Principes fondamentaux des systèmes de files d'attente

La théorie des files d'attente est une branche des mathématiques qui étudie et modélise les files d'attente.

Ce module définit les éléments de base des systèmes de files d'attente.

Aperçu

Introduction

Terminologie de la théorie des files d'attente

- Distribution de Poisson
- Distribution exponentielle
- Distribution d’Erlang
- Processus d'entrées
- Processus de sorties

Notation de Kendall-Lee

La loi de Little

Système de file d'attente $M/M/1$

- $M/M/1$ à capacité limitée

Système de file d'attente $M/M/c$

Introduction

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?

Introduction

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.

Introduction

Supposons qu'il y ait dans une épicerie une seule ligne de caisse.

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.

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.

Mais ce n'est pas ce qui se passe en réalité; beaucoup d'acheteurs font la queue, et ils doivent attendre longtemps avant d'être servis.

Introduction

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;
- les **tâches de taille irrégulière** – les achats ne sont pas tous traités en 4.5 minutes;
- le **gaspillage** – le temps perdu ne peut jamais être rattrapé; les clients se chevauchent parce que le deuxième client arrive avant que le premier n'ait eu le temps de finir.

Introduction

La file d'attente **s'aggrave** sous les conditions suivantes:

- une **utilisation élevée des serveurs** – plus la caissier est occupé, 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;
- 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.

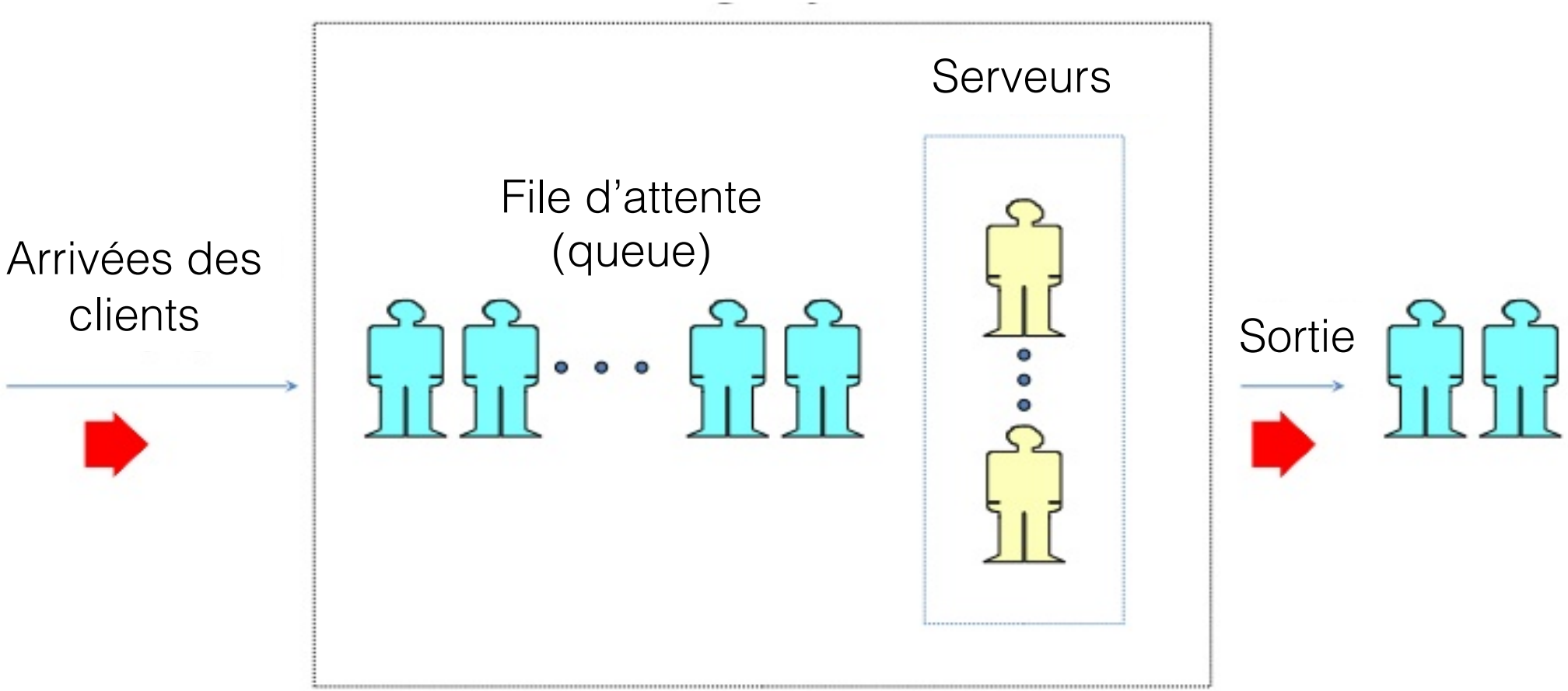
Introduction

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

- les **clients** (utilisateurs) sont les unités de travail désservies par le système (un client peut être une personne réelle, une requête web, une requête de base de données, une pièce à usiner, etc.);
- les **serveurs** sont les objets qui effectuent le traitement (le caissier de l'épicerie, un serveur web, un serveur de base de données, une fraiseuse, etc.);
- les **queues** (files d’attente) sont les endroits où les unités de travail attendent lorsque le serveur est occupé.

Terminologie des files d'attente

Composantes d’un système de file d’attente



Terminologie des files d’attente

Afin de décrire les files d'attente, nous devons d'abord connaître et comprendre

- certaines distributions utiles,
- les processus d’entrée, et
- les processus de sortie.

Distributions – Poisson, exponentielle

Deux distributions jouent un rôle primordial dans la théorie et les applications des systèmes de files d'attente:

- la distribution de Poisson **compte le nombre d'événements discrets** se produisant dans une période de temps fixe;
- la distribution exponentielle **mesure l'intervalle de temps entre deux arrivées**.

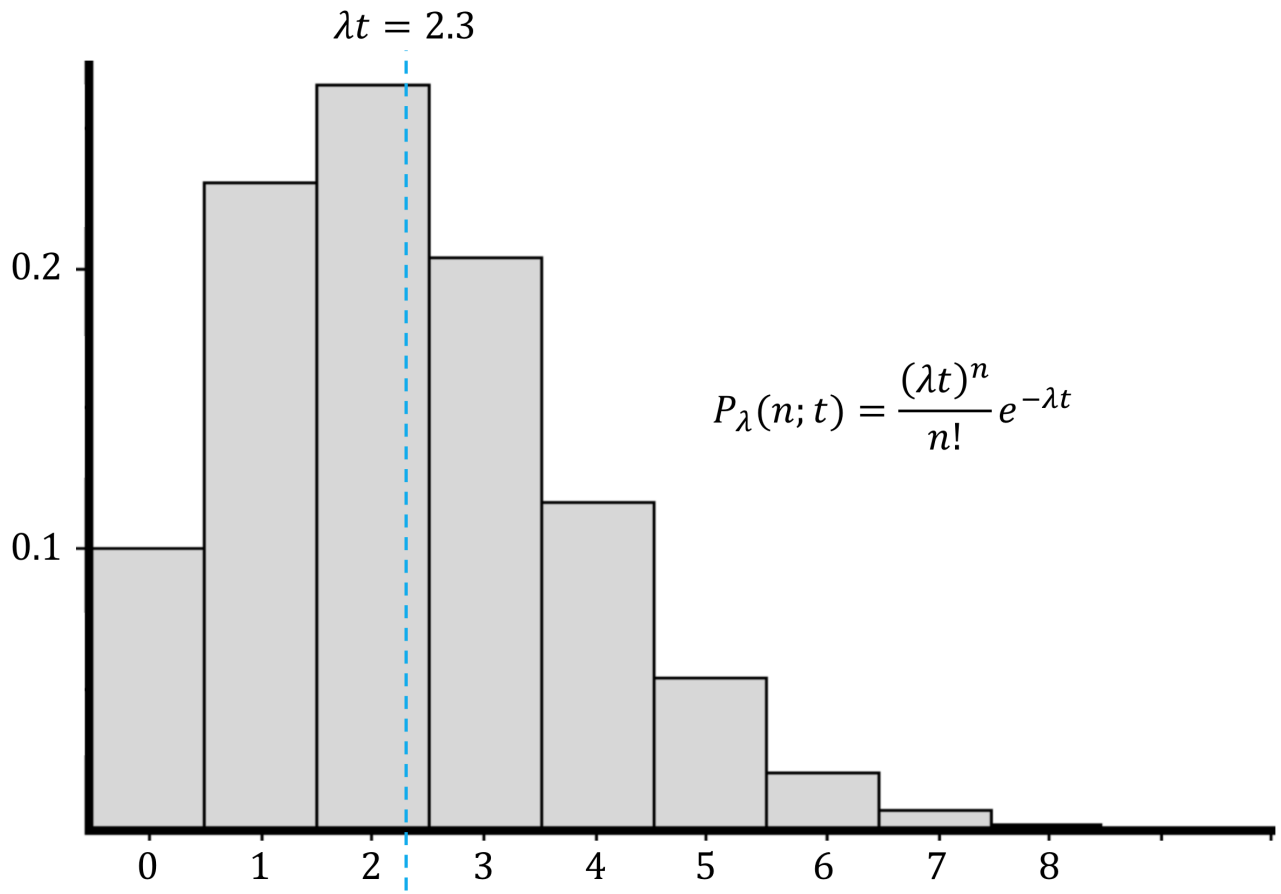
Distribution de Poisson

Si le taux moyen d'arrivées est λ (mesuré en secondes, minutes, heures, jours, etc.), la probabilité d'oberver n arrivées dans une intervalle de temps t est:

$$P(n,t) = \frac{(\lambda t)^n}{n!}\, e^{-\lambda t}$$

où $n = 0,1,2, \ldots$

Distribution de Poisson



Distribution de Poisson

En moyenne, 50 clients se pointent au café du coin à chaque heure.

Si les arrivées ont un profil de distribution de Poisson, quelle est la probabilité d'observer exactement 20 clients dans une période de 30 minutes?

Distribution de Poisson

Nous avons $\lambda = 50$ clients/h, $t = 30$ minutes $= 0.5$ h, et $n = 20$, d'où

$$P(20,0.5) = \frac{(50 \times 0.5)^{20}}{20!} e^{-(50 \times 0.5)} \approx 5\%.$$

Distribution exponentielle

Le temps d'attente entre deux arrivées successives est l'**intervalle d'arrivée**.

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

$$f(t) = \mu e^{-\mu t}.$$

La probabilité que le temps d'attente W pour un client dans la file dure moins de t unités de temps est

$$P(W \leq t) = 1 - e^{-\mu t}.$$

Distribution exponentielle

Dans un fast-food, on sert en moyenne 9 clients par heure.

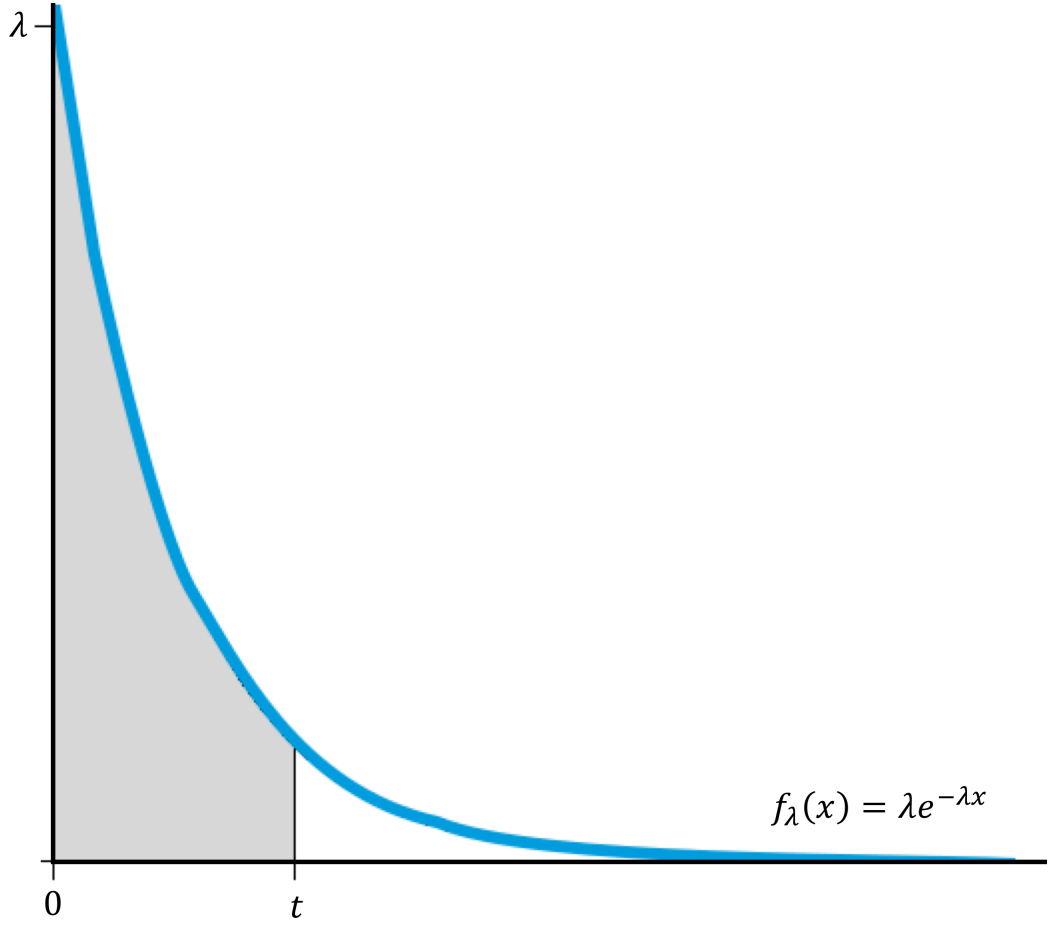
Si le temps de service suit une exponentielle, quelle est la probabilité qu'un client sera servi au plus 15 minutes après son arrivée dans la file?

Distribution exponentielle

Soit W le temps d'attente moyen. Nous avons $\mu = 9$ clients/h, et $t = 15$ minutes $= 0.25$ h, d'où

$$P(W \leq 15 \text{ minutes}) = 1 - e^{-9 \times 0.25} \approx 89\%.$$

Distribution exponentielle



Perte de mémoire

$$P(X > t + h \mid X > h) = P(X > t), \qquad \forall h$$

La distribution exponentielle est **sans mémoire**, c'est-à-dire que la distribution du temps d'attente jusqu'à la prochaine arrivée est indépendente du temps depuis la dernière arrivée …

(est-ce que c'est vrai pour le transport en commun?)

Perte de mémoire

Le temps d'attente w qu'un client passe dans la file d'attente d'une banque est distribué selon une exponentielled dont la moyenne est 10 minutes.

Alors

$$P(w > 15 | w > 10)$$
$$= P(w > 5) = e^{-5/10}$$
$$\approx 61\%$$

Ceux qui ont déjà attendu 10 minutes vont attendre plus de 15 minutes en tout avec 61% de probabilité.

Distribution d'Erlang

La distribution exponentielle n'est pas toujours un modèle approprié des intervalles d'arrivées; le temps d'attente n'est pas toujours sans mémoire, par exemple (cf. transport en commun).

Une approche alternative utilise la distribution d'**Erlang** $\mathcal{E}(R,k)$, une variable continue à 2 paramètres $R > 0, k \in \mathbb{Z}^+$ de densité:

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

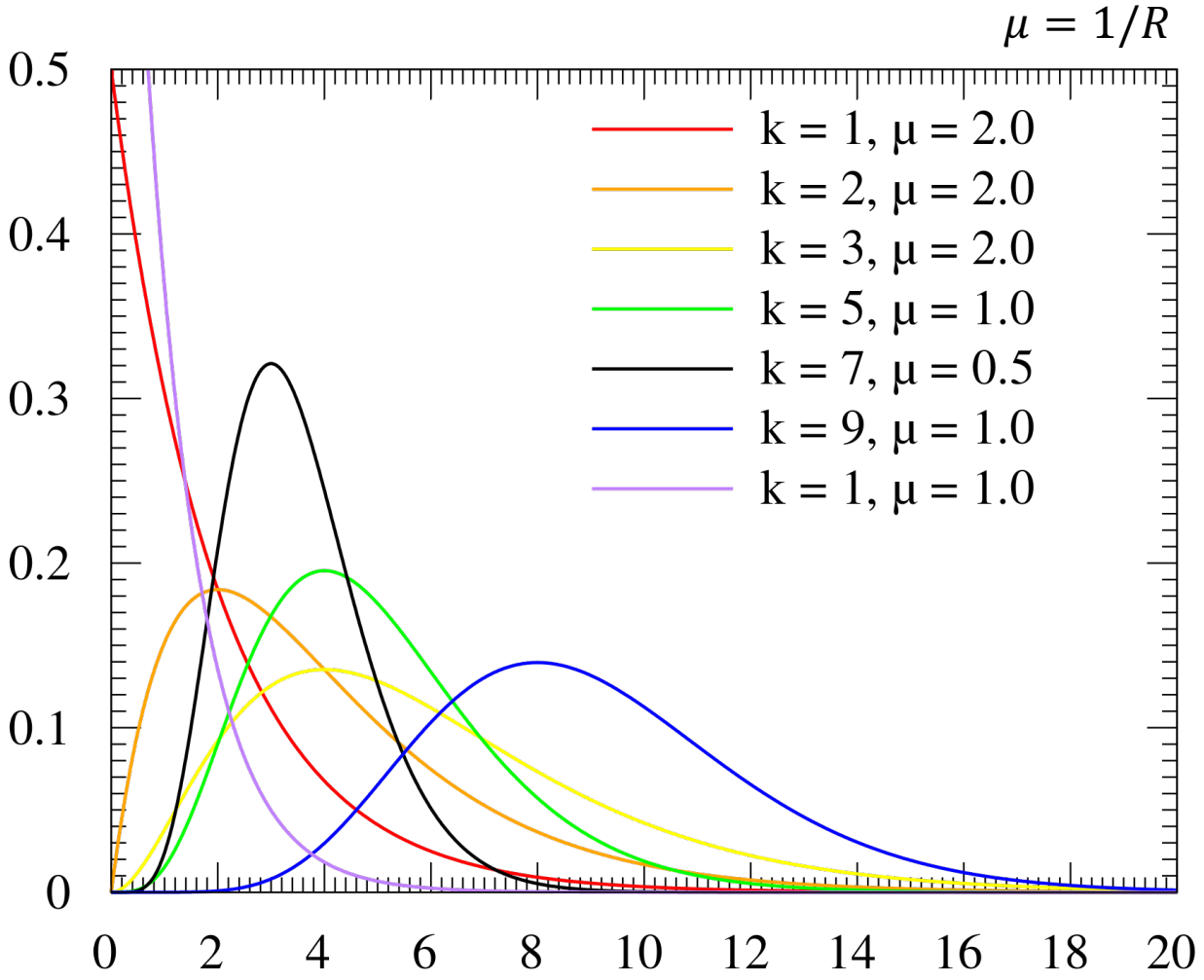
Distribution d'Erlang

Si $k = 1$, $\mathcal{E}(R, 1) = \mathrm{Exp}(R)$. En général, on écrit $R = k\lambda$, et on obtient une décomposition k exponentielles indépendentes:

$$\mathcal{E}(k\lambda, k) = \mathrm{Exp}(k\lambda) + \cdots + \mathrm{Exp}(k\lambda) = \sum_{i=1}^{k} \mathrm{Exp}(k\lambda)$$

Si les intervalles d'arrivées suivent une $\mathcal{E}(k\lambda, k)$, nous supposons que les utilisateurs passent par k phases sans mémoire avant d'être servis.

Distribution d'Erlang



Processus d’entrée

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.

On modélise souvent les arrivées à l’aide d’une distribution de Poisson (bien qu’on puisse aussi se servir d’autres distributions).

Processus de sortie

Combien de temps faut-il pour servir un client?

Dans la plupart des cas, on suppose que la distribution des temps de service est **indépendente du nombre de clients présent dans le système**.

- est-ce réaliste?

Un autre facteur: serveurs en parallèles ou en série?

On modélise souvent les sorties à l'aide d'une distribution exponentielle

- situations de maintenance ou de service non-planifié ⇒ bonne idée.

Exemples

Situation	Arrivées	Sorties
banque	clients arrivent à la banque	caissiers servent les clients
pizzéria	demandes de livraisons sont reçues	pizzéria envoient les pizzas pour la livraison
banque de sang à l'hôpital	pintes de sang arrivent à l'hôpital	patients utilisent les pintes de sang selon le groupe sanguin
chantier naval	navires arrivent au chantier naval pour se faire réparer	navires sont réparées et renvoyés à l'eau

Notation de Kendall-Lee

Notation de Kendall-Lee

On décrit les systèmes de file d'attente à l'aide de 6 attributs:

$$1/2/3/4/5/6$$

La 1e caractéristique précise la nature du **processus d'arrivée**.

On utilise les abréviations standard suivantes:

- $M =$ intervalles d'arrivées exponentiels, ind. et ident. dist. (iid);
- $D =$ intervalles d'arrivées déterministe, iid;
- $E_k =$ intervalles d'arrivées $\mathcal{E}(R, k)$, iid
- $G =$ intervalles d'arrivées généraux, iid

Notation de Kendall-Lee

La 2e caractéristique précise la nature du **temps de service**:

- M = temps de service exponentiels et iid;
- D = temps de service déterministes et iid, etc.

La 3e caractéristique représente le # de **serveurs parallèles**.

La 4e 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 générale.

Notation de Kendall-Lee

La 5e caractéristique précise le **nombre maximal d'utilisateurs** pouvant être accomodés par le système.

La 6e caractéristique donne la **taille de la population** dont sont issus les utilisateurs.

Dans de nombreux modèles importants, $4/5/6$ correspond à $GD/\infty/\infty$. Dans ce cas, ces attributs sont souvent omis de la description de la file d'attente.

Notation de Kendall-Lee

Nom (notation de Kendall)	Exemple
Système simple $(M/M/1)$	Service à la clientèle à l'épicerie
Système multi-serveurs $(M/M/c)$	Comptoir des billets d'avion
Service constant $(M/D/1)$	Lavage automatique des voitures
Service général $(M/G/1)$	Réparation d'automobile
Capacité limitée $(M/M/1/N)$	Salon de coiffure avec N places d'attente

Loi de Little

Loi de Little

$\lambda =$ nombre moyen d'arrivées dans le système par unité de temps

$L =$ nombre moyen d'utilisateurs présents dans le système

$L_q =$ nombre moyen d'utilisateurs qui font la queue

$L_s =$ nombre moyen d'utilisateurs en service

$W =$ temps moyen qu'un utilisateur passe dans le système

$W_q =$ temps moyen qu'un utilisateur passe dans la file d'attente

$W_s =$ temps moyen qu'un utilisateur passe dans la file d'attente

Little's Queuing Formula

Pour la plupart des systèmes de file d'attente pour lesquels un **régime stable existe**, la loi de Little se résume par:

- $L = \lambda W$
- $L_q = \lambda W_q$
- $L_s = \lambda W_s$

Exemple: si $\lambda = 46$ clients arrivent dans un restaurant durant chaque périod d'une heure, en moyenne, et si les clients attendent $W = 10$ minutes avant d'être servis, en moyenne, il y a $L = 46 \times 1/6 \approx 7.7$ clients en attente à tout moment, en moyenne.

Le système $M/M/1$

Le système $M/M/1$

Dans un système de file d'attente $M/M/1$, les temps de service et les intervalles d'arrivée sont exponentiels de taux respectifs μ et λ, il n'y a qu'un unique serveur, et le régime est stable.

Soit $\rho = \lambda/\mu$ l'**intensité du trafic**. Lorsque $0 \leq \rho < 1$, il y a exactement n utilisateurs dans le système avec probabilité

$$\rho^n (1 - \rho), \qquad n = 0,1,2, \ldots$$

Il n'y a aucun utilisateur dans le système avec probabilité $1 - \rho$.

Le système $M/M/1$

Nombre moyen d'utilisateurs en service: $L_s = \rho$

Nombre moyen d'utilisateurs faisant la queue: $L_q = \frac{\rho^2}{1-\rho}$

Nombre moyen d'utilisateurs dans le système: $L = L_q + L_s = \frac{\lambda}{\mu - \lambda}$

Temps moyen passé en service: $W_s = \frac{1}{\mu}$

Temps moyen passé dans la queue: $W_q = \frac{L_q}{\lambda} = \frac{\lambda}{\mu(\mu - \lambda)}$

Temps moyen passé dans le système: $W = W_q + W_s = \frac{L}{\lambda} = \frac{1}{\mu - \lambda}$

Le système $M/M/1$

Pour $1 \leq \rho$, le taux d'arrivée λ est supérieur au taux de service μ, et la file d'attente ne se dégage jamais (régime instable).

Lorsque $\rho \to 1$, $W, W_q \to \infty$. Lorsque $\rho \to 0$, $W_q \to 0$ et $W \to 1/\mu$, le **temps de service moyen**.

Station-service à 1 pompe

Supposons que tous les propriétaires de voitures fassent le plein lorsque leur réservoir est exactement à moitié vide.

En moyenne, 7.5 clients arrivent à la station-service à chaque heure (la station n'a qu'une pompe).

Il faut en moyenne à chaque voiture 4 minutes pour faire le plein. On suppose de plus que le système est $M/M/1$.

[Erickson, W., 1973]

Station-service à 1 pompe

Pour ce système, la sation-service reçoit $\lambda = 7.5$ clients par heure, et a la capacité de servir $\mu = 60/4 = 15$ clients par heure.

Ainsi,

- l'intensité du traffic: $\rho = \frac{\lambda}{\mu} = \frac{7.5}{15} = 0.5$;
- le nombre moyen d'utilisateurs dans le système: $L = \frac{\lambda}{\mu - \lambda} = \frac{7.5}{15 - 7.5} = 1$;
- le temps moyen passé dans le système: $W = \frac{L}{\lambda} = \frac{1}{7.5} = 0.13$ heure, ou $6\ \frac{2}{3}$ minutes.

Station-service à 1 pompe

Supposons que tous les propriétaires de voitures fassent le plein lorsque leur réservoir est exactement au ¾ plein en raison d'une pénurie de gaz.

En conséquence, le temps de service est réduit, à $3\ \frac{1}{3}$ minutes, mettons.

De quelle façon est-ce que cela affecte les valeurs de L et W?

[Erickson, W., 1973]

Station-service à 1 pompe

Dans ce cas, la station-service reçoit $\lambda = 2(7.5)=15$ clients par heure, et a la capacité de servir $\mu = \frac{60}{10/3} = 18$ clients par heure.

Ainsi,

- l'intensité du traffic: $\rho = \frac{\lambda}{\mu} = \frac{15}{18} = \frac{5}{6}$;
- le nombre moyen d'utilisateurs dans le système: $L = \frac{\lambda}{\mu - \lambda} = \frac{15}{18 - 15} = 5$;
- le temps moyen passé dans le système: $W = \frac{L}{\lambda} = \frac{5}{15} = 0.33$ heure.

Les achats dans la panique doublent (et +) le temps d'attente!

Station-service à 1 pompe

L'exemple précédent illustre le fait qu'à mesure que ρ se rapproche de 1, L (et par conséquent W) augmente rapidement.

ρ	L dans le système $M/M/1$
0.30	0.43
0.60	1.50
0.80	4.00
0.90	9.00
0.95	19.00
0.99	99.00

Capacité limitée dans un système $M/M/1$

La capacité d'une file d'attente ne peut être infinie; elle est limitée par les exigences du temps, de l'espace, de la politique de service.

Exemples: le stationnement des véhicules dans un supermarché est limité aux places de la zone de stationnement; la disposition des places assises dans un restaurant est limitée aux nombres de places.

La probabilité qu'il n'y ait aucun utilisateur dans un tel système est

$$p_0 = \frac{1-\rho}{1-\rho^{N+1}}$$

où N est le nombre maximal autorisé par le système.

Capacité limitée dans un système $M/M/1$

La probabilité qu'il y ait exactement n utilisateurs dans le système est

$$p_n = \begin{cases} \rho^n \, p_0, & n = 1,2, \dots, N \\ 0, & n = N+1, N+2, \dots \end{cases}$$

Le nombre moyen d'utilisateurs dans le système est

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

et $L_s = 1 - p_0$, $L_q = L - L_s$.

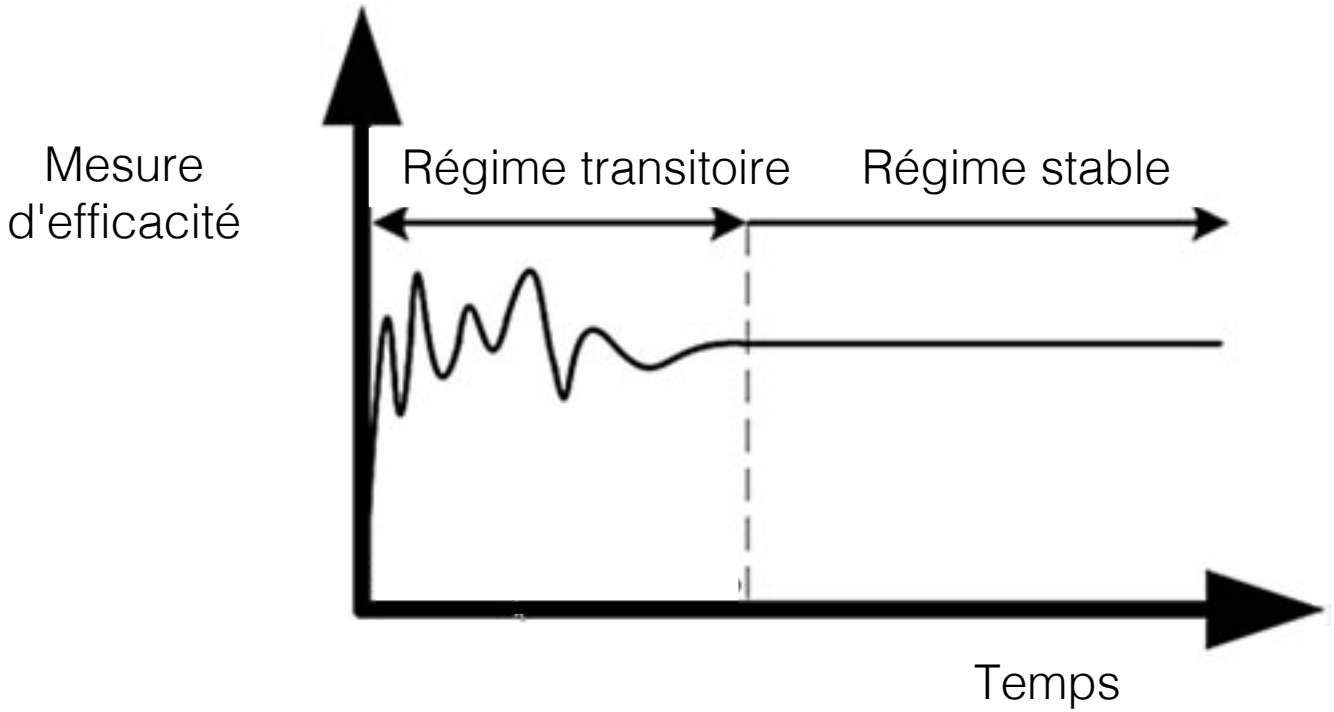
Capacité limitée dans un système $M/M/1$

Il y a $\lambda - \lambda p_N$ arrivées dans le système par unité de temps, en moyenne, en raison de sa capacité limite, d'où:

$$W = \frac{L}{\lambda(1 - p_N)}, \qquad W_q = \frac{L_q}{\lambda(1 - p_N)}.$$

Cette restriction impose l'existence d'un **régime stable** puisque même si $\lambda \geq \mu$, il ne peut y avoir plus de N clients dans le système.

Régime stable (équilibre de file)



Salon de coiffure

Un barbier travaille seul dans un salon de coiffure avec 10 sièges d'attente.

Les intervalles d'arrivée suivent une exponentielle; en moyenne, 20 clients potentiels se présentent à chaque heure au salon.

Le barbier prend en moyenne 12 minutes pour couper les cheveux des clients (le temps de coupe suit aussi une exponentielle).

En moyenne, combien de temps un client passe-t-il au salon?

Salon de coiffure

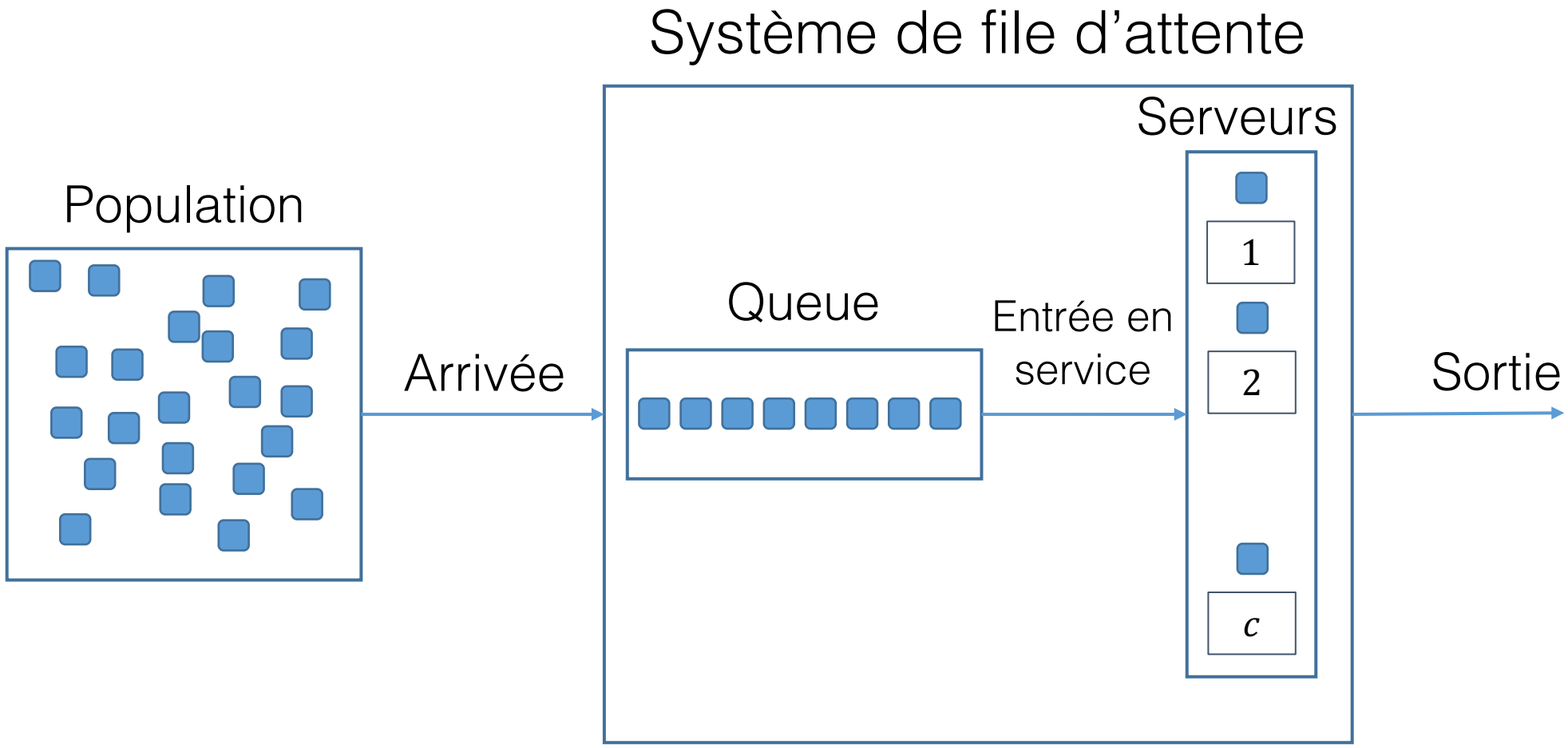
Selon l'énoncé du problème, $N = 10$, $\lambda = 20$ arrivées/h par heure, et $\mu = 60/12 = 5$ clients par heure. L'intensité du traffic est $\rho = \frac{20}{5} = 4$, et nous obtenons

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

d'où $W = \frac{L}{\lambda} = 1.93$ heures.

Ce salon est bondé, et le barbier serait bien avisé d'engager au moins un autre coiffeur – quel effet cela pourrait-t-il avoir sur L et W?

Le système $M/M/c$


Système de file d'attente
Population
Arrivée
Queue
Entrée en service
Serveurs
1
2
c
Sortie

Le système $M/M/c$

Les mêmes hypothèses que dans le cas $M/M/1$, mais avec c guichets pouvant servir des **clients provenant d'une seule file d'attente**, comme dans une banque, par exemple.

Si chaque serveur complète le service au taux μ, le taux de service du système est $c\mu$.

L'intensité du trafic est $\rho = \frac{\lambda}{c\mu}$, et on suppose toujours $\rho \leq 1$.

Si $\rho \geq 1$, il n'y a pas de régime stable et la file ne se vide jamais.

Le système $M/M/c$

Dans le **régime stable** (à long terme), la probabilité que tous les serveurs soient occupés est:

$$P(n \geq c) = \frac{(c\rho)^c}{c!\,(1-\rho)}\, p_0$$

où p_0 est la probabilité qu'il n'y ait pas de client dans le système (formule omise pour des raisons de simplicité). Nous avons alors

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

Le système $M/M/c$

$P(n \geq c)$ dans une variété de situations.

ρ	$c = 2$	$c = 3$	$c = 4$	$c = 5$	$c = 6$	$c = 7$
.10	.02	.00	.00	.00	.00	.00
.30	.14	.07	.04	.02	.01	.00
.50	.33	.24	.17	.13	.10	.08
.70	.57	.51	.43	.38	.34	.30
.80	.71	.65	.60	.55	.52	.49
.90	.85	.83	.79	.76	.74	.72
.95	.92	.91	.89	.88	.87	.85

Guichets de banque

Une banque a 2 guichets.

80 clients arrivent à chaque heure, en moyenne, et forment une ligne en attendant un guichet libre.

Le temps de service moyen est 1.2 minutes par client.

Quel est le nombre attendu de clients dans la banque à tout instant?

Quelle est la durée prévue du séjour dans la banque pour le client moyen?

Guichets de banque

Nous faisons affaire à un système $M/M/2$ où $\lambda = 80$ clients/h et $\mu = 50$ clients par heure, d'où $\rho = \frac{\lambda}{2\mu} = 0.8.$ En consultant la table, nous trouvons $P(n \geq 2) = 0.71.$

Ainsi

- $L_q = \frac{0.8}{1-0.8}(0.71) = 2.84$ clients/h
- $L = \frac{80}{50} + 2.84 = 4.44$ clients/h
- $W = \frac{L}{\lambda} = 0.055 \text{ heures} = 3.3 \text{ minutes}$