utilisées pour des modèles complexes où il

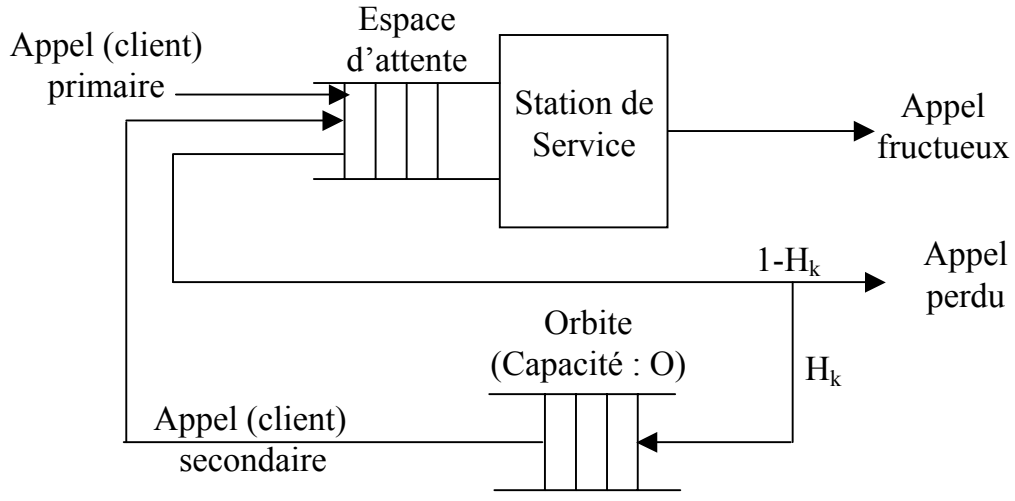
n'existe aucun résultat scientifique dans le cadre des différentes approximations existantes **[KIM 1986, 1991a, 1991b, 1992]**.

e) *Les méthodes de simulation :* sont devenues, malgré tout ce qu'elles peuvent engendrer comme inconvénients, l'outil majeur de la modélisation des systèmes GI/GI/1 en l'absence de résultats exacts et où la majorité des autres méthodes fait généralement défaut. Elles permettent de traiter par programme informatique la suite des événements susceptibles de se produire sur un système réel et de représenter l'évolution dans le temps des variables de ce système. Cette technique a l'avantage d'être infiniment moins limitée que la technique analytique quant aux hypothèses de fonctionnement, mais ceci sera nécessairement au coût d'un temps de calcul (et donc d'un temps de mise au point de programme) parfois assez fastidieux.

B. Files d'attente avec rappel

B. 1. Description du modèle

Un système de files d'attente avec rappel est un système classique dans lequel existe une *"orbite"* de capacité bien déterminée. Le client qui arrive peut trouver un ou plusieurs serveurs libres et être immédiatement pris en charge. Sinon, il rejoint la file d'attente s'il y'a une position libre. Dans le cas contraire, le client quitte le système définitivement avec une probabilité $(1-H_0)$ ou bien entre en orbite avec une probabilité H_0 pour rappeler ultérieurement après un temps aléatoire. Les clients qui rappelleront pour le service sont dits en orbite. Celle-ci peut être de capacité finie ou infinie. Dans le cas où elle est finie et pleine, le client qui trouve tous les serveurs et les positions d'attente de la file occupés, sera obligé de quitter le système définitivement sans être servi. Chaque client de l'orbite forme un processus *"d'arrivées secondaires"* de taux θ et est traité de la même manière qu'un *"client primaire"* qui arrive, trouve un serveur ou une position libre dans la file, ou quitte le système avec une probabilité $(1-H_k)$ si tous les serveurs et les positions d'attente sont occupés ou rejoint l'orbite (si elle n'est pas pleine) avec une probabilité H_k (H_k étant la probabilité de faire une $(k+1)^{ième}$ tentative de rappel après une $k^{ième}$ tentative infructueuse). Le schéma général d'un système de files d'attente avec rappel est donné par la figure I.2.



<u>Figure I.2</u> : Schéma descriptif d'une file d'attente avec rappel.

Terminologie et notation

Par analogie à la notation établie pour la file d'attente classique, celle du modèle de file d'attente avec rappel s'écrit en ajoutant deux autres symboles aux six symboles précédemment définis :

$$A / B / S / K / m / Z / O / H$$

Avec O : capacité de l'orbite.

H : fonction de persévérance.

Où H permet de définir le comportement du client devant une situation de blocage (serveurs occupés). Elle peut être décrite par une séquence $\{H_0, H_1, \ldots, H_k, \ldots\}$ de probabilités de rappel après un échec.

a) Quand $H_k = 1$ pour $k \geq 0$, le système devient *un système sans pertes*. Ainsi, chaque client reçoit éventuellement le service si O est infini. Dans ce cas, H = NL (*No Loss*).

b) Quand $H_k = \alpha$ ($\alpha < 1$) pour $k \geq 0$, le système est dit *un système à perte géométrique* et H = GL (Geometric *Loss*).

<u>Remarques</u> :

a) Si la capacité de l'orbite est infinie, elle est omise dans la notation. H est aussi omise dans le cas d'un système sans perte.

b) La distribution des temps de rappel est supposée généralement exponentielle de taux θ, $1/\theta$ étant la durée moyenne des intervalles de rappel. C'est la raison pour laquelle elle est omise dans la notation.

c) Lorsque $\theta \to \infty$, le système d'attente avec rappels se rapporte à un système d'attente classique, le temps de rappel étant nul. Ainsi, tout appel primaire bloqué (qui

trouve tous les serveurs occupés) restera dans le système jusqu'à ce qu'un serveur soit libéré.

d) Lorsque $\theta \to 0$, le système d'attente avec rappel est un système d'Erlang avec perte **[COH 1957]**. Cette hypothèse signifie que l'intervalle de temps entre deux rappels successifs du même client est infini. Ainsi, tout appel primaire bloqué est automatiquement perdu.

e) Lorsque la capacité de l'orbite est nulle (O = 0), il n'y a aucune position d'attente ou de rappel. Ainsi, tout appel (primaire) bloqué est perdu pour toujours. Ce modèle correspond aussi au système d'Erlang avec perte **[BRO 1948]**. Il peut être considéré comme un modèle dans lequel la probabilité de rappel $H_k = 0$ ou dans lequel le taux de rappel $\theta \to 0$ **[COH 1957]**.

On peut donc dire que le modèle de file d'attente avec rappel occupe une situation intermédiaire entre le modèle d'Erlang avec perte est le modèle classique avec la discipline FIFO, dans le cas de faible et de forte intensité de rappel respectivement.

B.2. Quelques exemples modélisés par des systèmes de files d'attente avec rappel

Il existe aujourd'hui des centaines de publications sur les systèmes avec rappel où des exemples concrets ont été cités **[YAN 1987 ; FAL 1990 ; AIS 1994, 1996]** en rapport avec les nouveaux développements technologiques dont l'intérêt porté s'accroît de jour en jour.

Dans cette section, nous présentons quelques exemples de problèmes (extraits de **[YAN 1987]**) pouvant être modélisés avec ces systèmes. Ceux-ci vont du cas le plus simple de réservation à d'autres cas plus complexes comme les systèmes informatiques à temps réel.

B.2.1. Problème de réservation

C'est l'exemple le plus simple d'un client qui sollicite une réservation par téléphone dans un restaurant. Il y'a une ligne unique qui est consacrée à répondre aux requêtes des réservations. Ainsi, si un client appelle et trouve la ligne occupée, il renouvellera sa tentative après une certaine période de temps aléatoire avec la probabilité H_k qui, en pratique, est strictement inférieure à 1 car le client ne peut rappeler indéfiniment.

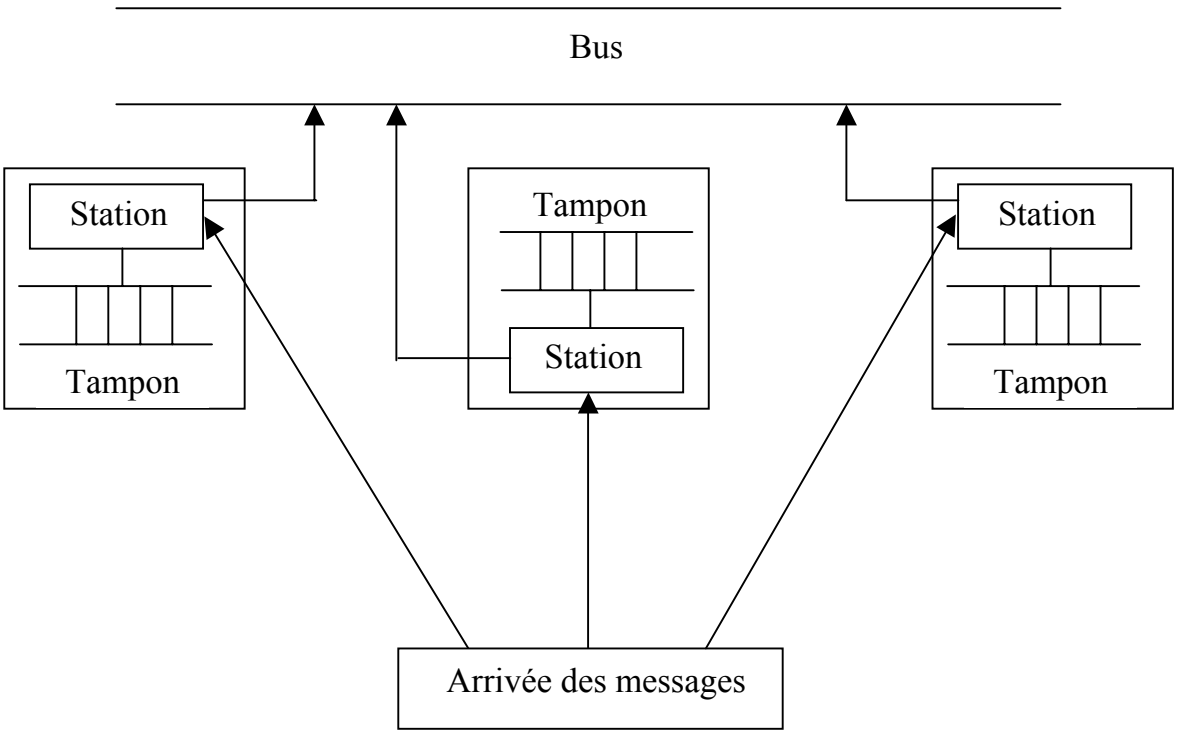
Cet exemple peut être modélisé par une file d'attente M/G/1 avec rappel et avec perte en considérant que le processus d'arrivée des appels est poissonnien. L'étude de ce genre de problème permet de prédire le temps d'attente du client, le nombre de clients perdus du à ce blocage, …

B.2.2. Réseaux locaux CSMA

Le modèle de file d'attente avec rappel à un seul serveur peut être utilisé pour la modélisation des réseaux d'ordinateurs locaux **[RUS 1984]**.

Un réseau local simple est composé de stations ou terminaux interconnectés par un bus unique qui constitue le canal de communication. Les stations communiquent entre elles via le bus qui ne peut être utilisé que par une seule station. Une telle architecture de réseau d'ordinateurs local est appelée *"architecture en bus"*. Un des protocoles de communication le plus généralement utilisé dans les réseaux locaux est celui dit *"Non Persistant CSMA (Carrier Sense Multiple Access)"* : des messages (de longueurs variables) arrivent de l'extérieur vers une station donnée, celle-ci les divise en un certain nombre de paquets (de longueurs fixes) et vérifie immédiatement la disponibilité du bus. Si ce dernier est libre, un des paquets sera alors transmis à la station de destination. Les autres paquets restants seront stockés dans le tampon (orbite) pour une prochaine transmission. Si, par contre, le bus est occupé, tous les paquets seront alors stockés dans le tampon et la station tentera la connexion (transmission) après un certain temps aléatoire.

La modélisation de ce type de problème permet de prévoir le temps moyen d'attente du paquet (le client), le nombre moyen de clients dans le tampon de la station. Le système de file d'attente considéré sera alors de type G/G/1 avec rappel : le serveur est le bus et l'espace du tampon dans toutes les stations est l'orbite. Si les paquets sont envoyés avec un flux poissonnien, le système devient de type M/G/1. Ce système est décrit dans la figure I.3.



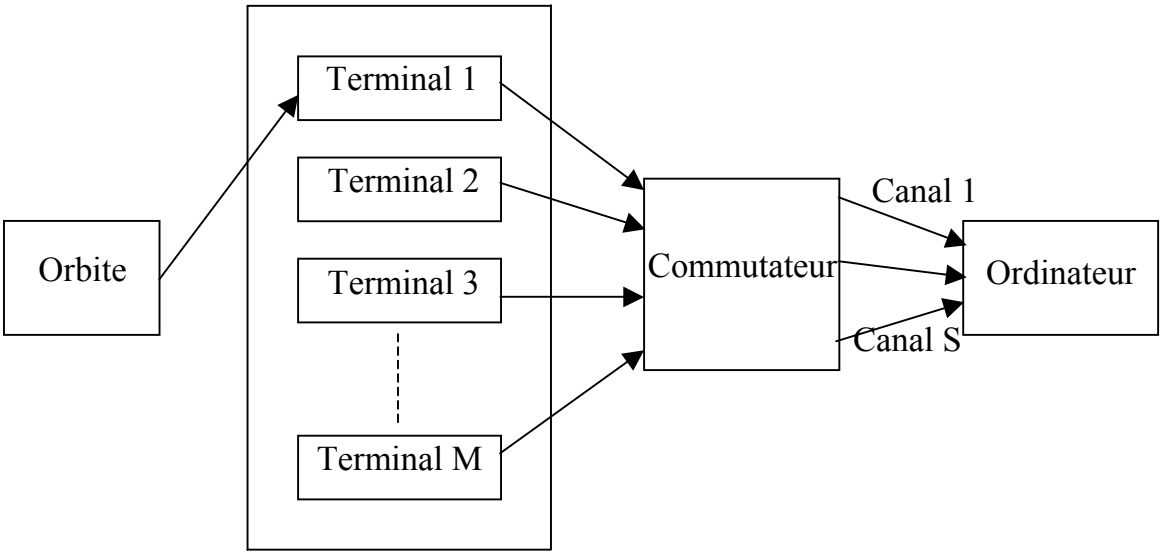
<u>Figure I.3</u> : Schéma d'un réseau local.

B.2.3. Système informatique à temps réel

Dans un système informatique à temps réel, on trouve M terminaux et S canaux de transmission tels que M > S. Pour qu'un terminal soit connecté à l'ordinateur, il suffit d'un canal de transmission libre. L'illustration de ce genre de système est le centre de calcul où arrive un étudiant pour utiliser l'ordinateur pendant une période de temps aléatoire. Celui-ci doit d'abord trouver un terminal libre pour se connecter. S'il n'y a aucun terminal disponible, il retentera sa chance après un temps aléatoire. Sinon,

il envoie sa demande au commutateur central pour se connecter à l'ordinateur. Le terminal est alors connecté selon que le canal serait disponible ou pas. Dans ce dernier cas, la demande est mise dans la file par le commutateur en attente de libération d'un canal.

Ce système peut être modélisé par une file d'attente G/G/S avec rappel, avec un tampon (espace d'attente) de capacité M et une orbite de taille infinie, où les canaux de transmission correspondent aux serveurs et les terminaux au tampon. Un schéma descriptif est donné par la figure I.4.



<u>Figure I.4</u> : Schéma d'un système informatique à temps réel.

B.3. Système M/G/1 avec rappel

Les premiers travaux avec des résultats analytiques concrets sur les files d'attente M/G/1 avec rappel sont ceux entrepris par Keilson **[KEI 1968]**, alors que la majorité des travaux précédents ne traitaient que des files M/M/s avec rappel **[WIL 1956 ; COH 1957 ; RIO 1962 ; ELL 1967]**. Dans **[KEI 1968]**, et par la suite dans **[ALE 1974]**. Des solutions pour la fonction génératrice du nombre de clients en orbite, ainsi que le temps d'attente moyen d'un client ont été obtenues. Quelques années plus

tard, ces résultats ont été étendus aux distributions du temps d'attente, temps d'oisiveté du système, de la période d'activité du système, …**[CHO 1979 ; FAL 1978, 1979]**. En résumé, des résultats analytiques très intéressants sont obtenus pour le modèle M/G/1 avec rappel **[YAN 1987 ; FAL 1990]**.

Dans ce système, le flux des arrivées primaires est poissonnien de taux λ, la durée de service suit une loi générale de distribution B de moyenne $1/\mu$ et la durée entre deux rappels successifs est exponentielle de paramètre θ.

Comme pour le cas classique, le système (semi markovien) est décrit par le processus bidimensionnel $\{X(t), Y(t)\}$ tel que :

$X(t) = \{N(t), C(t)\}$ avec $N(t)$ le nombre de clients en régime de rappel à l'instant t et $C(t)$ le nombre de clients en cours de service au même instant t.

$Y(t)$ est la durée de service écoulée à l'instant t.

Par la méthode de la CMI, le processus devient unidimensionnel discret par élimination de la variable continue $Y(t)$ en considérant les instants t_n de départ du $n^{ième}$ client $(Y(t_n) = 0)$.

Les probabilités de transition de la chaîne de Markov $X(t_n)$ sont données par :

$$P_{i,j} = \frac{i\theta}{\lambda + i\theta} q_{j-i+1} + \frac{\lambda}{\lambda + i\theta} q_{j-i} \qquad \text{(I.23)}$$

où q_i est donnée par (I.11).

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

$$Q(z) = f(z)\frac{\phi(z)}{\phi(1)} \qquad \text{(I.24)}$$

où f(z) est la fonction génératrice dans le cas classique, donnée par (I.13) et :

$$\phi(z) = \exp\left\{\frac{-\lambda}{\theta}\int_0^z \frac{1-B^*(\lambda-\lambda x)}{x-B^*(\lambda-\lambda x)}dx\right\} \qquad (I.25)$$

Les paramètres de performance sont :

a) Nombre moyen de clients dans le système :

$$\overline{N}_s = \rho + \frac{\rho^2(1+Cs^2)}{2(1-\rho)} + \frac{\lambda\rho}{\theta(1-\rho)} \qquad (I.26)$$

b) Nombre moyen de clients dans l'orbite :

$$\overline{N}_O = \overline{N}_s - \rho = \frac{\rho^2(1+Cs^2)}{2(1-\rho)} + \frac{\lambda\rho}{\theta(1-\rho)} \qquad (I.27)$$

c) Temps moyen d'attente dans le système :

$$\overline{W}_s = \frac{\rho}{\lambda} + \frac{\rho^2(1+Cs^2)}{2\lambda(1-\rho)} + \frac{\rho}{\theta(1-\rho)} \qquad (I.28)$$

d) Temps moyen d'attente dans la file (orbite) :

$$\overline{W}_q = \frac{\rho^2(1+Cs^2)}{2\lambda(1-\rho)} + \frac{\rho}{\theta(1-\rho)} \qquad (I.29)$$

On remarquera que le nombre moyen de clients (temps moyen d'attente) est la contribution de deux termes : le premier caractérisant la file classique de fonction génératrice f (z) et le second caractérisant l'influence des rappels de fonction génératrice $\frac{\phi(z)}{\phi(1)}$.

B.4. Autres modèles de systèmes avec rappel

B.4.1. Modèle d'attente avec des clients persistants

Le modèle avec clients persistants est un modèle infini sans perte où tout client ne peut quitter le système définitivement que s'il est servi. Dans ce cas, la fonction de persévérance H_k ($k \geq 1$) est toujours égale à 1.

Le système est stable si la condition $\rho < 1$ est vérifiée. Dans ce cas, le nombre moyen de clients dans l'orbite est donné par **[FAL 1990]** :

$$\overline{N}_O = \frac{\lambda^2}{1-\rho}\left(\frac{\beta_1}{\theta} + \frac{\beta_2}{2}\right) \qquad \text{(I.30)}$$

où :

$\beta_k = (-1)^k \beta^{(k)}(0)$ est le moment d'ordre k de la distribution des temps de service B(x).

$\beta(s) = \int_0^\infty \exp(-s\,x)\,dB(x)$ est la transformée de Laplace-Stieltjes de B(x).

Dans le système M/G/1 avec rappel, l'égalité (I.30) est équivalente à l'égalité (I.27).

B.4.2. Modèle d'attente avec des clients non persistants (impatients)

Ce modèle est le plus réaliste qu'on rencontre notamment dans les réseaux téléphoniques où le client qui rappelle après un certain nombre de tentatives décide d'y renoncer. Ceci se traduit par une fonction de persistance H_k ($k \geq 1$) ≤ 1. Il est admis que la probabilité de rappel ne dépend pas du nombre de tentatives précédentes (i.e $H_2 = H_3 = \ldots$) **[FAL 1990]**. Cependant, les cas $H_2 < 1$ et $H_2 = 1$ conduisent à des solutions différentes du problème :

a) Cas $H_2 = 1$:

Si $\rho H_1 < 1$ et la file est dans son état stable, alors selon Lubacz et Roberts **[LUB 1984]** on a le nombre moyen de clients en orbite donné par :

$$\overline{N}_O = \frac{\lambda^2 H_1}{1-\rho H_1}\left(\frac{\beta_1}{\theta} + \frac{\beta_2}{2(1+\rho(1-H_1))}\right) \qquad (I.31)$$

b) cas $H_2 < 1$:

Ce cas est très compliqué et on ne dispose de résultats que pour des cas particuliers où la distribution des temps de service est exponentielle **[FAL 1980]**.

Dans le régime stationnaire (qui pour le cas $H_2 < 1$ existe toujours), le nombre moyen de clients dans la file est :

$$\overline{N}_O = \frac{\lambda H_2 + (\lambda H_1 - \mu H_2)\Lambda}{\theta(1-H_2)(1+\Lambda)} \qquad (I.32)$$

où : $$\Lambda = \rho \frac{\phi(a+1, c, \gamma)}{\phi(a, c, \gamma)} \qquad (I.33)$$

et $\phi(a,c,\gamma) = \sum_{n=0}^{\infty} \frac{\gamma^n}{n!} \prod_{i=0}^{n-1} \frac{a+i}{c+i}$ est la fonction hypergéométrique à trois paramètres :

$$a = \frac{\lambda}{\theta} \quad , \quad c = \frac{\mu + (1-H_2)(\lambda+\theta)}{(1-H_2)\theta} \quad \text{et} \quad \gamma = \frac{\lambda H_1}{(1-H_2)\theta}$$

Dans le cas de fonction de distribution générale B(x), différentes approximations ont été proposées telles la généralisation du modèle classique avec des clients impatients qui quittent le système avec un certain taux δ **[COH 1957 ; WIL 1968]** ou l'utilisation d'une transformation algébrique simple **[GRE 1987]**…

B.4.3. Modèle d'attente avec des temps de rappel généraux

Un modèle plus généralisé des files d'attente classique et avec rappel (exponentiel) est celui dans lequel les temps de rappel ou encore les temps entre rappels successifs du même client sont distribués selon une fonction générale.

L'analyse de ce type de modèle s'inspire de l'observation des phénomènes de rappel dans les systèmes informatiques, téléphoniques et les réseaux de télécommunication où les temps de rappels peuvent difficilement être modélisés par une distribution exponentielle.

La recherche dans ce domaine reste très limitée. Le premier à s'en intéresser fut Kapyrin **[KAP 1977]** qui a essayé de déduire une solution analytique exacte pour la file M/G/1 avec rappel général. Cette méthode se révéla incorrecte et ses résultats totalement erronés **[FAL 1986]**. Plus tard, Pourbabai **[POU 1987]** s'intéressa au sujet en traitant le modèle G/M/s/N avec rappel non exponentiel à l'aide de méthodes d'approximation. Quelques années plus tard, Choi, Park et Pearce **[CHO 1993]** ont considéré le système M/M/1 dans lequel le temps de rappel a une distribution générale et où seulement le client en tête de file est autorisé à rappeler pour le service. Une fonction génératrice de la distribution du nombre de clients dans la file ainsi que la transformée de Laplace de la distribution du temps d'attente moyen à l'état stationnaire ont été données en accord avec des résultats connus pour des cas particuliers (rappel exponentiel). Yang, Posner, Templeton et Li **[YAN 1994]** ont étudié le système M/G/1 avec rappel général en considérant la propriété de décomposition stochastique (le nombre de clients dans le système est la contribution de deux variables aléatoires indépendantes : le nombre de clients dans la file classique M/G/1 et le nombre de clients dans la file M/G/1 avec rappel exponentiel sachant que le serveur est libre) **[YAN 1987 ; ART 1994]** pour proposer une méthode d'approximation comparable aux résultats numériques de certains modèles avec rappel exponentiel.

B. Approximation par interpolation des systèmes GI/GI/s

L'approche des deux moments par interpolation des systèmes developpée dans cette section appartient au type heuristique. Cette méthode a été largement utilisée pour les systèmes classiques de files d'attente **[BJÖ 1964 ; COS 1974, 1976, 1977 ; BOX 1979 ; TIJ 1986]**. La formulation par interpolation des systèmes présente un grand intérêt pratique dans les problèmes de conception et de décision (design, pour reprendre la terminologie anglaise plus significative), d'abord en raison de sa simplicité, mais surtout parce que dans les situations réelles, la seule information disponible est celle concernant les deux premiers moments.

Son utilisation pour l'approximation des systèmes GI/GI/s est utile, non seulement pour l'analyse des files individuelles GI/GI/s mais également pour la conception et / ou l'évaluation de systèmes plus complexes tels les réseaux non markoviens de files d'attente qui se prêtent mal aux analyses classiques (numériques ou autres). Cette approche a été intégrée dans des logiciels ou langages de simulation tels que le *Q. N. A* (Queueing Network Analyser) qui a été développé pour calculer les approximations des paramètres de performance des réseaux de files d'attente **[WHI 1983a, 1983b ; KIM 1984]**.

L'interpolation peut-être linéaire ou harmonique, mais peut également mettre en œuvre une base d'approximation pouvant contenir plusieurs systèmes connus. Cette approche permet, d'une part, de retrouver la méthode des deux moments en tant que cas particuliers et, d'autre part, de justifier plus rigoureusement les arguments heuristiques à l'origine de leur usage. La méthode d'approximation par interpolation des deux moments a fait l'objet de diverses extensions. Nous présentons dans cette section les principaux résultats obtenus par Kimura **[KIM 1994]** qui donna une revue unifiée de la méthode d'interpolation des deux moments pour les systèmes GI/GI/s en commençant par présenter la méthodologie utilisée qui sera appliquée au cas particulier M/G/s avant d'être étendue au système GI/GI/s.

B.1. Méthodologie de l'interpolation

L'interpolation heuristique du système est introduite pour approximer le temps moyen d'attente dans une file GI/GI/s (avec s serveurs homogènes en parallèle, un espace d'attente illimité et une politique de service du type FCFS) par une combinaison des temps d'attente de files plus simples.

Soit à considérer ξ et τ les variables génériques des temps d'inter arrivées et de service respectivement, et F et B les lois de distribution correspondantes :

$F(x) = P[\xi \le x]$ de moyenne λ^{-1} et $B(x) = P[\tau \le x]$ de moyenne μ^{-1}.

Soient Ca^2 (Cs^2) le coefficient de variation au carré de ξ (τ) et $\rho = \lambda/s\mu$ l'intensité du trafic de la file qu'on supposera strictement inférieure à 1 de telle sorte que le système soit stable et dans le régime stationnaire.

Pour obtenir des approximations convenables pour les caractéristiques de la file (le temps d'attente moyen dans la file $\overline{W}$ dans notre cas), on a souvent besoin de connaître leurs comportements dans les cas limites $s \to \infty$, $\rho \to 0$ ou $\rho \to 1$. Pour le temps moyen d'attente $\overline{W}$, il est évident que $\overline{W} \to 0$ quand $s \to \infty$ ou $\rho \to 0$ ou que $\overline{W} \to \infty$ quand $\rho \to 1$. Il est, en effet, bien admis que les interpolations entre les limites de trafic chargé et léger aboutissent à des approximations assez précises **[BUR 1983, 1986 ; REI 1988 ; WHI 1984a, 1984b.]**. Kimura propose une approche unifiée en considérant le quotient :

$$Q(GI/GI/s) \equiv \frac{\overline{W}(GI/GI/s)}{\overline{W}(GI/GI/1)} \qquad \text{(II.14)}$$

où le temps moyen d'attente est normalisé à celui d'une file à un seul serveur GI/GI/1 ayant la même distribution du temps de service, la même intensité de trafic et une distribution des temps inter-arrivées $F_s(t) \equiv F(t/s)$ $(t \ge 0)$.

La quantité Q (GI/GI/s) est approximée par une fonction des quantités correspondantes pour des systèmes analysables connus tels M/M/s, GI/M/s,… :

$$Q(GI/GI/s) \approx f\,(Q(\beta_1),\ldots,Q(\beta_n)) \qquad \text{(II.15)}$$

où l'ensemble des systèmes simples $B = \{\beta_1,\ldots,\beta_n\}$ constitue la base de l'approximation.

Il y a différentes façons de définir la fonction f mais deux d'entre elles seront utilisées dans le cadre de cette approche :

$$Q(GI/GI/s) \approx \sum_{i=1}^{n} l_i\, Q(\beta_i) \qquad \text{(II.16)}$$

$$Q(GI/GI/s) \approx \left\{\sum_{i=1}^{n} \frac{h_i}{Q(\beta_i)}\right\}^{-1} \qquad \text{(II.17)}$$

où $l_i \equiv l_i$ (GI/GI/s) et $h_i \equiv h_i$ (GI/GI/s) sont les coefficients de poids du $i^{ème}$ système (i= 1, …, n).

Les approximations (II.16) et (II.17) sont les combinaisons linéaires (de type L) et harmoniques (de type H) de la quantité Q(GI/GI/s) et sont l'extension naturelle des travaux de Cosmetatos **[COS 1974, 1976, 1977, 1980]** qui a développé la combinaison de type L pour l'approximation des temps moyens d'attente dans les systèmes GI/M/s, M/G/s, $E_m/E_k/s$ et $H_2/M/s$.

Pour une meilleure convenance analytique, nous supposerons que les coefficients l_i et h_i ne dépendent pas de l'intensité du trafic ρ. De plus, nous avons :

<u>Théorème II.1</u> [KIM 1994]

Les approximations (II.16) et (II.17) sont asymptotiquement correctes lorsque $\rho \to 1$ si :

$$\sum_{i=1}^{n} l_i = 1 \qquad \text{et} \qquad \sum_{i=1}^{n} h_i = 1 . \tag{II.18}$$

Preuve

D'après le théorème limite en régime chargé de Köllerstrom **[KÖL 1974]**, on a :

$$\lim_{\rho \to 1} (1-\rho)\, \overline{W}\,(GI/GI/s) = \frac{Ca^2 + Cs^2}{2s\mu} \tag{II.19}$$

qui nous permet de déduire que $\lim_{\rho \to 1} Q(.) = 1/s$.

Ainsi, en multipliant les deux cotés de (II.16) et (II.17) par (1-ρ) et en faisant tendre $\rho \to 1$, on aboutit aux résultats désirés.

□

Les approximations (II.16) et (II.17) s'écrivent :

$$\overline{W}\,(GI/GI/s) = \overline{W}\,(GI/GI/1) \sum_{i=1}^{n} l_i\, Q(\beta_i) \tag{II.20}$$

et

$$\overline{W}\,(GI/GI/s) = \overline{W}\,(GI/GI/1) \left\{ \sum_{i=1}^{n} \frac{h_i}{Q(\beta_i)} \right\}^{-1} \tag{II.21}$$

A partir de là, trois questions viennent automatiquement à l'esprit :

- Comment approximer $\overline{W}\,(GI/GI/1)$.
- Comment choisir la base propre pour ces approximations.
- Comment déterminer les coefficients des poids pour une base donnée.

Pour le cas M/G/s, la réponse à la première question est évidente du moment que la quantité $\overline{W}\,(M/G/1)$ est donnée de façon exacte par la formule de Pollaczek-Khintchine. Aussi est-il facile de choisir les candidats naturels pour la base

d'approximation. Nous montrerons dans le paragraphe suivant que la réponse à la troisième question peut se faire par une analyse asymptotique.

B.2. Temps moyen d'attente dans la file M/G/s

La base d'approximations la plus naturelle pour $\overline{W}(M/G/s)$ paraît être la base $B_1 \equiv \{M/M/s\}$ ou $B_2 \equiv \{M/M/s, M/D/s\}$ en accord avec les précédentes approximations déjà établies **[STO 1976 ; TAK 1977 ; HOK 1978 ; MIY 1986, KIM 1986, 1987, 1991a, 1992]**

Pour la base B_1, il est clair que les approximations (II.20) et (II.21) aboutissent à la même approximation :

$$\overline{W}(M/G/s) \approx \frac{1+Cs^2}{2}\overline{W}(M/M/s) \qquad \text{(II.22)}$$

et cela à l'aide de la formule de Pollaczek-Khintchine :

$$\overline{W}(M/G/1) \approx \frac{1+Cs^2}{2}\overline{W}(M/M/1) \qquad \text{(II.23)}$$

Nous obtenons ainsi la même approximation que celle obtenue par extension heuristique de (II.23) par Lee et Longton **[LEE 1957]** et bien d'autres approches différentes **[MAA 1973 ; STO 1976 ; SAK 1977 ; HOK 1978 ; MIY 1986]**.

Malgré sa simplicité d'analyse, l'approximation (II.22) présente l'inconvénient de sous-estimer (sauf pour le cas s = 1) la vraie valeur lorsque $Cs^2 < 1$.

Avec la base B_2, l'approximation (II.20) s'écrit :

$$\overline{W}(M/G/s) \approx \frac{1+Cs^2}{2}\left\{l\,\overline{W}(M/M/s) + 2(1-l)\,\overline{W}(M/D/s)\right\} \qquad \text{(II.24)}$$

où $l = l_1 = 1 - l_2$, $\beta_1 \equiv M/M/s$ et $\beta_2 \equiv M/D/s$.

De la même manière, à partir de l'approximation (II.21), on obtient :

$$\overline{W}(M/G/s) \approx \frac{1+Cs^2}{\frac{2h}{\overline{W}(M/M/s)} + \frac{1-h}{\overline{W}(M/D/s)}} \qquad \text{(II.25)}$$

où $h = h_1 = 1\text{-}h_2$.

La détermination des coefficients l et h se fait par l'utilisation de quelques propriétés asymptotiques de $\overline{W}(M/G/s)$ en introduisant une autre quantité normalisée $R_G \equiv R_G(s, \rho)$ définie par le rapport :

$$R_G(s,\rho) \equiv \frac{\overline{W}(M/G/s)}{\overline{W}(M/M/s)} \qquad \text{(II.26)}$$

De là, les approximations (II.24) et (II.25) peuvent se réécrire comme :

$$R_G \approx \frac{1+Cs^2}{2}\{l + 2(1-l)R_D\} \qquad \text{(II.27)}$$

et

$$R_G \approx \frac{(1+Cs^2)R_D}{(2R_D - 1)h + 1} \qquad \text{(II.28)}$$

respectivement, où R_D est donné par :

$$R_D(s,\rho) \equiv \frac{\overline{W}(M/D/s)}{\overline{W}(M/M/s)} \qquad \text{(II.29)}$$

On énoncera les deux théorèmes suivant:

Théorème II.2 **[KIM 1994]**

Pour $0 < \rho < 1$, on a :

$$R_G(1,\rho) = \frac{1+Cs^2}{2} \tag{II.30}$$

et

$$\lim_{s\to\infty} R_G(s,\rho) = 1 \tag{II.31}$$

Théorème II.3 **[KIM 1994]**

Pour la base $B_2 \equiv \{M/M/s, M/D/s\}$, les approximations (II.24) et (II.25) sont asymptotiquement correctes lorsque $s \to \infty$ si :

$$l = \frac{2Cs^2}{1+Cs^2} \tag{II.32}$$

et

$$h = Cs^2 \tag{II.33}$$

En vertu des théorèmes II.2 et II.3, on peut obtenir deux approximations pour $\overline{W}(M/G/1)$ en remplaçant l et h à partir de (II.32) et (II.33) dans les formules (II.24) et (II.25) respectivement :

$$\overline{W}(M/G/s) \approx Cs^2\,\overline{W}(M/M/s) + (1-Cs^2)\,\overline{W}(M/D/s) \tag{II.34}$$

$$\overline{W}(M/G/s) \approx \frac{1+Cs^2}{\frac{2Cs^2}{\overline{W}(M/M/s)} + \frac{1-Cs^2}{\overline{W}(M/D/s)}} \tag{II.35}$$

B.3. Temps moyen d'attente dans la file GI/GI/s

Un bref aperçu sur les différentes approximations de $\overline{W}(GI/GI/1)$ (par la méthode de diffusion et heuristique) a été présenté dans la partie A de ce chapitre. Il est clair que toutes ces approximations sont exactes pour le cas M/G/1 et asymptotiquement correctes lorsque $\rho \to 1$.

Comme pour le cas M/G/1, l'approximation de $\overline{W}(GI/GI/1)$ repose sur le choix adéquat d'une base propre et de ses coefficients de poids. Kimura propose une extension naturelle de la base B_2 par la base $B_3 \equiv \{M/M/s, M/D/s, D/M/s\}$. L'extension heuristique du résultat (II.32) dans le théorème II.3 pour le cas GI/GI/s avec la base B_3 donne :

$$l_1 = \frac{2(Ca^2 + Cs^2 - 1)}{Ca^2 + Cs^2}, \qquad l_2 = \frac{1 - Cs^2}{Ca^2 + Cs^2}, \qquad l_3 = \frac{1 - Ca^2}{Ca^2 + Cs^2} \tag{II.36}$$

et

$$l_1 = \frac{2\,Ca^2\,Cs^2}{Ca^2 + Cs^2}, \qquad l_2 = \frac{Ca^2(1 - Cs^2)}{Ca^2 + Cs^2}, \qquad l_3 = \frac{(1 - Ca^2)\,Cs^2}{Ca^2 + Cs^2} \tag{II.37}$$

De la même manière, le résultat (II.33) dans le théorème II.3, heuristiquement étendu dans la base B_3, donne :

$$h_1 = Ca^2 + Cs^2 - 1, \qquad h_2 = 1 - Cs^2, \qquad h_3 = 1 - Ca^2 \tag{II.38}$$

La combinaison de ces coefficients avec les différentes approximations citées plus haut (voir partie A) permet d'établir d'autres approximations existantes ou nouvelles. Un exemple est illustré par la combinaison des formules (II.20), (II.37) et l'approximation de Kingman **[KIN 1964]** donnée par :

$$\overline{W}(GI/GI/1) \approx \frac{Ca^2 + Cs^2}{2}\,\overline{W}(M/M/1) \tag{II.39}$$

et utilisée pour approximer $\overline{W}(D/M/1)$ dans (II.20). On aboutit, dans ce cas au résultat suivant :

$$\overline{W}(GI/GI/s) \approx Ca^2\, Cs^2\, \overline{W}(M/M/s) + Ca^2\,(1-Cs^2)\,\overline{W}(M/D/s) + (1-Ca^2)\,Cs^2\,\overline{W}(D/M/s) \tag{II.40}$$

Ce résultat n'est autre que l'approximation heuristique (II.9) donnée par Page **[PAG 1972]** lorsque $s \to 1$.

Une autre combinaison de (II.21), (II.39) et (II.38) aboutit à :

$$\overline{W}(GI/GI/s) \approx \frac{Ca^2 + Cs^2}{\dfrac{2(Ca^2+Cs^2-1)}{\overline{W}(M/M/s)} + \dfrac{1-Cs^2}{\overline{W}(M/D/s)} + \dfrac{1-Ca^2}{\overline{W}(D/M/s)}} \tag{II.41}$$

Sur la base des calculs numériques entrepris pour comparer les différentes approximations de $\overline{W}(GI/GI/1)$ obtenues par les différentes combinaison des coefficients de poids, Kimura **[KIM 1991a]** propose l'approximation unifiée suivante :

$$\overline{W}(GI/GI/s) \approx \begin{cases} \dfrac{g(Ca^2+Cs^2)}{\dfrac{2(Ca^2+Cs^2-1)}{\overline{W}(M/M/s)} + \dfrac{1-Cs^2}{\overline{W}(M/D/s)} + \dfrac{g_{01}(1-Ca^2)}{\overline{W}(D/M/s)}}, & Ca^2 \le 1 \\ (Ca^2+Cs^2-1)\,\overline{W}(M/M/s) + (1-Cs^2)\,\overline{W}(M/D/s) & \\ \quad + \dfrac{1-Ca^2}{g_{01}}\,\overline{W}(D/M/s), & Ca^2 > 1 \end{cases} \tag{II.42}$$

Où :

$$g_{01} \equiv g(\rho, 0, 1) = \exp\left(\frac{-2(1-\rho)}{3\rho}\right) \tag{II.43}$$

et g est défini dans (II.7).

Ces différentes approximations ont été comparées avec les valeurs exactes (obtenues à partir des tables **[SEE 1985]**) du nombre moyen de clients pour quelques systèmes de type $E_2/G/5$ **[KIM 1994]**. Il s'est avéré que l'approximation (II.42) était la plus performante pour $Cs^2 < 1$ alors que l'approximation de Page (II.40) était meilleure pour $Cs^2 > 1$ et spécialement lorsque $Cs^2 \geq 2.5$. Toutes ces observations soutiennent la validité (même partielle) des approximations à deux moments, notamment pour les faibles valeurs des coefficients de variabilité (plus particulièrement lorsque $Ca^2 < 1$).

CHAPITRE III : APPROXIMATIONS HEURISTIQUES DU MODELE GI/GI/1 AVEC RAPPEL EXPONENTIEL

I. Introduction–Position du problème

Par analogie aux systèmes classiques GI/GI/1, où de nombreuses études ont été menées avec plus ou moins de succès, le but de ce travail est d'essayer d'étendre ces travaux au système GI/GI/1 avec rappel exponentiel.

Un examen détaillé de la littérature montre que des progrès ont été réalisés en théorie des files d'attente avec rappel **[YAN 1987 ; FAL 1990 ; AIS 1994 ; M'BA 1995 ; BEN 1995]** et leurs applications **[LeG 1970 ; HAS 1979 ;KEL 1985]**. Cependant, peu de travaux ont été entrepris pour les cas de systèmes présentant des arrivées et des services (ainsi que des rappels) généraux qui restent très mal connus. La complexité de ce type de systèmes est liée au manque de résultats exacts. Les méthodes heuristiques ont été utilisées pour ce type de modèles (où des approches analytiques, algorithmiques et autres sont absentes), notamment pour les systèmes de type GI/G/m/m **[POU 1987, 1988]** ainsi que pour les réseaux en tandem dans le cas avec rappel **[POU 1989, 1990]**.

La contribution de ce travail s'inscrit dans la continuité de l'intérêt récent et croissant porté aux files d'attente GI/GI/1 en raison de leurs applications dans la modélisation de systèmes complexes tels que présentés dans le chapitre I (§ B.2).

Nous nous proposons dans le cadre de ce travail de présenter quelques approximations heuristiques des deux moments basées sur la méthode d'interpolation utilisée par Kimura **[KIM 1994]** pour le cas classique.

II. Approximations heuristiques du système GI/GI/1

En s'inspirant du modèle de Kimura pour le système classique, on considère le rapport :

$$Qr(GI/GI/s) \equiv \frac{\overline{Wr}(GI/GI/s)}{\overline{Wr}(GI/GI/1)} \qquad \text{(III.1)}$$

où $\overline{Wr}(GI/GI/s)$ et $\overline{Wr}(GI/GI/1)$ sont les temps moyens d'attente dans la file pour les systèmes GI/GI/s et GI/GI/1 avec rappel respectivement.

Cette quantité est ensuite écrite comme une combinaison d'interpolation (linéaire et harmonique) dans la base $B \equiv \{M/M/s, M/D/s, D/M/s\}$ de systèmes simples avec rappels :

$$Qr(GI/GI/s) \approx \sum_{i=1}^{n} l_i \, Qr(\beta_i) \qquad \text{(III.2)}$$

$$Qr(GI/GI/s) \approx \left\{ \sum_{i=1}^{n} \frac{h_i}{Qr(\beta_i)} \right\}^{-1} \qquad \text{(III.3)}$$

où l_i et h_i sont les coefficients de poids du $i^{ème}$ système (i= 1, …, n) de la base B et vérifient le théorème II.1 appliqué au cas avec rappel :

$$\sum_{i=1}^{n} l_i = 1 \quad \text{et} \quad \sum_{i=1}^{n} h_i = 1. \qquad \text{(III.4)}$$

II.1. Approximations de type linéaire

Dans la base B, l'approximation (III.2) s'écrit :

$$\begin{aligned}\overline{Wr}(GI/GI/s) &\approx \overline{Wr}(GI/GI/1)\sum_{i=1}^{n} l_i\, Qr(\beta_i) \\ &\approx \overline{Wr}(GI/GI/1)\left\{ l_1 Qr(M/M/s) + l_2 Qr(M/D/s) + l_3 Qr(D/M/s) \right\}\end{aligned} \tag{III.5}$$

Cette approximation est équivalente à :

$$\begin{aligned}\overline{Wr}(GI/GI/s) \approx \frac{\overline{Wr}(GI/GI/1)}{\overline{\overline{Wr}}(M/M/1)} \Bigg\{ & l_1 \overline{Wr}(M/M/s) + l_2 \frac{\overline{Wr}(M/M/1)}{\overline{\overline{Wr}}(M/D/1)} \overline{Wr}(M/D/s) \\ & + l_3 \frac{\overline{Wr}(M/M/1)}{\overline{\overline{Wr}}(D/M/1)} \overline{Wr}(D/M/s) \Bigg\}\end{aligned} \tag{III.6}$$

Où, à partir de la formule (I.29) du temps moyen d'attente dans la file M/G/1 avec rappel, on a :

$$\begin{aligned}\overline{Wr}(M/M/1) &= \frac{\rho^2}{\lambda(1-\rho)} + \frac{\rho}{\theta(1-\rho)} \\ \overline{Wr}(M/D/1) &= \frac{\rho^2}{2\,\lambda(1-\rho)} + \frac{\rho}{\theta(1-\rho)}\end{aligned} \tag{III.7}$$

Pour $\overline{Wr}(GI/GI/1)$, par extension heuristique du résultat exact de la file M/G/1 avec rappel, on a :

$$\overline{Wr}(GI/GI/1) = \frac{\theta\rho(Ca^2 + Cs^2) + 2\lambda}{2(\theta\rho + \lambda)}\, \overline{Wr}(M/M/1) \tag{III.8}$$

$\overline{Wr}(GI/GI/1)$ peut- aussi être obtenu heuristiquement par analogie avec la formule de Kramer et Lagenbach-Belz (II.6) :

$$\overline{Wr}(GI/GI/l) = \frac{\theta\rho(Ca^2 + Cs^2) + 2\lambda}{2(\theta\rho + \lambda)} g(Ca^2, Cs^2, \rho)\, \overline{Wr}(M/M/l) \qquad \text{(III.9)}$$

Où g (Ca^2, Cs^2, ρ) est donné par la formule (II.7).

Pour $\overline{Wr}(D/M/l)$, il n'existe aucune formule explicite mais il peut être approximé par les formules (III.8) et (III.9) en remplaçant Ca^2 par 0 et Cs^2 par 1 :

$$\overline{Wr}(D/M/l) = \frac{\theta\rho + 2\lambda}{2(\theta\rho + \lambda)} \overline{Wr}(M/M/l) \qquad \text{(III.10)}$$

et

$$\overline{Wr}(D/M/l) = \frac{\theta\rho + 2\lambda}{2(\theta\rho + \lambda)} g_{01}\ \overline{Wr}(M/M/l) \qquad \text{(III.11)}$$

où g_{01} est donné par la formule (II.43).

A partir de (III.7), (III.8) et (III.10), l'approximation (III.6) s'écrit :

$$\overline{Wr}(GI/GI/s) \approx \frac{\theta\rho(Ca^2 + Cs^2) + 2\lambda}{2(\theta\rho + \lambda)} \left\{ l_1 \overline{Wr}(M/M/s) + l_2 \frac{2(\theta\rho + \lambda)}{\theta\rho + 2\lambda} \overline{Wr}(M/D/s) + l_3 \frac{2(\theta\rho + \lambda)}{\theta\rho + 2\lambda} \overline{Wr}(D/M/s) \right\} \qquad \text{(III.12)}$$

A partir de (III.7), (III.8) et (III.11), l'approximation (III.6) s'écrit :

$$\overline{Wr}(GI/GI/s) \approx \frac{\theta\rho(Ca^2 + Cs^2) + 2\lambda}{2(\theta\rho + \lambda)} \left\{ l_1 \overline{Wr}(M/M/s) + l_2 \frac{2(\theta\rho + \lambda)}{\theta\rho + 2\lambda} \overline{Wr}(M/D/s) + l_3 \frac{2(\theta\rho + \lambda)}{(\theta\rho + 2\lambda) g_{01}} \overline{Wr}(D/M/s) \right\} \qquad \text{(III.13)}$$

Appliquées au système M/G/1 avec rappel, les approximations (III.12) et (III.13) s'écrivent :

$$R_G \approx \frac{\theta\rho(1+Cs^2)+2\lambda}{2(\theta\rho+\lambda)}\left\{1+2(1-l)\frac{(\theta\rho+\lambda)}{\theta\rho+2\lambda}R_D\right\} \qquad \text{(III.14)}$$

tel que :

$$R_G \equiv \frac{\overline{Wr}(M/G/s)}{\overline{Wr}(M/M/s)}, \qquad R_D \equiv \frac{\overline{Wr}(M/D/s)}{\overline{Wr}(M/M/s)} \quad \text{et} \quad l = l_1 = 1 - l_2 \qquad \text{(III.15)}$$

Dans la limite asymptotique $s \to \infty$, on a $R_G(R_D) \to 1$ et à partir de (III.14), on obtient :

$$l = \frac{2(\theta\rho+\lambda)}{\theta\rho(1+Cs^2)+2\lambda}Cs^2 \qquad \text{(III.16)}$$

En étendant ce résultat (III.16) pour le cas GI/GI/s avec la base B, on propose les deux cas suivant dans le choix des coefficients de poids :

$$\begin{cases} l_1 = \dfrac{2(\theta\rho+\lambda)(Ca^2+Cs^2-1)}{\theta\rho(Ca^2+Cs^2)+2\lambda}, \\ l_2 = \dfrac{(\theta\rho+2\lambda)(1-Cs^2)}{\theta\rho(Ca^2+Cs^2)+2\lambda}, \\ l_3 = \dfrac{(\theta\rho+2\lambda)(1-Ca^2)}{\theta\rho(Ca^2+Cs^2)+2\lambda} \end{cases} \qquad \text{(III.17)}$$

et

$$\begin{cases} l_1 = \dfrac{2(\theta\rho+\lambda)\,Ca^2\,Cs^2}{\theta\rho\,(Ca^2+Cs^2)+2\lambda}, \\ l_2 = \dfrac{(\theta\rho+2\lambda)(1-Cs^2)\,Ca^2}{\theta\rho\,(Ca^2+Cs^2)+2\lambda}, \\ l_3 = \dfrac{(\theta\rho\,Cs^2+2\lambda)(1-Ca^2)}{\theta\rho\,(Ca^2+Cs^2)+2\lambda} \end{cases} \qquad \text{(III.18)}$$

En remplaçant les l_i de (III.17) et (III.18) dans (III.12) et (III.13) respectivement, on aboutit aux approximations suivantes :

$$\begin{aligned} \overline{Wr}(GI/GI/s) \approx\ & (Ca^2+Cs^2-1)\overline{Wr}(M/M/s) + (1-Cs^2)\,\overline{Wr}(M/D/s) \\ & + (1-Ca^2)\overline{Wr}(D/M/s) \end{aligned} \qquad \text{(III.19)}$$

$$\begin{aligned} \overline{Wr}(GI/GI/s) \approx\ & Ca^2Cs^2\,\overline{Wr}(M/M/s) + (1-Cs^2)\,Ca^2\,\overline{Wr}(M/D/s) \\ & + \frac{(\theta\rho\,Cs^2+2\lambda)(1-Ca^2)}{\theta\rho+2\lambda}\,\overline{Wr}(D/M/s) \end{aligned} \qquad \text{(III.20)}$$

$$\begin{aligned} \overline{Wr}(GI/GI/s) \approx\ & (Ca^2+Cs^2-1)\overline{Wr}(M/M/s) + (1-Cs^2)\,\overline{Wr}(M/D/s) \\ & + \frac{(1-Ca^2)}{g_{01}}\,\overline{Wr}(D/M/s) \end{aligned} \qquad \text{(III.21)}$$

$$\begin{aligned} \overline{Wr}(GI/GI/s) \approx\ & Ca^2Cs^2\,\overline{Wr}(M/M/s) + (1-Cs^2)\,Ca^2\,\overline{Wr}(M/D/s) \\ & + \frac{(\theta\rho\,Cs^2+2\lambda)(1-Ca^2)}{(\theta\rho+2\lambda)\,g_{01}}\,\overline{Wr}(D/M/s) \end{aligned} \qquad \text{(III.22)}$$

Pour le cas simple s = 1, on déduit alors pour le temps d'attente moyen dans la file les approximations heuristiques suivantes :

$$\begin{aligned} \overline{Wr}(GI/GI/1) \approx\ & (Ca^2+Cs^2-1)\overline{Wr}(M/M/1) + (1-Cs^2)\,\overline{Wr}(M/D/1) \\ & + (1-Ca^2)\overline{Wr}(D/M/1) \end{aligned} \qquad \textbf{(app.1)}$$

$$\overline{Wr}(GI/GI/1) \approx Ca^2 Cs^2\, \overline{Wr}(M/M/1) + (1-Cs^2)\, Ca^2\, \overline{Wr}(M/D/1) + \frac{(\theta\rho\, Cs^2 + 2\lambda)(1-Ca^2)}{\theta\rho + 2\lambda}\, \overline{Wr}(D/M/1) \qquad \textbf{(app.2)}$$

$$\overline{Wr}(GI/GI/1) \approx (Ca^2 + Cs^2 - 1)\overline{Wr}(M/M/1) + (1-Cs^2)\, \overline{Wr}(M/D/1) + \frac{(1-Ca^2)}{g_{01}}\, \overline{Wr}(D/M/1) \qquad \textbf{(app.3)}$$

$$\overline{Wr}(GI/GI/1) \approx Ca^2 Cs^2\, \overline{Wr}(M/M/1) + (1-Cs^2)\, Ca^2\, \overline{Wr}(M/D/1) + \frac{(\theta\rho\, Cs^2 + 2\lambda)(1-Ca^2)}{(\theta\rho + 2\lambda)\, g_{01}}\, \overline{Wr}(D/M/1) \qquad \textbf{(app.4)}$$

II.2. Approximations de type harmonique

Dans la base B, l'approximation (III.3) s'écrit :

$$\overline{Wr}(GI/GI/s) \approx \overline{Wr}(GI/GI/1)\left\{\sum_{i=1}^{n} \frac{h_i}{Qr(\beta_i)}\right\}^{-1} \approx \overline{Wr}(GI/GI/1)\left\{\frac{h_1}{Qr(M/M/s)} + \frac{h_2}{Qr(M/D/s)} + \frac{h_3}{Qr(D/M/s)}\right\}^{-1} \qquad \text{(III.23)}$$

Cette approximation est équivalente à :

$$\overline{Wr}(GI/GI/s) \approx \frac{\overline{Wr}(GI/GI/1)}{\overline{Wr}(M/M/1)}\left\{\frac{h_1}{\overline{Wr}(M/M/s)} + \frac{h_2}{\overline{Wr}(M/D/s)}\frac{\overline{Wr}(M/D/1)}{\overline{Wr}(M/M/1)} + \frac{h_3}{\overline{Wr}(D/M/s)}\frac{\overline{Wr}(D/M/1)}{\overline{Wr}(M/M/1)}\right\}^{-1} \qquad \text{(III.24)}$$

A partir de (III.7), (III.8) et (III.10), l'approximation (III.24) s'écrit :

$$\overline{Wr}(GI/GI/s) \approx \frac{\dfrac{\theta\rho(Ca^2+Cs^2)+2\lambda}{2(\theta\rho+\lambda)}}{\dfrac{h_1}{\overline{Wr}(M/M/s)}+\dfrac{h_2}{\overline{Wr}(M/D/s)}\dfrac{\theta\rho+2\lambda}{2(\theta\rho+\lambda)}+\dfrac{h_3}{\overline{Wr}(D/M/s)}\dfrac{\theta\rho+2\lambda}{2(\theta\rho+\lambda)}} \qquad \text{(III.25)}$$

A partir de (III.7), (III.8) et (III.11), l'approximation (III.24) s'écrit :

$$\overline{Wr}(GI/GI/s) \approx \frac{\dfrac{\theta\rho(Ca^2+Cs^2)+2\lambda}{2(\theta\rho+\lambda)}}{\dfrac{h_1}{\overline{Wr}(M/M/s)}+\dfrac{h_2}{\overline{Wr}(M/D/s)}\dfrac{\theta\rho+2\lambda}{2(\theta\rho+\lambda)}+\dfrac{h_3}{\overline{Wr}(D/M/s)}\dfrac{(\theta\rho+2\lambda)g_{01}}{2(\theta\rho+\lambda)}} \qquad \text{(III.26)}$$

Appliquées au système M/G/1 avec rappel, les approximations (III.25) et (III.26) s'écrivent :

$$R_G \approx \frac{\left[\theta\rho(1+Cs^2)+2\lambda\right]R_D}{\left[2(\theta\rho+\lambda)R_D-(\theta\rho+2\lambda)\right]h+(\theta\rho+2\lambda)} \qquad \text{(III.27)}$$

tel que $h = h_1 = 1 - h_2$.

Dans la limite asymptotique $s \to \infty$, on a $R_G(R_D) \to 1$ et à partir de (III.27), on obtient :

$$h = Cs^2 \qquad \text{(III.28)}$$

En étendant ce résultat (III.28) pour le cas GI/GI/s avec la base B, on propose les coefficients de poids suivants :

$$\begin{cases} h_1 = Ca^2 + Cs^2 - 1, \\ h_2 = 1 - Cs^2, \\ h_3 = 1 - Ca^2 \end{cases} \qquad \text{(III.29)}$$

et

$$\begin{cases} h_1 = Cs^2\, Ca^2, \\ h_2 = (1 - Cs^2)\, Ca^2, \\ h_3 = 1 - Ca^2 \end{cases} \qquad \text{(III.30)}$$

En remplaçant les h_i de (III.29) et (III.30) dans (III.25) et (III.26) respectivement, on aboutit aux approximations suivantes :

$$\overline{Wr}(GI/GI/s) \approx \frac{\theta\rho(Ca^2 + Cs^2) + 2\lambda}{\frac{2(\theta\rho+\lambda)(Ca^2 + Cs^2 - 1)}{\overline{Wr}(M/M/s)} + \frac{(1 - Cs^2)(\theta\rho + 2\lambda)}{\overline{Wr}(M/D/s)} + \frac{(1 - Ca^2)(\theta\rho + 2\lambda)}{\overline{Wr}(D/M/s)}} \qquad \text{(III.31)}$$

$$\overline{Wr}(GI/GI/s) \approx \frac{\theta\rho(Ca^2 + Cs^2) + 2\lambda}{\frac{2(\theta\rho+\lambda)(Ca^2 + Cs^2 - 1)}{\overline{Wr}(M/M/s)} + \frac{(1 - Cs^2)(\theta\rho + 2\lambda)}{\overline{Wr}(M/D/s)} + \frac{(1 - Ca^2)(\theta\rho + 2\lambda)g_{01}}{\overline{Wr}(D/M/s)}} \qquad \text{(III.32)}$$

$$\overline{Wr}(GI/GI/s) \approx \frac{\theta\rho(Ca^2 + Cs^2) + 2\lambda}{\frac{2(\theta\rho+\lambda)Cs^2Ca^2}{\overline{Wr}(M/M/s)} + \frac{Ca^2(1 - Cs^2)(\theta\rho + 2\lambda)}{\overline{Wr}(M/D/s)} + \frac{(1 - Ca^2)(\theta\rho + 2\lambda)}{\overline{Wr}(D/M/s)}} \qquad \text{(III.33)}$$

$$\overline{Wr}(GI/GI/s) \approx \frac{\theta\rho(Ca^2 + Cs^2) + 2\lambda}{\frac{2(\theta\rho+\lambda)Cs^2Ca^2}{\overline{Wr}(M/M/s)} + \frac{Ca^2(1-Cs^2)(\theta\rho+2\lambda)}{\overline{Wr}(M/D/s)} + \frac{(1-Ca^2)(\theta\rho+2\lambda)g_{01}}{\overline{Wr}(D/M/s)}}$$

(III.34)

Ainsi, pour le cas simple s = 1, on déduit pour le temps d'attente moyen dans la file les approximations heuristiques suivantes :

$$\overline{Wr}(GI/GI/1) \approx \frac{\theta\rho(Ca^2 + Cs^2) + 2\lambda}{\frac{2(\theta\rho+\lambda)(Ca^2+Cs^2-1)}{\overline{Wr}(M/M/1)} + \frac{(1-Cs^2)(\theta\rho+2\lambda)}{\overline{Wr}(M/D/1)} + \frac{(1-Ca^2)(\theta\rho+2\lambda)}{\overline{Wr}(D/M/1)}}$$

(**app.5**)

$$\overline{Wr}(GI/GI/1) \approx \frac{\theta\rho(Ca^2 + Cs^2) + 2\lambda}{\frac{2(\theta\rho+\lambda)(Ca^2+Cs^2-1)}{\overline{Wr}(M/M/1)} + \frac{(1-Cs^2)(\theta\rho+2\lambda)}{\overline{Wr}(M/D/1)} + \frac{(1-Ca^2)(\theta\rho+2\lambda)g_{01}}{\overline{Wr}(D/M/1)}}$$

(**app.6**)

$$\overline{Wr}(GI/GI/1) \approx \frac{\theta\rho(Ca^2 + Cs^2) + 2\lambda}{\frac{2(\theta\rho+\lambda)Cs^2Ca^2}{\overline{Wr}(M/M/1)} + \frac{Ca^2(1-Cs^2)(\theta\rho+2\lambda)}{\overline{Wr}(M/D/1)} + \frac{(1-Ca^2)(\theta\rho+2\lambda)}{\overline{Wr}(D/M/1)}}$$

(**app.7**)

$$\overline{Wr}(GI/GI/1) \approx \frac{\theta\rho(Ca^2 + Cs^2) + 2\lambda}{\frac{2(\theta\rho+\lambda)Cs^2Ca^2}{\overline{Wr}(M/M/1)} + \frac{Ca^2(1-Cs^2)(\theta\rho+2\lambda)}{\overline{Wr}(M/D/1)} + \frac{(1-Ca^2)(\theta\rho+2\lambda)g_{01}}{\overline{Wr}(D/M/1)}}$$

(**app.8**)

II.3. Autre approximation heuristique

Toujours dans le cas d'un rappel exponentiel, nous proposons une dernière approximation pour le temps d'attente moyen dans la file en se basant sur les coefficients de poids (II.37) obtenus par Kimura dans le système GI/GI/1 classique

$$l_1 = \frac{2\,Ca^2\,Cs^2}{Ca^2 + Cs^2}, \qquad l_2 = \frac{Ca^2\,(1-Cs^2)}{Ca^2 + Cs^2}, \qquad l_3 = \frac{(1-Ca^2)\,Cs^2}{Ca^2 + Cs^2} \tag{III.35}$$

L'interpolation dans la base $\{M/M/1, M/D/1, D/M/1\}$ des files d'attente classiques avec ces coefficients de poids, aboutit à l'approximation de Page (II.40). L'idée serait d'étendre cette interpolation dans la base $B' \equiv \{M/M/1, M/D/1, D/M/1\}$des files d'attente avec rappel exponentiel. On obtient ainsi l'approximation heuristique suivante :

$$\begin{aligned} \overline{Wr}(GI/GI/1) \approx\; & Ca^2\,Cs^2\,\overline{Wr}(M/M/1) + Ca^2\,(1-Cs^2)\,\overline{Wr}(M/D/1) \\ & + (1-Ca^2)\,Cs^2\;\overline{Wr}(D/M/1) \end{aligned} \tag{**app.9**}$$

Dans le cas d'une file M/G/1 (Ca^2 =1) avec rappel exponentiel, l'approximation (app.9) devient :

$$\begin{aligned} \overline{Wr}(M/G/1) &\approx Cs^2\,\overline{Wr}(M/M/1) + (1-Cs^2)\,\overline{Wr}(M/D/1) \\ &\approx Cs^2\left\{\frac{\rho^2}{\lambda(1-\rho)} + \frac{\rho}{\theta(1-\rho)}\right\} + (1-Cs^2)\left\{\frac{\rho^2}{2\,\lambda(1-\rho)} + \frac{\rho}{\theta(1-\rho)}\right\} \\ &\approx \frac{(1+Cs^2)\,\rho^2}{2\,\lambda(1-\rho)} + \frac{\rho}{\theta(1-\rho)} \end{aligned} \tag{III.37}$$

Ce qui correspond au résultat exact pour une file M/G/1 (voir (I.29)) et justifie ainsi le choix d'une telle interpolation.

CHAPITRE IV : SIMULATION ET VALIDATION DES APPROXIMATIONS

Introduction

La validité de la méthode utilisée dans le troisième chapitre est testée à partir de la simulation statistique dans le but de sélectionner l' (les) approximation (s) la (les) plus appropriée (s) pour les différents domaines des coefficients de variations des temps d'inter arrivées et de service.

A. Simulation des systèmes de files d'attente

La simulation est un procédé d'imitation artificielle du comportement d'un système réel à travers le temps. C'est une expérimentation informatique qui permet de récolter différentes données sur le système simulé, comme s'il était réellement soumis à une observation. Elle remplace ainsi l'expérimentation réelle difficile et coûteuse. Cette *expérimentation indirecte* connaît un essor considérable grâce au besoin croissant de simuler par ordinateur des systèmes de plus en plus complexes

A.1. Simulation par évènements discrets

L'une des méthodes de simulation les plus utilisées, notamment dans les systèmes de files d'attente, est la *simulation par évènements discrets* **[ERA 1996 ; BAN 1996]**. Cette méthode, largement utilisée durant cette dernière décennie, se base sur la modélisation de systèmes dans lesquels le changement d'état s'effectue à des instants discrets de l'axe temporel. Elle permet ainsi d'étudier le comportement d'un système à travers quelques périodes, en construisant un système (modèle) ayant la même structure que l'original (objet de l'étude) mais plus simple à manipuler.

A.1.1. Génération de variables aléatoires

L'outil de base dans la simulation d'un phénomène stochastique est la génération des nombres aléatoires, c'est à dire une suite $\{U_1, U_2,\ldots, U_i,\ldots\}$ de variables aléatoires indépendantes et uniformément distribuées sur l'intervalle [0, 1] (voir annexe A.3.1). On pourra, à partir d'une telle suite, générer toute autre suite de variables aléatoires (les instants d'arrivées, de services et de rappels dans notre cas) distribuées suivant une loi arbitraire.

La plupart des langages informatiques possèdent des programmes de génération (*générateurs*) de suites aléatoires uniformes sur [0,.1] : le programme RANDOM en Turbo Pascal, RAND en Basic, RND en Fortran, …. Dans le cas de notre travail, nous travaillons avec le langage MATLAB 5.3 et son générateur RAND qui peut générer 2^{1492} nombres à virgule flottante uniformément distribués dans l'intervalle fermé $\left[2^{-53}, 1-2^{-53}\right]$.

A.1.2. Méthodes de génération des nombres aléatoires

La génération d'une suite de variables aléatoires, distribuées suivant une loi de probabilité arbitraire, s'obtient à partir d'une suite préalable $\{U_1, U_2,\ldots, U_i,\ldots\}$ de nombres issus de la loi uniforme U [0, 1], moyennant une certaine transformation.

Il existe différentes techniques de transformation qui se basent sur certaines propriétés propres à la distribution qu'on cherche à générer. Les plus connues d'entre elles sont celles dites d'*inversion*, de *convolution* **[BAN 1996]** et de *composition* **[BAN 1996 ; ERA 1996]**.

- La méthode d'inversion a pour principe de générer une variable aléatoire X de fonction de répartition F à partir de l'égalité U = F(x) où U est une variable aléatoire uniformément distribuée sur [0, 1]. La variable X peut ainsi être déterminée à l'aide de la relation $X = F^{-1}(U)$.

Cette technique est utilisée dans MATLAB 5.3 pour la génération de nombres aléatoires suivant la loi exponentielle (EXPRND) (voir annexe A.3.2), la loi de Weibull (WEIBRND), …. Bien que cette technique soit assez simple en théorie, elle est cependant très complexe pour certaines lois de probabilités (loi normale, loi gamma,…).

- La méthode de convolution consiste à sommer deux (ou plusieurs) variables aléatoires pour obtenir une variable aléatoire distribuée selon la loi de probabilité désirée. La loi d'Erlang (somme de plusieurs lois exponentielles) en est le parfait exemple (voir annexe A.3.3). Celle-ci étant un cas particulier de la loi Gamma, elle est générée dans MATLAB 5.3 par la commande GAMRND avec les paramètres appropriés.

- La méthode de composition est utilisée comme représentation de la fonction de répartition F(X) par une mixture de deux distributions ou plus :

$$F(X) = P * F_1(X) + (1 - P) * F_2(X)$$

la variable X est obtenue après génération d'une variable uniforme U. Si U < P, alors X = U avec la probabilité $F_1(X)$ sinon X est rejeté avec la probabilité $F_2(X)$.
Cette procédure est directement applicable pour la génération des lois de Cox (mélange de lois d'Erlang) et de la loi Hyper exponentielle (mélange de lois exponentielles).

A.2. Programme de simulation

Un système de files d'attente avec rappel est décrit par un flot d'arrivées (primaires et secondaires), une durée de service, la capacité de l'orbite et la fonction de persévérance. La modélisation pour simulation de ce genre de système nécessite l'assimilation de certains concepts tels *l'état du système*, *l'événement* et *l'horloge*.

- L'état d'un système se caractérise par le nombre de clients dans le système (orbite plus serveur) et l'état du serveur (occupé ou libre).

- Un événement est l'ensemble de circonstances produisant un changement instantané dans l'état du système. Cet évènement fait évoluer le système dans le temps d'un état vers un autre.

Dans le cas des files d'attente avec rappel à un seul serveur, les trois évènements possibles pouvant influer sur l'état du système sont : l'entrée du client (*événement arrivée primaire*), la fin du service d'un client (*événement départ*) et le rappel d'un client (*événement arrivée secondaire ou rappel*).

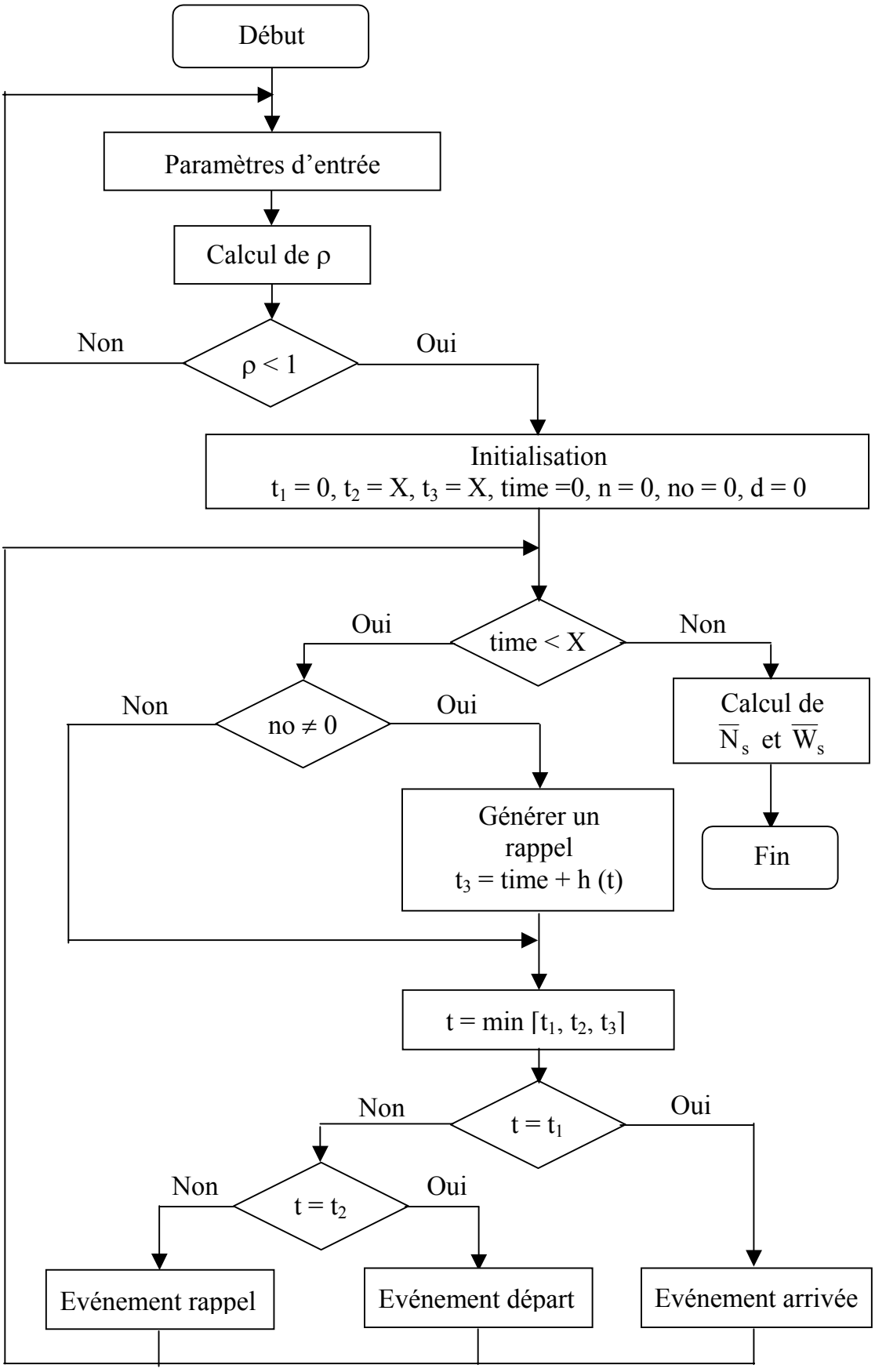
- L'horloge est un compteur qui sert à situer les événements dans le temps (instants d'occurrence des différents événements dans le système). Ces derniers se réalisent généralement d'une manière aléatoire.

A.2.1 Description du modèle

Le modèle de simulation, précédemment utilisé pour l'estimation de la stabilité forte dans les systèmes de files d'attente G/M/1 **[BEN 2000]** et M/G/1 avec rappel **[BER 2000]**, a été adapté dans le cadre de ce travail afin de simuler le comportement des files d'attente GI/GI/1 avec rappel exponentiel et une fonction de persévérance $H_k = 1$ pour tout k (clients persistant ou *orbite infinie sans perte*).

Dans ce système GI/GI/1, on s'intéresse à la durée séparant deux arrivées consécutives (variable aléatoire distribuée selon la loi de probabilité F), à la durée de service d'un client (variable aléatoire distribuée selon la loi de probabilité B) et à la durée séparant deux rappels consécutifs (variable aléatoire distribuée selon la loi exponentielle).

L'algorithme du programme de simulation, développé pour suivre l'évolution du temps d'attente des clients dans le système, est illustré dans l'organigramme de la



<u>Figure IV.1</u> : Organigramme du programme de simulation du système GI/GI/1 avec rappel.

figure IV.1. Cet algorithme consiste en une suite d'occurrence d'événements à partir d'un état initial vers un état final stationnaire au bout duquel le temps moyen d'attente (le nombre moyen de clients dans le système) est mesuré. Chaque simulation est entreprise de 10 à 20 fois dans le but d'améliorer la statistique de l'échantillon et de tester la reproductibilité des résultats.

A.2.2 Paramètres d'entrée

Le programme de simulation a été élaboré de manière à permettre la variation d'un certain nombre de paramètres d'entrée qui influent sur l'évolution du système. Dans le cadre de ce travail, nous avons été amenés à simuler les systèmes $E_k(\mu a)/E_q(\mu s)/1$, $E_k(\mu a)/\gamma(\alpha s, \beta s)/1$, $\gamma(\alpha a, \beta a)/E_q(\mu s)/1$ et $\gamma(\alpha a, \beta a)/\gamma(\alpha s, \beta s)/1$.

Le choix des différents paramètres des lois d'arrivée et de service a pour objectif de balayer le domaine des coefficients de variations $0 \le Ca^2 \le 1$ et $0 \le Cs^2 \le 1$, tout en respectant la condition de stabilité $\rho < 1$ et la condition de stationnarité de la file en définissant un temps de simulation X d'autant plus grand que ρ est proche de 1.

En plus du paramètre θ (taux de rappel exponentiel) et du nombre d'itérations du programme, nous nous trouvons en présence de sept paramètres d'entrée qui rendent difficile l'étude simultanée de leur influence sur les résultats de simulation. Nous avons donc été amené à faire les restrictions (raisonnables) suivantes:

- Le taux de rappel θ ainsi que le coefficient de variabilité Cs^2 ont été fixés pour étudier l'évolution du système pour différentes valeurs de Ca^2. Ce travail est entrepris pour différentes valeurs de ρ (ρ = 0.4, 0.8).

- Ce même travail est fait pour différentes valeurs de Cs^2.

- Le temps de simulation X (instant de fin de simulation) est fixé entre 100000 et 300000 (dans l'horloge de MATLAB 5.3) selon les valeurs de ρ (suffisant pour se considérer dans l'état stationnaire).

- Le nombre d'itérations est fixé entre 10 et 20 de telle manière que la variance de l'échantillon soit inférieure à 1% de la valeur mesurée (moyenne de l'échantillon).

A.2.3. Initialisation du système

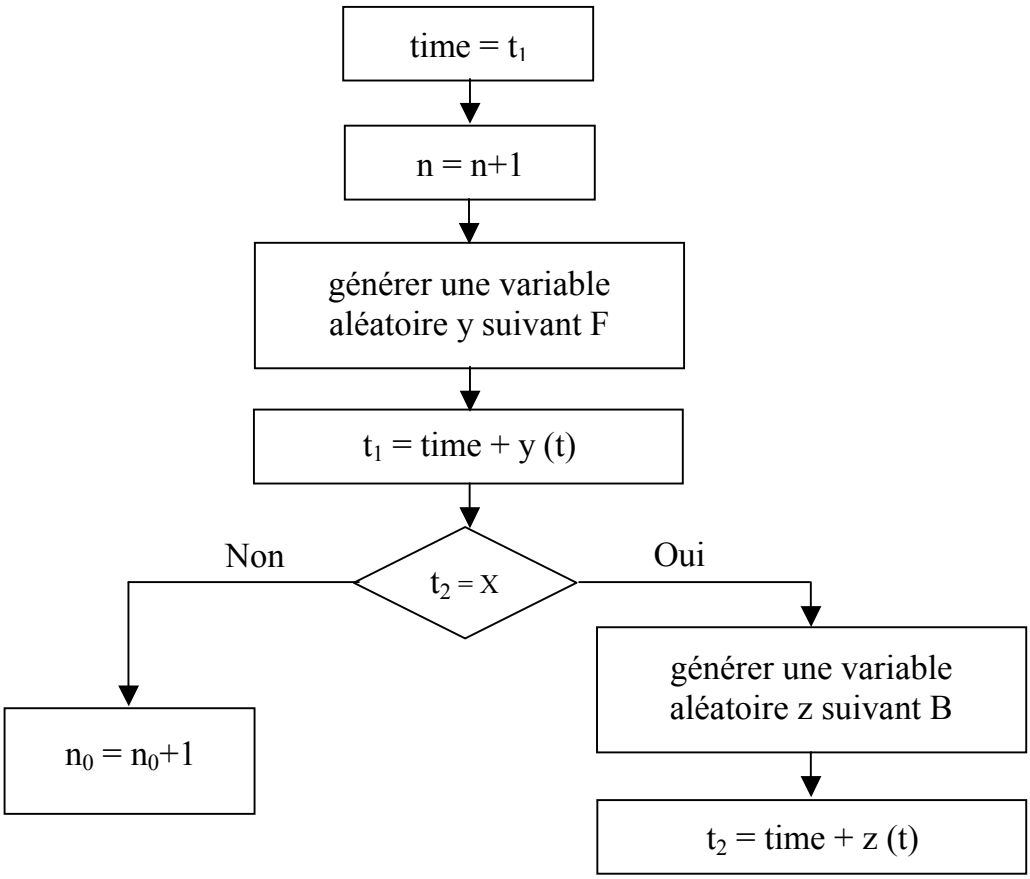
Le système de simulation est caractérisé par les variables représentant l'état du système, les différents événements ainsi que l'horloge qui situe ces événements. L'état initial du système est donné par:

t_1 = instant de la prochaine arrivée = 0
t_2 = instant du prochain départ = X (serveur libre)
t_3 = instant du prochain rappel = X (orbite vide)
n = nombre de clients dans le système = 0
n_0 = nombre de clients dans l'orbite = 0
d = nombre de départs du système = 0
time = horloge (ou compteur) = 0

A.2.4 Evolution du système

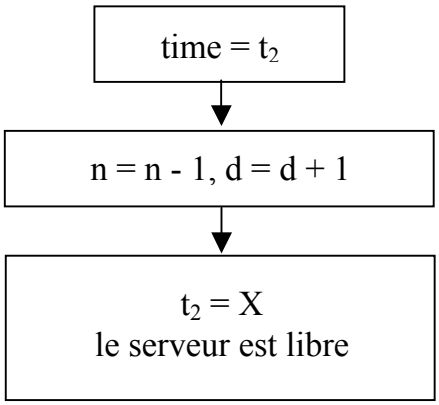
Une fois l'état initial défini, le système de files d'attente évolue sous l'influence des trois événements possibles (événement arrivée, événement départ et événement rappel).

A l'occurrence d'un événement arrivée, le serveur est soit occupé ou bien libre. Par conséquent, le client entre en orbite ou bien en service. La simulation se poursuit selon l'organigramme représenté sur la figure IV.2.



<u>Figure IV.2</u> : Organigramme de l'événement arrivée.

A l'occurrence d'un départ, le nombre de clients dans le système diminue d'une unité et le serveur redevient libre. La simulation de cet événement se fait suivant L'organigramme de la figure IV.3.



<u>Figure IV.3</u> : Organigramme de l'événement départ.

A l'occurrence d'un rappel, si le serveur est libre, le nombre de clients dans l'orbite diminue d'une unité. Si le nombre de clients restant en orbite est nul, le rappel est alors bloqué, sinon le serveur est sollicité pour un autre rappel selon sa disponibilité. L'organigramme de la procédure de rappel est illustré dans la figure IV.4.

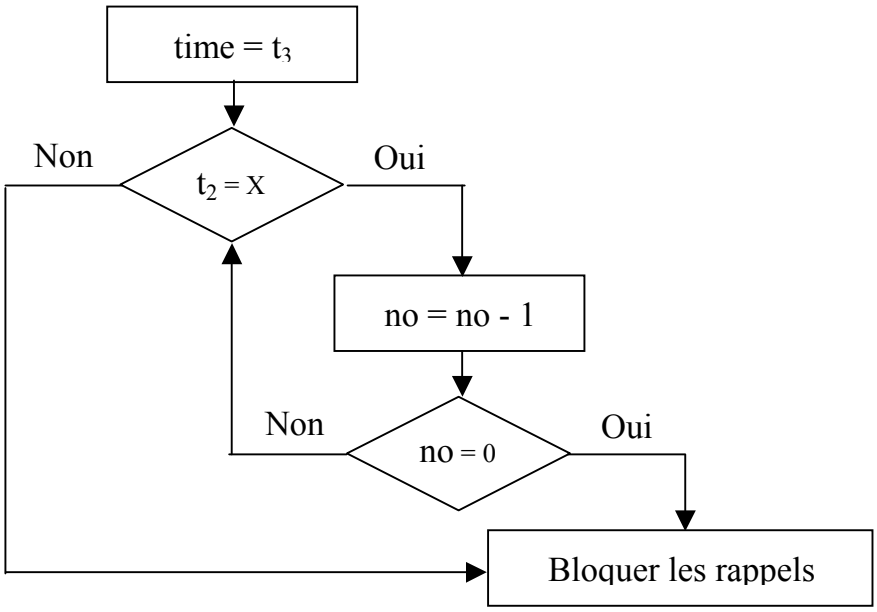


***Figure IV.4** : Organigramme de l'événement rappel.*

A.2.5. Calcul des paramètres de performance

Nous nous intéressons dans l'application de ce programme à la mesure du nombre moyen de clients dans le système et, par l'intermédiaire de la formule de Little, au temps moyen d'attente des clients dans le système. Cette mesure se base sur le principe selon lequel la probabilité d'état pour qu'il y ait n clients dans le système en état stationnaire ($\pi_n = \lim_{t\to\infty} P(X_t = n)$) puisse être également interprétée comme la proportion du temps où il y' aura n clients dans le système en état stationnaire:

$$\pi_n = \frac{\Delta t_n}{X}$$

Le nombre moyen de clients est donné par:

$$\overline{N}_s = \sum_n n\,\pi_n = \sum_n n\,\frac{\Delta t_n}{X} = \frac{\sum_n n\,\Delta t_n}{X}$$

Dans le programme, nous calculons à chaque instant de départ la somme des durées d'attente dans le système multipliées par le nombre de clients présents à cette instant ($s = \sum_n n\,\Delta t_n$). A la fin de la simulation, le nombre moyen de clients dans le système est obtenu en divisant s par le temps de simulation ($\overline{N}_s = \frac{s}{X}$).

CONCLUSION GENERALE

Dans cette thèse, nous avons présenté une méthode de génération et d'évaluation d'approximations pour l'estimation des principaux paramètres de performance dans le système GI/GI/1 avec rappel exponentiel. Notre étude était focalisée sur l'estimation du temps moyen d'attente dans le système, même si pour les systèmes avec rappel, on connaît plus de choses sur le nombre de clients dans le système ou en orbite que sur le temps d'attente.

La méthode d'approximation des deux moments est établie selon le principe d'interpolation en combinant (de manière linéaire et harmonique) des solutions analytiques de systèmes simples que sont les files M/M/1, M/D/1 et D/M/1. Plusieurs approximations du temps moyen dans le système sont alors présentées.

Une comparaison de ces approximations avec la simulation de différents modèles GI/GI/1 fournit un guide de sélection des approximations les plus appropriées pour les différentes valeurs des coefficients de variation des distributions des temps d'inter arrivées (Ca^2) et de service (Cs^2) et selon que le système soit en régime chargé ou peu chargé ($\rho = 0.4$, $\rho = 0.8$). Il en résulte deux approximations principales :

$$\overline{Wr}(GI/GI/1) \approx Ca^2 Cs^2 \overline{Wr}(M/M/1) + (1 - Cs^2) Ca^2 \overline{Wr}(M/D/1) + \frac{(\theta\rho Cs^2 + 2\lambda)(1 - Ca^2)}{(\theta\rho + 2\lambda) g_{01}} \overline{Wr}(D/M/1)$$

$$\overline{Wr}(GI/GI/1) \approx Ca^2 Cs^2 \overline{Wr}(M/M/1) + Ca^2 (1 - Cs^2) \overline{Wr}(M/D/1) + (1 - Ca^2) Cs^2 \overline{Wr}(D/M/1)$$

qui donnent, en alternance, des résultats satisfaisants dans le domaine $0 \leq Ca^2 (Cs^2) \leq 1$. Ceci confirme la validité de la méthode d'approximation des deux premiers moments dans ce domaine. De plus, le résultat positif obtenu pour $Cs^2 = 1.5$ et $Ca^2 = 2$ suppose l'extension de cette validité au delà de la valeur Cs^2 (Ca^2) = 1. Cette supposition reste, néanmoins, à confirmer par des calculs plus détaillés afin de délimiter le domaine de validité.

Ce travail constitue un pas important dans la modélisation et l'évaluation de systèmes présentant des arrivées et des services non exponentiels. Les futures recherches pourraient s'orienter vers une unification des différentes approximations proposées dans cette thèse à l'image de ce qui a été déjà établi pour les systèmes classiques **[KIM 1994]**. Des investigations plus détaillées sont souhaitables dans le but d'élargir le domaine de validité de l'approximation des deux moments. Enfin, cette même approche est à encourager afin de connaître le degré de validité de la méthode en considérant un coefficient de variation relatif à la distribution des inter rappels (système GI/GI/1 avec rappel général).

Références

[A]

[AIS 1994] A. Aissani, "A survey on retrial queueing models". Journées de statistiques appliquées, 1-11, Alger (16-18 Avril 1994).

[AIS 1996] A. Aissani, "Propriétés de second ordre des modèles d'attente, application à la conception des systèmes de production". 4ème Rencontre de Recherche Opérationnelle, USTHB Alger, 5-10 (06-08 Octobre 1996).

[ALE 1974] A. M. Aleksandrov, "A queueing system with repeated orders". Eng. Cybernet. Rev. 12 (N°3), 1-4 (1974).

[ART 1994] J. R. Artalejo et G. I. Falin, "Stochastic decomposition for retrial queues". Top 2 (N°2), 1-14 (1994).

[ART 1995] J. R Artalejo, "A queueing system with returning customers and waiting line". Oper. Res. Letters 17, 191-199 (1995)

[B]

[BAN 1996] J. Banks, J. S. Carson et B. L. Nelson, "Discret-event system simulation". Prentice Hall , New Jersey (1996).

[BAY 1990] B. Baynat et Y. Dallery. “A unified view of product-form approximation techniques for general closed queueing networks”. Technical Report 90. 48, Institut Blaise Pascal, Paris (Octobre 1990).

[BAY 2000] B. Baynat, “Théorie des files d’attente: des chaînes de Markov aux réseaux à forme produit”. Hermes Sciences Publications, Paris (2000).

[BEN 1995] E. Benyoucef, “Approximation discrète du modèle M(t)/G/1 avec rappels”. Mémoire d’Ingéniorat en recherche opérationnelle, Blida (Juin 1995).

[BEN 2000] M. Benaouicha, “Estimation de la stabilité forte dans un système d’attente G/M/1”. Mémoire de Magister en Mathématiques Appliquées, Université A/Mira de Béjaia (Octobre 2000).

[BEN 2002] B. Benameur, “Systèmes de files d’attente avec arrivées négatives et rappels”. Thèse de Magister, USTHB Alger (Mai 2002).

[BER 2000] L. Berdjoudj, “Stabilité forte dans les systèmes de files d’attente avec rappels”. Mémoire de Magister en Mathématiques Appliquées, Université A/Mira de Béjaia (Octobre 2000).

[BJÖ 1964] M. Björklund et A. Elldin, “A practical method of calculation for certain types of complex common control systems”. Ericsson Technics 20, 3-75 (1964).

[BOX 1979] O. J. Boxma, J. W. Cohen et N. Huffels, “Approximation of the mean waiting time in an M/G/s queueing system”. Oper. Res. 27, 1115-1127 (1979).

[BRO 1948] E. Brockmeyer, H. L. Halstrom et A. Jensen, “The life and works of A. K. Erlang”. Trans. Dan. Acad. Techn. Sci. N°2, Copenhagen, 138 (1948).

[BUR 1983] D. Y. Burman et D. R. Smith, “Asymptotic analysis of a queueing model with bursty traffic”. Bell Syst. Tech. J. 62, 1433-1453 (1983).

[BUR 1986] D. Y. Burman et D. R. Smith, “Asymptotic analysis of a queueing model with Markov modulated arrivals”. Oper. Res. 34, 105-119 (1986).

[BUZ 1992] J. A. Buzacott et J. G. Shanthikumar, “Design of manufacturing systems using queueing models”. Queueing Systems 12, 135-214 (1992).

[C]

[CHO 1979] Q. H. Choo et B. Conolly, “New results in the theory of repeated orders queueing systems”. J. Appl. Probab. 16, 631-640 (1979).

[CHO 1993] B. D. Choi, K. K. Park et C. E. M. Pearce, “An M/M/1 retrial queue with control policy and general retrial times”. Queueing Systems 14, 275-292 (1993).

[CLO 1948] C. Clos, “An aspect of the dialing behaviour of subscribers and its effect on the trank plant”. Bell Syst. Tech. J. 27, 424-445 (1948).

[COH 1957] J. W. Cohen, “Basic problems of telephone traffic theory and the influence of repeated calls”. Philips Telecommunication Rev. 18 (N°2), 49-100 (1957).

[COS 1974] G. P. Cosmetatos, “Approximate equilibrium results for the multi-server queue GI/M/r”. Oper. Res. Quarterly 25, 625-634 (1974).

[COS 1976] G. P. Cosmetatos, “Some approximate equilibrium results for the multi-server queue M/G/r”. Oper. Res. Quarterly 27, 615-620 (1976).

[COS 1977] G. P. Cosmetatos, “Some approximate equilibrium results for the multi-server queue $E_m/E_k/r$”. Opsearch 14, 108-117 (1977).

[COS 1980] G. P. Cosmetatos et S. A. Godsave, “Approximations in the multi-server queue with hyper-exponential inter-arrival times and exponential service times”. J. Oper. Res. Soc. 31, 57-62 (1980).

[COX 1961] D. R. Cox et W. L. Smith, “Queues”. Chapman and Hall editions, Londres (1961).

[D]

[DSH 1995] J. H. Dshalalow, “Theory, methods and open problems”, CRC Press (1995).

[E]

[ELL 1967] A. Elldin, “Approach to the theoretical description of repeated call attempts ”. Ericsson Technics 23 (N°3), 346-407 (1967).

[ENG 1918] T. Engset, “The probability theory of computing the number of switching equipments in automatic telephone exchanges”. E. T. Z. 31, 304-305 (1918).

[ERA 1996] P. J. Erard et P. Deguenon, “Simulation par événements discrets”. Presses Polytechniques et universitaires romandes (1996).

[ERL 1917] A. K. Erlang, “Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges”. Post Office Electrical Engineer’s Journal 10, 189-197 (1917).

[F]

[FAL 1978] G. I. Falin, “The output flow of a single-line queueing system when there are secondary orders”. Eng. Cybernet. Rev. 16 (N°5), 64-67 (1978).

[FAL 1979] G. I. Falin, “A single-line system with secondary orders”. Eng. Cybernet. Rev. 17 (N°2), 76-83 (1979).

[FAL 1980] G. I. Falin, “An M/M/1 queue with repeated calls in the presence of persistence function”. Paper # 1606-80, All-Union Institute for Scientific and Technical Information, Moscou (1980).

[FAL 1986] G. I. Falin, “A probabilistic model for investigation of load of subscriber’s lines with waiting places”. In: Probability Theory, Stochastic Processes and Functional Analysis, Moscow State University (Moscou 1985).

[FAL 1990] G. I. Falin, “A survey on retrial queues”. Queueing systems 7, 127-168 (1990).

[FAL 1994] G. I. Falin, M. Martin-Diaz et J. R. Artalejo, “Information theoretic approximations for the M/G/1 retrial queue”. Acta Informatica 31, 559-571 (1994).

[G]

[GAV 1959] D. B. Gaver, “Imbedded Markov chain analysis of waiting line process in continous time”. Ann. of Math. Stat. 30, 698-720 (1959).

[GEL 1976] E. Gelende, “A non-markovian process and its application to the approximation of queueing and computer system behavior”. TR-158 (IRIA), (1976).

[GRE 1987] B. S .Greeenberg et R. W. Wolff, “An upper bound on the performance of queues with returning costumers”. J. Appl. Probab. 24, 466-475 (1987).

[GRE 1989] B. S .Greeenberg, “M/G/1 queueing systems with returning costumers”. J. Appl. Probab. 26, 152-163 (1989).

[H]

[HAS 1979] O. Hashida et K. Kawashima, “Buffer Behavior with Repeated Calls”. Electronics and Communications in Japan, Vol. 62-B N°3, 27-35 (1979).

[HEY 1975] D. P. Heyman, “A diffusion model approximation for the GI/G/1 queues in heavy trafic”. Bell Syst. Tech. J. 54, 1637 (1975).

[HOK 1978] P. Hokstad, “Approximations for the M/G/m queue”. Oper. Res. 26, 510-523 (1978).

[K]

[KAP 1977] V. A. Kapyrin, “A study of the stationary distributions of a queueing system with recurring demands”. Cybernet. 13, 584-590 (1977).

[KEI 1968] J .Keilson, J. Cozzolino et H. Young, “A service system with unfilled requests repeated”. Oper. Res. 16, 1126-1137 (1968).

[KEL 1985] F. P. Kelly, “Auto-repeat facilities and telephone networks performance”. J. Roy. Statist. Soc. B48, 123-132 (1985).

[KEN 1953] D. G. Kendall, “Stochastic processes occurring in the theory of queues and their analysis by means of the imbedded Markov Chain”. Ann. Math. Statist. 24, 338-354 (1953).

[KIM 1984] T. Kimura, “The queueing network analyser : a survey (1)-(3) [in Japanese]”. Commun. Oper. Res. Soc. Japan 29, 366-371, 431-439, 494-500 (1984).

[KIM 1986] T. Kimura, “A two moment approximation for the mean waiting time in the GI/G/s queue”. Manag. Sci. 32, 751-763 (1986).

[KIM 1987] T. Kimura, “Heuristic approximation for the mean delay in the GI/G/s queue”. Econ. J. Hokaido Univ. 16, 87-98 (1987).

[KIM 1991a] T. Kimura, “Approximations for the waiting time in the GI/G/s queue”. J. Oper. Res. Soc. Japan 34, 173-186 (1991).

[KIM 1991b] T. Kimura, “Approximating the mean waiting time in the GI/G/s queue”. J. Oper. Res. Soc. Japan 42,959-970 (1991).

[KIM 1992] T. Kimura, “Interpolation approximations for the mean waiting time in a multi-server queue”. J. Oper. Res. Soc. Japan 35, 77-92 (1992).

[KIM 1994] T. Kimura, “Approximations for multi-server queues: system interpolations”. Queueing Systems 17, 347-382 (1994).

[KIN 1962] J. F. C. Kingman, “Some inequalities for the GI/G/1 queue”. Biometrica 49, 315 (1962).

[KIN 1964] J. F. C. Kingman, “The heavy traffic approximation in the theory of queues”. In: Proc. Symp. on Congestion Theory, eds: W. L. Smith et R. I. Wilkinson, 137-169 (Chapel Hill 1964).

[KLE 1975] L. Kleinrock, “Queueing systems, vol. 1: Theory”. John Wiley Interscience Publication, New York (1975).

[KLE 1976] L. Kleinrock, "Queueing systems, vol. 2: Computer application". John Wiley Interscience Publication, New York (1976).

[KÖL 1974] J. Köllerstrom, "Heavy traffic theory for queues with several servers I". J. Appl. Prob. 11, 544-552 (1974).

[KOS 1947] L. Kosten, "On the influence of repeated calls in the theory of probabilities of blocking". De Ingenieur 59, p 01 (1947).

[KRA 1976] W. Kramer et M. Langeberch-Belz, "Approximate formula for the delay in the queueing system GI/G/1". Congressbook, 8th International Tele Congress, 235.1-235.8 (Melbourne) (1976).

[KUM 1994] S. Kumar et P. R. Kumar, "Performance bounds for queueing networks and scheduling policies". IEEE Transactions on Automatic Control, Vol. AC-39, 1600-1611 (Aout 1994).

[L]

[LAR 1997] R. C. Larson et A. R. Odoni, "Introduction to queueing theory and its applications, Chap. 4". Massachusetts Institute of Technology (1997).

[LAZ 1984] E. D. Lazowska, J. Zahorjan, G. S. Graham et K. C. Sevcik, "Quantitative system performance: Computer system analysis using queueing network models". Prentice Hall, Englewood Cliffs, N. J. (1984).

[LEE 1957] A. M .Lee et P. A. Longton, "Queueing process associated with airline passenger check-in". Oper. Res. Quarterly 10, 56-71 (1957).

[LeG 1970] P. Le Gall, "L'influence des répétitions d'appel dans l'écoulement du trafic téléphonique". Ann. Telecom. Paris, Vol. 25 (N°9), 339-348 (1970).

[LIN 1952] D. V. Lindley, "The theory of queues with a single server". Proc. Camb. Phil. Soc. Math. Phys. Sci. 48, 277 (1952).

[LUB 1984] J. Lubacz et J. Roberts, "A new approach to the single server repeat attempts system with balking". Proc. 3rd Int. Seminar on Teletraffic Theory, 290-293 (Moscou 1984).

[M]

[MAA 1973] E. Maaløe, "Approximation formulae for estimation of waiting time in multiple-channel queueing systems". Manag. Sci. 19, 703-710 (1973).

[M'BA 1995] S. M'barki, "Approximations des files d'attente avec rappels: Interpolation de systèmes". Mémoire d'Ingéniorat (Math. Appl. Rech. Oper.), Blida (juin 1995).

[MAR 1976] W. G. Marchal, "En approximate formula for waiting time in single server queues ". A. I. I. E. Trans. 8, 473 (1976).

[MIY 1986] M. Miyazawa, "Approximation for the queue-length distribution of an M/GI/s queue by the basic equations". J. Appl. Prob. 23, 443-458 (1986).

[P]

[PAG 1972] E. Page, "Queueing theory in OR". Opnl Res. Series, Ed : K. B. Haley (1972).

[PEL 1994] J. Pellaumail, P. Boyer et P. Leguesdron, "Reseaux ATM et P-simulation". Hermes Paris (1994).

[POU 1987] B. Pourbabai, "Analysis of a G/M/k/o queueing loss system with heterogeneous servers and retrials". Inter. Journal of Systems Science 18, 985-992 (1987).

[POU 1988] B. Pourbabai, "Asymptotic analysis of G/G/k queueing loss system with retrials and heterogeneous servers". Inter. Journal of Systems Science 19 (N°6), 1047-1052 (1988).

[POU 1989] B. Pourbabai, "Tandem behaviour of a telecommunication system with repeated calls: A markovian case with buffers". J. Oper. Res. Soc. 40 (N°7), 671-680 (1989).

[POU 1990] B. Pourbabai, "Tandem behaviour of a telecommunication system with finite buffers and repeated calls". Queueing systems 6, 89-108 (1990).

[R]

[REI 1974] M. Reiser et H. Kobayachi, "Accurasy of the diffusion approximation for some queueing systems". IBM Jl. Res. Dev. 18, 110 (1974).

[REI 1988] M. I. Reiman et B. Simon, "An interpolation approximation for queueing systems with Poisson input". Oper. Res. 36, 454-469 (1988).

[RIO 1962] J. Riordan, "Stochastic service systems". J. Wiley, New York (1962).

[RUS 1984] P. Ruskov, K. Yanev, B. Dimitrov et K. Boyanov, "A model for investigating local area computer networks". Control Systems and Machines Vol 5, 37-40 (1984).

[S]

[SAA 1961] T. Saaty, "Elements of queueing theory and applications with applications". McGraw-Hill Book Company (1961).

[SAK 1977] H. Sakasegawa, "An approximation formula $L_q \approx \alpha\rho^{\beta}/(1-\rho)$". Anl. Inst. Statist. Math. Tokyo 29 A, 67 (1977).

[SAK 1978] M. Sakarovitch, "Techniques mathématiques de la recherche opérationnelle, Vol. 5 : Processus aléatoires ". E.N.S.I.M.A.G., Grenoble (1978).

[SAV 1981] C. H. Saver et K. M. Chandy, "Computer systems performance modelling". Prentice Hall, Englewood Cliffs, N. J. (1981).

[SEE 1985] L. P. Seelen, H. C. Tijms et M. H. Van Hoorne, "Tables of multi-server queues". North-Holland (1985).

[SHA 1980] J. G. Shanthikumar et J. A. Buzacott, "On the approximations to the single server queue". Int. J. Prod. Res. 18 (N°6), 761-773 (1980).

[STE 1983] S. N. Stepanov, "Numerical methods of calculation for systems with repeated calls". Nauka, Moscou (1983).

[STO 1976] D. Stoyan, "Approximations for M/G/s queues". Math. Operationsforsch. Statistik 7, 587-594 (1976).

[T]

[TAK 1977] Y. Takahashi, "An approximation formula for the mean waiting time of an M/G/c queue". J. Oper. Res. Soc. Japan 20, 150-163 (1977).

[TIJ 1986] H. C. Tijms, “Stochastic modelling and analysis: a computationnal approach”. Wiley Series in Probability and Mathematical Statistics (1986).

[W]

[WHI 1983a] W. Whitt, “The queueing network analyser”. Bell System Tech. J. 62, 2779-2815 (1983).

[WHI 1983b] W. Whitt, “Performance of the queueing network analyser”. Bell System Tech. J. 62, 2817-2843 (1983).

[WHI 1984a] W. Whitt, “On approximations for queues I: Extremal distributions ”. AT & T Bell Lab. Tech. J. 63, 115-138 (1984).

[WHI 1984b] W. Whitt, “On approximations for queues III: Mixture of exponential distributions ”. AT & T Bell Lab. Tech. J. 63, 163-175 (1984).

[WIL 1956] R. I. Wilkinson, “Theories for toll traffic engineering in the U. S. A”. Bell System Techn. J. 35 (N°2), 421-507 (1956).

[WIL 1968] R. I. Wilkinson et R. Radnik, “Customers’ retrials in toll circuit operation”. IEEE Int. Conf. On Communications (1968).

[Y]

[YAN 1987] T. Yang et J. G. C. Templeton, “A survey on retrial queues”. Queueing Systems 2, 201-233 (1987).

[YAN 1994] T. Yang, M. J. M. Posner, J. G. C. Templeton et H. Li, “An approximation method for the M/G/1 retrial queue with general retrial times”. Europ. J. of Oper. Res. 76, 552-562 (1994).

[YU 1977] T. S. Yu, “On accurasy improvemet and applicability conditions of diffusion approximation with applications to modelling of computer systems”. TR-129 (Digital Systems Laboratory, Saintford University) (1977).

Annexe A : Rappels de probabilités

A.1. Formule de Little

Considérons un système général, caractérisé par (voir figure A.1):

- Un nombre moyen d'arrivées par unité de temps : λ
- Le temps moyen d'attente d'un client dans le système : $\overline{W}_s$
- Le nombre moyen de clients présents dans le système : $\overline{N}_s$.

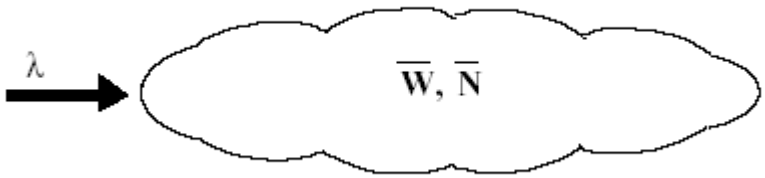


***Figure A.1* : file d'attente.**

Le résultat très général établi par Little stipule que les variables λ, $\overline{N}_s$ et $\overline{W}_s$ sont liées par la relation :

$$\overline{N}_s = \lambda \overline{W}_s \qquad (A.1)$$

Cette formule, qui ne sera pas démontrée formellement ici, est intéressante pour différentes raisons :

- La théorie des files d'attente donne assez facilement des résultats sur les probabilités d'état, desquelles on déduit le nombre moyen de clients dans un système. La formule de Little permet donc de calculer le temps de réponse (ou le temps d'attente) d'un système à partir de l'évaluation du nombre moyen.

- Il est extrêmement difficile de mesurer sur un système réel le temps de réponse (il faut pour cela mémoriser le temps d'arrivée de chaque client pour calculer à sa sortie du système des statistiques de temps de réponse). Par contre les données sur le nombre moyen d'arrivées par unité de temps d'une part et sur le nombre moyen de clients présents s'obtiennent facilement (par des mesures de comptage ou par échantillonnage). On peut donc évaluer le temps de réponse en faisant des mesures sur des variables beaucoup plus faciles à évaluer.

- Le résultat de Little ne dépend pas de la variance du processus d'arrivée mais seulement de sa moyenne. Il ne dépend pas non plus du temps de service d'un client dans le système.

- Cette formule est valable sous les seules conditions de stationnarité et d'existence des moyennes stationnaires et d'existence des moyennes stationnaires $\lambda, \overline{N}_S, \overline{W}_S$. En particulier, elle ne dépend pas de la discipline de service, pourvu que celle-ci ne modifie pas la durée de service des clients **[SAA 1961]**.

A.2. Processus de naissance et de mort

Ce processus est la fusion de deux processus stochastiques à temps continu et à états discrets n=1, 2, …que sont le processus de naissance et le processus de mort.

On qualifie un processus de processus de naissance, si celui-ci est caractérisé par l'apparition d'un individu, au sein d'une population, selon une certaine loi. Le processus de mort définit le phénomène de disparition au sein d'une population. Ces deux processus sont caractérisés par deux propriétés importantes :

- Ils sont sans mémoire.
- A partir d'un état donné n, les transitions ne sont possibles que vers l'un des états voisins n+1 et n-1 (si $n \geq 1$).

Les processus de Poisson sont des exemples simples de processus de naissance (arrivée) et de mort (départ).

Définition : Soit un processus stochastique $((X(t))_{t\geq 0}$, à états discrets n = 0,1,2,… et homogène dans le temps (i.e. la probabilité $P[X(t+s)=j/X(s)=i]=p_{ij}(t), s\geq 0, t\geq 0$ ne dépend pas de s), alors $(X(t))_{t\geq 0}$ est un processus de naissance et de mort s'il satisfait les postulats suivants :

$$\begin{aligned} p_{i,i+1}(\Delta t) &= \lambda_i\,\Delta t + o(\Delta t) && (i\geq 0)\\ p_{i,i-1}(\Delta t) &= \mu_i\,\Delta t + o(\Delta t) && (i\geq 1)\\ p_{i,i}(\Delta t) &= 1-(\lambda_i+\mu_i)\Delta t + o(\Delta t) && (i\geq 0) \end{aligned} \qquad (A.2)$$

où λ_i et μ_i sont les taux de transition dit généralement *taux de naissance* et *taux de mort*.

A.3. Lois de probabilités

Dans ce paragraphe, nous présentons quelques lois de probabilités et leurs propriétés. Nous nous limitons à celles utilisées dans le cadre de ce travail.

A.3.1. Loi uniforme sur [0,1]

Soit U une variable aléatoire distribuée uniformément sur [0,a]. Sa densité de probabilité est définie par :

$$f(x)=\begin{cases}\dfrac{1}{a} & \text{si } x\in[0,a],\\ 0 & \text{sin on.}\end{cases} \qquad (A.3)$$

et sa fonction de répartition par :

$$F(x) = \begin{cases} 0 & \text{si } x \in]-\infty, 0[, \\ \dfrac{x}{a} & \text{si } x \in [0, a], \\ 1 & \text{si } x \in [a, +\infty[. \end{cases} \qquad \text{(A.4)}$$

Son espérance est $E(U) = \dfrac{a}{2}$ et sa variance $V(U) = \dfrac{a^2}{12}$.

A partir des variables aléatoires uniformément distribuées, on peut générer toute variable aléatoire de loi arbitraire. D'où l'importance de la loi uniforme en simulation (voir **[BAN 1996]**)

A.3.2. Loi exponentielle

Soit X une variable aléatoire distribuée suivant une loi exponentielle de paramètre λ. Sa densité de probabilité est définie par :

$$f(x) = \begin{cases} \lambda \exp(-\lambda x) & \text{si } x \in [0, +\infty[, \\ 0 & \text{sin on.} \end{cases} \qquad \text{(A.5)}$$

et sa fonction de répartition par :

$$F(x) = \begin{cases} 1 - \exp(-\lambda x) & \text{si } x \in [0, +\infty[, \\ 0 & \text{sin on.} \end{cases} \qquad \text{(A.6)}$$

Son espérance est $E(X) = \dfrac{1}{\lambda}$, sa variance est $V(X) = \dfrac{1}{\lambda^2}$ et son coefficient de variation $\dfrac{V(X)}{E(X)^2} = 1$.

La génération de cette variable aléatoire se fait à l'aide de la technique d'inversion, i.e :

i) Poser F(X) = U où U est le nombre aléatoire uniformément distribué sur l'intervalle [0,1].

ii) Résoudre l'équation F(X)=U :

$$1 - \exp(-\lambda X) = U$$
$$\exp(-\lambda X) = 1 - U$$
$$X = \frac{-1}{\lambda} \ln(1 - U) \qquad \text{(A.7)}$$

L'équation (A.7) est appelée *générateur aléatoire pour la distribution exponentielle*.

iii) Générer des nombres aléatoires $U_1, U_2, \ldots, U_i, \ldots$ et calculer

$$X_i = \frac{-1}{\lambda} \ln(1 - U_i), \qquad i = 1, 2, 3, \ldots \qquad \text{(A.8)}$$

L'équation (A.8) peut être remplacée par $X_i = \frac{-1}{\lambda} \ln U_i$ puisque U_i et $(1\text{-}U_i)$ sont uniformément distribués sur [0,1].

A.3.3. Lois Gamma et d'Erlang

La variable aléatoire X est distribuée suivant une loi Gamma de paramètres α et β positifs. Si sa densité de probabilité est de la forme :

$$f(x) = \begin{cases} \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x) & \text{si } x \in]0,+\infty[, \\ 0 & \text{sin on.} \end{cases} \qquad \text{(A.9)}$$

sa fonction de répartition est alors :

$$F(x)=\begin{cases}1-\int_{x}^{+\infty}\frac{\beta^{\alpha}}{\Gamma(\alpha)}t^{\alpha-1}\exp(-\beta t)\,dt & \text{si } x\in\,]0,+\infty[,\\ 0 & \text{sin on.}\end{cases} \qquad (A.10)$$

où $\Gamma(\alpha)$ est la fonction Gamma, définie par :

$$\Gamma(\alpha)=\int_{0}^{+\infty}x^{\alpha-1}\exp(-x)dx \qquad (A.11)$$

et possédant les propriétés suivantes :

$$\begin{cases}\Gamma(\alpha)=(\alpha-1)\Gamma(\alpha-1) & \forall\,\alpha,\\ \Gamma(\alpha)=(\alpha-1)! & \text{si } \alpha\in N\end{cases} \qquad (A.12)$$

L'espérance de X est $E(X)=\frac{\alpha}{\beta}$, sa variance $V(X)=\frac{\alpha}{\beta^2}$ et son coefficient de variation $\frac{V(X)}{E(X)^2}=\frac{1}{\alpha}$.

Si $\alpha=k$ est un entier positif non nul et $\beta=\mu$, la relation (A.9) s'écrit alors :

$$f(x)=\begin{cases}\frac{\mu(\mu x)^{k-1}}{(k-1)!}\exp(-\mu x) & \text{si } x\in\,]0,+\infty[,\\ 0 & \text{sin on.}\end{cases} \qquad (A.13)$$

Dans ce cas, on dira que X est distribuée sur loi d'Erlang à k étapes de paramètre μ, notée *k-Erlang (μ)* ou E_k *(μ)*.

La loi d'Erlang est très utilisée dans la théorie des files d'attente. Une loi de service d'Erlang d'ordre k se modélise comme un ensemble de serveurs exponentiels placés en série. Cela signifie que le service est constitué par la somme de k services indépendants exponentiels de même taux (ce qui signifie qu'ils ont le même

paramètre, pas que la durée de service de chacun est la même), comme montré sur la figure A.2. Cette loi modélise des services plus réguliers qu'une exponentielle, donc dont le coefficient de variation est inférieur à 1.

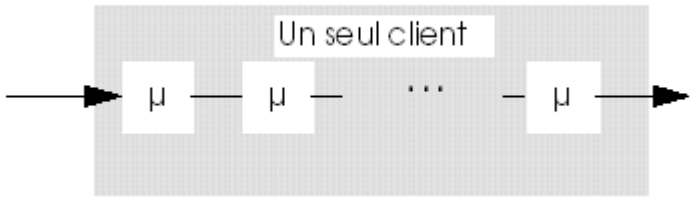


Figure A.2: service d'Erlang

Annexe B : Introduction aux méthodes par diffusion

Parmi les méthodes approximatives, les techniques dites de diffusion permettent des approches simples, et souvent très correctes sur le plan de la précision. Le principe de base est de remplacer le processus de base (un processus discrétisé avec sauts) par un processus de Markov continu.

Les équations du système se mettent sous la forme d'équations aux dérivées partielles, dont la complexité ne dépend pas (ou peu) des paramètres fondamentaux (comme les caractéristiques des lois).

B.1. Généralité sur le processus de diffusion

L'idée de base est d'utiliser des approximations dites *fluides*, de la variation dans le temps de la moyenne d'un processus. Les variations sur cette moyenne sont introduites en tenant compte des deux premiers moments des distributions des arrivées et des services. Les variables aléatoires que nous voulons approcher sont obtenues par des sommes de variables aléatoires indépendantes, donc dont la distribution tend asymptotiquement vers une loi normale.

Cette méthode sera utilisée pour obtenir le nombre de clients $N_s(t)$ et le temps d'attente $W_s(t)$. Dans le cas du nombre de clients, $N_s(t)$ est la différence entre le nombre de clients arrivés et le nombre de clients ayant achevé leur service, depuis l'origine. Cette superposition de processus indépendants tend vers un processus de distribution normale. La notion d'indépendance doit être prise avec quelques précautions : les arrivées et les services ne sont pas réellement indépendants (il n'y a pas de service s'il n'y a pas eu

d'arrivées). Il est donc naturel que ce type d'approximation soit particulièrement adapté aux systèmes chargés.

Le processus $N_s(t)$ sera approché par un processus continu X(t), dont les variations $\Delta X(t)$ sont normalement distribuées, de moyenne $\beta\Delta t$ et de variance $\alpha\Delta t$. Le processus X(t) est défini par l'équation différentielle stochastique $dX(t) = \beta\, dt + z(t)(\alpha\, dt)^{1/2}$ où z(t) est un bruit blanc (processus gaussien de moyenne nulle et de variance unité).

S'il n'y a pas de condition de frontière imposée au processus X(t), celui-ci décrit alors un mouvement brownien dont la distribution de probabilité $f(x_0, x, t)$ doit satisfaire l'équation :

$$\frac{\delta f(x_0, x, t)}{\delta t} = \frac{\alpha}{2}\frac{\delta^2 f(x_0, x, t)}{\delta x^2} - \beta\frac{\delta f(x_0, x, t)}{\delta x} \qquad \text{(B.1)}$$

Où x_0 est la valeur initiale et $f(x_0,x,t) = \text{Prob}\ \{x \le X(t) \le x+dx \mid X(0) = x_0\}$

Il s'agit de l'équation de diffusion ou *équation de Fokker-Plank*.

Dans la suite, nous supposerons que le système est vide à l'origine. Un des problèmes qui se pose est le comportement à l'origine (pour X=0). En effet, le processus est par nature positif, mais les équations ne le reflètent pas. Il faut donc ajouter des hypothèses sur le comportement à la frontière (et un problème analogue apparaît dans le cas de files à capacité limitée). Deux classes de solutions sont utilisées :

- des barrières réfléchissantes : le processus continue dans le domaine positif (comme si on ne considérait que sa valeur absolue).

- des barrières absorbantes : dans ce cas, le processus reste à l'origine un certain temps, puis saute directement de la frontière à un point de l'axe réel. L'équation

de diffusion doit alors être modifiée et complétée pour tenir compte de l'absorption et du retour instantané :

$$\frac{\delta f(x,t)}{\delta t} = \frac{\alpha}{2}\frac{\delta^2 f(x,t)}{\delta x^2} - \beta\frac{\delta f(x,t)}{\delta x} + \Lambda\, Q(t)\,\delta(x-\ell) \qquad \text{(B.2)}$$

où $\delta(x-\ell)$ est la fonction de Dirac au point ℓ (point de retour).

Λ^{-1} est le temps moyen de la période oisive.

Q(t) est la probabilité d'être sur la frontière au temps t.

L'intégration de l'équation étant faite, reste la question de la discrétisation de la distribution de probabilité continue.

B.2 Cas d'une file unique

Soit une file d'attente GI/GI/1 avec la discipline de service FIFO. Les notations utilisées sont les suivantes :

$1/\lambda$ = moyenne des inter-arrivées.

$1/\mu$ = moyenne des temps de service.

Va = variance des inter-arrivées.

Vs = variance des temps de service.

$Ca^2 = Va\ \lambda^2$ et $Cs^2 = Vs\ \mu^2$, sont les carrés des coefficients de variation (CCV) respectivement des inter-arrivées et des temps de service.

Appelons A (t) et D (t) le nombre de clients entrés et sortis de la file entre les instants 0 et t. Le nombre de clients présents est alors N_s (t) = A (t) – D (t). Si on regarde l'évolution de ce nombre entre t et $t+\Delta t$, qu'on appellera ΔN_s (t), nous obtenons :

$$\Delta N_s(t) = N_s(t+\Delta t) - N_s(t) = \Delta A(t) - \Delta D(t) \qquad \text{(B.3)}$$

Suivant la théorie expliquée ci-dessus, ΔN_s (t) est bien approché par une loi normale de moyenne $\beta\,\Delta t = (\lambda-\mu)\,\Delta t$ et de variance $\alpha\,\Delta t = (\lambda\, Ca^2 + \mu\, Cs^2)\,\Delta t$ (dans la mesure où la durée Δt est suffisamment grande pour qu'un nombre important d'entrées et de sorties se produisent). La densité de probabilité f (x,t) doit alors satisfaire l'équation :

$$\frac{\delta f(x,t)}{\delta t} - \frac{\alpha}{2}\frac{\delta^2 f(x,t)}{\delta x^2} + \beta\frac{\delta f(x,t)}{\delta x} = 0 \qquad \text{(B.4)}$$

La solution doit satisfaire la condition f(x,t) = 0 pour x > 0. La voie la plus naturelle pour obtenir une solution est de résoudre l'équation en prenant x = 0 comme une barrière réfléchissante. Cependant d'autres techniques ont été proposées. Dans ce cas, et dans l'hypothèse $t \to \infty$ (régime stationnaire), la solution est :

$$f(x) = -\gamma \exp(\gamma x) \qquad \text{(B.5)}$$

où $\gamma = 2\beta/\alpha < 0$ si $\rho = \lambda/\mu < 1$

A partir de cette distribution continue, reste à évaluer les probabilités discrètes. Plusieurs méthodes ont été proposées :

1. En intégrant la fonction f(x) sur l'intervalle [k, k+1].
2. En intégrant f(x) sur [k-1, k], avec p(0) = 0
3. En posant P(k) = f(k)

Dans tous les cas, il faudra vérifier que la somme des P(k) est bien égale à 1 (ce qui est vérifié dans les cas 1 et 2).

Dans le cas de la barrière réfléchissante, il est préférable d'utiliser la première méthode. Nous obtenons :

$$P(k) = \int_{k}^{k+1} f(x)\,dx = (1-\hat{\rho})\,\hat{\rho}^{\,k}, \qquad k = 0,1,2,... \tag{B.6}$$

où $\hat{\rho} = \exp(2\beta/\alpha) = \exp[2(\lambda-\mu)/(\lambda\,Ca^2 + \mu\,Cs^2]$.

La forme de cette distribution est bien connue, et d'un type analogue à celle de la M/M/1, et les résultats sont satisfaisants dans ce cas. Cependant la valeur de P(0) s'altère rapidement lorsque les CCV s'éloignent de 1. Il est recommandé d'ajuster la distribution de la façon suivante (Kobayashi) :

$$P(k) = \begin{cases} 1-\rho & \text{si } k = 0 \\ \rho\,(1-\hat{\rho})\,\hat{\rho}^{\,k-1} & \text{si } k \geq 1 \end{cases} \tag{B.7}$$

Une autre approche à ce problème est proposée par Gelenbe. Il place une barrière absorbante au point x=0, qui introduit une masse en ce point dans la distribution de probabilité. Le temps de résidence au point x=0 est exponentiellement distribué de paramètre λ ; une fois ce temps écoulé, le processus saute au point x=1. La masse au point 0 correspond à la probabilité que le système soit vide et le saut au point 1 correspondant à l'arrivée d'un client. La distribution de probabilité à l'état d'équilibre devient :

$$f(x) = \begin{cases} \rho\,(\exp[-\gamma]-1)\exp[\gamma x] & \text{si } x \geq 1 \\ \rho\,(1-\exp[\gamma x]) & \text{si } 0 \leq x \leq 1 \\ P(0) = 1-\rho \end{cases} \tag{B.8}$$

Les trois politiques de discrétisation ont été essayés. La politique n°3 redonne exactement les mêmes résultats que la solution ajustée par Kobayashi. On obtient également :

$$\bar{n}_1 = \frac{\rho}{1-\hat{\rho}} \tag{B.9}$$

La méthode n° 2 de discrétisation donne les valeurs suivantes :

$$P(k) = \begin{cases} P(0) = 1-\rho \\ P(1) = \dfrac{\rho\,(\rho\,Ca^2 + Cs^2)}{2\,(1-\rho)} \\ P(k) = \dfrac{\rho\,(\rho\,Ca^2 + Cs^2)}{2\,(1-\rho)} \exp[k\gamma]\ (1-\exp[-\gamma])^2 \qquad \text{pour } k \geq 2 \end{cases} \tag{B.10}$$

On obtient alors le nombre moyen de clients :

$$\bar{n}_2 = \sum_{k=1}^{\infty} k\,P(k) = \rho\left[1 + \frac{\rho\,Ca^2 + Cs^2}{2\,(1-\rho)}\right] \tag{B.11}$$

Dans le cas $Ca^2 = 1$, la forme est similaire à celle de Polaczek-Khintchine (avec une erreur de $1/2\,\rho\,Cs^2$).