L'ESSENTIEL DES SYSTÈMES DE FILES D'ATTENTE

Patrick Boily[1,2,3,4], Ehssan Ghashim[3]

Résumé

La **théorie des files d'attente** est une branche des mathématiques qui étudie et modélise le comportement des files d'attente. En tant que volet de la recherche opérationnelle, elle combine des éléments de diverses disciplines quantitatives, mais elle ne fait que rarement partie de la boîte à outils de l'analyste de données.

Dans ce rapport, nous présentons la terminologie et le contexte de base des modèles de files d'attente (y compris la notation de Kendall-Lee, les processus de naissance-mort, et la formule de Little), ainsi que le modèle de file d'attente le plus couramment utilisé : $M/M/c$. Nous décrivons également une application au contrôle de pré-embarquement dans les aéroports canadiens.

Mots-clés

Systèmes de files d'attente, procédés naissance-mort, $M/M/c$, formule de Little.

Reconnaissance de financement

Certaines sections de ce rapport ont été financées par l'entremise d'un octroi de l'Université d'Ottawa visant le développement de matériel pédagogique en français (2019-2020).

[1]Département de mathématiques et de statistique, Université d'Ottawa, Ottawa
[2]Sprott School of Business, Carleton University, Ottawa
[3]Data Action Lab, Ottawa
[4]Idlewyld Analytics and Consulting Services, Wakefield, Canada
Courriel: pboily@uottawa.ca

Table des matières

1 Introduction 1
2 Terminologie des files d'attente 2
2.1 Distributions exponentielle et de Poisson . . . 3
2.2 Distribution d'Erlang . . . 4
2.3 Arrivées et entrées . . . 4
2.4 Sorties et services . . . 4
2.5 Discipline de file d'attente . . . 5
2.6 Méthodes pour joindre la file d'attente . . . 5
3 Cadre théorique 5
3.1 Notation de Kendall-Lee . . . 5
3.2 Processus de naissance et de mort . . . 6
3.3 Loi de Little . . . 6
4 Le système $M/M/1$ 7
4.1 Principes fondamentaux . . . 7
4.2 Capacité limite . . . 8
5 Le système $M/M/c$ 9
6 Application: temps d'attente dans les aéroports 11

1. Introduction

La **théorie des files d'attente** est une branche des mathématiques qui étudie et modélise le comportement des files d'attente. L'article fondateur sur la théorie des files d'attente [1] a été publié en 1909 par le mathématicien danois A.K. Erlang; il y étudiait

> le problème de déterminer le nombre de circuits téléphoniques nécessaires afin de fournir un service qui empêcherait les clients d'attendre trop longtemps avant qu'un circuit se libère. En élaborant une solution à ce problème, il a réalisé que le problème de la minimisation du temps d'attente était applicable à de nombreux domaines, et a commencé à développer davantage sa théorie. Le **problème du tableau téléphonique** d'Erlang a ouvert la voie à la théorie moderne des files d'attente [2].

On cherche à répondre à des questions telles que:

- Est-il probable que des objets/unités/personnes fassent la queue et attendent en file?

- Quelle sera la taille de la file d'attente?
- Combien de temps faudra-t-il attendre?
- Quel sera le niveau d'occupation du système?
- Quelle capacité est requise pour répondre au niveau de demande attendu?

C'est en réfléchissant à ce genre de questions que les analystes et les parties prenantes pourront anticiper les **blocages** ("bottlenecks"). On pourra alors mettre en place des systèmes et des équipes plus efficaces et plus flexibles, plus performants et moins dispendieux et, en fin de compte, offrant un meilleur service aux clients et aux utilisateurs.

La théorie des files d'attente permet également de traiter les blocages (et leur effet sur la performance du système) de manière quantitative. On sera en mesure de donner une réponse à une question telle que "combien de temps devra-t-on attendre, en moyenne", tout comme à une multitude d'autres questions concernant la variabilité des temps d'attente, leur distribution, la probabilité qu'un client reçoive un service médiocre, extrêmement médiocre, etc [11].

Prenons un exemple simple. Supposons qu'il y ait dans une épicerie une seule ligne de caisse et un seul caissier. Si, en moyenne, un client arrive à la caisse pour payer son épicerie toutes les 5 minutes et si le scannage, l'emballage et le paiement prennent 4.5 minutes en moyenne, est-ce que l'on s'attendrait à ce que les clients doivent faire la queue ? Lorsque le problème est présenté de cette manière, notre intuition nous dit qu'il ne devrait pas y avoir de file d'attente et que le caissier devrait rester inactif, en moyenne, 30 secondes toutes les 5 minutes, n'étant occupé que 90% du temps. Personne n'aura alors besoin d'attendre avant d'être servi !

Si vous avez déjà fréquenté une épicerie, vous savez que ce n'est pas ce qui se passe dans la réalité; beaucoup d'acheteurs font la queue, et ils doivent attendre longtemps avant d'être servis. Fondamentalement, le phénomène de **files d'attente** se produit pour trois raisons :

- les **arrivées irrégulières** – les clients n'arrivent pas à la caisse selon un horaire régulier; ils le font parfois éloignés et parfois rapprochés les uns des autres, de sorte à qu'ils y a chevauchement (qui entraînent automatiquement des files d'attente);
- les **tâches de taille irrégulière** – les achats ne sont pas tous traités en 4.5 minutes; un client qui fait les courses pour une famille nombreuse aura besoin de beaucoup plus de temps qu'une personne qui ne fait les courses que pour elle-même, par exemple (lorsque cela se produit, le chevauchement est à nouveau un problème car de nouveaux clients arriveront pour payer leurs courses pendant que les clients actuels sont encore en train de se faire traîter), et
- le **gaspillage** – le temps perdu ne peut jamais être rattrapé; les clients se chevauchent parce que le deuxième client est arrivé trop tôt, avant que le premier n'ait eu le temps de finir de se faire traîter; mais ce n'est peut-être pas la faute du deuxième client; peut-être le premier client serait arrivé à la caisse plus tôt, mais il a perdu du temps à lire un magazine à potins pendant que le caissier était inactif! Ils ont manqué leur chance d'être servis rapidement et, par conséquent, ont fait attendre le deuxième client.

Les arrivées irrégulières et les tâches de taille irrégulière provoquent automatiquement des files d'attente. La seule manière de ne pas avoir de file d'attente est de s'assurer que les tâches soient uniformes, que les arrivées soient régulières et que le caissier ait exactement assez de travail pour faire face aux arrivées. Même lorsque le caissier est à peine occupé, les arrivées irrégulières ou les arrivées **en rafales** peuvent provoquer de l'attente.

En général, les files d'attente **s'aggravent** lorsque les conditions suivantes sont présentes:

- une **utilisation élevée des serveurs** – plus la caissière est occupée, plus il lui faut de temps pour se remettre du temps perdu;
- une **variabilité élevée** – plus la variabilité des arrivées ou de la taille des tâches est importante, plus il y a de gaspillage et de chevauchement (files d'attente), et
- un **nombre insuffisant de serveurs** – Moins de caissiers signifie moins de capacité à absorber les rafales à l'arrivée, ce qui entraîne une plus grande perte de temps et une plus forte utilisation.

2. Terminologie des files d'attente

La théorie des files d'attente étudie les systèmes et les processus d'attente en fonction de trois concepts clés:

- les **clients** sont les unités de travail désservies par le système – un client peut être une personne réelle, ou il peut s'agir de tout ce que le système est censé traiter et compléter: une requête web, une requête de base de données, une pièce à usiner par une machine, etc;
- les **servers** sont les objets qui effectuent le traitement – un serveur peut être le caissier de l'épicerie, un serveur web, un serveur de base de données, une fraiseuse, etc., et
- les **queues** (ou files d'attente) sont les endroits où les unités de travail attendent lorsque le serveur est occupé et ne peut pas commencer le travail dès leur arrivée – une file d'attente peut être une ligne d'attente physique, elle peut résider en mémoire, etc.

Afin de décrire les files d'attente, nous devons d'abord connaître et comprendre certaines distributions utiles, ainsi que les processus dît d'entrée-sortie.

2.1 Distributions exponentielle et de Poisson

Deux distributions jouent un rôle important dans la théorie des files d'attente. La distribution **de Poisson** compte le nombre d'événements discrets se produisant dans une période de temps fixe; elle est étroitement liée à la **distribution exponentielle** (et à la distribution Gamma), qui (entre autres applications) mesure le temps depuis l'arrivée d'un dernier événement. La distribution de Poisson est discrète; la variable aléatoire ne peut prendre que des valeurs entières non négatives. La distribution exponentielle peut prendre n'importe quelle valeur réelle non négative.

Considérons le problème de déterminer la probabilité que n arrivées soient observées sur un intervalle de temps de longueur t, en supposant de plus que:

- la probabilité qu'une arrivée soit observée sur un petit intervalle de temps (disons de longueur ν) est proportionnelle à la longueur de l'intervalle et que la constante de proportionnalité soit λ, de sorte que la probabilité devienne $\lambda\nu$;
- la probabilité de deux ou plusieurs arrivées sur ce même petit intervalle est nulle, et
- le nombre d'arrivées dans un intervalle de temps donné est indépendant du nombre d'arrivées dans un intervalle qui ne le chevauche pas (par exemple, le nombre d'arrivées survenant entre 5 et 25 minutes depuis le départ ne fournit pas de renseignements sur le nombre d'arrivées survenant entre 30 et 50 depuis le départ).

Soit $P(n;t)$ la probabilité d'observer n arrivées dans un intervalle de temps de longueur t. La fonction de masse de la distribution du nombre d'arrivées est

$$P_\lambda(n;t) = \frac{(\lambda t)^n}{n!} e^{-\lambda t},\ n = 0, 1, 2, \cdots$$

pour un $\lambda > 0$ spécifique au problème, c'est-à-dire que la distribution est Poisson (consulter la figure 1). Dans un système de file d'attente, ces arrivées sont dîtes **de Poisson**.

Le temps d'attente entre deux arrivées successives est l'**intervalle d'arrivées**. Si le nombre d'arrivées dans un intervalle de temps donné suit une distribution de Poisson avec paramètre λt, les intervalles d'arrivées suivent une distribution exponentielle, dont la fonction densité est donnée par

$$f_\lambda(t) = \lambda e^{-\lambda t},\ \text{pour } t > 0,$$

et la probabilité $P(W \leq t)$ que le temps d'attente W pour un utlisateur est inférieur à t est

$$P(W \leq t) = 1 - e^{-\lambda t} \quad \text{(consulter la Figure 1)}.$$

De manière générale, si le taux d'arrivée est **stationnaire** et si il n'y a pas d'arrivées **en vrac** (c'est à dire qu'il n'y a pas d'arrivées simultanées), et si les arrivées passées n'affectent pas les arrivées futures, alors les intervalles d'arrivées suivent une distribution exponentielle avec paramètre λ, et le nombre d'arrivées dans tout intervalle de longueur t est Poisson avec paramètre λt.

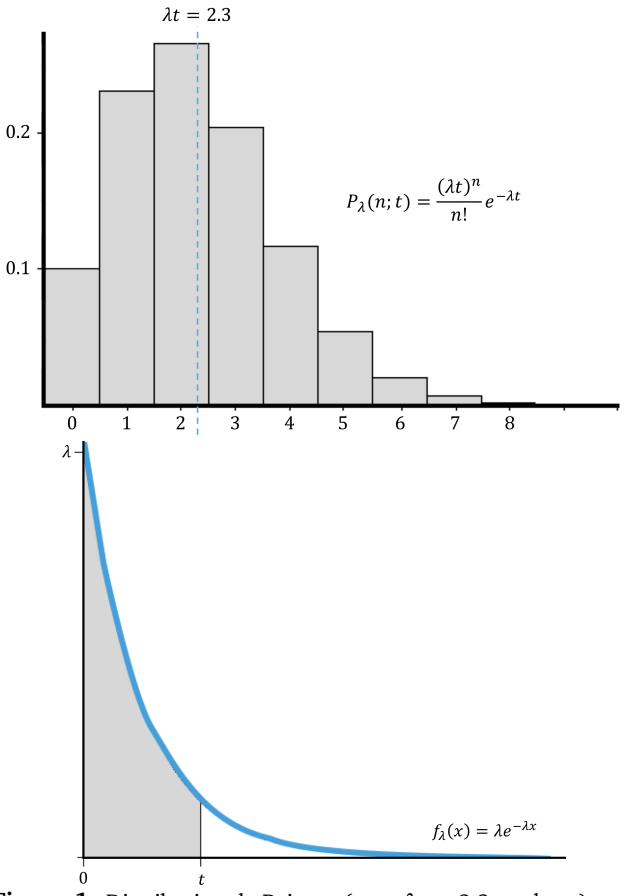

Figure 1. Distribution de Poisson (avec $\lambda t = 2.3$, en haut) et distribution exponentielle (avec paramètre λ, en bas). La zone ombrée sur cette dernière représente la probabilité d'une attente de t unités de temps si les arrivées sont Poisson de paramètre λ.

L'une des caractéristiques les plus intéressantes de la distribution exponentielle relative aux intervalles d'arrivées est qu'elle est **sans mémoire** – si une variable aléatoire X suit une distribution exponentielle, alors pour tout $t, h \geq 0$,

$$P(X \geq t + h | X \geq t) = P(X \geq h). \tag{1}$$

C'est la seule fonction de densité qui satisfait à cette propriété [4]. Cette propriété est importante car elle implique que la distribution des intervalles d'arrivées est indépendante du temps écoulé depuis la dernière arrivée – imaginez si c'était le cas dans les transports publics!

Par exemple, si nous savons qu'au moins t d'unités de temps se sont écoulées depuis la dernière arrivée, alors l'intervalle h jusqu'à la prochaine arrivée est indépendante de t. Si $h = 4$, disons, alors (1) donne

$$P(X > 9 | X > 5) = P(X > 7 | X > 3) = P(X > 4).$$

2.2 Distribution d'Erlang

La distribution exponentielle n'est pas toujours un modèle approprié des intervalles d'arrivées; il n'est pas difficile d'imaginer une situation où le temps d'attente ne devrait pas être sans mémoire, par exemple. Une approche alternative utilise la distribution d'**Erlang** $\mathscr{E}(R,k)$, une variable aléatoire continue à deux paramètres $R>0$ $k\in\mathbb{Z}^+$, dont la fonction de densité est

$$f_{R,k}(t)=\frac{R(Rt)^{k-1}e^{-Rt}}{(k-1)!},\ t\geq 0.$$

Lorsque $k=1$, la distribution d'Erlang se réduit à la distribution exponentielle Exp(R). On peut aussi montrer que si $X\sim\mathscr{E}(k\lambda,k)$, alors $X\sim X_1+X_2+\cdots+X_k$, où chaque X_i est une variable aléatoire exponentielle indépendent, de paramètre $k\lambda$.

En modélisant les intervalles d'arrivées selon une distribution d'Erlang $\mathscr{E}(k\lambda,k)$, nous disons de façon équivalente que les utilisateurs passent par k **phases** (dont chacune est sans mémoire) avant d'être servi. Pour cette raison, le paramètre de forme est souvent appelé le nombre de phases de la distribution d'Erlang [14].

2.3 Arrivées et entrées

Le processus de saisie est généralement appelé **processus d'arrivée**, les arrivées sont appelées **clients** (ou utilisateurs). Dans les modèles que nous considérerons, on suppose qu'il n'y a pas d'arrivées simultanées (ce qui peut être irréaliste lorsque l'on modélise les arrivées dans un restaurant, par exemple). Si des arrivées simultanées sont possibles (en théorie et/ou en pratique), nous disons des arrivées en vrac qu'elles sont **autorisées**.

En général, nous supposons que le processus d'arrivée **n'est pas affecté par le nombre de clients** dans le système. Dans le contexte d'une banque, cela impliquerait que le processus régissant les arrivées reste inchangé, qu'il y ait 500 ou 5 personnes attendant qu'un guichet se libère.

Il existe deux situations courantes dans lesquelles le processus d'arrivée peut dépendre du nombre de clients présents. La première se produit lorsque les arrivées sont issues d'une petite population – les modèles dits de **source limitée** – si tous les membres de la population sont déjà dans le système, il ne peut y avoir une autre arrivée!

Une autre situation de ce type se produit lorsque le taux auquel les clients arrivent dans l'établissement diminue lorsque celui-ci devient trop encombré. Par exemple, lorsque les clients constatent que le stationnement d'un restaurant est plein, ils peuvent très bien décider d'aller dans un autre restaurant ou de renoncer complètement à manger à l'extérieur. Si un client arrive mais ne parvient pas à entrer dans le système, nous disons que l'accès au système lui a été **bloqué**.

2.4 Sorties et services

Pour décrire le processus des sorties d'un système de file d'attente (souvent appelé **processus de service**), nous spécifions généralement une **distribution du temps de service** qui régit le temps de service pour les utilisateurs.

Dans la plupart des cas, on suppose que cette distribution est indépendante du nombre de clients présents dans le système. Cela signifie, par exemple, que le serveur ne fonctionne pas plus vite lorsque lorsque le nombre de clients augmente.

On distingue deux types de serveurs: les serveurs **en parallèle** et les serveurs **en série**. Les serveurs en parallèle fournissent tous le même type de service et les clients ne doivent passer que par l'un d'entre eux pour obtenir un service complet. Les guichets d'une banque, par exemple, sont généralement disposés en parallèle; typiquement, les clients sont servis que par un seul guichet, et n'importe quel des guichets peut offrir le service souhaité.

Les serveurs sont en série si un client doit passer par plusieurs serveurs avant de terminer son service. Une chaîne de montage est un exemple d'un tel système de mise en file d'attente.

On retrouve de tels processus dans diverses situations:

- **situation:** achater des billets pour les Blue Jays au centre Rogers
  arrivées: les partisans arrivent au guichet
  sorties: les guichetiers servent les acheteurs;

- **situation:** pizzéria
  arrivées: les demandes de livraison de pizzas sont reçues
  output: la pizzéria prépare et cuit des pizzas, et les envoie pour être livrées;

- **situation:** centre de services publics
  input: les citoyens/résidents entrent dans le centre de services
  sorties: la réceptionniste les affecte à une file d'attente spécifique en fonction de leurs besoins
    arrivées: les citoyens/résidents se joignent à une file d'attente spécifique
    sorties: un fonctionnaire répond à leurs besoins;

- **situation:** banque de sang à l'hôpital
  arrivées: les pintes de sang arrivent à l'hôpital
  sorties: les patients utilisent les pintes de sang selon leur type sanguin;

- **situation:** garage
  arrivées: les voitures tombent en panne et sont envoyées au garage afin d'être réparées
  sorties: les voitures sont réparées par des mécaniciens et renvoyées sur les rues.

Les calculs pertinents sont assez faciles à exécuter, comme le montrent les exemples suivants.

Exemple 1. En moyenne, on s'attend à ce que 4.6 clients entrent dans un café durant chaque heure où il est ouvert. Si les arrivées respectent un processus de Poisson, la probabilité qu'au plus deux clients entrent pendant une période de 30 minutes (0.5 heures) est de

$$\begin{aligned} P_{4.6}(n \leq 2; 0.5) &= P_{4.6}(0, 0.5) + P_{4.6}(1, 0.5) + P_{4.6}(2, 0.5) \\ &= e^{-4.6 \cdot 0.5} \left[\frac{(4.6 \cdot 0.5)^0}{0!} + \frac{(4.6 \cdot 0.5)^1}{1!} + \frac{(4.6 \cdot 0.5)^2}{2!} \right] \\ &\approx 0.5960; \end{aligned}$$

la distribution de Poisson correspondante peut être consultée à la Figure 1.

Exemple 2. Dans un fast-food, un caissier sert en moyenne 9 clients par heure. Si le temps de service suit une distribution exponentielle, 77.7% et 1.1% des clients seront servis en 10 minutes ou moins, et après 30 minutes, respectivement. En effet,

$$\begin{aligned} P(W \leq 10/60) &= 1 - e^{-9 \cdot 10/60} \approx 0.7769 \\ P(W > 30/60) &= e^{-9 \cdot 30/60} \approx 0.0111. \end{aligned}$$

2.5 Discipline de file d'attente

Lorsque l'on cherche à décrire complètement un système de file d'attente, il faut aussi décrire sa **discipline** (ou de sa politique de service) et la manière dont les utilisateurs se **joignent aux lignes**. La discipline de la file d'attente décrit la méthode utilisée pour déterminer l'ordre dans lequel les clients sont servis :

- la politique de service la plus courante est celle dite du **premier arrivé, premier servi** ("first come, first served", FCFS), dans laquelle les clients sont servis dans l'ordre de leur arrivée, comme on s'attendrait à le voir dans un café d'Ottawa, par exemple;
- sous la politique dite du **dernier arrivé, premier servi** ("last come, first served", LCFS), les derniers arrivés sont les premiers à entrer en service, comme c'est le cas dans un ascenseur bondé où tous les utilisateurs se rendent au même étage, par exemple;
- parfois, l'ordre dans lequel les clients arrivent n'a aucun effet sur l'ordre dans lequel ils sont servis; ce serait le cas si le prochain client à entrer en service est choisi au hasard parmi les clients en attente de service, une situation appelée **service dans un ordre aléatoire** ("service in random order", SIRO);
- enfin, la politique de **priorisation** place chaque arrivée dans une catégorie, chacune d'entre elles se voyant attribuer un niveau de priorité (un processus de **triage**); à l'intérieur de chaque niveau de priorité, les clients entrent dans la file d'attente sur une base FCFS; une telle politique de service est souvent utilisée dans les salles d'urgence afin de déterminer l'ordre dans lequel les clients sont traités, ou dans les services de reprographie et de partage de temps informatique, où la priorité est généralement accordée aux travaux dont les délais de traitement sont plus courts.

2.6 Méthodes pour joindre la file d'attente

Un autre facteur crucial dans le comportement du système de file d'attente est la **méthode** utilisée par les utilisateurs afin de déterminer à quelle ligne ils se joindront. Par exemple, dans certaines banques, les clients doivent se joindre à une seule ligne, mais dans d'autres, les clients peuvent choisir la ligne qu'ils souhaitent rejoindre.

Lorsqu'il y a plusieurs lignes, les clients rejoignent souvent la ligne la plus courte. Malheureusement, dans de nombreuses situations (comme au supermarché, par exemple), il est difficile de définir quelle ligne est la plus courte, puisque le temps d'attente est aussi une fonction du nombres d'items dans les paniers des utilisateurs se trouvant déjà dans la ligne. S'il y a plusieurs lignes dans une installation de file d'attente, il est important de savoir si les clients sont autorisés à **changer** de ligne. Dans la plupart des systèmes de files d'attente à lignes multiples, le passage d'une ligne à l'autre est autorisé, mais il n'est pas recommandé de le faire aux douanes, par exemple.

3. Cadre théorique

La notation de **Kndall-Lee** permet de décrire de grandes familles de systèmes de files d'attente [10].

3.1 Notation de Kendall-Lee

On décrit les systèmes de file d'attente à l'aide de six attributs:

$$x_1/x_2/x_3/x_4/x_5/x_6.$$

La première caractéristique précise la nature du **processus d'arrivée**. On utilise les abréviations standard suivantes:

M	=	intervalles d'arrivées exponentiels indépendants et identiquement distribués (iid)
D	=	intervalles d'arrivées déterministes, iid
E_k	=	intervalles d'arrivées suivent Erlang $\mathscr{E}(R,k)$, iid
G	=	intervalles d'arrivées iid et régis par une distribution générale.

La deuxième caractéristique précise la nature du **temps de service**:

M	=	temps de services exponentiels, iid
D	=	temps de services déterministes, iid
E_k	=	temps de services Erlang $\mathscr{E}(R,k)$, iid
G	=	temps de service iid et régis par une distribution générale.

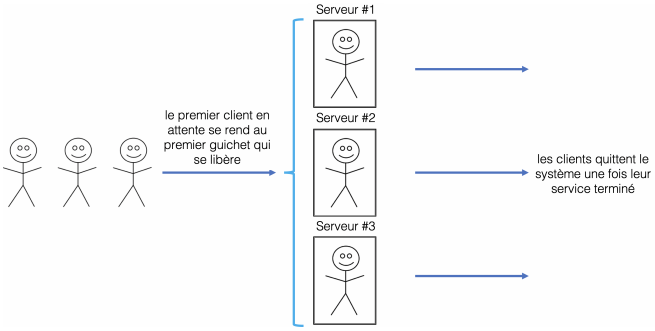

Figure 2. Une file d'attente avec 3 guichets – $M/M/3/\text{FCFS}/20/\infty$.

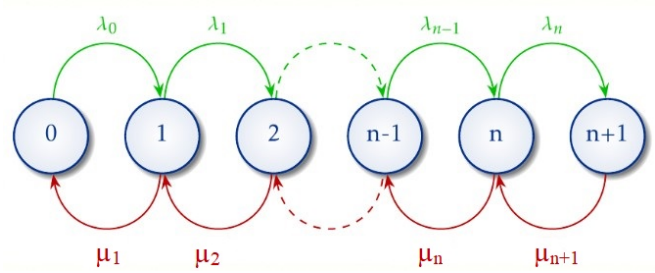

Figure 3. Processus de naissance et de mort; les taux de natalité et de mortalité sont indiqués par λ_n et μ_m, respectivement (source inconnue).

La troisième caractéristique représente le **nombre de serveurs parallèles**; c'est un nombre entier positif.

La quatrième caractéristique décrit la **politique de service**:

FCFS	=	premier arrivé, premier servi
LCFS	=	dernier arrivé, dernier servi
SIRO	=	service dans un ordre aléatoire
GD	=	politique de service geénérale.

La cinquième caractéristique précise le **nombre maximal d'utilisateurs** pouvant être accommodé par le système, tandis que la sixième caractéristique donne le **taille de la population** dont sont issus les utilisateurs. À moins que le nombre de clients potentiels ne soit du même ordre de grandeur que le nombre de serveurs, la taille de la population est considérée comme infinie.

Dans de nombreux modèles importants, $x_4/x_5/x_6$ correspond à $\text{GD}/\infty/\infty$; dans ce cas, ces attributs sont souvent omis de la description de la file d'attente.

Par exemple, $M/M/3/\text{FCFS}/20/\infty$ pourrait représenter une banque avec 3 guichets, des temps d'arrivée et de service exponentiels, une politique de service "premier arrivé, premier servi", une capacité totale de 20 clients et un bassin de population infini dans lequel puiser. La situation est partiellement illustrée à la Figure 2.

3.2 Processus de naissance et de mort

L'**état** d'un système de file d'attente au temps t est le nombre de clients dans le système de file d'attente (soit en attente en ligne ou en service) au temps t. Lorsque $t = 0$, l'état du système est tout simplement le nombre initial de clients dans le système. Cet état vaut la peine d'être noté car il affecte l'état du système pour les autres instants t.

En conséquence, nous définissons $P_{i,j}(t)$ comme la probabilité que l'état soit j au temps t, étant donné que l'état à $t = 0$ était i. Pour des valeurs élevées de t, $P_{i,j}(t)$ devient indépendant de i et se rapproche d'une limite π_j. Cette limite est le **régime stable** de l'état j.

Il est en général difficile de déterminer les étapes des arrivées et des services qui mènent à un état stable π_j. De même, à partir d'un t près du début, il est difficile de déterminer exactement quand un système atteindra son état stable π_j, et même si un tel état existe.

Par souci de simplicité, lorsqu'on étudie un système de file d'attente, on commence par supposer que l'état d'équilibre a déjà été atteint.

Un **processus de naissance et de mort** est un processus de Markov dans lequel les états sont indexés par des entiers non négatifs, et les transitions d'états ne sont autorisées qu'entre états "voisins." Après une "naissance," l'état du système passe de n à $n+1$; après une "mort," de m à $m-1$. En général, les taux de natalité et de mortalité sont représenté spar λ_n et μ_m, respectivement (comme on peut le constater à la Figure 3). Les processus de nature **purement naissance** sont ceux pour lesquels $\mu_m = 0$ pour tout m; pour ceux de nature **purement mort**, on a $\lambda_n = 0$ pour tout n. La **solution d'état stable** d'un processus de naissance-mort, c'est-à-dire la probabilité π_n que le système se retrouve dans l'état n, *peut* en fait être calculée:

$$\pi_n = \pi_0 \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n}, \quad \text{pour } n = 1, 2, \cdots, \tag{2}$$

où π_0 est la probabilité que le système se retrouve dans l'état 0 (c'est-à-dire sans utilisateurs). On peut également montrer [11] que:

$$\pi_0 = \frac{1}{1 + \sum_{n=1}^{\infty} \prod_{j=0}^{n-1} \frac{\lambda_j}{\mu_{j+1}}}.$$

3.3 Loi de Little

Les analystes et les clients souhaitent souvent déterminer combien temps l'utilisateur moyen passe dans le système de file d'attente. Soit W le **temps d'attente prévu passé dans le système de file d'attente**, y compris le temps passé dans la file et le temps de service, et W_q le **temps d'attente prévu d'un client dans la file**. Les deux valeurs W et W_q sont calculés en supposant que l'état d'équilibre du système a été atteint. En utilisant un puissant résultat connu sous le nom de **loi de Little**, W et W_q sont facilement liés au

nombre de clients dans la file d'attente et à ceux qui font la queue.

Pour tout système de file d'attente (ou tout sous-ensemble d'un système de file d'attente), on considère les quantités suivantes:

- λ = nombre moyen d'arrivées dans le système par unité de temps;
- L = le nombre moyen d'utilisateurs présents dans le système de file d'attente;
- L_q = le nombre moyen d'utilisateurs qui font la queue;
- L_s = le nombre moyen d'utilisateurs en service;
- W = le temps moyen qu'un utilisateur passe dans le système;
- W_q = le temps moyen qu'un utilisateur passe dans la file d'attente, et
- W_s = le temps moyen qu'un utilisateur passe en service.

Les utilisateurs du système ne peuvent que se trouver dans la file d'attente ou en service, de sorte que $L = L_q + L_s$ et $W = W_q + W_s$ (dans ces définitions, toutes les moyennes sont des moyennes de régime permanent ("steady-state")). Pour la plupart des systèmes de file d'attente dans lesquels un tel régime existe, la loi de Little se résume par

$$L = \lambda W, \quad L_q = \lambda W_q, \quad \text{et} \quad L_s = \lambda W_s.$$

Exemple 3. Si, en moyenne, 46 clients entrent dans un restaurant à chaque heure où il est ouvert, et s'ils passent, en moyenne, 10 minutes à attendre d'être servis, nous devrions nous attendre à ce que $46 \cdot 1/6 \approx 7.7$ clients se retrouve dans la file d'attente à tout moment, en moyenne.

4. Le système $M/M/1$

Nous allons maintenant discuter du système de file d'attente non trivial le plus simple.

4.1 Principes fondamentaux

Un système de file d'attente $M/M/1/\mathrm{GD}/\infty/\infty$ dispose d'intervalles d'arrivée et de temps de service dont les distributions respectives sont exponentielles, et un serveur unique. On peut le modéliser à l'aide d'un processus de naissance et de mort où

$$\begin{aligned} &\lambda_j = \lambda, \ j = 0, 1, 2, \dots \\ &\mu_0 = 0 \\ &\mu_j = \mu, \ j = 1, 2, 3, \dots \end{aligned}$$

En substituant ces taux de natalité et de mortalité dans (2), on obtient

$$\pi_j = \frac{\lambda^j \pi_0}{\mu^j} = \rho^j \pi_0,$$

où $\rho = \lambda/\mu$ représente l'**intensité du trafic** dans le système.

Étant donné que le système doit se trouver exactement dans l'un des états à tout moment, la somme de toutes les probabilités se doit d'être 1:

$$\pi_0 + \pi_1 + \pi_2 + \cdots = \pi_0(1 + \rho + \rho^2 + \cdots) = 1.$$

Lorsque $0 \le \rho < 1$, la série converge vers $\frac{1}{1-\rho}$, d'où l'on dérive

$$\pi_0 \cdot \frac{1}{1-\rho} = 1 \implies \pi_0 = 1-\rho \implies \pi_j = \rho^j \pi_0 = \rho^j(1-\rho);$$

c'est la **probabilité de retrouver le système dans le j^{e} état dans le régime stable**. Si au contraire $\rho \ge 1$, la série diverge et le système n'atteint pas de régime stable. Intuitivement, cela se produit lorsque $\lambda \ge \mu$: lorsque le taux d'arrivée est supérieur au taux de service, l'état du système croît sans cesse et la file d'attente ne se dégage jamais.

À partir de maintenant, nous allons présumer que $\rho < 1$ afin de garantir l'existence des probabilités π_j dans le régime stable, à partir desquelles nous pouvons déterminer plusieurs quantités d'intérêt.

En supposant que l'état d'équilibre a été atteint, on peut montrer que L, L_s, et L_q sont, respectivement:

$$\begin{aligned} L &= \frac{\lambda}{\mu-\lambda} = \frac{\rho}{1-\rho} \\ L_s &= \rho \\ L_q &= \frac{\rho^2}{1-\rho}. \end{aligned}$$

À l'aide de la loi de Little, nous pouvons également obtenir W, W_s, et W_q en divisant chacune des valeurs correspondantes de L par λ:

$$\begin{aligned} W &= \frac{1}{\mu-\lambda} \\ W_s &= \frac{1}{\mu} \\ W_q &= \frac{\lambda}{\mu(\mu-\lambda)}. \end{aligned}$$

On note comme prévu que $W, W_q \to +\infty$ quand $\rho \to 1$. En revanche, $W_q \to 0$ et $W \to \frac{1}{\mu}$ (le **temps moyen de service**) lorsque $\rho \to 0$.

Exemple 4. (selon [3]) En moyenne, dix clients arrivent à un guichet à toutes les heures. Si le client moyen est servi en 4 minutes, et si les intervalles d'arrivée et le temps de service suivent tous deux des distributions exponentielles, alors:

(a) Quelle est la probabilité que le guichet se retrouve au repos?
(b) Sans compter le client qui se fait servir, combien de clients font la queue au guichet, en moyenne?

(c) Combien de temps, en moyenne, un client passe-t-il dans le système de file d'attente (y compris le temps de service)?
(d) En moyenne, combien de clients seront servis par le caissier à chaque heure?

Solution: nous faisons affaire à un système de file d'attente

$$M/M/1/\mathrm{GD}/\infty/\infty$$

pour lequel $\lambda = 10$ clients/h et $\mu = 15$ clients/h, d'où $\rho = 10/15 = 2/3$.

(a) Puisque $\pi_0 = 1-\rho = 1/3$, le guichet se retrouve au repos $1/3$ du temps.
(b) Il y a $L_q = \rho^2/(1-\rho) = 4/3$ clients en ligne pour le guichet, en moyenne.
(c) On sait que $L = \lambda/(\mu-\lambda) = 10/(15-10) = 2$,, d'où $W = L/\lambda = 0.2$ h $= 12$ min.
(d) Si le guichet est toujours occupé, on sert en moyenne $\mu = 15$ clients par heure. Selon (a), nous savons que la caissière n'est occupée que les deux tiers du temps, c'est-à-dire qu'à chaque heure, elle ne sert en moyenne que $15 \cdot 2/3 = 10$ clients. C'est un résultat raisonnable car, dans le régime stable, 10 clients entrent dans le système et 10 clients quittent le système à chaque heure.

Exemple 5. (selon [17]) Supposons que tous les propriétaires de voitures fassent le plein lorsque leur réservoir est exactement à moitié vide. Supposons également que 7.5 clients arrivent en moyenne à chaque heure à une station-service qui n'a qu'une seule pompe et qu'il faut en moyenne 4 minutes par voiture pour faire le plein. Supposons finalement que les intervalles d'arrivées et les temps de service suivent tous deux des distributions exponentielles.

(a) Quelles valeurs prennent L et W dans ce scénario?
(b) Supposons que lorsqu'il y a pénurie de gaz, les achats d'essence se fassent dans la panique. On modélise ce phénomène en imaginant que tous les propriétaires de voitures achètent désormais de l'essence lorsque leur réservoir est rempli aux trois quarts exactement. Comme chaque propriétaire de voiture met désormais moins d'essence dans le réservoir à chaque visite à la station, nous supposons que la durée moyenne du temps de service a été réduite à 10/3 minutes. De quelle façon est-ce que cela affecte les valeurs de L et W?

Solution: on prend pour acquis que le systme de file d'attente prend la forme

$$M/M/1/\mathrm{GD}/\infty/\infty,$$

où $\lambda = 7.5$ voitures/h et $\mu = 60/4 = 15$ voitures/h. Nous avons alors $\rho = 7.5/15 = 1/2$.

(a) Par définition, $L = \lambda/(\mu-\lambda) = 7.5/(15-7.5) = 1$ et $W = 1/7.5 \approx 0,13$ h $= 7.8$ min. Dans cette situation, tout est sous contrôle et de longues files d'attente semblent peu probables.
(b) Pour le scénario où les achats se font dans la panique, $\lambda = 2(7.5) = 15$ voiture/h puisque les propriétaires de voiture font le plein 2 fois plus souvent qu'au préalable, et $\mu = 60 \cdot 3/10 = 18$ voitures/h, d'où $\rho = \lambda/\mu = 5/6$. Dans ce cas,

$$L = \frac{\rho}{1-\rho} = 5 \text{ voitures, et } W = \frac{L}{\lambda} = \frac{5}{15} = 20 \text{ min.}$$

Ainsi, les achats dans la panique ont pour effet de plus que doubler le temps d'attente dans la file d'attente.

Dans un système $M/M/1$, on a obligatoirement

$$L = \frac{\rho}{1-\rho} = -1 + \frac{1}{1-\rho},$$

et on constate aisément que $L \to \infty$ comme $\rho \to 1$. La multiplication par 5 de la valeur de L lorsque ρ passe de $1/2$ à $5/6$ (avec des sauts correspondants en W) illustre ce fait.

4.2 Capacité limite

En réalité, la capacité d'une file d'attente ne saurait être infinie – elles est limitées par les exigences de l'espace et/ou du temps, ou encore par la politique d'exploitation des services. Un modèle qui tient compte des ces aspects est du ressort des **files d'attente finies**.

Ces modèles limitent le nombre de clients autorisés dans le système de service, soit N. Si le système est à **capacité**, l'arrivée d'un $(N+1)^{\mathrm{e}}$ client entraîne l'impossibilité d'entrer dans la file d'attente – l'accès au système est bloqué pour ce client, qui doit quitter sans recevoir de service.

Les files d'attente finies peuvent également être modélisées par un processus de naissance-mort, mais avec une légère modification de ses paramètres:

$$\begin{aligned} &\lambda_j = \lambda,\ j = 0,1,2,\ldots,N-1 \\ &\lambda_N = 0,\ \mu_0 = 0 \\ &\mu_j = \mu,\ j = 1,2,3,\ldots,N \end{aligned}$$

La restriction $\lambda_N = 0$ distingue ce modèle de $M/M/1/\infty$. Elle rend impossible l'accès à un état supérieur à N. En conséquence, les modèles de file d'attente finie ont toujours un régime stable puisque même si $\lambda \geq \mu$, il ne peut jamais y avoir plus de N clients dans le système.

Du côté mathématique, on remplace la série infinie reliant les π_j par une série géométrique finie, qui converge quelque soit la valeur de ρ:

$$\pi_0 + \pi_1 + + \cdots + \pi_N = \pi_0(1 + \rho + \cdots + \rho^N) = 1,$$

d'où

$$\pi_0 \cdot \frac{1-\rho^{N+1}}{1-\rho} = 1$$
$$\implies \pi_0 = \frac{1-\rho}{1-\rho^{N+1}}$$
$$\implies \pi_j = \begin{cases} \rho^j \pi_0 & \text{lorsque } j = 0,\ldots,N \\ 0 & \text{lorsque } j > N \end{cases}$$

Puisque $L = \sum_{j=0}^{N} j\pi_j$ (est-ce évident?),

$$L = \frac{\rho[1+N\rho^{N+1}-(N+1)\rho^N]}{(1-\rho)(1-\rho^{N+1})}$$

lorsque $\lambda \neq \mu$. Comme c'est le cas dans une file d'attente $M/M/1/\infty$, $L_s = 1-\pi_0$ et $L_q = L - L_s$.

Dans un modèle à capacité finie, seulement $\lambda - \lambda\pi_N = \lambda(1-\pi_N)$ arrivées par unité de temps entrent effectivement dans le système, en moyenne (il y a en fait λ arrivée, mais pour $\lambda\pi_N$ d'entre elles, le système est rempli). On a alors

$$W = \frac{L}{\lambda(1-\pi_N)} \quad \text{et} \quad W_q = \frac{L_q}{\lambda(1-\pi_N)}.$$

A quoi cela ressemble-t-il, concrètement?

Exemple 6. Imaginons un salon de coiffure, tenu par un seul barbier, contenant un total de 10 sièges. Supposons de plus, comme nous l'avons toujours fait jusqu'à présent (quoique cela ne soit pas nécessaire), que les intervalles d'arrivée sont réparties de manière exponentielle, avec une moyenne de 20 clients potentiels arrivant chaque heure au salon. Les clients qui trouvent le magasin plein n'y entrent pas (peut-être n'aiment-ils pas rester debout). Le barbier prend en moyenne 12 minutes pour couper les cheveux de chaque client (supposons que les temps de coupe soient également répartis de manière exponentielle).

(a) En moyenne, combien de coupes de cheveux le barbier réalise-t-il par heure?
(b) En moyenne, combien de temps un client qui entre dans le magasin passe-t-il dans le salon?

Solution: il n'y a pas tant de chose à dire. Allons-y!

(a) Le salon sera plein pour une fraction π_{10} de toutes les arrivées. Ainsi, $\lambda(1-\pi_{10})$ clients sont introduits dans la file d'attente à toutes les heures, en moyenne. Puisque tous les clients qui entrent dans la file reçoivent une coupe de cheveux, le barbier donne en moyenne $\lambda(1-\pi_{10})$ coupes de cheveux par heure. Dans ce scénario, $N = 10$, $\lambda = 20$ clients/h, et $\mu = 60/12 = 5$ clients/h. Ainsi, $\rho = 20/5 = 4$ et nous avons

$$\pi_0 = \frac{1-\rho}{1-\rho^{N+1}} = \frac{1-4}{1-4^{11}} \approx 7.15 \times 10^{-7} \text{ et}$$
$$\pi_{10} = 4^{10}\pi_0 = \frac{3}{4}.$$

Il y a ainsi $20(1-3/4) = 5$ clients qui reçoivent une coupe de cheveux par heure, en moyenne. Cela signifie que l'accès à la file d'attente sera bloqué pour $20-5 = 15$ clients potentiels à chaque heure, en moyenne.

(b) On dfermine W en calculant tout d'abord

$$L = \frac{4[1+(10)4^{11}-(11)4^{10}]}{(1-4)(1-4^{11})} = 9.67.$$

À l'aide de la formule donnée précedemment, nous obtenons

$$W = \frac{L}{\lambda(1-\pi_{10})} = \frac{9.67}{5} = 1.93 \text{ h}.$$

Le salon de coiffure est bondé - le barbier serait bien avisé d'engager au moins un autre barbier !

Quel effet l'embauche d'un deuxième coiffeur aurait-elle sur le systme de file d'attente? Pour répondre à cette question, nous devons étudier les systèmes de files d'attente $M/M/c$.

5. Le système $M/M/c$

Dans un système $M/M/c/\text{GD}/\infty$, les intervalles d'arrivées et les temps de service suivent également des distributions exponentielles, de paramètre respectif λ et μ. Ce qui distingue ce système de celui que nous avons déjà étudié, c'est qu'il y a maintenant $c > 1$ serveurs/guichets prêts à servir une ligne de clients, comme on en trouve dans une banque, par exemple (cf. Figure 4).

Avec $j \leq c$ utilisateur dans le système, chaque client est servi et il n'y a pas de temps d'attente; mais s'il y $j > c$ utlisateurs dans le système, c se font servir et $j-c$ clients restent en attente dans la file.

On peut modéliser la situation à l'aide d'un processus de naissance-mort; il suffit de remarquer que le taux de mortalité dépend du nombre effectif de serveurs utilisés.

Si chaque serveur fournit un service à un taux μ (ce qui peut ne pas être le cas dans la pratique car il peut y avoir des variations dans les serveurs, tout du moins pour les serveurs humains), alors le taux de mortalité réel est μ multiplié par le nombre de clients desservis. Les paramètres de ce processus sont

$$\lambda_n = \lambda,\ n = 0,1,2,\ldots$$
$$\mu_n = \begin{cases} n\mu,\ n = 0,1,2,\ldots,c \\ c\mu,\ n = c+1, c+2,\ldots \end{cases}$$

L'intensité du trafic d'un système $M/M/c$ est $\rho = \lambda/(c\mu)$ et la solution de régime stable est

$$\pi_n = \begin{cases} \frac{(c\rho)^n}{n!}\pi_0, & 1 \leq n \leq c \\ \frac{c^c\rho^n}{c!}\pi_0, & n \geq c \end{cases}$$

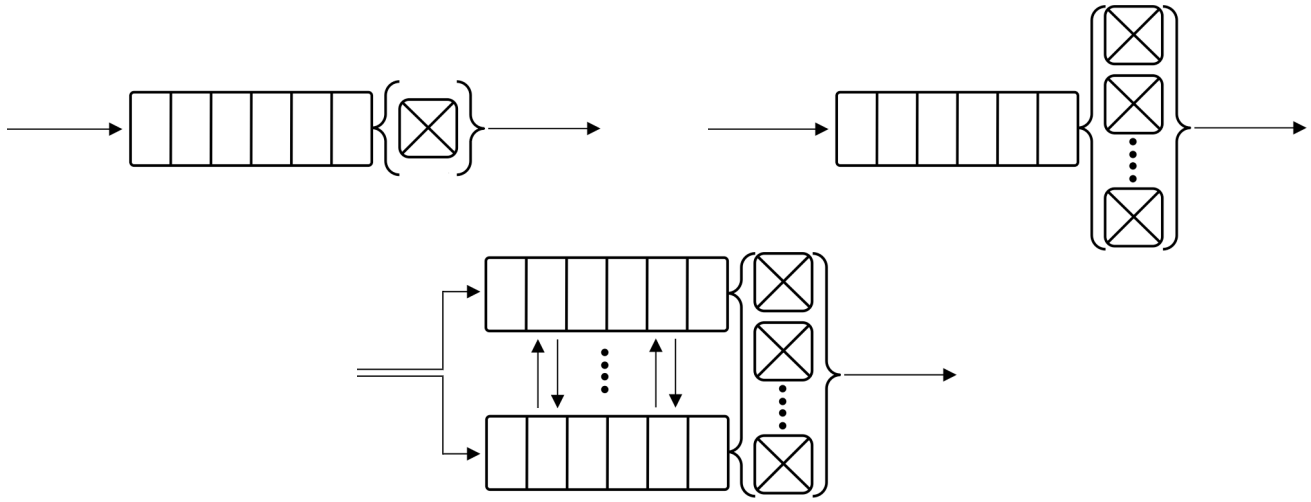

Figure 4. Schéma de systèmes de file d'attente ($M/M/1$, $M/M/c$, en tandem); les clients arrivent par la gauche, entrent dans la file d'attente, y demeurent jusqu'à ce qu'ils soient servis, puis sortent de la file.

où

$$\pi_0 = \left[1 + \frac{(c\rho)^c}{c!(1-\rho)} + \sum_{n=1}^{c-1} \frac{c\rho^n}{n!}\right]^{-1}.$$

Notons que, comme c'était le cas dans un système $M/M/1$, si $\rho \geq 1$, il ne peut y avoir d'état stable – c'est-à-dire que si le taux d'arrivée est au moins aussi important que le taux de service maximum possible ($\lambda \geq c\mu$), le système "explose".

Du point de vue du propriétaire, on peut vouloir s'assurer que les clients ne font pas la queue pendant un temps excessif, tout en réduisant au minimum le temps pendant lequel les serveurs restent en veille. Dans un système $M/M/c$, la probabilité que tous les guichets soient occupés (dans le régime stable) est donnée par

$$P(n \geq c) = \frac{(c\rho)^c}{c!(1-\rho)}\pi_0.$$

Le tableau 1 donnent les valeur de $P(n \geq c)$ dans diverses situations. On peut effectuer des calculs (assez difficiles, au demeurant) à l'aide de $W_s = \frac{1}{\mu}$, afin de montrer que

$$L_q = P(n \geq c)\frac{\rho}{1-\rho},\ W_q = \frac{L_q}{\lambda},\ W = \frac{1}{\mu} + W_q,\ L = \frac{\lambda}{\mu} + L_q.$$

Exemple 7. Une banque a deux guichets. À chaque heure, 80 clients arrivent à la banque en moyenne et attendent dans une file unique qu'un guichet se libère. Supposons que le temps de service moyen est de 1.2 minute, et que les intervalles d'arrivée et les temps de service suivent des distributions enponentielles. Que sont:

(a) le nombre prévu de clients dans la banque;
(b) la durée prévue du séjour dans la banque pour le client moyen, et
(c) la proportion du temps pendant laquelle les caissiers son inactifs.

ρ	$c=2$	$c=3$	$c=4$	$c=5$	$c=6$	$c=7$
.10	.02	.00	.00	.00	.00	.00
.20	.07	.02	.00	.00	.00	.00
.30	.14	.07	.04	.02	.01	.00
.40	.23	.14	.09	.06	.04	.03
.50	.33	.24	.17	.13	.10	.08
.55	.39	.29	.23	.18	.14	.11
.60	.45	.35	.29	.24	.20	.17
.65	.51	.42	.35	.30	.26	.21
.70	.57	.51	.43	.38	.34	.30
.75	.64	.57	.51	.46	.42	.39
.80	.71	.65	.60	.55	.52	.49
.85	.78	.73	.69	.65	.62	.60
.90	.85	.83	.79	.76	.74	.72
.95	.92	.91	.89	.88	.87	.85

Table 1. Probabilités $P(n \geq c)$ que tous les serveurs/guichets soient occupés dans un système $M/M/c$ pour $c = 2, \ldots, 7$ et $\rho \in (0.1, 0.95)$ [3, p.1088].

Solutions: nous faison affaire à un système $M/M/2$ avec $\lambda = 80$ clients/h et $\mu = 50$ clients/h. Ainsi, l'intensité de trafic est $\rho = \frac{80}{2 \cdot 50} = 0.80 < 1$ et le système admet un régime stable.

(a) En consultant la table 1, on constate que

$$P(n \geq 2) = 0.71$$

pour $\rho = 0.8$, d'où

$$L_q = P(n \geq 2) \cdot \frac{.8}{1-.8} = 2.84 \text{ clients}$$
$$L = \frac{80}{50} + L_q = 4.44 \text{ clients}.$$

(b) On a $W = \frac{L}{\lambda} = \frac{4.44}{80} = 0.055$ hr $= 3.3$ min.
(c) On détermine la proportion du temps pendant lequel un serveur est inactif en notant que les guichets sont

inactifs à tout instant où $n = 0$, et la moitié du temps lorsque $n = 1$ (par symétrie) . La probabilité qu'un serveur reste inactif est donc donnée par $\pi_0 + 0,5\pi_1$. Mais

$$\pi_0 = \left[1 + \frac{(2 \cdot .8)^2}{2!\,(1-.8)} + \sum_{n=1}^{2-1} \frac{2 \cdot .8^n}{n!}\right]^{-1} = \frac{1}{9}$$

et

$$\pi_1 = \frac{1.6}{1!}\pi_0 = 0.176,$$

d'où la probabilité qu'un caissier spécifique est inactif est de $0.111 + 0.5(0.176) = 0.199$.

Remarque importante: en général, les modèles de files d'attente ne sont pas compris dans la même mesure que le $M/M/1$ (et le $M/M/c$ dans une moindre mesure), et les mesures de performance ne sont souvent que des approximations fortement dépendantes des spécificités du problème en question.

C'est pourquoi les modèles $M/M/c$ sont parfois utilisés même lorsque leur emploi n'est pas justifié par les données et observations (la situation n'est pas sans rappeler l'utilisation répandue de la distribution normale dans divers problèmes de probabilité et de statistiques).

Dans de nombreuses applications, les distributions empiriques des arrivées et des temps de service approchent une distribution de Poisson et une distribution exponentielle, respectivement, de sorte que l'hypothèse n'est pas entièrement erronée, mais les simulations numériques ne devraient pas être évitées lorsque les écarts par rapport au modèle $M/M/c$ sont trop prononcés.

6. Application: temps d'attente dans les aéroports

En assurant un **contrôle pré-embarquement** efficace et efficient, l'*Administration canadienne de la sûreté du transport aérien* (ACSTA) assure la sécurité de tous les passagers et de l'équipage à bord des vols quittant les aéroports canadiens, tout en maintenant un équilibre approprié entre les effectifs du personnel de contrôle et le temps d'attente des passagers.

Le nombre de postes de contrôle actifs et le nombre de passagers ont une incidence sur les temps d'attente et, par conséquent, les réductions budgétaires ont un impact important sur le système.

De nombreux facteurs influencent le temps d'attente aux points de contrôle dans les aéroports canadiens: l'intensité des horaires des vols, le volume de passagers sur ces vols, le nombre de serveurs et les taux de traitement à un point de contrôle donné, etc.

L'un des objectifs de l'ACSTA est de faire en sorte que l'expérience du contrôle préembarquement dans les aéroports canadiens soit la plus efficace possible en réduisant au minimum le temps d'attente aux points de contrôle. Dans cette optique, le **modèle d'impact sur l'attente** ("Wait-Time Impact Model" — WTIM) a été conçu afin de réaliser les tâches suivantes:

1. fournir des estimés des taux d'arrivée des passagers (λ), des taux de traitement (μ) et du nombre de serveurs (c) à chaque point de contrôle, en utilisant les données disponibles sur le terrain;
2. calculer le niveau de qualité de service (QdS) (p_x, x) et déterminer quel niveau de service peut être atteint à chaque point de contrôle (c'est-à-dire, le pourcentage p de passagers qui attendront moins de x minutes, pour x fixe) pour un taux d'arrivée donné λ, taux de traitement μ, nombre de serveurs c;
3. obtenir le nombre moyen de serveurs c^* requis pour atteindre un niveau de qualité de service prescrit (p_x, x), compte tenu d'un profil d'arrivée λ^*;
4. fournir des courbes de niveau de QdS $(p_x(x), x)$ (c'est-à-dire des courbes de distribution cumulative) en fonction de divers taux d'arrivée et du nombre de serveurs actifs pour chaque point de contrôle (où x varie).

La structure de la file d'attente permet d'obtenir des informations intéressantes (comme on peut le constater à la Figure 5). Plus de détails sont disponibles dans la présentation qui accompagne ce rapport.

Références

[1] Erlang, A.K. [1909], "The Theory of Probabilities and Telephone Conversations", Nyt Tidsskrift for Matematik B, 20, 33.

[2] Berry, R. [2002], "Queueing Theory and Applications", 2e éd., PWS, Kent Publishing.

[3] Winston, W.L. [2004], Operations Research: Applications and Algorithms, 2e éd., PWS-Kent Publishing, Boston.

[4] Ross, S.M. [2014], Introduction to Probability Models, 11e éd., Academic Press.

[5] NPTEL – **Services Operations Management**, cours en ligne, module 9, lesson 4.

[6] NPTEL – **Services Operations Management**, cours en ligne, module 9, lesson 5.

[7] Nicol, D.M., **Introduction to Queueing Theorey**, notes de cours.

[8] **Quantitative Techniques for Management: Poisson and Exponential Distributions**, sur wisdomjobs.com

[9] Schwartz, B. [2016], **The Essential Guide to Queueing Theory**, Vivid Cortex.

[10] Kendall, D.G. [1953], "Stochastic Processes Occurring in the Theory of Queues and their Analysis by the Method of the Imbedded Markov Chain", Ann.Math.Stat. 24 (3): 338.

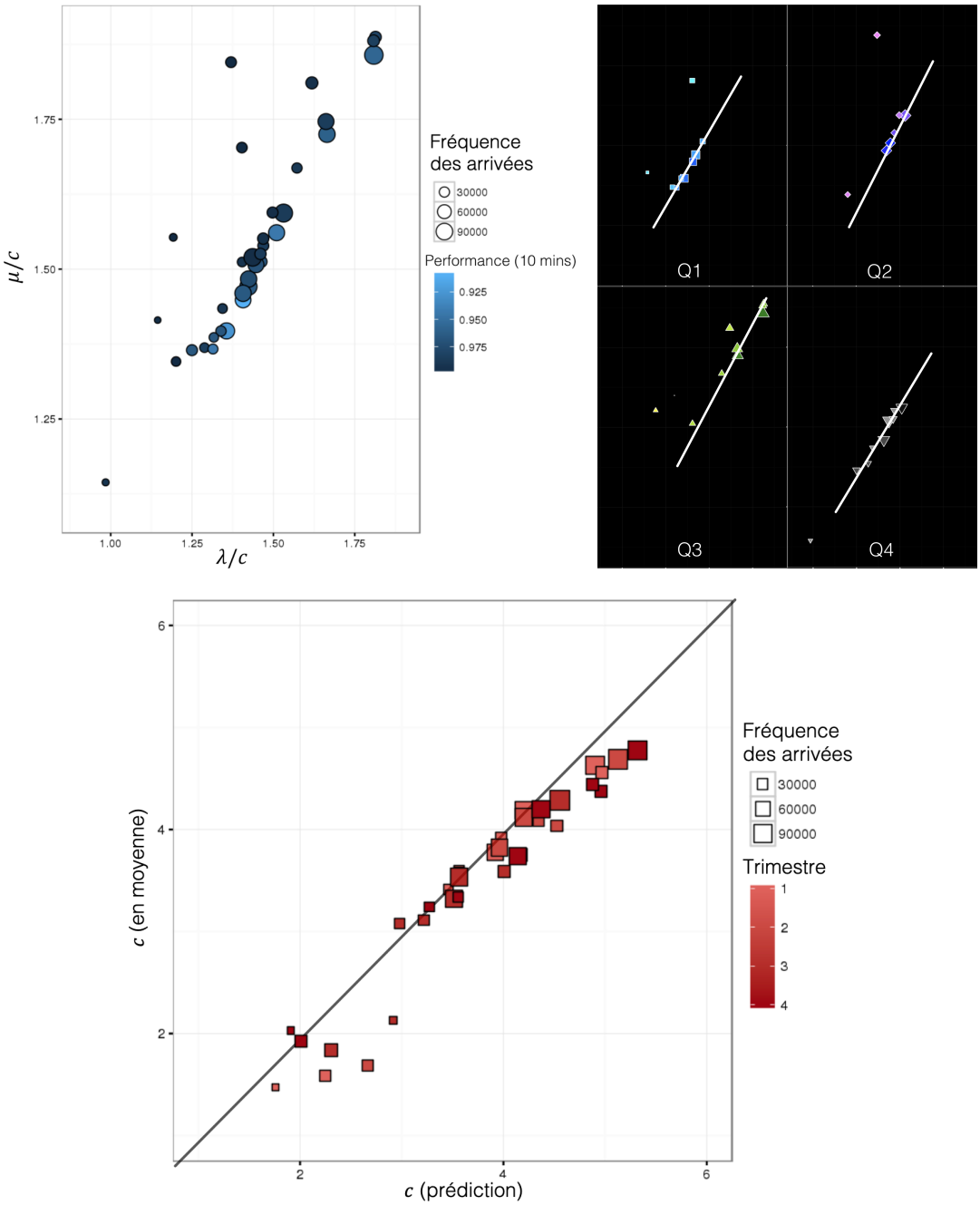

Figure 5. En haut: visualisation des paramètres de file d'attente d'un point de contrôle spécifique – λ, μ, $\bar{c}$, nombre de passagers et performance (pourcentage de voyageurs attendant moins de 15 minutes pour être contrôlés); la relation entre $\lambda/\bar{c}$ et $\mu/\bar{c}$ est pratiquement linéaire (à gauche), ce qui est plus facile à voir au niveau des trimestres (à droite). En bas: prédiction du nombre moyen de serveurs par rapport au nombre réel de serveurs nécessaires pour maintenir les performances prescrites, avec le nombre de passagers, par trimestre. La ligne de prédiction idéale est ajoutée pour faciliter la comparaison.

[11] Kleinrock, L. [1975], Queueing Systems, vol. 1, Wiley.

[12] Gross, D., Shortle, J.F., Thompson, J.M., Harris, C.M. [2008], Fundamentals of Queueing Theory, Wiley.

[13] Burke, P.J. [1956], The Output of a Queueing System, Operations Reserach vol 4 (6): 699704.

[14] Newell, G.F. [1971], Applications of Queueing Theory, Chapman and Hall.

[15] Walrand, J. [1983], A probabilistic look at networks of quasi-reversible queues, IEEE Transactions on Information Theory, vol 29 (6): 825831.

[16] **Queueing Theory** sur Wikipedia.

[17] Erickson, W. [1973], Management Science and the Gas Shortage, Interfaces 4:47–51.