Evaluation des performances des systèmes de communication

Cours

Prof. Djellab Natalia

Chapitre 1

Introduction

1.1 Qu'est-ce que l'évaluation des performances ?

En raison de leur évolution continue et rapide, les systèmes informatiques et réseaux de communication deviennent de plus en plus complexes et le besoin d'outils et de techniques permettant d'analyser et d'étudier leur comportement augmente sans cesse.

Les études de performance sont nécessaires pour fournir des réponses aux questions de coût, performance, sécurité, surgissant durant la vie du système. Par exemple, 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 exemple, le taux de hits, c'est-à-dire, le pourcentage de documents qui sont dans le cache lorsqu'on les demande. Ces questions sont de grande importance pour les organismes impliqués. Des réponses incorrectes engendrent de potentielles répercussions : problèmes de sécurité, performance, etc.

L'évaluation des performances intervient à deux niveaux : en conception et en exploitation.
En conception, le système n'existe pas encore et il s'agit de le créer en respectant le cahier des charges. Il est alors judicieux d'étudier le comportement du système avant son déploiement sur le terrain afin de comprendre et régler les éventuels problèmes qui pourront affecter le système. Par exemple, pour concevoir un réseau de communication, on doit être en mesure de connaître le débit souhaité, capable de transférer différents types de médias, tels que la voix, la vidéo, ou des données tout en respectant les contraintes temps réel 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 nombreux, pourtant, il est absolument important, pour une configuration donnée, 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 inversement un système sur-dimensionné entraînera un gaspillage d'argent inutile.
En exploitation, 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. Ceci rejoint un peu la conception. Il s'agira, par exemple, de modifier un serveur dont le taux d'exécution est insuffisant, en remplaçant son processeur par un autre deux fois plus puissant ou de modifier un réseau de communication en remplaçant la bande passante par une autre de capacité plus importante de telle manière à satisfaire la demande.

Cependant, pour réaliser l'évaluation de performance d'un système informatique ou d'un réseau, il faut disposer d'outils adéquats considérant les différents aspects du système. Les différentes techniques d'évaluation des performances peuvent être réunies en deux groupes : *mesures directes sur le système réel* et *l'évaluation des performances sur un modèle du système*. Le modèle d'un système est une abstraction pour laquelle seules certaines caractéristiques du système ont été retenues. La construction des modèles s'appuie généralement sur la modélisation analytique et la simulation. Chacune des techniques a ses

possibilités spécifiques et ses restrictions. Souvent l'évaluation des performances d'un même système demande l'utilisation de différentes méthodes.

1.2 Types d'analyse des systèmes

On distingue deux grands types d'analyse : *analyse qualitative* et *analyse quantitative.*
L'analyse qualitative consiste à définir les propriétés structurelles et comportementales du système, telles que l'absence de blocage (vivacité), les invariants du système, l'équité, l'inévitabilité, etc. Par exemple, dans un système informatique où deux processus s'exécutent en parallèle et partagent une ressource critique (par exemple, deux utilisateurs qui se partagent une imprimante), on veut vérifier que, quelque soit l'état du système, la ressource critique ne peut être utilisée par plus d'un processus (exclusion mutuelle). Cette propriété est une propriété d'invariance. De même, dans un réseau de communication, si une machine M1 attend un message d'une machine M2 pour poursuivre son processus et en même temps la machine M2 attend un message de la machine M1 pour poursuivre son exécution, le système est dans un état d'inter-blocage (deadlock) et ne peut plus évoluer. Ainsi, l »étude qualitative nous renseignera sur l'éventualité d'un tel état. Pour palier à ce blocage, la modification du protocole de communication s'impose. Le formalisme *Réseaux de Petri* et *Algèbres de processus* sont des formalismes les plus importants utilisés pour l'évaluation qualitative.

L'analyse quantitative concerne le calcul des mesures que l'on veut effectuer sur un système informatique, permettant de quantifier ses performances : débit, temps de réponse, nombre moyen de ressources occupées, taux d'utilisation de ses ressources, … Ces paramètres sont des critères principaux qui seront utilisés par la suite pour optimiser le fonctionnement du système ou prédire ses performances. Ils sont obtenus en exploitant les relations fondamentales qui opèrent sur les données du système. L'analyse quantitative doit toujours être précédée d'une analyse qualitative. En effet, il est inutile de s'intéresser aux indices de performance d'un système qui est dans un état de blocage et l'évaluation de performance peut être achevée sans se rendre compte qu'un tel état de blocage peut se produire.

1.3 Critères de l'évaluation des performances

Du point de vue de l'utilisateur, de l'administrateur, du concepteur du système informatique ou du réseau de communication, les critères d'évaluation des performances diffèrent en fonction de leurs besoins et leurs espoirs. Ces personnes peuvent avoir des besoins communs et contradictoires.

L'utilisateur s'intéresse à son expérience avec le réseau et non à la performance technique de celui-ci, comme c'est le cas de l'administrateur. Le concepteur s'intéresse à ce qui touche l'utilisateur et l'administrateur.

Les critères de l'utilisateur peuvent être : la disponibilité du système à tout moment, la sécurité, « en équité ». Il veut le délai de transmission des messages le plus court possible (ceci implique une faible utilisation du réseau).

Toutefois, l'administrateur voudrait voir l'utilisation du système aussi grande que possible. Il peut aussi, s'intéresser par les caractéristiques du trafic du réseau (comment les chemins sont calculés, combien de fois les tables d'acheminement sont adaptées, :::) et par la fiabilité.

Le concepteur mesure la performance du réseau dans le but de vérifier comment la performance actuelle s'accorde avec la performance prédite. Il est surtout concerné par les critères : la capacité des buffers, l'efficacité du protocole, l'efficacité du contrôle des flux.

Certains critères, comme le temps de transmission, l'utilisation des ressources du réseau, peuvent être mesurés numériquement, les autres, comme la disponibilité, sont subjectifs.

1.4 Caractérisation de la charge de travail (workload)

On entend par charge du système, la quantité de requêtes imposées par un ensemble de tâches dans une application donnée. Le taux d'arrivée des tâches ou plus généralement le mécanisme d'arrivées des tâches est certainement un facteur qui détermine la charge de travail du système. Un autre problème de caractérisation de la charge consiste à représenter les requêtes par tâches individuelles. Par exemple, pour un ordinateur constitué de ressources multiples, les demandes de travail d'une tâche doivent être représentées par au moins : la demande de travail CPU, la demande d'espace mémoire, la demande de travail d'E/S et les demandes sur les composantes software (applications).

Les performances d'un système dépendent étroitement de la demande en ressources matérielles et logicielles de la charge à exécuter. La caractérisation de la charge procure des bases pour construire une charge synthétique exécutable et représentative afin de générer un système à mesurer ou pour obtenir des valeurs représentatives de paramètres pour les modèles analytiques ou de simulation.

1.5 Approches de l'analyse quantitative

1.5.1 Mesures directes

L'objectif principal de cette approche est de cumuler les statistiques sur les événements divers, de les interpréter en termes de la performance du système opérationnel et de régler les paramètres de ce dernier pour atteindre la performance la plus optimale possible. Une autre motivation d'obtenir les mesures sur le système réel est d'estimer les paramètres d'entrée, nécessaires pour les modèles.

Les mesures peuvent être classées en deux catégories : orientées *utilisateur, orientées système*. L'importante mesure orientée utilisateur est le temps de réponse : l'intervalle du temps entre l'arrivée de la demande et son achèvement dans le système. Les mesures orientées système sont le débit et l'utilisation. Le débit est défini comme le nombre moyen de jobs traités par unité de temps. L'utilisation d'une ressource est une fraction du temps pendant laquelle la ressource particulière est occupée.

La réalisation des mesures nécessite l'emploi des outils : *moniteurs matériels, moniteurs logiciels*. Un moniteur matériel, souvent une station particulière, sert à observer l'état des composantes matérielles du système. Il présente les statistiques comme le nombre d'événements spécifiques durant une période donnée du temps (le nombre de messages entrant et quittant le réseau), le débit, l'utilisation d'une ressource individuelle. L'avantage principal d'utilisation d'un moniteur matériel est la précision des résultats fournis.

Un moniteur logiciel est l'ensemble des routines du système. Il cumule l'information concernant le statut du système. Il y a deux types de moniteurs logiciels : *moniteurs de traçage d'événements, moniteurs-échantillonneurs*. Le moniteur de traçage d'événements fournit des statistiques concernant le fait : comment les demandes individuelles progressent dans le système. Parmi elles se trouvent : le nombre de nœuds qu'une demande visite avant d'arriver à sa destination, le temps de réponse d'une demande. Le moniteur-échantillonneur tient le statut instantané du système dans les intervalles périodiques. Il donne les statistiques momentanées, comme les longueurs des files d'attente aux niveaux des nœuds, les paramètres de contrôle de flux. Les moniteurs logiciels peuvent modifier les performances du système ; les performances mesurées deviennent alors différentes des performances réelles. Il est nécessaire de tester la validité d'un moniteur logiciel en le comparant à un moniteur matériel. Il est possible de combiner les avantages des moniteurs matériels (vitesse et précision) avec

ceux des moniteurs logiciels (flexibilité et accès aux informations logicielles) dans un *moniteur hybride.*

Les problèmes majeurs associés aux mesures directes sont :

- Le temps nécessaire pour cumuler assez de statistiques peut être très variable : à partir de quelque jours jusqu'à plusieurs semaines.
- Il est facile de trouver où le paramètre doit être incrémenté ou décrémenté pour arriver aux buts de la performance, mais la tâche de décider de combien il faut incrémenter ou décrémenter n'est évidente.

1.5.2 Modèles et modélisation

Une autre approche pour évaluer la performance d'un système informatique ou d'un réseau comprend la construction du modèle du système. La modélisation a un objectif bien déterminé qui consiste à fédérer en un seul objet des parties du système que l'on veut analyser. Le processus de modélisation s'apparente à une phase de transformation d'un système dans un symbolisme de représentation qui enveloppe les composantes principales caractérisant le comportement du système. Ce symbolisme de représentation s'appuie également sur des outils théoriques qui garantissent la structure des objets manipulés, et les opérations, que on veut leur appliquer.

La modélisation est la substitution d'un système par un modèle que l'on pourra résoudre. Il est souhaitable, que le modèle possède une structure modulaire hiérarchique afin d'exprimer la structure modulaire du système réel/ Ceci signifie que le modèle est décomposé en sous-modèles ou couches d'une manière « top to bottom » : la fonction de chaque couche est constituée des fonctions des couches sous-jacentes. Cette décomposition commence par la définition de la fonction globale et des exigences de la performance et est suivie par la détermination des fonctions et des niveaux de la performance, qui sont nécessaires pour attendre les objectifs. Cette procédure est répétée (donc une série de couches est produite) jusqu'à ce que le système soit considéré en termes de primitives (les éléments matériels du système).

L'avantage de la modélisation par rapport aux mesures directes est qu'elle peut être employée aussi bien pendant les phases de conception d'un système, que durant les phases de l'exploitation.

La première étape de modélisation consiste à identifier et à examiner les paramètres des composants 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 terminaux et équipements de communication. Il est également nécessaire de connaître les composants software : l'algorithme d'ordonnancement des tâches, l'algorithme de gestion de la mémoire, l'algorithme de distribution (dispatching) 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 trafic (de la charge) prévue pour chacune de ces composantes : le taux d'arrivée des tâches, le temps de CPU par tâche, les besoins en espace mémoire, le taux de défauts 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 décrivant le comportement du système.

Soient principalement deux types de modèles utilisés pour évaluer la performance des systèmes informatiques et réseaux : les modèles offrant les solutions analytiques et les modèles de simulation.

Le modèle analytique permet d'écrire la relation fonctionnelle entre les paramètres du système et le critère d'évaluation de performance choisi en termes d'équations, qui peuvent être résolues numériquement, ou fournit des moyens pour obtenir la solution analytique. Le plus souvent, le modèle analytique enveloppe les concepts de la théorie des filles d'attente. Les modèles analytiques sont rapides, économiques et faciles quand on travaille avec eux. Cependant, la construction du modèle exige de bonnes connaissances des mécanismes de modélisation. Dans le cas des systèmes complexes, plusieurs suppositions doivent être faites ; ceci implique l'obtention des résultats approchés. Par conséquent, les modèles analytiques sont bons pour un « gros design ».

Un modèle de simulation décrit le comportement dynamique réel d'un système même si l'analyste ne s'intéresse qu'à la valeur moyenne de quelques mesures de performance (temps de réponse, utilisation d'une CPU) à l'état stationnaire. Il s'agit d'implémenter un modèle simplifié du système à l'aide d'un programme de simulation adéquat. C'est une technique largement utilisée pour l'évaluation des performances de systèmes informatiques et réseaux de communication. Elle permet de traduire d'une manière plus réaliste le comportement du système à évaluer.

La technique de simulation constitue un outil très important pour la détermination des différences de performance entre les configurations alternatives (aussi bien hardware que software). La simulation permet en plus de visualiser les résultats sous formes de graphes faciles à analyser et à interpréter. Elle rend possible l'analyse systématique des systèmes lorsque les solutions analytiques ne sont pas disponibles et l'expérimentation sur le système considéré (mesures directes) est impossible ou non pratique.

Le modèle de simulation est conduit ou bien par la génération des données d'entrée (pseudo-aléatoires), on parle alors de *la simulation probabiliste ou Monté Carlo* ; ou bien par l'introduction des données d'entrée, on parle alors de *simulation déterministe ou par trace*. Encore, la simulation peut être *continue* (elle est utilisée pour construire un modèle d'un système continu ; le temps est représenté par des équations mathématiques et varie continuellement) et *à événements discrets* (le programme de simulation produit une liste d'événements à apparaître). Le simulateur simule donc le comportement dynamique actuel du système. En répétant le processus pour des configurations et paramètres du système alternatifs différents, on peut identifier une structure de système optimale. Cependant, le module de simulation prend beaucoup de temps pour l'élaboration, nécessite beaucoup de temps d'exécution et malgré tout fournit beaucoup moins d'informations. C'est pourquoi, la simulation est généralement considérée comme une technique de dernier recours.

Un problème de simulation peut être résolu en suivant les étapes suivantes :

1. construction du modèle de simulation ;
2. implémentation du modèle ;
3. création des expériences de simulation ;
4. validation du modèle de simulation ;
5. exécution du simulateur et analyse de données.

Les systèmes informatiques et réseaux devenant de plus en plus complexes, le besoin de développement des langages et outils de simulation de plus haut niveau augmente sans cesse. Ces langages sont basés sur la simulation orientée événements discrets. L'un des langages les plus largement utilisés est le GPSS (General Purpose System Simulator), interprété et développé par IBM. Sa première version a paru en 1961. La version GPSS V a également été développé en 1971 par IBM. Ce langage permet de décrire directement le flot fonctionnel des tâches (appelées transactions) à travers le système. Le langage SIMSCRIPT, basé sur Fortran

et initialement développé par Markowitw, Karr et Hauser en 1963 au Rand Corporation est l'un des langages les plus disponibles et usuels, peut être juste après le GPSS. La version SIMSCRIPT 1.5, développée par Karr, Kleine et Markowitz en 1965, SIMSCRIPT II par Kiviat, Villanueva et Markowitz en 1969 et SIMSCRIPT II.5 par Consolidated Analysis Center, Inc. En 1971 sont des extensions successives. Plusieurs autres langages ont été développés tels que SIMPL/1 par IBM, SIMULA par Dahl et Nygoard en 1966.

Récemment, plusieurs autres outils sont développés pour répondre aux besoins d'évaluation de performance des nouvelles architectures des systèmes informatiques et réseaux de communication.

QNAP/modline est un langage de description, de simulation et d'évaluation de systèmes à événements discrets (réseaux informatiques, de télécommunication, systèmes de production). La modélisation est basée sur le principe des files d'attente. *Modline* est une sur-couche graphique permettant de designer de tels systèmes. Il utilise un langage très semblable au Pascal pour la description des modèles.

L'outil NS-2/NAM (Network Simulator 2) est un simulateur développé à Lawrence Berkely National Laboratory et conçu principalement pour le monde de l'Internet. Il permet de simuler le comportement des protocoles TCP, IP, d'étendre le simulateur aux protocoles spécifiques de l'Internet (routage, transport, application) et aux nouvelles architectures de qualité de service (IntServ, DiffServ MPLS, RED). Le langage de base de NS est le C++. L'outil NAM (Network Animator) associé au NS-2 permet de visualiser des animations de la simulation (transfert des paquets d'un nœud à un autre, taille des paquets, remplissage des files d'attente, …).

OPNET (Optimum Network Performance) est un outil très puissant pour la simulation et l'évaluation des performances de réseaux. Il permet aussi à l'utilisateur de construire ses propres modèles du plus simple au plus complexe. Il possède trois niveaux d'abstraction pour construire les modèles et décrit les processus à l'aide des automates et intègre ces processus dans les nœuds formant un réseau ou un dispositif informatique. Le langage de base de OPNET est le C.

Il existe plusieurs autres outils de simulation tels que PARSEC (Parallel Simulation Environment for Complex System), GloMoSim (Global Mobile Simulator).

1.6 Quelques exemples d'application

Un système informatique moderne est une organisation complexe, à la fois de par l'architecture matérielle et de par le type de programmes qui y sont traités. Même centralisé, il comporte de nombreuses entités distinctes 'contrôleurs, disques, coprocesseurs, etc.), qui coopèrent. Les mesures de performance et de prédiction diffèrent d'un système à un autre.

Exemple 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 CPU, d'un disque de pagination (DiscP) et d'un disque fichier (DiscF). L'objectif de l'utilisation de la MV est d'augmenter le taux d'utilisation de la CPU. 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 principale (MC) virtuelle. Il s'agit de donner à chaque programme qui se présente une place en MC et de mettre tout ce qui ne peut entrer dans cette place sur un disque ou mémoire secondaire appelé *disque de pagination*. La MC est découpée en plusieurs pages (variant de 500 à K octets). Lorsque l'information nécessaire pour l'exécution d'un programme n'est pas située sur l'une des 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 le disque DiscP 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 millisecondes pour un chargement et pendant ce temps là, l'UC ne fonctionne pas. Il est bien clair qu'il existe un degré 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 infinie de mémoire. A chaque instruction, il va falloir effectuer un remplacement de page.
Problématique : On souhaite connaître le degré de multiprogrammation 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 (ces deux grandeurs sont dépendantes). Ce même problème peut se poser pour tous les systèmes fonctionnant en multiprogrammation, tels qu'un *serveur de messagerie.*

Exemple 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'accès 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 accès dès que ceux-ci se présenteront.
Problématique : 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 de données est en mode normal de fonctionnement ? Quel est le nombre moyen d'utilisateurs connectés à la base de données ? etc.

Exemple 3 : Serveur Web.
On considère un serveur Web équipé d'une mémoire cache de capacité c Mo. Le fonctionnement d'un tel serveur est le suivant : un document demandé et se trouvant dans la 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, lorsqu'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.

1.7 Notion d'un système

Un *système* est l'interaction d'une collection d'objets dans un environnement fermé. Tout système contient des objets identifiables qui peuvent varier en nombre. Ces objets sont

appelés *entités*. Une entité possède un nombre de caractéristiques identificatrices appelées *attributs*. Ces derniers sont reliés les uns aux autres et à leur environnement de façon variable quoique prescrite. Les valeurs des attributs d'une entité définissent son *état*. La collection des états de toutes les entités du système définit *l'état du système*. Toute action qui provoque un changement dans le système (changement des valeurs des attributs) est appelée *activité*. Les changements de l'état du système (dus aux activités) sont appelés *événements*. On définit encore *processus* comme un ensemble d'activités logiquement reliées.

Tout système est caractérisé par des processus « dynamiques » prenant place dans une structure « statique ». La structure statique est une charpente indépendante du temps à l'intérieur de laquelle les états du système sont définis. Les processus agissent et réagissent à l'intérieur de la structure statique du système changeant ainsi son état à mesure que le système évolue. Un système peut aussi être affecté par des changements dans son environnement.

Les activités peuvent être :

- endogènes : elles se produisent à l'intérieur du système ;
- exogènes : elles se produisent dans l'environnement et affectent le système.

Un système, qui ne subit pas l'influence d'activités exogènes, s'appelle *un système fermé*, par opposition à *un système ouvert.*

Un système peut être :

- discret : les changements de l'état du système interviennent à des instants discrets du temps ;
- continu : les changements de l'état du système s'effectuent d'une manière continue ;
- stable : on observe de faibles changements de l'état du système ;
- instable ;
- déterministe : les entités du système sont liées entre elles d'une manière bien déterminée ;
- stochastique : un élément de hasard est inclus.

Références

1. J.Y.L. Boudec. *Performance Evaluation of Computer and Communications systems.* Ecole Polytechnique Féderale de Laisanne, 2004.
2. B.R. Haverkort. *Performance of Computer Communication Systems: A Model-Based Approach.* John Wiley and Sons Ltd, 1998.
3. H. Kobayachi. *Modeling and Analysis, an Introduction to System Performance Evaluation Methodology.* Addison-Wesley Publishing Company, 1981.

Chapitre 2
Introduction aux systèmes de files d'attente

2.1Processus stochastiques : quelques définitions

Définition 1

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

Un processus est à temps continu lorsque l'ensemble T est non dénombrable (le plus souvent R^+). On le dénote par $\{X(t), t \geq 0\}$. Un processus est à temps discret lorsque T est fini ou tout au moins dénombrable (le plus souvent $T = Z_+$). On le dénote par $\{X_n, n \geq 0\}$.

Définition 2

L'ensemble de toutes les valeurs que peuvent prendre les variables définissant un processus stochastique est appelé *l'espace d'états* du processus et sera noté S. Si cet ensemble est fini ou dénombrable, le processus est appelé *une chaîne*.

Définition 3

Un processus stochastique $\{X(t), t \geq 0\}$ défini sur un espace d'états S satisfait *la propriété de Markov* si, pour tout instant $t \geq 0$ et tout sous ensemble d'états $I \subseteq S$, il est vraie que

$P(X(t+\Delta) \in I / X(u), 0 \leq u \leq t) = p(X(t+\Delta) \in I / X(t))$, $\forall \Delta \geq 0$.

Un processus stochastique vérifiant la propriété précédente est appelé *processus de Markov*.

Exemple : chaîne de Markov à temps discret, chaîne de Markov à temps continu.

2.2Processus 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 ($S = \{0,1,2...\}$). 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 $n-1$ ($n \geq 1$).

Soit $\{N(t), t \geq 0\}$, où $N(t)$ est le nombre d'individus dans la population à la date t, avec $S = \{0,1,2,...\}$. Le processus de naissance et de mort est caractérisé par l'apparition et la disparition d'un individu au sein de la population. Il est homogène dans le temps si la *probabilité* :

(a) d'apparition d'un individu pendant l'intervalle Δt sachant qu'il existe déjà *k* individus au sein de la population, qui est $\lambda_k \Delta t + 0(\Delta t)$,

(b) d'apparition d'aucun individu pendant l'intervalle Δt sachant qu'il existe déjà *k* individus au sein de la population, qui est $1 - \lambda_k \Delta t + 0(\Delta t)$,

(c) d'apparition de deux ou plus individus pendant l'intervalle Δt sachant qu'il existe déjà *k* individus au sein de la population, qui est $0(\Delta t)$,

(d) de disparition d'un individu pendant l'intervalle Δt sachant qu'il existe déjà *k* individus au sein de la population, qui est $\mu_k \Delta t + 0(\Delta t)$,

(e) de disparition d'aucun individu pendant l'intervalle Δt sachant qu'il existe déjà k individus au sein de la population, qui est $1-\mu_k \Delta t + 0(\Delta t)$,

(f) de disparition de deux ou plus individus pendant l'intervalle Δt sachant qu'il existe déjà k individus au sein de la population, qui est $0(\Delta t)$,

est indépendante de la position de Δt sur l'axe des temps. Ici, λ_k est le taux d'apparition (de croissance), μ_k est le taux de disparition (de décroissance). Encore, $P(N(t+s)=j / N(s)=i) = p_{ij}(t)$ ne dépend pas de s. Alors,

$p_{i,i+1}(\Delta t) = \lambda_i \Delta t + 0(\Delta t)$, $i \geq 0$;

$p_{i,i-1}(\Delta t) = \mu_i \Delta t + 0(\Delta t)$, $i \geq 1$;

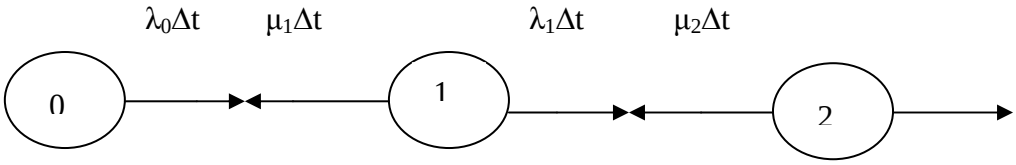
$p_{i,i}(\Delta t) = 1-(\lambda_i + \mu_i)\Delta t + 0(\Delta t)$, $i \geq 0$;

$p_{i,j}(\Delta t) = 0(\Delta t)$ si $|i-j| \geq 2$; $p_{ij}(0) = \delta_{ij} = \begin{cases} 1 & i=j \\ 0 & i \neq j \end{cases}$.

On a également que $\lambda_i > 0$, $\mu_i > 0$, $\mu_0 = 0$.

Régime transitoire

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



La matrice des transitions correspondante est

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

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

$p_0(t+\Delta t) = (1-\lambda_0 \Delta t)\, p_0(t) + \mu_1 \Delta t p_1(t)$;

(1.1)

$p_n(t+\Delta t) = \lambda_{n-1}\Delta t p_{n-1}(t) + [1-(\lambda_n+\mu_n)\Delta t] p_n(t) + \mu_{n+1}\Delta t p_{n+1}(t)$, $n \geq 1$.

On déduit les *équations de Kolmogorov*

$p_0'(t) = -\lambda_0 p_0(t) + \mu_1 p_1(t)$;

$p_n'(t) = \lambda_{n-1} p_{n-1}(t) - (\lambda_n + \mu_n) p_n(t) + \mu_{n+1} p_{n+1}(t)$, $n \geq 1$.

Remarques

1. Si $S = \{0,1,...,K\}$, 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 (1) en prenant $\lim_{t\to\infty}$) :

$\lambda_0 p_0 = \mu_1 p_1$;

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

(1.2) avec l'équation de normalisation $\sum_{n=0}^{\infty} p_n = 1$.

De (1.2), on obtient

$p_1 = \frac{\lambda_0}{\mu_1} p_0$;

Pour $n = 1$: $(\lambda_1 + \mu_1)p_1 = \mu_2 p_2 + \lambda_0 p_0$, $p_2 = \frac{\lambda_0 \lambda_1}{\mu_1 \mu_2} p_0$;

---------- ;

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

(1.3)

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

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

(1.4)

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, $p_n = 0 \quad \forall n \geq 0$.

2.3 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 v.a. continues, positives et en général indépendantes et identiquement distribuées.

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

- 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 : $p_k(t) = P(X(t) = k)$.
- Pour tout système d'intervalles disjoints, les nombres d'événements s'y produisant sont des variables aléatoires indépendantes.
- La probabilité $p_k(\Delta t) = \begin{cases} 0(\Delta t) & si \quad k \geq 2 \\ \lambda \Delta t + 0(\Delta t) & si \quad k = 1 \\ 1 - \lambda \Delta t + 0(\Delta t) & si \quad k = 0 \end{cases}$, où λ est la densité ou intensité du processus (le nombre moyen d'événements qui apparaissent par unité de temps).

Théorème 1 Pour un processus de Poisson, on a :

$P(X(t)=k)=p_k(t)=\frac{(\lambda t)^k}{k!}e^{-\lambda t}, \lambda>0, k\geq 0$;

$E[X(t)]=\lambda t$ et $Var[X(t)]=\lambda t$.

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

2.4 Définition et classification des systèmes de files d'attente

La formation des files d'attente est un phénomène fréquent qui apparaît chaque fois que les demandes de service dépassent la capacité admise des dispositifs de service.

Le modèle général d'un phénomène d'attente (système d'attente) peut être résumé comme suit : les demandes (clients) arrivent à un certain endroit et réclament un certain service. Si un dispositif de service est libre, la demande qui arrive se dirige immédiatement vers ce dispositif où elle est servie. Dans le cas contraire, on a deux possibilités : soit la demande quitte le système (systèmes à demandes refusées), soit elle prend place dans une file d'attente (système de files d'attente). A un moment donné, la demande est sélectionnée pour service selon une discipline. Après l'achèvement du service, la demande quitte le système.

Un système de files d'attente comprend donc un espace de service avec un ou plusieurs dispositifs de service (serveurs) et un espace d'attente dans lequel se forme une éventuelle file d'attente.

Pour identifier un système de files d'attente, on a besoin de spécifier le flux d'entrée, le mécanisme de service et la discipline d'attente.

La dimension d'une population des clients potentiels (ou le nombre de sources) peut être finie ou infinie. Un modèle avec la population finie est analytiquement plus compliqué parce que le nombre de clients déjà dans le système à n'importe quel instant affecte le nombre de clients potentiels composant la population. Le processus (flux) des arrivées peut être régulier ou aléatoire. Dans le premier cas, les arrivées des clients se suivent à des intervalles de temps déterminés. Dans les systèmes réels, on rencontre rarement de processus des arrivées de ce genre. Pour les systèmes plus typiques, le flux de demandes est aléatoire (la durée de temps entre deux arrivées successives suit une loi de probabilité).

Le mécanisme de service comprend le nombre de serveurs et la distribution des durées de service.

Les clients peuvent être choisis et servis dans l'ordre d'arrivée (FIFO), ou LIFO, ou choisis au hasard (RANDOM). La capacité de l'espace d'attente peut être illimitée ou non. Dans le deuxième cas, certains clients qui arrivent vers le système n'ont pas la possibilité d'y entrer.

Puisque les instants d'arrivée et les durées de service sont généralement des quantités aléatoires, le processus décrivant le fonctionnement d'un système de files d'attente est processus aléatoire (stochastique). Par ailleurs, on suppose généralement que toutes les variables aléatoires introduites pour décrire un système d'attente sont mutuellement indépendantes.

Pour la classification des systèmes de files d'attente, on a recours à une notation symbolique (notation de Kendall) comprenant 4 symboles rangés dans l'ordre **A/B/c/m**, où **A** et **B** décrivent respectivement la distribution des temps entre deux arrivées successives et la distribution des temps de service, **c** est le nombre de serveurs (montés en parallèle), **m** est la

capacité du système (le nombre de serveurs plus le nombre de position d'attente). Le dernier symbole peut être supprimé si **m=∞.**

Remarque

Pour spécifier les distributions **A** et **B**, on introduit les symboles suivants :

M – distribution exponentielle ;

E_k – distribution d'Erlang de degré k ;

H_k – distribution hyperexponentielle de degré k ;

D – déterministe ;

G – distribution générale.

Les caractéristiques d'exploitation du système: le temps d'attente d'un client *W*, le temps de séjour d'un client dans le système W_s, le taux d'occupation des dispositifs de service, la durée de la période d'activité, le nombre de clients dans le système *N*, nombre de clients dans la files d'attente N_f.

Les mesures de performance sont:

- 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 $\overline{W}_s$.

Soient encore des relations (formules de Little) :

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

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

ρ *= temps moyen de service/temps moyen entre deux arrivées successives.*

Chapitre 3

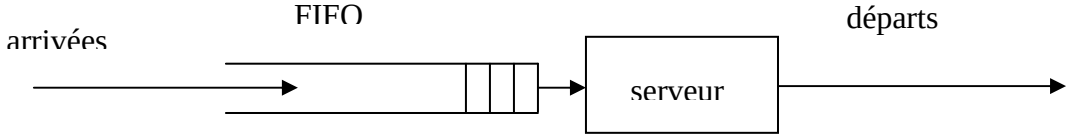
Systèmes de files d'attente régis par un modèle markovien de naissance et de mort

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

$$\{N(t), t \geq 0\}. \qquad (3.1).$$

3.1Système de files d'attente M/M/1

Description du modèle : Les clients arrivent vers le système selon un processus de Poisson de taux $\lambda > 0$ (nombre moyen de clients arrivant pendant une unité de temps) ; c'est-à-dire l'intervalle de temps entre deux arrivées successives 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 $1/\mu$. 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 (3.1).

Régime transitoire

Soit $p_n(t) = P(N(t) = n)$. Le graphe des transitions se présente de la manière suivante (figure1).

A partir du graphe des transitions, on obtient

$p_0(t+\Delta t) = \mu \Delta t p_1(t) + [1 - \lambda \Delta t] p_0(t)$;

$p_n(t+\Delta t) = \mu \Delta t p_{n+1}(t) + \lambda \Delta t p_{n-1}(t) + [1 - (\lambda + \mu)\Delta t] p_n(t), \quad n \geq 1.$

Puis, les équations de Kolmogorov :

$$\begin{cases} p_0'(t) = -\lambda p_0(t) + \mu p_1(t) \\ p_n'(t) = -(\lambda + \mu) p_n(t) + \lambda p_{n-1}(t) + \mu p_{n+1}(t), n \geq 1 \end{cases}. \qquad (3.2)$$

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 (3.1) ; et $\lim_{t\to\infty} p_n'(t) = 0, \quad n \geq 0$. De (2.2), on obtient le système d'équations de balance suivant

$$\begin{cases} \mu p_1 = \lambda p_0 \\ \lambda p_{n-1} + \mu p_{n+1} = (\lambda + \mu) p_n, n \geq 1 \end{cases} \tag{3.3}$$

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

La résolution du système (3.3) (la résolution du modèle) s'effectue de la manière suivante :

$p_1 = \frac{\lambda}{\mu} p_0$;

pour $n = 1$, $\lambda p_0 + \mu p_2 = (\lambda + \mu) p_1$, $p_2 = \left(\frac{\lambda}{\mu}\right)^2 p_0$;

pour $n > 1$, $p_n = \left(\frac{\lambda}{\mu}\right)^n p_0$.

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

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

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

où $1 + \frac{\lambda}{\mu} + \left(\frac{\lambda}{\mu}\right)^2 + ...$ est une progression géométrique de raison $\frac{\lambda}{\mu}$. Elle convergente si $\frac{\lambda}{\mu} < 1$, et est égale à $\frac{1}{1 - \frac{\lambda}{\mu}}$. Alors, $p_0 = 1 - \frac{\lambda}{\mu}$. D'où $p_n = \left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^n$.

$\rho = \frac{\lambda}{\mu}$ est l'intensité du trafic. $\rho < 1$ 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.

Remarques

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. La simplicité de la formule $p_0 = 1 - \frac{\lambda}{\mu}$ s'explique par la notion de conservation des clients : en régime stationnaire le taux des arrivées dans le système M/M/1 est égal à λ, celui de sortie serait égal à μ si le serveur était occupé en permanence, mais celui-ci n'est occupé qu'avec la probabilité $1 - p_0$. Le taux des départs valant alors $\mu(1 - p_0)$ est égal, en régime stationnaire, au taux des arrivées, soit 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 : Soit $N = \lim_{t\to\infty} N(t)$.

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

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

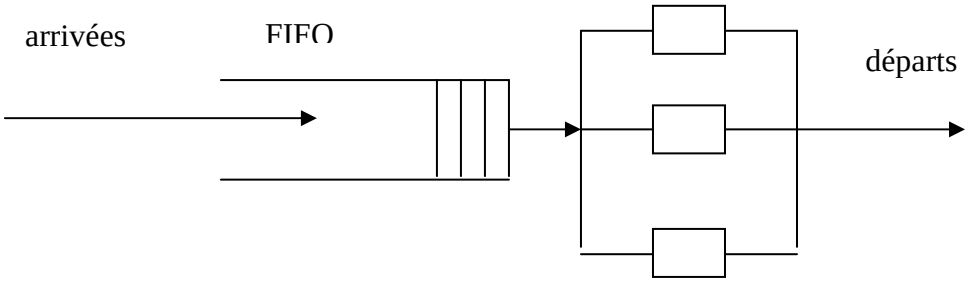
Soit $N_f = \lim_{t\to\infty} N_f(t)$, où $N_f(t)$ est le nombre de clients dans la file d'attente à la date t. La variable N_f est définie de la manière suivante : $N_f = \begin{cases} 0 & N = 0 \\ N-1 & N \geq 1 \end{cases}$.

$$\bar{n}_f = E\left[N_f\right] = \sum_{n=1}^{\infty} (n-1)p_n = \frac{\lambda^2}{\mu(\mu-\lambda)} \text{ ou bien } \bar{n}_f = \bar{n} - \rho \ .$$

Le temps moyen d'attente $\overline{W}$ et le temps moyen de séjour $\overline{W_s}$ peuvent être calculés soit à l'aide de formules de Little, soit à partir de la distribution stationnaire du système.

3.2 Modèle 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 $1/\mu$. 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 (2.1), dont l'espace des états est $S = \{0,1,2,...\}$ Ce dernier est un processus de naissance et de mort dont les taux de transition sont :

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

Le graphe des transitions est (figure2).

Régime transitoire

Le système d'équations de Kolmogorov pour les probabilités d'état $p_n(t) = P(N(t) = n), n \geq 0$, se présente de la manière suivante :

$p_0'(t) = -\lambda p_0(t) + \mu p_1(t)$;

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

$p_n'(t) = \lambda p_{n-1}(t) - (\lambda + c\mu) p_n(t) + c\mu p_{n+1}(t), \quad n \geq c$.

Régime stationnaire

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

$0 = -\lambda p_0 + \mu p_1$;

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

$\lambda p_{n-1} - (\lambda + c\mu) p_n + c\mu p_{n+1}, \quad n \geq c$.

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

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

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

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 $\frac{1}{c!}\left(\frac{\lambda}{\mu}\right)^c \left[1 + \frac{\lambda}{\mu c} + \left(\frac{\lambda}{\mu c}\right)^2 + ...\right]$. La somme […] possède une limite égale à $\frac{1}{1 - \frac{\lambda}{\mu c}}$ si $\frac{\lambda}{\mu c} < 1$. Par conséquent, le système considéré est en régime stationnaire si $\rho = \frac{\lambda}{\mu c} < 1$, ρ est l'intensité globale du trafic. On obtient ainsi

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

Encore,

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

Mesures de performance :

$$\overline{n} = \sum_{n=0}^{\infty} np_n = \sum_{n=1}^{c-1} \frac{n(\lambda/\mu)^n}{n!} p_0 + \sum_{n=c}^{\infty} \frac{n(\lambda/\mu)^n}{c!c^{n-c}} p_0 = \frac{\lambda}{\mu} + \frac{(\lambda/\mu)^{c+1}}{cc!\left(1-\frac{\lambda}{\mu c}\right)^2} p_0 \ ;$$

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

$$\overline{n}_f = \sum_{k=0}^{\infty} kp_{c+k} = \frac{(\lambda/\mu)^c}{c!} \sum_{k=0}^{\infty} k\left(\frac{\lambda}{\mu c}\right)^k p_0 = \frac{(\lambda/\mu)^{c+1}}{cc!\left(1-\frac{\lambda}{\mu c}\right)^2} p_0 \ ;$$

$$\overline{W} = \frac{\overline{n}_f}{\lambda} = \frac{c\mu(\lambda/\mu)^c}{c!(c\mu-\lambda)^2} p_0 .$$

Remarque : la distribution stationnaire peut s'obtenir rapidement en appliquant la relation établie pour les processus de naissance et de mort. En effet $p_n = \frac{\lambda_0\lambda_1...\lambda_{n-1}}{\mu_1\mu_2...\mu_n} p_0$, pour $n \le c$ il vient : $p_n = \frac{\lambda \times \lambda \times ... \times \lambda}{\mu \times 2\mu \times ... \times n\mu} p_0$, soit $p_n = \frac{\lambda^n}{n!\mu^n} p_0$. Pour $n \ge c$:

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

3.3 Modèle M/M/c/K

A présent, supposons que dans le système M/M/c, le nombre de positions d'attente est limité (égal à K). A l'arrivée d'un client, si tous les serveurs et toutes les positions d'attente sont occupées, le client quitte le système définitivement sans recevoir le service.

Analyse du modèle : Le processus (3.1) décrivant l'état du système à l'étude à la date t est celui de naissance et de mort avec $\lambda_n = \lambda$ si $0 \le n < K$, et $\mu_n = \mu \times \min\{n, c\}$ si $1 \le n \le K$ ($\mu_0 = 0$). L'espace des états est $S = \{0,1,2,...,K\}$.

Le graphe des transitions est (figure3).

Régime stationnaire

Soient $p_n = \lim_{t\to\infty} p_n(t), 0 \le n \le K$. La distribution stationnaire p_n satisfait le système d'équations de balance suivant

$0 = -\lambda p_0 + \mu p_1$;

$0 = \lambda p_{n-1} - (\lambda + n\mu) p_n + (n+1)\mu p_{n+1}, \quad 1 \le n < c$;

$0 = \lambda p_{n-1} - (\lambda + c\mu) p_n + c\mu p_{n+1}, \quad c \le n < K$;

$0 = \lambda p_{K-1} - c\mu p_K$.

La résolution de ce système, nous donne

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

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

La mesure importante de ce système est la probabilité de perte, qui est la probabilité pour que le système se trouve dans l'état K :

$$p_K = \frac{1}{c!}\left(\frac{\lambda}{\mu c}\right)^{K-c}\left(\frac{\lambda}{\mu}\right)^{c} p_0 .$$

La probabilité p_0 s'obtient à partir de l'équation de normalisation $\sum_{n=0}^{K} p_n = 1$:

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

Dans le cas particulier où $K = c$ (système à demandes refusées), la distribution stationnaire du processus $\{N(t), t \geq 0\}$ correspondant (formule d'Erlang) est

$$p_n = \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^{n} p_0 , 0 \leq n \leq c , \text{ où } \quad p_0 = \left[\sum_{n=0}^{c}\frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^{n}\right]^{-1} .$$

On a également $P(\textit{perte}) = \frac{1}{c!}\left(\frac{\lambda}{\mu}\right)^{c} p_0$.

Mesures de performance : On démontre que

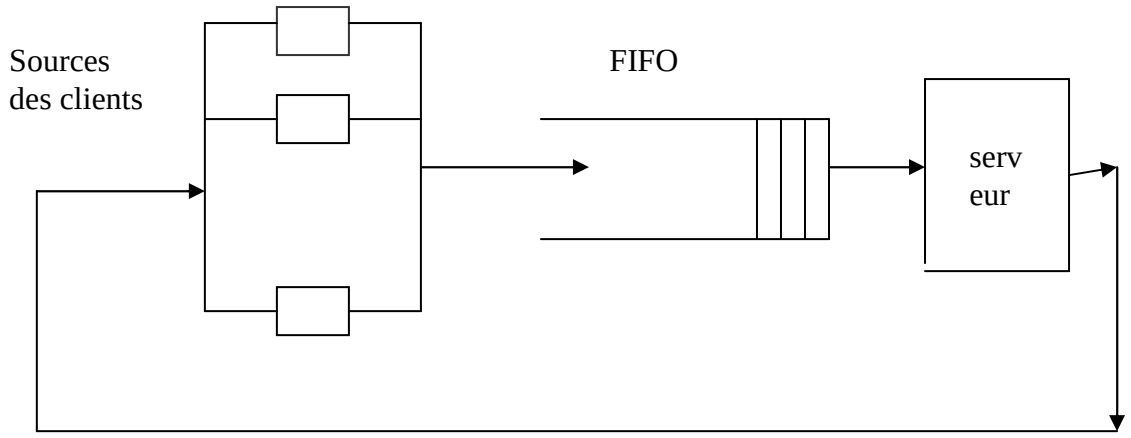
$$\bar{n}_f = \sum_{n=1}^{K-c} n p_{c+n} = \frac{\lambda}{\mu}\frac{(\lambda/\mu)^c}{c!c} p_0 \left[1 + 2\left(\frac{\lambda}{\mu c}\right) + 3\left(\frac{\lambda}{\mu c}\right)^2 + \ldots + (K-c)\left(\frac{\lambda}{\mu c}\right)^{K-c-1}\right]$$

$$= \frac{(\lambda/\mu)^{c+1}}{(c-1)!}\frac{1-\left(\frac{\lambda}{\mu c}\right)^{K-c}\left[1+(K-c)(1-\frac{\lambda}{\mu c}\right]}{\left(c-\frac{\lambda}{\mu}\right)^2} p_0 \ ;$$

L'application des relations de Little fournit d'autres mesures de performance

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

3.4 Système de files d'attente fréquenté par un nombre fini de clients



Une source peut être libre ou occupée (a engendré une demande de service).
Supposons que le flux d'entrée est quasi aléatoire :

- la probabilité pour qu'une source particulière engendre une demande de service durant l'intervalle $(t, t+\Delta t)$ sachant que cette source est libre à la date t est $\alpha\Delta t + 0(\Delta t)$;
- toutes les sources se comportent indépendamment les unes des autres.

Par conséquent, si une source est libre à la date t, la distribution de l'intervalle entre l'instant t et la date où cette source engendre la demande est exponentielle de paramètre $\alpha > 0$. Les durées de service suivent également une loi exponentielle de paramètre $\mu > 0$ (la durée moyenne de service est donc $1/\mu$). La discipline d'attente est FIFO. Ainsi, le processus $\{N(t), t \geq 0\}$ est un processus de naissance et de mort dont les taux de transition sont : $\lambda_n = \begin{cases} (K-n)\alpha & 0 \leq n \leq K \\ 0 & n > K \end{cases}$ (ici n est le nombre de clients déjà présents dans le système) ; $\mu_n = \mu, 1 \leq n \leq K$. L'espace des états du processus en question est $S = \{0,1,2,...,K\}$.

Le graphe des transitions est :

Régime stationnaire

Soient $p_n = \lim_{t\to\infty} p_n(t), 0 \leq n \leq K$. La distribution stationnaire p_n satisfait le système d'équations de balance suivant

$0 = -p_0\alpha K + p_1\mu$;

$0 = (K-n+1)\alpha p_{n-1} - [(K-n)\alpha + \mu]p_n + \mu p_{n+1}, \quad 1 \leq n < K$;

$0 = \alpha p_{K-1} - \mu p_K$;

$\sum_{n=0}^{K} p_n = 1$.

La résolution du système ci-dessus s'effectue de la manière suivante :

$p_1 = \frac{\alpha K}{\mu} p_0$;

pour $n = 1$: $\alpha K p_0 + \mu p_2 = (\alpha(K-1) + \mu)p_1$,

$$p_2 = \frac{\alpha(K-1)+\mu}{\mu} p_1 - \frac{\alpha K}{\mu} p_0 = \frac{\alpha(K-1)+\mu}{\mu}\frac{\alpha K}{\mu} p_0 - \frac{\alpha K}{\mu} p_0 = \frac{\alpha^2}{\mu^2}[K(K-1)]p_0 = \left(\frac{\alpha}{\mu}\right)^2 \frac{K!}{(K-2)!} p_0$$

pour $n = 2$: $\alpha(K-1)p_1 + \mu p_3 = [\alpha(K-2)+\mu]p_2$,

$$p_3 = \left[\frac{(\alpha(K-2)+\mu)\alpha^2 K(K-1)}{\mu^3} - \frac{\alpha(K-1)\alpha K}{\mu^2}\right] p_0 = \left(\frac{\alpha}{\mu}\right)^3 K(K-1)(K-2)p_0 = \left(\frac{\alpha}{\mu}\right)^3 \frac{K!}{(K-3)!} p_0$$

--

Finalement

$$p_n(K) = \frac{K!}{(K-n)!}\left(\frac{\alpha}{\mu}\right)^n p_0(K), \quad 0 < n \leq K \ ; p_0(K) = \left[\sum_{n=0}^{K} \frac{K!}{(K-n)!}\left(\frac{\alpha}{\mu}\right)^n\right]^{-1}.$$

Notons que l'intensité du trafic s'obtient en fonction de la taille de la population des clients potentiels $\rho(K) = 1 - p_0(K)$.

Chapitre 4

Modèles semi markoviens

4.1 Système de files d'attente M/G/1

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.

Description du modèle : Les clients arrivent dans le système selon un processus de Poisson ($\lambda>0$). Le service est assuré par un seul serveur. Les durées de service sont des variables aléatoires Se positives mutuellement indépendantes et distribuées selon une loi générale de fonction de répartition $B(x)$, de moyenne finie $E[Se] = \frac{1}{\mu}$ et de $E[Se^2]$. Les durées entre deux arrivées consécutives et les durées de service sont également mutuellement indépendantes.

Analyse du modèle : Soit $\{N(t), t \geq 0\}$. Montrons que $\{N(t)\}_t$ ne définit pas une chaîne de Markov. Soient t_d et t_f les dates de début et de fin d'un service, t_a l'instant d'arrivée d'un nouveau client. Si $t_d < t_a < t_f$, la probabilité qu'un départ s'effectue dans $]t_a, t_a + \Delta t]$ ne dépend pas seulement de Δt, mais de la date t_d à la quelle le service en cours a commencé. Comme le temps résiduel du service $(t_f - t_a)$ dépend du passé, alors la chaîne $\{N(t)\}_t$ n'est pas markovienne. Par conséquent, on utilise la méthode de ***la chaîne de Markov induite***. A cet effet, on considère $N(t)$ aux instants $\xi_1, \xi_2, ..., \xi_n, ...$ où les clients terminent leur service et quittent le système. On définit ainsi un processus stochastique à temps discret

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

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 *n-ème* client est servi. Les variables A_n sont indépendantes entre elles, leur distribution commune est

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

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

L'équation fondamentale de la chaîne vaut donc

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

où $\delta_n = \begin{cases} 1 & N_n > 0 \\ 0 & N_n = 0 \end{cases}$.

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},.... La suite $\{N_n, n \geq 1\}$ est une chaîne de Markov induite du processus $\{N(t), t \geq 0\}$. Ses probabilités de transition $p_{ij} = P(N_{n+1} = j / N_n = i)$ se calcule par

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

La matrice des transitions est

$$M = \begin{pmatrix} a_0 & a_1 & a_2 & a_3 & -- & -- & -- \\ a_0 & a_1 & a_2 & a_3 & -- & -- & -- \\ 0 & a_0 & a_1 & a_2 & -- & -- & -- \\ 0 & 0 & a_0 & a_1 & -- & -- & -- \\ -- & -- & -- & -- & -- & -- & -- \end{pmatrix}.$$

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

$$= \int_0^{\infty} e^{-\lambda t z} e^{\lambda t z} dB(t) = \int_0^{\infty} e^{-(\lambda - \lambda z)t} dB(t).$$

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

1) $|z| < 1 \quad 0 < a_k < 1 \quad \forall k$, on a $|a_k z^k| < |z^k|$;

2) $|z| = 1 \quad A(1) = 1$.

Remarques

1. Théorème des probabilités totales :

   Cas discret $P(A) = \sum_k P\left(A / Y = y_k \right) P(Y = y_k)$;

   Cas continu $P(A) = \int P\left(A / Y = y \right) g(y) dy$.

2. Probabilité que le nombre d'événements N qui ont lieu pendant un intervalle $U = u$ dont la densité de probabilité $f(u)$ est connue, est égal à n :

   $P\left(N = n / U = u\right) = e^{-\lambda u} \frac{(\lambda u)^n}{n!}$. D'où

   $$P(N = n) = \int_0^{\infty} P\left(N = n / U = u\right) f(u) du = \frac{1}{n!} \int_0^{\infty} e^{-\lambda u} (\lambda u)^n f(u) du .$$

$E[N] = \lambda E[U]$; $Var[N] = \lambda^2 Var[U] + \lambda E[U]$.

Supposons que $\rho < 1$. Le système se trouve dans un régime stationnaire. Soit $\Pi = [\pi_0, \pi_1, ...]$ la distribution stationnaire de la chaîne de Markov induite ($\pi_j = \lim_{n\to\infty} P(N(\xi_n) = j)$). Par conséquent, $\Pi = \Pi \cdot M$, ou $\pi_j = \sum_{i=0}^{\infty} \pi_i p_{ij} \quad j \geq 0$.

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

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 :

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

Finalement, on obtient

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

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

On a que $\Pi(1) = 1$. Cependant, $\Pi(1) = \lim_{z\to 1} \Pi(z) = {}^{0}\!/_{0}$. 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 :

$$\Pi(z) = \frac{(1-\rho)A(z)(z-1)}{z - A(z)} = \frac{(1-\rho)\tilde{B}(\lambda - \lambda z)(z-1)}{z - \tilde{B}(\lambda - \lambda z)} \quad . \qquad (4.3)$$

La condition d'existence d'un régime stationnaire est $\rho = \frac{\lambda}{\mu} < 1$.

Considérons les probabilités suivantes :

$p_j = \lim_{t\to\infty} P(N(t) = j), \quad j \geq 0$;

$\pi_j = \lim_{n\to\infty} P(N(\xi_n) = j), \quad j \geq 0$;

$r_j = \lim_{n\to\infty} P(N(\varsigma_n) = j), \quad j \geq 0$, ς_n est l'instant d'arrivée de *n-ème* client.

Vu que le processus des arrivées est poissonien, et le nombre *N(t)* subit des changements discontinus de taille 1 (± 1), on obtient $p_j = r_j = \pi_j$. Comme suite logique, la distribution stationnaire du processus à temps continu $\{N(t), t \geq 0\}$ est identique à celle de la chaîne de Markov induite. Par conséquent $Q(z) = \sum_{j=0}^{\infty} p_j z^j = \Pi(z)$.

Caractéristiques de performance

Formule de Pollaczek-Khintchine pour le nombre moyen de clients dans le système :
Considérons l'équation fondamentale (4.2). Vu que $\delta_n^{\,2} = \delta_n$ et $\delta_n N_n = N_n$, on trouve

$$N_{n+1}^2 = N_n^2 + \delta_n + A_{n+1}^2 - 2N_n - 2\delta_n A_{n+1} + 2N_n A_{n+1}.$$

On a que : A_{n+1} est indépendante de N_n et de δ_n ; $E[N_{n+1}^2] = E[N_n^2]$; $E[A_n] = \rho = \frac{\lambda}{\mu}$.

Alors, $E[N_{n+1}^2] = E[N_n^2] + E[\delta_n] + E[A^2{}_{n+1}] - 2E[N_n] + 2E[A_{n+1}]E[N_n - \delta_n]$, ou bien $0 = \rho + E[A_{n+1}^2] - 2E[N_n] + 2\rho\left[E[N_n] - \rho\right]$. D'où

$$E[N_n] = \frac{\rho + E[A_{n+1}^2] - 2\rho^2}{2(1-\rho)}. \qquad (4.4)$$

Pour trouver $E[A_{n+1}^2]$, considérons le régime stationnaire.

$$\lim_{n\to\infty} E[A_{n+1}^2] = E[A^2] = \int_0^\infty E\left[A^2 \middle/ T = t\right] dB(t) = \lambda \int_0^\infty t\, dB(t) + \lambda^2 \int_0^\infty t^2\, dB(t)$$

$$= \frac{\lambda}{\mu} + \lambda^2 \left(Var[Se] + \left(1 \middle/ \mu \right)^2 \right).$$

Enfin, la formule (4.4) devient

$$\lim_{n\to\infty} E[N_n] = E[N] = \rho + \frac{\rho^2 + \lambda^2 Var[Se]}{2(1-\rho)}. \qquad (4.5)$$

Le nombre moyen de clients dans le système peut être également trouvé à partir de la fonction génératrice $\Pi(z)$: $E[N] = \bar{n} = \lim_{z\to 1} \Pi'(z)$. Ici, le calcul de la limite donne une indétermination. Par conséquent, il est nécessaire d'appliquer la règle de l'Hôpital deux fois.

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* ($t \to \infty$), on a
$t \approx n\left[E[U] + E[V]\right]$. 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

Modèle M/E$_k$/1 : Dans ce système, la durée de service suit une loi d'Erlang d'ordre *k* et de moyenne finie *1/μ*. Les fonctions de densité de probabilités et de répartition sont données par :

$$b(t) = \frac{k\mu(k\mu t)^{k-1}}{(k-1)!} e^{-k\mu t} \text{ et } \quad B(t) = 1 - e^{-k\mu t} \sum_{j=0}^{k-1} \frac{(k\mu t)^j}{j!}, t \geq 0.$$

On démontre que $A(z) = \tilde{B}(\lambda - \lambda z) = \left[1 + \frac{\rho(1-z)}{k}\right]^{-k}$. Alors, l'équation (4.3) devient

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

Modèle M/H$_2$/1 : La durée de service suit une loi hyperexponentielle d'ordre 2 dont la fonction de répartition est donnée par $B(t) = 1 - p_1 e^{-\mu_1 t} - p_2 e^{-\mu_2 t}$, où p_1, p_2, μ_1, 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 $A(z) = \tilde{B}(\lambda - \lambda z) = \frac{p_1}{1 + \rho_1(1-z)} + \frac{p_2}{1 + \rho_2(1-z)}$. Par conséquent, l'équation (4.3) devient

$$\Pi(z) = \frac{(1-\rho)[1 + (\rho_1 + \rho_2 - \rho)(1-z)]}{\rho_1\rho_2 z^2 - (\rho_1 + \rho_2 + \rho_1\rho_2)z + 1 + \rho_1 + \rho_2 - \rho}$$, 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 $1/\mu$. 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)*) $\alpha = \int_0^\infty e^{-\mu t(1-\alpha)} a(t) dt$ 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 : $\bar{n} = \frac{\rho}{1-\alpha}$; $\bar{n}_f = \frac{\rho\alpha}{1-\alpha}$; $\bar{W}_s = \frac{1}{\mu(1-\alpha)}$; $\bar{W} = \frac{\alpha}{\mu(1-\alpha)}$.

Chapitre 5

Systèmes de files d'attente avec rappels

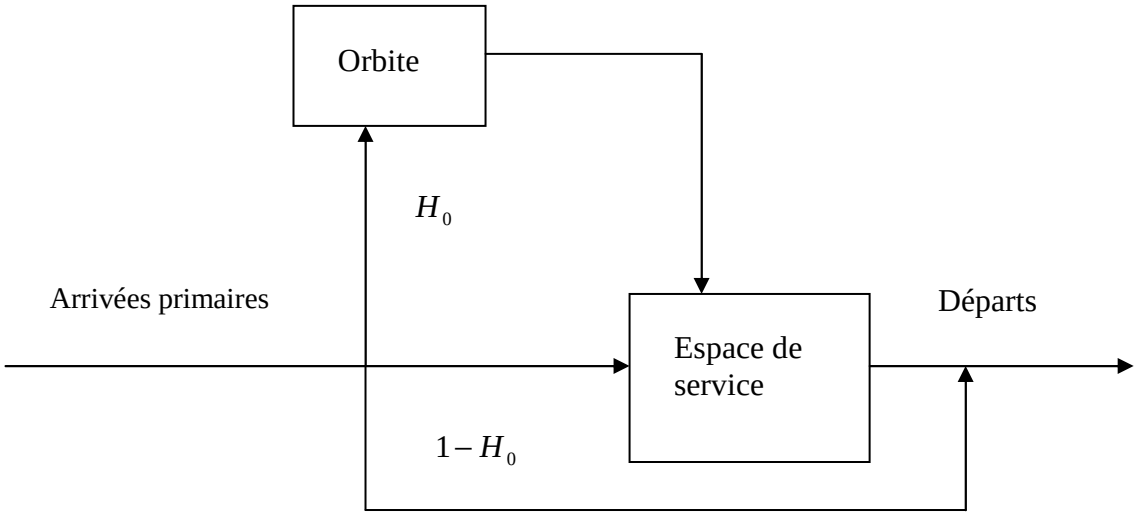
5.1 Introduction

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 et la discipline d'attente ainsi que 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 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.



5.2 Description du modèle de type M/G/1

On considère une population (des clients potentiels) très importante, afin d'avoir un flux d'entrée poissonien. En effet, les clients primaires arrivent dans le système selon un processus homogène de Poisson de taux $\lambda > 0$. La durée de temps entre deux arrivées primaires consécutives suit une loi exponentielle de fonction de répartition $A(t) = 1 - e^{-\lambda t}, t \geq 0$. Le service des clients est assuré par un 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 source de tentatives répétés (devient source de clients secondaires). Les durées de service suivent une loi générale commune de fonction de répartition *B(t),* de transformée de Laplace-Stieltjes $\tilde{B}(s)$, $\mathrm{Re}(s) > 0$, et de moyenne finie $1/\gamma$. Soient les moments $\beta_k = (-1)^k \tilde{B}^{(k)}(0)$, le taux de service $\gamma = \frac{1}{\beta_1}$ et l'intensité du trafic $\rho = \lambda\beta_1$ 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*) ($T(t) = 1 - e^{-\theta t}$) de moyenne finie $1/\theta$. Les trois variables aléatoires introduites sont supposées mutuellement indépendantes.

Le système évolue de la manière suivante : On suppose que le *(n-1)-ème* client termine son service à l'instant ξ_{n-1} (les clients sont numérotés dans l'ordre de service) et le serveur devient libre. Même s'il y a des clients dans le système, ils ne peuvent pas occuper le dispositif de service immédiatement à cause de leur ignorance de l'état de ce dernier. C'est pourquoi le *n-ème* client suivant n'entre en service qu'après un intervalle de temps R_n durant lequel le serveur est libre. A l'instant $\eta_n = \xi_{n-1} + R_n$ le *n-ème* client débute son service qui durera un temps $t_{service}$. Les clients primaires arrivant dans le système pendant ce temps deviennent sources de clients secondaires. Tous les rappels de l'orbite qui arrivent durant ce temps de service n'influent pas sur le processus. A l'instant $\xi_n = \eta_n + t_{service}$ le *n-ème* client achève son service, le serveur devient libre et ainsi de suite.

5.3 Modèle M/M/1 avec rappels

Soit $B(t)=1-e^{-\gamma t}, t\geq 0$. Supposons que les durées inter-rappels suivent une loi exponentielle : $T(t)=1-e^{-\theta t}, t\geq 0$. L'état du système à la date *t* peut être décrit par le processus stochastique suivant :

$$\{C(t), N_o(t), t\geq 0\}\ ; \tag{1}$$

où *C(t)* est 1 ou 0 selon le fait que le serveur est occupé ou non, $N_o(t)$ est le nombre de clients en orbite à la date *t*. Il s'agit d'un processus de Markov. Supposons que le régime stationnaire existe, c'est-à-dire $\rho=\frac{\lambda}{\gamma}<1$.

Théorème 1: *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* $p_{in}=\lim_{t\to\infty} P(C(t)=i, N_o(t)=n)$, $i=0,1$ *et* $n\geq 0$, *est donnée par*

$$p_{0n}=\frac{\rho^n}{n!\theta^n}\prod_{k=0}^{n-1}(\lambda+k\theta)(1-\rho)^{1+\frac{\lambda}{\theta}}\ ; \tag{2}$$

$$p_{1n}=\frac{\rho^{n+1}}{n!\theta^n}\prod_{k=1}^{n}(\lambda+k\theta)(1-\rho)^{1+\frac{\lambda}{\theta}}\ . \tag{3}$$

Les fonctions génératrices partielles correspondantes sont données par

$$P_0(z)=\sum_{n=0}^{\infty} z^n p_{0n}=(1-\rho)\left(\frac{1-\rho}{1-\rho z}\right)^{\frac{\lambda}{\theta}}\ ; \tag{4}$$

$$P_1(z)=\sum_{n=0}^{\infty} z^n p_{1n}=\rho\left(\frac{1-\rho}{1-\rho z}\right)^{\frac{\lambda}{\theta}+1}\ . \tag{5}$$

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

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

Les équations d'équilibre statistique sont

$$(\lambda+n\theta)p_{0n}=\gamma p_{1n}\ ; \tag{6}$$

$$(\lambda+\gamma)p_{1n}=\lambda p_{0n}+(n+1)\theta p_{0,n+1}+\lambda p_{1,n-1}\ . \tag{7}$$

A l'aide de fonctions génératrices, telles que $P_0(z)=\sum_{n=0}^{\infty} z^n p_{0n}$ et $P_1(z)=\sum_{n=0}^{\infty} z^n p_{1n}$, les équations (6)-(7) deviennent

$$\lambda P_0(z)+\theta z P_0'(z)=\gamma P_1(z)\ ; \tag{8}$$

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

D'où $P_0'(z)=\frac{\lambda\rho}{\theta(1-\rho z)}P_0(z)$. La solution de cette dernière équation est

$$P_0(z)=const\times(1-\rho z)^{-\frac{\lambda}{\theta}}\ . \tag{9}$$

Des équations (8), on a

$$P_1(z) = \rho P_0(z) + \frac{\theta z}{\gamma} P_0'(z) = P_0(z)\frac{\rho}{1-\rho z} = \frac{\rho \times const}{(1-\rho z)^{\frac{\lambda}{\theta}+1}}. \quad (10)$$

Vu que $\sum_{n=0}^{\infty}(p_{0n} + p_{1n}) = P_0(1) + P_1(1) = 1$, on obtient $const = (1-\rho)^{\frac{\lambda}{\theta}+1}$. (11)

A partir des équations (9)-(11), on déduit les équations (4) et (5).

A présent, à l'aide de l'équation (6), on élimine p_{1n} de l'équation (7). De cette manière, on trouve

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

Ceci implique que $n\theta\gamma p_{0n} - \lambda(\lambda + (n-1)\theta)p_{0,n-1} = 0$. D'où

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

De l'équation (6), 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 $\sum_{n=0}^{\infty} p_{0n} + \sum_{n=1}^{\infty} p_{1n} = 1$,

$$p_{00} = \left[\sum_{n=0}^{\infty}\frac{\rho^n}{n!\theta^n}\prod_{k=0}^{n-1}(\lambda + k\theta) + \sum_{n=0}^{\infty}\frac{\rho^{n+1}}{n!\theta^n}\prod_{k=1}^{n}(\lambda + k\theta)\right]^{-1}. \quad (12)$$

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 (2) et (3).

Fin de preuve

Conséquences

1. La distribution stationnaire de processus (1) existe si $\rho = \frac{\lambda}{\gamma} < 1$.
2. La fonction génératrice de la distribution stationnaire marginale du nombre de clients en orbite $N_o = \lim_{t\to\infty} N_o(t)$ 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 $N = \lim_{t\to\infty}(N(t) = C(t) + N_o(t))$ 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

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

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

5.4 Modèle M/G/1 avec rappels

Chaîne de Markov induite : Considérons le processus $\{C(t), N_o(t), t \geq 0\}$, qui n'est pas un processus de Markov. On dénote par $q_n = N_o(\xi_n)$ le nombre de clients en orbite après le *n*-ème départ (ξ_n est l'instant de départ du *n*-ème client). La suite de variables aléatoires q_n forme une chaîne de Markov induite, dont l'équation fondamentale est :

$$q_{n+1} = q_n - \delta_n + A_{n+1} .$$

La variable aléatoire A_{n+1} représente le nombre de clients primaires arrivant dans le système durant le service du (n+1)-ème client. Elle ne dépend pas des événements qui se sont produits avant l'instant du début de service du (n+1)-ème client. Sa distribution est donnée par : $P(A_n = i) = a_i = \int_0^\infty \exp(-\lambda t)\frac{(\lambda t)^i}{i!} dB(t)$, où $a_i > 0, i \geq 0$; avec les résultats suivants : Si $A = \lim_{n\to\infty} A_n$, $E[A] = \rho$ et $A(z)\sum_{i=0}^{\infty} a_i z^i = \tilde{B}(\lambda - \lambda z)$.

La variable aléatoire δ_n est égale à 0 ou 1 selon le fait que le (n+1)-ème client servi est primaire ou provient de l'orbite. Elle dépend de q_n et sa distribution est donnée par :

$$P(\delta_n = 1 / q_n = k) = \frac{k\theta}{\lambda + k\theta} \; ; \; P(\delta_n = 0 / q_n = k) = \frac{\lambda}{\lambda + k\theta}.$$

Les probabilités de transition à une étape de la chaîne sont :

$$p_{ij} = \frac{i\theta}{\lambda + i\theta} a_{j-i+1} + \frac{\lambda}{\lambda + i\theta} a_{j-i} .$$

Théorème 2 *Soit* $\rho < 1$. *La distribution stationnaire* $\pi_k = \lim_{n\to\infty} P(q_n = k)$ *de la chaîne de Markov induite possède la fonction génératrice suivante* :

$$\varphi(z) = \sum_{k=0}^{\infty} \pi_k z^k = \frac{(1-\rho)(1-z)\tilde{B}(\lambda - \lambda z)}{\tilde{B}(\lambda - \lambda z) - z} \exp\left\{ \frac{\lambda}{\theta} \int_1^z \frac{1 - \tilde{B}(\lambda - \lambda u)}{\tilde{B}(\lambda - \lambda u) - u} du \right\}.$$

Distributions stationnaires de l'état du système : L'état du système peut être décrit par le processus $\{C(t), N_o(t), R(t), t \geq 0\}$, où *R(t)* est une variable supplémentaire à valeurs dans R^+ et désignant la durée de service écoulée à la date *t*. Notons
$p_{0n} = \lim_{t\to\infty} P(C(t) = 0, N_o(t) = n)$ et $p_{1n} = \lim_{t\to\infty} P(C(t) = 1, N_o(t) = n)$.

Théorème 3 *Les fonctions génératrices de la distribution conjointe de l'état du serveur et de la taille de l'orbite en régime stationnaire ($\rho < 1$) sont données par :*

$$P_0(z) = \sum_{n=0}^{\infty} p_{0n} z^n = (1-\rho)\exp\left\{ \frac{\lambda}{\theta} \int_1^z \frac{1 - \tilde{B}(\lambda - \lambda u)}{\tilde{B}(\lambda - \lambda u) - u} du \right\} ;$$

$$P_1(z) = \sum_{n=0}^{\infty} p_{1n} z^n = \frac{1 - \tilde{B}(\lambda - \lambda z)}{\tilde{B}(\lambda - \lambda z) - z} P_0(z).$$

Conséquences

1. La distribution marginale du nombre de serveurs occupés s'exprime de la manière suivante :

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

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

$$P(z) = P_0(z) + P_1(z) = \frac{1-z}{\tilde{B}(\lambda-\lambda z)-z} P_0(z) \, .$$

3. La fonction génératrice du nombre de clients dans le système se présente comme :

$$Q(z) = P_0(z) + zP_1(z) = \frac{(1-\rho)(1-z)\tilde{B}(\lambda-\lambda z)}{\tilde{B}(\lambda-\lambda z)-z} \exp\left\{\frac{\lambda}{\theta}\int_1^z \frac{1-\tilde{B}(\lambda-\lambda u)}{\tilde{B}(\lambda-\lambda u)-u} du\right\} .$$

Mesures de performance

1. Nombre moyen de clients dans le système

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

2. Nombre moyen de clients en orbite

$$\bar{n}_o = E[N_o] = \bar{n} - \rho = P'(1) = \frac{\lambda^2 \beta_2}{2(1-\rho)} + \frac{\lambda\rho}{\theta(1-\rho)} \ ;$$

3. Temps moyen d'attente d'un client

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

4. Nombre moyen de rappels par client

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

Dans le cas d'un service exponentiel, $\beta_1 = E[t_{service}] = \frac{1}{\gamma}$ et $\beta_2 = E[t_{service}^2] = \frac{2}{\gamma^2}$.

5.5 Modèle M/M/c avec rappels

Nous considérons un système de files d'attente avec rappels où l'espace de service comprend $c > 1$ serveurs. Les clients primaires arrivent selon un processus de Poisson de taux $\lambda > 0$. Si un client primaire arrivant trouve au moins un serveur libre, il commence son service. Sinon, il entre en orbite. Nous admettons que la durée de service et la durée entre deux rappels consécutives sont exponentiellement distribuées de moyennes finies, respectivement, $\frac{1}{\gamma}$ et $\frac{1}{\theta}$.

Le processus stochastique $\{C(t), N_o(t), t \geq 0\}$ est celui de Markov, d'espace d'états $S = \{0,1,...,c\} \times N$. Les probabilités d'état sont

$$p_{ij}(t) = P(C(t) = i, N_o(t) = j),\ (i, j) \in S\ .$$

Les taux de transition en régime stationnaire sont données par

- pour $0 \le i \le c-1$

$$q_{ij}(k,l) = \begin{cases} \lambda & si \quad (k,l) = (i+1, l) \\ i\gamma & si \quad (k,l) = (i-1, j) \\ j\theta & si \quad (k,l) = (i+1, j-1) \end{cases};$$

$$q_{ij}(k,l) = \begin{cases} -(\lambda + i\gamma + j\theta) & si \quad (k,l) = (i, j) \\ \\ 0 \qquad\qquad si \quad ailleurs \end{cases};$$

- pour $i=c$

$$q_{cj}(k,l) = \begin{cases} \lambda & si \quad (k,l) = (c, j+1) \\ c\gamma & si \quad (k,l) = (c-1, j) \\ -(\lambda + c\gamma) & si \quad (k,l) = (c, j) \\ 0 & si \quad ailleurs \end{cases}.$$

La condition d'existence d'un régime stationnaire est $\rho = \dfrac{\lambda}{c\gamma} < 1$. Les caractéristiques importantes de la qualité de service sont :

- *probabilité que tous les serveurs sont occupés* $P_c = \lim_{t\to\infty} P(C(t) = c)$;
- *nombre moyen de clients en orbite* $\overline{n}_o$;
- *nombre moyen de serveurs occupés* $\overline{c} = \lim_{t\to\infty} E[C(t)]$.

La résolution des modèles à multiserveurs fait appel aux approches basées sur le principe de troncation de l'espace d'états ainsi que aux méthodes numériques de résolution des systèmes d'équation algébriques.

5.6 Quelques applications

L'étude des modèles avec rappels est motivée par l'avènement de nouvelles technologies, notamment de l'Internet ou des réseaux modernes de télécommunications (réseaux à commutation par paquets, réseaux locaux de type bus à conflit d'accès ou réseaux ATM).

Voici quelques exemples de problèmes qui peuvent être modélisés comme les systèmes de files d'attente avec rappels.

Réseaux de commutation par paquet

Considérons un réseau de communications d'ordinateurs dans lequel on trouve un ensemble d'interfaces IMPs (Interface Message Processors) reliées entre elles par des câbles**.** Un ordinateur principal est connecté à l'une de ces interfaces. Si l'ordinateur veut envoyer le message à un autre ordinateur principal, il doit en premier lieu envoyer le message avec l'adresse de destination à l'interface à laquelle il est connecté. L'interface à son tour envoie le message à l'ordinateur destinataire directement si elle y est connectée, ou indirectement via d'autres interfaces.

Considérons une interface à laquelle un ordinateur principal est connecté. Les messages arrivent de l'extérieur selon un processus aléatoire. Après la réception du message, l'ordinateur l'envoie immédiatement à son interface. S'il y a un tampon libre au niveau de l'interface, le message est accepté par cette dernière. Dans le cas contraire, le message est rejeté et l'ordinateur doit réessayer une autre fois après une période de temps. Le message rejeté par l'interface sera stocké dans un tampon de l'ordinateur principal (pour réaliser une autre tentative). Dans le cas contraire, le message est considéré comme perdu. On peut poser les questions suivantes :

1. Quelles sont les probabilités pour qu'un message soit rejeté par l'interface et par l'ordinateur principal ?
2. Quel est le nombre moyen de messages dans les tampons de IMP ?
3. Quel est le nombre moyen de messages dans les tampons de l'ordinateur principal ?
4. Quel est le temps moyen d'attente d'un message dans le tampon de l'ordinateur principal ?

Le problème présenté peut être modélisé comme un système avec rappels à serveur unique (interface IMP) possédant des tampons (positions d'attente). Les tampons de l'ordinateur principal constituent l'orbite et leur nombre est la capacité de l'orbite. Les tampons de l'interface présente une file d'attente classique. Donc il s'agit d'un modèle avec rappels, espace d'attente limité et orbite de capacité finie. En général, le nombre de tampons de l'ordinateur principal est largement supérieur au nombre de tampons de IMP. Donc on peut considéré l'orbite de capacité illimitée.

Description du modèle : Supposons que les clients (messages) arrivent dans le système selon un processus de Poisson de taux $\lambda > 0$. Le service est assuré par un seul serveur. L'espace de service comprend également *m*-1 positions d'attente. Supposons que les durées de service suivent une loi exponentielle de fonction de répartition $B(t) = 1 - e^{-\gamma t}, t \geq 0$, de moyenne finie $1/\gamma$.L'intensité du trafic est $\rho = \frac{\lambda}{\gamma}$. La durée entre deux rappels consécutifs d'une même source secondaire est exponentiellement distribuée de paramètre $\theta > 0$: $T(t) = 1 - e^{-\theta t}, t \geq 0$. La capacité de l'orbite est *K* (un nombre assez grand). Les temps inter rappels, les temps de service ainsi que les temps entre deux arrivées primaires successives sont supposés mutuellement indépendants. Par conséquent, nous avons un système de files d'attente M/M/1/m/K avec rappels.

Analyse du modèle : L'état du système à la date *t* peut être décrit par le processus stochastique de Markov $\{C(t), N_0(t), t \geq 0\}$, où $C(t)$ est le nombre de clients dans l'espace de service (en attente et dans le serveur), $N_o(t)$ est le nombre de clients en orbite. L'espace d'états est donc $S = \{0,1,...,m\} \times \{0,1,...,K\}$. Vu que *S* est fini, le régime stationnaire du processus introduit existe toujours. Soit $\rho = \frac{\lambda}{\gamma} < 1$. La distribution stationnaire conjointe de l'état de l'espace de service et du nombre de clients en orbite $p_{ij} = \lim_{t \to \infty} P(C(t) = i, N_o(t) = j), (i, j) \in S$, vérifie le système d'équation d'équilibre statistique suivant

$(\lambda + l\gamma + j\theta) p_{ij} = \lambda p_{1-1,j} + (j+1)\theta p_{i-1,j+1} + \gamma p_{i+1,j}$, $0 \leq i \leq m-1, 0 \leq j \leq K-1$;

$(\lambda + l\gamma + K\theta) p_{iK} = \lambda p_{i-1,K} + \gamma p_{i+1,K}$, $0 \leq i \leq m-1, j = K$;

$$(\lambda + 1\gamma)p_{mj} = \lambda p_{m-1,j} + (j+1)\theta p_{m-1,j+1} + \lambda p_{m,j-1},\ \ i = m, 0 \le j \le K-1\ ;$$

$$\gamma p_{mK} = \lambda p_{m-1,K} + \lambda p_{m,K-1}\ ;\quad \text{où } l = \begin{cases} 0 & si \quad i = 0 \\ & \\ 1 & si \quad 1 \le i \le c \end{cases}.$$

L'équation de normalisation est $\sum_{i=0}^{m}\sum_{j=0}^{K} p_{ij} = 1$.

Mesures de performance : Les mesures de performance s'expriment en termes des probabilités p_{ij} comme suit :

- *Nombre moyen de clients en orbite :* $\bar{n}_o = \sum_{i=0}^{m}\sum_{j=0}^{K} jp_{ij}$;
- *Nombre moyen de clients en attente :* $\bar{n}_f = \sum_{i=0}^{m}\sum_{j=0}^{K} ip_{ij}$;
- *Probabilité de blocage :* $B = \lim_{t\to\infty} P(C(t) = m) = \sum_{j=0}^{K} p_{mj}$;
- *Probabilité de perte des clients :* p_{mK} ;
- *Temps moyen de séjour en orbite :* $\overline{W} = \frac{\bar{n}_o}{\lambda}$.

CSMA non persistant

Dans les réseaux locaux (LAN) d'ordinateurs, l'un des protocoles de communication les plus utilisés est CSMA (Carrier-Sence Multiple Acces) non persistant. Supposons qu'un réseau local est composé de *n* stations connectées par un seul bus. La communication entre les stations est réalisée au moyen de ce bus. Les messages de longueurs 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 ou non. Si le bus est libre, l'un des paquets est transmis via ce bus à la station de destination, et les autres paquets qui constituent le message sont stockés dans les tampons pour une transmission ultérieure. Autrement, tous les paquets sont stockés dans les tampons et la station peut consulter le bus après une certaine durée aléatoire. Les questions concernant ce problème sont :

1. Quel est le temps moyen d'attente d'un paquet ?
2. Quel est le nombre moyen de paquets dans le tampon d'une station ?

Si les messages arrivent selon un processus de Poisson, le système peut être modélisé comme un système $M^X/G/1$ avec rappels. Le serveur est le bus, et les tampons de la station représentent l'orbite. Si le nombre de tampons est suffisamment grand, on a un système de files d'attente avec rappels et capacité de l'orbite infinie.

Description du modèle : Les clients (paquets) arrivent dans le système par groupes (par messages), et selon un processus de Poisson de taux $\lambda > 0$. Le nombre de clients constituant un groupe est une variable aléatoire X. Supposons que $c_k = P(X = k)$. Le service est assuré par un serveur. Si à l'arrivée d'un groupe, le serveur est libre, l'un des clients du groupe en question est immédiatement pris en charge, et les autres entrent en orbite. Les temps de service suivent une loi générale de fonction de répartition $B(t)$, de transformée de Laplace-

Stieltjes $\tilde{B}(s), \mathrm{Re}(s) > 0$. Soient les moments initiaux $\beta_k = E\left[t_{service}^{k}\right] = (-1)^k \tilde{B}^{(k)}(0)$, le taux de service $\gamma = \frac{1}{\beta_1}$. La durée entre deux rappels consécutifs d'une même source secondaire est exponentiellement distribuée de paramètre $\theta > 0$. Soient $C(z) = \sum_{k=1}^{\infty} z^k c_k$ la fonction génératrice de la distribution stationnaire de la taille des groupes et $\bar{c} = E[X] = C'(1)$ la taille moyenne d'un groupe. L'intensité du trafic est $\rho = \lambda \beta_1 \bar{c}$. Le flux de clients primaires (les instants d'arrivée et les tailles des groupes), les durées entre les rappels consécutifs ainsi que les temps de service sont supposés mutuellement indépendants.

Analyse du modèle : On décrit l'état du système à la date *t* par le processus stochastique

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

Où *C*(*t*) est 0 ou 1 selon le fait que le serveur est libre ou occupé, $N_o(t)$ est le nombre de clients en orbite, $\zeta(t)$ est une variable aléatoire supplémentaire à valeurs dans R^+ et désignant la durée de service écoulée à la date *t*.

Le modèle peut être résolu (la distribution stationnaire du processus peut être trouvée), lorsque $\rho < 1$, à l'aide d'une approche basée sur la méthode des variables supplémentaires.

Mesures de performance

- *Nombre moyen de clients en orbite*

$$\bar{n}_o = \frac{\lambda^2 (C'(1))^2 \beta_2 + \rho C''(1)/C'(1)}{2(1-\rho)} + \frac{\lambda}{\theta} \times \frac{\rho + C'(1) - 1}{1-\rho} \ ;$$

- *Temps moyen d'attente*

$$\overline{W} = \frac{\bar{n}_o}{\lambda \bar{c}} = \frac{\lambda \bar{c} \beta_2}{2(1-\rho)} + \frac{\beta_1 (\bar{c}^2 - \bar{c})}{2(1-\rho)\bar{c}} + \frac{\rho + \bar{c} - 1}{\theta (1-\rho)\bar{c}} \ ;$$

- *Nombre moyen de rappels par client*

$$\overline{R} = \theta \overline{W} = \frac{\theta \lambda \bar{c} \beta_2}{2(1-\rho)} + \frac{\theta \beta_1 (\bar{c}^2 - \bar{c})}{2(1-\rho)\bar{c}} + \frac{\rho + \bar{c} - 1}{(1-\rho)\bar{c}}.$$

Conclusion

Conçue au début du siècle passé par Erlang pour modéliser les phénomènes de congestion dans les réseaux téléphoniques, la théorie des files d'attente connaît un regain d'intérêt depuis plusieurs années. Ce nouvel essor a été déclenché par l'avènement des ordinateurs actuels, de plus en plus complexes, dont il s'agit d'estimer et d'améliorer les performances, ainsi que par la création d'Internet dont il s'agit de concevoir, dimensionner et opérer les réseaux afin d'acheminer de manière efficace les énormes masses d'information concernées vers leurs destinataires. Au fil du temps, les modèles simples de files d'attente proposés par Erlang ont été généralisés pour les rendre plus réalistes et ont ainsi donné naissance à une riche théorie et à une méthodologie évoluée de traitement numérique.

Bibliographie

1. J. Artalejo and A. Gomez-Corral. *Retrial Queueing Systems: A Computational Approach.* Springer, 2008.
2. L. Breuer and D. Baum. *An Introduction to Queueing Theory and Matrix-Analytic Methods.* Springer, 2005.
3. D. Gross, J.F. Shortle, J.M. Thompson and C.M. Harris. *Fundamentals of Queueing Theory*. John Wiley and Sons, 2008.
4. G.I. Falin and J.G.C. Templeton. *Retrial Queues.* Chapman and Hall, London, 1997.
5. L. Kleinrock. *Queueing Systems, Volume 1: Theory.* John-Wiley and Sons, 1975.
6. P. Robert. *Réseaux et Files d'Attente : Méthodes Probabilistes.* Springer-Verlag Berlin Heidelberg, 2000.
7. S. Ross. *Stochastic Processes.* John-Wiley and Sons, New York, 2e éd., 1996.
8. T.L. Saaty. *Elements of Queueing Theory.* M.C. Graw-Hill, New York, 1961.

Complément 1 Chapitre 3

Systèmes particuliers de files d'attente (modèles markoviens)

C1 : 3.1 Système de files d'attente $M^X/M/1$

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 : $P(X(t)=x)=c_x(t)$. 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.

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 $S=\{0,1,2,...\}$.

Régime transitoire

Soient $p_n(t)=P(N(t)=n),\ n\geq 0$. Le graphe des transitions est figure7.

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

$p_0(t+\Delta t)=p_0(t)(1-\lambda\Delta t)+\mu\Delta t p_1(t)$;

$p_1(t+\Delta t)=\lambda\Delta t c_1(t)p_0(t)+[1-(\lambda+\mu)\Delta t]p_1(t)+\mu\Delta t p_2(t)$;

$p_2(t+\Delta t)=\lambda\Delta t c_1(t)p_1(t)+\lambda\Delta t c_2(t)p_0(t)+[1-(\lambda+\mu)\Delta t]p_2(t)+\mu\Delta t p_3(t)$;

$$p_n(t+\Delta t)=\lambda\Delta t\sum_{k=1}^{n}p_{n-k}(t)c_k(t)+[1-(\lambda+\mu)\Delta t]p_n(t)+\mu\Delta t p_{n+1}(t),\ n>2.$$

Alors le système d'équations de Kolmogorov est :

$p_0'(t)=-\lambda p_0(t)+\mu p_1(t)$;

$$p_n'(t)=\lambda\sum_{k=1}^{n}p_{n-k}(t)c_k(t)-(\lambda+\mu)p_n(t)+\mu p_{n+1}(t),\quad n\geq 1.$$

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

$\lambda p_0=\mu p_1$;

$$0=\lambda\sum_{k=1}^{n}p_{n-k}c_k-(\lambda+\mu)p_n+\mu p_{n+1},n\geq 1. \qquad \text{(C1 : 3.1)}$$

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 ») :

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

En multipliant les équations (C3.1) par z^n et en sommant l'ensemble, on obtient

$$\lambda \sum_{n=0}^{\infty} p_n z^n + \mu \sum_{n=1}^{\infty} p_n z^n = \frac{\mu}{z} \sum_{n=1}^{\infty} p_n z^n + \lambda \sum_{n=1}^{\infty} \sum_{k=1}^{\infty} p_{n-k} c_k z^n . \qquad \text{(C1 : 3.2)}$$

A l'aide des propriétés des fonctions génératrices, l'équation (C3.2) devient

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

D'où

$$P(z) = \frac{\mu p_0 (1-z)}{\mu(1-z) - \lambda z[1 - C(z)]}, \text{si} |z| < 1. \qquad \text{(C1 : 3.3)}$$

Pour obtenir p_0, il faut utiliser l'équation de normalisation $P(1) = 1$. La relation (3.3) pour $z \to 1$ (en appliquant la règle de l'Hôpital $0/0$) devient

$1 = \frac{-\mu p_0}{-\mu + \lambda E[X]}$. D'où, $p_0 = 1 - \frac{\lambda E[X]}{\mu}$. L'intensité du trafic $\rho = \frac{\lambda E[X]}{\mu}$.

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

Remarque : Convolution de deux distributions est définie par

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

C1 : 3.2 Système de files d'attente M/M^{Y}/1

A présent, supposons que les clients sont servis par groupes de taille *K* maximum, sauf si le nombre de clients présents dans la file d'attente ne permet pas d'atteindre cette taille. Dans ce dernier cas, tous les clients de la file sont servis ensemble. Supposons également que la distribution des temps de service ne dépend pas de la taille du groupe servi et qu'elle est exponentielle ayant la moyenne finie $1/\mu$. La discipline est toujours FIFO.

En régime stationnaire, les équations de balance sont :

$(\lambda + \mu) p_{n+1} = \lambda p_n + \mu p_{n+K+1}$;

$\lambda p_0 = \mu p_1 + ... + \mu p_{K-1} + \mu p_K$.

C1 : 3.3 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.

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 (λ>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 $(1-\alpha)$ ($\lambda = \lambda\alpha + \lambda(1-\alpha) = \lambda_1 + \lambda_2$). 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.

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 $S = \{(n_1, n_2), n_1 \in N, n_2 \in N\}$. Le graphe des transitions est (figure8).

En régime stationnaire, les équations de balance sont :

$(\lambda_1 + \lambda_2)p_{0,0} = \mu_1 p_{1,0} + \mu_2 p_{0,1}$, $n_1 = 0$ et $n_2 = 0$;

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

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

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

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

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

$$F(x,y) = \left(\frac{1-\rho_1-\rho_2}{1-\eta-y\rho_2}\right)\left(\frac{1-\eta}{1-\eta x}\right), \text{ où } \rho_1 = \frac{\lambda_1}{\mu_1},\ \rho_2 = \frac{\lambda_2}{\mu_2},$$

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

De la fonction génératrice, on déduit

$$\bar{n}_1 = \left.\frac{\partial F(x,y)}{\partial x}\right|_{x=y=1} = \frac{\rho_1}{1-\rho_1}\ ;\quad \bar{n}_2 = \left.\frac{\partial F(x,y)}{\partial y}\right|_{x=y=1} = \frac{\rho_2 + E[n_1]\rho_2}{1-\rho_1-\rho_2}.$$

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

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

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

Complément 2 Chapitre 3

C2 :3.1 Autres modèles

Système de files d'attente avec service variable

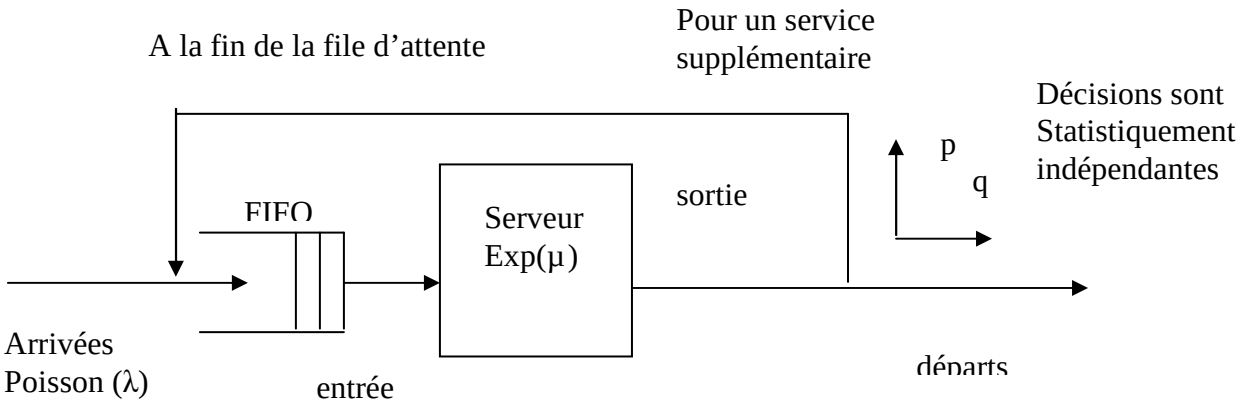
Considérons un système de files d'attente où le taux de service dépend du nombre de clients dans le système. Par exemple, un second serveur peut être ajouté lorsque la taille de la file d'attente dépasse une valeur critique L. Si les arrivées sont de Poisson et les temps de service sont sans mémoire, alors les taux de transition du processus de naissance et de mort sont

$\lambda_n = \lambda,\ n \geq 0\ ;\quad \mu_n = \begin{cases} \mu & n \leq L \\ 2\mu & n > L \end{cases}.$

De manière générale : le processus des arrivées est un processus de Poisson avec $\lambda_n = \lambda$, $n \geq 0$; les taux de service sont arbitraires μ_n, $n \geq 1$. La distribution stationnaire dans ce cas est

$$p_n = \frac{\lambda^n}{\mu_1 \mu_2 \ldots \mu_n} p_0,\ n \geq 1, \text{ où } p_0 = \left[1 + \sum_{k=1}^{\infty} \frac{\lambda^k}{\mu_1 \mu_2 \ldots \mu_k}\right]^{-1}.$$

Système de files d'attente avec feedback



Le processus $\{N(t), t \geq 0\}$ est un processus de naissance et de mort avec

$$\lambda_n = \lambda,\quad n \geq 0\ ;$$
$$\mu_n = q\mu,\quad n \geq 1,\quad q = 1 - p\ .$$

La distribution stationnaire $p_n = \lim_{t \to \infty} P(N(t) = n)$ existe si $\frac{\lambda}{q\mu} < 1$. Elle est donnée par

$$p_n = \left(1 - \frac{\lambda}{q\mu}\right)\left(\frac{\lambda}{q\mu}\right)^n,\quad n \geq 0\ .$$

Système de files d'attente avec découragement

Considérons le système de files d'attente M/M/c où le temps d'attente est limité à T_{at} (si avant d'expiration de T_{at} le client n'est pas servi, il quitte le système sans recevoir son

service). Supposons que T_{at} est exponentiellement distribué de paramètre v. Il faut noter que : si $v \to \infty$, on a un système à demandes refusées ; si $v \to 0$, on a un système à capacité illimitée . L'état du système à la date t est toujours décrit par le processus (2.1) ayant l'espace des états $S = \{0,1,2,...,c,c+1,...,c+i\}$, $i > 1$.

Le graphe des transitions est figure5.

En régime stationnaire, on a des équations de balance suivantes :

$0 = -\lambda p_0 + \mu p_1$;

$0 = \lambda p_{n-1} - (\lambda + n\mu) p_n + (n+1)\mu p_{n+1}, \quad 1 \le n < c$;

$0 = \lambda p_{c-1} - (\lambda + c\mu) p_c + (c\mu + v) p_{c+1}, \quad n = c$;

$0 = \lambda p_{c+i-1} - (\lambda + c\mu + iv) p_{c+i} + (c\mu + (i+1)v) p_{c+i+1}, \quad i \ge 1 \ (n > c)$;

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

La résolution de ce système fournit la distribution stationnaire $p_n = \lim_{t\to\infty} P(N(t) = n)$. En effet,

$$p_n = \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n p_0, \quad 1 \le n \le c \ ; \quad p_{c=i} = \frac{\lambda^{c+i}}{c!\mu^c \prod_{n=1}^{i}(c\mu + nv)} p_0 \ ;$$

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

Les mesures de performance de ce modèle sont :

-le nombre moyen de clients dans la file d'attente

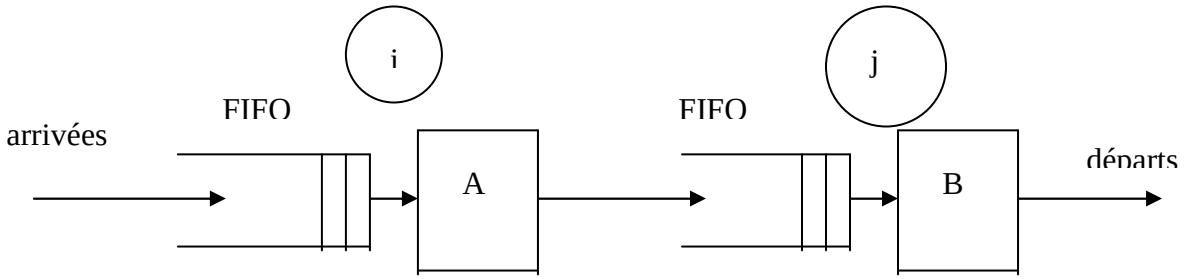
$$\bar{n}_f = \sum_{i=1}^{\infty} i p_{c+i} = \frac{\frac{(\lambda/\mu)^c}{c!} \sum_{r=1}^{\infty} \frac{r(\lambda/\mu)^r}{\prod_{n=1}^{r}(c\mu + nv)}}{\sum_{k=0}^{c} \frac{(\lambda/\mu)^k}{k!} + \frac{(\lambda/\mu)^c}{c!} \sum_{r=1}^{\infty} \frac{(\lambda/\mu)^r}{\prod_{n=1}^{r}(c\mu + nv)}} \ ;$$

-la probabilité pour qu'un client parte sans être servi

$$P_{dep} = \frac{v\bar{n}_f}{\lambda} .$$

Système de files d'attente en cascade

On considère le système de files d'attente suivant : Le client arrive selon un processus de Poisson (λ>0), et se présente à un serveur A, puis à un serveur B. Les temps de service en A et en B sont deux variables aléatoires indépendantes, distribuées selon une loi exponentielle de moyennes $\frac{1}{\mu}$ et $\frac{1}{v}$ respectivement. Les disciplines de service sont FIFO pour les deux sous systèmes.



L'état du système à la date t est décrit par un processus stochastique $\{N_A(t), N_B(t), t \geq 0\}$, où $N_A(t)$ ($N_B(t)$) est le nombre de clients dans le sous système A (sous système B) à la date t. Ce processus est une chaîne de Markov homogène avec $S = \{(i, j)\}$, $i \in N$ et $j \in N$. Legraphe des transitions est figure6.

Théorème

Soit $\{X(t), t \geq 0\}$ le processus de naissance et de mort avec $\lambda_n = \lambda$, $n \geq 0$, et μ_n, $n \geq 1$, quelconque. Supposons que la distribution stationnaire $p_n = \lim_{t\to\infty} P(X(t) = n)$ existe. Soit $D(t)$ le nombre de morts durant $[0, t]$. Alors

$$P(X(t) = n, D(t) = j) = P(X(t) = n)P(D(t) = j) = p_n \frac{(\lambda t)^j}{j!} e^{-\lambda t}, \quad n, j \geq 0.$$

Remarque : Les départs du système de files d'attente avec les arrivées poissoniennes et le service exponentiel forment un processus de Poisson.

Pour le sous système A (M/M/1) :

$$\lim_{t\to\infty} P(N_A(t) = n) = p_n = \left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^n, \; n \geq 0.$$

Pour le sous système B qui est statistiquement indépendant de A et est également M/M/1 :

$$\lim_{t\to\infty} P(N_B(t) = m) = p_m = \left(1 - \frac{\lambda}{\nu}\right)\left(\frac{\lambda}{\nu}\right)^m, \; n \geq 0 \; .$$

Par conséquent,

$$p_{nm} = \lim_{t\to\infty} P(N_A(t) = n, N_B(t) = m) = p_n q_m = (1 - \rho_A)\rho_A^n (1 - \rho_b)\rho_B^m \; .$$

La probabilité de trouver k clients dans le système :

$$\sum_{n+m=k} p_{nm} = \sum_{n=0}^{k} p_n q_{k-n} = \left(1 - \frac{\lambda}{\mu}\right)\left(1 - \frac{\lambda}{\nu}\right)\lambda^k \sum_{n=0}^{k} \frac{1}{\mu^i \nu^{k-i}}.$$

En ce qui concerne les mesures de performance,

$$\bar{n}_A = \frac{\rho_A}{1 - \rho_A} \; ; \quad \bar{n}_B = \frac{\rho_B}{1 - \rho_B} \; ; \quad \bar{n} = \bar{n}_A + \bar{n}_B \; ;$$

de même pour les autres mesures.

Remarque : Si le nombre de sous systèmes est >2, on subdivise la cascade en autant de sous systèmes qu'il en contient et on étudie chaque sous système séparément.

Transformée de Laplace-Stieltjes (T L-S)

Soit X une variable aléatoire continue, F(x) est sa fonction de répartition.

$\tilde{F}(s) = E\left[e^{-sX}\right] = \int_0^{\infty} e^{-sx} f(x)dx = \int_0^{\infty} e^{-sx} dF(x)$, où s est une variable complexe. Cette intégrale est définie lorsque $\mathrm{Re}(s) \geq 0$.

Propriétés

1. Si X et Y sont des variables aléatoires indépendantes, alors
$E\left[e^{-s(X+Y)}\right] = E\left[e^{-sX}\right]E\left[e^{-sY}\right]$.
2. $TLS\left(\frac{df(t)}{dt}\right) = s\tilde{F}(s)$; $TLS\left(\frac{d^n f(t)}{dt^n}\right) = s^n \tilde{F}(s)$; $TLS\left(\int_0^t f(x)dx\right) = \tilde{F}(s)/s$.
3. $\tilde{F}(0) = \int_0^{\infty} f(t)dt$.
4. $\alpha_k = (-1)^k \tilde{F}^{(k)}(0)$.

Exemples

1. Loi exponentielle
$f(x) = \lambda e^{-\lambda x}, x \geq 0$.
$\tilde{F}(s) = \frac{\lambda}{\lambda + s}$.
2. Loi gamma G(n, λ)
$g(n,\lambda) = \frac{\lambda(\lambda x)^{n-1}}{(n-1)!} e^{-\lambda x}$, $x \geq 0$.
$\tilde{F}(s) = \left(\frac{\lambda}{\lambda + s}\right)^n, n = 1,2,\ldots$.

Z-transformée (fonction génératrice)

Soit X une variable aléatoire discrète à valeurs entières :
$X = n, n \in N$, $p_n = P(X = n)$.

$F(z) = \sum_{n=0}^{\infty} z^n p_n = E\left[z^X\right]$, où z est une variable complexe. F(z) est définie si $|z| \le 1$ et $F(0) = p_0$ ainsi que $F(1) = 1$.

Propriétés

1. $E[X] = F'(1)$;
2. $E\left[X^2\right] = F''(1) + F'(1)$;
3. Si X et Y sont des variables aléatoires indépendantes à valeurs entières, alors $E\left[z^{X+Y}\right] = E\left[z^X\right]E\left[z^Y\right]$.

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. $\sum_{n=0}^{\infty} a^n p_n z^n = F(az)$.

Exemples

1. Loi binomiale

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

2. Loi de Poisson

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

3. Loi géométrique

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

Transformée inverse de la fonction génératrice

Soit $P(z) = \sum_{k=0}^{\infty} p_k z^k$ la fonction génératrice de la distribution p_k, $k \geq 0$;

Méthode 1 : Série de Taylor dans la région *z=0.*

$$p_k = \frac{P^{(k)}(z)}{k!}\Big|_{z=0} .$$

Méthode 2 : Méthode récursive.

En général, une fonction génératrice est de la forme suivante $P(z) = \frac{\sum_{i=0}^{n} a_i z^i}{\sum_{j=0}^{m} b_j z^j}$.

Alors, $\sum_{k=0}^{\infty} p_k z^k \sum_{j=0}^{m} b_j z^j = \sum_{k=0}^{\infty}\sum_{j=0}^{m} p_k b_j z^{k+j} = \sum_{i=0}^{n} a_i z^i$. En comparant les coefficients de z^i, on obtient $\sum_{j=0}^{\min(m,i)} p_{i-j} b_j = \begin{cases} a_i & i \geq 0 \\ 0 & \end{cases}$. Alors, $p_i = \frac{1}{b_0}\left[a_i - \sum_{j=1}^{\min(m,i)} b_j p_{i-j} \right]$, $i \geq 0$. Il faut noter que $a_i = 0$ pour $i > n$.