Aide mémoire sur les réseaux de files d'attente markoviennes

1 Introduction

Nous venons de voir dans la section précédente les principaux résultats concernant les files d'attente simples. Le problème maintenant est de savoir si ces résultats sont robustes. Que se passe-t-il si on fait varier un peu les processus d'arrivée et de service ? Est-ce que les résultats sur les indices de performances varient beaucoup ? La réponse empirique est :
"en général le comportement a la même allure et lorsque l'on n'est pas dans des conditions extrêmes les indices calculés sont de bonnes approximations"
(on trouvera dans [3] des indications sur des techniques de comparaison de files d'attente, en particulier avec les modèles Markoviens).

Le modèle d'une file d'attente étant robuste il devient naturel de les utiliser comme briques de base pour construire des systèmes où les clients se partagent plusieurs ressources. Par exemple, dans un système informatique, un processus peut être actif en mémoire centrale, interrompu pour des entrées/sorties ou un défaut de page. Dans un réseau de communication les paquets circulent sur une succession de liens et partagent ces liens avec d'autres messages...

Nous allons montrer dans cette section comment étendre les résultats obtenus sur une file d'attente à des réseaux de files d'attente. Pour comprendre précisément le comportement de tels réseaux nous étudierons d'abord les réseaux acycliques à routage probabiliste, puis les réseaux dits de Jackson et enfin les réseaux à routage prédéterminé.

2 Les réseaux acycliques

Le premier paramètre que l'on cherche à étudier dans les réseaux de files d'attente est le processus de sortie d'une file d'attente. On a montré ([2]) que le processus de sortie d'une file d'attente $M/M/1$ en régime stationnaire est un processus de Poisson de même intensité que le processus d'entrée des clients dans la file d'attente. Ce résultat, obtenu en utilisant un concept de réversibilité, (voir [2]) permet immédiatement de dire que les comportements de 2 files d'attente $M/M/1$ en cascade sont indépendants. Par suite, asymptotiquement les indices de performance du réseau se calculent "comme si" on disposait de 2 files d'attentes indépendantes.

Ceci se généralise au cas de réseaux acycliques de la manière suivante : on étudie N files d'attente de taux de service respectifs $\mu_1, \cdots, \mu_N$, chaque file d'attente reçoit des clients de l'extérieur selon des processus d'arrivée de Poisson, indépendants, de paramètres respectifs λ_i^0 pour la file i. A la sortie de la file i, le client choisit sa destination suivante (la station j) avec la probabilité $r_{i,j}$ et avec la probabilité $1 - \sum_j r_{i,j}$ il sort du réseau. $R = ((r_{i,j}))$ est appelée la <u>matrice de routage</u> du réseau. Lorsque le réseau est acyclique, en utilisant les propriétés de composition/décomposition des processus, on calcule l'intensité du processus de Poisson entrant dans chaque file (λ_i pour la file i).

L'état du réseau est déterminé par le nombre respectif de clients $(X_1, \cdots, X_N)$ dans chacune des files. On montre ensuite que l'état du système au cours du temps est un processus de Markov et qu'il est stable si

$$\forall i \ \ \rho_i = \frac{\lambda_i}{\mu_i} < 1.$$

Dans ce cas, à l'état stationnaire, on a :

$$\mathbb{P}(X_1 = x_1, \cdots, X_N = x_N) = \prod_{i=1}^{N} (1 - \rho_i) \rho_i^{x_i}.$$

Tout se passe comme si on avait N files d'attente "indépendantes" de charge respectives ρ_i. On utilise alors tout le formulaire obtenu pour les files $M/M/1$ pour calculer les performances du réseau.

3 Les réseaux de Jackson

En fait la propriété trouvée ci-dessus se généralise aux routage probabiliste quelconque. Soient donc N files d'attente connectées par un réseau de communication connexe. On utilise les mêmes notations que pour les réseaux acycliques, mais maintenant le routage est quelconque. Soit $R = ((r_{i,j}))$ la matrice de routage (matrice $N \times N$). En quittant la file i un client va vers la file j avec la probabilité $r_{i,j}$ ou quitte le réseau avec la probabilité :

$$1 - \sum_{j=1}^{N} r_{i,j}.$$

On montre d'abord le résultat sur les flots de clients :

Proposition 1
Dans le réseau, si de toute file il existe un routage de probabilité non nulle vers l'extérieur alors il existe une unique solution $(\lambda_1, \cdots, \lambda_N)$ *au système d'équations suivant :*

$$\lambda_j = \lambda_j^0 + \sum_i \lambda_i r_{i,j}.$$

Ceci s'interprète en disant que ce qui doit sortir de la file j (λ_j) doit être égal à ce qui entre dans la file j de l'extérieur (λ_j^0) plus ce qui est routé des autres files ($\sum_i \lambda_i r_{i,j}$). Nous pouvons maintenant déterminer la condition de stabilité du réseau comme :

$$\forall i \quad \rho_i = \frac{\lambda_i}{\mu_i} < 1.$$

Dans ce cas, à l'état stationnaire on a :

Théorème 1 (Jackson)
Lorsque le système est stable, la distribution stationnaire est à forme produit. Autrement dit, si X_i *est le nombre de clients dans la file* i *à l'état stationnaire on a :*

$$\mathbb{P}(X_1 = x_1, \cdots, X_N = x_N) = \prod_{i=1}^{N} (1 - \rho_i) \rho_i^{x_i}.$$

Tout se passe comme si on avait N files d'attente "indépendantes" de charge respectives ρ_i. Ce théorème est donc extrêmement utile car il permet d'obtenir des résultats analytiques simples sur des modèles Markoviens multi-dimensionnels en général intraitables par des méthodes numériques standard.

Contrairement aux réseaux de files d'attente acycliques, il faut remarquer que le processus d'arrivée de clients dans une file (de l'extérieur et des autres files) n'est pas un processus de Poisson. Cependant le processus de sortie des clients du réseau est un processus de Poisson.

4 Les extensions des réseaux de Jackson

En fait, ce qui est très utile c'est la forme algébrique du résultat. Il s'avère que des affaiblissements conséquents des hypothèses de ce résultat conduisent encore à des formes produit (donc numériquement calculables).

En remplaçant les taux de service μ_i par $\mu_i(x_i)$ la structure de la forme produit est conservée. Dans ce cas, le taux de service d'une file dépend de sa "charge de travail". On peut ainsi prendre par exemple des files multi-serveur.

De même on peut jouer sur le taux global d'arrivée des clients dans le réseau en prenant

$$\lambda_i^0 = \alpha(x_1 + \cdots + x_N) r_{0,i},$$

où $r_{0,i}$ représente la probabilité pour qu'un client arrivant de l'extérieur arrive à la station i.

On peut également fermer le système. Dans ce cas, il n'y a ni arrivées de l'extérieur, ni départs vers l'extérieur. On calcule alors une solution non nulle de l'équation d'équilibre des flots et on obtient le même style de résultats.

Malheureusement, le routage probabiliste ne permet pas toujours de modéliser les systèmes de communication (le choix du routage est déterminé à l'arrivée du paquet dans le réseau). Kelly a montré que pour des réseaux à routage déterministe, on a encore une forme produit (voir [2]).

En faisant la synthèse de tous ces résultats et en faisant varier les lois de service, les disciplines de service, les routages, on obtient le fameux théorème BCMP (Baskett, Chandy, Muntz et Palacios) qui explicite une forme produit pour de multiples configurations de réseaux [1].

Références

[1] F. Baskett, K.M. Chandy, R.R. Muntz, and F. Palacios. Open, closed and mixed networks of queues with different class of customers. *JACM*, 22, 1975.

[2] F.P. Kelly. *Reversibility and stochastic networks*. J. Wiley & Sons, 1979.

[3] D. Stoyan. *Comparaison Methods for Queues and other Stochastic Models*. J. Wiley and Son, 1976.

5 Exercices sur les réseaux de files d'attente

1. **Stations en cascades**
   On considère deux files d'attente en tandem. C'est à dire que le processus d'arrivée dans la deuxième file est égal au processus de départ de la première file. (voir figure 1).

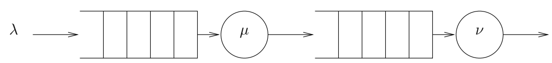

FIG. 1 – Stations en cascade

On note $N_t = (n^1_, t, n_t^2)$ le processus décrivant le nombre de clients dans le système à l'instant t le nombre de clients dans la file i est noté n_t^i.

(a) Montrez que le processus N_t est un processus de Markov homogène.

(b) Déterminez le diagramme d'états de N_t

(c) Montrez que le processus de Markov est irréductible. Que peut on en conclure ?

(d) Cherchez une solution stationnaire à ce processus, quelle est la condition nécessaire et suffisante d'ergodicité du système.

(e) Calculez le nombre moyen de clients dans le système ainsi que le temps de réponse moyen.

2. **Le modèle serveur central saturé**
On considère le modèle de files d'attente suivant (voir figure 2).

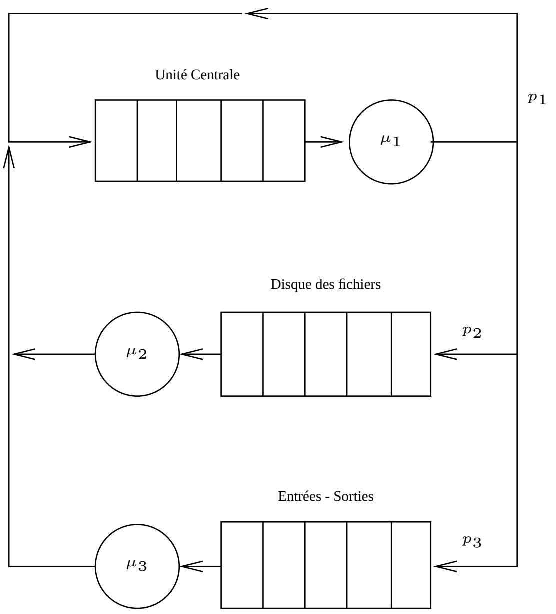

FIG. 2 – Modèle d'un système serveur central saturé

Le modèle est composé d'une unité centrale (file 1) et de $2..M$ périfériques. Soit p_i la probabilité qu'un client sortant de l'UC se dirige vers la station i. On suppose les serveurs des stations de loi exponentielle de taux μ_i pour la file i.

Écrire la matrice de routage et résoudre l'équation d'équilibre des flots.

Dans la suite de l'exercice on supposera les valeurs numériques suivantes :

$$M = 4 \quad \mu_1 = 12 \quad \mu_2 = \mu_3 = 1 \quad \mu_4 = 3.$$

On suppose que les p_i sont choisis tels que :

$$p_1 = 5/12 \quad p_2 = p_3 = 1/6.$$

Déterminer la capacité du réseau N de sorte que chaque taux d'activité U_i soit inférieur à 0.8. Le reste du temps étant utilisé pour la gestion du système, interruption de programme, gestion de contexte, reprise de résultats...

En déduire le nombre moyen de programmes entièrement exécutés par unité de temps, puis calculer la charge moyenne de chaque file, vérifier que $\mathbb{E}n_1 + \mathbb{E}n_2 + \mathbb{E}n_3 + \mathbb{E}n_4 = N$.

3. **Réseau fermé de files d'attente**

   Soit un réseau fermé constitué de M stations. On suppose que N le nombre de clients dans ce réseau reste constant. Les stations sont composées de files d'attente de serveur exponentiel de taux μ_i pour la station i.

   On définit la matrice de routage du réseau $P = ((p_{i,j}))$. On a $p_{i,j}$ la probabilité pour un client sortant de i d'aller dans la file j. On obtient donc $\sum_j p_{i,j} = 1$ car le réseau est fermé. On supposera de plus que de chaque station il existe un chemin vers toute autre station de probabilité non nulle.

   (a) Montrer qu'il existe une unique solution, à un coefficient multiplicatif près, à l'équation matricielle

   $$\underline{\lambda} = \underline{\lambda} P,$$

   avec

   $$\underline{\lambda} = (\lambda_1, ...\lambda_M) \quad et \quad \lambda_i > 0,$$

   Interpréter cette équation matricielle comme des équations d'équilibre des flots de clients.

   (b) Soit N_t le vecteur représentant le nombre de clients dans chaque file à l'instant t, $N_t = (n_{1,t}, ..., n_{M,t})$. Montrer que N_t est un processus de Markov. Calculer son générateur infinitésimal.

   En déduire les équations d'équilibre statistique, montrer par un argument simple qu'il existe toujours une solution stationnaire.

   (c) Soit $\rho_i = \frac{\lambda_i}{\mu_i}$. A l'état stationnaire montrer que :

   $$\Pr(n_1, ..., n_M) = \frac{1}{G(N, M)} \prod_{i=1}^{M} \rho_i^{n_i},$$

   où $G(N, M)$ constante de normalisation vérifie :

   $$G(N, M) = \sum_{n_1+\cdots+n_M=N} \prod_{i=1}^{M} \rho_i^{n_i}.$$

   (d) Montrer que

   $$\Pr(n_i \geqslant k) = \rho_i^k \frac{G(N-k, M)}{G(N, M)}.$$

   En déduire $\mathbb{E} n_i$ ainsi que le taux d'utilisation du serveur i à l'état stationnaire.

   (e) Calcul de $G(N, M)$

      i. Donner la relation liant $G(N, M)$ à $G(N, M-1)$,$G(N-1, M)$ et ρ_M.

      ii. En déduire un algorithme de calcul de $G(N, M)$.

      iii. Retrouver ces résultats de manière analytiques en utilisant des séries génératrices.