Gestion des files d'attentes

A. Bazeniar

Enseignant chercheur
Centre universitaire Abdelhafid Boussouf
Mila, Algérie

Se reporter à des manuels de base et à certaines livres de spécialités

Septembre 2024

Chapitre 3

Modèles de files d'attente

3.1 Modèles M/M/c

Le système $M/M/c$ est un système d'attente à arrivées poissoniennes et à nombre de serveurs c, chacun ayant une durée de service exponentielle à paramètre μ. Le nombre de clients est illimité et la file d'attente est supposée unique et « ordonnée. » Même s'il y a plusieurs files il suffit que les clients n'aient aucune préférence pour l'un ou l'autre des serveurs et que la circulation soit possible entre les files pour que le cas soit facilement assimilable à un système avec une seule file d'attente. Pour ce système les hypothèses de base sont les suivantes :

— Les arrivées suivent un processus de Poisson avec un taux λ.

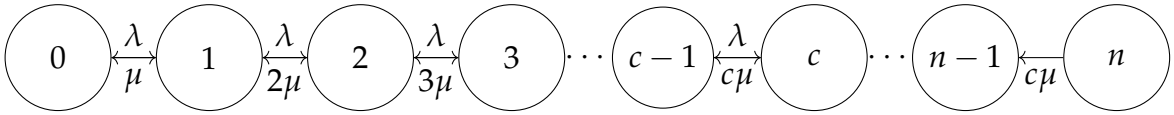
— Chaque serveur offre un service de durée moyenne $1/\mu$, ce qui signifie que le temps de service est distribué exponentiellement avec un taux μ.

— La charge du système est $\rho = \frac{\lambda}{c \cdot \mu}$.

Un régime stationnaire existe seulement si la charge est inférieure au nombre de serveurs, c'est-à-dire $\rho < 1$ (ce qui équivaut à $\frac{\lambda}{\mu} < c$).

3.1.1 Probabilités stationnaires de nombre de clients dans le système

En régime stationnaire, les taux d'entrée et de sortie de chaque état doivent être égaux pour garantir une distribution d'équilibre. Rappelant que $P(Q_\infty = n)$ est la probabilité d'avoir n clients dans le système. On est en présence de deux cas :



Cas 1 : $0 < n < c$

Pour $n < c$, il y a suffisamment de serveurs pour que tous les clients soient servis sans attente. Le flux entrant dans l'état n (clients arrivant au taux λ depuis l'état $n-1$) doit être égal au flux sortant (clients se faisant servir au taux $n\mu$, car il y a n serveurs disponibles).

L'équation d'équilibre devient donc pour $0 \leq n < c$:

$$\lambda P(Q_\infty = n-1) = n\mu P(Q_\infty = n).$$

Cela donne,

$$P(Q_\infty = n) = \frac{\lambda}{n\mu} P(Q_\infty = n-1).$$

En itérant cette relation à partir de π_0, on obtient,

$$\pi_n = \pi_0 \frac{(c\rho)^n}{n!}, \quad \text{pour } 0 \leq n < c.$$

Cas 2 : $n \geq c$

Pour $n \geq c$, tous les serveurs sont occupés, donc le taux de service total est constant et égal à $c\mu$. Dans ce cas, l'équation d'équilibre devient :

$$\lambda P(Q_\infty = n-1) = c\mu P(Q_\infty = n).$$

On résout cette équation en fonction de $P(Q_\infty = n-1)$, ce qui donne :

$$\begin{aligned} P(Q_\infty = n) &= \frac{\lambda}{c\mu} P(Q_\infty = n-1) \\ &= (\frac{\lambda}{c\mu})^{n-c} P(Q_\infty = c) \\ &= \rho^{n-c} \frac{(c\rho)^c}{c!} \cdot P(Q_\infty = 0) \end{aligned}$$

Donc, on obtient :

$$\pi_n = \pi_0 \frac{c^c}{c!} \rho^n, \quad \text{pour } n \geq c. \tag{3.1}$$

3.1.1.1 Calcul de π_0

La constante $\pi_0 = P(Q_\infty = 0)$ est déterminée en normalisant la distribution, c'est-à-dire en assurant que la somme des probabilités pour tous les états est égale à 1 :

$$\sum_{n=0}^{\infty} P(Q_\infty = n) = 1.$$

Cela donne :

En simplifiant la seconde somme comme une série géométrique, on trouve que :

$$\begin{aligned} \pi_0 &= \left(\sum_{n=0}^{c-1} \frac{(c\rho)^n}{n!} + \sum_{n=c}^{\infty} \frac{\rho^n c^c}{c!} \right)^{-1} \\ &= \left(\frac{(c\rho)^c}{c!} \frac{1}{1-\rho} + \sum_{n=0}^{c-1} \frac{(\rho)^n}{n!} \right)^{-1} \end{aligned}$$

Donc, La distribution d'équilibre de π_n dans un système de file d'attente $M/M/c$ est donnée par :

$$\pi_n = P(Q_\infty = n) = \begin{cases} \pi_0 \frac{(c\rho)^n}{n!}, & \text{si } 0 < n < c, \\ \pi_0 \frac{c^c}{c!}\rho^n & \text{si } n \geq c. \end{cases}$$

3.1.2 Nombre moyen de clients en attente

Le nombre moyen de clients en attente $L_q = E(Q_\infty)$ est la différence entre le nombre dans le système est celui en services :

$$L_q = E(Q_\infty) = \sum_{n=c}^{\infty}(n-c)\pi_n.$$

Remplaçant π_n, on aura

$$L_q = \sum_{n=c}^{\infty}(n-c)\pi_n = \sum_{n=c}^{\infty}(n-c)\frac{c^c}{c!}\rho^n\pi_0.$$

Et comme $\rho = \frac{\lambda}{\mu c}$, où $\rho < 1$, et avec un changement de variable, on aura

$$L_q = \pi_0\frac{c^c}{c!}\rho^c\sum_{n=1}^{\infty} n\rho^n.$$

En utilisant le résultat connu :

$$\sum_{n=1}^{\infty} nx^n = \frac{x}{(1-x)^2}, \quad \text{pour } |x| < 1,$$

on obtient finalement :

$$L_q = \frac{(c\rho)^c}{c!}\cdot\frac{\rho}{(1-\rho)^2}\pi_0. \tag{3.2}$$

3.1.3 Nombre moyen de serveurs occupés – Nombre moyen de serveurs inoccupés

Le nombre moyen de serveurs occupés dépend du nombre de clients en attente :
— Si les serveurs ne sont pas tous occupés : $\sum_{n=0}^{c-1} n\pi_n$.
— Si les serveurs sont tous occupés : $\sum_{n=c}^{\infty} c\pi_n$.
Après certains calculs, le nombre moyen de serveurs occupés est,

$$L_s = E(S_\infty) = \pi_0 \cdot \pi_0^{-1} \cdot c\rho = \frac{\lambda}{\mu}, \tag{3.3}$$

et le nombre moyen de serveurs inoccupés est $\overline{L_s} = (1-\rho)c$.

Le nombre moyen de clients dans le système (attente et service) est donné par :

$$L = L_q + \frac{\lambda}{\mu} = \frac{(c\rho)^c}{c!} \cdot \frac{\rho}{(1-\rho)^2}\pi_0 + c\rho. \tag{3.4}$$

3.1.4 Probabilité de temps d'attente

Un client se voit contraint de rejoindre la queue s'il trouve tous les guichets occupés, soit $n \geq c$. Alors,

$$P(W_q > 0) = P(n \geq c) = \sum_{n=s}^{\infty} p_n = \sum_{n=c}^{\infty} \frac{\rho^n c^c}{c!}\pi_0,$$

où $\rho = \frac{\lambda}{\mu c}$.

En utilisant le résultat connu :

$$\sum_{n=c}^{\infty} x^n = \frac{x^c}{1-x}.$$

Ce qui donne, en définitive :

$$P(W_q > 0) = \frac{c^c}{c!}\frac{\rho^c}{1-\rho)}\pi_0.$$

D'autre part, soit t l'instant de d'arrivée du client c et s le temps l'instant de départ du client c, alors $[t, s]$ n'est autre que le temps d'attente du client, qui suit une lois exponentielle avec une moyenne égale à $\frac{1}{c\mu-\lambda}$, on écrit

$$p(W_q > t/n \geq c) = e^{-(c\mu-\lambda)t}.$$

Et d'une manière générale

$$p(W_q > t) = p(W_q > 0) \cdot e^{-(c\mu-\lambda)t}.$$

La probabilité que le temps d'attente soit inférieur à un certain seuil t est donnée par :

$$P(W_q \leq t) = 1 - p(W_q > t).$$

Finalement, on obtient

$$P(W_q \leq t) = 1 - \pi_0 \frac{(c\rho)^c}{c!(1-\rho)} e^{-(c\mu-\lambda)}. \tag{3.5}$$

Temps moyen d'attente des clients : Le temps moyen d'attente pour un client dans le système est :

$$W_q = \frac{L_q}{\lambda} = \mathbb{E}(W_\infty) = \frac{1}{c\mu}\frac{(c\rho)^c}{c!\,(1-\rho)^2}\pi_0. \tag{3.6}$$

Cela représente le temps d'attente moyen pour les clients arrivant lorsque tous les serveurs sont occupés. **Temps moyen de séjour des clients :** Le temps moyen de séjour des clients est :

$$W = \frac{L}{\lambda} = \frac{1}{c\mu} \frac{(c\rho)^c}{c!\,(1-\rho)^2} \pi_0 + \frac{1}{\mu}. \tag{3.7}$$

Exercice : Prouver les formules de temps moyen décrites au dessus.

3.2 Modèles M/M/1/k

Le modèle **M/M/1/K** est une extension du modèle M/M/1 où la capacité du système est limitée à K clients. Lorsqu'un client arrive et trouve le système plein (i.e., K clients présents), il est rejeté. Ce modèle est utilisé pour modéliser des systèmes avec capacité limitée tels que les files d'attente avec espace contraint.

3.2.1 Probabilités d'État

1. **Équation de récurrence** : La probabilité d'état P_n est donnée par :

$$P_n = \rho^n \pi_0, \quad n = 0, 1, \ldots, K,$$

où $\rho = \frac{\lambda}{\mu}$ est le facteur de charge.

2. **Probabilité** π_0 (système vide) : Pour normaliser, on obtient :

$$\pi_0 = \frac{1-\rho}{1-\rho^{K+1}}, \quad \text{si } \rho \neq 1.$$

Si $\rho = 1$, alors $\pi_0 = \frac{1}{K+1}$.

3. **Probabilité** π_n : En remplaçant π_0, on a :

$$\pi_n = \frac{\rho^n(1-\rho)}{1-\rho^{K+1}}, \quad n = 0, 1, \ldots, K.$$

3.2.2 Performances du Système

1. **Taux d'utilisation du serveur (U)** : Le serveur est occupé si le système n'est pas vide :

$$U = 1 - \pi_0.$$

2. **Nombre moyen de clients (L)** :

$$L = \sum_{n=0}^{K} n\pi_n.$$

En développant :

$$L = \frac{\rho}{1-\rho} - \frac{(K+1)\rho^{K+1}}{1-\rho^{K+1}}.$$

3. **Nombre moyen de clients en file (L_q)** :

$$L_q = L - (1 - \pi_0).$$

4. **Temps moyen dans le système (W)** : Par la loi de Little ($L = \lambda_{\text{eff}} W$), où :

$$\lambda_{\text{eff}} = \lambda(1 - \pi_K),$$

on a :

$$W = \frac{L}{\lambda_{\text{eff}}}.$$

5. **Temps moyen en file (W_q)** :

$$W_q = W - \frac{1}{\mu}.$$

3.3 Système M/G/1

Le modèle M/G/1 est un modèle de file d'attente à un seul serveur largement utilisé en recherche opérationnelle et en théorie des files d'attente. Voici une explication des notations et des composants clés :

3.3.1 Composants du modèle M/G/1

M (Arrivées markoviennes) : Les temps entre les arrivées suivent une distribution markovienne (exponentielle). Les arrivées sont modélisées comme un processus de Poisson avec un taux λ (taux moyen d'arrivée).

G (Temps de service de distribution générale) : Les temps de service peuvent suivre n'importe quelle distribution de probabilité générale avec une moyenne $E[S] = \mu^{-1}$ (temps de service moyen) et une variance $\text{Var}(S) = \sigma^2$.

1 (Serveur unique) : Le système comporte un seul serveur qui traite les tâches ou clients un par un.

3.3.2 Taux d'occuation

Dans un intervalle de temps T, le nombre moyen de clients arrivant est $\lambda \cdot T$ (car les arrivées suivent un processus de Poisson).

Chaque client nécessite en moyenne un temps de service $E[S]$. Ainsi, le temps total de service requis par tous les clients dans l'intervalle T est :

$$\text{Temps total requis} = (\lambda \cdot T) \cdot E[S].$$

Le taux d'occupation est défini comme la fraction de temps où le serveur est occupé. Cela correspond au ratio du temps total de service requis au temps total disponible T :

$$\rho = \frac{\text{Temps total requis}}{T}$$

En substituant la valeur du temps total requis :

$$\rho = \frac{(\lambda \cdot T) \cdot E[S]}{T}.$$

Après simplification, on obtient :

$$\rho = \lambda \cdot E[S].$$

3.3.3 Probabilité d'avoir n clients dans le système (π_n)

La probabilité π_n est donnée par :

$$\begin{aligned} \pi_n &= \pi_0 \cdot \rho^n, \\ &= (1-\rho)\rho^n \quad \text{pour } n \geq 0. \end{aligned}$$

Cette formule résulte du fait que le modèle $M/G/1$, comme le modèle $M/M/1$, obéit à une distribution géométrique en régime stationnaire.

3.3.4 Indicateurs de performance du modèle $M/G/1$

Le théorème de Pollaczek-Khinchine est un concept clé dans l'étude des files d'attente. Il s'applique au modèle $M/G/1$ et permet de calculer précisément les indicateurs de performance d'un système avec un seul serveur, comme le temps d'attente ou la longueur de la file.

a) Temps moyen d'attente dans la file ($\mathbb{E}[W_\infty]$)

Le temps moyen d'attente avant service dans la file est donné par :

$$\mathbb{E}[W_\infty] = \frac{\lambda \cdot \text{Var}(S)}{2(1-\rho)}.$$

b) Temps moyen de séjour dans le système $(\mathbb{E}[T_\infty])$

Le temps total passé dans le système (file d'attente + service) est :

$$\mathbb{E}[T_\infty] = \mathbb{E}[W_\infty] + \mathbb{E}[S].$$

En remplaçant $\mathbb{E}[W_\infty]$, on obtient :

$$\mathbb{E}[T_\infty] = \frac{\lambda \cdot \mathrm{Var}(S)}{2(1-\rho)} + \mathbb{E}[S].$$

c) Longueur moyenne de la file ($\mathbb{E}[Q_\infty]$)

En utilisant la loi de Little ($\mathbb{E}[Q_\infty] = \lambda \mathbb{E}[W_\infty]$), on obtient :

$$\mathbb{E}[Q_\infty] = \frac{\lambda^2 \cdot \mathrm{Var}(S)}{2(1-\rho)}.$$

d) Longueur moyenne totale dans le système ($\mathbb{E}[L_\infty]$)

En utilisant à nouveau la loi de Little ($\mathbb{E}[L_\infty] = \lambda \mathbb{E}[T_\infty]$), on a :

$$\mathbb{E}[L_\infty] = \rho + \frac{\lambda^2 \cdot \mathrm{Var}(S)}{2(1-\rho)}.$$