Files d'attente

Anaïs Vergne et Céline Comte

30 octobre 2018

1 Introduction

Notation de Kendall

$$A/B/C[/D/E]$$

A Processus d'arrivée des clients dans la file
M pour *Markovian* ou *memoryless*, correspondant à un processus d'arrivée Poisson, D pour *determinist*, G pour *generally distributed*.

B Distribution du temps de service
M pour *Markovian* ou *memoryless*, correspondant à un temps de service distribué selon la loi exponentielle, D pour *determinist*, G (ou GI) pour *generally distributed* (et *independent*).

C Nombre de serveurs

D Nombre de places dans la file, en comptant les clients en service sur un serveur
Valeur par défaut : ∞.

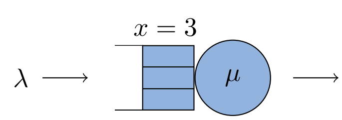
E Discipline de service
FIFO pour *first-in, first-out* (premier arrivé, premier servi), LIFO pour *last-in, first-out* (qui peut être préemptif ou non), PS pour *processor-sharing*.
Valeur par défault : FIFO.

Dans ce cours, on suppose que les temps d'inter-arrivée des clients successifs sont indépendants et identiquement distribués. De même, on suppose que le temps de service d'un client est indépendant du temps de service d'un autre client.

Exemples On se contente de deux exemples qui seront revus plus tard dans le cours.

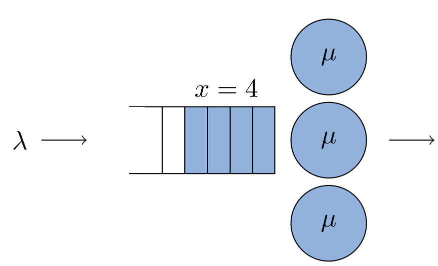
— File M/M/1/∞/PS

M Processus d'arrivée : Poisson de paramètre $\lambda > 0$.
M Distribution du temps de service :
exponentielle de paramètre $\mu > 0$.
1/∞ 1 serveur, une infinité de places dans la file.
PS Discipline de service processor-sharing.



— File M/M/3[/∞/FIFO]

M Processus d'arrivée : Poisson de paramètre $\lambda > 0$.
M Distribution du temps de service :
exponentielle de paramètre $\mu > 0$.
3/∞ 3 serveurs, une infinité de places dans la file.
FIFO Discipline de service first-in, first-out.

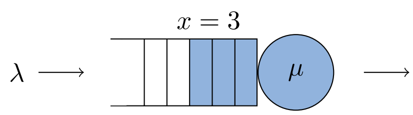


2 La file M/M/1

Modèle

- ***M*** Les clients arrivent selon un processus de Poisson de paramètre $\lambda > 0$.
- ***M*** Le temps de service de chaque client suit une loi exponentielle de paramètre $\mu > 0$ (et est indépendant du temps de service des autres clients).
- ***1/∞*** Un seul serveur, une infinité de places dans la file.
- ***FIFO*** Discipline de service first-in, first-out.

La charge de la file est notée $\rho = \frac{\lambda}{\mu}$. C'est une grandeur sans unité, exprimée en *Erlang*. On suppose dans la suite que $\rho < 1$. Cette condition va s'avérer nécessaire et suffisante pour garantir la stabilité de la file. On suppose de plus que la discipline de service est FIFO, même si la plupart des résultats sont valables pour d'autres disciplines de service.

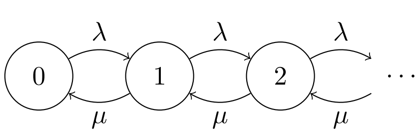


(a) File M/M/1/∞/FIFO (b) File M/M/1/∞/PS

FIGURE 1 – Représentation d'une file M/M/1 sous plusieurs disciplines de service

État de la file Pour chaque $t \in \mathbb{R}_+$, on note $\mathbf{X}_t$ la variable aléatoire qui compte le nombre de clients dans la file à l'instant t. On suppose que la file est vide à l'instant 0, c'est-à-dire que $\mathbf{X}_0 = 0$. $(\mathbf{X}_t)_{t \geq 0}$ est un processus de Markov homogène sur $\mathbb{N}$. Plus précisément, $(\mathbf{X}_t)_{t \geq 0}$ est un processus de naissance et de mort dont le taux de naissance est constant égal à λ et le taux de mort est constant égal à μ.
On peut construire une réalisation du processus pas-à-pas (sur un intervalle de temps fini) en tirant, pour chaque instant t où se produit une transition, le temps avant la prochaine transition et son type (arrivée ou départ d'un client) :

- Si $\mathbf{X}_t = 0$, alors la prochaine arrivée se produit après un temps $\sim \mathrm{Exp}(\lambda)$.
- Si $\mathbf{X}_t > 0$, alors la prochaine arrivée se produit après un temps $\sim \mathrm{Exp}(\lambda)$ et le prochain départ se produit après un temps $\sim \mathrm{Exp}(\mu)$. De façon équivalente,
  - → le prochain événement se produit après un temps $\sim \mathrm{Exp}(\lambda + \mu)$;
  - → indépendamment du temps qui s'est écoulé, il s'agit d'une arrivée avec probabilité $\frac{\lambda}{\lambda+\mu}$ et d'un départ avec probabilité $\frac{\mu}{\lambda+\mu}$.



(a) Diagramme de transition

$$A = \begin{bmatrix} -\lambda & \lambda & & & \\ \mu & -(\lambda+\mu) & \lambda & & \\ & \mu & -(\lambda+\mu) & \lambda & \\ & & \ddots & \ddots & \ddots \end{bmatrix}$$

(b) Générateur infinitésimal (les entrées vides valent 0)

FIGURE 2 – Processus de Markov décrivant l'état de la file M/M/1 représentée en Figure 1

Distribution stationnaire On cherche les mesures π, définies sur $\mathbb{N}$, qui sont solutions de l'équation matricielle $\pi A = 0$, où A est le générateur infinitésimal du processus $(\mathbf{X}_t)_{t\geq 0}$, rappelé en FIGURE 2b. Cela revient à résoudre le système d'équations :

$$\begin{cases} -\lambda\pi(0) + \mu\pi(1) = 0, \\ \lambda\pi(i-1) - (\lambda+\mu)\pi(i) + \mu\pi(i+1) = 0, \quad \forall i \geq 1. \end{cases}$$

On peut remarquer que les deux premières équations

$$-\lambda\pi(0) + \mu\pi(1) = 0 \quad \text{et} \quad \lambda\pi(0) - (\lambda+\mu)\pi(1) + \mu\pi(2) = 0,$$

impliquent que

$$-\lambda\pi(1) + \mu\pi(2) = 0.$$

En injectant cette égalité dans l'équation obtenue pour $i = 2$, on obtient

$$-\lambda\pi(2) + \mu\pi(3) = 0.$$

En raisonnant par récurrence sur $i \geq 1$, on peut ainsi montrer que le système ci-dessus est équivalent à

$$-\lambda\pi(i-1) + \mu\pi(i) = 0, \quad \forall i \geq 1.$$

On déduit directement que $\pi(i) = \pi(0)(\frac{\lambda}{\mu})^i = \pi(0)\rho^i$ pour chaque $i \in \mathbb{N}$. On obtient la valeur de $\pi(0)$ par normalisation (en utilisant le fait que $\rho < 1$) :

$$1 = \sum_{i\in\mathbb{N}} \pi(i) = \pi(0) \sum_{i\in\mathbb{N}} \rho^i = \pi(0)\frac{1}{1-\rho},$$

d'où $\pi(0) = 1 - \rho$ et $\pi(i) = (1-\rho)\rho^i$ pour chaque $i \in \mathbb{N}$.
La distribution stationnaire de l'état d'une file M/M/1 de charge ρ est donc *géométrique* de paramètre ρ. On peut remarquer que cette distribution stationnaire ne dépend que du rapport $\rho = \frac{\lambda}{\mu}$, et pas des valeurs exactes de λ et de μ. Augmenter ou diminuer ces quantités en gardant leur rapport constant revient à “accélérer” ou à “ralentir” l'évolution de la file, mais n'a pas d'impact sur son comportement à l'état stationnaire.
Dans la suite, on note $\mathbf{X}$ une variable aléatoire de distribution π. On s'intéresse au comportement de la file lorsqu'elle est à l'état stationnaire.

Loi de conservation La probabilité qu'il y ait au moins un client dans la file est donnée par

$$\mathbb{P}(\mathbf{X} \geq 1) = \sum_{i\geq 1} \pi(i) = 1 - \pi(0) = 1 - (1-\rho) = \rho.$$

La charge ρ donne donc la probabilité que le serveur soit actif (à un instant donné). Par ergodicité, c'est aussi la fraction du temps où le serveur est actif.
Remarquons que l'on peut réécrire l'égalité $\rho = 1 - \pi(0)$ sous la forme

$$\lambda = \mu(1 - \pi(0)).$$

On obtient la *loi de conservation* selon laquelle, en régime stationnaire, le taux d'arrivée des clients dans la file (donné par λ) doit être égal au taux de service effectivement fourni par le serveur (donné par $\mu(1-\pi(0))$).

Nombre moyen de clients dans la file L'espérance L du nombre de clients dans la file est donnée par

$$L = \mathbb{E}(\mathbf{X}) = \sum_{i=0}^{+\infty} i\pi(i) = \sum_{i=1}^{+\infty} i(1-\rho)\rho^i = \rho(1-\rho)\sum_{i=1}^{+\infty} i\rho^{i-1} = \rho(1-\rho)\frac{1}{(1-\rho)^2},$$

ce qui donne

$$L = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu-\lambda}.$$

Comme on peut s'y attendre, lorsque la charge ρ de la file augmente :

— la fraction de temps $1-\pi(0) = \rho$ où le serveur est actif augmente, i.e., le serveur est utilisé plus souvent ;

— le nombre moyen de clients dans la file $\mathbb{E}(\mathbf{X}) = \frac{\rho}{1-\rho} = \frac{1}{1-\rho} - 1$ augmente aussi.

Lorsque $\rho \to 1$, la fraction du temps où le serveur est actif tend vers 1 et le nombre moyen de clients dans la file tend vers $+\infty$.

Nombre moyen de clients *en attente* dans la file Seul le premier client est en service, donc le nombre de clients en attente dans la file est donné par $\mathbf{X}-1$ dès que $\mathbf{X} \geq 1$:

$$\mathbb{E}((\mathbf{X}-1)_+) = \sum_{i\geq 1}(i-1)\pi(i) = \underbrace{\sum_{i\geq 1} i\pi(i)}_{\mathbb{E}(\mathbf{X})} - \underbrace{\sum_{i\geq 1}\pi(i)}_{1-\pi(0)} = \frac{\rho}{1-\rho} - \rho = \frac{\rho^2}{1-\rho} = \rho\mathbb{E}(\mathbf{X}).$$

Ce résultat peut être interprété de la façon suivante. La probabilité qu'il y ait au moins un client dans la file est égale à ρ. Sachant cela, le reste de la file se comporte comme une file M/M/1 de charge ρ. Plus précisément, pour chaque $i \in \mathbb{N}$, on a

$$\mathbb{P}(\mathbf{X}-1 = i|\mathbf{X}\geq 1) = \mathbb{P}(\mathbf{X} = i+1|\mathbf{X}\geq 1) = \frac{\mathbb{P}(\mathbf{X}=i+1)}{\mathbb{P}(\mathbf{X}\geq 1)} = \frac{\pi(i+1)}{1-\pi(0)} = \frac{(1-\rho)\rho^{i+1}}{1-(1-\rho)} = (1-\rho)\rho^i = \pi(i).$$

Propriété PASTA (Poisson arrivals see time averages) On applique la formule des transitions conditionnelles. Pour chaque $i \in \mathbb{N}$, la probabilité qu'un client entrant trouve la file dans l'état i est donnée par

$$\frac{\lambda\pi(i)}{\sum_{j\in\mathbb{N}}\lambda\pi(j)} = \pi(i).$$

Cette propriété est vraie en général, dès que les clients entrent dans la file selon un processus de Poisson. On l'appelle la propriété PASTA, pour *Poisson arrivals see time averages*.

Loi de Little Le temps moyen δ qu'un client passe dans la file est la somme de son temps d'attente et de son temps de service. On a donc

$$\delta = \sum_{i\in\mathbb{N}} \underbrace{\pi(i)}_{\substack{\text{(PASTA) Probabilité}\\ \text{qu'un client trouve } i \text{ clients}\\ \text{devant lui à son arrivée}}} \times \underbrace{(i+1)\frac{1}{\mu}}_{\substack{\text{Temps moyen pour}\\ \text{servir les } i \text{ clients déjà}\\ \text{présents, plus celui qui arrive}}}.$$

On obtient

$$\delta = \frac{1}{\lambda}\sum_{i\in\mathbb{N}}(i+1)\rho\pi(i) = \frac{1}{\lambda}\sum_{i\in\mathbb{N}}(i+1)\pi(i+1) = \frac{1}{\lambda}\sum_{i\geq 1} i\pi(i) = \frac{L}{\lambda}.$$

On peut réécrire cette égalité sous la forme $L = \lambda\delta$. Il s'agit de la loi de Little, vraie dans n'importe quelle file d'attente à l'état stationnaire.

3 Systèmes à attente et à perte

On distingue les systèmes à perte des systèmes à attente. Les *systèmes à perte* sont ceux dont la file d'attente a un nombre limité de places, de sorte que des clients sont perdus (rejetés) s'ils arrivent alors que la file est déjà pleine. Les *systèmes à attente* sont ceux dont la file d'attente a un nombre infini de places.
En pratique, il n'existe pas de système à attente. On modélisera cependant par un système à attente tout système dont le dimensionnement (du nombre de serveurs et de la file d'attente) est tel que les pertes sont négligeables.

Système à attente Puisque les clients peuvent arriver et trouver une file arbitrairement longue, le critère principal de performance est le *temps moyen d'attente*.

Théorème (Formule de Little)**.** *Dans un système à attente qui se trouve à l'état stationnaire, le nombre moyen L de clients dans la file (en attente ou en service) est égal au produit du taux d'arrivée λ des clients dans la file par leur délai moyen δ. Autrement dit, on a $L = \lambda\delta$.*

Cela revient à dire que le temps moyen de séjour d'un client dans la file est égal au nombre moyen de clients dans la file, divisé par leur taux d'arrivée.

Exemple (File M/M/1)*.* La loi de Little redonne directement le résultat énoncé plus tôt :

$$\delta = \frac{\rho}{1-\rho}\frac{1}{\lambda} = \frac{\frac{\lambda}{\mu}}{1-\frac{\lambda}{\mu}}\frac{1}{\lambda} = \frac{1}{\mu - \lambda}.$$

Système à perte La taille de la file d'attente étant finie, on dispose automatiquement d'une borne sur le temps moyen de séjour d'un client. En contrepartie, les clients peuvent être rejetés à leur arrivée. Le critère principal de performance est donc la *probabilité de perte*, définie comme la probabilité qu'un client soit perdu lorsqu'il arrive dans le système. D'après la formule des transitions conditionnelles, on a informellement

$$\text{Probabilité de perte} = \frac{\sum_{i\in E, i \text{ état bloquant}} \pi(i) \times \text{taux d'arrivée des clients dans l'état } i}{\sum_{i\in E} \pi(i) \times \text{taux d'arrivée des clients dans l'état } i}.$$

Le dénominateur est appelé le *taux d'arrivée*. Il s'agit de la fréquence moyenne à laquelle les clients arrivent dans la file. Le numérateur est appelé le *taux de perte*. Il s'agit de la fréquence moyenne à laquelle les clients sont rejetés de la file. On a ainsi

$$\text{Probabilité de perte} = \frac{\text{Taux de perte}}{\text{Taux d'arrivée}}.$$

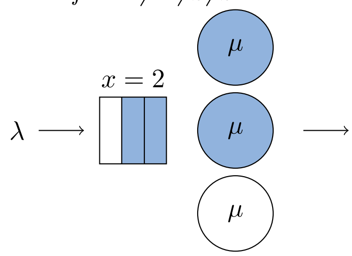
Lorsque les clients arrivent selon un processus de Poisson de paramètre λ, le taux d'arrivée des clients dans l'état i est égal à λ quel que soit l'état $i \in E$ considéré. En simplifiant l'égalité ci-dessus, on obtient alors que la probabilité de perte est égale à la *probabilité de blocage*, définie comme la probabilité que le système soit dans un état où les nouveaux clients sont rejetés. Plus généralement :

Théorème (Propriété PASTA (Poisson arrivals see time averages))**.** *Quand le processus d'arrivée est un processus de Poisson, un point du processus (c'est-à-dire un client entrant) voit la distribution stationnaire.*

Intuitivement, la propriété PASTA signifie que les instants d'arrivée des clients dans la file sont indépendants de l'état de la file. Autrement dit, la probabilité qu'un nouveau client trouve i clients devant lui à sont arrivée est égale à $\pi(i)$.

4 Modèle d'Erlang à pertes

Modèle Le modèle d'Erlang à perte a été introduit en 1917 par le mathématicien et ingénieur danois Agner Krarup Erlang alors qu'il travaillait pour la *Copenhagen Telephone Company*. Ce modèle correspond à une *file M/M/S/S* :



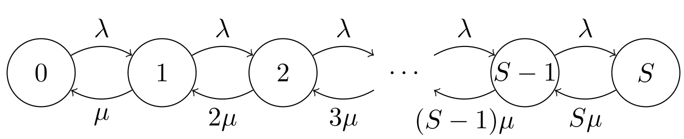
- ***M*** Les clients arrivent selon un processus de Poisson de paramètre $\lambda > 0$.
- ***M*** Le temps de service de chaque client suit une loi exponentielle de paramètre $\mu > 0$ (et est indépendant du temps de service des autres clients).
- ***S*** S serveurs.
- ***S*** Il y a autant de places dans la file que de serveurs. Un client qui arrive alors qu'il y a déjà S clients (en service) dans la file est perdu.

La charge de la file est à nouveau donnée par $\rho = \frac{\lambda}{\mu}$.
Ce modèle a été introduit pour modéliser les allocations de lignes téléphoniques dans les réseaux d'opérateurs. Dans ce contexte, chaque serveur correspond à une ligne téléphonique et chaque client représente un appel entrant. Un appel est rejeté si toutes les lignes sont occupées. L'objectif est de calculer la probabilité de perte, c'est-à-dire la probabilité qu'un appel entrant soit perdu.

État de la file Pour chaque $t \in \mathbb{R}_+$, on note $\mathbf{X}_t$ la variable aléatoire qui compte le nombre de clients dans la file à l'instant t. On suppose que la file est vide à l'instant 0, c'est-à-dire que $\mathbf{X}_0 = 0$. $(\mathbf{X}_t)_{t\geq 0}$ est un processus de Markov homogène de type naissance et mort. Son espace d'état $E = \{0, 1, 2, \ldots, S\}$ est fini. Les transitions qui régissent l'évolution du processus sont représentées sur le diagramme de la Figure 3a :

- Arrivées. La prochaine arrivée se produit après un temps de loi exponentielle de paramètre λ. S'il y a déjà m clients dans la file, le client entrant est perdu et l'état de la file reste inchangé.
- Départs. Lorsque i serveurs sont occupés, le temps avant le prochain départ est le minimum de i variables aléatoires indépendantes de loi exponentielle de paramètre μ, donc il suit une loi exponentielle de paramètre $i\mu$.

Le processus $(\mathbf{X}_t)_{t\geq 0}$ est irréductible (on peut relier n'importe quel couple d'états en suivant le chemin le plus direct dans le diagramme de transition) et récurrent (l'espace d'état est fini).



(a) Diagramme de transition

$$A = \begin{bmatrix} -\lambda & \lambda & & & & & \\ \mu & -(\lambda+\mu) & \lambda & & & & \\ & 2\mu & -(\lambda+2\mu) & \lambda & & & \\ & & 3\mu & -(\lambda+3\mu) & \lambda & & \\ & & & \ddots & \ddots & \ddots & \\ & & & & (S-1)\mu & -(\lambda+(S-1)\mu) & \lambda \\ & & & & & S\mu & -S\mu \end{bmatrix}$$

(b) Générateur infinitésimal

Figure 3 – Processus de Markov décrivant l'état d'une file M/M/S/S

Distribution stationnaire On cherche les mesures π, définies sur E, qui sont solutions de l'équation matricielle $\pi A = 0$, où A est le générateur infinitésimal du processus $(\mathbf{X}_t)_{t\geq 0}$, donné en FIGURE 3b. Cela revient à résoudre le système d'équations :

$$\begin{cases} -\lambda\pi(0) + \mu\pi(1) = 0, \\ \lambda\pi(i-1) - (\lambda + i\mu)\pi(i) + (i+1)\mu\pi(i+1) = 0, \quad \forall i = 1, \dots, S-1, \\ \lambda\pi(S-1) - S\mu\pi(S) = 0. \end{cases}$$

Comme pour la file M/M/1, on peut montrer par récurrence sur $i = 1, \dots, S-1$ que le système ci-dessus est équivalent à

$$-\lambda\pi(i-1) + i\mu\pi(i) = 0, \quad \forall i = 1, \dots, S,$$

de sorte que

$$\pi(i) = \pi(0)\frac{\rho^i}{i!}, \quad \forall i = 0, 1, \dots, S.$$

La condition de normalisation impose la valeur de $\pi(0)$:

$$1 = \sum_{i=0}^{S} \pi(0)\frac{\rho^i}{i!}, \quad \text{soit} \quad \pi(0) = \frac{1}{\sum_{i=0}^{S} \frac{\rho^i}{i!}}.$$

La distribution stationnaire du processus $(\mathbf{X}_t)_{t\geq 0}$ est donc donnée par

$$\pi(i) = \frac{\frac{\rho^i}{i!}}{\sum_{j=0}^{S} \frac{\rho^j}{j!}}, \quad \forall i = 1, \dots, S.$$

Performance et dimensionnement Comme il s'agit d'un système à perte, on s'intéresse à la probabilité de perte. Les clients arrivent selon un processus de Poisson. D'après la propriété PASTA, la probabilité de perte est donc égale à la probabilité de blocage, donnée par

$$\pi(S) = \frac{\frac{\rho^S}{S!}}{\sum_{j=0}^{S} \frac{\rho^j}{j!}}.$$

L'égalité ci-dessus s'appelle la *formule d'Erlang-B*, notée

$$E(\rho, S) = \frac{\frac{\rho^S}{S!}}{\sum_{j=0}^{S} \frac{\rho^j}{j!}}.$$

Cette formule a été beaucoup utilisée pour dimensionner les réseaux téléphoniques. En général, on connaît la charge ρ, on fixe une probabilité de blocage à ne pas dépasser, et on cherche le nombre minimum S de serveurs garantissant que $E(\rho, S)$ est inférieur ou égal à cette probabilité.
Pour une charge ρ fixée, on peut calculer $E(\rho, S)$ par récurrence sur le nombre S de serveurs. Son inverse est donné par

$$\frac{1}{E(\rho, S)} = \frac{\sum_{j=0}^{S} \frac{\rho^j}{j!}}{\frac{\rho^S}{S!}} = \frac{\frac{\rho^S}{S!} + \sum_{j=0}^{S-1} \frac{\rho^j}{j!}}{\frac{\rho^S}{S!}} = \frac{\frac{\rho^S}{S!}}{\frac{\rho^S}{S!}} + \frac{\sum_{j=0}^{S-1} \frac{\rho^j}{j!}}{\frac{\rho}{S} \times \frac{\rho^{S-1}}{(S-1)!}} = 1 + \frac{S}{\rho}\frac{1}{E(\rho, S-1)}.$$

En réinversant, on obtient

$$E(\rho, S) = \frac{E(\rho, S-1)}{\frac{S}{\rho} + E(\rho, S-1)}.$$

On peut utiliser cette formule pour démontrer que $E(\rho, S)$ décroît bien lorsque S augmente [1]. En pratique, on initialise la récurrence avec $E(\rho, 0) = 1$ (perte sûre) pour $S = 0$ puis on incrémente S jusqu'à ce que $E(\rho, S)$ passe sous le seuil souhaité.

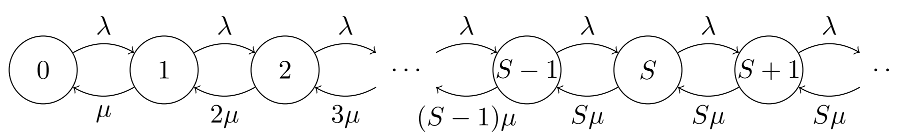
Insensibilité Nous avons supposé la distribution des temps de service exponentielle. On peut montrer que la formule d'Erlang-B reste valable sous des hypothèses plus générales. On dit que le modèle d'Erlang à perte est *insensible* à la distribution des temps de service des clients.

5 Modèle d'Erlang à attente

Modèle Le modèle d'Erlang à attente correspond à une file M/M/S[/∞/FIFO]. Il est similaire au modèle d'Erlang à perte, sauf que la taille de la file d'attente est infinie. La partie 1 donne un exemple avec $S = 3$.

État de la file Soit $(\mathbf{X}_t)_{t\geq 0}$ le processus qui compte le nombre de clients dans la file à chaque instant, avec $\mathbf{X}_0 = 0$. $(\mathbf{X}_t)_{t\geq 0}$ est un processus de Markov homogène de type naissance et mort. Son espace d'état $E = \mathbb{N}$ est infini. Les transitions qui régissent l'évolution de la file sont représentées sur le diagramme de la Figure 4a :

- Arrivées. La prochaine arrivée se produit après un temps de loi exponentielle de paramètre λ. Il n'y a pas de blocage.
- Départs. Lorsque i clients sont dans la file, le nombre de serveurs occupés est donné par $\min(i, k)$. Le temps avant le prochain départ suit donc une loi exponentielle de paramètre $\min(i, k) \times \mu$.



(a) Diagramme de transition de l'état de la file M/M/S/S

$$A = \begin{bmatrix} -\lambda & \lambda & & & & & & & \\ \mu & -(\lambda+\mu) & \lambda & & & & & & \\ & 2\mu & -(\lambda+2\mu) & \lambda & & & & & \\ & & 3\mu & -(\lambda+3\mu) & \lambda & & & & \\ & & & \ddots & \ddots & \ddots & & & \\ & & & & S\mu & -(\lambda+S\mu) & \lambda & & \\ & & & & & S\mu & -(\lambda+S\mu) & \lambda & \\ & & & & & & \ddots & \ddots & \ddots \end{bmatrix}$$

(b) Générateur infinitésimal

Figure 4 – Processus de Markov décrivant l'état d'une file M/M/S

1. On peut montrer par récurrence que $\frac{S}{\rho} + E(\rho, S) \geq 1$ pour n'importe quelle valeur de S. L'initialisation est immédiate puisque $E(\rho, 0) = 1$ (perte sûre). L'hérédité s'écrit

$$\frac{S}{\rho} + E(\rho, S) = \frac{S}{\rho} + \frac{E(\rho, S-1)}{\frac{S}{\rho} + E(\rho, S-1)} = \frac{S}{\rho} + 1 - \frac{\frac{S}{\rho}}{\frac{S}{\rho} + E(\rho, S-1)} \geq \frac{S}{\rho} + 1 - \frac{\frac{S}{\rho}}{\frac{S-1}{\rho} + E(\rho, S-1)} \geq \frac{S}{\rho} + 1 - \frac{S}{\rho} = 1,$$

où la deuxième inégalité découle de l'hypothèse de récurrence.

Le processus $(\mathbf{X}_t)_{t\geq 0}$ est irréductible. On va montrer qu'il est récurrent positif si et seulement si $\lambda < S\mu$, c'est-à-dire $\rho < S$.

Distribution stationnaire On cherche les mesures π, définies sur $E = \mathbb{N}$, qui sont solutions de l'équation matricielle $\pi A = 0$, où A est le générateur infinitésimal du processus $(\mathbf{X}_t)_{t\geq 0}$, donné en FIGURE 4b. Cela revient à résoudre le système d'équations :

$$\begin{cases} -\lambda\pi(0) + \mu\pi(1) = 0, & \\ \lambda\pi(i-1) - (\lambda + i\mu)\pi(i) + (i+1)\mu\pi(i+1) = 0, & \forall i = 1, \ldots, S-1, \\ \lambda\pi(i-1) - (\lambda + S\mu)\pi(i) + S\mu\pi(i+1) = 0, & \forall i \geq S. \end{cases}$$

Après résolution, on obtient

$$\pi(i) = \begin{cases} \pi(0)\dfrac{\rho^i}{i!} & \text{si } i = 0, 1, \ldots, S, \\ \pi(0)\dfrac{\rho^i}{S!S^{i-S}} & \text{si } i \geq S, \end{cases}$$

Lorsque $\rho < S$, le processus de Markov est récurrent positif et la probabilité que le système soit vide est donnée par

$$\pi(0) = \frac{1}{\sum_{j=0}^{S} \frac{\rho^j}{j!} + \frac{\rho^S}{S!}\frac{\frac{\rho}{S}}{1-\frac{\rho}{S}}},$$

car

$$\sum_{j=S+1}^{+\infty} \frac{\rho^j}{S!S^{j-S}} = \frac{\rho^S}{S!}\sum_{j=1}^{+\infty}\left(\frac{\rho}{S}\right)^j = \frac{\rho^S}{S!}\frac{\frac{\rho}{S}}{1-\frac{\rho}{S}}.$$

Lorsque $\rho \geq S$, le processus de Markov $(\mathbf{X}_t)_{t\geq 0}$ n'est pas récurrent positif : la file d'attente est instable. Dans la suite, on supposera que $\rho < S$.

Remarque. On peut retrouver la distribution stationnaire de l'état du système dans modèle d'Erlang à perte en tronquant la distribution stationnaire de l'état du système dans le modèle d'Erlang à attente au-delà de l'état S et en renormalisant.

Performance et dimensionnement Les clients arrivent selon un processus de Poisson. D'après la propriété PASTA, la probabilité d'attente est donc égale à la probabilité d'être dans un état où les serveurs sont pleins. On obtient

$$C(S,\rho) = \sum_{i=S}^{+\infty}\pi(i) = \sum_{i=S}^{+\infty}\pi(0)\frac{\rho^i}{S!S^{i-S}} = \pi(0)\frac{\rho^S}{S!}\sum_{j=0}^{+\infty}\left(\frac{\rho}{S}\right)^j = \pi(0)\frac{\rho^S}{S!}\frac{1}{1-\frac{\rho}{S}}.$$

En injectant l'expression de $\pi(0)$ trouvée plus haut, il vient

$$C(S,\rho) = \frac{\frac{\rho^S}{S!}\frac{1}{1-\frac{\rho}{S}}}{\left(\sum_{j=0}^{S}\frac{\rho^j}{j!}\right) + \frac{\rho^S}{S!}\frac{\frac{\rho}{S}}{1-\frac{\rho}{S}}}.$$

Il s'agit de la formule d'*Erlang-C*. On peut la réécrire autrement pour l'exprimer en fonction de la probabilité $E(\rho, S)$ de blocage dans le modèle d'Erlang à perte :

$$C(S,\rho) = \frac{\frac{\rho^S}{S!}}{\left(1-\frac{\rho}{S}\right)\left(\sum_{j=0}^{S}\frac{\rho^j}{j!}\right) + \frac{\rho^S}{S!}\times\frac{\rho}{S}} = \frac{\frac{\frac{\rho^S}{S!}}{\sum_{j=0}^{S}\frac{\rho^j}{j!}}}{1-\frac{\rho}{S} + \frac{\frac{\rho^S}{S!}}{\sum_{j=0}^{S}\frac{\rho^j}{j!}}\times\frac{\rho}{S}} = \frac{E(\rho,S)}{1-\frac{\rho}{S}+\frac{\rho}{S}E(\rho,S)}.$$

On obtient :

$$C(S,\rho) = \frac{E(\rho,S)}{1-\frac{\rho}{S}(1-E(\rho,S))}.$$

6 Modèle d'Engset

Le modèle d'Engset ressemble au modèle d'Erlang à pertes dans le sens où il y a autant de places que de serveurs dans la file. Cependant, le modèle d'Engset suppose que la population de clients est finie. Plus précisément, le modèle est décrit de la façon suivante :

— S serveurs et S places dans la file d'attente.

— $M > S$ clients dans la population.

— Chaque client est alternativement "inactif" et "actif".

  — Un client inactif tente de devenir actif après un temps exponentiel de paramètre λ ; il est rejeté si tous les serveurs sont déjà occupés, auquel cas il retente sa chance après un temps exponentiel de paramètre λ, et ainsi de suite jusqu'à ce qu'il soit accepté.

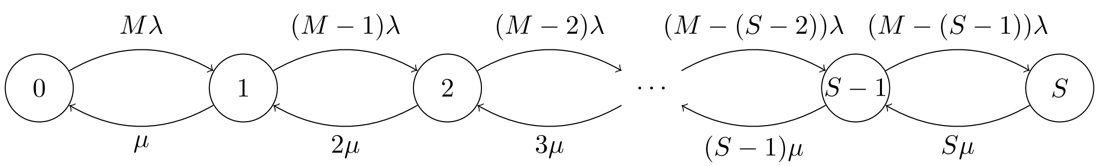
  — Une fois qu'un client est accepté, il reste actif (en service) pendant un temps exponentiel de paramètre μ avant de redevenir inactif.

État de la file Comme avant, on considère le processus $(\mathbf{X}_t)_{t\geq 0}$ qui compte le nombre de clients (actifs) dans la file à chaque instant. $(\mathbf{X}_t)_{t\geq 0}$ est un processus de Markov homogène. Son espace d'état $E = \{0, 1, 2, \ldots, S\}$ est fini. Les transitions qui régissent l'évolution de la file sont représentées sur le diagramme de la FIGURE 5a. Lorsque i clients sont en service (i.e., i serveurs sont occupés) :

— Arrivées. Il y a $M - i$ clients susceptibles d'arriver. Le temps avant la prochaine arrivée est le minimum de $M - i$ variables aléatoires indépendantes de loi exponentielle de paramètre λ, donc il suit une loi exponentielle de paramètre $(M - i)\lambda$.

— Départs. Comme dans le modèle d'Erlang à perte, le temps avant le prochain départ suit une loi exponentielle de paramètre $i\mu$.

Le processus $(\mathbf{X}_t)_{t\geq 0}$ est irréductible et récurrent.



(a) Diagramme de transition

$$A = \begin{bmatrix} -M\lambda & M\lambda & & & & \\ \mu & -((M-1)\lambda+\mu) & (M-1)\lambda & & & \\ & 2\mu & -((M-2)\lambda+2\mu) & (M-2)\lambda & & \\ & & \ddots & \ddots & \ddots & \\ & & & (S-1)\mu & -((M-S+1)\lambda+(S-1)\mu) & (M-S+1)\lambda \\ & & & & S\mu & -S\mu \end{bmatrix}$$

(b) Générateur infinitésimal

FIGURE 5 – Processus de Markov décrivant l'état de la file d'attente associée au modèle d'Engset

Distribution stationnaire En suivant le même raisonnement qu'avant, on obtient

$$\pi(i) = \pi(0)\frac{M(M-1)\ldots(M-i+1)\rho^i}{i!} = \pi(0)\frac{M!}{i!(M-i)!}\rho^i = \pi(0)\binom{M}{i}\rho^i, \quad \forall i = 0, 1, \ldots, M.$$

Par normalisation, on a

$$\pi(0) = \frac{1}{\sum_{j=0}^{S} \binom{M}{j} \rho^j},$$

d'où

$$\pi(i) = \frac{\binom{M}{i} \rho^i}{\sum_{j=0}^{S} \binom{M}{j} \rho^j}, \quad \forall i = 0, 1, \ldots, S.$$

Performance et dimensionnement Les arrivées ne forment pas un processus de Poisson, donc la propriété PASTA ne s'applique pas. La probabilité de perte est donnée par la formule d'Engset :

$$\begin{aligned}
\mathrm{Eng}(\rho, S, M) &= \frac{\text{taux de perte}}{\text{taux d'arrivée}}, \\
&= \frac{(M-S)\lambda\pi(S)}{\sum_{j=0}^{S}(M-j)\lambda\pi(j)} \left(= \frac{\text{taux d'arrivée dans l'unique état bloquant } S}{\text{somme de tous les taux d'arrivée}} \right), \\
&= \frac{(M-S)\lambda\binom{M}{S}\rho^S\pi(0)}{\sum_{j=0}^{S}(M-j)\lambda\binom{M}{j}\rho^j\pi(0)},
\end{aligned}$$

ce qui donne

$$\mathrm{Eng}(\rho, S, M) = \frac{\binom{M-1}{S} \rho^S}{\sum_{j=0}^{S} \binom{M-1}{j} \rho^j}.$$

On peut remarquer que $\mathrm{Eng}(\rho, S, M)$ est égal à la probabilité d'être dans l'état S dans un système avec une population de $M-1$ clients, c'est-à-dire à la probabilité de blocage dans ce système.
Comme pour la formule d'Erlang-B, on peut calculer $\mathrm{Eng}(\rho, S, M)$ par récurrence sur S. En effet, pour chaque $S = 1, \ldots, M-1$, on a

$$\frac{1}{\mathrm{Eng}(\rho, S, M)} = \frac{\binom{M-1}{S}\rho^S + \sum_{j=0}^{S-1}\binom{M-1}{j}\rho^j}{\binom{M-1}{S}\rho^S} = 1 + \frac{\sum_{j=0}^{S-1}\binom{M-1}{j}\rho^j}{\frac{M-S}{S}\rho \times \binom{M-1}{S-1}\rho^{S-1}} = 1 + \frac{S}{(M-S)\rho}\frac{1}{\mathrm{Eng}(\rho, S-1, M)}.$$

On a donc

$$\mathrm{Eng}(\rho, S, M) = \begin{cases} 1 & \text{si } S = 0, \\ \dfrac{\mathrm{Eng}(\rho, S-1, M)}{\frac{S}{(M-1)\rho} + \mathrm{Eng}(\rho, S-1, M)} & \text{si } S = 1, \ldots, M-1, \\ 0 & \text{si } S \geq M. \end{cases}$$

On peut constater que $\mathrm{Eng}(\rho, S, M) > E(\rho, S)$ puisque $\frac{S}{(M-1)\rho} \leq \frac{S}{\rho}$.

Références

[1] T. Bonald and M. Feuillet. *Performances des réseaux et des systèmes informatiques*. Collection Télécom. Hermes Science Publications, 2011. Existe aussi en anglais. Les chapitres 6 "Files d'attente" et 8 "Trafic circuit" couvrent le contenu de cette séance.

[2] L. Decreusefond and P. Moyal. *Modélisation et analyse stochastiques des réseaux de télécommunications*. Hermès, 2011. Existe aussi en anglais. Les chapitres 8 "Systèmes à attente" et 9 "Modèles à pertes" couvrent le contenu de cette séance.