Université Badji Mokhtar-Annaba-

Département d'Informatique

Réseau et Sécurité Informatique

Evaluation de Performance

Dr. STIHI Nadjet

13 octobre 2021

Table des matières

Table des figures

Introduction générale

L'objectif de ce module est de sensibiliser les étudiants aux problèmes de modélisation et d'évaluation des performances des systèmes réels tels que les systèmes informatiques, les réseaux de communication et les systèmes de production. Il se propose de répondre aux questions suivantes : Pourquoi évaluer les performances d'un système ? Dans quels cas cela est-il nécessaire ? Comment modéliser un système ? Quel type de modèle utiliser ? Comment analyser le modèle ? Ce cours est une introduction à la modélisation de systèmes complexes et l'évaluation de leurs performances.

Chapitre 1

Introduction à la modélisation et l'évaluation de performance

Il s'agit dans ce chapitre de donner quelques éléments de réponse aux questions telles que : Qu'est ce que l'évaluation de performance ? Pourquoi évaluer les mesures de performance d'un système ? Comment évaluer ces mesures d'un système ? Comment analyser les résultats de cette évaluation ? On va aussi citer les différentes méthodes d'évaluation de performance.

1.1 Principes généraux

Le problème d'évaluation de performance se pose lors de la conception du design de l'architecture du système que pendant son fonctionnement.

Les questions que l'on doit se poser lors de l'évaluation de performance sont : quel est le meilleur choix de l'organisation ou du design de la machine ? quel est le système opérationnel à supporter et quelles sont les fonctions qu'il pourrait fournir ?

L'évaluation de performance d'un système de communication ou d'un système informatique nécessite une très bonne connaissance de celui-ci, aboutissant à une description prenant en compte le paramètres les plus importants permettant de déboucher sur les

critères de performance du système.

La première étape consiste à définir l'objectif de cette étude (dimensionnement d'un réseau, comparaison de deux protocoles d'accès pour le réseau sans fil, étude du compor-tememt d'un serveur en cas de surcharge, etc).

Il faut identifier les facteur agissant sur le système (paramètres d'entrée) et difinir la charge du système (le nombre moyen de transactions par seconde et le modèle de charge).

La dernière étape consiste à définir les mesures de performance à déterminer (taux d'occupation du serveur, débit effectif, pire temps de réponse, etc) en choisissant un outil d'évaluation adéquat.

Les études de performance sont nécessaires pour fournir des réponses aux questions de coût, de performance, de qualité de service et de sécurité, surgissant durant la vie d'un système.

Exemple 1.1 *Pour un système informatique à mémoire virtuelle, connaissant les paramètres d'entrée (vitesse du disque de pagination, taille de la page, puissance de la CPU, la capacité de la mémoire, ...), il est souhaitable d'évaluer le taux d'utilisation de la CPU. Pour un serveur web, il est important d'implémenter une bonne politique de gestion de cache. Dans ce cas, il est souhaitable d'évaluer, par exemle, le taux de hits, c'est à dire, le pourcentage de documents qui sont dans le cache lorsqu'on demande. Ces questions sont de grandes importances pour les organismes impliqués. Des réponces incorrectes engendrent de potentielles répercussions : problèmes de sécurité, performance, coû, etc.*

Pour faire de l'évaluation de performances, nous devons disposer de deux éléments :

Le système : c'est l'entité dont on évalue les performances. Il est considéré comme étant un ensemble de ressources partagées entre différentes tâches. La caractéristique commune pour de tels systèmes est la présence d'un temps d'attente pour l'accès à ces ressources partagées.

La charge : la charge du système est une description sans ambiguïté des inputs qui contiennent suffisamment de détails qui permettront l'évaluation de performances du système considéré.

La bonne connaissance de ces deux éléments permet l'évaluation de performances du système.

1.2 Le rôle de l'évaluation de performances

Le rôle essentiel de toute application informatique est d'accomplir les fonctions pour lesquelles elle a été conçue, et d'offrir une performance adaptée à un coût raisonnable. La performance d'une application est donc un facteur clé de son succès.

L'évaluation de performance peut intervenir à deux niveaux :

En conception : Le système n'existe pas et il s'agit de le créer en respectant le cahier de charges et de le dimensionner. Il est alors judicieux d'étudier le comportememt du système avant son déploiement sur le terrain afin de comprndre et régler les éventuels problèmes qui pourraient affecter le système.

Exemple 1.2 *pour concevoir un réseau de communication, on doit être en mesure de connaître le débit souhaite (*$800Mb/s$ *par exemle), capable de transférer différents tupes de médias, tels que la voix, la vidéo, ou des données tout en respectant les contraintes temps réels ainsi que les délais de transmission et assurant que les informations seront transmises avec une certaine qualité de service (avec un taux de perte prédéterminé). Du fait que les systèmes sont de plus en plus complexes, de calculer les indices de performances, afin de vérifier leur conformité avec le cahier des charges. En effet, un système sous-dimensionné n'est pas utilisable et inversememt un système sur-dimensionné entraînera un gaspillage d'argent inutile.*

En exploitation : À ce niveau, le système existe, mais on souhaite le tester ou le

modifier de telle manière à améliorer son fonctionnement. Il s'agit de concevoir un nouveau système répondant à de nouveaux objectifs.

Exemple 1.3 *modifier un serveur dont le taux d'exécution est insuffisant, en remplaçant son processeur par un autre deux fois plus puissant ou modifier un réseau de communication en remplaçant la bande passante par une autre de capacité plus importante de telle manière à satisfaire la demande.*

1.3 Différents types d'annalyse des systèmes

On distingue deux grands types d'analyse :

L'analyse qualitative : consiste à définir les propriétés structurelles et comportememtales du système, telles que l'absence de blocage (vivacité), les invariants du système, le comportememt fini

Exemple 1.4 *1. Dans un système un système informatique où deux processus s'exécutent en parallèle et partagent une ressource critique (deux utilisateurs qui se partagentune imprimemte), le modèle permettra alors de vérifier que, quelque soit l'état du système, la ressource critique ne peut être utilisée par plus d'processus (exclusion mutuelle).*

2. *De même, dans un réseau de communication, si la machine M_1 attend un message d'une machine M_2 pour poursuivre son precessus et en même temps la machine M_2 attend un message de la machine M_1 pour poursuivre son exécution, le système est dans un état d'inter-blocage (deadlock) et ne peut plus évaluer. Pour palier à ce blocage,la modification du protocolede communication s'impose.*

Ainsi, l'étude qualitative nous renseignera sur l'éventualité d'un tel état. Plusieurs autres propriétés qualitatives importantes existent : propriété d'équaté, d'inévitabilité, le formalisme *Réseau de Pétri*.

L'analyse quantitative : concerne le calcul des mesures que l'on veut effectuer sur un système informatique, permettant de quantifier ses performances : début, temps de réponce, taux d'utilisation de ses ressources ....

1.4 Les paramètres de performance

Parmi les paramètres de performance les plus fréquents, on trouve :

Le débit (de sortie) X **:** La vitesse à laquelle les clients quittent le système ;

Le temps de séjour R **:** Le temps écoulé entre l'instant d'arrivée d'un client et la fin de son service ;

Le nombre de clients Q **:** Le nombre de clients présents dans le système ;

Taux d'utilisation d'une ressource U **:** La proportion de temps pendant laquelle la ressource est occupée ;

La probabilité de rejet Pr **:** Le taux de clients refusés par le système.

On peut citer, à titre d'exemple :

Exemple 1.5 *Dans* ***les réseaux de communication****, le paramètre le plus important est* ***le temps de réponse*** *(délai d'acheminement) qui mesure le temps qui sépare l'émission d' un message de sa réception par le destinataire.*

Exemple 1.6 *Dans* ***les systèmes de production****, le paramètre de performance le plus retenu est* ***le débit en produit fini****.*

Exemple 1.7 *Dans un domaine plus pratique (celui* ***d'un guichet d'un organisme****), on peut distinguer deux paramètre importants :****temps d'attente, le nombre de clients****.*

1.5 Les étapes d'évaluation de performance

L'évaluation de performance d'un système se résume en trois étapes :

Etape 1 : Comprendre le fonctionnement du système.Substituer le système à un modèle que l'on pourra résoudre soit par des méthodes analytiques (mathématiques) soit par des simulations.

Etape 2 : Elaborer un modèle plus fidèle aux caractéristiques et fonctionnements du système.Appliquer une méthode de résolution ou de simulation (au modèle donné). L'objectif étant d'obtenir les performances du système (temps d'attente, temps de réponse... etc.)

Etape 3 : Evaluer les mesures de performance du système selon le formalisme du modèle.Analyser les résultats obtenus par une méthode quelconque.

L'évaluation de performance d'un système peut être schématisée de la façon suivante :

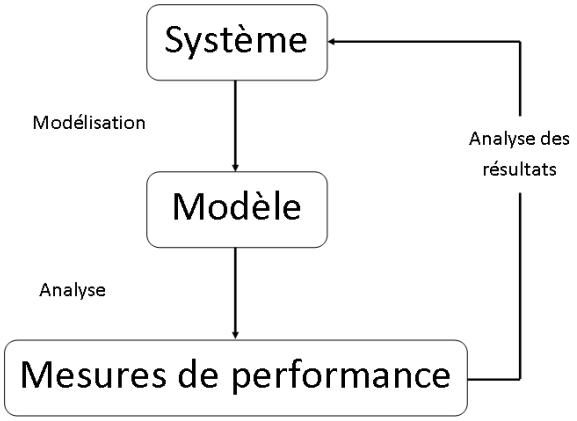

FIGURE 1.1 *Etapes de l'évaluation de performances d'un système.*

Ce schéma se décompose en une étape de modélisation permettant de passer du système au modèle, et une étape d'analyse de performance du modèle.

1.6 Modélisation et performance d'un système

Un modèle est la repeésentation du fonctionnemnt d'un système en s'appuyant sur des outils mathématique ou autres (telles que les simulations) permettant d'approcher le comportement du système. Suivant les objectifs souhaités, les modèles tendent à concentrer

les comportements en les paramètres permettant de cerner au mieux le fonctionnement à étudier.

La modélisation est substitution d'un système par un modèle que l'on pourra résoudre. La modélisation nécessite un support théorique afin de disposer d'outils indépendants des langages et des machines. Néanmoins, la modélisation d'un système est subjective et fait appel à l'expérience et à l'analyse de l'expert dans le domaine. Il est important de décomposer le problème global en problèmes simples correspondant à des sous systèmes faciles à modéliser. En effet, les systèmes informatiques devenant de plus en plus comlexes, il est pratiquement impossible de dimemsionner correctement de tels systèmes. Or, le dimensionnement des systèmes informatiques est nécessitéindustrielle. Une erreur dans le paramétrage peut d'avérer onéreux : mauvaise utilisation d'une CPU, d'une ligne de transmission,...

L'évaluation de performance nécessite la maîtrise du système (côté statique) et la connaissance de sa charge (côté dynamique). Le système est l'entité dont on évalue les performances. Il est considéré comme étant un ensemble de ressourses partagées entre eux différentes tâches (ordinateurs, serveurs de données(web, multimédia,...), commutateur,...). Le premier critère de performance du système est l'ensemble des conditions de stabilité. Un système non stable ne peut pas exploité ; il est alors inutile de déterminer ses mesures de performances. Par ailleurs, la charge de travail du système représente généralement le traffic en entrée qui pourrait être l'ensemble de message servis par un dispositif du réseau, le nombre de tâche à exécuter par un processeur. Il est généralement décrit par des lois de probabilités.

La première étape de la modélisation consiste à identifier et àexaminerles paramètres des compossants du système : la vitesse de la CPU, le temps d'accès et le taux de transfert de données des mémoires de stockage, la capacité d'une ligne de transmission,... ; ainsi que les types et les caractéristiques des ternimaux et équipements de communication. Il est également nécessaire de connaître les compossants software : l'algorithme d'ordonnacementdes tâches, l'algorithme de gestion de la mémoire, l'algorithme de distribution de

la CPU, l'algorithme d'ordonnancement du disque et du disque de pagination, les tailles de la page et du bloc et l'organisation des fichiers. De même, il est intéressant de trouver la quantité du traffic prévue pour chacune composantes : le taux d'arrivée des tâches, le temps de CPU par tâche, les besoins en espase mémoire, le taux de défaut de pages, le nombre de mouvements rotatifs du disque par seconde, le taux de demande du disque de pagination et le taux de transfert de données requis entre la mémoire centrale et les mémoires de stockage auxiliaires.

L'établissement de telles listes contenant les composants et les paramètres du système, pouvant avoir un impact sur la performance du système, est relativement facile. Néanmoins, il est difficile d'identifier un ensemble de paramètres critiques et il est encore plus difficile de déterminer les relations ou les équations décrivant le comportement du système qui permet d'évaluer les performances du système global en fonction de ses paramètres.

1.6.1 Choix du modèle

L'un des problèmes majeurs qui se pose lors de la modélisation d'un système est le choix du modèle. Ceci intervient aussi bien au niveau des entités (entités dynamiques) qu'au niveau des composants du système (entités statiques).

— **Niveau de détails des entités :** dans un réseau de communication numérique, l'information s'change sous forme de messages. L'entité de base du modèle serait donc le message. Mais généralement (selon le réseau considéré), ces messages peuvent se décomposer en paquets. Chaque paquet est constitué d'octets, lesquels sont constitués de bits, qui eux-mêmes voyagent sous la forme de signal électromagnétique. Quand on s'intéresse à l'évaluation de performance, on n'a pas à aller plus loin.

— **Niveau de détails des composants :** dans un réseau à commutation de paquets, les ressources peut être des nœuds de commutation et les liens de communica-

tion. Selon le niveau de détail du modèle, la ressource NŒUDS DE COMMUTATION peut se décomposer en des composants plus élémentaires, comme tampons, processeurs, bus, unités de traitement du signal,.... Dans l'exemple du serveur WEB, les ressources qui sont naturalement à modéliser sont plutô les unités centrales, mémoires, caches, disque,...

Cependant, le modèle choisi dépend également d'autres facteurs, comme :

— de la mesure de performance qu'on cherche à obtenir.

— du degré de connaissance qu'on a du système. Par exemple, si on sait que transmettre un paquet standard prend $1ms$, il est inutile de construire un modèle au niveau des octets pour trouver cette quantité ;

— du temps qu'on veut passer à analyser le modèle. En effet, s'il est vrai qu'un modèle très détaillé de la réalitédoit fournir des prédictions plus fines, il sera aussi plus compliqué et donc plus difficile à construire et àmettre au point et évidemment plus difficile à analyser. En effet, quelle que soit la méthode d'analyse choisie, analyse mathématique ou simulation sur ordinateur, plus le modèle sera gros, plus celui-cidevient fidèle et donc les performances obtenues seront proches de celles du système réel ; mais plus l'analyse du modèle est complexe, et plus on a besoin d'information sur le système initial et le modèle produira des données qu'il s'agira d'interpréter. Dans bien des cas, un modèle plus simple sera plus adéquat, même s'il est moins précis.

d'un côté, plus simplifie le modèle, moins celui-ci fidèle, mais plus son analyse est aisée.

1.7 Méthodes et techniques d'évaluation de performances

L'évaluation de performances peut se réaliser par deux manières :

1.7.1 Méthodes analytiques

Elles consistent à réduire le système en un modèle mathématique et à l'analyser numériquement.

Un modèle analytique est une relation fonctionnelle entre les paramètres du système et un critère de performance choisi en termes d'équations pouvant être résolues d'une manière analytique. Il existe de nombreux outils mathématiques permettant l'évaluation de performances :Les réseaux de pétri, Réseaux de files d'attente, les chaines de Markov, les automates.... etc.

Ces méthodes sont très efficaces et beaucoup plus rapides que la simulation, mais la solution n'existe que pour une classe très restreinte des systèmes.

Dans ces méthodes, l'évaluation de performance est alors réalisée, par la détermination des équations analytiques, qui l'est suivi d'une résolution analytique.

Il existe de nombreux outiles mathématiques permettant de telles évaluluations. Citons brièvement les approches les plus importantes :

— Les approches probabilistes (chaîne de Markov, files d'attente,...) ;

— Les réseaux de Petri.

Chaîne de Markov

Les processus stochastiques : Un processus stochastique décrit l'évolution temporelle de l'état d'un système à l'aide de variables aléatoires et de lois de probabilités.

Un processus stochastique $\{X(t), t \in T\}$ est une collection de variables aléatoires indexées par un paramètre t et définies sur un même espace de probabilités. Le paramètre est généralement interprété comme le temps et appartient à un ensemble ordonné T.

Il existe deux classes de processus stochastiques en fonction de l'ensemble T :

— processus stochastique à temps discret si $T \subset IN$ (entiers naturels).

— processus stochastique à temps continu si $T \subset IR^+$ (réels positifs).

Chaîne de Markov : Une chaine de Markov est un modèle fournissant un outil simple de modélisation et d'analyse de performances d'une classe particulière de système à événements discrets.

Un processus stochastique vérifiant la propriété suivante est appelé un processus de Markov ou processus markovien.

La propriété : pour toute séquence d'instant $t_{n+1} > t_n > t_{n-1} > t_{n-2} > \ldots > t_0$ et tout sous ensemble d'états $I \in S$, il est vrai que

$$\Pr\left(X\left(t+\Delta\right) \in I / X\left(u\right), 0 \leq u \leq t\right) = \Pr\left(X\left(t+\Delta\right) \in I / X\left(t\right)\right), \forall \Delta \geq 0 \qquad (1.1)$$

Files d'attente

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

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

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

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

La théorie des files d'attente est une technique de la recherche opérationnelle qui

permet de modéliser un système admettant un phénomène d'attente, de calculer ses performances et de déterminer ses caractéristiques, pour aider les décideurs dans leurs prises de décisions.

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

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

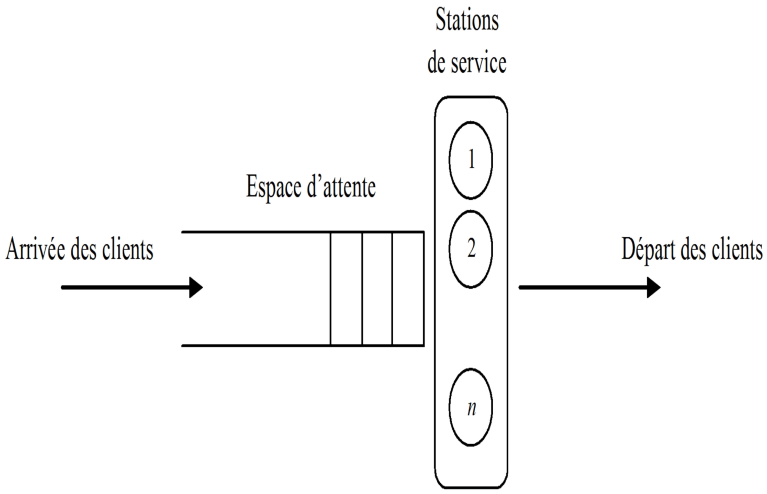

FIGURE 1.2 *Représentation d'une file d'attente*

Application des files d'attente

La théorie des files d'attente a de nombreuses applications dans :

— La gestion de trafic (réseaux de communication, compagnies aériennes, embouteillages , . . .).

— La planification (opérations sur des machines de production, programmes sur des ordinateurs, l'ordonnancement, par exemple les patients dans les hôpitaux, . . .).

— Le dimensionnement d'infrastructures (usines, . . .).

Les réseaux de Petri

Les réseaux de Pétri constituent un outil mathématique de modélisation développé au début des années soixante par le mathématicien allemand Carl Adam Petri. Les réseaux de Petri décrivent des relations existantes entre des conditions et des événements et ils modélisent le comportement des systèmes dynamiques à événements discrets. Les réseaux de Petri présentent des caractéristiques intéressantes à savoir le parallélisme, la synchronisation, le partage des ressources,. . .

Réseaux de Petri simples : On appelle Réseau de Petri simple places-transitions le quadruplet $Q = \langle P; T; I; O \rangle$ où :

— P est un ensemble fini non vide de places ;

— T est un ensemble fini non vide de transitions ;

— $P \cap T = \emptyset$;

— $I(T_i)$ représente l'ensemble de places d'entrée de la transition T_i (transitions-d'entrée) ;

— T_i représente l'ensemble de places qui sont en sortie de la transition T_i (sotie des places).

Réseaux de Petri marqués : On appelle Réseaux de Petri marqués, le couple $R = \langle Q; M_0 \rangle$, où M_0 est le marquage initial du réseau et tel que la fonction de marquage M est définie de P dans l'ensemble des entiers naturels.

Exemple 1.8 ***Modélisation d'un atelier de coupe de bois :***

Un atelier est constitué d'une machine de coupe et d'un stock. Quand une commande arrive et que la machine de coupe est disponible, la commande peut être traitée (Opération de découpe). Une fois le traitement terminé, la commande qui a été traitée est stockée.

Sinon, la commande doit attendre que la machine se libère avant de pouvoir être traitée.

1.7.2 La simulation

Il s'agit d'implémenter un modèle simplifié du système à l'aide d'un programme de simulation adéquat. Elle présente l'avantage de traduire de manière plus réaliste le comportement du système.

C'est une technique largement utilisée pour l'évaluation des performances des systèmes informatiques et réseaux de communications.

Les étapes de la simulation

Les étapes de la simulation :

1. **Formulation du modèle :** cette étape consiste à identifier et analyser le problème, en déterminant ses composantes, leurs relations et les frontières entre le système et son environnement.
2. **Elaboration du modèle :** cette étape consiste à extraire un modèle aussi fidèle que possible du système réel.
3. **Identification du modèle et collecte de données :** la collecte de données est indispensable pour l'estimation des paramètres du modèle. Ceci requiert une connaissance des méthodes statistiques et des tests d'hypothèses.
4. **Validation du modèle :** cette étape consiste à évaluer les performances du modèle puis les comparer à celles du système réel.
5. **Exécution de la simulation :** pour permettre à l'épreuve le modèle. Le concepteur doit effectuer plusieurs exécutions et recueillir les résultats.
6. **Analyse et interprétation des résultats :** une fois les résultats obtenus, le concepteur passe à l'analyse et l'interprétation de ces résultats pour donner des recommandations et des propositions.

7. Conclusion : cette dernière étape consiste à évaluer les perspectives d'exploitation du modèle pour d'autres préoccupations

Les techniques de simulation

On distingue trois techniques de simulation :

Simulation à événements discrets : Cette approche représente typiquement la simulation d'un modèle de files d'attente qui est dirigé par une séquence de nombres aléatoires (ou semi-aléatoires) dont la distribution est spécifiée par l'utilisateur. Ces séquences aléatoires sont utilisées pour obtenir les temps de réponse, les probabilités de branchement (routage,...). Ce type de technique a été largement utilisé pour l'évaluation de performances des ordinateurs. La simulation à événements discrets représentant une expérimentation statistique, les résultats fournis nécessitent une interprétation statistique effectuée avec soin.

Simulation conduite par trace : Cette technique doit être utilisée avec une grande précaution. Elle est très appropriée quand on veut faire des comparaisons entre deux alternatives dans le même environnement. Si le système existe et s'il peut être observé, les entrées du modèle à simuler peuvent provenir d'échantillons obtenus directement du système par des techniques de mesure au lieu d'employer des valeurs aléatoires. Cela peut donner un peu plus de certitude quant aux résultats obtenus.

Simulation continue Si le modèle du système contient en grande partie des variables continues qui évoluent dans le temps, ce modèle conduit à un système d'équations différentielles. Fréquemment, il se trouve que le système d'équations ne peut pas être résolu analytiquement. On utilise alors la simulation dite continue, où des combinaisons de méthodes numériques essaient d'aboutir à la solution du modèle.

Domaines d'application de la simulation

Les domaines d'application de la simulation sont nombreux et variés, nous trouvons, entre autres :

— Concevoir et analyser des systèmes industriels ;

— Evaluer du hardware et du logiciel pour un système d'exploitation d'ordinateur ;

— Déterminer des politiques d'ordonnancement pour un système de production ;

— Concevoir des systèmes de communications et leurs protocoles ;

— Evaluer des politiques de gestion pour les organisations de service ;

— Analyser des systèmes financiers ou économiques.

Avantage de la simulation

1. Observations des états du système.
2. Études des points de fonctionnement d'un système.
3. Études de systèmes à échelles de temps variable.
4. Études de l'impact de variables sur les performances du système.

Inconvénients de simulation

1. La conception de modèle peut nécessiter des compétences spéciales.
2. Une autre forme d'analyse plus proche de la réalité est peut être nécessaire.
3. Résultats difficillement interprétables.
4. Résultats pas forcément généralisables.

1.8 Quelques exemples d'application

1.8.1 Un modèle d'ordinateur avec mémoire virtuelle (MV)

Considérons un modèle d'ordinateur à MV fonctionnant en multiprogrammation. Le système est constitué d'une UC, d'un disque de pagination (DiskP) et d'un disque fichier (DiskF). L'objectif de l'utilisation de la MV est d'augmenter le taux d'utilisation de la UC. Le principe du fonctionnement des ordinateurs à MV est de mettre le plus grand nombre possible de programmes simultanément dans l'ordinateur. L'idée consiste à rendre la mémoire centrale (MC)virtuelle. Il s'agit de donner à chaque programme qui se présente une plase en MC et de mettre tout ce qui ne entrer dans cette place sur un disque ou mémoiresecondaire appelé *disque de pagination*. La MC est découpée en plusieurs pages. Lorsque l'information nécessaire pour l'exécution d'un programme n'est pas située sur l'une de pages en MC, il faut charger la page contenant la bonne information en MC depuis le disque de pagination. Le cas échéant, il faut décharger une page de la MC pour la mettre sur diskP afin de laisser la place à une autre page.L'utilisation de cette technique permet le stockage d'un nombre quelconque de programmes. Cependant, charger un maximum de programmes en MC n'augmentera pas toujours le taux d'utilisation de l'UC. En effet, un disque de pagination va demander plusieurs dizaines de milli-secondes pour un chargement et pendant ce temps là, L'UC ne fonctionne pas. Il est bien claire qu'il existe un dégré optimal de multiprogrammation qui n'est pas infini. En effet, si le nombre de programmes en MC est trop grand, chaque tâche ne va posséder qu'une fraction infini de mémoire. À chaque instruction, il va falloir effectuer un remplacement de page.

Problèmatique : On souhate connaître le dégre de mutiprogrammation optimal du système. On pourrait s'intéresser à l'évaluation du taux d'utilisation du disque de pagination en fonction du degré de multiprogrammation, du taux d'utilisation du disque de pagination. Ce même problème peut se poser pour tous les système fonctionnant en mutiprogrammation, tels qu'un ***serveur messagerie***.

1.8.2 Système de traitement de base de données

On considère un modèle d'ordinateur gérant une base de données assez volumineuse. Dans ce genre d'applications, il est nécessaire de faire régulièrement des sauvetages d'informations. Le modèle va donc traiter les requêtes d'accés à la base de données et, de temps en temps, va s'interrompre pour effectuer une opération de sauvetage qui paralyse l'utilisation normale de la machine.

Lorsque la machine décide de lancer une sauvegarde, elle se place en mode d'alerte. Elle bloque alors l'accès à la base en refusant toute nouvelle requête se présentant et avertit les utilisateurs en cours qu'ils doivent se déconnecter rapidement. Au bout d'un temps donné, elle suppose que les utilisateurs ont eu le temps de se déconnecter, coupe sans préavis toutes les connexions et commence son processus de sauvetage. A l'issue de ce temps, la machine est prête à traiter de nouveaux accés que ceux-ci se présenteront.

Problèmatique :les questions que l'on doit se poser sont :

- Quelle est la proportion de temps pendant laquelle la base de données est en mode normal de fonctionnement (c-à-d ni en alerte, ni en sauvegarde) ?
- Quel est le nombre moyen d'utilisateurs connectés lorsque la base donnée est en mode normal de fonctionnement ?
- Quel est le nombre moyen d'utilisateurs connectés à la base de données ?

La résolution de ce problème peut se faire à la base d'un ***modèle Markovien*** sous certaines hypothèses.

1.8.3 Serveur Web

On considère un serveur Web équipé d'une mémoire cache de capacité α Mo. Le fonctionnement d'un tel serveur est le suivant : un document demandé et se trouvant dans le cache est aussitôt transmis au demandeur. Si le document demandé n'est pas dans le cache, il doit être recherché dans la mémoire centrale du serveur, ramené dans le cache et transmis au demandeur. Lorsque le cache est plein, la politique de gestion du cache

doit décider quel(s) document(s) ôter pour faire de la place à un nouveau document. La politique de gestion de cache la plus répandue sur le Web consiste à ôter du cache, lorqu'il est plein, les documents les moins récemment demandés. Cette politique est nommée LRU, pour "Least Recently Used". Mais il y a d'autres politiques (LFU pour "Least Frequently Used" où la page la moins fréquemment référencée est ôtée,etc.). La recherche de politiques plus efficaces que LRU est d'ailleurs un domaine de recherche très actif en ce moment. Pour une politique de gestion de cache donnée on peut, par exemple, essayer d'évaluer le taux de "hits", c'est à dire le pourcentage de documents qui sont dans le cache lorsqu'on les demande.

Chapitre 2

Modèles d'attente markoviens

La théorie de files d'attente a de nombreuses applications, en particulier dans les réseaux de communication et les réseaux informatique. Nous insisterons dans ce chapitre, sur les modèles markoviens, en supossant acquises les notions de base sur les chaines de Markov et les processus markoviens.

2.1 Processus de naissance et de mort

Les processus en question permettent de façon générale d'écrire l'évolution temporelle de la taille d'une population d'un type donné. Il s'agit des processus stochastiques à temps continu et à espace d'états discret ($S = \{0, 1, 2, \ldots\}$). Ils sont caractérisés par deux conditions importantes :

— sans mémoire ;

— à partir d'un état donné n, des transitions ne sont possibles que vers l'un ou l'autre des états voisins $n+1$ et $n-1$ $(n \geq 0)$.

2.1.1 Processus de naissance

Soit $\{N(t), t \geq 0\}$, où $N(t)$ est le nombre d'individus dans la population à la date t, et $S = \{0, 1, 2, \ldots\}$ est l'espace d'états. Un processus de naissance est caractérisé par l'ap-

parition d'un individu, au sein d'une population, selon une certaine loi. Il est **homogène**, si la probabilité d'apparition

1. d'un individu pendant l'intervalle Δt, sachant qu'il existe déjà n individus au sein de la population est $\lambda_n \Delta t + 0(\Delta t)$ (indépendante de la position de sur l'axe des temps),
2. d'aucun individu est $1 - \lambda_n \Delta t + 0(\Delta t)$;
3. de deux ou plus est $0(\Delta t)$.

Ici, λ_n représente le taux d'apparition ou le taux de croissance.

Soient $p_n(t) = \Pr(N(t) = n)$ les probabilités d'ètats, ($N(t)$ durant $[0,t]$) ; et p_{ij} la probabilité que à l'instant t l'effectif est j sachant qu'il y avait déjà i individus dans la population (les probabilités de transitions).

On a que $p_{ij} = 0$ si $j < i$. Encore, $p_n(t) = p_{0,n}(t) = p_{1,n+1}(t) = \ldots$ (le processus est homogène).

La matrice des transitions se présente de la manière suivante :

$$M = \begin{pmatrix} 1-\lambda_0\Delta t & \lambda_0\Delta t & 0 & \ldots \\ 0 & 1-\lambda_1\Delta t & \lambda_1\Delta t & 0 \\ 0 & 0 & 1-\lambda_2\Delta t & \lambda_2\Delta t \\ \ldots & \ldots & \ldots & \ldots \end{pmatrix}$$

Le graphe de transitions est (2.1)

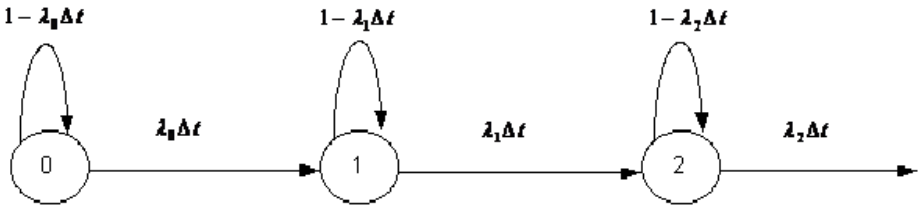

FIGURE 2.1 *Graphe de transitions de processus de naissance*

Le régime transitoire est décrit par $P(t + \Delta t) = P(t) \times M$:

$$[p_0(t+\Delta t), p_1(t+\Delta t), \ldots, p_n(t+\Delta t), \ldots] = [p_0(t), p_1(t), \ldots, p_n(t), \ldots] \times M,$$

ou bien

$$\left\{\begin{array}{l} p_0(t+\Delta t) = p_0(t)(1-\lambda_0\Delta t) \\ p_1(t+\Delta t) = p_0(t)\lambda_0\Delta t + p_1(t)(1-\lambda_1\Delta t) \\ ---------- \\ p_n(t+\Delta t) = p_{n-1}(t)\lambda_{n-1}\Delta t + p_n(t)(1-\lambda_n\Delta t) \\ ---------- \end{array}\right.$$

A présent, on trouve

$$\left\{\begin{array}{l} \dfrac{p_0(t+\Delta t)-p_0(t)}{\Delta t} = -\lambda_0 p_0(t); \\ ---------- \\ \dfrac{p_n(t+\Delta t)-p_n(t)}{\Delta t} = \lambda_{n-1}p_{n-1}(t) - \lambda_n p_n(t), n \geq 1 \end{array}\right.$$

Faisons $\Delta t \to 0$. Alors, on obtient le systèmes d'équations différentielles (équations du futur) suivant :

$$\left\{\begin{array}{l} p_0^{'}(t) = -\lambda_0 p_0(t); \\ ---------- \\ p_n^{'}(t) = \lambda_{n-1}p_{n-1}(t) - \lambda_n p_n(t), n \geq 1 \end{array}\right.$$

On a que $p_0(0) = 1$ (population est vide) et $p_n(0) = 0 \ \forall n \geq 1$.

Ici, le processus est en régime transitoire.

2.1.2 Processus de mort

Soit $\{N(t), t \geq 0\}$. L'espace d'états est $S = \{0, 1, 2, \ldots\}$. Ce processus est caractérisé par la disparition d'un individu au sein d'une population :

$$\left\{\begin{array}{l} \Pr(N(t+\Delta t) - N(t) = -1/N(t) = k) = \mu_k\Delta t + 0(\Delta t); \\ \Pr(N(t+\Delta t) - N(t) = 0/N(t) = k) = 1 - \mu_k\Delta t + 0(\Delta t); \\ \Pr(N(t+\Delta t) - N(t) = 2/N(t) = k) = 0(\Delta t); \end{array}\right.$$

μ_k est le taux de décroissance.($\mu_k > 0, k > 0; \mu_0 = 0$).

Le graphe de transitions est figure(2.2)

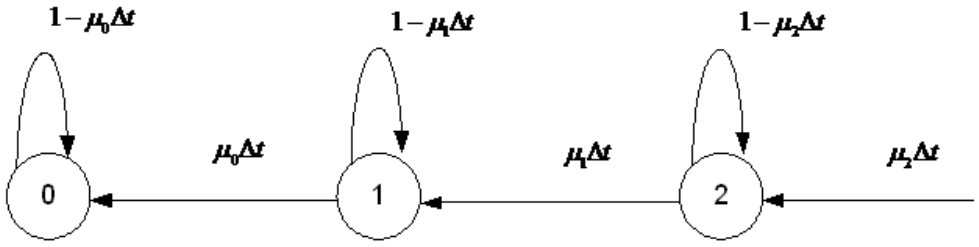

FIGURE 2.2 *Graphe de transitions de processus de mort*

2.1.3 Processus de naissance et de mort

Soit $\{N(t), t \geq 0\}$ avec $S = \{0,1,2,\ldots\}$. Ce processus est homogène dans le temps : les probabilités de transitions $p_{ij} = \Pr(N(t+\Delta t) = j / N(\Delta t) = i)$ ne dépend pas de Δt.

Il est de naissance et de mort si

$$\left\{ \begin{array}{l} p_{i,i+1}(\Delta t) = \lambda_i \Delta t + 0(\Delta t), i \geq 0; \\ p_{i,i-1}(\Delta t) = \mu_i \Delta t + 0(\Delta t), i \geq 1; \\ p_{i,i}(\Delta t) = 1 - (\lambda_i + \mu_i)\Delta t + 0(\Delta t); i \geq 0 \\ p_{i,j}(\Delta t) = 0(\Delta t); |i-j| \geq 2 \end{array} \right|$$

On a également que

$$\lambda_i > 0, \mu_i > 0, \mu_0 = 0, p_{ij}(0) = \delta_{ij} = \left\{ \begin{array}{l} 1; i = j \\ 0; i \neq j \end{array} \right|$$

Régime transitoire

Soient $p_n(t) = \Pr(N(t) = n), n \geq 0$, les probabilités d'état.

Le graphe de transitions est donnée dans la figure (2.3)

La matrice des transitions correspondante est

$$M = \begin{pmatrix} 1-\lambda_0\Delta t & \lambda_0\Delta t & 0 & \ldots \\ \mu_1\Delta t & 1-(\lambda_1+\mu_1)\Delta t & \lambda_1\Delta t & 0 \\ 0 & \mu_2\Delta t & 1-(\lambda_2+\mu_2)\Delta t & \lambda_2\Delta t \\ \ldots & \ldots & \ldots & \ldots \end{pmatrix}$$

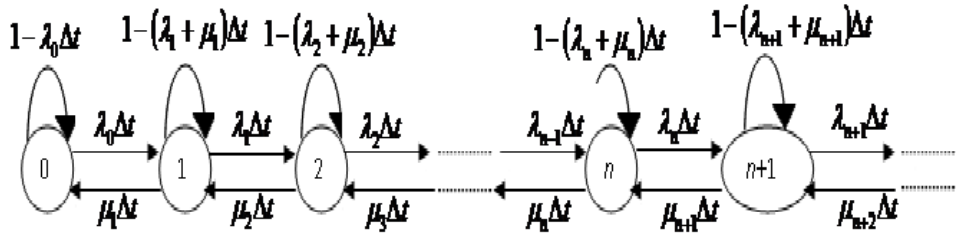

FIGURE 2.3 *Graphe de transitions de processus de naissance et de mort*

En appliquant $P(t+\Delta t) = P(t) \times M$, on trouve

$$\left\{ \begin{array}{l} p_0(t+\Delta t) = p_0(t)(1-\lambda_0\Delta t) + p_1(t)\mu_1\Delta t \\ p_n(t+\Delta t) = p_{n-1}(t)\lambda_{n-1}\Delta t + p_n(t)[1-(\lambda_n+\mu_n)\Delta t] + p_{n+1}(t)\mu_{n+1}\Delta t, n \geq 1 \end{array} \right|$$

On déduit les équations de Kolmogorov

$$\left\{ \begin{array}{l} p_0^{'}(t) = -\lambda_0 p_0(t) + \mu_1 p_1(t); \\ p_n^{'}(t) = \lambda_{n-1} p_{n-1}(t) - (\lambda_n+\mu_n) p_n(t) + \mu_{n+1} p_{n+1}(t), n \geq 1 \end{array} \right| \tag{2.1}$$

Remarque 2.1

1. *Si* $S = \{0,1,\ldots,K\}$, *alors* $\lambda_K = 0$. *D'où,*

   $p_K^{'}(t) = \lambda_{K-1} p_{K-1}(t) - \mu_K p_K(t)$.

2. *Les équations de Kolmogorov, complétées par des conditions initiales, gouvernent le régime transitoire du processus* $\{N(t), t \geq 0\}$.

Régime stationnaire

Soit $p_n = \lim_{t\to\infty} p_n(t)$, qui est la distribution stationnaire du processus étudié. Ces probabilités satisfont le système d'équations de balance suivant (obtenu à partir de (2.1) en prenant $\lim_{t\to\infty}$) :

$$\lambda_0 p_0 = \mu_1 p_1; \tag{2.2}$$

$$p_n(\lambda_n+\mu_n) = \mu_{n+1} p_{n+1} + \lambda_{n-1} p_{n-1}, n \geq 1 \tag{2.3}$$

avec l'équation de normalisation

$$\sum_{n=0}^{\infty} p_n = 1$$

De l'équation (2.2), on obtient

$$p_1 = \frac{\lambda_0}{\mu 1} p_0;$$

Pour $n = 1$:(l'équation (2.3))

$p_1(\lambda_1 + \mu_1) = \mu_2 p_2 + \lambda_0 p_0$; $p_2 = \dfrac{\lambda_0 \lambda_1}{\mu_1 \mu_2} p_0$

$p_n = \dfrac{\lambda_0 \lambda_1 \dots \lambda_{n-1}}{\mu_1 \mu_2 \dots \mu_n} p_0$

Pour déduire p_0, on utilise l'équation de normalisation. On obtient le résultat suivant

$$p_0 = \left[1 + \frac{\lambda_0}{\mu 1} + \frac{\lambda_0 \lambda_1}{\mu_1 \mu_2} + \dots + \frac{\lambda_0 \lambda_1 \dots \lambda_{n-1}}{\mu_1 \mu_2 \dots \mu_n} + \dots \right]^{-1}$$

Par conséquent, pour qu'une distribution stationnaire existe, il faut donc que la somme [] converge. Ceci a toujours lieu si l'espace d'états du processus à l'étude est fini.

Lorsque la somme en question n'est pas convergente, $p_n = 0, \forall n \geq 0$.

2.2 La théorie de files d'attente

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

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

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

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

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

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

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

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

Définition 2.1 *Une file d'attente est composé d'un certain nombre de places d'attente, d'un ou plusieurs serveurs et de clients qui arrivent, attendent, se font servir selon des règles de priorité, puis quittent le système .*

2.2.1 Origine de la théorie des files d'attente

— La théorie des files d'attente fut développée pour fournir des modèles permettant de prévoir le comportement de systèmes répondant é des demandes aléatoires.

— Les premiers problèmes étudiés en utilisant la théorie des files d'attennte et sur la congestion du trafic téléphonique en 1909, grâce à l'article du mathématicien danois A.K. Erlang "The theory of probabilities and telephone conversations".

Les premiers résultats sont variés : Erlang observe le caractère poissonnien des arrivées des appels à un central téléphonique, et le caractère exponentiel des durées des appels ; il réussit à calculer de manière relativement simple la probabilité d'avoir un appel rejeté. La notion d'équilibre stationnaire d'un système d'attente est introduite.

— A partir des années 30, les travaux de plusieurs mathématiciens tels que Molina, Fry, Pollaczek aux Etats-Unis, Kolmogorov et Khintchine en Russie, Palm en Suède, ou Crommelin en France permettent à la théorie des files d'attente de se développer lentement. Ce sont ensuite les années 50 qui verront l'essor important de la théorie.

— Actuellement ce sont les applications dans le domaine de l'analyse de performance des réseaux (téléphone mobile, Internet, multimédia, ...) qui suscitent le plus de travaux.

2.2.2 Application des files d'attente

La théorie des files d'attente a de nombreuses applications dans :

— La gestion de trafic (réseaux de communication, compagnies aériennes, embouteillages , . . .).

— La planification (opérations sur des machines de production, programmes sur des ordinateurs, l'ordonnancement, par exemple les patients dans les hôpitaux, . . .).

— Le dimensionnement d'infrastructures (usines, . . .).

2.2.3 Caractéristiques des files d'attente simples

On peut caractériser les files d'attente par plusieurs critères :

— Le processus d'arrivée des clients.

— Le temps de service.

— La discipline de service.

— La capacité du système.

Notation de Kendall

La notation suivante est standard dans la théorie des files d'attente $A/B/c/m$ avec

— A La distribution de probabilité du temps inter-arrivées.

— B La distribution de probabilité du temps de service.

— c Le nombre de serveurs.

— m La capacité de la système.

Processus d'arrivée des clients et temps de service

A et B pourra prendre les valeurs suivantes :

— M : markovien (i.e. exponentiel).

— G : loi générale.

— D : loi déterministe.

— E_k : loi de Erlang de de.

— H_k ; loi hyperexponentielle.

Si il y a des arrivées groupées on pourra utiliser la notation $A^{[X]}$ où X est la variable aléatoire mesurant le nombre de clients à chaque arrivée

$$\Pr(X = k) = \Pr(k \text{ clients arrivent en même temps}).$$

Discipline de la file

Il y a plusieurs discipline de service :

— FIFO : First in First out.

— LIFO : Last in First out.

— RANDOM : aléatoire.

— HL (Hold On Line) : si un client important arrive, il prend la première place dans la file d'attente.

— PR (Preemption) : si un client important arrive, il est servi directement et le client moins important (en train d'être servi) est remis dans la file.

— ...

Notation abrégée

On utilisera la notation abrégée $A/B/c$ quand on considère une file où :

— La capacité du système est infinie.

— La discipline est FIFO.

— Il ya c serveurs dans le système.

Exemple 2.1 *La notation $M/D/1/4$ définit un système d'attente comprenant une station service et pour lequel la longueur maximale de la file d'attente vaut 3. Le processus d'arrivée est un processus de Poisson et la durée du service est constante.*

Exercice 2.1 *les notations $M/M/2/5$, $M/G/5/\infty$ et $M/E_1/\infty/\infty$.*

2.2.4 Analyse mathématique et mesures de performance

L'étude mathématique d'un système de files d'attente se fait par l'introduction d'un processus stochastique défini de façon appropriée $\{N(t), t \geq 0\}$. En général, on s'intéresse au nombre $N(t)$ de clients se trouvant dans le système à l'instant $t \geq 0$. En fonction des quantités qui déterminent la structure du système.

On cherche à calculer :

1. **Les probabilités d'état** $p_n(t) = \Pr(N(t) = n)$ qui définissent le régime transitoire du processus $N(t), t \geq 0$, les probabilités p_n doivent évidemment dépendre de l'état initial ou la distribution initiale du processus,

2. **Les probabilités d'état** qui définissent le régime stationnare du processus en question, est défini par

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

La distribution stationnaire du processus stochastique introduit permet d'obtenir les caractéristiques d'exploitation du système, telles que :

— Le temps d'attente d'un client,

— Le temps de séjour d'un client dans le système,

— Le taux d'occupation des dispositifs de service,

— La durée de la période d'activité ;

et également les mesures de performance :

— Le nombre moyen de clients dans le système $\overline{n}$;

— Le nombre moyen de clients dans la file d'attente $\overline{n}_f$;

— Le temps moyen d'attente d'un client $\overline{W}$;

— Le temps moyen de séjour d'un client dans le système $\overline{W}_f$.

On encore des relations (formules de Little) :

$$\overline{n} = \lambda \overline{W}_s ; \overline{n}_f = \lambda \overline{W} ; \overline{W}_s = \overline{W} + \frac{1}{\mu} ; \overline{W} = \frac{\overline{n}_f}{\lambda} ; \overline{n} = \overline{n}_f + \frac{\lambda}{\mu}.$$

où λ est le taux d'entrée des clients dans le système, $\frac{1}{\mu}$ est la durée moyenne de service ($\mu > 0$).

Une autre mesure importante d'un système de files d'attente, celle qui mesure le degré de saturation du système, est **l'intensité du trafic** ρ. Elle est définie par

$$\rho = \frac{\text{temps moyen de service}}{\text{temps moyen entre deux arrivées successives}}$$

2.3 Système de files d'attente markoviens

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

$$\{N(t), t \geq 0\} \tag{2.4}$$

Les modèles Markoviens ont en commun que le processus (2.4) constitue une chaîne de Markov à temps continu homogène, ce qui signifie que le comportement futur ne

dépend que de l'état présent et non pas de l'évolution dans le passé :

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

où $t_1 < t_2 < \cdots < t_n < t$. Supposons que la discipline d'attente est $FIFO$, alors (2.4) est un processus de naissance et de mort, qui est une chaîne de Markov à temps continu et à états discrets.

2.3.1 Système de files d'attente $M/M/1$

Description du modèle :

Les clients arrivent vers le système selon un processus de Poisson de taux λ (nombre moyen de clients arrivant pendant une unité de temps) ; c'est-à-dire l'intervalle de temps entre deux arrivée successifs suit une loi exponentielle de paramètre $\lambda > 0$. Le service est assuré par un seul serveur. A l'arrivée d'un client, si le serveur est libre, il est immédiatement pris en charge. Dans le cas contraire, le client en question est placé en attente. La capacité d'attente est illimitée (le nombre de positions est infini et aucune autre restriction n'est imposée). La discipline d'attente est FIFO. Les durées de service suivent une loi exponentielle de paramètre $\mu > 0$. Par conséquent, le taux de service est μ (nombre moyen de clients servis pendant une unité de temps), et le temps moyen de service d'un client est $\frac{1}{\mu}$. Les variables aléatoires représentant les durées entre deux arrivées consécutives et les durées de service sont mutuellement indépendantes.

Analyse du modèle :

L'état du système à la date t peut être décrit par le processus stochastique (2.4). Grâce aux propriété fondamentales du processus de Poisson et de la loi exponentielle, on a pour

un petit intervalle du temps Δt les probabilités suivantes :

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

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

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

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

Régime transitoire

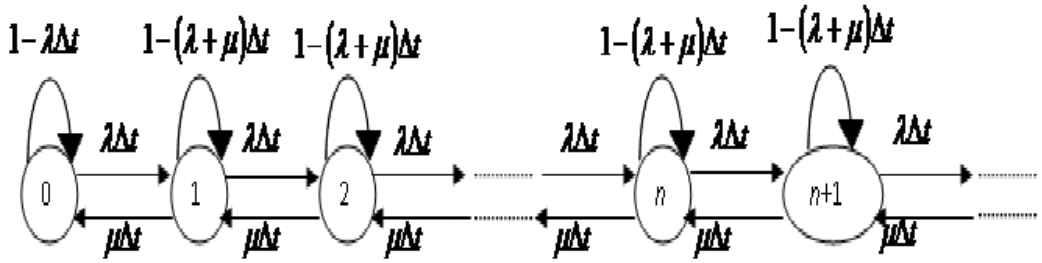

FIGURE 2.4 *Graphe de transitions de la file $M/M/1$*

Soit $p_n(t) = \Pr(N(t) = n), n = 0,1,2,\ldots$. Le graphe des transitions est la figure (2.4)

A partir du graphe des transitions, on obtient

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

Puis, les équations de Kolmogorov :

$$\begin{cases} p_0'(t) = -\lambda p_0(t) + \mu p_1(t) \\ p_n'(t) = -(\lambda+\mu)p_n(t) + \lambda p_{n-1}(t) + \mu p_{n+1}(t), n \geq 1 \end{cases} \tag{2.5}$$

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

Régime stationnaire

Il est démontré que $\lim_{t\to\infty} p_n(t) = p_n, n \geq 0$, existent et sont indépendantes de l'état initial du processus (2.4) ; et $\lim_{t\to\infty} p_n'(t) = p_n, n \geq 0$.

De (2.5), on obtient le système d'équations de balance suivant :

$$\begin{cases} \mu p_1 = \lambda p_0 \\ \lambda p_{n-1} + \mu p_{n+1} = (\lambda+\mu)p_n, n \geq 1 \end{cases} \tag{2.6}$$

La résolution du système (2.6)(la résolution du modèle) s'effectue de la manière suivante :

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

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

$$p_0 + \frac{\lambda}{\mu} p_0 + (\frac{\lambda}{\mu})^2 p_0 + \cdots = 1$$

$$p_0 = \frac{1}{1 + \dfrac{\lambda}{\mu} + (\dfrac{\lambda}{\mu})^2 + \ldots}$$

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

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

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

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

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

Remarque 2.2

1. *Si $\lambda \geq \mu$, on a $\lim_{t\to\infty} p_n(t) = 0, n \geq 0$. Ceci signifie que la longueur de la file d'attente dépasse toute limite.*
2. *En ce qui concerne le processus de sortie du système $M/M/1$, il est (en régime stationnaire) évident que le taux de sortie est égal au taux d'arrivée λ. Il est démontré que le processus de sortie d'un système $M/M/1$ est à nouveau de type poissonien.*

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

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

Le nombre moyen de clients dans le système

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

L'expression entre les crochets est une dérivée de $B = \rho + \rho^2 + \rho^3 + \ldots$. En effet, $B = \rho[1+\rho^1+\rho^2+\ldots] = \frac{\rho}{1-\rho}$ et $B' = \frac{1}{(1-\rho)^2}$.

Le nombre moyen de clients dans la file d'attente

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

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

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

Le temps moyen d'attente d'un client, Le temps moyen de séjour d'un client dans le système

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

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

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

2.3.2 Système de files d'attente $M/M/c$

Description du modèle

Les clients arrivent vers le système selon un processus de Poisson de taux $\lambda > 0$.

Le service est assuré par $c \geq 1$ serveurs montés en parallèle.

A l'arrivée d'un client, si l'un des serveurs est libre, le client commence immédiatement son service. Dans le cas contraire (tous les serveurs sont occupés par le service), le client prend place dans la file d'attente, commune pour tous les serveurs.

La capacité d'attente est illimitée (le nombre de positions d'attente est infini).

Lorsqu'un serveur se libère, le client en tête de la file d'attente occupe le serveur libéré. Par conséquent, la discipline d'attente est FIFO.

Les temps de service sont exponentiellement distribués de moyenne finie $\frac{1}{\mu}$. Les durées entre deux arrivées consécutives et les durées de service sont mutuellement indépendantes.

Analyse du modèle

L'état du système à la date t peut être décrit à l'aide du processus (2.4), dont l'espace des états est $S = \{0, 1, 2, \ldots\}$. Ce dernier est un processus de naissance et de mort dont les taux de transition sont :

$$\lambda_n = \lambda, n \geq 0$$

$$\mu_n = \min\{n, c\}\mu, n \geq 1$$

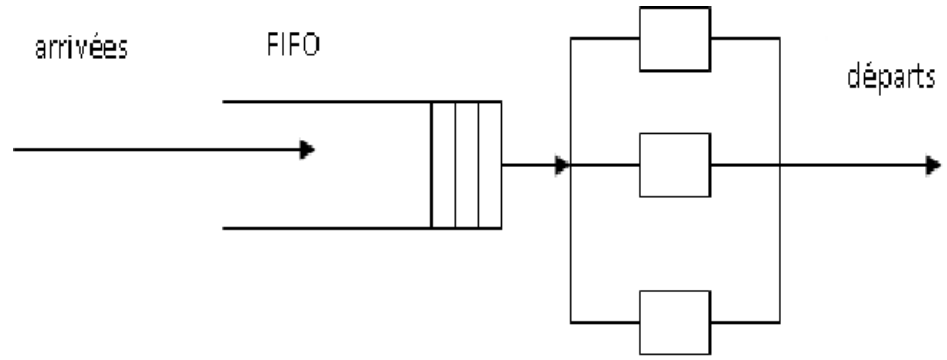

FIGURE 2.5 *Modèle d'attente $M/M/c$*

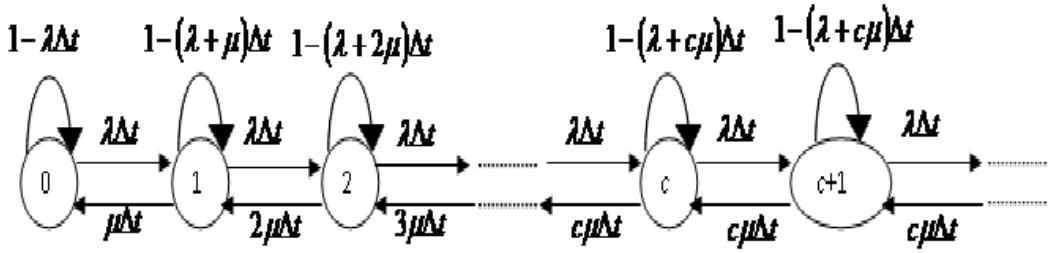

FIGURE 2.6 *Graphe de transitions de la file $M/M/c$*

Régime transitoire

Le graphe des transitions est donné dans la figure (2.6).

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

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

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

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

Régime stationnaire

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

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

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

$$p_n = \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n p_0, 1 \leq n \leq c$$

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

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

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

La deuxième somme peut être réécrite de la manière suivante

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

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

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

Encore,

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

et

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

Mesures de performance

Le nombre moyen de clients dans le système

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

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

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

Le nombre moyen de clients dans la file d'attente

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

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

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

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

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

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

Le temps moyen d'attente d'un client

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

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

Chapitre 3

Systèmes particuliers de files d'attente (Modèles Markoviens)

Il existe de nombreux modèles d'attente de type markovien pour lesquels le processus stochastique (2.4) possède une structure complexe : les modèles où les arrivées (ou le service) se font par groupe, les modèles avec classes des clients,.... Le processus (2.4), associé à ces modèles n'est plus de naissance et de mort parce que les transitions à partir d'un état donné peuvent se produire à n'importe quel autre état.

3.1 Système de files d'attente $M^X/M/1$

3.1.1 Description du modèle

Les clients arrivent dans le système par groupe selon un processus de Poisson ($\lambda > 0$).

Le nombre de clients par groupe est une variable aléatoire X strictement positive : $\Pr(X(t) = x) = c_x(t)$.

Le service est assuré par un seul serveur. Les durées de service suivent une loi commune, exponentielle en occurrence de moyenne finie $\frac{1}{\mu}$.

A l'arrivée d'un groupe, si le serveur est libre, le premier client du groupe en question

est immédiatement pris en charge et les autres clients sont placés en attente (la capacité d'attente est illimitée). Dans le cas contraire, tous les clients du groupe rejoignent la file d'attente.

La discipline est FIFO.

Nous supposons que toutes les variables introduites (la durée entre deux arrivées consécutives, la durée de service, la longueur du groupe) sont mutuellement indépendantes.

3.1.2 Analyse du modèle :

L'état du système à la date t peut être décrit par le processus stochastique $\{N(t), t \geq 0\}$, où $N(t)$ est le nombre de clients dans le système à l'instant t. L'espace des états est $S = \{0, 1, 2, \ldots\}$.

Régime transitoire

Soient $p_n(t) = \Pr(N(t) = n), n \geq 0$. Le graphe des transitions est donné par la figure (3.1)

Système d'équations des probabilités d'état s'obtient à partir du graphe des transitions. En effet,

$$
\begin{aligned}
&p_0(t+\Delta t) = p_0(t)(1-\lambda\Delta t) + \mu\Delta t\, p_1(t);\\
&p_1(t+\Delta t) = \lambda\Delta t c_1(t) p_0(t) + [1-(\lambda+\mu)\Delta t]\, p_1(t) + \mu\Delta t\, p_2(t);\\
&p_2(t+\Delta t) = \lambda\Delta t c_1(t) p_1(t) + \lambda\Delta t c_2(t) p_0(t) + [1-(\lambda+\mu)\Delta t]\, p_2(t) + \mu\Delta t\, p_3(t);\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&p_n(t+\Delta t) = \lambda\Delta t \sum_{k=1}^{n} c_k(t) p_{n-k}(t) + [1-(\lambda+\mu)\Delta t]\, p_n(t) + \mu\Delta t\, p_{n+1}(t); n > 2.
\end{aligned}
$$

Alors le système d'équations de Kolmogorov est :

$$
\begin{aligned}
&p_0'(t) = -\lambda p_0(t) + \mu p_1(t);\\
&p_n'(t) = \lambda \sum_{k=1}^{n} c_k(t) p_{n-k}(t) - (\lambda+\mu) p_n(t) + \mu p_{n+1}(t); n \geq 1.
\end{aligned}
$$

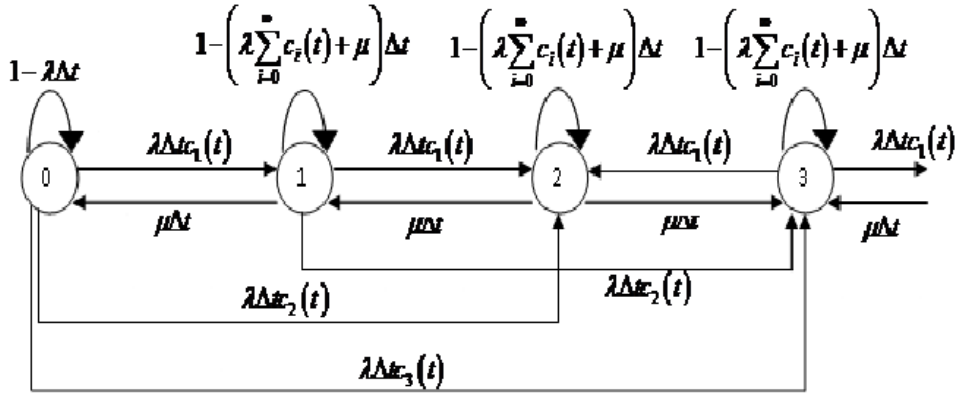

FIGURE 3.1 *Graphe de transition de file $M^X/M/1$*

Régime stationnaire

Soient $p_n = \lim_{t\to\infty} p_n(t), n \geq 0$, et $c_n = \lim_{t\to\infty} c_n(t), n \geq 0$, $c_0 = 0$.

Les équations de balance se présentent de la manière suivante

$$\lambda p_0 = \mu p_1;$$

$$0 = \lambda \sum_{k=1}^{n} c_k p_{n-k} - (\lambda+\mu)p_n + \mu p_{n+1}; n \geq 1. \tag{3.1}$$

Pour résoudre le système ci-dessus, nous utilisons les fonctions génératrices suivantes (nous allons utiliser une méthode qui porte le nom *la méthode des fonctions génératrices*) :

$$P(z) = \sum_{n=0}^{\infty} p_n z^n \text{ et } C(z) = \sum_{n=0}^{\infty} c_n z^n.$$

En multipliant les équations (3.1) par z^n et en sommant l'ensemble, on obtient

$$\lambda \sum_{n=0}^{\infty} p_n z^n + \mu \sum_{n=1}^{\infty} p_n z^n = \frac{\mu}{z} \sum_{n=1}^{\infty} p_n z^n + \lambda \sum_{n=1}^{\infty} \sum_{k=1}^{\infty} p_{n-k} c_k z^n. \tag{3.2}$$

A l'aide des propriétés des fonctions génératrices, l'équation (3.2) devient

$$\lambda P(z) + \mu[P(z) - p_0] = \frac{\mu}{z}[P(z) - p_0] + \lambda C(z)P(z).$$

D'où

$$P(z) = \frac{\mu p_0 (1-z)}{\mu(1-z) - \lambda z[1-C(z)]}. \text{ si } |z| < 1 \tag{3.3}$$

Pour obtenir p_0, il faut utiliser l'équation de normalisation $P(1) = 1$.

La relation (3.3) pour $z \longrightarrow 1$ (en appliquant la règle de l'Hôpital $\frac{0}{0}$) devient

$$1 = \frac{-\mu p_0}{-\mu + \lambda E[X]}$$

$$p_0 = 1 - \frac{\lambda E[X]}{\mu}$$

L'intensité du trafic $\rho = \dfrac{\lambda E[X]}{\mu}$.

La condition pour que le régime stationnaire existe est $\rho < 1$.

Remarque 3.1 *Convolution de deux distributions est définie par*

$$p_n \otimes q_n \Longrightarrow \sum_{n=0}^{\infty} (p_n \otimes q_n) z^k = \sum_{n=0}^{\infty} \sum_{k=0}^{n} p_{n-k} q_k z^{n-k} z^k = \sum_{k=0}^{\infty} \sum_{n=k}^{\infty} p_{n-k} q_k z^{n-k} z^k = Q(z)P(z).$$

Mesures de performance

Le nombre moyen de clients dans le système

$$\overline{n} = P^{'}(1);$$

$$\overline{n} = \frac{\rho + \dfrac{\lambda E\left[X^2\right]}{\mu}}{2(1-\rho)}$$

Exemples d'application

Dans les réseaux locaux (LAN), l'un des protocoles de communication les plus utilisés est CSMA (Carrier-Sence Multiple Acces). Supposons qu'un réseau local est composé de n stations connectées par un seul bus. La communication entre les stations est réalisée au moyen de ce bus. Les messages de longueur variables arrivent aux stations du monde extérieur. En recevant le message, la station le découpe en un nombre fini de paquets de longueur fixe et consulte le bus pour voir s'il est occupé. Si le bus est libre,

l'un des paquets est transmis via ce bus à la station de destination, et les autres paquets sont stockés dans les tampons pour transmission ultérieure. Autrement, tous les paquets sont stockés dans le tampon et la station peut consulter le bus après une certaine durée aléatoire.

Les questions concernant ce problème sont : Quel est le temps moyen d'attente d'un paquet ? Quel est le nombre moyen de messages (paquets) dans le tampon d'une station ?

3.2 Système de files d'attente $M/M/1$ avec différentes classes de clients et priorité absolue

En général, on admet que la population des clients est homogène, c'est-à-dire que les durées de service des clients sont identiquement distribuées selon une loi de probabilité commune.

Dans les cas plus complexes où les clients sont divisés en classes, chaque classe peut être identifiée par sa propre distribution du temps de service (population hétérogène). De plus, il y a des systèmes où certains clients jouissent d'une priorité de service.

La priorité peut être absolue ou relative. Par **priorité absolue**, on entend qu'un client moins prioritaire est remis en tête de file d'attente lorsqu'un client plus prioritaire se présente devant la file d'attente. Ce dernier venu commence son service immédiatement.

Si la **priorité est relative**, un nouveau client plus prioritaire attend la fin du service avant de pouvoir commencer le sien.

Dans le cas de priorité absolue, deux nouvelles possibilités se présentent : soit le client suspendu reprend son service là où il a été interrompu, soit il le reprend depuis le début.

3.2.1 Description du modèle

Considérons un système de files d'attente de type $M/M/1$.

Cependant supposons qu'il y a deux classes de clients qui arrivent toujours selon un processus de Poisson $(\lambda > 0)$:

la proportion des clients de la première classe (clients plus prioritaires) est α, la proportion des clients de la deuxième classe (clients moins prioritaires) est alors $(1-\alpha)$,(on a $\lambda = \lambda\alpha + \lambda(1-\alpha) = \lambda_1 + \lambda_2$).

Les durées de service des clients de la première classe suivent une loi exponentielle de paramètre μ_1, tandis que celles des clients de la deuxième classe sont réparties selon une loi exponentielle de paramètre μ_2.

La priorité est absolue. Par conséquent, les clients de la première classe ne sont pas perturbés par les clients de la deuxième classe. Encore, pour les clients de la deuxième classe, vu que la distribution du temps de service est exponentielle (sans mémoire), reprendre le service là où il a été interrompu est équivaut à recommencer depuis le début.

3.2.2 Analyse du modèle

Le processus stochastique associé est $\{N_1(t), N_2(t), t \geq 0\}$, où $N_1(t)$ est le nombre de clients de la première classe et $N_2(t)$ est le nombre de clients de la deuxième classe dans le système à la date t. Son espace des états est
$S = \{(n_1, n_2), n_1 \in \mathbb{N}, n_2 \in \mathbb{N}\}$.

Le graphe des transitions est donné par la figure (3.2)

En régime stationnaire, les équations de balance sont :

$$(\lambda_1 + \lambda_2) p_{0,0} = \mu_1 p_{1,0} + \mu_2 p_{0,1}, \text{ si } n_1 = 0 \text{ et } n_2 = 0$$
$$(\lambda_1 + \lambda_2 + \mu_1) p_{n_1,0} = \lambda_1 p_{n_1-1,0} + \mu_1 p_{n_1+1,0}, \text{ si } n_1 > 0 \text{ et } n_2 = 0$$
$$(\lambda_1 + \lambda_2 + \mu_2) p_{0,n_2} = \mu_1 p_{1,n_2} + \lambda_2 p_{0,n_2-1} + \mu_2 p_{0,n_2+1}, \text{ si } n_1 = 0 \text{ et } n_2 > 0$$
$$(\lambda_1 + \lambda_2 + \mu_1) p_{n_1,n_2} = \mu_1 p_{n_1+1,n_2} + \lambda_2 p_{n_1,n_2-1} + \lambda_1 p_{n_1-1,n_2}, \text{ si } n_1 > 0 \text{ et } n_2 > 0$$

où $p_{n_1,n_2} = \lim_{t\to\infty} \Pr(N_1(t) = n_1, N_2(t) = n_2), n_1, n_2 \geq 0$.

Nous allons décrire les calculs au lieu de les effectuer. En effet, ils sont longs et

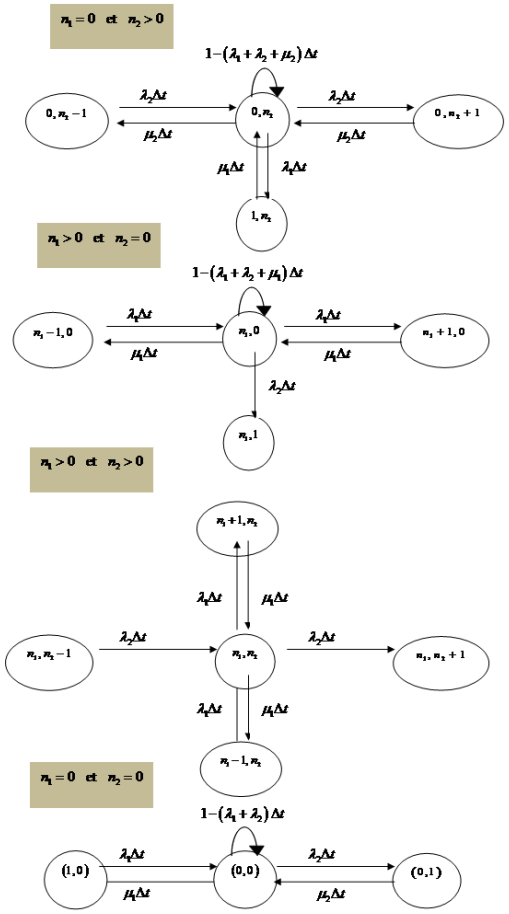

FIGURE 3.2 *Graphe de transition de file d'attente avec priorité*

fastidieux. Il faut calculer la fonction génératrice

$$F(x,y) = \sum_{n_1=0}^{\infty}\sum_{n_2=0}^{\infty} p_{n_1,n_2} x^{n_1} y^{n_2}.$$

A partir du système d'équations de balance, on trouve

$$F(x,y) = \left(\frac{1-\rho_1-\rho_2}{1-\eta-y\rho_2}\right)\left(\frac{1-\eta}{1-\eta x}\right)$$

où

$$\rho_1 = \frac{\lambda_1}{\mu_1}, \rho_1 = \frac{\lambda_2}{\mu_2} \text{ et}$$

$$\eta = \frac{1}{2\mu_1}\left[\mu_1+\lambda_1+\lambda_2(1-y) - \sqrt{[\mu_1+\lambda_1+\lambda_2(1-y)]^2 - 4\lambda_1\mu_1}\right]$$

De la fonction génératrice, on déduit

$$\overline{n}_1 = \frac{\partial F(x,y)}{\partial x}|_{x=y=1} = \frac{\rho_1}{1-\rho_1}$$
$$\overline{n}_2 = \frac{\partial F(x,y)}{\partial y}|_{x=y=1} = \frac{\rho_2 + E[n_1]\rho_2}{1-\rho_1-\rho_2}$$

La distribution stationnaire s'obtient également à partir de la fonction génératrice

$$p_{n_1,n_2} = \frac{1}{n_1!n_2!}[\frac{\partial^{n_1+n_2}F(x,y)}{(\partial x)^{n_1}(\partial y)^{n_2}}]_{x=y=0}$$

en particulier, $p_{0,0} = 1-\rho_1-\rho_2$. Pour que la distribution stationnaire existe, $p_{0,0} > 0$: $\rho_1 + \rho_2 < 1$.

3.2.3 Exemple d'application

Systèmes téléphoniques modernes : Dans les échanges téléphoniques modernes, les lignes des abonnés sont connectées aux modules qui traitent les appels arrivants et ceux sortants. Dans le cas du blocage (les canaux sont occupés), les appels sortants sont placés en attente dans un tampon (de capacité infinie) tandis que les appels arrivants sont refusés et plus tard réinitialisés pour établir la connexion. Lorsque l'un des canaux se libère, un appel sortant (s'il se trouve dans le tampon) l'occupe immédiatement. De ce fait, les appels arrivants ne réussissent pas à établir la connexion aussi longtemps qu'il y a des appels dans le tampon. Ce comportement implique que les appels sortants possèdent une priorité relative sur les appels arrivants.

Réseaux LAN avec le protocole CSMA non persistant et CSMA persistant : Supposons qu'un LAN possède des utilisateurs non persistants et des utilisateurs persistants connectés par un bus. Les utilisateurs persistants sont contrôlés par l'unité centrale de façon que chaque fois lorsque le canal est libre, un utilisateur persistant l'occupe pour envoyer ces paquets. Les utilisateurs non persistants essayent la retransmission indépendamment après une durée de temps. Ce système se présente comme un système avec rappels et priorité, ou les clients type 1 sont ceux persistants, les clients type 2 sont ceux non persistants et le serveur est le bus.

Réseau cellulaire mobile : On considère un réseau cellulaire mobile, où chaque cellule est servie par une station de base différente. Une cellule particulière peut s'occuper de plusieurs communications actives simultanément. La station de base dans une cellule manipule deux types d'appels. Le premier est un appel d'origine (initié dans la cellule en question). Habituellement, un abonné avec un appel d'origine bloqué refait sa tentative après une durée de temps. Un autre type d'appels, "handoff", apparait lorsqu'un abonné détenteur de la ligne de communication entre en cellule en question des cellules adjacentes. Si la station de base ne parvient pas à attribuer un canal libre jusqu'à ce que l'abonne sort de la région de recouvrement des cellules, il souffre de la panne durant la conversation. La dégradation de la qualité de service téléphonique dans ce cas est plus sérieuse que dans le cas d'un appel d'origine. Par conséquent, la station de base peut donner la priorité aux appels "handoff" en attribuant une file. Dans un réseau cellulaire mobile, la perte d'un appel "handoff" est un facteur important pour la qualité de service. Ce système peut être modélisé comme un système avec rappels, deux types de clients et perte géométrique.

Téléphone dans une banque : Prenant l'exemple d'un banquier qui doit s'occuper d'une ligne d'attente et d'un téléphone. Un tel employé prête attention au téléphone seulement quand il n'y a aucun client dans la banque. Un client arrivant et trouvant ce banquier inoccupé sera servi immédiatement ; autrement, il attend avec une probabilité q dans une file d'attente ou avec la probabilité complémentaire p=1-q décide de téléphoner plus tard jusqu'à ce qu'il obtienne son service. Ce système peut être modelé comme un système avec rappels et priorité relative où les clients attendant dans la banque sont considérés comme des clients prioritaires et les clients faisant des appels téléphoniques comme des clients non prioritaires.

3.3 Modèle $M/M/1$ avec rappels

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

Pour identifier un système de files d'attente avec rappels, on a besoin des spécifications suivantes :

— La nature stochastique du processus des arrivées,

— La distribution du temps de service,

— Le nombre de serveurs qui composent l'espace de service,

— La capacité du système

— La discipline de service,

— La spécification concernant le processus de répétition d'appels.

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

A l'arrivée d'un client primaire, s'il y a un ou plusieurs serveurs libres, le client sera immédiatement pris en charge. Sinon, s'il y a une position d'attente libre, le client rejoint la file d'attente.

Dans le cas contraire, il quitte l'espace de service temporairement avec une probabilité H_0 pour tenter sa chance après une durée de temps aléatoire, ou il quitte le système définitivement avec une probabilité $(1-H_0)$.

Entre les tentatives, le client est en **orbite** et devient une source de clients secondaires ou de clients répétés.

La capacité de l'orbite O peut être finie ou infinie. Dans le cas où O est finie et si l'orbite est pleine, le client quitte le système pour toujours.

Lorsqu'un client (secondaire) est rappelé de l'orbite, il est traité de la même manière

qu'un client primaire avec une probabilité H_k (s'il s'agit de la k-ème tentative échouée).

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

— A et B décrivent respectivement la distribution du temps entre deux arrivées consécutives et la distribution du temps de service,

— c est le nombre de serveurs identiques et indépendants,

— $(m-c)$ est la capacité d'attente,

— O est la capacité de l'orbite,

— H est la fonction de persistance :$H = \{H_k, k \geq 0\}$.

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

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

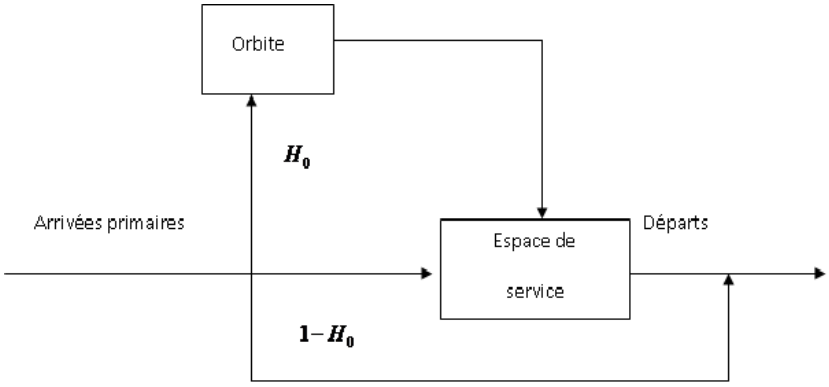

FIGURE 3.3 *Modèles avec rappels simples*

3.3.1 Description du modèle

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

Le service des clients est assuré par un seul serveur.

A l'arrivée d'un client primaire, si le serveur est libre, il est immédiatement pris en

charge. Dans le cas contraire, le client en question entre en orbite et devient une source de tentatives répétés (devient source de clients secondaires).

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

La durée de temps entre deux tentatives consécutives (rappels) d'un même client secondaire est distribuée selon une loi de probabilité de fonction de répartition $T(t) = 1 - \exp\{-\theta t\}, t \geq 0$, de moyenne finie $\frac{1}{\theta}$.

Les trois variables aléatoires introduites sont supposées mutuellement indépendantes.

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

$$\{C(t), N_0(t), t \geq 0\}; \tag{3.4}$$

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

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

Théorème 3.1 *Pour un système de files d'attente $M/M/1$ avec rappels, la distribution stationnaire conjointe de l'état du serveur et du nombre de clients en orbite $p_{in} = \lim_{t\to\infty} \Pr\{C(t) = i, N_0(t) = n\}, i = 0,1$ et $n \geq 0$, est donnée par*

$$p_{0n} = \frac{\rho^n}{n!\theta^n} \prod_{k=0}^{n-1} (\lambda + k\theta) (1-\rho)^{1+\frac{\lambda}{\theta}}; \tag{3.5}$$

$$p_{1n} = \frac{\rho^{n+1}}{n!\theta^n} \prod_{k=1}^{n} (\lambda + k\theta) (1-\rho)^{1+\frac{\lambda}{\theta}}. \tag{3.6}$$

Les fonctions génératrices partielles correspondantes sont données par

$$P_0(z) = \sum_{n=0}^{\infty} z^n p_{0n} = (1-\rho) \left(\frac{1-\rho}{1-\rho z}\right)^{\frac{\lambda}{\theta}}; \tag{3.7}$$

$$P_1(z) = \sum_{n=0}^{\infty} z^n p_{1n} = \rho \left(\frac{1-\rho}{1-\rho z}\right)^{\frac{\lambda}{\theta}+1}. \tag{3.8}$$

Preuve 1 *Le processus* (3.4) *a pour espace d'états* $S = \{0,1\} \times \mathbb{N}$. *Les transitions possibles sont :*

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

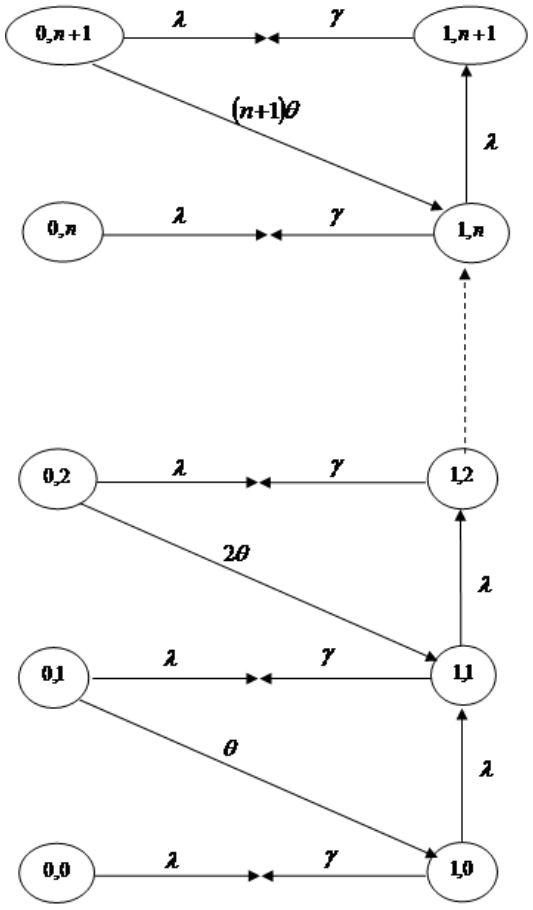

FIGURE 3.4 *Graphe de transitions de la file* $M/M/1$ *avec rappels*

Les équations d'équilibre statistique (de balences) sont

$$(\lambda + n\theta)\, p_{0n} = \gamma p_{1n}; \tag{3.9}$$

$$(\lambda+\gamma)\,p_{1n} = \lambda\,p_{0n} + (n+1)\,\theta\,p_{0n+1} + \lambda\,p_{1n-1}. \tag{3.10}$$

A l'aide de fonctions génératrices, telles que

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

les équations (3.9)-(3.10) *deviennent*

$$\lambda P_0(z) + \theta z P_0^{'}(z) = \gamma P_1(z); \tag{3.11}$$

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

D'où

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

La solution de cette dernière équation est

$$P_0(z) = const\,(1-\rho z)^{-\frac{\lambda}{\theta}}. \tag{3.12}$$

Des équations (3.11)*, on a*

$$P_1(z) = \rho P_0(z) + \frac{\theta z}{\gamma} P_0^{'}(z) = P_0(z)\,\frac{\rho}{1-\rho z} = \frac{\rho\,const}{(1-\rho z)\,\frac{\lambda}{\theta}+1}. \tag{3.13}$$

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

$$const = (1-\rho)\,\frac{\lambda}{\theta}+1. \tag{3.14}$$

A partir des équations (3.12)-(3.14)*, on déduit les équations* (3.7) *et* (3.8)*.*

A présent, à l'aide de l'équation (3.9)*, on élimine* p_{1n} *de l'équation* (3.10)*. De cette manière, on trouve*

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

Ceci implique que

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

D'où

$$p_{0n} = \frac{\lambda\left(\lambda + (n-1)\,\theta\right)}{n\theta\gamma} p_{0n-1} = \frac{\rho^n}{n!\theta^n} \prod_{k=0}^{n-1} (\lambda + k\theta)\, p_{00}.$$

De l'équation (3.9)*, on a*

$$p_{1n} = \frac{\rho^{n+1}}{n!\theta^n} \prod_{k=1}^{n} (\lambda + k\theta)\, p_{00}.$$

La probabilité p_{00} sera trouvée à l'aide de l'équation de normalisation

$\sum_{n=0}^{\infty} p_{0n} + \sum_{n=0}^{\infty} p_{1n} = 1$

$$p_{00} = \left[\sum_{n=0}^{\infty} \frac{\rho^n}{n!\theta^n} \prod_{k=0}^{n-1} (\lambda + k\theta) + \sum_{n=0}^{\infty} \frac{\rho^{n+1}}{n!\theta^n} \prod_{k=1}^{n} (\lambda + k\theta)\right]^{-1}.$$

A l'aide de la formule binomiale

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

on obtient

$$p_{00} = (1-\rho)\frac{\lambda}{\theta} + \rho\,(1-\rho)\frac{\lambda}{\theta}^{+1} = (1-\rho)\frac{\lambda}{\theta}^{+1}.$$

En fin, on peut former les équations (3.5) *et* (3.6)*.*

Fin de preuve

3.3.2 Conséquences

1. La distribution stationnaire de processus (3.4) existe si $\rho = \dfrac{\lambda}{\gamma} < 1$.
2. La fonction génératrice de la distribution stationnaire marginale du nombre de clients en orbite $N_0 = \lim_{t\to\infty} N_0(t)$ est définie par

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

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

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

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

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

3.3.3 Mesures de performance

Nombre moyen de clients dans le système

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

Nombre moyen de clients en orbite

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

Temps moyen d'attente d'un client

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

Nombre moyen de rappels par client

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

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

Chapitre 4

Modèle semi-Markoviens

Dans les chapitres précédents, nous avons examiné les système de files d'attente Markoviens, dont les arrivées forment un processus de Poisson, et la durée de services suit une loi exponentielle.

dans nombreux cas, les clients arrivants dans le système peut être réaliser par un processus de Poisson, mais le temps de service d'un client est distribué selon unr loi qui n'est plus supposée exponentielle..

Dans ce chapitre, nous étudierons un système de file d'attente avec un seul serveur et des temps de service indépendants identiquement distribués de loi générale, et le processus d'arrivée des clients dans la file est toujours supposé poissonien. Ce système est connu sous le nom de file $M/G/1$.

4.1 Système de files d'attente $M/G/1$

4.1.1 Description du modèle

Les clients arrivent dans le système selon un processus de Poisson de taux $\lambda > 0$. De ce fait, le temps entre deux arrivées successives suit une loi exponentielle de moyenne $\frac{1}{\lambda}$. Le service est assuré par un seul serveur. A l'arrivée d'un client, si le serveur est libre, le client sera pris en charge immédiatement. Dans le cas contraire, il rejoint la

file d'attente (de capacité illimitée et discipline FIFO) les durées de service *Se* sont des variables aléatoires indépendantes et identiquement distribuées de loi générale dont la fonction de répartition $B(x)$ et la transformée de Laplace-Stieltjes $\widetilde{B}(s)$. Soient $E[Se] = \frac{1}{\mu}$ et $E\left[Se^2\right] =$.

4.1.2 Analyse du modèle

Pour décrire l'état d'un système de type $M/G/1$ à la date t, il faut connaître non seulement le nombre de clients qui se trouvent dans le système à la date t, mais également le temps de service déjà écoulé $R(t)$ du client qui est en train d'être servi. On peut alors montrer que le processus bidimensionnel $\{N(t), R(t), t \geq 0\}$ est à nouveau du type markovien. Cependant, le calcul de son régime transitoire ferait intervenir des équations aux dérivées partielles.

Par conséquent, on choisit une autre méthode qui ramène l'étude du processus non markovien $\{N(t), t \geq 0\}$ à celle d'une chaîne de Markov à temps discret associée au processus considéré dont elle permet de calculer le régime stationnaire.

Soit le processus $\{N(t), t \geq 0\}$ qui n'est pas un processus de Markov. Pour le rendre markovien, on utiliserons la méthode de **la chaîne de Markov induite**.

A cet effet, on considère $N(t)$ aux instants $\xi_1, \xi_2, \ldots, \xi_n, \ldots$ où les clients terminent leur service et quittent le système. On définit ainsi un processus stochastique à temps discret

$$\{N_n = N(\xi_n), n \geq 1\}. \tag{4.1}$$

(Nombre de clients restants dans le système immédiatememt après le départ du $n^{'}ime$ client à l'instant ξ_n)

Pour vérifier que cette suite de variables aléatoires est une chaîne de Markov à temps discret, on considère le nombre A_n de clients qui entrent dans le système pendant que le *nème* client est servi.

Les variables A_n sont indépendantes entre elles, leur distribution commune est

$$\Pr(A_n = k) = a_k = \int_0^\infty \exp(-\lambda t) \frac{(\lambda t)^k}{k!} dB(t), \text{ où } a_k > 0 \text{ et } k > 0.$$

Alors

$$N_{n+1} = \begin{cases} N_n - 1 + A_{n+1} \text{ si } N_n \geq 1 \\ A_{n+1} \text{ si } N_n = 0 \end{cases} ; n \geq 1$$

L'équation fondamentale de la chaîne vaut donc

$$N_{n+1} = N_n - \delta_n + A_{n+1}. \tag{4.2}$$

où

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

N_{n+1} ne dépend que de N_n et de A_{n+1} et non pas des valeurs prises par $N_{n-1}, N_{n-2}, \ldots$. La suite $\{N_n, n \geq 1\}$ est une chaîne de Markov induite du processus $\{N(t), t \geq 0\}$. Ses probabilités de transition $p_{ij} = \Pr(N_{n+1} = j / N_n = i)$ se calculent par

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

Le graphe est :

La matrice des transitions est

$$M = \begin{pmatrix} p_{00} & p_{01} & p_{02} & p_{03} & -- & -- \\ p_{10} & p_{11} & p_{12} & p_{13} & -- & -- \\ p_{20} & p_{21} & p_{22} & p_{23} & -- & -- \\ p_{30} & p_{31} & p_{32} & p_{33} & -- & -- \\ -- & -- & -- & -- & -- & -- \end{pmatrix} = \begin{pmatrix} a_0 & a_1 & a_2 & a_3 & -- & -- \\ a_0 & a_1 & a_2 & a_3 & -- & -- \\ 0 & a_0 & a_1 & a_2 & -- & -- \\ 0 & 0 & a_0 & a_1 & -- & -- \\ -- & -- & -- & -- & -- & -- \end{pmatrix}.$$

Vu qu'on peut passer de chaque état vers n'importe quel autre état, il s'agit d'une chaîne de Markov irréductible. De plus, la matrice n'est pas décomposable (est apériodique). La chaîne est donc ergodique.

La distribution stationnaire de la chaîne existe si $\rho = \frac{\lambda}{\mu} < 1$.

Pour les variables aléatoires A_n, nous disposons de quelques résultats importants :

$$E[A_n] = \lambda E[Se] = \frac{\lambda}{\mu} = \rho$$

La fonction génératrice

$$A(z) = \sum_{k=0}^{\infty} a_k z^k = \sum_{k=0}^{\infty} z^k \int_0^{\infty} \exp(-\lambda t) \frac{(\lambda t)^k}{k!} dB(t) = \int_0^{\infty} \exp(-\lambda t) \left(\sum_{k=0}^{\infty} \frac{(\lambda t z)^k}{k!} \right) dB(t)$$

$$A(z) = \int_0^{\infty} \exp(-\lambda t) \exp(-\lambda t z) \, dB(t) = \int_0^{\infty} \exp(-(\lambda - \lambda z) t) \, dB(t)$$

Soit $\widetilde{B}(s) = \int_0^{\infty} \exp(-st) \, dB(t)$. Alors, $A(z) = \widetilde{B}(\lambda - \lambda z)$. Encore, la série $A(z)$ converge pour $|z| \leq 1$:

1. Si $|z| < 1 : \ 0 < a_k < 1, \forall k,$ on a $|a_k z^k| < |z^k|$;
2. Si $|z| < 1 : \ A(1) = 1$

Théorème 4.1 *Soit $\rho < 1$. La distribution stationnaire de la chaîne de Markov induite $\{N_n = N(\xi_n, n \geq 1)\}$ possède la fonction génératrice suivante*

$$\Pi(z) = \frac{(1-\rho) A(z) (z-1)}{z - A(z)} = \frac{(1-\rho) \widetilde{B}(\lambda - \lambda z) (z-1)}{z - \widetilde{B}(\lambda - \lambda z)} \tag{4.3}$$

Soient les probabilités suivantes :

$$p_j = \lim_{t \to \infty} \Pr(N(t) = j), j \geq 0$$
$$\pi_j = \lim_{n \to \infty} \Pr(N(\xi_n) = j), j \geq 0$$
$$r_j = \lim_{n \to \infty} \Pr(N(\varsigma_n) = j), j \geq 0.$$

où ς_n est l'instant d'arrivée du nme client. Comme le processus des arrivées est celui de Poisson de paramètre λ et le nombre de clients dans le système $N(t)$ est discontinu avec un changement de taille ∓ 1, alors $p_j = r_j = \pi_j$.

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

$$Q(z) = \sum_{j=0}^{\infty} p_j z^j = \Pi(z)$$

Preuve 2 *Supposons que* $\rho < 1$. *Le système se trouve dans un régime stationnaire.*

Soit $(\pi_0, \pi_1, \pi_2, \dots)$ *la distribution stationnaire de la chaîne de Markov induite* ($\pi_j = \lim_{n\to\infty} \Pr(N(\xi_n) = j)$).

Par conséquent, $(\pi_0, \pi_1, \pi_2, \dots) = (\pi_0, \pi_1, \pi_2, \dots) M$, *ou*

$$\pi_j = \sum_{i=0}^{\infty} \pi_i p_{ij}, j \geq 0.$$

$$\pi_j = a_j \pi_0 + \sum_{i=1}^{j+1} a_{j-i+1} \pi_i$$
$$\pi_j = a_j \pi_0 + \sum_{i=0}^{j+1} a_{j-i+1} \pi_i - a_{j+1} \pi_0, j \geq 0$$

A présent, on applique la méthode des fonctions génératrices. En effet,

$$\sum_{j=0}^{\infty} \pi_j z^j = \pi_0 \sum_{j=0}^{\infty} a_j z^j + \frac{1}{z} \sum_{j=0}^{\infty} c_{j+1} z^{j+1} - \frac{\pi_0}{z} \sum_{j=0}^{\infty} a_{j+1} z^{j+1} \text{ où } c_{j+1} = \sum_{i=0}^{j+1} a_{j-i+1} \pi_i.$$

On introduit les fonctions génératrices suivantes :

$$\Pi(z) = \sum_{i=0}^{\infty} \pi_i z^i, A(z) = \sum_{i=0}^{\infty} a_i z^i, C(z) = \sum_{j=0}^{\infty} c_j z^j = \Pi(z) A(z).$$

Finalement, on obtient

$$\Pi(z) = \pi_0 A(z) + \frac{1}{z} [C(z) - c_0] - \frac{\pi_0}{z} [A(z) - a_0].$$

ou bien

$$\Pi(z) = \frac{\pi_0 A(z)(z-1)}{z - A(z)}, \text{ pour } |z| < 1 \text{ et } |z| \neq 0$$

On a que $\Pi(1) = 1$. *Cependant,* $\Pi(1) = \lim_{z\to 1} \Pi(z) = \frac{0}{0}$. *En appliquant la règle de l'Hôpital, on obtient* $\frac{\pi_0}{1 - A'(1)} = 1$. *Alors* $\pi_0 = 1 - A'(1) = 1 - \lambda E[Se] = 1 - \rho$.

Le résultat final est ***la première équation de Pollaczek-Khintchine*** *pour le nombre de clients dans le système :*

$$\Pi(z) = \frac{(1-\rho) A(z)(z-1)}{z - A(z)} = \frac{(1-\rho) \widetilde{B}(\lambda - \lambda z)(z-1)}{z - \widetilde{B}(\lambda - \lambda z)}$$

4.1.3 Mesures de performance

Formule de Pollaczek-Khintchine pour le nombre moyen de clients dans le système :

$$\overline{n} = \lim_{t\to\infty} E[N_n] = E[N] = \rho + \frac{\rho^2 + \lambda^2 Var[Se]}{2(1-\rho)}. \tag{4.4}$$

Temps moyen de séjour d'un client dans le système

$$\overline{W_s}(s) = \frac{s(1-\rho)}{s-\lambda+\lambda\widetilde{B}(s)}\widetilde{B}(s). \tag{4.5}$$

Temps moyen d'attente d'un client

$$\overline{W}(s) = \frac{s(1-\rho)}{s-\lambda+\lambda\widetilde{B}(s)}. \tag{4.6}$$

Période d'activité

Soit U la durée de la période d'activité du système $M/G/1$ (l'intervalle de temps pendant lequel le dispositif de service est continuellement occupée). Admettons que pendant une longue durée t, le système d'attente passe par n cycles d'exploitation complets dont chacun est composé d'une période d'activité U et d'une période d'inactivité V. Pour les grandes valeurs de t ($t \longrightarrow \infty$), on a $t \approx n[E[U]+E[V]]$.

D'autre part, la probabilité que le système soit vide est

$$\pi_0 = p_0 = \frac{E[V]}{E[U]+E[V]}.$$

Mais $p_0 = 1-\rho$ et $E[V] = \frac{1}{\lambda}$. Il en résulte que $E[U] = \frac{1}{\mu-\lambda}$, si $\lambda < \mu$. Ce résultat est valable et pour le système de files d'attente $M/M/1$.

4.2 Cas particuliers du modèle $M/G/1$

4.2.1 Modèle $M/M/1$

Supposons que le temps de service est exponentiellement distribué avec une moyenne de $\frac{1}{\mu}$. La fonction de TLS (voir Annexe C) est

$$\widetilde{B}(s) = \frac{\lambda}{\lambda + s},$$

ainsi

$$\Pi(z) = \frac{(1-\rho)\,\widetilde{B}(\lambda - \lambda z)\,(z-1)}{z - \widetilde{B}(\lambda - \lambda z)}$$

4.2.2 Modèle $M/E_k/1$

Dans ce système, la durée de service suit une loi d'Erlang d'ordre k et de moyenne finie $\frac{1}{\mu}$. Les fonctions de densité de probabilités et de répartition sont données par :

$$\begin{cases} b(t) = \frac{k\mu(k\mu t)^{k-1}}{(k-1)!}\exp\{-k\mu t\} \\ B(t) = 1 - \exp\{-k\mu t\}\sum_{j=0}^{k-1}\frac{(k\mu t)^j}{j!} \end{cases} ,t \geq 0.$$

On démontre que $A(z) = \widetilde{B}(\lambda - \lambda z) = \left[1 + \frac{\rho(1-z)}{k}\right]^{-k}$.

Alors, l'équation (4.3) devient

$$\Pi(z) = \frac{(1-\rho)\,(z-1)}{z\left[1 + \frac{\rho(1-z)}{k}\right]^{k} - 1}.$$

4.2.3 Modèle $M/H_2/1$

La durée de service suit une loi hyperexponentielle d'ordre 2 dont la fonction de répartition est donnée par $B(t) = 1 - p_1\exp\{-\mu_1 t\} - p_2\exp\{-\mu_2 t\}$, où p_1, μ_1, p_2, μ_2 vérifient $p_1 + p_2 = 1$ et $\frac{1}{\mu_1} + \frac{1}{\mu_2} = \frac{1}{\mu}$ (ici, $\frac{1}{\mu}$ est la durée moyenne de service). Pour une telle distribution, on démontre que

$$A(z) = \widetilde{B}(\lambda - \lambda z) = \frac{p_1}{1 + \rho_1(1-z)} + \frac{p_2}{1 + \rho_2(1-z)}.$$

Par conséquent, l'équation (4.3) devient

$$\Pi(z) = \frac{(1-\rho)\left[1+(\rho_1+\rho_1-\rho)(1-z)\right]}{\rho_1\rho_2 z^2-(\rho_1+\rho_2+\rho_1\rho_2)z+1+\rho_1+\rho_2-\rho}.$$

où $\rho_i = \frac{\lambda}{\mu_i}, i = 1,2$ et $\rho = \frac{\lambda}{\mu}$.

4.3 Système de files d'attente $G/M/1$

Le système $G/M/1$ peut être considéré comme symétrique du système $M/G/1$, et traité de façon analogue. On étudie ce système aux instants où un client arrive. Le processus sous-jacent ainsi défini est alors une chaîne de Markov à temps discret. Cependant, contrairement au cas $M/G/1$, la distribution de la chaîne de Markov induite du système $G/M/1$ n'est pas identique à celle du processus $\{N(t), t \geq 0\}$.

Soit A la variable aléatoire modélisant les temps entre deux arrivées successives, $a(t)$ sa densité de probabilité. Le temps moyen entre deux arrivées successives est $E[A] = \int_0^\infty t a(t)\,dt$, le taux d'arrivée est $\lambda = \frac{1}{E[A]}$.

Les durées de service suivent une loi exponentielle de moyenne finie $\frac{1}{\mu}$. Sous la condition pour que le régime stationnaire existe $\rho = \frac{\lambda}{\mu} = \frac{1}{\mu E[A]} < 1$, on démontre que l'équation fonctionnelle (la transformée de Laplace de la densité de probabilités $a(t)$)$\alpha = \int_0^\infty \exp\{-\mu t(1-\alpha)\}\, a(t)\,dt$ possède une unique solution α comprise entre 0 et 1 (généralement, la résolution est possible à l'aide des méthodes numériques). Cette solution permet d'avoir la probabilité stationnaire qu'un client arrivant dans le système y trouve k clients

$$\pi_k = (1-\alpha)\,\alpha^k, k \geq 0$$

La probabilité en question permet, à son tour, de calculer les mesures de performance

$$\overline{n} = \frac{\rho}{1-\alpha}$$
$$\overline{n}_f = \frac{\rho\alpha}{1-\alpha}$$
$$\overline{W}_s = \frac{1}{\mu(1-\alpha)}$$
$$\overline{W} = \frac{\alpha}{\mu(1-\alpha)}$$

Chapitre 5

Les réseaux de files d'attente

La modélisation d'un système à l'aide d'un seul système de files d'attente n'offre qu'un champ d'applications restreint. Bien souvent, un client a besoin de recevoir plusieurs traitement consécutives et de différentes natures avant de quitter un système. ***Les réseaux de files d'attente*** permettent de modéliser de telles situations et correspondent à des systèmes composés de plusieurs systèmes de files d'attente reliés entre eux. Lorsqu'un client quitte un système de files d'attente, il peut se diriger vers une nouvelle ***station*** (un nouveau système de files d'attente) du réseau (qui peut, éventuellement, être celle qu'il vient de quitter) ou sortir définitivement du système. Ces décisions sont dictées par des ***règles de routage*** qui peuvent être déterministes ou stochastiques.

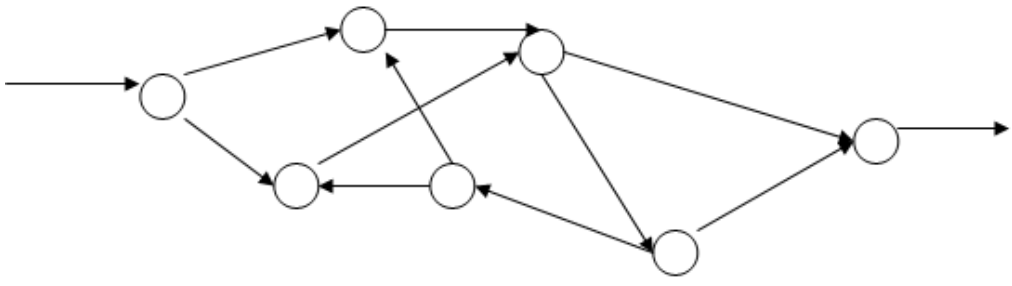

FIGURE 5.1 *Un réseau de files d'attente*

Un réseau de files d'attente est dit ***ouvert*** lorsque des arrivées de clients depuis l'extérieur du système sont possibles et lorsque les clients peuvent quitter le système. A priori,

les arrivées externes peuvent avoir lieu dans n'importe quelle station et il en est de même des départs, l'important étant que tout client entrant ou présent dans le système ait une possibilité de le quitter un jour. Lorsque aucun client ne peut ni entrer dans le système ni le quitter, le réseau est dit ***fermé***. Le nombre de clients dans un réseau fermé est donc constant au cours du temps. Finalement, il existe également des réseaux ***mixtes*** qui ne sont, cependant, qu'une juxtaposition de systèmes ouverts et fermés.

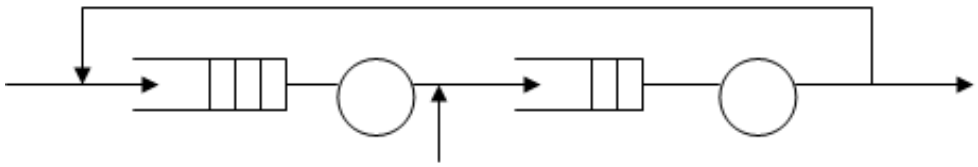

FIGURE 5.2 *Un réseau de files d'attente* ***ouvert***

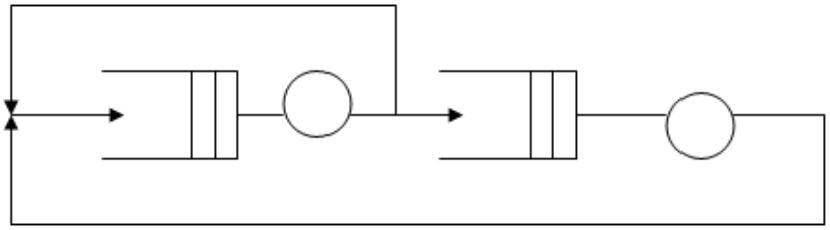

FIGURE 5.3 *Un réseau de files d'attente* ***fermé***

Un réseau de files d'attente traite parfois plusieurs types de clients répartis en autant de classes. Celles-ci sont caractérisées par leurs processus d'arrivée et de service dans les différentes stations, leurs règles de routage ou, encore, leurs priorités qui permettent de modéliser le comportement à adopter lorsque les clients de plusieurs classes cherchent à accéder à un même serveur. Si de tels réseaux offrent une très grande flexibilité, leur étude en est d'autant plus complexe. Nous considérons les réseaux de files d'attente à une seule classe de clients. Les réseaux de Jackson entrent dans cette catégorie et présentent la propriété de posséder des distributions stationnaires à ***forme produit*** égales au produit de

distributions marginales associées à chacune des stations formant le réseau. Les résultats de Jackson ont été généralisés à plusieurs reprises, en particulier les ***réseaux BCMP***, dus à Baskett, Chandy, Muntz et Palacios, présentent également une distribution stationnaire à forme produit.

5.1 Les réseaux de Jackson ouverts

Un réseau de Jackson ouvert est un réseau formé par l'interconnexion de n stations de type $-/M/c$ (que l'on supposera numérotées de 1 à n) où les clients arrivent selon des processus de Poisson et se déplacent en suivant de règles de routage markoviennes. Plus précisément, les clients arrivent de l'extérieur du système selon des processus de Poisson indépendants, le taux d'arrivée dans la station i étant constant et égal à γ_i, $1 \leq i \leq n$. Chaque station i du réseau est régie par une discipline FIFO, possède un nombre fini c_i de serveurs et fournit des traitements dont les durées sont indépendantes et identiquement distribuées selon une loi exponentielle de paramètre μ_i constant. Après avoir complété son service dans une station i, un client est envoyé à la station j avec la probabilité r_{ij} et quitte le système avec la probabilité $r_{i0} = 1 - \sum_{j=1}^{n} r_{ij}, i = 1,2,3,\ldots,n$.

Remarque 5.1 *La notation $-/M/c$ est utilisée car, si les clients arrivent dans le réseau selon des processus de Poisson, les flux de clients à l'intérieur du réseau ne sont, en général plus des processus de Poisson.*

L'état d'un réseau de Jackson à l'instant t est donné par le processus $\{N(t) = (N_1(t), N_2(t), \ldots, N_n(t))\}$, où $N_i(t)$ est le nombre de clients présents dans la station $i, 0 \leq i \leq n$. Vu les hypothèses faites sur le processus des arrivées et de service ainsi que sur les règles de routage, il n'est pas difficile de vérifier que le processus $\{N(t)\}$ est une chaîne de Markov à temps continu ayant pour espace d'états $S = \mathbb{Z}_+^n$.

Intuitivement, pour qu'un réseau de Jackson soit en régime stationnaire et possède une distribution stationnaire, il faut que, pour chacune des n stations du réseau, le taux

effectif d'arrivée des clients y soit inférieur au taux maximal de service. Soit λ_i le taux effectif d'arrivée dans la station i. Si le système est en régime stationnaire (stable), ce taux d'arrivée est égal au taux de sortie de la station et vérifier les équations de *conservation du flot*

$$\lambda_i = \gamma_i + \sum_{j=1}^{n} \lambda_j r_{ij}, i = 1,2,\ldots,n \tag{5.1}$$

Sous forme matricielle, ces équations d'équilibre s'écrivent

$$\Lambda = \Gamma + \Lambda R. \tag{5.2}$$

La matrice de routage R a tous ces éléments compris entre 0 et 1 (il s'agit des probabilités r_{ij}) et la somme des termes de chacune de ses lignes ne dépasse pas l'unité. De plus, le réseau étant ouvert, il existe au moins une ligne où cette somme est strictement inférieure à 1.

Les réseaux de Jackson doivent leur nom à J. Jackson qui fut le premier à obtenir la distribution stationnaire de ce type de système. Le résultat central de ces travaux fait l'objet du théorème qui suit.

Théorème 5.1 *Un réseau de Jackson ouvert, stable et formé de n stations possède une distribution stationnaire unique donnée par*

$$\begin{gathered} \Pi(N) = \prod_{j=1}^{n} \pi_i(N_i), \\ \forall N = (N_1, N_2, \ldots, N_n) \in \mathbb{Z}_+^n \end{gathered} \tag{5.3}$$

où $\pi_i(N_i)$ est la distribution stationnaire d'une station $M/M/c_i$= de taux d'arrivée λ_i et de taux de service μ_i.

Ainsi, en régime stationnaire, un réseau de Jackson se comporte comme autant de systèmes de files d'attente M/M/c isolés recevant leurs clients selon des processus Poisson indépendants les uns des autres. Il est important de souligner le « comme » dans la phrase précédente car les flux à l'intérieur du réseau ne sont en général ni indépendants

ni des processus de Poisson.

Une distribution de la forme (5.3) est dite à *forme produit* et les réseaux associés, des *réseaux à forme produit*.

5.2 Les réseaux de Jackson fermés

Les travaux initiaux de Jackson qui ne portaient que sur des réseaux ouverts ont été étendus et généralisés à plusieurs reprises, en particulier par Gordon et Newell qui ont montré que les distributions à forme produit s'appliquaient également aux réseaux fermés formés de stations $-/M/c$. Dans de tels réseaux, $\gamma_i = 0$ et r_{i0} quel que soit i et le nombre de clients présents dans le système est constant. Si ce nombre est égal à K, l'état $N(t)$ du système au temps t vérifie $K = \sum_{i=1}^{K} N_i(t)$ quel que soit $t \geq 0$. Le nombre d'états possibles du processus $\{N(t), t \geq 0\}$ est donc fini et égal au coefficient binomial $\begin{pmatrix} n+K-1 \\ K \end{pmatrix}$ décrivant le nombre de manières de répartir K clients dans n stations.

Le résultat montré par Gordon et Newell est : un réseau de Jackson fermé est non seulement toujours stable mais possède une distribution stationnaire à forme produit donnée par

$$\Pi(N) = \Pi(N_1, N_2, \ldots, N_n) = \frac{1}{G(K)} \prod_{i=1}^{n} F_i(N_i), \tag{5.4}$$

où $G(K)\,K$ est une *constante de normalisation* telle que la somme des probabilités sur tous les états du réseau soit égale à 1 :

$$\begin{aligned} G(K) &= \sum_{N \in S_K} \prod_{i=1}^{n} F_i(N_i) \\ \text{où } S_K &= \left\{ N \in \mathbb{Z}_+^n, N = \sum_{i=1}^{n} \right\} \end{aligned} \tag{5.5}$$

Les fonctions $F_i(N_i)$ correspondent aux probabilités stationnaires $\pi_i(N_i)$ de la station

i d'un réseau ouvert et sont données par

$$F_i(N_i) = \begin{cases} \left(\frac{\lambda_i}{\mu_i}\right)^{N_i} \times \frac{1}{N_i!} & 1 \leq N_i \leq c_i \\ \left(\frac{\lambda_i}{\mu_i}\right)^{N_i} \times \frac{1}{c_i! c^{N_i - c_i}} & c_i \leq N_i \leq K \end{cases} \qquad i = 1, 2, \ldots, n \tag{5.6}$$

Les valeurs λ_i apparaissant dans (5.6) sont, comme pour un réseau ouvert, solutions des équations d'équilibre (5.2). Cependant, dans le cas d'un réseau fermé, le vecteur est nul. Le système (5.2) possède alors une infinité de solutions toutes égales à une constante multiplicative près. La solution à ce problème consiste simplement à fixer arbitrairement une des valeurs λ_i à 1 (ou à tout autre valeur positive).

Les principales difficultés lors de l'étude d'un réseau de Jackson fermé ne se situent pas au niveau théorique mais apparaissent lors du calcul de la constante de normalisation $G(K)$. Le calcul de cette constante à partir de sa définition (5.5) n'est, en effet, pas réaliste même pour des valeurs modestes de n et K. Il existe cependant des méthodes efficaces, telles que les algorithmes de convolution ou des méthodes d'analyse en valeur moyenne, permettant le calcul de la distribution stationnaire de réseaux de tailles respectables.

Annexe A

Processus de Markov

Le processus de Markov fournit un outil simple de modélisation d'une classe particulièrede systèmes à espace d'états discret. L'analyse des processus de Markov est un préliminaire nécessaire à l'étude des systèmes de files d'attente.

Définition A.1 *Etant donné un espace fondamental* Ω *et une probabilité* Pr,
on appelle ***variable aléatoire*** *sur cet espace, toute application X de* Ω *dans* $\mathbb{R}$ *telle que :*

$$X : \Omega \longrightarrow \mathbb{R}$$
$$\omega \longmapsto X(\omega)$$

1. *Une variable aléatoire est dite* ***discrète*** *si elle ne prend que des valeurs discontinues dans un intervalle donné (borné ou non borné). L'ensemble des nombres entiers est discret. En règle générale, toutes les variables qui résultent d'un dénombrement ou d'une numération sont de type discrètes.*
2. *Une variable aléatoire est dite* ***continue*** *si elle peut prendre toutes les valeurs dans un intervalle donné (borné ou non borné). En règle générale, toutes les variables qui résultent d'une mesure sont de type continu.*

Définition A.2 ***Un processus stochastique*** $\{X(t), t \in T\}$ *est une collection de variables aléatoires indexées par un paramètre t et définies sur un même espace de probabilités* $\{\Omega, F, P\}$. *Le paramètre t est généralement interprété comme le temps.*

Définition A.3 *La variable $X(t)$ représente **l'état** du processus au temps t et l'ensemble de toutes les valeurs possibles pour cette variable est appelé **l'espace des états** du processus et sera noté S.*

Définition A.4 *Un processus stochastique est à **temps continu** lorsque l'ensemble T est non dénombrable (le plus souvent $T = \mathbb{R}^+$). On le dénote par $\{X(t), t \geq 0\}$.*

Définition A.5 *Un processus stochastique est à **temps discret** lorsque l'ensemble T est fini ou bien dénombrable (le plus souvent $T = \mathbb{Z}^+$). On le dénote par $\{X_n, n \geq 0\}$.*

Définition A.6 *Un processus stochastique dont l'ensemble des états S est fini ou dénombrable est appelé une **chaîne**.*

Définition A.7 *Le processus stochastique $\{X(t), t \geq 0\}$ est un processus de Markov si et seulement si pour tout instant $t \geq 0$ et tout sous ensemble d'états $I \subseteq S$, il est vraie que*

$$\Pr(X(t+\Delta) \in I / X(u), 0 \leq u \leq t) = \Pr(X(t+\Delta) \in I / X(t)), \forall \Delta \geq 0 \tag{A.1}$$

Remarque A.1 *L'expression(A.1) est appelée **la propriété markovienne**, cette propriété signifie que l'état présent du processus, c'est à dire son état à l'instant t, résume toute l'information utile pour connaître son évaluation future. En d'autres termes, la prévision de cette dernière ne peut être améliorée par une connaissance supplémentaire du passé du processus.*

Définition A.8 *La chaîne $\{X_n, n \geq 0\}$ à valeurs dans S est **une chaîne de Markov** si et seulement si pour tout $n \in \mathbb{N}$, la loi de X_{n+1} sachant X_0, X_1,..., X_{n+1} est égale à la loi conditionnelle sachant X_n, i.e. $\forall i_0, i_1, ..., i_{n-1}, j \in S$*

$$p_{ij} = \Pr(X_{n+1} = j / X_0 = i_0, ..., X_{n-1} = i_{n-1}, X_n = i_n) = \Pr(X_{n+1} = j / X_n = i_n = i) \tag{A.2}$$

Remarque A.2 *La probabilité conditionnelle donnée dans (A.2) est appelée **probabilité de transition**.*

Définition A.9 *Une matrice carrée $M = (p)_{ij}$ est **stochastique** si :*

$$\begin{cases} 0 \leq p_{ij} \leq 1, \forall i, j \\ \sum_{j=1}^{n} p_{ij} = 1, \forall i = 1, 2, ..., n \end{cases}$$

Définition A.10 ***Le graphe des transitions*** *est formé de points représentant les états du processus et d'arcs correspondant aux transitions possibles c-à-d pour lesquelles les probabilités p_{ij} existent.*

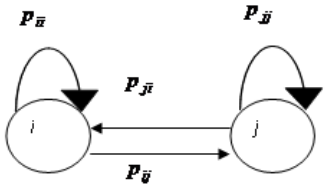

FIGURE A.1 *Graphe de transitions*

Définition A.11 *On définit **les probabilités d'état** définit comme suit : $\pi_k(n) = \Pr(X_n = k)$. On pose :*

$$\begin{cases} \pi(n) = (\pi_0(n), \pi_1(n), ...) \\ p_{ij} = \Pr(X_{n+k} = j / X_n = i) \\ M = (p_{ij})_{i,j} \end{cases}$$

Propriété A.1 *Pour tout m, n entiers : $\pi(n+1) = \pi(n)M$.*

Classifications des états d'une chaîne de Markov

1. On dit que deux états i et j **communiquent** si l'on peut passer de l'état i à l'état j ainsi de l'état j à l'état i avec des probabilités non nulles. on note $i \longleftrightarrow j$.
2. Ce concept de communication est une **relation d'équivalence** sur l'ensemble des états S. Donc on peut définir des **classes d'équivalence** sur l'ensemble S.

3. Une classe est dite **transitoire** s'il est possible d'en sortir, mais dans ce cas, le processus ne pourra plus jamais y retourner.
4. Une classe est dite **récurrente** s'il est impossible de la quitter.
5. Une chaîne de Markov est **irréductible** s'il existe une seule classe (tous les états communiquent).

Distribution stationnaire

1. Une chaîne de Markov est dite **stationnaire** si la distribution $\pi(n)$ de la variable aléatoire X_n est indépendante du temps $(n = 0, 1, 2, ...)$, en d'autre termes si $\pi(0)$ est une distribution stationnaire du processus en question.
2. Une Chaîne de Markov est dite **ergodique** si la limite $\lim_{n\to\infty} \pi(n)$ existe et ne dépend pas du vecteur stochastique initial $\pi(0)$.
3. Pour calculer les composantes d'une matrice-ligne stationnaire $\pi = (\pi_0, \pi_1, \pi_2, \dots)$ d'une chaîne de Markov finie, on a les deux approches suivantes :
   — On résout le système d'équations linéaires formés de $\pi M = \pi$, et la condition de normalisation $\sum_{k\in S} \pi_k = 1$.
   — Dans le graphe des transitions, on interprète les probabilités π_k comme des masses associées aux états $k \in S$ et les produits $\pi_k p_{kj}$ comme des flux de masses entre les deux états k et j. la répartition des masses π_k est stationnaire si lors d'une transition, le flux d'entrée est égal au flux de sortie pour chacun des états. On les appellera équations de blance.

Annexe B

Processus de Poisson

Le processus en question est utilisé pour décrire la réalisation dans le temps d'événements aléatoires d'un type donné. La description mathématique d'un flux d'événements aléatoires peut se faire de deux manières différentes :

1. On considère le nombre d'événements $X(t)$ se produisant dans$[0,t]$ et on cherche à déterminer la loi de probabilité de cette variable aléatoire discrète. Le processus $\{X(t), t \geq 0\}$ est appelé *processus de comptage*.
2. On considère les intervalles de temps qui séparent les instants d'apparition de deux événements consécutifs. Ce sont des variables aléatoires continues, positives et en général indépendantes et identiquement distribuées.

Définition B.1 *On dit qu'un processus de comptage $\{X(t), t \geq 0\}$ est un processus de Poisson s'il satisfait aux 3 conditions suivantes :*

1. *Le processus est homogène dans le temps : la probabilité d'avoir k événements dans un intervalle de longueur t ne dépend que de t et non pas de la position de l'intervalle par rapport à l'axe temporel : $p_k(t) = \Pr(X(t) = k)$.*
2. *Pour tout système d'intervalles disjoints, les nombres d'événements s'y produisant sont des variables aléatoires indépendantes.*

3. La probabilité

$$p_k(\Delta t) = \begin{cases} 0(\Delta t) & si \;\; k \geq 2 \\ \lambda \Delta t + 0(\Delta t) & si \;\; k = 1 \\ 1 - \lambda \Delta t + 0(\Delta t) & si \;\; k = 0 \end{cases} \tag{B.1}$$

où λ est la densité ou intensité du processus (le nombre moyen d'événements qui apparaissent par unité de temps).

Théorème B.1 *Pour un processus de Poisson, on a :*

$$\Pr(X(t) = k) = p_k(t) = \frac{(\lambda t)^k}{k!} \exp\{-\lambda t\}, \lambda > 0, k \geq 0 \tag{B.2}$$

$$E[X(t)] = \lambda t \tag{B.3}$$

$$Var[X(t)] = \lambda t \tag{B.4}$$

Ces relations définissent le régime transitoire du processus de Poisson. Aucun régime stationnaire n'existe vu que $p_k = \lim_{t \to \infty} p_k(t) = 0, \forall k \geq 0$.

Théorème B.2 *Le temps V qui sépare un instant quelconque du prochain événement est une variable aléatoire répartie selon une loi* $\exp\{\lambda\}$.

La loi exponentielle

Une variable aléatoire X absolument continue est dite variable exponentielle de paramètre α (α est une constante positive) si la densité de probabilité est :

$$f(x) = \begin{cases} \alpha \exp(-\alpha x) \text{ si } x \geq 0 \\ 0 \text{ si } x < 0 \end{cases}$$

On a :

$$E[X] = \frac{1}{\alpha}$$
$$Var[X] = \frac{1}{\alpha^2}$$

Annexe C

Transformée de Laplace-Stieltjes (T L-S)

Soit X une variable aléatoire continue, $F(x)$ est sa fonction de répartition. Sa transformée de Laplace-Stieltjes est donnée par :

$$\widetilde{F}(s) = E\left[e^{-sX}\right] = \int_0^\infty e^{-sx} f(x)\, dx = \int_0^\infty e^{-sx} dF(x)$$

où s est une variable complexe. Cette intégrale est définie lorsque $\Re(s) \geq 0$.

Propriété C.1 *Si X et Y sont des variables aléatoires indépendantes, alors :*

1. $E\left[e^{-s(X+Y)}\right] = E\left[e^{-sX}\right] E\left[e^{-sY}\right]$.
2. $TLS\left(\frac{df(t)}{dt}\right) = s\widetilde{F}(s)$.
3. $TLS\left(\frac{d^n f(t)}{dt^n}\right) = s^n \widetilde{F}(s)$.
4. $TLS\left(\int_0^t f(x)\, dx\right) = \frac{\widetilde{F}(s)}{s}$.
5. $\widetilde{F}(0) = \int_0^\infty f(t)\, dt$.
6. $\alpha_k = (-1)^k F^{(k)}(0)$.

Exemple C.1 ***Loi exponentielle***

Fonction de densité : $f(x) = \lambda e^{-\lambda x}, x \geq 0$

Sa fonction de TLS : $\widetilde{F}(s) = \frac{\lambda}{\lambda+s}$.

Exemple C.2 ***Loi gamma*** $G(n,\lambda)$

Fonction de densité : $g(n,\lambda,x) = \frac{\lambda(\lambda x)^{n-1}}{(n-1)!}e^{-\lambda x}, x \geq 0$

Sa fonction de TLS : $\widetilde{F}(s) = \left(\frac{\lambda}{\lambda+s}\right)^n, n = 1,2,....$

Annexe D

$Z-$transformée (fonction génératrice)

Soit X une variable aléatoire discrète à valeurs entières :

$$X = n, n \in \mathbb{N}, p_n = \Pr(X = n)$$

La fonction génératrice de la variable X est donnée par la formule suivante :

$$F(z) = \sum_{n=0}^{\infty} z^n p_n = E\left[z^X\right],$$

où z est une variable complexe. $F(z)$ est définie si $|z| \leq 1$ et $F(0) = p_0$ ainsi que $F(1) = 1$

Propriété D.1 *Si X et Y sont des variables aléatoires indépendantes à valeurs entières, alors :*

1. $E[X] = F^{'}(1)$.

2. $E[X^2] = F^{''}(1) + F^{'}(1)$.

3. $E[z^{X+Y}] = E[z^X]E[z^Y]$.

Formules usuelles

(1) $\sum_{n=0}^{\infty} z^n \frac{\partial p_n}{\partial a} = \frac{\partial F(z)}{\partial a}$	(2) $\sum_{n=0}^{\infty} n p_n z^n = z\frac{dF(z)}{dz}$
(3) $\sum_{n=0}^{\infty} \sum_{k=0}^{n} p_k z^n = \frac{F(z)}{1-z}$	(4) $\sum_{n=0}^{\infty} z^n p_{n-1} = zF(z)$
(5) $\sum_{n=0}^{\infty} z^n p_{n+1} = \frac{1}{z}[F(z) - p_0]$	(6) $F(1) = \sum_{n=0}^{\infty} p_n$
(7) $F(-1) = \sum_{n=0}^{\infty} (-1)^n p_n$	(8) $F(0) = p_0$
(9) $\sum_{n=0}^{\infty} z^n (a p_n + b q_n) = aF(z) + bQ(z)$	(10) $\sum_{n=0}^{\infty} a^n p_n z^n = F(az)$

Exemple D.1 *La fonction génératrice de la loi binomiale est*

$$F(z) = \sum_{k=0}^{n} C_n^k p^k (1-p)^{n-k} z^k$$

Exemple D.2 *La fonction génératrice de la loi Poisson est*

$$F(z) = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} z^k = e^{\lambda z} e^{-\lambda} = e^{\lambda(z-1)}$$

Exemple D.3 *La fonction génératrice de la loi géométrique est*

$$F(z) = \sum_{k=0}^{\infty} (1-p)^k p z^k = \frac{p}{1-(1-p)z}$$