

OBJECTIFS D'APPRENTISSAGE

Après avoir terminé l'étude de ce chapitre, vous pourrez :

1. Expliquer pourquoi des files d'attente se forment dans des systèmes non congestionnés.
2. Identifier l'objectif de l'analyse des files d'attente.
3. Énoncer les mesures de performance utilisées dans le cadre des files d'attente.
4. Formuler les hypothèses des principaux modèles de base.
5. Résoudre des problèmes classiques.



Chapitre 19

LES FILES D'ATTENTE

Plan du chapitre

Lecture : Attendre... un passe-temps populaire :
M^{me} Bonnes Manières 692

19.1 Introduction 693
19.2 Pourquoi y a-t-il de l'attente? 693
19.3 L'objectif de l'analyse des files d'attente 694
19.4 Les caractéristiques du système de files d'attente 695
 19.4.1 La population 695
 19.4.2 Le nombre de serveurs 696
 19.4.3 Les tendances quant à l'arrivée et au service 696
 19.4.4 La discipline de la file d'attente 699
19.5 Les mesures de performance 699
19.6 Les principaux modèles de files d'attente 699
 19.6.1 Modèles avec population infinie 699
 19.6.1.1 Les relations de base 700
 19.6.1.2 Modèle 1 : serveur unique, temps de service exponentiel 702

 19.6.1.3 Modèle 2 : serveur unique, temps de service constant 703
 19.6.1.4 Modèle 3 : serveurs multiples, temps de service exponentiel 703
 19.6.1.5 Optimisation des files d'attente 707
 19.6.1.6 Capacité maximale de la file d'attente 709
 19.6.1.7 Modèle 4 : serveurs multiples et règles de priorité 710
19.6.2 Modèle avec population finie 714
19.7 Autres approches d'analyse 720
19.8 Conclusion 721
 Terminologie 721
 Problèmes résolus 721
 Questions de révision et de discussion 723
 Problèmes 723
 Bibliographie 727



LECTURE

ATTENDRE... UN PASSE-TEMPS POPULAIRE : M^{me} BONNES MANIÈRES

(*Judith Martin, adaptation de Hocine Bourenane*)

Il y a plusieurs choses dans la vie pour lesquelles cela vaut la peine d'attendre, mais pas très longtemps. M^{me} Bonnes Manières limiterait, par exemple, le temps passé par les vendeurs à discuter avant de prendre une commande ou le temps pris par un mari qui a quitté sa femme pour se rendre compte de sa terrible erreur.

Quoiqu'il en soit, l'attente et le travail sont devenus des passe-temps populaires. Un spécialiste de l'analyse de l'attente a déterminé qu'une personne adulte passait au minimum le dixième de son temps à attendre. On attend les autobus, les ascenseurs, dans les banques, les magasins, les cinémas, les stations d'essence, les tribunaux, pour l'obtention du permis de conduire, chez le dentiste, etc.

On pourrait passer toute sa vie à subir ce genre d'attente classique. Cependant, il y a d'autres types d'attente : l'attente intermédiaire, comme attendre que la pluie cesse ou, encore, l'attente plus fébrile, comme attendre que votre bateau rentre à bon port. Il fut un temps où toute l'Amérique attendait d'être découverte dans un supermarché par un cinéaste et, aujourd'hui, chacun attend qu'une

caméra de télévision arrive pour lui demander de donner son avis au monde entier. C'est l'attente classique, l'attente à court terme, qui intéresse M^{me} Bonnes Manières. Si vous voulez en savoir plus sur les autres types d'attente, vous n'avez qu'à... attendre !

Il est tout à fait correct, malgré le fait que très peu de gens le comprennent, de refuser d'attendre au téléphone. Quand on demande à M^{me} Bonnes Manières de patienter quelques minutes au téléphone, elle réplique souvent par la négative. Malheureusement, la personne au bout du fil l'a déjà mise automatiquement en attente, car elle n'a pas attendu la réponse de M^{me} Bonnes Manières.

On devrait toujours refuser d'attendre pour un service inefficace et qui prend un temps infini. Au restaurant, on devrait être capable de vous donner le temps d'attente et de ne pas vous laisser attendre, excepté pour servir les clients arrivés avant vous. En effet, c'est inconvenant de refuser d'attendre en annonçant que nos besoins ont une primauté sur ceux des autres. M^{me} Bonnes Manières ne peut concevoir aucune situation d'attente ordinaire où une personne pourrait

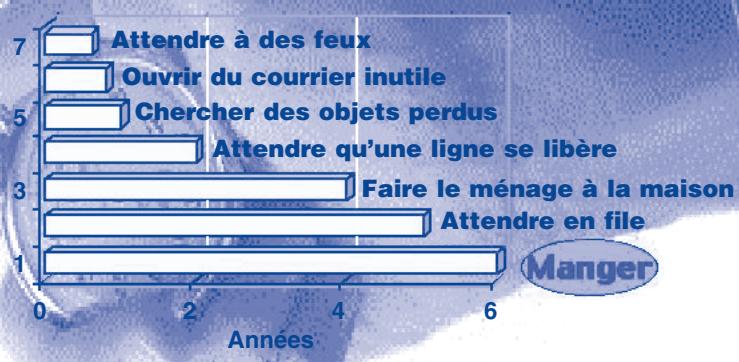
légitimement passer avant les autres. «Laissez-moi passer, s'il vous plaît, je suis enceinte, j'ai des douleurs, je pense que je vais accoucher !» Peut-être, mais que faites-vous donc dans un magasin au moment des soldes ?

La seule manière polie d'attendre est d'emporter avec soi du travail ou de quoi se divertir. Toute personne inoccupée dans une file d'attente est, par définition, un maniaque qui peut se déchaîner à tout moment. Un bon roman de Jane Austen a pu préserver l'esprit naturellement tranquille de M^{me} Bonnes Manières. Selon elle, même le fait d'engager la discussion pour passer le temps peut être dangereux. C'est tout simplement une honte de voir deux personnes qui attendent en train de discuter tranquillement. Par définition, ce sont des comploteurs potentiels.

Pour conclure et en attendant, relaxez-vous en écoutant de la musique ! Pourquoi pas Georges Moustaki ? «Passe, passe le temps, il n'y en a plus pour très longtemps. Pendant que j'attendais...»

©1980 Judith Martin. Reproduit avec autorisation.

PASSE, PASSE LE TEMPS....



Source : U.S. News & World Report, 30 janvier 1989, p.81

19.1 INTRODUCTION

L'article portant sur M^{me} Bonnes Manières parodie une des réalités de la vie : l'attente en file. Pour ceux qui attendent en file, la solution est très simple : ajouter des ressources ou bien agir, faire n'importe quoi pour accélérer le service. C'est l'évidence même. Cependant, ce n'est pas aussi simple, car il faut tenir compte de certaines subtilités. Premièrement, sur une longue période, la majorité des processus de service ont une capacité de traitement supérieure à celle qui est nécessaire. Par conséquent, le problème des files d'attente ne survient que pendant de courtes périodes. Deuxièmement, il ne faut pas perdre de vue le fait qu'à certains moments, le système est vide : les employés sont inoccupés et attendent que les clients se présentent. En augmentant la capacité, on ne fait qu'augmenter le temps d'inoccupation des employés. Donc, si on veut concevoir un système de service, il faut comparer le coût associé au niveau de service (capacité) mis en place et le coût associé à l'attente des clients. La planification et l'analyse de la capacité de service sont des thèmes traités par la théorie des files d'attente. Cette théorie est une approche mathématique permettant d'analyser les files d'attente. Elle est basée sur l'étude des équipements téléphoniques automatiques réalisée au début du xx^e siècle par l'ingénieur danois en télécommunication, A. K. Erlang. L'application de cette théorie n'a été généralisée à divers types de problèmes qu'après la Seconde Guerre mondiale.

La théorie mathématique des files d'attente étant assez complexe, on ne s'attardera dans ce chapitre qu'aux concepts et aux hypothèses relatifs à la résolution des problèmes d'attente. On utilisera les formules et les tables disponibles.

Les files d'attentes se forment lorsque les clients arrivent de façon aléatoire pour se faire servir. Les exemples les plus courants de la vie de tous les jours sont les caisses des supermarchés, les établissements de restauration rapide, les billetteries des aéroports, les cinémas, les bureaux de poste, les banques. Toutefois, lorsqu'on parle d'attente, on pense souvent à des personnes. Or, les «clients» en attente sont aussi des commandes en attente de traitement, des camions en attente de chargement ou de déchargement, des machines en attente de réparation, des programmes d'ordinateur qui attendent d'être exécutés, des avions qui attendent l'autorisation de décoller, des bateaux qui attendent les remorqueurs pour accoster, les voitures aux panneaux d'arrêt, les patients dans les salles d'urgence, etc.

Généralement, les clients voient dans l'attente une activité sans valeur ajoutée et, s'ils attendent trop longtemps, ils associent cette perte de temps à une mauvaise qualité de service. De la même façon, au sein de l'entreprise, des employés inoccupés ou des équipements inutilisés représentent des activités sans valeur ajoutée. Pour éviter ces situations, la majorité des entreprises ont mis en place des processus d'amélioration continue dont le but ultime est l'élimination de toute forme de gaspillage, notamment l'attente. Tous ces exemples révèlent l'importance de l'analyse des files d'attente. Commençons par une question fondamentale : pourquoi y a-t-il de l'attente ?

théorie des files d'attente
Approche mathématique servant à l'analyse des files d'attente.

19.2 POURQUOI Y A-T-IL DE L'ATTENTE ?

Il est surprenant d'apprendre que des files d'attente se forment même dans les systèmes non congestionnés. Par exemple, un établissement de restauration rapide qui peut traiter en moyenne 200 commandes à l'heure voit malgré tout se former des files d'attente avec un nombre moyen de 150 commandes à l'heure. L'expression clé est «en moyenne». Le problème vient du fait que les arrivées des clients ont lieu à intervalles aléatoires plutôt qu'à intervalles fixes. De plus, certaines commandes requièrent un temps de traitement plus long. En d'autres termes, les processus d'arrivée et de service ont un degré de variabilité élevé. Par conséquent, le système est soit temporairement congestionné, ce qui crée des files d'attente, soit vide, parce qu'aucun client ne se présente. Donc, si le système n'est pas congestionné d'un point de vue macro, il l'est d'un point de vue micro. Par ailleurs, en cas de variabilité minimale ou inexistante (arrivée selon les rendez-vous et temps de service constant), aucune file d'attente ne se forme.

19.3 L'OBJECTIF DE L'ANALYSE DES FILES D'ATTENTE

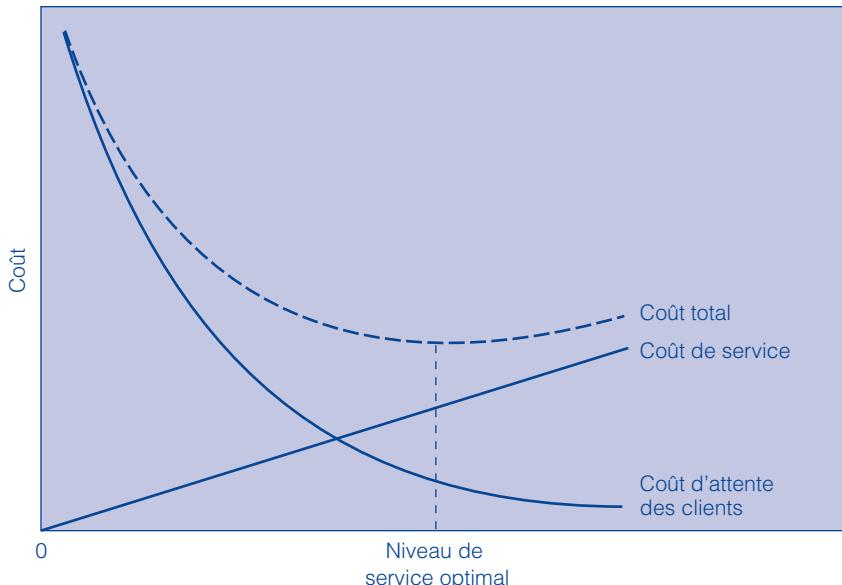
L'objectif de l'analyse des files d'attente est de minimiser le coût total, qui équivaut à la somme de deux coûts : le coût associé à la capacité de service mise en place (coût de service) et le coût associé à l'attente des clients (coût d'attente). Le coût de service est le coût résultant du maintien d'un certain niveau de service, par exemple le coût associé au nombre de caisses dans un supermarché, au nombre de réparateurs dans un centre de maintenance, au nombre de guichets dans une banque, au nombre de voies d'une autoroute, etc. En cas de ressources inoccupées, la capacité est une valeur perdue, car elle est non stockable. Les coûts d'attente sont constitués des salaires payés aux employés qui attendent pour effectuer leur travail (mécanicien qui attend un outil, chauffeur qui attend le déchargement du camion, etc.), du coût de l'espace disponible pour l'attente (grandeur de la salle d'attente dans une clinique, longueur d'un portique de lave-auto, kérozène consommé par les avions qui attendent pour atterrir) et, bien sûr, du coût associé à la perte de clients impatients qui vont chez les concurrents.

En pratique, lorsque le client est externe à l'entreprise, le coût d'attente est difficile à évaluer, car il s'agit d'un impact plutôt que d'un coût pouvant être comptabilisé. Cependant, on peut considérer les temps d'attente comme un critère de mesure du niveau de service. Le gestionnaire décide du temps d'attente acceptable, «tolérable», et il met en place la capacité susceptible de fournir ce niveau de service.

Lorsque le client est interne à l'entreprise — les clients sont les machines et les commis, l'équipe d'entretien —, on peut établir directement certains coûts se rapportant au temps d'attente des clients (machines). Par ailleurs, il ne faut pas conclure trop rapidement que pour l'entreprise, le coût du temps d'attente d'un employé qui attend est égal à son salaire durant le temps d'attente ; cela impliquerait que la baisse nette des gains de l'entreprise, du fait de l'inactivité d'un employé, est égale au salaire de ce dernier, ce qui, *a priori*, n'est pas évident. L'employé, qu'il travaille ou qu'il attende, reçoit le même salaire. Par contre, sa contribution aux gains de l'entreprise est réellement perdue, car la productivité baisse. Quand un opérateur de machine est inactif parce qu'il attend, sa force productive (qui peut comprendre, outre son salaire, une proportion des coûts fixes de l'entreprise) est perdue. En d'autres termes, il faut tenir compte non pas de la ressource physique en attente, mais plutôt de la valeur (coût) de toutes les ressources économiques inactives, et évaluer ensuite la perte de profit à partir de la perte de productivité.

L'objectif de l'analyse des files d'attente est de trouver un compromis entre le coût associé à la capacité de service et le coût d'attente des clients. La figure 19.1 illustre bien ce concept. Notez que lorsque la capacité de service augmente, le coût de service augmente. Par souci de simplicité, nous avons illustré un coût de service linéaire. Cela n'affecte en rien la démonstration. Lorsque la capacité de service augmente, le nombre de clients en attente et le temps d'attente tendent à diminuer, donc les coûts d'attente diminuent. Le coût total (la somme des coûts de service et d'attente) est représenté sur le graphique par une courbe en forme de U. Graphiquement, il suffit de déterminer le niveau de service se traduisant par le coût total minimum. (Contrairement au modèle de la quantité économique utilisé dans la gestion des stocks, le minimum n'est pas nécessairement atteint au point d'intersection de la droite et de la courbe.)

Dans le cas d'une clientèle externe à l'entreprise, les files d'attente donnent une image négative de la qualité du service offert. Dans cette situation, les entreprises auront tendance à augmenter la rapidité du service plutôt que d'augmenter le nombre d'employés. Le fait d'abaisser le coût d'attente aura pour effet de déplacer vers le bas la courbe en U, qui représente le coût total.

**Figure 19.1**

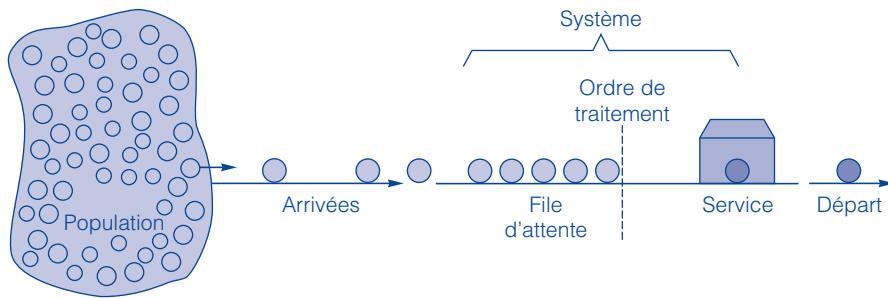
L'objectif de l'analyse des files d'attente est de minimiser la somme de deux coûts : le coût d'attente des clients et le coût de service.

19.4 LES CARACTÉRISTIQUES DU SYSTÈME DE FILES D'ATTENTE

Dans le cadre de la théorie des files d'attente, on a conçu plusieurs modèles d'analyse. Le succès de l'analyse des files d'attente repose surtout sur le choix du modèle approprié. Plusieurs caractéristiques sont à prendre en considération :

1. La population.
2. Le nombre de serveurs.
3. Les tendances quant à l'arrivée et au service.
4. L'ordre de traitement des clients.

La figure 19.2 illustre un système de file d'attente.

**Figure 19.2**

Système de file d'attente simple

19.4.1 La population

Dans la théorie des files d'attente, on appelle « population » la source de clients potentiels.

Il y a deux situations possibles. Dans le premier cas, la **population** est **infinie**, c'est-à-dire que le nombre potentiel de clients est infiniment grand en tout temps. C'est le cas des clients des supermarchés, des banques, des restaurants, des cinémas, des centres d'appels, etc. De plus, les clients proviennent de toutes les régions possibles. Dans la deuxième situation, la **population** est **finie**, ce qui signifie que le nombre de clients potentiels est limité.

Un bon exemple est le nombre de machines, d'avions, etc., en réparation dans le centre de maintenance d'une entreprise. L'entreprise en question possède un nombre fini de machines, d'avions, etc. Voici d'autres situations semblables : une infirmière

population infinie

Le nombre de clients qui arrivent est illimité.

population finie

Le nombre de clients qui arrivent est limité.

ayant la charge de 10 patients, un employé de banque chargé de remplir et de vider 4 guichets automatiques, une secrétaire qui s'occupe de 5 représentants, un contrôleur de la navigation aérienne qui dirige l'atterrissement ou le décollage de 5 avions, etc.

19.4.2 Le nombre de serveurs

serveur

Entité qui fournit le service.

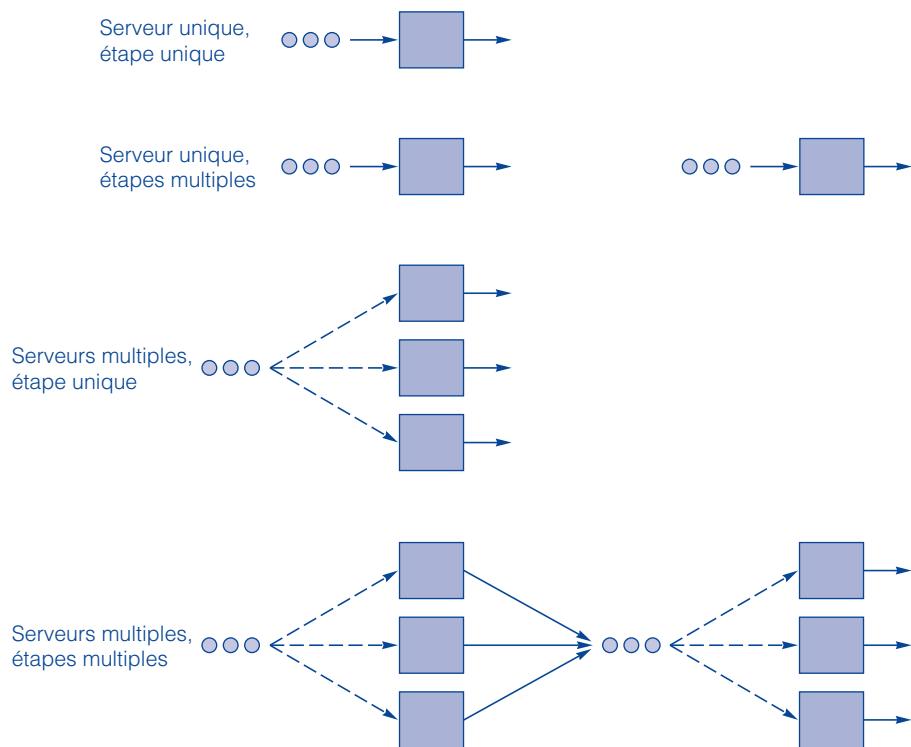


La capacité de service dépend de la capacité de chaque serveur et du nombre de serveurs disponibles. Le terme «serveur» représente ici la ressource et, en général, on suppose qu'un serveur ne traite qu'un client à la fois.

Les systèmes de files d'attente fonctionnent avec serveur unique ou serveurs multiples (plusieurs serveurs travaillant en équipe constituent un serveur unique, par exemple une équipe chirurgicale). Les exemples de systèmes de files d'attente avec serveur unique sont nombreux : les petits magasins avec une seule caisse, tels que les dépanneurs, certains cinémas, certains lave-autos et établissements de restauration rapide avec guichet unique. Les systèmes à multiples serveurs sont les banques, les billetteries d'aéroports, les garages et les stations-service. La figure 19.3 illustre les systèmes de files d'attente les plus courants. Pour des raisons pratiques, ceux qui sont étudiés dans le cadre de ce chapitre comprennent une seule étape.

Figure 19.3

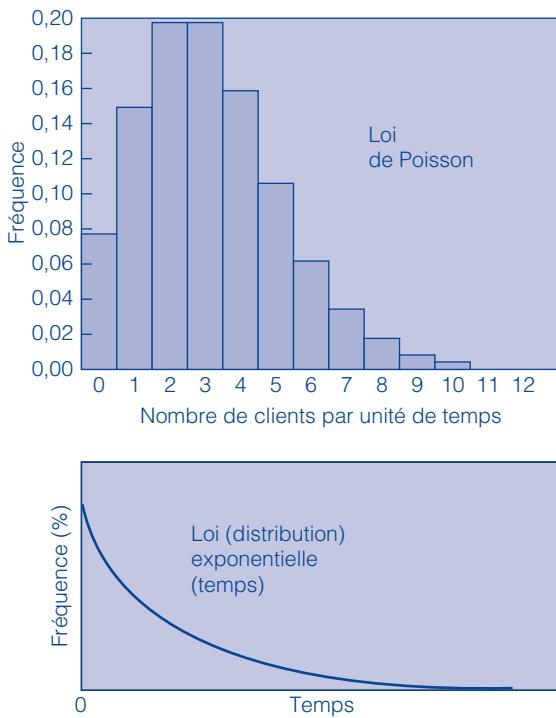
Quatre types de systèmes de files d'attente



19.4.3 Les tendances quant à l'arrivée et au service

Les files d'attente résultent de la variabilité des tendances d'arrivée et de service. Elles se forment parce que le degré élevé de variation dans les intervalles entre les arrivées et dans les temps de service cause des congestions temporaires. Dans plusieurs cas, on peut représenter ces variations par des distributions théoriques de probabilités. Dans les principaux modèles utilisés, on suppose que le nombre d'arrivées dans un intervalle donné suit la loi de Poisson, alors que le temps de service suit une loi exponentielle. La figure 19.4 illustre ces deux distributions.

En général, la distribution de Poisson donne un assez bon aperçu du nombre de clients qui arrivent par unité de temps (par exemple le nombre de clients à l'heure). La figure 19.5 A illustre les arrivées distribuées selon la loi de Poisson (par exemple des accidents) pendant une période de trois jours. Durant certaines heures, on note de trois à quatre accidents ; à d'autres, un ou deux, et pour certaines, aucun.

**Figure 19.4**

Distributions de Poisson
et exponentielle

La distribution exponentielle, quant à elle, donne une assez bonne approximation des temps de service (par exemple avant l'arrivée des premiers secours auprès des victimes d'accidents). La figure 19.5 B illustre le temps de service pour des clients qui arrivent selon le processus illustré à la figure 19.5 A. Remarquez que la plupart des temps de service sont très courts — certains sont proches de zéro — et quelques-uns, assez longs. C'est la caractéristique typique de la distribution exponentielle. Par exemple, les opérations traitées au guichet d'une banque prennent approximativement le même temps (assez court), alors qu'un nombre limité de clients requièrent un temps de traitement assez long.

Les files d'attente se forment plus souvent lorsque les arrivées se font en groupe ou que les temps de service sont particulièrement longs ; elles se créent presque à coup sûr si ces deux facteurs se manifestent. Par exemple, notez, à la figure 19.5 B, le temps de service particulièrement long pour le client n° 7 au jour 1. À la figure 19.5 A, le client n° 7 arrive à 10 heures et les 2 clients suivants arrivent juste après, ce qui crée alors une file d'attente. Une situation similaire s'est présentée le jour 3 avec les 3 derniers clients : le temps de service assez long pour le client n° 13 (figure 19.5 B) combiné au temps relativement court entre les deux arrivées suivantes (figure 19.5 A, jour 3) va certainement engendrer (ou augmenter) une file d'attente.

Remarquez qu'il existe une relation entre la distribution de Poisson et la distribution exponentielle. En d'autres termes, si le temps de service suit la loi exponentielle, le taux de service (nombre de clients servis par unité de temps) suit la loi de Poisson. De la même manière, si le taux d'arrivée suit la loi de Poisson, le **temps interarrivées** (temps entre deux arrivées successives) suit une loi exponentielle. Par exemple, si un centre de service a la capacité de traiter en moyenne 12 clients à l'heure (taux de service), le temps moyen de service est de cinq minutes. Si le taux moyen d'arrivée est de 10 clients à l'heure, le temps moyen entre 2 arrivées successives est de 6 minutes. Ainsi, les modèles de files d'attente décrits dans ce chapitre ont généralement comme processus d'arrivée un processus de Poisson ou, de façon équivalente, des temps interarrivées exponentiels et des temps de service distribués selon une loi exponentielle. En pratique, avant d'utiliser un modèle, il faut vérifier ces caractéristiques. Dans certains cas, on peut le faire en colligeant des données et en les représentant graphiquement. On ajuste ensuite la distribution observée à la distribution théorique.

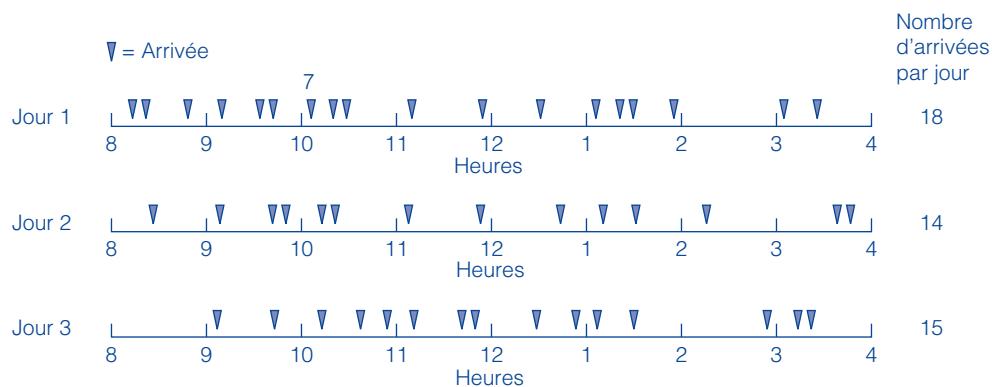
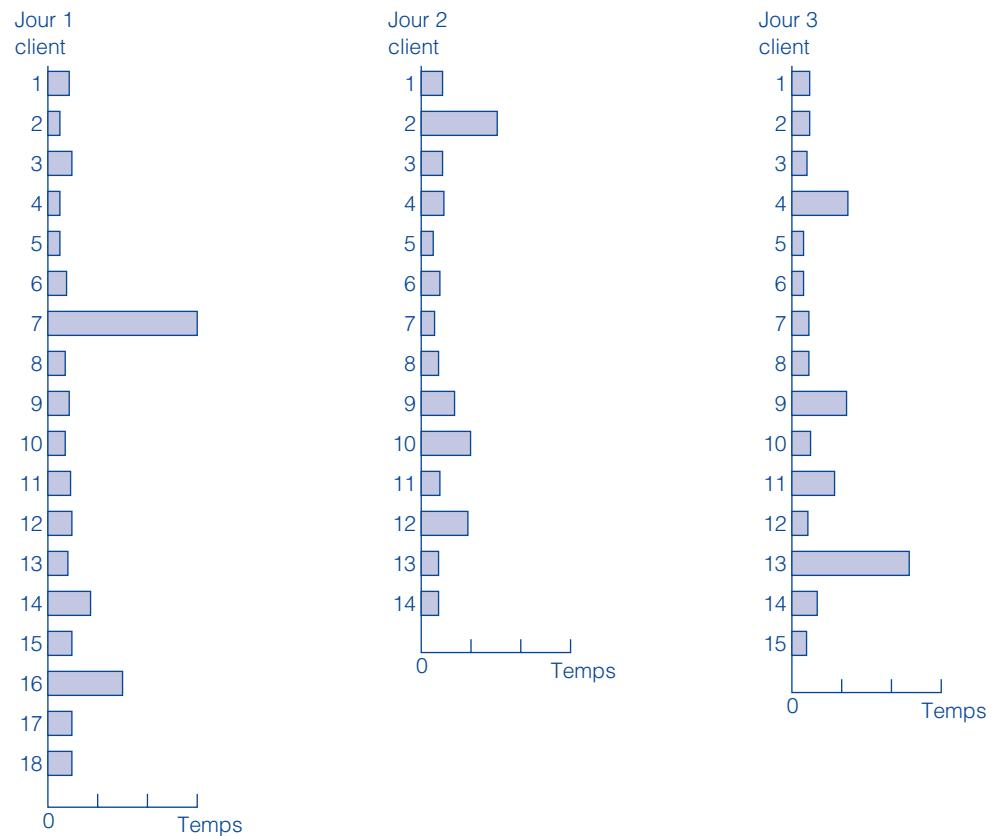


temps interarrivées

Temps entre deux arrivées successives.

Figure 19.5

Arrivées distribuées selon la loi de Poisson et temps de service distribués de façon exponentielle

A. Processus d'arrivée*B. Temps de service*

Il est toutefois préférable pour ce type de problème d'utiliser le test du chi-deux (χ^2) : ce sujet ne fait pas l'objet d'étude de ce chapitre et il est développé dans la majorité des ouvrages de statistique. Aujourd'hui, il existe des logiciels très puissants qui permettent d'ajuster très rapidement une série d'observations à une distribution théorique de probabilité. L'un des plus complets est ExpertFit, conçu par Averill Law et associés.

Par ailleurs, les recherches ont démontré que si ces hypothèses sont généralement appropriées pour le processus d'arrivée, elles le sont moins pour le processus de service. Dans ce cas, les solutions à considérer consistent à : 1) mettre au point un modèle plus approprié ; 2) utiliser un meilleur modèle (généralement plus complexe) ; 3) avoir recours à la simulation numérique. Ces solutions requièrent généralement plus d'efforts, de temps et d'argent que les modèles de files d'attente présentés dans ce chapitre.

19.4.4 La discipline de la file d'attente

La discipline de la file d'attente concerne l'ordre de traitement des clients. Dans tous les modèles décrits dans les pages suivantes, on suppose que la règle de priorité est : premier entré, premier servi (PEPS). C'est la règle la plus communément utilisée dans les entreprises de services ; elle procure aux clients un sentiment de justice, bien qu'elle pénalise les clients dont le temps de service est court. Elle est appliquée dans les banques, les magasins, les cinémas, les restaurants, les intersections avec arrêt obligatoire, les contrôles douaniers, etc. Certains systèmes ne s'en servent pas : les salles d'urgence des hôpitaux, en général, utilisent trois niveaux de priorité (les cas graves étant traités en priorité) ; les usines traitent les commandes urgentes et les ordinateurs centraux traitent les tâches par ordre d'importance. Certains clients devront donc attendre plus longtemps, même s'ils sont arrivés plus tôt. Prenons un exemple. Vous venez d'avoir un bébé et vous êtes une personne plutôt anxieuse. Si vous allez à l'urgence de l'hôpital Sainte-Justine de Montréal à la moindre petite fièvre de votre bébé, armez-vous de patience et priez pour qu'il n'y ait pas trop de cas graves ce jour-là. Les autres règles de priorité susceptibles d'être appliquées sont les temps d'opération les plus courts, les commandes ou les clients les plus importants, les urgences, les réservations en priorité, les délais de livraison les plus courts, etc.

discipline de la file d'attente

Ordre dans lequel les clients sont traités.



19.5 LES MESURES DE PERFORMANCE

Les gestionnaires ont à leur disposition cinq outils de mesure ou indices pour évaluer la performance d'un système de production de biens ou de services existant ou celle d'un système qu'ils veulent concevoir. Ces mesures sont :

1. Le nombre moyen de clients qui attendent en file ou dans le système¹.
2. Le temps moyen d'attente en file et dans le système.
3. Le taux d'utilisation du système, c'est-à-dire le pourcentage de la capacité utilisée.
4. Le coût associé au niveau de service (capacité) mis en place.
5. La probabilité qu'un client potentiel attende pour être servi.

Parmi ces cinq outils de mesure, le taux d'utilisation du système nécessite quelques éclaircissements. Il reflète l'étendue de l'occupation des serveurs plutôt que leur inactivité. Il est logique de penser qu'une bonne gestion des ressources implique un taux d'utilisation de 100 %. Cependant, comme le montre la figure 19.6, le fait d'augmenter le taux d'utilisation revient à augmenter à la fois le nombre de clients qui attendent et le temps moyen d'attente. En fait, ces deux mesures augmentent indéfiniment lorsque le taux d'utilisation approche de 100 %. Si tous les serveurs sont occupés, il est certain que les clients potentiels qui arrivent vont attendre. Cela implique que dans des conditions normales d'opération, un taux d'utilisation de 100 % est irréaliste. Le gestionnaire devrait plutôt essayer d'équilibrer le système de telle sorte que la somme des coûts de service et d'attente soit minimale, tel qu'illustré à la figure 19.11.

19.6 LES PRINCIPAUX MODÈLES DE FILES D'ATTENTE

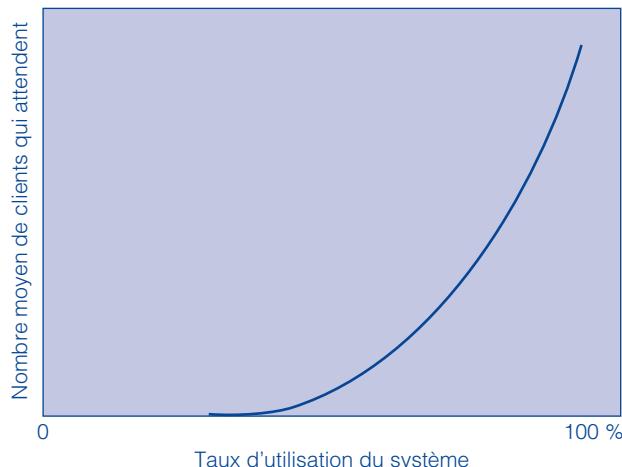
19.6.1 Modèles avec population infinie

Plusieurs modèles de files d'attente sont à la disposition des gestionnaires pour leur permettre de concevoir des systèmes de production de biens ou de services ou de représenter un système réel afin d'en analyser la performance. Dans ce chapitre, nous

1. Voir la figure 19.2.

Figure 19.6

Le nombre moyen de clients qui attendent en file et le temps moyen d'attente des clients en file augmentent de façon exponentielle à mesure que le taux d'utilisation augmente.



présentons les quatre modèles de base les plus utilisés. Le but n'est pas d'étudier de façon exhaustive les modèles, mais plutôt d'analyser un certain nombre d'entre eux. Tous ont pour hypothèse que le taux d'arrivée est distribué selon la loi de Poisson. On suppose aussi que le système étudié est en **régime permanent** (stationnaire), c'est-à-dire que les taux d'arrivée et de service sont stables. Les quatre modèles présentés sont :

1. Serveur unique, temps de service exponentiel.
2. Serveur unique, temps de service constant.
3. Serveurs multiples, temps de service exponentiel.
4. Serveurs multiples, règles de priorité multiples, temps de service exponentiel.

Afin de faciliter l'utilisation des modèles, le tableau 19.1 présente les symboles et la terminologie utilisés pour les modèles avec population infinie.

TABLEAU 19.1

Symboles (modèles avec population infinie)

Symbol	Signification
λ	Taux d'arrivée des clients
μ	Taux de service
ρ	Taux d'utilisation du système
\bar{n}_w	Nombre moyen de clients qui attendent d'être servis
\bar{n}_s	Nombre moyen de clients dans le système (clients qui attendent et clients qui sont en train d'être servis)
$\frac{1}{\mu}$	Temps de service
\bar{t}_w	Temps moyen d'attente en file
\bar{t}_s	Temps moyen d'attente dans le système (temps d'attente en file, plus le temps de service)
P_0	Probabilité qu'il y ait zéro unité (client) dans le système
P_n	Probabilité qu'il y ait n unités (clients) dans le système
M	Nombre de serveurs

19.6.1.1 Les relations de base

Dans les modèles de files d'attente avec population infinie, il existe certaines relations de base (entre certains paramètres et les mesures de performance) qui permettent de déterminer les mesures de performance désirées grâce à quelques valeurs clés. Les principales relations sont présentées ci-dessous :

Remarque : les taux d'arrivée (λ) et de service (μ) doivent être exprimés dans la même unité de mesure (clients à l'heure, clients par minute, etc.).

Le taux d'utilisation du système: il représente le rapport entre la demande (mesurée grâce au taux d'arrivée, λ) et la capacité de service (produit du nombre de serveurs M par le taux de service, μ).

$$\rho = \frac{\lambda}{M\mu} \quad (19-1)$$

Le nombre moyen de clients en train d'être servis si M = 1 :

$$\rho = \frac{\lambda}{\mu} \quad (19-2)$$

Le nombre moyen de clients en file :

\bar{n}_l est obtenu à partir d'une table ou de la formule appropriée, selon le modèle en question.

Le nombre de clients dans le système :

$$\bar{n}_s = \bar{n}_l + \rho \quad (19-3)$$

Le temps moyen d'attente en file :

$$\bar{t}_l = \frac{\bar{n}_l}{\lambda} \quad (19-4)$$

Le temps moyen d'attente dans le système :

$$\bar{t}_s = \bar{t}_l + \frac{1}{\mu} = \frac{\bar{n}_s}{\lambda} \quad (19-5)$$

Pour ces modèles, le taux d'utilisation du système doit être inférieur à 1 ($\lambda < M\mu$). D'autre part, ces modèles ne s'appliquent qu'à des systèmes non congestionnés. Il n'y a aucune utilité à analyser les systèmes dans lesquels $\lambda > M\mu$, car il est évident que dans de tels cas, ils sont congestionnés.

Le nombre moyen de clients qui attendent en file (\bar{n}_l) est l'élément clé qui sert à déterminer les autres mesures de performance du système, tels le nombre moyen de clients dans le système, le temps moyen passé en file et le temps moyen passé dans le système. Par conséquent, lorsqu'on veut résoudre des problèmes de files d'attente, la première mesure de performance à considérer est \bar{n}_l .

Les clients d'une boulangerie se présentent généralement en matinée (calcul effectué selon la loi de Poisson), à raison de 18 clients en moyenne à l'heure. On estime que chaque vendeur au comptoir peut servir un client (temps distribué selon une loi exponentielle) en moyenne en quatre minutes.

- Quels sont les taux d'arrivée et de service? Calculez le nombre moyen de clients en train d'être servis (supposez que le taux d'utilisation du système est inférieur à 1).
- En supposant que le nombre moyen de clients qui attendent en file est égal à 3,6, déterminez le nombre moyen de clients dans le système, le temps moyen passé en file et le temps moyen passé dans le système.
- Déterminez le taux d'utilisation du système lorsque $M = 2, 3, 4$ serveurs.
- Le taux d'arrivée est donné dans l'énoncé du problème: $\lambda = 18$ clients à l'heure. Il faut traduire le temps moyen de service en heures, puis déduire le taux sachant que pour une distribution exponentielle, la moyenne est égale à $1/\mu$. Donc, $1/\mu = 4$ minutes par client / 60 minutes par heure = $1/15$, ce qui donne un taux moyen de service $\mu = 15$ clients à l'heure.

$$\rho = \frac{\lambda}{\mu} = \frac{18}{15} = 1,2 \text{ client}$$

Exemple 1

Solution

b) Nombre moyen de clients dans le système :

$$\bar{n}_s = \bar{n}_l + \rho = 3,6 + 1,2 = 4,8 \text{ clients (puisque } \bar{n}_l \text{ est égal à 3,6)}$$

Temps moyen d'attente en file :

$$\bar{t}_l = \frac{\bar{n}_l}{\lambda} = \frac{3,6}{18} = 0,20 \text{ heure par client, ou } 0,20 \times 60 \text{ minutes} = 12 \text{ minutes}$$

Temps moyen passé dans le système :

$$\bar{t}_s = \bar{t}_l + \frac{1}{\mu} = \frac{\bar{n}_l}{\lambda} + \frac{1}{\mu} = 0,20 + \frac{1}{15} = 0,267 \text{ heure, soit environ 16 minutes}$$

c) Taux d'utilisation du système $\rho = \lambda/M\mu$:

$$M = 2, \quad \rho = \frac{18}{2(15)} = 0,60$$

$$M = 3, \quad \rho = \frac{18}{3(15)} = 0,40$$

$$M = 4, \quad \rho = \frac{18}{4(15)} = 0,30$$

Par conséquent, lorsque la capacité de service augmente, le taux d'utilisation du système diminue.

19.6.1.2 Modèle 1 : serveur unique, temps de service exponentiel

Le modèle classique (le plus simple) d'analyse des files d'attente concerne les systèmes comptant un seul serveur (ou une seule équipe). La règle de priorité est « premier entré, premier servi (PEPS) »; on suppose que le processus d'arrivée suit une loi de Poisson et que le temps de service suit une loi exponentielle. Il n'y a aucune restriction quant à la longueur de la file proprement dite.

Le tableau 19.2 présente les formules servant à calculer les mesures de performance pour un modèle avec serveur unique. On les utilise conjointement avec les formules des tableaux 19.1 à 19.5.

TABLEAU 19.2

Formules pour le modèle de base (serveur unique, temps de service exponentiel)

Mesure de performance	Équation
Nombre moyen de clients en file	$\bar{n}_l = \frac{\lambda^2}{\mu(\mu - \lambda)} \quad (19-6)$
Nombre moyen de clients dans le système	$\bar{n}_s = \frac{\lambda}{(\mu - \lambda)}$
Temps moyen d'attente en ligne	$\bar{t}_l = \frac{\lambda}{\mu(\mu - \lambda)}$
Temps moyen passé dans le système	$\bar{t}_s = \frac{1}{(\mu - \lambda)}$
Probabilité qu'il y ait zéro unité dans le système	$P_0 = 1 - \left(\frac{\lambda}{\mu}\right) \quad (19-7)$
Probabilité qu'il y ait n unités dans le système	$P_n = P_0 \left(\frac{\lambda}{\mu}\right)^n \quad (19-8a)$
Probabilité qu'il y ait moins de n unités dans le système	$P_{<n} = 1 - \left(\frac{\lambda}{\mu}\right)^n \quad (19-8b)$

Exemple 2

Une compagnie aérienne envisage d'ouvrir un point de vente dans un nouveau centre commercial. Elle compte y faire travailler un agent qui sera responsable des réservations et de la vente de billets. On prévoit un achalandage de 15 clients à l'heure en moyenne; on estime aussi que la distribution des arrivées peut être calculée selon la loi de Poisson et que le temps de service sera de 3 minutes en moyenne par client (distribution exponentielle). Déterminez les mesures de performance suivantes :

- a) Taux d'utilisation du système.
- b) Pourcentage d'inactivité de l'agent.
- c) Nombre moyen de clients qui attendent pour être servis.
- d) Temps moyen passé par un client dans le système.
- e) Probabilité qu'il n'y ait aucun client dans le système et probabilité qu'il y ait quatre clients dans le système.

$$\lambda = 15 \text{ clients à l'heure et } 3 \text{ minutes/client} = \text{temps de service} = \frac{1}{\mu}$$

Solution

$$\text{donc } \mu = \left(\frac{1}{3} \frac{\text{client}}{\text{minute}} \right) \times 60 \text{ minutes par heure} = 20 \text{ clients à l'heure}$$

- a) $\rho = \frac{\lambda}{M\mu} = \frac{15}{1(20)} = 0,75$
- b) Pourcentage d'inactivité $= 1 - \rho = 1 - 0,75 = 0,25$, c'est-à-dire 25 % du temps
- c) $\bar{n}_l = \frac{\lambda^2}{\mu(\mu - \lambda)} = \frac{15^2}{20(20 - 15)} = 2,25 \text{ clients}$
- d) $\bar{t}_s = \frac{1}{(\mu - \lambda)} = \frac{1}{(20 - 15)} = 0,20 \text{ heure ou 12 minutes}$
- e) $P_0 = 1 - \frac{\lambda}{\mu} = 1 - \frac{15}{20} = 0,25$ et $P_4 = P_0 \left(\frac{\lambda}{\mu} \right)^4 = 0,25 \left(\frac{15}{20} \right)^4 = 0,079$

19.6.1.3 Modèle 2 : serveur unique, temps de service constant

Comme nous l'avons signalé précédemment, les files d'attente sont la conséquence directe de phénomènes aléatoires et du degré élevé de variabilité des taux d'arrivée et de service. Si, dans un système donné, on arrive à diminuer ou à réduire les variations d'un taux ou des deux, on peut également raccourcir les files d'attente de façon significative. Toutefois, dans le cas où les temps de service sont constants, le nombre moyen de clients qui attendent en file diminue de moitié.

$$\bar{n}_l = \frac{\lambda^2}{2 \mu(\mu - \lambda)} \quad (19-9)$$

Le temps d'attente en file est aussi réduit de moitié.

On retrouve ce modèle dans plusieurs situations, notamment lorsque le serveur est une machine automatique. Les lave-autos en sont un exemple.

Un lave-auto avec file unique a été programmé pour laver une automobile en cinq minutes. En fin de semaine, particulièrement le samedi, il arrive (selon un processus de Poisson) huit voitures à l'heure en moyenne. Déterminez :

Exemple 3

- a) Le nombre moyen de voitures dans la file d'attente.
- b) Le temps moyen passé dans la file et le temps moyen passé dans le système.

$$\mu = 1 \text{ toutes les 5 minutes ou encore 12 voitures à l'heure ; } \lambda = 8 \text{ voitures à l'heure}$$

$$\text{a) } \bar{n}_l = \frac{\lambda^2}{2\mu(\mu - \lambda)} = \frac{8^2}{2(12)(12 - 8)} = 0,667 \text{ voiture}$$

Solution

$$\text{b) } \bar{t}_l = \frac{\bar{n}_l}{\lambda} = \frac{0,667}{8} = 0,083 \text{ heure ou 5 minutes}$$

$$\bar{t}_s = \bar{t}_l + \frac{1}{\mu} = \frac{\bar{n}_s}{\lambda} + \frac{1}{\mu} = \frac{0,667}{8} + \frac{1}{12} = 0,167 \text{ heure ou 10 minutes}$$

19.6.1.4 Modèle 3 : serveurs multiples, temps de service exponentiel

Un tel modèle existe lorsqu'il y a deux serveurs ou plus qui travaillent en parallèle, de façon indépendante. Il faut tout d'abord vérifier les hypothèses suivantes :

1. Le processus d'arrivée est distribué selon une loi de Poisson et le processus de service, selon une loi exponentielle.
2. Le taux de service moyen est identique pour tous les serveurs.
3. Les clients sont traités selon l'ordre d'arrivée : premier entré, premier servi (règle PEPS).

Dans le tableau 19.3, vous trouverez les formules permettant de calculer les mesures de performance de ce modèle. Vous constaterez qu'elles sont beaucoup plus complexes que celles du modèle 1, particulièrement celles qui déterminent \bar{n}_l et P_0 . Nous vous les présentons pour montrer leur complexité et compléter la description de ce modèle, mais on utilise plutôt le tableau 19.4, qui donne les valeurs de \bar{n}_l et de P_0 pour différentes valeurs de λ/μ et de M .

Pour se servir du tableau 19.4, on commence par calculer la valeur de λ/μ (arrondie aux décimales près comme dans le tableau), puis on lit tout simplement les valeurs de \bar{n}_l et de P_0 correspondant au nombre approprié de serveurs, M . Par exemple, si $\lambda/\mu = 0,50$ et $M = 2$, on peut lire : $\bar{n}_l = 0,033$ et $P_0 = 0,600$. On peut se servir de ces valeurs pour déterminer d'autres mesures de performance. Notez que les formules du tableau 19.3 et les valeurs du tableau 19.4 donnent des moyennes. On peut utiliser le tableau 19.4 pour le modèle 1 (serveur unique, temps de service exponentiel) en prenant $M = 1$.

TABLEAU 19.3

Formules pour le modèle de files d'attente (serveurs multiples, temps de service exponentiel)

Mesure de performance	Équation
Nombre moyen de clients en file	$\bar{n}_l = \frac{\lambda\mu \left(\frac{\lambda}{\mu}\right)^M}{(M-1)!(M\mu - \lambda)^2} P_0 \quad (19-10)$
Probabilité qu'il y ait zéro unité dans le système	$P_0 = \left[\sum_{n=0}^{M-1} \frac{\left(\frac{\lambda}{\mu}\right)^n}{n!} + \frac{\left(\frac{\lambda}{\mu}\right)^M}{M! \left(1 - \frac{\lambda}{M\mu}\right)} \right]^{-1} \quad (19-11)$
Temps moyen d'attente pour un client potentiel non servi immédiatement	$\bar{t}_a = \frac{1}{M\mu - \lambda} \quad (19-12)$
Probabilité qu'un client potentiel attende avant d'être servi	$P_W = \frac{\bar{t}_l}{\bar{t}_a} \quad (19-13)$

TABLEAU 19.4

Valeurs de \bar{n}_l et de P_0 pour des valeurs de λ/μ et de M données

λ/μ	M	\bar{n}_l	P_0	λ/μ	M	\bar{n}_l	P_0	λ/μ	M	\bar{n}_l	P_0
0,15	1	0,026	,850	1,3	3	0,130	,264	2,7	3	7,354	,025
	2	0,001	,860		4	0,023	,271		4	0,811	,057
0,20	1	0,050	,800	1,4	5	0,004	,272	5	5	0,198	,065
	2	0,002	,818		2	1,345	,176		6	0,053	,067
0,25	1	0,083	,750	1,4	3	0,177	,236	7	0,014	,067	
	2	0,004	,778		4	0,032	,245		8	12,273	,016
0,30	1	0,129	,700	1,5	5	0,006	,246	4	1,000	,050	
	2	0,007	,739		2	1,929	,143		5	0,241	,058
0,35	1	0,188	,650	1,6	3	0,237	,211	6	0,066	,060	
	2	0,011	,702		4	0,045	,221		7	0,018	,061
0,40	1	0,267	,600	1,6	5	0,009	,223	2,9	3	27,193	,008
	2	0,017	,667		2	2,844	,111		4	1,234	,044
0,45	1	0,368	,550	1,6	3	0,313	,187	5	5	0,293	,052
	2	0,024	,633		4	0,060	,199		6	0,081	,054
0,50	1	0,500	,400	1,6	5	0,012	,201	7	7	0,023	,055

TABLEAU 19.4

(suite)

λ/μ	M	\bar{n}_I	P_0	λ/μ	M	\bar{n}_I	P_0	λ/μ	M	\bar{n}_I	P_0
0,50	1	0,500	,500	1,7	2	4,426	,081	3,0	4	1,528	,038
	2	0,033	,600		3	0,409	,166		5	0,354	,047
	3	0,003	,606		4	0,080	,180		6	0,099	,049
0,55	1	0,672	,450	1,8	5	0,017	,182	3,1	7	0,028	,050
	2	0,045	,569		2	7,674	,053		8	0,008	,050
	3	0,004	,576		3	0,532	,146		4	1,902	,032
0,60	1	0,900	,400	1,9	4	0,105	,162	3,2	5	0,427	,042
	2	0,059	,538		5	0,023	,165		6	0,120	,044
	3	0,006	,548		2	17,587	,026		7	0,035	,045
0,65	1	1,207	,350	2,0	3	0,688	,128	3,2	8	0,010	,045
	2	0,077	,509		4	0,136	,145		4	2,386	,027
	3	0,008	,521		5	0,030	,149		5	0,513	,037
0,70	1	1,633	,300	2,1	6	0,007	,149	3,3	6	0,145	,040
	2	0,098	,481		3	0,889	,111		7	0,043	,040
	3	0,011	,495		4	0,174	,130		8	0,012	,041
0,75	1	2,250	,250	2,2	5	0,040	,134	3,4	4	3,027	,023
	2	0,123	,455		6	0,009	,135		5	0,615	,033
	3	0,015	,471		3	1,149	,096		6	0,174	,036
0,80	1	3,200	,200	2,3	4	0,220	,117	3,4	7	0,052	,037
	2	0,152	,429		5	0,052	,121		8	0,015	,037
	3	0,019	,447		6	0,012	,122		4	3,906	,019
0,85	1	4,817	,150	2,4	3	1,491	,081	3,5	5	0,737	,029
	2	0,187	,404		4	0,277	,105		6	0,209	,032
	3	0,024	,425		5	0,066	,109		7	0,063	,033
0,90	1	8,100	,100	2,5	6	0,016	,111	3,5	8	0,019	,033
	2	0,229	,379		3	1,951	,068		4	5,165	,015
	3	0,030	,403		4	0,346	,093		5	0,882	,026
0,95	1	18,050	,050	2,6	5	0,084	,099	3,6	6	0,248	,029
	2	0,277	,356		6	0,021	,100		7	0,076	,030
	3	0,037	,383		5	0,105	,089		4	7,090	,011
1,0	1	0,004	,406	2,7	6	0,027	,090	3,6	5	1,055	,023
	2	0,333	,333		7	0,007	,091		6	0,295	,026
	3	0,045	,364		3	3,511	,045		7	0,019	,027
1,1	1	0,007	,367	2,8	4	0,533	,074	3,7	8	0,028	,027
	2	0,477	,290		5	0,130	,080		9	0,008	,027
	3	0,066	,327		6	0,034	,082		4	10,347	,008
1,2	1	0,011	,367	2,9	7	0,009	,082	3,8	5	1,265	,020
	2	0,675	,250		3	4,933	,035		6	0,349	,023
	3	0,094	,294		4	0,658	,065		7	0,109	,024
1,3	1	0,016	,300	3,0	5	0,161	,072	3,8	8	0,034	,025
	2	0,003	,301		6	0,043	,074		9	0,010	,025
	3	0,951	,212		7	0,011	,074		4	16,937	,005
3,8	5	1,519	,017	3,1	5	9,289	,004	5,3	8	0,422	,005
	6	0,412	,021		6	1,487	,008		9	0,155	,005
	7	0,129	,022		7	0,453	,009		10	0,057	,005
	8	0,041	,022		8	0,156	,010		11	0,021	,005
	9	0,013	,022		9	0,054	,010		12	0,007	,005
3,9	4	36,859	,002	3,2	10	0,018	,010	5,3	12	0,007	,005
	5	1,830	,015		5	13,382	,003		7	1,444	,004
	6	0,485	,019		6	1,752	,007		8	0,483	,004
	7	0,153	,020		7	0,525	,008		9	0,178	,004
	8	0,050	,020		8	0,181	,008		10	0,066	,004
3,9	9	0,016	,020		9	0,064	,009		11	0,024	,005

TABLEAU 19.4

(suite)

	λ/μ	M	\bar{n}_I	P_0		λ/μ	M	\bar{n}_I	P_0		λ/μ	M	\bar{n}_I	P_0
	4,0	5	2,216	,013		10	0,022	,009			12	0,009	,005	
		6	0,570	,017	4,8	5	21,641	,002	5,5		6	8,590	,002	
		7	0,180	,018		6	2,071	,006			7	1,674	,003	
		8	0,059	,018		7	0,607	,008			8	0,553	,004	
		9	0,019	,018		8	0,209	,008			9	0,204	,004	
	4,2	5	2,703	,011		9	0,074	,008			10	0,077	,004	
		6	0,668	,015		10	0,026	,008			11	0,028	,004	
		7	0,212	,016	4,9	5	46,566	,001			12	0,010	,004	
		8	0,070	,016		6	2,459	,005	5,6		6	11,519	,001	
		9	0,023	,017		7	0,702	,007			7	1,944	,003	
	4,2	5	3,327	,009		8	0,242	,007			8	0,631	,003	
		6	0,784	,013		9	0,087	,007			9	0,233	,004	
		7	0,248	,014		10	0,031	,007			10	0,088	,004	
		8	0,083	,015		11	0,011	,077			11	0,033	,004	
		9	0,027	,015	5,0	6	2,938	,005			12	0,012	,004	
		10	0,009	,015		7	0,810	,006	5,7		6	16,446	,001	
	4,3	5	4,149	,008		8	0,279	,006			7	2,264	,002	
		6	0,919	,012		9	0,101	,007			8	0,721	,003	
		7	0,289	,130		10	0,036	,007			9	0,266	,003	
		8	0,097	,013		11	0,013	,007			10	0,102	,003	
		9	0,033	,014	5,1	6	3,536	,004			11	0,038	,003	
		10	0,011	,014		7	0,936	,005			12	0,014	,003	
	4,4	5	5,268	,006		8	0,321	,006	5,8		6	26,373	,001	
		6	1,078	,010		9	0,117	,006			7	2,648	,002	
		7	0,337	,012		10	0,042	,006			8	0,823	,003	
		8	0,114	,012		11	0,015	,006			9	0,303	,003	
		9	0,039	,012	5,2	6	4,301	,003			10	0,116	,003	
		10	0,013	,012		7	1,081	,005			11	0,044	,003	
	4,5	5	6,862	,005		8	0,368	,005			12	0,017	,003	
		6	1,265	,009		9	0,135	,005	5,9		6	56,300	,000	
		7	0,391	,010		10	0,049	,005			7	3,113	,002	
		8	0,133	,011		11	0,017	,006			8	0,939	,002	
		9	0,046	,011	5,3	6	5,303	,003			9	0,345	,003	
		10	0,015	,011		7	1,249	,004			10	0,133	,003	

Exemple 4

La compagnie Taxi-Air possède sept taxis stationnés à l'aéroport de Dorval. Les statistiques de la compagnie indiquent que durant les heures tardives des jours ouvrables de la semaine, les clients se présentent pour prendre un taxi (selon un processus de Poisson) à une cadence moyenne de 6,6 clients à l'heure. Le service, quant à lui, suit une distribution exponentielle de 50 minutes en moyenne. Le service consiste à prendre un client à l'aéroport, à le conduire à destination et à revenir à l'aéroport pour se placer en file, dans l'attente d'autres clients. Déterminez chacune des mesures de performance présentées dans le tableau 19.3, ainsi que le taux d'utilisation du système.

Solution

$\lambda = 6,6$ clients à l'heure et $M = 7$ voitures (serveurs)

$$\mu = \frac{1 \text{ client/voyage}}{(50 \text{ min/voyage} \div 60 \text{ min/h})} = 1,2 \text{ client à l'heure}$$

À partir du tableau 19.4: en considérant $\lambda/\mu = 5,5$ et $M = 7$, on peut lire:

a) $\bar{n}_I = 1,674$

b) $P_0 = 0,003$

c) $\bar{t}_a = \frac{1}{M\mu} - \lambda = \frac{1}{7 \cdot 1,2} - 6,6 = 0,556$ heure ou 33,36 minutes

d) $\bar{t}_q = \frac{\bar{n}_q}{\lambda} = \frac{1,674}{6,6} = 0,2536$ heure ou 15,22 minutes, donc :

$$P_w = \frac{\bar{t}_q}{\bar{t}_a} = \frac{0,2536}{0,556} = 0,456; \quad \text{il y a } 45,6 \% \text{ de chances qu'un client potentiel attende avant d'être servi.}$$

e) $\rho = \frac{\lambda}{M\mu} = \frac{6,6}{7} (1,2) = 0,786$; le système est utilisé à 78,6 % de sa capacité.

Avec Excel, la solution de l'exemple 4 apparaît comme suit :

T19-3: Modèle de files d'attente avec serveurs multiples				
1				
3	Taux moyen d'arrivée	lambda =	6,6	
4	Nombre de serveurs	M =	7	
5	Taux moyen de service	mu =	1,2	
6	Nombre moyen de clients servis	r =	5,500	
7	Nombre moyen en file	\bar{n}_q =	1,674	
8	Nombre moyen de clients dans le système	\bar{n}_s =	7,174	
9	Temps moyen d'attente en file	\bar{t}_q =	0,254	
10	Temps moyen dans le système	\bar{t}_s =	1,087	
11	Taux d'utilisation du système	rho =	0,786	
12	P(zero unités dans le système)	P0 =	0,003	
13	Temps moyen d'attente (client potentiel)	t_a =	0,556	
14	P(d'attente d'un client potentiel)	Pw =	0,456	
15				

Le processus de résolution peut être inversé, c'est-à-dire que l'analyste peut déterminer la capacité requise pour satisfaire à des niveaux spécifiés de performance. L'exemple ci-dessous illustre cette approche.

La compagnie Taxi-Air envisage de desservir une nouvelle gare. Le taux moyen d'arrivée des clients à la gare est de 4,8 clients à l'heure et le taux de service (aller-retour) est de 1,5 client à l'heure. Combien de taxis seront nécessaires pour obtenir un temps d'attente moyen tolérable de 20 minutes ou moins?

$\lambda = 4,8$ clients à l'heure, $\mu = 1,5$ client à l'heure, $M = ?$

$$\rho = \frac{\lambda}{\mu} = \frac{4,8}{1,5} = 3,2$$

$\bar{t}_q = 20$ minutes ou 0,333 heure (attente moyenne désirée)

$\bar{n}_q = \lambda \times \bar{t}_q = 4,8 \times 0,333 = 1,6$ unité. Donc, le nombre moyen de clients qui attendent ne doit pas dépasser 1,6. À partir du tableau 19.4, avec $\lambda/\mu = 3,2$, $\bar{n}_q = 2,386$ pour $M = 4$ et $\bar{n}_q = 0,513$ pour $M = 5$.

Taxi-Air a besoin de 5 voitures pour obtenir 20 minutes comme temps d'attente moyen tolérable.

Exemple 5

Solution

19.6.1.5 Optimisation des files d'attente

Pour concevoir un système, on calcule et on compare le coût associé au niveau de service (capacité de service) et le coût d'attente des clients (coût encouru par l'entreprise en raison de l'attente des clients dans le système). Par exemple, lorsqu'on conçoit un quai de chargement pour un entrepôt, on étudie le coût du quai plus le coût des

équipes de chargement par rapport au coût associé à l'attente des camions (chargement et déchargement). Même chose pour le coût du mécanicien qui attend des outils devant un centre d'outillage : il doit être équilibré avec le coût du serveur du centre d'outillage. Dans le cas où les clients sont externes à l'entreprise (commerces de détail, par exemple), les coûts vont inclure les ventes perdues à cause du refus du client d'attendre, le coût associé à l'espace d'attente mis en place par l'entreprise et le coût associé à la congestion du système (perte de clients, vols à l'étalage, etc.). La capacité optimale de service (en général, le nombre de serveurs qui travaillent en parallèle) est celle qui permet de réduire le coût total de gestion de l'attente. Ce coût total est la somme du coût d'attente des clients et du coût de la capacité de service.

L'objectif est donc de minimiser le coût total.

$$\text{Coût total (CT)} = \text{coût d'attente (C}_a\text{)} + \text{coût de service (C}_s\text{)}$$

L'approche d'optimisation consiste à calculer le coût total du système en fonction de différentes valeurs correspondant au nombre de serveurs. Après un certain nombre d'itérations, on établit la capacité qui minimise le coût total. Comme la courbe représentant le coût total est en forme de U, le fait d'augmenter le nombre de serveurs va faire en sorte que le coût total va diminuer jusqu'à atteindre le minimum. À partir de là, le fait d'augmenter la capacité va plutôt engendrer une augmentation du coût total. C'est donc à ce point que se situe la capacité optimale.

Le coût d'attente se calcule en fonction du nombre moyen de clients dans le système. Cela n'est peut-être pas intuitivement évident et on serait plutôt tenté de considérer le temps moyen d'attente dans le système. Or, ce serait ne tenir compte que d'un seul client. Cela ne donnerait pas d'informations concernant le nombre de clients qui attendent pendant ce temps. Il est évident que le coût engendré par la présence de cinq clients en moyenne qui attendent va être moindre que celui d'en avoir neuf. Par conséquent, il est nécessaire de se concentrer sur le nombre de clients en attente. Par ailleurs, si on a en moyenne deux clients dans le système, cela équivaut à avoir exactement deux clients dans le système en tout temps, malgré le fait qu'en réalité, on aura à certains moments zéro, un, deux, trois clients ou plus dans le système.

Exemple 6

Les camions arrivent à un entrepôt durant les jours ouvrables de la semaine selon un processus de Poisson, à raison de 15 camions à l'heure. Les équipes de manutention déchargent 5 camions à l'heure ; le processus de service suit une distribution exponentielle. Le taux élevé de déchargement est dû au fait que le transport se fait par conteneurs, ce qui rend le processus plus facile. La mise en application de la dernière convention syndicale étant prévue pour très bientôt, le directeur de la logistique voudrait réexaminer son processus de chargement/déchargement, notamment le nombre de manutentionnaires requis au quai. Les nouveaux coûts sont le salaire d'un manutentionnaire, auquel s'ajoute le coût d'exploitation du quai, estimé à 100 dollars l'heure, alors que le coût d'attente d'un chauffeur et de son camion est estimé à 120 dollars l'heure.

À partir du tableau 19.4, on détermine \bar{n}_l en utilisant : $\lambda / \mu = 15 / 5 = 3$.

Solution

Taille de l'équipe	Coût de service $C_s = 100 \$ \times M$	Nombre moyen de clients dans le système	Coût d'attente $C_a = 120 \$ \times \bar{n}_l$	Coût total (CT)
4	400	$1,528 + 3 = 4,528$	543,36	943,36
5	500	$0,354 + 3 = 3,354$	402,48	902,48
6	600	$0,099 + 3 = 3,099$	371,88	971,88
7	700	$0,028 + 3 = 3,028$	363,36	1063,36

La configuration optimale est une équipe de manutentionnaires composée de cinq personnes. Puisque le coût total va continuer d'augmenter une fois le minimum atteint, il n'est pas nécessaire de calculer les coûts totaux correspondant à des équipes de huit personnes ou plus. On voit bien que le coût total correspondant à la solution optimale est de 902,48 dollars et qu'il ne cesse d'augmenter. Après le coût total de 902,48 dollars, on entame la phase ascendante de la courbe en U.

Remarque: Soulignons que lorsqu'on fait de l'optimisation, les coûts d'attente et de service sont des estimations, donc la solution optimale obtenue peut ne pas être la vraie. Le fait de calculer le coût total au cent près ou même au dollar près semble indiquer un degré élevé de précision, ce qui n'est pas corroboré par les estimations des coûts. Cela est également compliqué par le fait que les approximations des taux d'arrivée et de service par les distributions de Poisson et exponentielle peuvent être fausses. Une autre solution serait d'estimer les coûts par intervalles (par exemple, le coût d'attente des clients serait compris entre 40 et 50 dollars l'heure). Dans ce cas, on devrait calculer le coût total pour chacune des limites afin de vérifier si la solution optimale est affectée. Si oui, le gestionnaire devra décider s'il est nécessaire de faire des efforts supplémentaires pour obtenir plus de précision dans les estimations des coûts ou tout simplement choisir une des deux solutions optimales obtenues. Le gestionnaire choisira probablement cette dernière approche si les variations dans le coût total pour différents niveaux de capacité sont minimales par rapport aux solutions optimales obtenues.

19.6.1.6 Capacité maximale de la file d'attente

Un autre point important est à considérer: la capacité maximale de la file d'attente proprement dite, c'est-à-dire la longueur maximale en termes d'espace disponible. Théoriquement, dans le cas d'une population infinie, la file d'attente peut devenir indéfiniment longue, et l'espace disponible peut être insuffisant pour accueillir tous les clients. Par exemple, les clients qui arrivent pour laver leur automobile dans une station libre-service proviennent d'une population infinie, et l'espace disponible est limité au nombre de voitures qui peuvent attendre en file sans perturber la circulation. Par contre, dans le cas des voitures qui arrivent de l'État de New York et qui se présentent au contrôle frontalier de Lacolle, au Québec, la longueur de la file correspond à toute l'autoroute 87.

D'un point de vue pratique, on peut toujours déterminer la longueur de la file d'attente qui ne sera pas dépassée pour un certain pourcentage de temps spécifié. Par exemple, un analyste pourrait déterminer la longueur de la file qui ne sera pas dépassée 98 % ou 99 % du temps.

Pour fixer la longueur de la file d'attente, on utilise les équations suivantes :

$$n = \frac{\log K}{\log \rho} \text{ ou } \frac{\ln K}{\ln \rho} \text{ où } K = \frac{1 - \text{pourcentage spécifié}}{\bar{n}(1 - \rho)} \quad (19-14)$$

La valeur de n n'est généralement pas un nombre entier; il faudra donc arrondir le nombre. Cependant, en pratique, si la valeur de n est inférieure à 0,10 au-dessus du nombre entier le plus petit, on arrondit vers le bas. Par exemple, si $n = 15,2$, alors $n = 16$; si $n = 15,06$, alors $n = 15$, n étant le nombre d'unités à servir.

Déterminez la longueur maximale de la file permettant d'atteindre des niveaux de satisfaction de 95 % et de 98 %. Les caractéristiques du système sont :

$M = 2$, $\lambda = 8$ à l'heure, $\mu = 5$ à l'heure

$$\rho = \frac{\lambda}{\mu} = \frac{8}{5} = 1,6 \quad \text{et} \quad \rho = \frac{\lambda}{M\mu} = \frac{8}{2(5)} = 0,80$$

À partir du tableau 19.4, on obtient $\bar{n} = 2,844$ clients.

Exemple 7

Solution

Si on utilise la formule 19-4, on obtient, pour 95 % :

$$K = \frac{1 - \text{pourcentage spécifié}}{\bar{n}(1 - \rho)} = \frac{1 - 0,95}{2,844(1 - 0,80)} = 0,088$$

$$n = \frac{\ln K}{\ln \rho} = \frac{\ln 0,088}{\ln 0,80} = \frac{-2,4304}{-0,2231} \approx 10,89 \approx 11$$

Pour 98 % :

$$K = \frac{1 - 0,98}{2,844(1 - 0,80)} \approx 0,035$$

$$n = \frac{\ln 0,035}{\ln 0,80} = \frac{-3,352}{-0,2231} = 15,02 \approx 15$$

19.6.1.7 Modèle 4 : serveurs multiples et règles de priorité

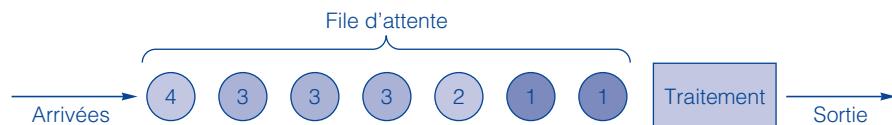


Dans la majorité des systèmes de files d'attente, particulièrement ceux des services, la règle de priorité pour le traitement des clients est la règle du premier entré, premier servi (PEPS). Cependant, dans plusieurs situations, cette règle est inapplicable, car le coût ou les conséquences qui en résultent ne sont pas les mêmes. Par exemple, dans les salles d'urgence des hôpitaux, où les clients sont malades ou accidentés, la rapidité de la prise en charge des patients dépend de la gravité de la situation. Certains patients peuvent être traités assez rapidement par l'infirmière, alors que d'autres, dont la vie est en danger, ont besoin de plusieurs intervenants. C'est pourquoi, dans les hôpitaux, il existe trois niveaux de priorité, qui vont de l'urgence (intervention immédiate) au cas le plus simple. Même chose pour le traitement des programmes à exécuter sur un ordinateur central, qui se fait selon la règle donnant la priorité au temps d'opération le plus court.

Ces exemples illustrent l'importance des modèles de files d'attente qui prennent en considération plusieurs règles de priorité.

Dans ces systèmes, on attribue aux clients qui se présentent une des **règles de priorité**² disponibles. Par règle de priorité, on entend l'ordre de traitement des clients (dans une salle d'urgence, une personne inconsciente ou ayant une crise cardiaque aura la priorité la plus élevée, celle qui a subi une blessure mineure aura la priorité la plus faible, et les autres auront une priorité intermédiaire). Ainsi, les clients sont classés par catégories en fonction de la règle de priorité qui leur est attribuée. Dans chaque classe ou catégorie, le traitement se fait selon la règle du premier entré, premier servi (PEPS), puisque les clients d'une même catégorie ont la même importance. Lorsque les clients d'une classe ont tous été servis, on passe à la classe inférieure. Si un client de la classe supérieure se présente, deux situations sont possibles, selon qu'il y a préséance ou non. S'il n'y a pas priorité, son traitement ne commence que lorsque le client en traitement a fini de se faire servir ; dans le cas contraire, il est traité immédiatement.

Quant aux hypothèses, ce sont les mêmes que celles du modèle 3 (serveurs multiples avec temps de service exponentiel), excepté que ce modèle utilise des règles de priorité de traitement autres que la règle PEPS. On attribue aux clients qui arrivent une priorité (priorité 1 à n). Une file d'attente organisée selon des règles de priorité aurait l'allure de celle qui est représentée ci-dessous :



Chaque client est traité selon la règle PEPS dans chacune des catégories. On commence par servir le client n° 1 de la classe 1, puis le n° 2 de la classe 1, puis le n° 1 de

2. Pour plus de détails sur ce sujet, voir le chapitre 17.

la classe 2, et ainsi de suite. À ce point, si un client de la classe 1 ou 2 se présente, on le placera devant le premier client de la classe 3. Si un client de la classe 4 se présente, il sera placé à la fin de la file, juste après le seul client de la classe 4. Il est évident que les clients dont la priorité est la moins élevée pourraient attendre assez longtemps, ce qui serait intolérable. Dans ce cas, on leur attribue une priorité plus élevée. Le tableau 19.5 donne les formules permettant de calculer les principales mesures de performance de ce modèle.

Mesure de performance	Formule	Référence
Taux d'utilisation	$\rho = \frac{\lambda}{M\mu}$	(19-15)
Mesures intermédiaires (\bar{n}_l à déterminer à partir du tableau 19.4)	$A = \frac{\lambda}{(1 - \rho)\bar{n}_l}$ $B_k = 1 - \sum_{o=1}^k \frac{\lambda}{M\mu}$ ($B_0 = 1$)	(19-16) (19-17)
Temps moyen d'attente en file pour les clients de la classe k (priorité k)	$\bar{t}_k = \frac{1}{A * B_{k-1} * B_k}$	(19-18)
Temps moyen d'attente dans le système pour les clients de la classe k (priorité k)	$\bar{t}_s = \bar{t}_k + \frac{1}{\mu}$	(19-19)
Nombre moyen de clients de la classe k (priorité k) qui attendent en file	$\bar{n}_k = \lambda_k * t_k$	(19-20)

TABLEAU 19.5

Formules pour le modèle avec règles de priorité multiples

Une entreprise dispose de son propre centre de maintenance, où sont réparés les équipements et les outils de l'entreprise. Chaque fois qu'un équipement ou qu'un outil arrive au centre, on y attribue une priorité en fonction de l'urgence du besoin. Le taux de demandes de réparations peut être établi avec une distribution de Poisson. Les taux d'arrivée sont : $\lambda_1 = 2$ à l'heure, $\lambda_2 = 2$ à l'heure, et $\lambda_3 = 1$ à l'heure. Le taux de service est de un équipement ou outil à l'heure par réparateur et il y a six réparateurs dans le centre de maintenance. Déterminez les mesures de performance suivantes :

- a) Le taux d'utilisation du système.

Pour chaque catégorie de priorité, déterminez :

- b) Le temps moyen d'attente pour la réparation.
c) Le temps moyen passé dans le système pour chaque équipement ou outil.
d) Le nombre moyen d'équipements ou d'outils en attente d'être réparés.

$$\lambda = \sum \lambda_k = 2 + 2 + 1 = 5 \text{ à l'heure}$$

$$M = 6 \text{ serveurs}$$

$$\mu = 1 \text{ client à l'heure}$$

a) $\rho = \frac{\lambda}{M\mu} = \frac{5}{6(1)} = 0,833$

b) Valeurs intermédiaires pour $\frac{\lambda}{\mu} = \frac{5}{1} = 5$; à partir du tableau 19.4, $\bar{n}_l = 2,938$

$$A = \frac{5}{(1 - 0,833)2,938} = 10,19$$

Exemple 8

Solution

$$B_0 = 1$$

$$B_1 = 1 - \frac{2}{6(1)} = \frac{2}{3} = 0,667$$

$$B_2 = 1 - \frac{2+2}{6(1)} = \frac{1}{3} = 0,333$$

$$B_3 = 1 - \frac{2+2+1}{6(1)} = \frac{1}{6} = 0,167$$

$$\bar{t}_1 = \frac{1}{A * B_0 * B_1} = \frac{1}{10,19(1)(0,667)} = 0,147 \text{ heure}$$

$$\bar{t}_2 = \frac{1}{A * B_1 * B_2} = \frac{1}{10,19(0,667)(0,333)} = 0,442 \text{ heure}$$

$$\bar{t}_3 = \frac{1}{A * B_2 * B_3} = \frac{1}{10,19(0,333)(0,167)} = 1,765 \text{ heure}$$

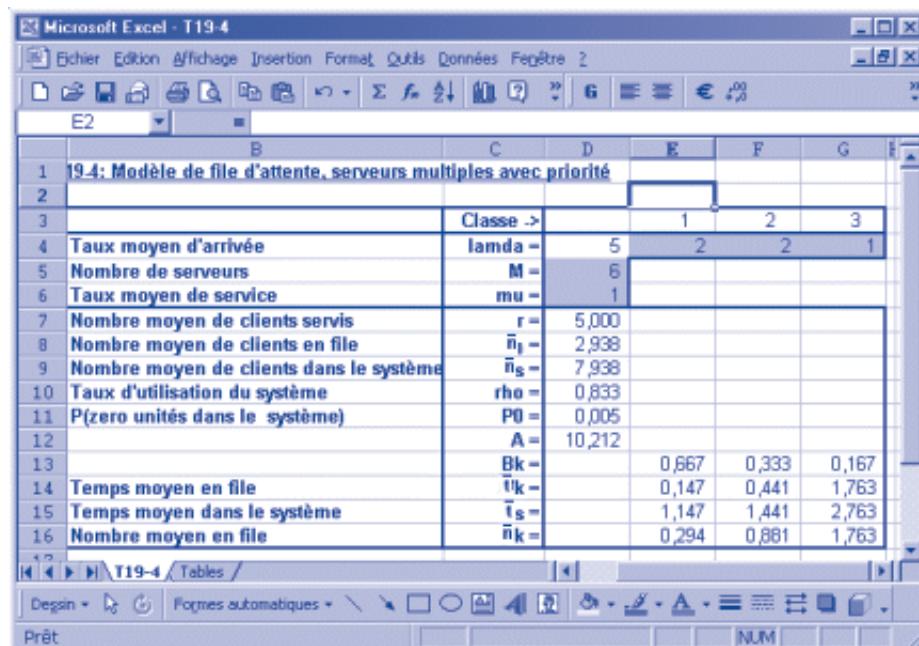
- c) Temps moyen dans le système = $\bar{t}_s = \bar{t}_k + \frac{1}{\mu}$; dans ce cas-ci, $\frac{1}{\mu} = \frac{1}{1} = 1$

Catégorie	$\bar{t}_s = \bar{t}_k + 1$
1	$0,147 + 1 = 1,147$
2	$0,442 + 1 = 1,442$
3	$1,765 + 1 = 2,765$

- d) Le nombre moyen d'unités qui attendent d'être réparées :

Catégorie	$\lambda_k \cdot \bar{t}_k = \bar{n}_k$
1	$2(0,147) = 0,294$
2	$2(0,442) = 0,884$
3	$1(1,765) = 1,765$

La solution de l'exemple 8 établie avec le tableur Excel est présentée ci-dessous :



Si les gestionnaires jugent trop longs les temps d'attente calculés dans l'exemple 8 (par exemple le temps moyen d'attente de 0,147 heure, soit environ 9 minutes pour les équipements de priorité 1), ils peuvent choisir d'autres options. L'une d'elles serait d'augmenter le nombre de serveurs. Une autre option serait d'essayer d'augmenter le taux de service, par exemple en introduisant de nouvelles méthodes de travail. Si toutes ces tentatives s'avèrent infructueuses, ils devraient revoir l'attribution de l'ordre de priorité et ramener certaines demandes de réparation de la classe de priorité 1 à la classe inférieure. Cela aura pour effet de diminuer le temps moyen d'attente de la classe de priorité 1, tout simplement parce que le taux d'arrivée aura diminué.

L'exemple 9 illustre des résultats intéressants concernant cette approche. On constate que la réduction du taux d'arrivée de la classe supérieure — grâce à l'attribution d'une cote de priorité inférieure à certains clients — a pour conséquence de réduire le temps moyen d'attente de cette classe. On constate aussi que le temps moyen d'attente de la classe inférieure a diminué, même si on a augmenté le nombre de clients de cette classe. Notez que le temps total d'attente (quand toutes les arrivées sont prises en considération) restera inchangé. On peut le vérifier en comparant le nombre moyen de clients qui attendent (exemple 8d) : $0,294 + 0,884 + 1,765 = 2,943$ avec le nombre d'unités en attente dans les trois classes, qui est (à partir des temps moyens d'attente de chaque classe de l'exemple 9) :

$$\sum_{k=1}^3 \lambda_k * \bar{t}_k = 1,5(0,131) + 2,5(0,393) + 1,0 (1,765) = 2,944$$

Les totaux sont pratiquement identiques, à part une différence négligeable due aux nombres arrondis.

On peut faire une autre observation intéressante : le temps moyen d'attente des clients de la troisième classe n'a pas changé par rapport à l'exemple précédent. Par conséquent, les unités ayant priorité la plus faible vont toujours être en compétition avec le taux d'arrivée combiné de 4 des deux autres classes supérieures.

Après avoir analysé les besoins de ses clients internes, le directeur de la logistique voudrait maintenant réviser la liste des outils classés dans la catégorie ayant la priorité la plus élevée. Ce besoin se traduit par la révision des taux d'arrivée. Les nouveaux taux sont : $\lambda_1 = 1,5$; $\lambda_2 = 2,5$; λ_3 reste inchangé, égal à 1. Déterminez les mesures de performance suivantes :

- a) Le taux d'utilisation du système.
- b) Le temps moyen d'attente pour chacune des classes.

$$\lambda = \sum \lambda_k = 1,5 + 2,5 + 1 = 5 \text{ à l'heure}$$

$M = 6$ serveurs

$\mu = 1$ client à l'heure

Notez que ces valeurs sont les mêmes que celles de l'exemple précédent.

- a) $\rho = \frac{\lambda}{M\mu} = \frac{5}{6(1)} = 0,833$, le même que dans l'exemple précédent.
- b) La valeur de A est la même que dans l'exemple précédent, puisqu'elle dépend de M, λ et μ ; donc $A = 10,19$.

$B_0 = 1$ (toujours)

$$B_1 = 1 - \frac{1,5}{6(1)} = 0,75$$

$$B_2 = 1 - 1,5 + \frac{2,5}{6(1)} = 0,333$$

$$B_3 = 1 - 1,5 + 2,5 + \frac{1,0}{6(1)} = 0,167$$

Exemple 9

Solution

Alors

$$\bar{t}_1 = \frac{1}{10,19(1)(0,75)} = 0,131 \text{ heure}$$

$$\bar{t}_2 = \frac{1}{10,19(0,75)(0,333)} = 0,393 \text{ heure}$$

$$\bar{t}_3 = \frac{1}{10,19(0,333)(0,167)} = 1,765 \text{ heure}$$

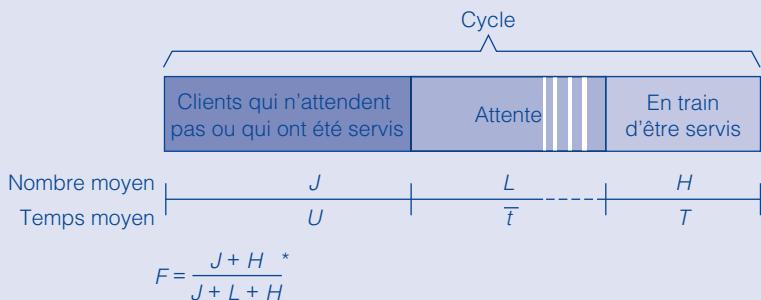
19.6.2 Modèle avec population finie



TABLEAU 19.6

Formules et notation pour le modèle de files d'attente avec population finie

Description	Formule	Référence	Notation
Facteur de service	$X = \frac{T}{(T + U)}$	(19-21)	D = Probabilité qu'un client potentiel attende en file F = Facteur d'efficience : $1 - \text{pourcentage d'attente en file}$
Nombre moyen de clients en attente	$L = N(1 - F)$	(19-22)	H = Nombre moyen de clients en train d'être servis J = Nombre moyen de clients qui ne sont pas en file ou en train d'être servis
Temps moyen d'attente	$\bar{t} = \frac{L(T + U)}{(N - L)}$ $= \frac{T(1 - F)}{XF}$	(19-23)	L = Nombre moyen de clients qui attendent d'être servis M = Nombre de serveurs
Nombre moyen de clients servis ou qui n'attendent pas	$J = NF(1 - X)$	(19-24)	N = Nombre de clients potentiels T = Temps moyen de service
Nombre moyen de clients servis	$H = FNX$	(19-25)	U = Temps moyen entre chaque demande de service \bar{t} = Temps moyen d'attente en file
Taille de la population	$N = J + L + H$	(19-26)	X = Facteur de service



*Le but de cette formule est de permettre de mieux comprendre F. Puisque la valeur de F est requise pour calculer J, L et H, les formules ne peuvent être utilisées pour calculer F. Les tableaux conçus pour les files d'attente avec population finie doivent être utilisés à cette fin.

Comme dans le cas des modèles avec population infinie, les processus d'arrivée et de service doivent respectivement suivre une distribution de Poisson et une distribution exponentielle. Il existe toutefois une différence majeure. Dans le cas d'une population finie, le taux d'arrivée est affecté par le nombre de clients qui attendent en file. Il diminue à mesure que le nombre de clients en attente augmente, tout simplement parce que si le nombre de clients en file augmente, la proportion de clients susceptibles de se présenter va diminuer, la majorité des clients étant en train d'attendre. Lorsque tous les clients (toute la population) sont en train d'attendre, le taux d'arrivée est forcément nul.

Pour analyser les systèmes de files d'attente avec population finie, on utilise une liste de formules clés et de définitions (tableau 19.6). Le graphique représentant un cycle a été ajouté pour une meilleure compréhension du modèle. Le tableau 19.7 est un tableau non exhaustif que l'on utilise pour déterminer les valeurs de D et F (la plupart des formules nécessitent la connaissance de F). Pour s'en servir, suivre la procédure suivante :

1. Noter les valeurs de :
 - a) N , la taille de la population ;
 - b) M , le nombre de serveurs ;
 - c) T , le temps moyen de service ;
 - d) U , le temps moyen entre chaque service.
2. Calculer le facteur de service $X = \frac{T}{(T + U)}$.
3. Localiser N sur le tableau.
4. En utilisant la valeur de X comme point de repère, déterminer les valeurs de D et de F qui correspondent à M .
5. En utilisant les valeurs de N , M , X , D et F , déterminer les mesures de performance désirées.

TABLEAU 19.7

Modèle avec population finie (valeurs de X , M , N , D , F)

X	M	D	F													
Population 5				1	0,229	0,984	0,085	2	0,040	0,998	1	0,473	0,920			
0,012	1	0,048	0,999	0,060	2	0,020	0,999		1	0,332	0,965	0,130	2	0,089	0,933	
0,019	1	0,076	0,998		1	0,237	0,983	0,090	2	0,044	0,998		1	0,489	0,914	
0,025	1	0,100	0,997	0,062	2	0,022	0,999		1	0,350	0,960	0,135	2	0,095	0,933	
0,030	1	0,120	0,996		1	0,245	0,982	0,095	2	0,049	0,997		1	0,505	0,907	
0,034	1	0,135	0,995	0,064	2	0,023	0,999		1	0,368	0,955	0,140	2	0,102	0,992	
0,036	1	0,143	0,994		1	0,253	0,981	0,100	2	0,054	0,997		1	0,521	0,900	
0,040	1	0,159	0,993	0,066	2	0,024	0,999		1	0,386	0,950	0,145	3	0,011	0,999	
0,042	1	0,167	0,992		1	0,260	0,979	0,105	2	0,059	0,997		2	0,109	0,991	
0,044	1	0,175	0,991	0,068	2	0,026	0,999		1	0,404	0,945		1	0,537	0,892	
0,046	1	0,183	0,990		1	0,268	0,978	0,110	2	0,065	0,996	0,150	3	0,012	0,999	
0,050	1	0,198	0,989	0,070	2	0,027	0,999		1	0,421	0,939		2	0,115	0,990	
0,052	1	0,206	0,988		1	0,275	0,977	0,115	2	0,017	0,995		1	0,553	0,885	
0,054	1	0,214	0,987	0,075	2	0,031	0,999		1	0,439	0,933	0,155	3	0,013	0,999	
0,056	2	0,018	0,999		1	0,294	0,973	0,120	2	0,076	0,995		2	0,123	0,989	
	1	0,222	0,985	0,080	2	0,035	0,998		1	0,456	0,927		1	0,568	0,877	
0,058	2	0,019	0,999		1	0,313	0,969	0,125	2	0,082	0,994	0,160	3	0,015	0,999	

TABLEAU 19.7

(suite)

X	M	D	F	X	M	D	F	X	M	D	F	X	M	D	F
	2	0,130	0,988	0,290	4	0,007	0,999		3	0,238	0,960		2	0,950	0,568
	1	0,582	0,869		3	0,079	0,992		2	0,652	0,807		1	0,999	0,286
0,165	3	0,016	0,999		2	0,362	0,932		1	0,969	0,451	0,750	4	0,316	0,944
	2	0,137	0,987		1	0,856	0,644	0,460	4	0,045	0,995		3	0,763	0,777
	1	0,597	0,861	0,300	4	0,008	0,999		3	0,266	0,953		2	0,972	0,532
0,170	3	0,017	0,999		3	0,086	0,990		2	0,686	0,787	0,800	4	0,410	0,924
	2	0,145	0,985		2	0,382	0,926		1	0,975	0,432		3	0,841	0,739
	1	0,611	0,853		1	0,869	0,628	0,480	4	0,053	0,994		2	0,987	0,500
0,180	3	0,021	0,999	0,310	4	0,009	0,999		3	0,296	0,945	0,850	4	0,522	0,900
	2	0,161	0,983		3	0,094	0,989		2	0,719	0,767		3	0,907	0,702
	1	0,683	0,836		2	0,402	0,919		1	0,980	0,415		2	0,995	0,470
0,190	3	0,024	0,998		1	0,881	0,613	0,500	4	0,063	0,992	0,900	4	0,656	0,871
	2	0,117	0,980	0,320	4	0,010	0,999		3	0,327	0,936		3	0,957	0,666
	1	0,665	0,819		3	0,103	0,988		2	0,750	0,748		2	0,998	0,444
0,200	3	0,028	0,998		2	0,422	0,912		1	0,985	0,399	0,950	4	0,815	0,838
	2	0,194	0,976		1	0,892	0,597	0,520	4	0,073	0,991		3	0,989	0,631
	1	0,689	0,801	0,330	4	0,012	0,999		3	0,359	0,927	Population 10			
0,210	3	0,032	0,998		3	0,112	0,986		2	0,779	0,728	0,016	1	0,144	0,997
	2	0,211	0,973		2	0,442	0,904		1	0,988	0,384	0,019	1	0,170	0,996
	1	0,713	0,783		1	0,902	0,583	0,540	4	0,085	0,989	0,021	1	0,188	0,995
0,220	3	0,036	0,997	0,340	4	0,013	0,999		3	0,392	0,917	0,023	1	0,206	0,994
	2	0,229	0,969		3	0,121	0,985		2	0,806	0,708	0,025	1	0,224	0,993
	1	0,735	0,765		2	0,462	0,896		1	0,991	0,370	0,026	1	0,232	0,992
0,230	3	0,041	0,997		1	0,911	0,569	0,560	4	0,098	0,986	0,028	1	0,250	0,991
	2	0,247	0,965	0,360	4	0,017	0,998		3	0,426	0,906	0,030	1	0,268	0,990
	1	0,756	0,747		3	0,141	0,981		2	0,831	0,689	0,032	2	0,033	0,999
0,240	3	0,046	0,996		2	0,501	0,880		1	0,993	0,357		1	0,285	0,988
	2	0,265	0,960		1	0,927	0,542	0,580	4	0,113	0,984	0,034	2	0,037	0,999
	1	0,775	0,730	0,380	4	0,021	0,998		3	0,461	0,895		1	0,301	0,986
0,250	3	0,052	0,995		3	0,163	0,976		2	0,854	0,670	0,036	2	0,041	0,999
	2	0,284	0,955		2	0,540	0,863		1	0,994	0,345		1	0,320	0,984
	1	0,794	0,712		1	0,941	0,516	0,600	4	0,130	0,981	0,038	2	0,046	0,999
0,260	3	0,058	0,994	0,400	4	0,026	0,997		3	0,497	0,883		1	0,337	0,982
	2	0,303	0,950		3	0,186	0,972		2	0,875	0,652	0,040	2	0,050	0,999
	1	0,811	0,695		2	0,579	0,845		1	0,996	0,333		1	0,354	0,980
0,270	3	0,064	0,994		1	0,952	0,493	0,650	4	0,179	0,972	0,042	2	0,055	0,999
	2	0,323	0,944	0,420	4	0,031	0,997		3	0,588	0,850		1	0,371	0,978
	1	0,827	0,677		3	0,211	0,966		2	0,918	0,608	0,044	2	0,060	0,998
0,280	3	0,071	0,993		2	0,616	0,826		1	0,998	0,308		1	0,388	0,975
	2	0,342	0,938		1	0,961	0,471	0,700	4	0,240	0,960	0,046	2	0,065	0,998
	1	0,842	0,661	0,440	4	0,037	0,996		3	0,678	0,815		1	0,404	0,973

TABLEAU 19.7

(suite)

X	M	D	F												
0,048	2	0,071	0,998	0,100	3	0,056	0,998		3	0,169	0,987		4	0,142	0,988
	1	0,421	0,970		2	0,258	0,981		2	0,505	0,928		3	0,400	0,947
0,050	2	0,076	0,998		1	0,776	0,832		1	0,947	0,627		2	0,791	0,794
	1	0,437	0,967	0,105	3	0,064	0,997	0,160	4	0,044	0,998		1	0,995	0,434
0,052	2	0,082	0,997		2	0,279	0,978		3	0,182	0,986	0,240	5	0,044	0,997
	1	0,454	0,963		1	0,800	0,814		2	0,528	0,921		4	0,162	0,986
0,054	2	0,088	0,997	0,110	3	0,072	0,997	0,165	4	0,049	0,997		3	0,434	0,938
	1	0,470	0,960		2	0,301	0,974	0,165	4	0,049	0,997		2	0,819	0,774
0,056	2	0,094	0,997		1	0,822	0,795		3	0,195	0,984		1	0,996	0,416
	1	0,486	0,956	0,115	3	0,081	0,996		2	0,550	0,914	0,250	6	0,010	0,999
0,058	2	0,100	0,996		2	0,324	0,971		1	0,961	0,594		5	0,052	0,997
	1	0,501	0,953		1	0,843	0,776	0,170	4	0,054	0,997		4	0,183	0,983
0,060	2	0,106	0,996	0,120	4	0,016	0,999		3	0,209	0,982		3	0,469	0,929
	1	0,517	0,949		3	0,090	0,995		2	0,571	0,906		2	0,844	0,753
0,062	2	0,113	0,996		2	0,346	0,967		1	0,966	0,579		1	0,997	0,400
	1	0,532	0,945		1	0,861	0,756	0,180	5	0,013	0,999	0,260	6	0,013	0,999
0,064	2	0,119	0,995	0,125	4	0,019	0,999		4	0,066	0,996		5	0,060	0,996
	1	0,547	0,940		3	0,100	0,994		3	0,238	0,978		4	0,205	0,980
0,066	2	0,126	0,995		2	0,369	0,962		2	0,614	0,890		3	0,503	0,919
	1	0,562	0,936		1	0,878	0,737		1	0,975	0,890		2	0,866	0,732
0,068	3	0,020	0,999	0,130	4	0,022	0,999	0,190	5	0,016	0,999		1	0,998	0,384
	2	0,133	0,994		3	0,110	0,994		4	0,078	0,995	0,270	6	0,015	0,999
	1	0,577	0,931		2	0,392	0,958		3	0,269	0,973		5	0,070	0,995
0,070	3	0,022	0,999		1	0,893	0,718		2	0,654	0,873		4	0,228	0,976
	2	0,140	0,994	0,135	4	0,025	0,999		1	0,982	0,522		3	0,537	0,908
	1	0,591	0,926		3	0,121	0,993	0,200	5	0,020	0,999		2	0,886	0,712
0,075	3	0,026	0,999		2	0,415	0,952		4	0,092	0,994		1	0,999	0,370
	2	0,158	0,992		1	0,907	0,699		3	0,300	0,968	0,280	6	0,018	0,999
	1	0,627	0,913	0,140	4	0,028	0,999		2	0,692	0,854		5	0,081	0,994
0,080	3	0,031	0,999		3	0,132	0,991		1	0,987	0,497		4	0,252	0,972
	2	0,177	0,990		2	0,437	0,947	0,210	5	0,025	0,999		3	0,571	0,896
	1	0,660	0,899		1	0,919	0,680		4	0,108	0,992		2	0,903	0,692
0,085	3	0,037	0,999	0,145	4	0,032	0,999		3	0,333	0,961		1	0,999	0,357
	2	0,196	0,988		3	0,144	0,990		2	0,728	0,835	0,290	6	0,022	0,999
	1	0,692	0,883		2	0,460	0,941		1	0,990	0,474		5	0,093	0,993
0,090	3	0,043	0,998		1	0,929	0,662	0,220	5	0,030	0,998		4	0,278	0,968
	2	0,216	0,986	0,150	4	0,036	0,998		4	0,124	0,990		3	0,603	0,884
	1	0,722	0,867		3	0,156	0,989		3	0,366	0,954		2	0,918	0,672
0,095	3	0,049	0,998		2	0,483	0,935		2	0,761	0,815		1	0,999	0,345
	2	0,237	0,984		1	0,939	0,644		1	0,993	0,453	0,300	6	0,026	0,998
	1	0,750	0,850	0,155	4	0,040	0,998	0,230	5	0,037	0,998		5	0,106	0,991

TABLEAU 19.7

(suite)

X	M	D	F												
0,310	4	0,304	0,963	0,400	7	0,026	0,998	0,520	3	0,972	0,598	0,700	6	0,651	0,878
	3	0,635	0,872		6	0,105	0,991		2	0,999	0,400		5	0,882	0,759
	2	0,932	0,653		5	0,292	0,963		8	0,026	0,998		4	0,980	0,614
	1	0,999	0,333		4	0,591	0,887		7	0,115	0,989		3	0,999	0,461
	6	0,031	0,998		3	0,875	0,728		6	0,316	0,958		9	0,040	0,997
	5	0,120	0,990		2	0,991	0,499		5	0,606	0,884		8	0,200	0,979
	4	0,331	0,957	0,420	7	0,034	0,993		4	0,864	0,752		7	0,484	0,929
	3	0,666	0,858		6	0,130	0,987		3	0,980	0,575		6	0,772	0,836
	2	0,943	0,635		5	0,341	0,954		2	0,999	0,385		5	0,940	0,711
	6	0,036	0,998		4	0,646	0,866	0,540	8	0,034	0,997		4	0,992	0,571
0,320	5	0,135	0,988		3	0,905	0,700		7	0,141	0,986	0,750	9	0,075	0,994
	4	0,359	0,952		2	0,994	0,476		6	0,363	0,949		8	0,307	0,965
	3	0,695	0,845	0,440	7	0,045	0,997		5	0,658	0,867		7	0,626	0,897
	2	0,952	0,617		6	0,160	0,984		4	0,893	0,729		6	0,870	0,792
	6	0,042	0,997		5	0,392	0,943		3	0,986	0,555		5	0,975	0,666
	5	0,151	0,986		4	0,698	0,845	0,560	8	0,044	0,996		4	0,998	0,533
	4	0,387	0,945		3	0,928	0,672		7	0,171	0,982	0,800	9	0,134	0,988
	3	0,723	0,831		2	0,996	0,454		6	0,413	0,939		8	0,446	0,944
	2	0,961	0,600	0,460	8	0,011	0,999		5	0,707	0,848		7	0,763	0,859
	7	0,010	0,999		7	0,058	0,995		4	0,917	0,706		6	0,939	0,747
0,340	6	0,049	0,997		6	0,193	0,979		3	0,991	0,535		5	0,991	0,625
	5	0,168	0,983		5	0,445	0,930	0,580	8	0,057	0,995		4	0,999	0,500
	4	0,416	0,938		4	0,747	0,822		7	0,204	0,977	0,850	9	0,232	0,979
	3	0,750	0,816		3	0,947	0,646		6	0,465	0,927		8	0,611	0,916
	2	0,968	0,584		2	0,998	0,435		5	0,753	0,829		7	0,879	0,818
	7	0,014	0,999	0,480	8	0,015	0,999		4	0,937	0,684		6	0,978	0,705
	6	0,064	0,995		7	0,074	0,994		3	0,994	0,517		5	0,998	0,588
	5	0,205	0,978		6	0,230	0,973	0,600	9	0,010	0,999	0,900	9	0,387	0,963
	4	0,474	0,923		5	0,499	0,916		8	0,072	0,994		8	0,785	0,881
	3	0,798	0,787		4	0,791	0,799		7	0,242	0,972		7	0,956	0,777
0,380	2	0,978	0,553		3	0,961	0,621		6	0,518	0,915		6	0,995	0,667
	7	0,019	0,999		2	0,998	0,417		5	0,795	0,809	0,950	9	0,630	0,938
	6	0,083	0,993	0,500	8	0,020	0,999		4	0,953	0,663		8	0,934	0,841
	5	0,247	0,971		7	0,093	0,992		3	0,996	0,500		7	0,994	0,737
	4	0,533	0,906		6	0,271	0,966	0,650	9	0,021	0,999				
	3	0,840	0,758		5	0,553	0,901		8	0,123	0,988				
	2	0,986	0,525		4	0,830	0,775		7	0,353	0,954				

Source : L. G. Peck et R. N. Hazelwood, *Finite Queuing Tables*, New York, John Wiley & Sons, 1958. Reproduit avec autorisation.

Un opérateur est responsable du chargement et du déchargement de cinq machines. Le temps de service est distribué, selon une loi exponentielle, à raison de 10 minutes en moyenne par machine et par cycle (un cycle correspond à la période de fonctionnement de la machine + le temps d'attente pour le service + le temps de service). Les machines fonctionnent pendant 70 minutes en moyenne entre chaque chargement et déchargement; ce temps est aussi distribué selon une loi exponentielle. Déterminez:

- Le nombre moyen de machines qui attendent l'opérateur.
- Le nombre moyen de machines qui fonctionnent.
- Le temps moyen d'arrêt des machines.
- La probabilité qu'une machine n'attende pas pour le service.

$$N = 5$$

$$T = 10 \text{ minutes}$$

$$M = 1$$

$$U = 70 \text{ minutes}$$

$$X = \frac{T}{T+U} = \frac{10}{10+70} = 0,125$$

À partir du tableau 19.7, avec $N = 5$, $M = 1$, $X = 0,125$, $D = 0,473$ et $F = 0,920$:

- Nombre moyen de machines en attente: $L = N(1 - F) = 5(1 - 0,920) = 0,40$ machine
- Nombre moyen de machines en marche: $J = NF(1 - X) = 5(0,92)(1 - 0,125) = 4,025$ machines
- Temps moyen d'arrêt: temps moyen d'attente + temps moyen de service

$$\bar{t} = \frac{L(T+U)}{N-L} = \frac{0,40(10+70)}{5-0,40} = 6,957 \text{ minutes}$$

$$\text{Temps moyen d'arrêt} = 6,957 \text{ minutes} + 10 \text{ minutes} = 16,957 \text{ minutes}$$

- Probabilité de ne pas attendre = 1 – probabilité d'attendre
 $= 1 - D = 1 - 0,473 = 0,527$

Supposez maintenant que les opérateurs sont payés 10 \$ l'heure et que 1 heure d'arrêt des machines coûte à l'entreprise 16 \$ par machine. Est-ce qu'on devrait rajouter un opérateur, alors que l'objectif visé est la minimisation des coûts?

Comparez le coût total de la situation actuelle à celle qui est proposée.

Exemple 10

Solution

M	Nombre moyen d'unités inoccupées <i>N - J</i>	Coût moyen d'inoccupation <i>(N - J) × 16,00 \$</i>	Coût horaire (opérateurs)	Coût total
1	0,975	15,60 \$	10,00 \$	25,60 \$
2	0,651	10,42 \$	20,00 \$	30,42 \$

Si le critère de choix est la minimisation du coût, il est préférable de garder le système actuel, car il est moins coûteux.

Exemple 11

Solution



19.7 AUTRES APPROCHES D'ANALYSE

Dans ce chapitre, nous avons mis l'accent sur la conception de systèmes basée sur le coût de service et le coût d'attente des clients. Cela implique que le gestionnaire peut déterminer le niveau de service approprié (en termes de capacité). Mais dans certaines

situations, cette approche n'est pas applicable, et ce, pour diverses raisons. L'une d'elles est que le système est déjà en activité et que les changements suggérés ne sont pas réalisables parce qu'ils sont trop coûteux ou que l'espace disponible est une contrainte majeure. La solution est d'avoir recours à une forme quelconque de distraction, de telle sorte que l'attente devienne plus tolérable pour les clients. Par exemple, des journaux et des magazines peuvent être mis à la disposition des clients, comme c'est le cas chez les médecins et les dentistes. Les garages ont installé des radios, des télévisions et des machines à café dans leurs salles d'attente; les compagnies aériennes offrent des repas et des boissons pour rendre les vols plus agréables, et projettent aussi des films pour faire passer le temps plus vite.

D'autres mesures consistent à mettre des miroirs près des ascenseurs ou bien à demander aux clients de remplir des formulaires, ce qui rend l'attente plus agréable.

De plus, l'aménagement a aussi un effet sur la réaction face à l'attente. Prenons l'exemple de la Société de l'assurance automobile du Québec (SAAQ). Elle a décidé de ne pas avoir recours à de « vraies » files d'attente : les clients qui viennent pour obtenir divers services (renouvellement du permis de conduire, tests, etc.) sont tous assis. La SAAQ met à leur disposition des numéros qui correspondent à des services. Cela permet aux clients de constater que malgré le nombre élevé de personnes présentes, leur attente sera minime, puisque qu'elle dépend du numéro attribué.

Dans certaines situations, on peut tirer profit de l'attente. Les supermarchés installent tout près des caisses des articles qui sont généralement achetés de manière impulsive ; les banques affichent les taux d'intérêts et placent des brochures publicitaires à portée des clients ; les restaurants ont généralement des bars où les clients peuvent consommer en attendant d'être dirigés vers leur table.

En résumé, l'imagination et la créativité sont importantes pour quiconque veut concevoir un système et gérer l'attente de façon optimale. On ne devrait pas tenir compte uniquement des approches mathématiques.

Quelques recommandations pour gérer les files d'attente³:

- *Déterminer un temps d'attente tolérable pour les clients.* Combien de temps vos clients peuvent-ils attendre? Fixez vos objectifs en fonction de ce qui est acceptable.
- *Essayer de divertir les clients pendant l'attente.* Musique, café, magazines, télévision sont autant de sources de distraction qui font patienter les clients.
- *Informier les clients de la durée de l'attente.* Ce point est particulièrement important lorsque l'attente risque d'être longue. Expliquez aux clients pourquoi l'attente est anormalement longue, et ce que vous êtes en train de faire pour y remédier.
- *Éloigner les employés visibles qui ne sont pas concernés par le service.* Il n'y a rien de plus frustrant pour une personne qui attend en file que de voir un employé occupé à faire autre chose que de venir répondre aux clients qui attendent.
- *Segmenter la clientèle.* Si un groupe de clients peut être servi rapidement, créez une file d'attente spéciale pour ne pas les faire attendre plus que nécessaire.
- *Former et sensibiliser le personnel à la gentillesse.* En plus du sourire quotidien et de l'accueil chaleureux, et même personnalisé, le personnel doit être capable d'affronter les situations difficiles et de réagir de manière à détendre l'atmosphère lorsque les clients s'impatientent.
- *Encourager les clients à venir durant les périodes mortes.* Informez les clients sur les périodes moins achalandées. Le directeur d'une succursale bancaire de ville Saint-Laurent a proposé à ses clients âgés de venir le mercredi, la journée la moins chargée, afin qu'il puisse leur consacrer plus de temps.

3. KATZ, K. L., B. M. LARSON et R. C. LARSON. « Prescriptions for the Waiting-in-Line Blues: Entertain, Enlighten and Engage », *Sloan Management Review*, vol. 32, n° 2, hiver 1991, p. 44-53.

- Avoir une vision à long terme concernant la gestion de l'attente.* Mettez en place un processus d'amélioration continue concernant la réduction de l'attente. Réfléchissez sur les moyens d'accélérer le processus de traitement des clients. Automatisez lorsque cela est possible sans pour autant éliminer le contact personnalisé. On a toujours besoin d'un peu d'attention.

19.8 Conclusion

L'analyse des files d'attente peut être un aspect important de la conception des systèmes. Les files d'attente ont tendance à se former, bien que, d'un point de vue macro, les systèmes ne soient pas congestionnés. Les arrivées aléatoires des clients combinées à la variabilité des temps de service créent temporairement des congestions dans le système, d'où la création de files d'attente. Dans certains cas, il arrive également que les serveurs soient inactifs.

Pour analyser des files d'attente, il est important de définir si la population de clients potentiels est infinie ou bien si elle se limite à un nombre fini de clients. Il existe cinq modèles de base pour analyser les files d'attente; quatre pour une population infinie et un pour une population finie. En général, les hypothèses émises dans le cadre de ces modèles sont que les taux d'arrivée des clients sont distribués selon une loi de Poisson, alors que les temps de service suivent une loi exponentielle.

Terminologie

Discipline de la file d'attente	Serveur
Modèle avec règles de priorité	Temps inter-arrivées
Population finie	Théorie des files d'attente
Population infinie	

Problèmes résolus

Le directeur de la logistique voudrait déterminer le nombre de magasiniers à affecter au magasin nouvellement implanté dans l'usine qui fournit les ouvriers en outils et en pièces. Les magasiniers reçoivent un salaire de 9 \$ l'heure (incluant les avantages sociaux). Une heure de travail d'un ouvrier (généralement des mécaniciens) est évaluée à 30 \$, ce qui inclut les avantages sociaux ainsi que le temps perdu à attendre les outils ou les pièces. Par expérience, le directeur de la logistique estime que les demandes des ouvriers sont de l'ordre de 18 à l'heure, alors que la capacité de service est de 20 demandes à l'heure par magasinier. Combien de magasiniers devrait-on affecter au magasin, si on suppose que les taux d'arrivée et de service sont distribués selon une loi de Poisson? (Hypothèse: le nombre d'ouvriers qui se présentent au magasin est très élevé.)

$$\lambda = 18 \text{ à l'heure}$$

$$\mu = 20 \text{ à l'heure}$$

On obtient la solution par essais et erreurs en calculant le coût total correspondant à une solution réalisable (c'est-à-dire avec un taux d'utilisation inférieur à 100 %) et en choisissant la solution comportant le coût total le plus faible. Notez que la courbe représentant le coût total est en forme de U; on augmente donc le nombre de serveurs jusqu'à ce que cette augmentation donne une augmentation du coût total par rapport à la solution précédente. La solution optimale sera la solution précédente.

Le tableau ci-dessous résume les calculs requis.

Problème 1

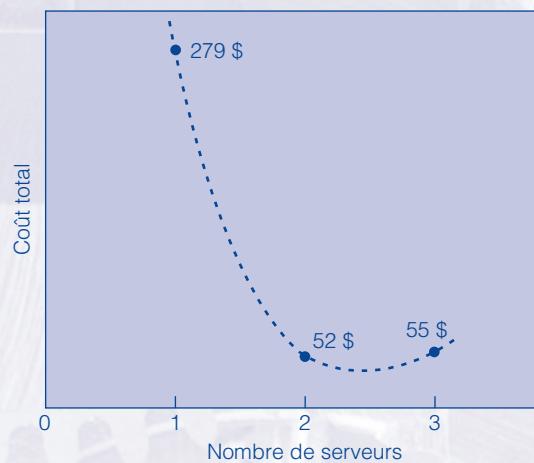
Solution

Nombre de serveurs (M)	Nombre moyen de clients dans le système		Coût de service $C_s = 9,00 \$ \times M$	Coût d'attente $C_a = 30 \$ \times \bar{n}_s$	Coût total (CT) par heure
	\bar{n}_l^*	$\bar{n}_s = \bar{n}_l + \frac{\lambda}{\mu}$			
1	8,1	$8,1 + 0,9 = 9,00$	9,00 \$	270,00 \$	279,00 \$
2	0,229	$0,229 + 0,9 = 1,129$	18,00 \$	33,87 \$	52,00 \$**
3	0,03	$0,03 + 0,9 = 0,93$	27,00 \$	27,90 \$	55,00 \$**

* \bar{n}_l est déterminé à partir du tableau 19.4, avec $\rho = \lambda / \mu = 18 / 20 = 0,90$.

** Valeurs arrondies.

D'après les calculs, il faudra deux magasiniers.



Problème 2

Le tableau ci-dessous présente les temps de service de trois opérations différentes :

Opération	Temps de service
A	8 minutes
B	1,2 heure
C	2 jours

- Déterminez les taux de service de chaque opération.
- Est-ce que les taux seraient différents si les temps de service étaient plutôt des temps inter-arrivées ?
- $\mu_A = 1/8 \text{ par minute} = 0,125 \text{ par minute ou } 0,125 / \text{min} \times 60 \text{ min/h} = 7,5 \text{ à l'heure}$
 $\mu_B = 1/1,2 \text{ à l'heure} = 0,833 \text{ à l'heure}$
 $\mu_C = 1/2 \text{ par jour} = 0,50 \text{ par jour}$
- Non, car dans les deux cas, il y a une équivalence entre le taux de service et le temps de service (taux de service = 1 / temps de service).

Solution

Problème 3

Un groupe de 10 machines est chargé et déchargé par 3 opérateurs. Les machines sont en marche 6 minutes en moyenne par cycle, alors que le temps moyen requis pour charger et décharger est de 9 minutes. Les temps suivent une distribution exponentielle. Lorsque les machines sont en marche, elles produisent à un taux de 16 unités à l'heure. Quel est le débit horaire moyen de chaque machine, quand l'attente et le service sont pris en considération ?

$$T = 9 \text{ minutes et } U = 6 \text{ minutes, donc } X = \frac{T}{T+U} = \frac{9}{9+6} = 0,60$$

$N = 10$ machines et $M = 3$ serveurs ; à partir du tableau 19.7, $F = 0,500$

- Le nombre moyen de machines en marche est :
 $J = NF(1 - X) = 10(0,500)(0,40) = 2$
- On détermine le pourcentage de machines en marche et on multiplie par le taux de production pour trouver le débit horaire de chaque machine.

$$\frac{J}{N} \times (16 \text{ unités à l'heure}) = \frac{2}{10} \times (16 \text{ unités à l'heure}) = 3,2 \text{ unités à l'heure}$$

Questions de discussion et de révision

1. Dans quelles situations l'analyse des files d'attente est-elle appropriée?
2. Expliquez pourquoi des files d'attente se forment même si le système n'est pas congestionné.
3. Énumérez les principales mesures de performance utilisées dans le cadre de l'analyse des files d'attente.
4. Quel est l'effet de la réduction de la variabilité dans les processus d'arrivée et de service sur la capacité effective d'un système?
5. Quelles sont les approches utilisées par les supermarchés pour contrecarrer les variations du trafic de la clientèle?
6. Expliquez la différence entre la population finie et la population infinie.
7. Est-ce que le fait de doubler le taux de service dans un système de files d'attente à serveur unique aura pour effet de réduire de moitié le temps moyen d'attente en file? Justifiez.
8. Expliquez la raison pour laquelle, dans les systèmes de files d'attente à serveurs multiples (par exemple dans les banques), on utilise une file d'attente unique plutôt que plusieurs files d'attente.
9. Dans un système de file d'attente à variabilité élevée, comment peut-on, du point de vue du nombre de clients qui attendent en file, atteindre un pourcentage élevé d'utilisation de la capacité?
10. En une dizaine de lignes au maximum, expliquez à votre directeur d'usine, Gaëtan Tremblay, les coûts et les bénéfices associés à deux options possibles concernant l'installation d'un magasin d'outillage dans l'usine. La première option consiste à installer un magasin central; la deuxième, à installer un magasin à chaque extrémité de l'usine⁴.
11. Expliquez à votre responsable du service à la clientèle, M. Jean Dumalheur, pourquoi il serait préférable d'utiliser un modèle de files d'attente avec règles de priorité pour gérer les plaintes des clients⁵.

Problèmes

1. Un service de maintenance de photocopieurs est sous la responsabilité d'un réparateur. Le temps de réparation, incluant le déplacement chez le client, est distribué selon une loi exponentielle et est de deux heures en moyenne. Les statistiques indiquent que les appels de demandes de service enregistrés sont au nombre de trois appels en moyenne par quart de travail de huit heures (selon une distribution de Poisson). Déterminez:
 - a) Le nombre moyen de photocopieurs qui attendent d'être réparés.
 - b) Le taux d'utilisation du système.
 - c) Le temps pendant le quart de travail où le réparateur ne travaille pas.
 - d) La probabilité qu'il y ait deux photocopieurs ou plus dans le système.
2. Le processus de préparation du café dans une machine automatique prend 30 secondes par tasse. Les clients se présentent devant la machine à un rythme de 80 à l'heure, selon un processus de Poisson.
Déterminez:
 - a) Le nombre moyen de clients en attente devant la machine.
 - b) Le temps moyen que passent les clients dans le système.
 - c) Le nombre moyen de clients dans le système.
3. Les guichets automatiques sont, de nos jours, de plus en plus utilisés par les clients, surtout depuis que les banques ont réduit leurs heures d'ouverture. En début de soirée, l'été, les clients se présentent au guichet automatique d'une des succursales de l'ouest de l'île de Montréal au taux moyen de un client toutes les deux minutes (selon la loi de Poisson). Les transactions durent en moyenne 90 secondes. Ce temps étant distribué selon une loi exponentielle, déterminez:

4. Voir la solution dans l'ouvrage de BOURENANE, H. et W. STEVENSON. *Guide de l'étudiant*, Montréal, Chenelière/McGraw-Hill, 2001.

5. *Ibid.*

- a) Le temps moyen que passent les clients dans le système.
 b) La probabilité qu'un client potentiel n'ait pas à attendre lorsqu'il se présente au guichet automatique.
 c) Le temps moyen d'attente des clients devant le guichet automatique.
4. Le service ambulancier de la Cité de la Santé de Laval dispose de deux ambulances. Les demandes d'ambulances, pendant les fins de semaines, arrivent à un rythme moyen de 0,8 appel à l'heure et peuvent être prévues grâce à une distribution de Poisson. La durée moyenne des secours, incluant le déplacement, est d'environ une heure par appel. Le temps d'assistance et de déplacement étant distribué selon une loi exponentielle, déterminez:
- a) Le taux d'utilisation du système.
 b) Le nombre moyen de clients qui attendent.
 c) Le temps moyen d'attente des clients pour l'ambulance.
 d) La probabilité que les deux ambulances soient occupées lors d'un appel.
5. Les informations fournies dans le tableau ci-dessous concernent les appels téléphoniques adressés au standard d'un motel pendant la journée du mardi.

Période	Taux d'arrivée (appels/minute)	Taux de service (appels/minute/ téléphoniste)	Nombre de téléphonistes
Matin	1,8	1,5	2
Après-midi	2,2	1,0	3
Soir	1,4	0,7	3

Pour chacune des périodes :

- a) Déterminez le temps moyen passé par les clients à attendre une réponse et la probabilité qu'un client potentiel attende.
 b) Déterminez la longueur maximale de la file d'attente qui ne sera pas dépassée 96 % du temps.
6. Des camionneurs se présentent à un poste de pesée pour le contrôle de la charge totale, afin de vérifier s'ils respectent la réglementation en vigueur. Les camions arrivent entre 7 h et 21 h, selon un processus de Poisson, à un taux de 40 à l'heure. Deux inspecteurs s'occupent du contrôle de la charge, chacun ayant la capacité d'inspecter 25 camions à l'heure. Le taux d'inspection est présumé suivre un processus de Poisson.
- a) Combien de camions peut-on s'attendre à voir en moyenne au poste de pesée, incluant ceux qui sont en train d'être inspectés?
 b) Si un camion vient juste d'arriver au poste de pesée, quel temps moyen devra-t-il passer au poste de pesée?
 c) Quelle est la probabilité que les inspecteurs soient tous deux occupés en même temps?
 d) Combien de minutes un camionneur devra-t-il attendre en moyenne avant d'être servi?
 e) Que se passerait-il s'il y avait un seul inspecteur?
 f) Quelle est la longueur maximale de la file d'attente, si la probabilité qu'elle ne soit pas dépassée est de 0,97?
7. La directrice d'un centre de distribution doit décider du nombre de quais de chargement à mettre en place dans une nouvelle installation. Son critère pour la prise de décision est de minimiser le coût total engendré par le coût d'attente des camions et le coût associé aux quais. Les coûts reliés au duo camion-chauffeur sont estimés à 300 \$ par jour, alors que les coûts d'exploitation associés à chaque quai, incluant les manutentionnaires, sont estimés à 1100 \$ par jour.
- a) Combien de quais devrait-on installer si les camions arrivent, selon un processus de Poisson, au taux moyen de quatre camions par jour et que chaque quai a la capacité d'accueillir cinq camions par jour, selon une distribution de Poisson?
 b) Un employé a proposé d'ajouter un nouvel équipement qui permettrait d'augmenter le taux de chargement à 5,71 camions par jour. L'équipement coûterait 100 \$ par quai pour chaque jour d'exploitation. Les responsables devraient-ils accepter la proposition?
8. Le service des pièces d'un important concessionnaire automobile de la rive nord de Montréal a un comptoir réservé aux mécaniciens du service à la clientèle. Le temps écouté entre chaque demande de pièces est distribué selon une loi exponentielle: il est, en moyenne, de 5 minutes. Le magasinier peut servir en moyenne 15 mécaniciens à l'heure; le taux de service suit une loi de Poisson et on suppose qu'il y a deux magasiniers en service.

- a) Combien trouve-t-on de mécaniciens en moyenne devant le comptoir, y compris ceux que l'on est en train de servir?
- b) Quelle est la probabilité qu'un mécanicien attende avant d'être servi?
- c) Quel est le temps moyen d'attente des mécaniciens?
- d) Quel est le pourcentage d'inactivité des magasiniers?
- e) Quel nombre optimal de magasiniers en service devrait-on avoir, pour minimiser le coût total, si un mécanicien revient à 30 \$ l'heure, alors qu'un magasinier coûte 20 \$ l'heure?
9. Un représentant du service à la clientèle d'une petite entreprise d'informatique est responsable de cinq clients. Ceux-ci demandent de l'aide en moyenne tous les quatre jours ouvrables. On peut estimer que la demande suit une loi de Poisson. Le représentant peut répondre en moyenne à un appel par jour. Déterminez:
- a) Le nombre moyen de clients qui attendent d'être servis.
- b) Le temps d'attente des clients entre le moment où ils appellent pour le service et le moment où le service a été rendu.
- c) Le pourcentage du temps où le représentant est inoccupé.
- d) De combien la réponse obtenue en a) serait-elle réduite si on décidait d'engager deux représentants pour les mêmes clients?
10. Deux opérateurs sont responsables du réglage de 10 machines. Le temps de réglage des machines est distribué selon une loi exponentielle: il est en moyenne de 14 minutes par machine. Les machines fonctionnent pendant en moyenne 86 minutes avant d'avoir besoin d'un réglage. Chaque machine en marche a la capacité de produire 50 pièces à l'heure. Déterminez:
- a) La probabilité qu'une machine attende un réglage.
- b) Le nombre moyen de machines qui attendent un réglage.
- c) Le nombre moyen de machines qui sont en train d'être réglées.
- d) La production moyenne de chaque machine en tenant compte du réglage.
- e) Quel doit être le nombre optimal d'opérateurs, si le temps mort des machines coûte 70 \$ l'heure par machine et que le coût d'un opérateur, incluant le salaire et les avantages sociaux, est de 15 \$ l'heure?
11. Un opérateur est responsable de la maintenance de cinq machines. Les temps de fonctionnement des machines et de maintenance suivent tous deux une distribution exponentielle. Les machines fonctionnent pendant 90 minutes avant de nécessiter une intervention de l'opérateur, et le temps d'intervention est, en moyenne, de 35 minutes. L'opérateur coûte 20 \$ l'heure, incluant le salaire et les avantages sociaux, et le temps mort des machines coûte 70 \$ l'heure par machine.
- a) Si la production de chaque machine en marche est de 60 pièces à l'heure, déterminez la production horaire de chaque machine en tenant compte des attentes pour la maintenance et du temps passé à l'entretien.
- b) Déterminez le nombre optimal d'opérateurs.
12. Un service de fraisage est constitué de 10 machines. Chaque machine fonctionne en moyenne huit heures avant d'avoir besoin d'un réglage. Celui-ci prend en moyenne deux heures. Les machines en marche ont la capacité de produire 40 pièces à l'heure chacune.
- a) Quelle est la production horaire moyenne par machine, lorsqu'il y a un seul opérateur qui s'occupe du réglage au service de fraisage?
- b) Quelle est la configuration optimale des opérateurs affectés au réglage, si le coût horaire des temps morts est de 80 \$, alors que le coût associé à l'opérateur en fonction est de 30 \$ l'heure?
13. Des camions arrivent au quai de chargement d'un grossiste en fruits et légumes à raison de 1,2 camion à l'heure. Une équipe de 2 employés s'occupe du chargement qui prend, en moyenne, 30 minutes. Le salaire des employés est de 10 \$ l'heure, incluant les avantages sociaux, alors que le coût associé aux chauffeurs et aux camions en attente est estimé à 60 \$ l'heure. Le responsable de la logistique envisage d'ajouter un employé à l'équipe. Le taux de service est estimé à 2,4 camions à l'heure. On suppose que les deux taux suivent approximativement la loi de Poisson.
- a) Est-il économique d'ajouter un employé à l'équipe?
- b) Est-ce qu'un quatrième employé serait nécessaire, si le taux de service était de 2,6 camions à l'heure?

14. Les clients qui arrivent à un centre de services sont classés selon trois catégories de priorité, la catégorie n° 1 étant la plus élevée. Les statistiques indiquent qu'il arrive en moyenne neuf clients à l'heure, dont un tiers est attribué à chacune des trois catégories. Il y a deux préposés au centre de services et chacun peut servir en moyenne cinq clients à l'heure. Les processus d'arrivée et de service sont distribués selon une loi de Poisson.

- a) Quel est le taux d'utilisation du système?
- b) Déterminez le temps moyen d'attente des clients de chacune des catégories.
- c) Déterminez le nombre moyen de clients qui attendent d'être servis dans chacune des catégories.

15. Le traitement des clients arrivant dans un centre de services se fait en fonction de leur appartenance à l'une des deux catégories. Celle qui a la priorité la plus élevée a un taux moyen d'arrivée de quatre clients à l'heure, alors que l'autre a un taux moyen de deux clients à l'heure. Les deux processus d'arrivée sont distribués selon une loi de Poisson. Le centre de services est constitué de deux serveurs ayant la capacité de traiter les clients en un temps moyen de six minutes. Les temps de service sont distribués selon une loi exponentielle.

- a) Quel est le taux d'utilisation du système?
- b) Déterminez le temps moyen d'attente des clients de chacune des classes.
- c) Déterminez le nombre moyen de clients qui attendent d'être servis dans chacune des catégories.

16. Un système de file d'attente utilise quatre catégories pour déterminer l'ordre de traitement des clients. Les taux moyens d'arrivée (selon un processus de Poisson) pour chacune des catégories sont donnés dans le tableau ci-dessous :

Catégories	1	2	3	4
Taux moyen	2	4	3	2

Le service est assuré par cinq serveurs ayant chacun la capacité de traiter en moyenne trois clients à l'heure (selon une distribution exponentielle).

a) Quel est le taux d'utilisation du système?
 b) Quel est le temps moyen d'attente pour le service dans chacune des catégories? Quel est le nombre moyen de clients en attente dans chacune des catégories?
 c) Si on réduisait le taux d'arrivée de la deuxième catégorie à trois clients en moyenne, en réorientant certains clients de la deuxième catégorie vers la troisième, quel serait l'effet sur le résultat de la question b)?

17. Répondez aux questions du problème n° 16 en prenant en considération le fait que chaque serveur peut traiter en moyenne quatre clients à l'heure plutôt que trois. Expliquez pourquoi l'impact de la réattribution des clients à la troisième catégorie est beaucoup moins important que dans le cas du problème n° 16.

18. Dans un centre d'appels, les appels des clients arrivent (selon un processus de Poisson) à raison de 40 à l'heure en moyenne. Les personnes auxquelles on ne peut répondre immédiatement sont mises en attente. Le système en place ne peut mettre en attente qu'un maximum de huit personnes. Lorsque ce nombre est atteint, les clients potentiels entendent une sonnerie indiquant que les agents du centre d'appels sont occupés. La communication avec les clients dure en moyenne trois minutes et il y a actuellement trois agents en fonction. La durée de la communication est distribuée selon une loi exponentielle.
 a) Quelle est la probabilité qu'un client potentiel tombe sur le signal «occupé»?
 b) Quelle est la probabilité qu'un client potentiel soit mis en attente?

Bibliographie

- BUFFA, Elwood. *Operations Management*, 3^e édition, New York, John Wiley & Sons, 1972.
- FRITZSIMMONS, James A. et Mona J. FRITZSIMMONS. *Service management: Operations, Strategy and Information Technology*, 3^e édition, New York, Irwin/McGraw-Hill, 2001.
- GRIFFIN, W. *Queueing: Basic Theory and Applications*, Columbus, Ohio, Grid Publishing, 1978.
- HILLIER, Frederick S., Mark S. HILLIER et Gerald J. LIEBERMAN. *Introduction to Management Science: A Modeling and Case Studies Approach with Spreadsheets*, New York, Irwin/McGraw-Hill, 2000.

KARTZ, K. L., B. M. LARSON et R. C. LARSON. « Prescriptions for the Waiting-in-Line Blues: Entertain, Enlighten, and Engage », *Sloan Management Review*, vol. 32, n° 2, hiver 1991, p. 44-53.

STEVENSON, William J. *Introduction to Management Science*, 2^e édition, Burr Ridge, IL., Richard D. Irwin, 1992.

