TD 9 : Analyse de files d'attente

lionel.rieg@ens-lyon.fr

Exercice 1 (Distribution du temps d'attente)

On considère une file d'attente M/M/1 de taux d'utilisation $\rho < 1$. On rappelle que sous l'hypothèse de régime permanent, on a les résultats suivants (vu au TD précédent) :

- la distribution du nombre de personnes dans le système suit une loi géométrique de paramètre ρ (décalée de 1 car elle peut prendre la valeur 0) ;
- l'espérance du nombre de personnes dans le système est $L = \frac{\rho}{1-\rho}$;
- l'espérance du nombre de personnes en attente est $L_q = L - (1 - p_0) = \frac{\rho^2}{1-\rho}$;
- pour obtenir la distribution du temps d'attente, il nous faut choisir une politique de service, ici PAPS ;
- dans le cas d'arrivée Poissonienne (c'est faux sinon), la probabilité q_n qu'un client arrivant dans le système trouve n personnes devant lui est égale à la probabilité p_n qu'il y ait n personnes dans le système dans l'absolu ;
- la variable aléatoire T_q représentant le temps d'attente est à densité pour $t > 0$ et possède un point de masse en $t = 0$;
- sa fonction de répartition $W_q(t) = F_{T_q}(t)$ vérifie $W_q(0) = q_0 = 1 - \rho$.

Le temps de service de n clients suit une distribution d'Erlang de paramètres n (paramètre de forme) et μ (paramètre d'intensité) dont la fonction de densité est

$$E(n, \mu, x) = \frac{\mu^n x^{n-1} \mathrm{e}^{-\mu x}}{(n-1)!}.$$

4. En utilisant la formule des probabilités totales, donner une expression pour $W_q(t)$.
5. Donner l'espérance W_q de T_q.
6. Adapter les deux questions précédentes pour traiter le cas de T, le temps de séjour total dans le système.
7. Quelle relation y a-t-il entre L_q et W_q d'une part, et L et W d'autre part ?

Exercice 2 (Serveur et miroir)

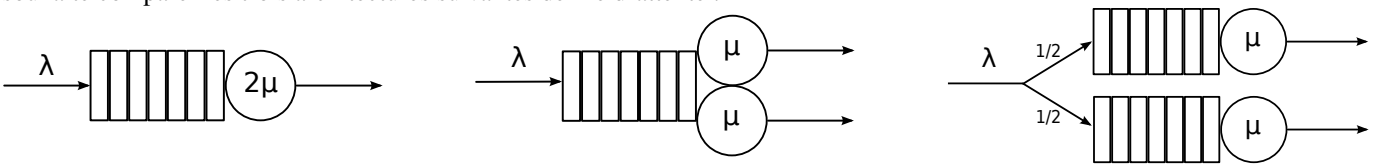
On dispose de deux serveurs Web : un serveur principal et un serveur miroir. Les requêtes arrivent à un routeur qui doit essayer de répartir le travail entre le serveur principal ou son miroir, qui est supposé moins performant. Le routeur ne connaissant pas la charge respectives des deux serveurs, il décide d'envoyer chaque requête au serveur principal avec probabilité p ou au serveur miroir avec probabilité $1 - p$. On cherche à optimiser le choix de p.

On suppose que les arrivées suivent une loi de Poisson de paramètre λ. Les temps de service des serveurs sont exponentiels de paramètres respectifs μ pour le serveur principal et ν pour le miroir.

1. Montrer que les arrivées des paquets dans les files du serveur principal et de son miroir sont des processus de Poisson de paramètres respectifs $p\lambda$ et $(1 - p)\lambda$.
2. Quelles sont les conditions de stabilité du système ?
3. Calculer le nombre de requêtes dans le serveur principal et dans le miroir à l'état stationnaire.
4. Donnez le temps de réponse moyen dans le système global.
5. Déterminez la valeur optimale de p.

Exercice 3 (Comparaison de trois modèles de serveur)

On souhaite comparer les trois architectures suivantes de file d'attente :



1. Intuitivement, quelle configuration semble la meilleure ?

Pour chacune des trois configurations :

2. Donnez sa condition de stabilité.
3. Écrire son générateur infinitésimal (en précisant les cas limites).
4. Calculer le nombre moyen de clients dans le système.
5. En déduire le temps moyen de réponse.

Où se trouve la différence entre les différentes configurations ? Laquelle préférer ?

Exercice 4 (Dimensionnement de buffer)

On considère un réseau d'entreprise composé de deux LANs, l'un constituant le réseau de production (LAN 1) et l'autre hébergeant la base de données de l'entreprise (LAN 2). Ces deux sous-réseaux sont connectés par un lien à $1,5$ Mbits/seconde. On suppose que :

- Les requêtes provenant du réseau de production pour la base de données arrivent au routeur d'interconnexion selon un processus de Poisson d'intensité égale à 15 requêtes par seconde.
- Les documents hébergés par la base de données ont une taille aléatoire de distribution exponentielle, de moyenne 100 kbits.
- Les délais de propagation et de traitement sont négligés, de sorte que les temps de transferts de ces documents entre les deux sites sont proportionnels à la taille des documents demandés.
- Les temps de traitement par la base de données sont négligés.
- La taille des requêtes est négligeable devant la taille des documents demandés.

1. En supposant les buffers de taille infinie, montrer que le processus d'arrivée des documents au routeur d'interconnexion du LAN 2 vers le LAN 1 peut être modélisé par une file M/M/1. Calculer ses taux d'arrivée et de service.
2. Le système est-il stable ? Décrivez l'évolution asymptotique du temps de réponse.
3. On suppose que l'on augmente le débit du lien d'interconnexion a 3 Mbits/seconde. Calculer la taille du buffer du routeur (mise en attente des requêtes) afin de garantir une probabilité de perte inférieure à 10^{-3}.

Solutions

► **Exercice 1**

4.

$$\begin{aligned}
W_q(t) &= \mathbb{P}\Big(T_q \leqslant t\Big) \\
&= W_q(0) + \sum_{n=1}^{+\infty} p_n \cdot \mathbb{P}\Big(\text{les client devant servis en moins de } t \,\Big|\, n \text{ client présents à l'arrivée dans la file}\Big) \\
&= 1-\rho + \sum_{n=1}^{+\infty} (1-\rho)\rho^n \cdot \int_0^t \frac{\mu^n x^{n-1} \mathrm{e}^{-\mu x}}{(n-1)!}\,\mathrm{d}x \\
&= 1-\rho + (1-\rho)\lambda \sum_{n=1}^{+\infty} \int_0^t \mathrm{e}^{-\mu x} \frac{(\lambda x)^{n-1}}{(n-1)!}\,\mathrm{d}x \\
&= 1-\rho + (1-\rho)\lambda \int_0^t \mathrm{e}^{-\mu x} \sum_{n=1}^{+\infty} \frac{(\lambda x)^{n-1}}{(n-1)!}\,\mathrm{d}x \\
&= 1-\rho + (1-\rho)\lambda \int_0^t \mathrm{e}^{-\mu x} \mathrm{e}^{\lambda x}\,\mathrm{d}x \\
&= 1-\rho + \frac{(1-\rho)\lambda}{\lambda-\mu}\Big[\mathrm{e}^{(\lambda-\mu)x}\Big]_0^t \\
&= 1-\rho + \frac{(1-\rho)\lambda}{-\mu(1-\rho)}\Big(\mathrm{e}^{(\lambda-\mu)t}-1\Big) \\
&= 1-\rho - \rho\Big(\mathrm{e}^{(\lambda-\mu)t}-1\Big) \\
&= 1-\rho\mathrm{e}^{(\lambda-\mu)t}
\end{aligned}$$

5. On a $f_{T_q}(t) = W_q'(t) = -\rho(\lambda-\mu)\mathrm{e}^{(\lambda-\mu)t}$, d'où

$$W_q = \mathbb{E}\Big(T_q\Big) = \int_0^{+\infty} t\cdot(-\rho)(\lambda-\mu)\mathrm{e}^{(\lambda-\mu)t}\,\mathrm{d}t = \Big[-\rho t\mathrm{e}^{(\lambda-\mu)t}\Big]_0^{+\infty} + \int_0^{+\infty} \rho\mathrm{e}^{(\lambda-\mu)t}\,\mathrm{d}t = 0 + \rho\left[\frac{\mathrm{e}^{(\lambda-\mu)t}}{\lambda-\mu}\right]_0^{+\infty} = \frac{\rho}{\mu-\lambda}.$$

6. Il faudrait compter $n+1$ temps de service lorsqu'il y a n personnes à l'arrivée d'un client. Pour cela, on supprime le premier terme (T est une variable à densité), comme $p_{n-1} = \frac{1}{\rho}p_n$, cela revient à diviser par ρ le deuxième terme de la somme :

$$W(t) = \mathbb{P}(T \leqslant t) = \frac{1}{\rho}\cdot(-\rho)\Big(\mathrm{e}^{(\lambda-\mu)t}-1\Big) = 1-\mathrm{e}^{(\lambda-\mu)t}$$

on reconnaît là la fonction de répartition d'une loi exponentielle de paramètre $\mu-\lambda$, d'où $W = \frac{1}{\mu-\lambda}$.

7. On découvre la loi de Little : $L = \lambda W$ et $L_q = \lambda W_q$ qui est en fait beaucoup plus générale.

► **Exercice 2**

1. On va calculer les fonctions de répartition des temps inter-arrivées. Entre deux paquets successifs arrivés dans la file secondaire, il peut y avoir un nombre arbitraire de paquets arrivés dans la file principale. En utilisant la formule des probabilités totales et le fait que l'arrivée de n paquets suit une loi de Erlang de paramètres n et λ, on a, en notant A l'événement « le n^{e} paquet est le premier qui va dans la bonne file » :

$$\begin{aligned}
\mathbb{P}(T_{n+1} - T_n \leqslant t) &= \sum_{n=1}^{+\infty} \mathbb{P}(A) \cdot \mathbb{P}\Big(n \text{ paquets arrivés en moins de } t \,\Big|\, A\Big) \\
&= \sum_{n=1}^{+\infty} \mathbb{P}(Geom(p) = n) \cdot \int_0^t E(n, \lambda, x)\,\mathrm{d}x \\
&= \sum_{n=1}^{+\infty} (1-p)^{n-1} p \int_0^t \frac{x^{n-1}\lambda^n \mathrm{e}^{-\lambda x}}{(n-1)!}\,\mathrm{d}x \\
&= \int_0^t p\mathrm{e}^{-\lambda x} \sum_{n=1}^{+\infty} (1-p)^{n-1} \frac{x^{n-1}\lambda^n}{(n-1)!}\,\mathrm{d}x \\
&= \int_0^t p\lambda \mathrm{e}^{-\lambda x} \sum_{n=1}^{+\infty} \frac{((1-p)\lambda x)^{n-1}}{(n-1)!}\,\mathrm{d}x \\
&= \int_0^t p\lambda \mathrm{e}^{-\lambda x} \mathrm{e}^{(1-p)\lambda x}\,\mathrm{d}x \\
&= \int_0^t p\lambda \mathrm{e}^{-p\lambda x}\,\mathrm{d}x \\
&= 1 - \mathrm{e}^{-p\lambda t}
\end{aligned}$$

On reconnaît la fonction de répartition d'une loi exponentielle de paramètre $p\lambda$, donc les temps d'arrivées dans la file principale suivent un processus de Poisson de paramètre $p\lambda$. Par symétrie, les temps d'arrivées dans la file secondaire suivent un processus de Poisson de paramètre $(1-p)\lambda$.

On peut aussi procéder différemment en comptant le nombre de paquets arrivés dans l'intervalle $[0, t[$ et en disant qu'au moins l'un d'entre eux doit aller dans la bonne file. Notons cette fois-ci B l'événement « n paquets arrivent dans l'intervalle $[0, t[$ ».

$$\begin{aligned}
\mathbb{P}(T_{n+1} - T_n \leqslant t) &= \sum_{n\geqslant 1} \mathbb{P}(B) \cdot \mathbb{P}\Big(\text{ au moins un paquet dans la bonne file} \,\Big|\, B\Big) \\
&= \sum_{n\geqslant 1} \mathbb{P}(\mathcal{P}(\lambda t) = n) \cdot \mathbb{P}(\mathcal{B}(n,p) \geqslant 1) \\
&= \sum_{n\geqslant 1} \mathrm{e}^{-\lambda t} \frac{(\lambda t)^n}{n!} \cdot \Big(1 - (1-p)^n\Big) \\
&= \sum_{n\geqslant 1} \mathrm{e}^{-\lambda t} \frac{(\lambda t)^n}{n!} - \sum_{n\geqslant 1} \mathrm{e}^{-\lambda t} \frac{((1-p)\lambda t)^n}{n!} \\
&= 1 - \mathrm{e}^{-\lambda t} - \mathrm{e}^{-\lambda t}\Big(\mathrm{e}^{(1-p)\lambda t} - 1\Big) \\
&= 1 - \mathrm{e}^{-p\lambda t}
\end{aligned}$$

2. Les deux conditions sont alors $p\lambda < \mu$ et $(1-p)\lambda < \nu$.
3. D'après l'étude du premier exercice, le nombre moyen de requêtes pour le serveur principal est $\frac{p\lambda}{\mu - p\lambda}$ et $\frac{(1-p)\lambda}{\nu - (1-p)\lambda}$ pour le miroir.
4. Le temps de réponse moyen dans le serveur principal est $\frac{1}{\mu - p\lambda}$ et dans le serveur secondaire $\frac{1}{\nu - (1-p)\lambda}$. Au total, on décompose par la formule des probabilités totales, d'où $\overline{W} = p\frac{1}{\mu - p\lambda} + (1-p)\frac{1}{\nu - (1-p)\lambda}$.
5. Elle vaut :

$$p_{opt} = \frac{\mu\lambda - 2\mu\nu + \sqrt{-2\mu^2\lambda\nu + 2\mu^2\nu^2 - 2\mu\nu^2\lambda + \mu\nu^3 + \nu\mu\lambda^2 + \mu^3\nu}}{(\mu - \nu)\lambda}$$

► **Exercice 3**

1. C'est subjectif mais … la deuxième ?
2. La condition de stabilité est toujours $\lambda < 2\mu$.
3. – d'après de TD de la semaines dernière : $q_{i+1,i} = \lambda$, $q_{i-1,i} = 2\mu$, $q_{ii} = -(\lambda + 2\mu)$

  - l'analyse donne $q_{i+1,i} = \lambda$, $q_{i-1,i} = 2\mu$, $q_{ii} = -(\lambda + 2\mu)$
  - chaque file a un générateur donné par $q_{i+1,i} = \frac{\lambda}{2}$, $q_{i-1,i} = \mu$, $q_{ii} = -(\frac{\lambda}{2} + \mu)$ donc par linéarité de la dérivation, le générateur final est $q_{i+1,i} = \lambda$, $q_{i-1,i} = 2\mu$, $q_{ii} = -(\lambda + 2\mu)$

4. – $\frac{\lambda}{2\mu-\lambda}$
   – $\frac{\lambda}{2\mu-\lambda}$
   – $2\frac{\frac{\lambda}{2}}{\mu-\frac{\lambda}{2}} = 2\frac{\lambda}{2\mu-\lambda}$
5. – d'après l'exercice 1 : $\frac{1}{2\mu-\lambda}$
   – presque pareil (même temps d'attente mais la loi de traitement par les serveurs est plus faible)
   – le temps moyen est celui de chaque file : $\frac{1}{\mu-\frac{\lambda}{2}} = \frac{2}{2\mu-\lambda}$

La différence entre les deux premières configurations se fait lorsqu'il n'y a qu'un client dans le système : il est servi à vitesse double dans la première. La meilleure configuration semble être la première mais il y a d'autres facteurs à prendre en compte dans la vie réelle : le coût, la résistance aux pannes, …

► **Exercice 4**

1. Comme on néglige la taille et la gestion de la requête, elle peut être directement vue comme l'envoi du fichier demandé, donc les arrivées suivent un processus de Poisson de paramètre 15. De même, les départs correspondent aux transmissions dont la durée (*i.e.* le temps inter-départ) est exponentielle de paramètre 15, il s'agit donc d'un processus de Poisson.
2. On a $\lambda = \mu$ donc le système est instable.
3. Notons K la taille du buffer. On peut croire qu'il s'agit d'une file M/M/1/K dont on connaît la probabilité de dépassement : $\mathbb{P}(\text{dépassement}) = \mathbb{P}(N \geqslant K) = \frac{\rho^K}{1-\rho}$. Il suffit alors de résoudre et on trouve $K = \frac{\ln(10^3(1-\rho))}{\ln\rho}$.
   Ceci n'est valable que si les paquets ont tous une taille fixée. Ici, la taille a une distribution exponentielle donc on s'intéresse à la probabilité que la somme de variables indépendantes de loi exponentielle dépasse K… ce qui fait penser à la loi d'Erlang ! En fait, on peut tout simplement dire que comme le débit est constant, l'espace $E(t)$ occupé du buffer est proportionnel au temps d'attente : $W(t) = E(t)/3$. On en tire $\mathbb{P}(E(t) \geqslant K) = \mathbb{P}(W(t) \geqslant 3K) = \mathrm{e}^{(\lambda-\mu)3K}$ et ainsi, $K = \frac{\ln(10^{-3})}{3(\lambda-\mu)} = \frac{\ln 10}{\mu(1-\rho)} \approx 307$ kbits.
   Cela semble peu mais mais on peut vérifier ce résultat par une simulation avec Network Simulator 2 : on découpe les fichiers en paquets de 1000 bits et on prend une file de capacité 300 paquets. Pour une simulation de 100000 secondes avec 19488074 paquets émis (donc en moyenne 194881 fichiers), seuls 4 ont été jetés de la file et ils faisaient partie du même ficher. Voir le fichier `mm1k.tcl` pour le code de la simulation.