Chapitre 8

Files d'attente

8.1 Résultats généraux

La théorie des files d'attente est utile pour évaluer la performance de systèmes informatiques, dans lesquels les ressources sont partagées par plusieurs tâches, ou de réseaux de communication, qu'ils soient à commutation par paquets ou par circuits.

8.1.1 Notations

Une *file d'attente* est constituée d'un ou plusieurs *serveurs* et d'un *tampon* ("buffer") où les "clients" attendent d'être servis. Elle peut être caractérisée par six paramètres, ce qui donne la notation de Kendall

$$A/B/s/K/C/DS$$

où

- A spécifie la loi de *distribution des temps entre les arrivées* des clients. Par exemple, dans le contexte des réseaux de communication, ces clients sont les messages arrivant au noeud considéré.
- B spécifie la loi de *distribution des durées de service*. Le(s) serveur(s), dans le contexte des réseaux de communication, est (sont) la (les) ligne(s) de sortie.
- s spécifie le *nombre de serveurs*.
- K spécifie la *capacité* de la file, c'est-à-dire le nombre maximum de clients, ou de messages, susceptibles d'être stockés dans la file d'attente. Ce nombre inclut les clients qui attendent d'être servis, dans le tampon, ainsi que ceux qui sont servis.
- C spécifie la *population des clients*.

- DS spécifie la *discipline de service*, c'est-à-dire l'ordre dans lequel les clients arrivants sont rangés, puis sortis de la file d'attente. La discipline la plus courante est FIFO (First In, First Out : Premier arrivé, premier sorti).

Les trois derniers paramètres ne sont pas explicités si leurs valeurs sont $K = \infty$, $C = \infty$ et $DS = FIFO$. Les distributions des temps entre arrivées et des temps de service sont, par exemple,

- la loi exponentielle (M),
- une loi d'Erlang-k (E_k),
- une constante déterministe (D),
- une loi générale (G ou GI (GI signifie que les v.a. sont i.i.d., G qu'elles ne sont pas nécessairement indépendantes. Beaucoup d'auteurs ommettent cette distinction dans la notation, et utilisent G lieu de GI).

Par exemple, le symbole $M/M/1$ représente une file à un serveur avec distribution exponentielle des temps entre arrivées et des temps de service, dans laquelle la distribution de service est FIFO, avec une capacité et une population des clients infinie. Une telle file peut être schématisée par le dessin de la figure 8.1.

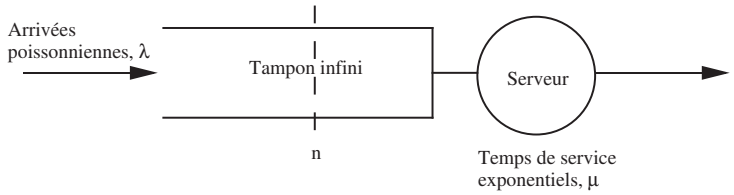


FIGURE 8.1 – File d'attente M/M/1.

Les grandeurs suivantes seront également utilisées :

- T est la v.a. décrivant le temps entre deux arrivées consécutives de clients.
- $1/\lambda$ est le temps moyen entre deux arrivées consécutives : $1/\lambda = E[T]$. Par conséquent, λ est le *taux d'arrivée moyen.*
- S est la v.a. décrivant le temps de service par client. Plus précisément, $S(n)$ est la v.a. décrivant le temps de service du $n$ième client.
- $1/\mu$ est la durée moyenne de service par client : $1/\mu = E[S]$. Par conséquent, μ est le *taux de service moyen par serveur.* Le taux de service moyen pour s serveurs est donc $s\mu$.
- Dans le contexte particulier d'un réseau de communication, le serveur est une ligne dont la *capacité* (fixe) vaut C bits/seconde. Le taux de service μ s'exprimant en messages/sec, on

introduit la notion de *longueur moyenne des messages*, $1/\mu'$, exprimée en bits/message, avec $\mu = \mu' C$. La durée moyenne de service par client : $1/\mu' C$ est alors le temps moyen de transmission par message à travers la ligne.

- N est la v.a. décrivant le nombre de clients séjournant dans la file d'attente. Ce nombre inclut les clients qui attendent d'être servis ainsi que ceux qui sont servis. Plus précisément, $N(t)$ est la v.a. décrivant le nombre de client présents dans le système au temps t.
- Q est la v.a. décrivant le nombre de clients séjournant dans le buffer . Ce nombre ne comprend donc que les clients qui attendent d'être servis.
- W est la v.a. décrivant le temps d'attente avant service
- R est la v.a. décrivant le *temps de réponse* du système, qui est la somme du temps d'attente avant service et du temps de service : $R = W + S$.

8.1.2 Stabilité ou ergodisme

Si le nombre de clients dans un système augmente continûment et devient infini, le système est instable. Pour que le système reste stable (ergodique), il faut que le taux moyen d'arrivées soit inférieur au taux moyen de service :

$$\lambda < s\mu. \tag{8.1}$$

Les processus $N(t)$, $R(t)$, etc sont alors ergodiques, dans le sens de la section 1.9 du module 7. Si la capacité de la file d'attente K ou la population des usagers N est *finie*, la file d'attente est toujours stable, et la condition (8.1) ne doit plus être nécessairement satisfaite.

8.1.3 Loi de Little

La loi de Little est valable pour tous les systèmes, ou les parties de systèmes, telles que le nombre de clients entrant dans le système est égal au nombre de clients sortant (aucun client n'est donc perdu à cause d'un tampon de taille insuffisante). Elle énonce simplement que

Nombre moyen de clients dans le système = Taux d'arrivée × temps de réponse moyen.

Cette formule est tout à fait générale pour les systèmes stables ou ergodiques. Même si certains clients sont perdus à cause de files de capacité limitée, cette loi peut être utilisée en ajustant le taux d'arrivées pour qu'il soit celui des clients entrant réellement dans le système.

Par exemple, si le système est sans perte, on a que

$$E[N] = \lambda E[R] \tag{8.2}$$
$$E[Q] = \lambda E[W]. \tag{8.3}$$

Pour justifier sommairement cette loi, supposons que le système soit initialement vide, et appelons $A(t)$ et $D(t)$ respectivement le nombre d'arrivées et de départs dans l'intervalle $[0, t]$. Alors le

nombre de clients dans le système au temps t est $N(t) = A(t) - D(t)$. La moyenne temporelle de $N(\cdot)$ sur l'intervalle $[0, t]$ est

$$< N >_t = \frac{1}{t} \int_0^t N(\tau) d\tau. \tag{8.4}$$

Cette intégrale est représentée à la figure 8.2 par l'aire comprise entre $A(t)$ et $D(t)$.

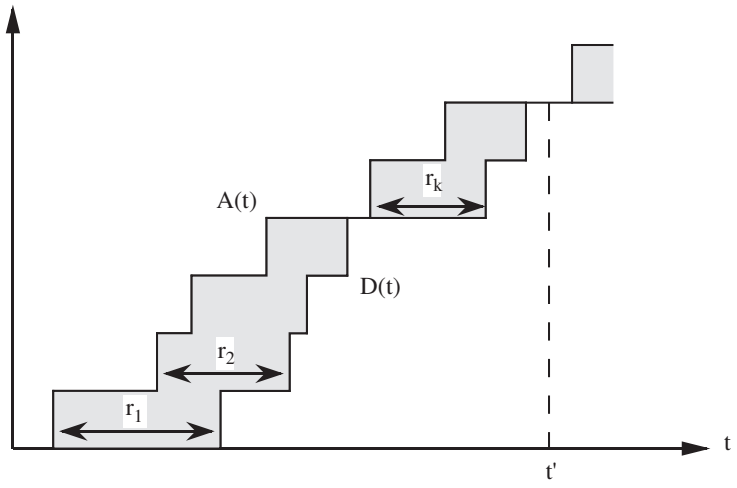


FIGURE 8.2 – Représentation de l'intégrale $t < N >_t$.

Soit r_k le temps passé par le $k$ième client dans le système. Plaçons-nous à un instant $t = t'$ où $N(t') = 0$, comme indiqué à la figure 8.2. Alors l'aire comprise entre $A(t)$ et $D(t)$, divisée par t', peut également s'exprimer comme

$$< N >_{t'} = \frac{1}{t'} \sum_{k=0}^{A(t')} r_k. \tag{8.5}$$

Le taux d'arrivée moyen pendant l'intervalle $[0, t']$ est

$$< \lambda >_{t'} = \frac{A(t')}{t'} \tag{8.6}$$

tandis que la moyenne temporelle du temps de séjour d'un client dans le système pendant ce même intervalle est

$$< R >_{t'} = \frac{1}{A(t')} \sum_{k=0}^{A(t')} r_k. \tag{8.7}$$

La combinaison des équations (8.5), (8.6) et (8.7) entraîne que

$$< N >_{t'} =< \lambda >_{t'} < R >_{t'} \tag{8.8}$$

On suppose le système ergodique de sorte qu'avec probabilité 1, les moyennes temporelles tendent vers les moyennes d'ensemble lorsque $t' \to \infty$, et donc (8.8) devient la loi de Little

$$E[N] = \lambda E[R]. \tag{8.9}$$

Finalement, comme on suppose le processus ergodique, il passera infiniment souvent par l'état $N = 0$, et donc le choix d'un temps t' tel que $N(t') = 0$ dans les formules précédentes n'est pas restrictif.

8.2 Files simples (uniques)

Les files les plus simples à analyser sont les files du types M/M/s/K, car ce sont des cas particuliers de processus de naissance et de mort, qui sont des chaînes de Markov à temps continu réversibles. Nous en avons rencontré un certain nombre au module précédent, nous nous bornons ici à résumer et compléter les résultats du module précédent, mais uniquement dans le cas stationnaire. Nous passons ensuite aux modèles plus généraux (et plus importants) où la loi de service n'est plus exponentielle. Toutes les files que nous considérons ici sont à l'état stationnaire.

8.2.1 File M/M/1

Nous avons vu au module précédent que la solution des équations de balance est

$$P(N = n) = \pi_n^\star = (1 - \rho)\rho^n \tag{8.10}$$

où $\rho = \lambda/\mu$ est appelé *intensité du trafic*. Rappelons que pour que le système soit stable, il faut que $\lambda < \mu$, ou encore que $\rho < 1$.

Connaissant la distribution des probabilités $\{\pi_n^\star\}$, il est aisé de calculer la fonction génératrice de N

$$G_N(z) = \sum_{n=0}^{\infty} \pi_n^\star z^n = \frac{1-\rho}{1-\rho z}. \tag{8.11}$$

Celle-ci aurait pu être également obtenue directement à partir des équations de balance sans passer par le calcul des probabilités $\pi_n^\star$.

Le nombre moyen $E[N]$ de clients dans la file d'attente est alors

$$E[N] = \left.\frac{dG_N(z)}{dz}\right|_{z=1} = \frac{\rho}{1-\rho}. \tag{8.12}$$

Pour déterminer la loi de probabilité de R, observons que si un client arrive quand la file comporte déjà n clients, son temps de séjour R dans la file sera

$$R = S(1) + S(2) + \cdots + S(n) + S(n+1)$$

ou les temps de service $S(n)$ sont une suite de v.a exponentielles i.i.d., de moyenne $1/\mu$ (même pour le client dans le serveur, car la v.a. exponentielle est sans mémoire (cfr module 1, exercice 4)). Comme la somme de $(n+1)$ v.a. exponentielles i.i.d. est une v.a. d'Erlang (cfr propriété P5 du module 5), la densité de probabilité conditionnelle de R s'il y a n clients dans la file est

$$f_{R|N=n}(r|N = n) = \frac{\mu(\mu r)^n e^{-\mu r}}{n!}.$$

En utilisant le théorème des probabilités totales, on en déduit que

$$f_R(r) = (\mu - \lambda)\, e^{-(\mu-\lambda)r}$$

avec $r > 0$, ce qui montre que les temps de réponse R sont distribués exponentiellement, avec une moyenne

$$E[R] = 1/(\mu - \lambda). \tag{8.13}$$

Remarquons que ce dernier résultat aurait pu être directement obtenu à partir de (8.12) grâce à la loi de Little, sans passer par le calcul de sa densité de probabilité. Pour un réseau de communication, on a

$$E[R] = 1/(\mu' C - \lambda). \tag{8.14}$$

On étudiera la distribution des autres v.a. Q, etc dans les exercices.

8.2.2 File M/M/s/K

Les files M/M/s/K sont des cas particuliers de processus de naissance et de mort, vus au module précédent, dans lesquels les taux de transition sont

$$\begin{aligned} \lambda_n &= \lambda \qquad \text{si } 0 \leq n \leq K-1 \\ \mu_n &= n\mu \qquad \text{si } 1 \leq n \leq s \\ &= s\mu \qquad \text{si } s+1 \leq n \leq K \end{aligned}$$

Nous avons déjà analysé au module précédent les cas M/M/s et M/M/s/s, qui sont les plus importants en pratique.

8.2.3 File M/GI/1 (M/G/1)

Le choix de lois exponentielles pour représenter les processus d'arrivée et de sortie permet de simplifier les calculs, mais peut ne pas toujours être très réaliste. Dans le cas d'un réseau de communication, si l'hypothèse d'arrivée des messages selon une loi de Poisson est proche de la réalité, car ils sont souvent produits par un grand nombre de sources indépendantes, l'hypothèse d'une distribution exponentielle de leur taille ne l'est pas, car leur longueur ne prend pas des valeurs quelconques, mais est toujours un multiple d'un bit ou d'un octet (byte). Il est donc très utile de connaître les caractéristiques d'un système M/G/1, où la densité de probabilité des temps de service $f_S(\sigma)$ est quelconque. Nous nous limiterons au cas où les temps de service sont indépendants les uns des autres (ainsi que des arrivées) : la dénomination exacte de cette file est en fait M/GI/1.

L'étude de ce cas est nettement plus compliquée, car $N(t)$ n'est plus une chaîne de Markov : les temps de séjour dans un état ne sont en effet plus des v.a. exponentielles. On peut néanmoins se construire une *chaîne de Markov induite* par la file M/G/1, en ne considérant que les temps $0 < S_d(1) < S_d(2) < \ldots < S_d(k) < \ldots$ auxquels un client quitte la file, le service étant terminé (ce n'est donc pas la chaîne induite par $N(t)$ en considérant les temps auxquels une transition se produit : ici on ne considère que les temps de départ, pas d'arrivée). Alors $\hat{N}(k) = N(S_d(k))$ est le nombre de clients laissés dans la file après le départ du $k$ième client. Nous allons maintenant montrer que le processus à temps discret $\{\hat{N}(k), k \in \mathbb{N}\}$ est effectivement une chaîne de Markov,

mais avant, il faut étudier les v.a. $A(k)$ décrivant le nombre de clients arrivant dans la file pendant que le $k$ième client est servi.

Les v.a. $S(k)$ décrivant la durée du service du $k$ième client sont indépendantes et ont la même densité de probabilité $f_S(s)$. Dès lors, comme les arrivées sont poissonniennes, toutes les v.a. $A(k)$ sont également indépendantes et identiquement distribuées, leur distribution de probabilité étant

$$a_n = P(A(k) = n) = \int_0^\infty P(A(k) = n|S(k) = s) f_S(s) ds = \int_0^\infty e^{-\lambda s} \frac{(\lambda s)^n}{n!} f_S(s) ds.$$

si $n \geq 0$ et $a_n = 0$ si $n < 0$. Leur fonction génératrice est

$$G_A(z) = \sum_{n=0}^{\infty} a_n z^n = \int_0^\infty e^{-\lambda s} \sum_{n=0}^{\infty} \left(\frac{(\lambda s z)^n}{n!} \right) f_S(s) ds = \int_0^\infty e^{\lambda s(z-1)} f_S(s) ds = \hat{\Phi}_A(\lambda(z-1)).$$

En d'autres termes, la fonction génératrice $G_A(z)$ des v.a. $A(k)$ est la fonction génératrice des moments, évaluée en $\lambda(z-1)$. La moyenne des v.a. $A(k)$ est obtenue en évaluant la dérivée de $G_A(z)$ en $z = 1$:

$$E[A(k)] = \left. \frac{dG_A(z)}{dz} \right|_{z=1} = \lambda \int_0^\infty s f_S(s) ds = \frac{\lambda}{\mu} = \rho. \tag{8.15}$$

A présent, on peut lier $\hat{N}(k+1)$ et $\hat{N}(k)$ par

$$\hat{N}(k+1) = \hat{N}(k) + A(k+1) - 1_{\{\hat{N}(k) \geq 1\}} = \begin{cases} \hat{N}(k) + A(k+1) - 1 & \text{si} \quad \hat{N}(k) \geq 1 \\ A(k+1) & \text{si} \quad \hat{N}(k) = 0. \end{cases} \tag{8.16}$$

Comme $\hat{N}(k+1)$ ne dépend que de $\hat{N}(k)$ et de la v.a. $A(k+1)$, $\hat{N}(k)$ est une chaîne de Markov. Les probabilités de transition de cette chaîne de Markov sont

$$\hat{q}_{mn} = P(\hat{N}(k+1) = n | \hat{N}(k) = m) = \begin{cases} P(A(k+1) = n - m + 1) = a_{n-m+1} & \text{si} \quad m \geq 1 \\ P(A(k+1) = n) = a_n & \text{si} \quad m = 0. \end{cases}$$

Comme on peut passer de tout état m à n'importe quel état n, la chaîne est irréductible. Comme $\hat{q}_{nn} > 0$, elle est apériodique. Enfin, on peut montrer que si la condition de stabilité est satisfaite, i.e. si $\rho = \lambda/\mu < 1$, tous ses états sont récurrents positifs.

En effet, en supposant la file vide intialement, on a

$$\hat{N}(k) = A(1) + \ldots + A(k) - k + Z(k) \tag{8.17}$$

où $Z(k)$ est le nombre de visites par l'état 0 avant le temps k :

$$Z(k) = \sum_{l=0}^{k-1} 1_{\{\hat{N}(l)=0\}}$$

Si l'état 0 est récurrent positif, une conséquence directe des théorèmes 2 et 5 du module 6 est que si $k \to \infty$

$$E[Z(k)]/k \to 1/E[T_0 \mid \hat{N}(0) = 0] > 0$$

où $E[T_0 \mid \hat{N}(0) = 0]$ est le temps moyen de premier retour à l'état 0. Par contre, si l'état 0 est transitoire, le nombre de visites par l'état 0 est par définition (presque sûrement) fini si bien que si $k \to \infty$,

$$E[Z(k)]/k \to 0.$$

Si les états sont récurrents nuls, on a de même

$$E[Z(k)]/k \to 1/E[T_0 \mid \hat{N}(0) = 0] = 0.$$

En prenant les espérances dans (8.17) et en divisant par k, on obtient

$$E[\hat{N}(k)/k] = \rho - 1 + E[Z(k)/k]$$

Comme $\hat{N}(k) \geq 0$, on en tire que si $\rho < 1$

$$E[Z(k)/k] \geq 1 - \rho > 0$$

ce qui entraîne que 0, et donc tous les états de cette chaîne irréductible, sont récurrents positifs. Dès lors, la chaîne $\hat{N}(k)$ est ergodique et toute distribution initiale de probabilité converge vers la distribution stationnaire (cfr module 6) $\{\hat{\pi}_n^\star\} = \{P(\hat{N} = n)\}$ donnée par

$$\begin{aligned}
\hat{\pi}_n^\star &= \sum_{m=0}^{\infty} \hat{q}_{mn}\hat{\pi}_m^\star = a_n\hat{\pi}_0^\star + \sum_{m=1}^{n+1} a_{n-m+1}\hat{\pi}_m^\star \\
&= a_n\hat{\pi}_0^\star + \sum_{m=0}^{n+1} a_{n-m+1}\hat{\pi}_m^\star - a_{n+1}\hat{\pi}_0^\star = a_n\hat{\pi}_0^\star + \sum_{m=0}^{\infty} a_{n-m+1}\hat{\pi}_m^\star - a_{n+1}\hat{\pi}_0^\star
\end{aligned}$$

où on a utilisé le fait que $a_{n-m+1} = 0$ pour $m > n + 1$.

En multipliant par z^n les membres de cette égalité, et en sommant pour tous les n, on obtient, après quelques manipulations, une expression de la fonction génératrice des probabilités $\hat{\pi}_n^\star = P(\hat{N} = n)$

$$\begin{aligned}
G_{\hat{N}}(z) &= \sum_{n=0}^{\infty} \hat{\pi}_n^\star z^n = \hat{\pi}_0^\star \sum_{n=0}^{\infty} a_n z^n + \sum_{n=0}^{\infty}\sum_{m=0}^{\infty} a_{n-m+1}\hat{\pi}_m^\star z^n - \hat{\pi}_0^\star \sum_{n=0}^{\infty} a_{n+1} z^n \\
&= \hat{\pi}_0^\star G_A(z) + \frac{1}{z}\sum_{m=0}^{\infty}\sum_{n=0}^{\infty} a_{n-m+1} z^{n-m+1}\hat{\pi}_m^\star z^m - \frac{\hat{\pi}_0^\star}{z}\sum_{n=0}^{\infty} a_{n+1} z^{n+1} \\
&= \hat{\pi}_0^\star G_A(z) + \frac{1}{z}\left[\sum_{m=0}^{\infty}\sum_{n=-1}^{\infty} a_{n-m+1} z^{n-m+1}\hat{\pi}_m^\star z^m - a_0\hat{\pi}_0^\star\right] - \frac{\hat{\pi}_0^\star}{z}\left[\sum_{n=-1}^{\infty} a_{n+1} z^{n+1} - a_0\right] \\
&= \hat{\pi}_0^\star G_A(z) + \frac{1}{z}\left[\sum_{m=0}^{\infty}\sum_{n'=0}^{\infty} a_{n'-m} z^{n'-m}\hat{\pi}_m^\star z^m\right] - \frac{\hat{\pi}_0^\star}{z}\sum_{n'=0}^{\infty} a_{n'} z^{n'} \\
&= \hat{\pi}_0^\star G_A(z) + \frac{1}{z}\sum_{m=0}^{\infty}\hat{\pi}_m^\star z^m \sum_{l=0}^{\infty} a_l z^l - \frac{\hat{\pi}_0^\star}{z} G_A(z) \\
&= \frac{\hat{\pi}_0^\star(z-1)}{z}\, G_A(z) + \frac{1}{z}\, G_A(z) G_{\hat{N}}(z)
\end{aligned}$$

d'où

$$G_{\hat{N}}(z) = \frac{\hat{\pi}_0^\star(z-1)G_A(z)}{z - G_A(z)}. \tag{8.18}$$

La valeur $\hat{\pi}_0^\star$ est déterminée par la propriété

$$1 = \lim_{z\to 1} G_{\hat{N}}(z) = \lim_{z\to 1} \hat{\pi}_0^\star \frac{G_A(z) + (z-1)\frac{dG_A}{dz}(z)}{1 - \frac{dG_A}{dz}(z)} = \hat{\pi}_0^\star \frac{1}{1 - E[A]}$$

et vaut donc, à partir de (8.15)

$$\hat{\pi}_0^\star = 1 - \rho. \tag{8.19}$$

Enfin, on peut montrer que les probabilités à l'équilibre $\pi_n^\star$ du processus à temps continu $N(t)$ sont identiques aux probabilités d'états $\hat{\pi}_n^\star$ de la chaîne à temps discret $\hat{N}(k)$ si bien que leurs fonctions génératrices coïncident et sont donc, en combinant (8.18) et (8.19)

$$G_N(z) = G_{\hat{N}}(z) = \frac{(1-\rho)(z-1)G_A(z)}{z - G_A(z)}. \tag{8.20}$$

Si la fonction $G_A(z)$ est calculable par (8.15), la formule (8.20) permet alors de déterminer les probabilités $\pi_n^\star$ en évaluant les dérivées d'ordre n de $G_N(z)$ à l'origine, ou en développant $G_N(z)$ en série de Mac-Laurin.

Le nombre moyen de clients dans la file est calculé après des manipulations simples mais fastidieuses par

$$E[N] = \lim_{z\to 1} \frac{dG_N}{dz}(z) = \frac{\rho}{1-\rho}\left[1 - \frac{\rho(1-\mu^2\sigma_S^2)}{2}\right], \tag{8.21}$$

où $\sigma_S^2 = E[(S-1/\mu)^2]$ est la variance du temps de service. On verra à l'exercice 7 une méthode plus directe permettant de trouver ce résultat, connu sous le nom de *formule de Pollaczek-Khintchine*, à moindres frais.

Dans le cas particulier où $G = M$, on a $\sigma_S^2 = 1/\mu^2$ et on retrouve bien le résultat obtenu à la section 8.2.1

8.2.4 File M/GI/∞ (M/G/∞)

Le système M/G/∞, ou plus précisément M/GI/∞, est un système dans lequel tout client entrant dans le système commence immédiatement à être "servi". Comme dans le cas précédent, les arrivées suivent un processus de Poisson de taux λ, mais les temps de service sont une suite de v.a. i.i.d. dont la densité de probabilité est $f_S(s)$. On va en fait étudier ce système – qui est plus simple que le système M/G/1, car les clients n'interagissent pas entre eux – non seulement à l'état stationnaire mais d'abord à l'état transitoire. Ce système est utilisé pour modéliser des délais aléatoires dans des réseaux de communication – avec l'hypothèse que les arrivées de paquets suivent un processus de Poisson.

Nous avons besoin du lemme suivant, qui complète les propriétés du processus de Poisson :

Lemme 5 *Soit $\{N(t), t \in \mathbb{R}^+\}$ un processus de Poisson homogène, de taux $\lambda > 0$. Sachant que $N(t) = n$,*

(i) la densité de probabilité jointe des temps de ces n arrivées $S(1)$, $S(2)$, ..., $S(n)$ est

$$f_{S(1)\ldots S(n)|N(t)=n}(s_1, s_2, \ldots, s_n | N(t) = n) = \frac{n!}{t^n} \tag{8.22}$$

avec $0 < s_1 < s_2 < \ldots < s_n < t$.

(ii) le temps auquel l'une de ces n arrivées, prise arbitrairement parmi celles-ci, a eu lieu, est distribué uniformément dans $[0, t]$.

Preuve: (i) Pour une séquence $0 < s_1 < s_2 < \ldots < s_n < t$, l'évènement $\{S(1) = s_1, S(2) = s_2, \ldots, S(n) = s_n\} \cap \{N(t) = n\}$ est équivalent à l'évènement $\{T(0) = s_1, T(1) = s_2 - s_1, \ldots, T(n-1) = s_n - s_{n-1}\} \cap \{T(n) > t - s_n\}$, où $\{T(n), n \in \mathbb{N}\}$ est la séquence des temps entre arrivées, qui sont des v.a. exponentielles i.i.d., de taux λ. Par conséquent, en prenant la limite de

$$\begin{aligned} & P(s_1 \leq S(1) < s_1 + \Delta s_1, s_2 \leq S(2) < s_2 + \Delta s_2, \ldots, s_n \leq S(n) < s_n + \Delta s_n \mid N(t) = n) \\ = \; & \frac{P(s_1 \leq S(1) < s_1 + \Delta s_1, s_2 \leq S(2) < s_2 + \Delta s_2, \ldots, s_n \leq S(n) < s_n + \Delta s_n, N(t) = n)}{P(N(t) = n)} \end{aligned}$$

pour $\Delta s_1, \ldots, \Delta s_n \to 0$, on déduit que

$$\begin{aligned}
& f_{S(1)\ldots S(n)|N(t)=n}(s_1, s_2, \ldots, s_n | N(t) = n) \\
&= \frac{f_{S(1)\ldots S(n),N(t)}(s_1, s_2, \ldots, s_n, n; t)}{P(N(t) = n)} \\
&= \frac{f_{T(0)\ldots T(n-1)}(s_1, s_2 - s_1, \ldots, s_n - s_{n-1}) P(T(n) > t - s_n)}{P(N(t) = n)} \\
&= \frac{f_{T(0)}(s_1) f_{T(1)}(s_2 - s_1) \ldots f_{T(n-1)}(s_n - s_{n-1}) P(T(n) > t - s_n)}{P(N(t) = n)} \\
&= \frac{\lambda e^{-\lambda s_1} \lambda e^{-\lambda(s_2 - s_1)} \ldots \lambda e^{-\lambda(s_n - s_{n-1})} e^{-\lambda(t - s_n)}}{(\lambda t)^n e^{-\lambda t}/n!} = \frac{\lambda^n e^{-\lambda t}}{(\lambda t)^n e^{-\lambda t}/n!} = \frac{n!}{t^n}
\end{aligned}$$

(ii) Pour cette seconde partie, il faut introduire la notion de *statistique d'ordre.*

Soient $U_1, \ldots, U_n$ n v.a. i.i.d. uniformément distribuées sur $[0, t]$, à partir desquelles on forme la suite de n v.a. $V_1, \ldots, V_n$ obtenues en rangeant les v.a. $U_1, \ldots, U_n$ dans l'ordre croissant : $V_1 = \min\{U_1, \ldots, U_n\}$, $V_2 = \min\{\{U_1, \ldots, U_n\} \setminus \{V_1\}\}$, ..., $V_n = \max\{U_1, \ldots, U_n\}$. On dit que la suite des v.a. $V_1, \ldots, V_n$ forme la *statistique d'ordre* correspondant aux v.a. $U_1, \ldots, U_n$. Leur densité jointe est

$$f_{V_1\ldots V_n}(v_1, \ldots, v_n) = n! f_{U_1}(v_1) \ldots f_{U_n}(v_n) = \frac{n!}{t^n} \tag{8.23}$$

En effet, $(V_1, \ldots, V_n) = (v_1, \ldots, v_n)$ si une des $n!$ permutations de $U_1, \ldots, U_n$ est égale à $(v_1, \ldots, v_n)$, et la densité de probabilité jointe de $(U_1, \ldots, U_n)$ est le produit des densités marginales car ces v.a. sont indépendantes. En comparant (8.22) avec (8.23), on constate que sachant que $N(t) = n$, les temps d'arrivées $S(1)$, $S(2)$, ..., $S(n)$ ont la même distribution que la statistique d'ordre de n v.a. indépendantes uniformément distribuées entre 0 et t, autrement dit qu'un échantillon ordonné de valeurs tirées suivant une loi uniforme dans $[0, t]$. Ceci entraîne que si on sait qu'il y a eu n arrivées en $[0, t]$, le temps à laquelle l'une d'elles, prise arbitrairement parmi ces n arrivées, a eu lieu, est distribué uniformément entre 0 et t. ■

Retournons maintenant à notre système M/G/∞. Supposons que le système soit initialement vide, et appelons $A(t)$ et $D(t)$ respectivement le nombre d'arrivées et de départs dans l'intervalle $[0, t]$. Alors le nombre de clients dans le système au temps t est $N(t) = A(t) - D(t)$.

Supposons à présent que m clients soient arrivés dans le système en $[0, t]$: $A(t) = m$. Si on prend un de ces m clients au hasard, le temps de son arrivée sera dès lors une v.a. uniformément distribuée entre 0 et t, en vertu du lemme 1 ci-dessus. Par conséquent, la probabilité qu'un de ces m clients, pris au hasard parmi ceux-ci, n'ait pas encore terminé son service au temps t vaut, à partir du théorème des probabilités totales au cas continu

$$\begin{aligned}
p(t) &= \int_0^t P(S > t - \tau \mid \text{cette arrivée a eu lieu au temps } \tau) \cdot \frac{1}{t}\, d\tau \\
&= \frac{1}{t} \int_0^t P(S > t - \tau) d\tau = \frac{1}{t} \int_0^t P(S > s) ds = \frac{1}{t} \int_0^t (1 - F_S(s)) ds
\end{aligned}$$

où $F_S(s)$ est la fonction de répartition de S.

Par conséquent, conditionnellement à $A(t) = m$, le nombre de clients qui n'ont pas encore terminé leur service au temps t est décrit par une v.a. binomiale Binom (m, p), si bien que

$$P(N(t) = n \mid A(t) = m) = C_m^n p^n (1-p)^{m-n}$$

avec $0 \leq n \leq m$ et avec p donné ci-dessus. Dès lors, comme $A(t)$ est un processus de Poisson de taux λ,

$$\begin{aligned}
\pi_n(t) &= P(N(t) = n) = \sum_{m=0}^{\infty} P(N(t) = n \mid A(t) = m) P(A(t) = m) \\
&= \sum_{m=n}^{\infty} C_m^n p^n (1-p)^{m-n} \frac{(\lambda t)^m}{m!} e^{-\lambda t} = e^{-\lambda t} \frac{(\lambda t p)^n}{n!} \sum_{m=n}^{\infty} \frac{((1-p)\lambda t)^{m-n}}{(m-n)!} \\
&= e^{-\lambda t} \frac{(\lambda t p)^n}{n!} e^{(1-p)\lambda t} = \frac{(\lambda t p(t))^n}{n!} e^{-\lambda t p(t)}
\end{aligned}$$

ce qui montre que $N(t)$ est une v.a. de Poisson de moyenne $\lambda p(t) t$. Observons que dans le cas où $G = M$, on a

$$p(t) = \frac{1}{t} \int_0^t e^{-\mu s} ds = \frac{1}{\mu t} \left(1 - e^{-\mu t}\right),$$

ce qui entraîne que

$$\pi_n(t) = \frac{1}{n!} \left[(1 - e^{-\mu t}) \frac{\lambda}{\mu}\right]^n \exp\left[-(1 - e^{-\mu t}) \frac{\lambda}{\mu}\right]$$

et on retrouve bien le résultat de l'exercice 7 du module 7.

Maintenant, dans le cas général à l'équilibre, on trouve que

$$\begin{aligned}
\lim_{t \to \infty} p(t) t &= \int_0^{\infty} P(S > s) ds = \int_0^{\infty} \left[\int_s^{\infty} f_S(s') ds'\right] ds \\
&= \int_0^{\infty} \left[\int_0^{s'} f_S(s') ds\right] ds' = \int_0^{\infty} s' f_S(s') ds' = E[S] = \frac{1}{\mu}
\end{aligned}$$

Par conséquent, en régime stationnaire, le nombre de clients dans le système suit une loi de Poisson de paramètre $\lambda/\mu = \rho$. On a donc

$$P(N = n) = \pi_n^{\star} = \frac{\rho^n}{n!} e^{-\rho} \tag{8.24}$$

$$E[N] = \rho \tag{8.25}$$

On peut recommencer le même raisonnement pour le processus de départ $D(t)$. On suppose tout d'abord que m clients soient arrivés dans le système en $[0, t]$: $A(t) = m$. Cette fois-ci la probabilité qui nous intéresse est la probabilité qu'un de ces clients, pris au hasard parmi les m clients, ait déjà terminé son service au temps t. Elle vaut donc

$$q(t) = 1 - p(t) = \frac{1}{t} \int_0^t (1 - P(S > s)) ds = \frac{1}{t} \int_0^t F_S(s) ds.$$

Conditionnellement à $A(t) = m$, le nombre de clients qui ont terminé leur service au temps t est décrit par une v.a. binomiale $Bin(m, q)$, si bien que

$$P(D(t) = n \mid A(t) = m) = C_m^n q^n (1-q)^{m-n}$$

avec $0 \leq n \leq m$ et avec q donné ci-dessus. Dès lors, comme $A(t)$ est un processus de Poisson de taux λ, on trouve de manière similaire que

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

ce qui montre que $D(t)$ est une v.a. de Poisson de moyenne $\lambda t q(t)$. Comme

$$\lim_{t\to\infty} q(t) = 1 - \lim_{t\to\infty} p(t) = 1 - 0 = 1,$$

la loi de $D(t)$ en régime stationnaire est

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

Par conséquent, en régime stationnaire, le nombre de clients quittant le système suit une loi de Poisson de paramètre λt. On peut montrer en fait que $\{D(t), t \in \mathbb{R}^+\}$ est un processus de Poisson de taux λ, et est indépendant de $N(t)$, ce qui étend le théorème de Burke aux files M/G/∞.

8.2.5 File M/G/s/s

Le système M/G/∞ est un cas limite du système M/G/s/s (en fait M/GI/s/s) lorsque $s \to \infty$. L'étude pour s fini est nettement plus difficile, mais produit la même probabilité de rejet (blocage) qu'un système M/M/s/s ! La formule d'Erlang-B reste donc valable pour des temps de services dont la distribution est quelconque, et ne dépend que leur valeur moyenne $1/\mu$:

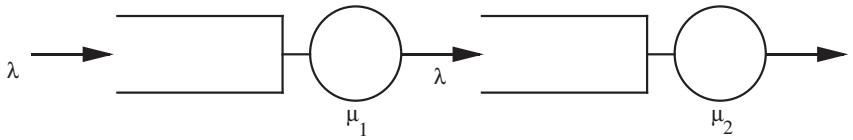
$$P(\text{rejet}) = \pi_s^\star = \frac{(\lambda/\mu)^s/s!}{\sum_{n=0}^{s} (\lambda/\mu)^n/n!} = B(s, \lambda/\mu). \tag{8.26}$$

8.3 Réseaux ouverts de files d'attente

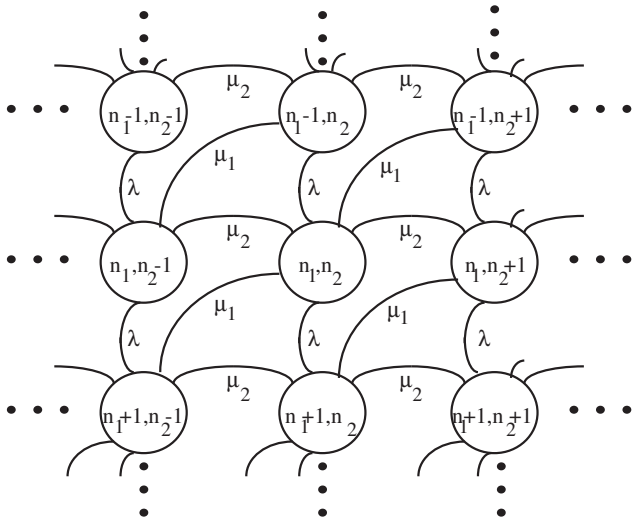
Nous allons survoler les *réseaux ouverts* de files d'attente sans pertes. On fera l'hypothèse très (trop) simplificatrice que tous les flux entrants et les temps de services sont exponentiels, ce qui mène à des solutions produits élégantes mais dont l'impact sur les réseaux de communication actuels reste cependant fort limité. Néanmoins, comme on l'a déjà observé avec les files simples, le calcul des files d'attentes devient ardu dès qu'on s'écarte des modèles exponentiels. L'analyse de performance de réseaux dans le cas général et réaliste reste donc largement ouvert...

8.3.1 Cascade de files M/M/1

On désire connaître les caractéristiques d'un réseau ouvert formé de la cascade de deux files M/M/1 comme celui représenté à la figure 8.3(a), où $\lambda < \mu_1, \mu_2$. L'état de ce système est à présent composé de deux variables, n_1 et n_2, qui sont le nombre de clients respectivement dans les files 1 et 2. Le diagramme des transitions entre états est donné à la figure 8.3(b).



(a)



(b)

FIGURE 8.3 – Cascade de deux files d'attente (a) et diagramme des transitions entre états (b).

En égalant les taux de transition vers et hors de l'état (n_1, n_2), on obtient les équations de

balance

$$\begin{aligned}
\lambda\pi^\star_{0,0} &= \mu_2\pi^\star_{0,1} \\
&\vdots \\
(\lambda+\mu_2)\pi^\star_{0,n_2} &= \mu_1\pi^\star_{1,n_2-1} + \mu_2\pi^\star_{0,n_2+1} \\
&\vdots \\
(\lambda+\mu_1)\pi^\star_{n_1,0} &= \lambda\pi^\star_{n_1-1,0} + \mu_2\pi^\star_{n_1,1} \\
&\vdots \\
(\lambda+\mu_1+\mu_2)\pi^\star_{n_1,n_2} &= \lambda\pi^\star_{n_1-1,n_2} + \mu_1\pi^\star_{n_1+1,n_2-1} + \mu_2\pi^\star_{n_1,n_2+1} \\
&\vdots
\end{aligned}$$

D'autre part, on sait par le théorème de Burke que les processus de départ de files M/M/1 sont des processus de Poisson homogènes, de même taux que les processus d'entrée et indépendants du nombre de clients dans le système. On peut donc considérer chaque file isolément de l'autre. On peut vérifier que la solution de ces équations de balances est de fait

$$\pi^\star_{n_1,n_2} = (1-\rho_1)\rho_1^{n_1}(1-\rho_2)\rho_2^{n_2}$$

où $\rho_1 = \lambda/\mu_1$ et $\rho_2 = \lambda/\mu_2$.

8.3.2 Théorème de Jackson

Le réseau de la section précédente n'est qu'un exemple très simple de réseau ouvert, pour lequel la solution a une forme produit, puisque $\pi^\star_{n_1,n_2} = \pi^\star_{n_1}\pi^\star_{n_2}$ où $\pi^\star_{n_i}$ est la probabilité d'avoir n_i clients dans une file M/M/1. Plus généralement, considérons un réseau ouvert,

- formé de M files d'attentes de type ./M/1 (On ne peut faire d'hypothèse sur la distribution des arrivées à chaque file), dont le vecteur des taux de service est $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_M)^T$
- avec des arrivées extérieures poissonniennes, de moyennes $\boldsymbol{\lambda_s} = (\lambda_{s1}, \ldots, \lambda_{sM})^T$,
- et une matrice de routage R, dont chaque composante r_{ij} est la probabilité qu'un client sortant de la file i aille dans la file j. La probabilité qu'il sorte du système est alors $1 - \sum_{i=1}^M r_{ij}$.

On peut alors calculer les taux moyens des arrivées dans chaque file $1 \leq i \leq M$. En effet,

$$\lambda_i = \lambda_{si} + \sum_{j=1}^{M} r_{ji}\lambda_j.$$

Sous forme vectorielle, cette équation devient

$$\boldsymbol{\lambda} = \boldsymbol{\lambda_s} + R^T\boldsymbol{\lambda}$$

ou encore

$$\boldsymbol{\lambda} = (I - R^T)^{-1}\boldsymbol{\lambda_s} \tag{8.27}$$

où I dénote la matrice unité d'ordre M.

Le théorème de Jackson énonce alors qu'à l'équilibre la probabilité que l'état du système soit $(n_1, n_2, \ldots, n_M)$ est

$$\pi^\star_{n_1,n_2,\ldots,n_M} = \pi^\star_{n_1}\pi^\star_{n_2}\cdots\pi^\star_{n_M} \tag{8.28}$$

où $\pi^\star_{n_i}$ est la probabilité que la file i prise isolément soit dans l'état n_i, étant soumise à un flux d'arrivées de Poisson de moyenne λ_i. Notons qu'on peut encore réécrire (8.28) sous la forme

$$\pi^\star_{n_1,n_2,\ldots,n_M} = \pi^\star_{0,0,\ldots,0}\prod_{i=1}^{M} f(n_i) \tag{8.29}$$

où $f(n_i)$ est une fonction ne dépendant que de n_i (Par exemple, dans le cas de serveurs ./M/1, $f(n_i) = (\lambda_i/\mu_i)^{n_i}$ tandis que dans le cas de serveurs ./M/∞, $f(n_i) = (\lambda_i/\mu_i)^{n_i}/n_i!$) et où la probabilité qu'il n'y ait aucun client dans le système est

$$\pi^\star_{0,0,\ldots,0} = \left(\sum_{n_1,\ldots,n_M}\prod_{i=1}^{M} f(n_i)\right)^{-1}. \tag{8.30}$$

Il est donc très facile d'analyser un réseau ouvert de files d'attente dans le cas exponentiel, car sa solution est la même que celle du cas particulier où toutes les files sont indépendantes les unes des autres ! De plus, ce réseau peut être sans rebouclage, comme la cascade de files étudiée à la section 1.1, dans laquelle tous les flux internes sont poissonniens, ou avec rebouclage, auquel cas tous les flux internes ne sont plus poissonniens. Ceci montre la puissance de ce théorème, qui est toutefois contrebalancée par le fait que les temps de services des files avec attente (buffer) doivent être exponentiels. Prenez un réseau ATM ou IP : c'est loin d'être le cas.

8.3.3 Analyse de performance

Le nombre moyen de clients $E[N]$ dans le système est tout simplement la somme du nombre moyen de clients $E[N_i]$ dans chaque file d'attente i :

$$E[N] = \sum_{i=1}^{M} E[N_i] \tag{8.31}$$

tandis que le temps de réponse moyen du réseau peut être obtenu par la formule de Little, si les flux moyens d'entrée et de sortie sont égaux à $\Lambda_s = \sum_{i=1}^{M}\lambda_{si}$:

$$E[R] = \frac{E[N]}{\Lambda_s} = \frac{1}{\Lambda_s}\sum_{i=1}^{M}\lambda_i E[R_i]. \tag{8.32}$$

Grâce au théorème de Jackson, l'analyse de performance d'un réseau ouvert de files d'attente se ramène à l'évaluation de performance de files d'attente simples, une fois les flux internes déterminés. Par exemple, si toutes les files sont du type ./M/1, on a

$$E[N] = \sum_{i=1}^{M} \frac{\rho_i}{1-\rho_i}, \tag{8.33}$$

avec $\rho_i = \lambda_i/\mu_i$, et

$$E[R] = \frac{1}{\Lambda_s} \sum_{i=1}^{M} \frac{\lambda_i}{\mu_i - \lambda_i}. \tag{8.34}$$

8.3.4 Application 1 : Modèle à serveur central d'un ordinateur

Le réseau ouvert de la figure 8.4, dans lequel il y a rebouclage, est appelé *modèle à serveur central*. Le processeur central (CPU) est le serveur qui distribue les tâches ("jobs") aux autres appareils (ici, les disques A et B). Après service, ces tâches retournent au CPU, et le quittent lorsqu'elles sont terminées ou lors d'une nouvelle interruption entrée-sortie (E/S). On veut analyser certaines performances de cet ordinateur, en déduire quels sont les appareils à améliorer, et notamment vérifier si un appareil en particulier est un goulot d'étranglement ("bottleneck") du système.

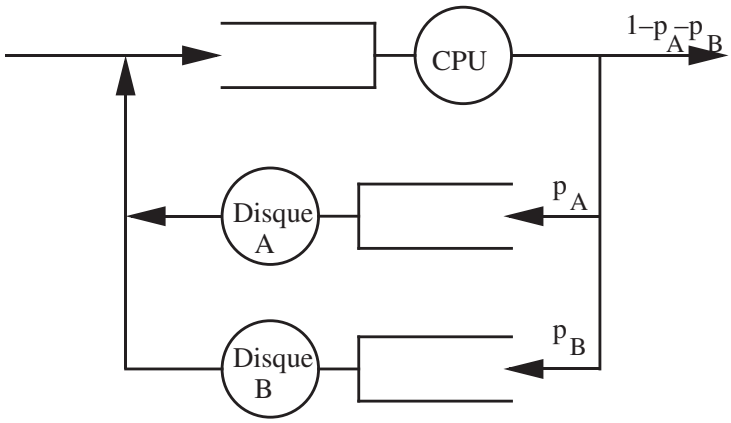


FIGURE 8.4 – Modèle à serveur central.

Le CPU et les deux disques sont des files ./M/1. Le nombre moyen de tâches que doit traiter l'ordinateur est $\Lambda_s = 3$ tâches par seconde. Les temps de service du CPU et des deux disques sont respectivement $1/\mu_{CPU} = 0.005$ sec, $1/\mu_A = 0.02$ sec et $1/\mu_B = 0.03$ sec. Après avoir reçu son service dans le CPU, une tâche a une probabilité $p_A = 0.4$ d'être envoyée vers le disque A et $p_B = 0.5$ vers le disque B.

La matrice de routage d'un tel système est

$$R = \begin{bmatrix} 0 & p_A & p_B \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix},$$

et on obtient les taux d'arrivées internes

$$\begin{bmatrix} \lambda_{CPU} \\ \lambda_A \\ \lambda_B \end{bmatrix} = \begin{bmatrix} 1 & -1 & -1 \\ -p_A & 1 & 0 \\ -p_B & 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} \Lambda_s \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 1/(1-p_A-p_B) \\ p_A/(1-p_A-p_B) \\ p_B/(1-p_A-p_B) \end{bmatrix} \Lambda_s = \begin{bmatrix} 30 \\ 12 \\ 15 \end{bmatrix}.$$

Le temps de réponse d'un tel système est alors donné par (8.34), et vaut

$$E[R] = \frac{1}{\Lambda_s}\left(\lambda_{CPU}E[R_{CPU}] + \lambda_A E[R_A] + \lambda_B E[R_B]\right) = \frac{1}{3}\left(0.176 + 0.316 + 0.818\right) = 0.437 \text{ sec.}$$

Il est clair que l'appareil formant le goulot d'étranglement du système est le disque B. Si on le change par un disque deux fois plus rapide, $E[R_B] = 0.019$ sec et le temps de réponse moyen du système devient $E[R] = 0.261$ sec, ce qui représente une amélioration d'environ 40% du temps de réponse. Si on avait divisé par deux le temps de service du disque A au lieu du disque B, cette amélioration n'aurait été que de 14%, ce qui montre l'intérêt de localiser le goulot d'étranglement du système.

Le concepteur peut se poser d'autres questions, comme par exemple de savoir si l'utilisation d'un seul disque, mettons le disque A, vers lequel toutes les opérations d'entrées-sorties sont dirigées, dégrade fortement les performances (La taille mémoire étant supposée suffisante). Si ce n'est pas le cas, cette solution serait évidemment plus économique (cfr exercice 10) !

8.3.5 Application 2 : conception d'un réseau de comunication

Une autre application du théorème de Jackson concerne l'optimisation (très grossière) de réseaux à commutation par paquets.

On cherche à assigner les capacités C_i des lignes de transit d'un réseau dont la topologie et les flux internes λ_i sont supposés fixés. Les arrivées de paquets dans le réseau, dont le taux moyen total d'arrivée vaut Λ_s paquets/sec, sont supposées poissonniennes, et leurs longueurs $1/\mu_i'$ sont distribuées exponentiellement (Hypothèse d'indépendance de Kleinrock : les paquets perdent leur identité à leur entrée dans un noeud, qui les reconstruit avec une longueur distribuée exponentiellement. Les longueurs des paquets sont donc indépendantes d'un noeud à l'autre, ce qui est approximativement vérifié en réalité).

Par conséquent, le temps de traversée moyen du réseau est donné par (8.34) avec $\mu_i = \mu_i' C_i$.

Le coût de chaque ligne de transit sera supposé être une fonction linéaire de sa capacité. Le coût total du réseau est alors

$$W = \sum_{i=1}^{M}(d_i C_i + f_i)$$

où d_i sont les coûts variables et f_i les coûts fixes.

Le problème qui se pose alors est de calculer les capacités C_i de chaque ligne qui minimisent le coût total du réseau P tout en maintenant un temps moyen de traversée $E[R]$ raisonnable,

c'est-à-dire inférieur à une certaine limite $R_{\max}$:

$$\frac{1}{\Lambda_s}\sum_{i=1}^{M}\frac{\lambda_i}{\mu'_i C_i - \lambda_i} \leq R_{\max}. \tag{8.35}$$

La méthode utilisée est celle des multiplicateurs de Lagrange. On cherche à minimiser la fonction

$$F = W - \alpha(E[R] - R_{\max}) = \sum_{i=1}^{M}(d_i C_i + f_i) - \alpha(\frac{1}{\Lambda_s}\sum_{i=1}^{M}\frac{\lambda_i}{\mu'_i C_i - \lambda_i} - R_{\max}),$$

où α est le *multiplicateur de Lagrange*.

Le minimum est atteint en annulant toutes les dérivées partielles de F par rapport à C_i

$$\frac{\partial F}{\partial C_i} = d_i - \frac{\alpha\lambda_i\mu'_i}{\Lambda_s(\mu'_i C_i - \lambda_i)^2} = 0,$$

ce qui entraîne que

$$\frac{1}{\mu'_i C_i - \lambda_i} = \sqrt{\frac{\Lambda_s d_i}{\alpha\lambda_i\mu'_i}}.$$

Après avoir multiplié les deux membres de cette relation par λ_i/Λ_s, et sommé sur toutes les lignes, on trouve

$$\sqrt{\alpha} = \frac{1}{\Lambda_s R_{\max}}\sum_{i=1}^{M}\sqrt{\frac{\Lambda_s d_i\lambda_i}{\mu'_i}}$$

où on a pris le signe d'égalité dans (8.35). On en déduit alors que les capacités optimales sont

$$C_i^\star = \frac{\lambda_i}{\mu'_i} + \frac{\sum_{i=1}^{M}\sqrt{\frac{d_i\lambda_i}{\mu'_i}}}{\Lambda_s R_{\max}}\sqrt{\frac{\lambda_i}{d_i\mu'_i}}$$

et le coût minimum

$$W^\star = \sum_{i=1}^{M}(d_i\frac{\lambda_i}{\mu'_i} + f_i) + \frac{1}{\Lambda_s R_{\max}}\left(\sum_{i=1}^{M}\sqrt{\frac{d_i\lambda_i}{\mu'_i}}\right)^2.$$

En pratique, les capacités ne peuvent pas prendre n'importe quelle valeur, mais seulement un nombre fini de valeurs différentes (32 kbps, 64 kbps, 1555 Mbps, ...), et les coûts ne sont pas linéaires avec la capacité (économies d'échelle). De plus, les flux et la topologie du réseau doivent être également déterminés pour optimiser les coûts. Le problème est donc plus compliqué, et ne peut être résolu que numériquement. Qui plus est, les flux ne sont pas Poissonniens, et les longueurs de paquest ne sont pas exponentielles.

8.4 Réseaux fermés de files d'attente

L'étude des *réseaux fermés* de files d'attente, dans lesquels circulent N clients, et formés d'un nombre quelconque M de files ./M/s/K/N, est plus complexe que l'analyse des réseaux ouverts vue à la section précédente.

8.4.1 Cascade de files ./M/1

On désire connaître les caractéristiques d'un réseau fermé composé de la cascade (rebouclée) de deux files ./M/1, comme celui représenté à la figure 8.5(a). Le système ne reçoit pas de clients venant de l'extérieur, aussi le nombre de clients N à l'intérieur du système est-il constant. Les deux variables formant l'état du système, n_1 et n_2, sont donc liées par la relation

$$N = n_1 + n_2. \tag{8.36}$$

Le diagramme des transitions entre états est donné à la figure 8.5(b).

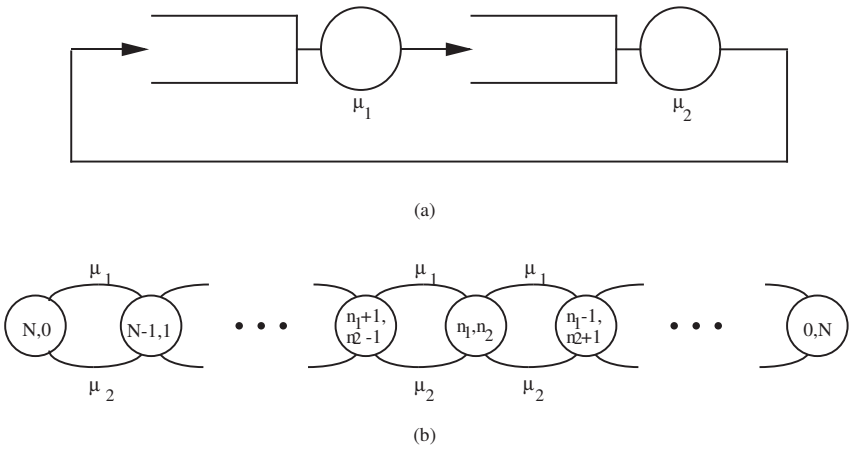


FIGURE 8.5 – Cascade rebouclée de deux files (a) et diagramme des transitions entre états (b).

Les équations de balance détaillée sont

$$\begin{aligned}
\mu_1 \pi^\star_{N,0} &= \mu_2 \pi^\star_{N-1,1} \\
&\vdots \\
\mu_1 \pi^\star_{n_1,n_2} &= \mu_2 \pi^\star_{n_1-1,n_2+1} \\
&\vdots \\
\mu_1 \pi^\star_{1,N-1} &= \mu_2 \pi^\star_{0,N}.
\end{aligned}$$

ou encore, en tenant compte de (8.36),

$$\pi^\star_{n_1,N-n_1} = (\mu_2/\mu_1)\pi^\star_{n_1-1,N-n_1+1} = \cdots = (\mu_2/\mu_1)^{n_1}\pi^\star_{0,N}.$$

La valeur de $\pi^\star_{0,N}$ est obtenue en sommant toutes les probabilités :

$$\pi^\star_{0,N} \sum_{n_1=0}^{N} (\mu_2/\mu_1)^{n_1} = 1$$

d'où $\pi^\star_{0,N} = (1-\mu_2/\mu_1)/(1-(\mu_2/\mu_1)^{N+1})$. Par conséquent, on a

$$\pi^\star_{n_1,n_2} = \frac{\mu_1^{N+1}\mu_2^N - \mu_1^N\mu_2^{N+1}}{\mu_1^{N+1}-\mu_2^{N+1}} \left(\frac{1}{\mu_1}\right)^{n_1} \left(\frac{1}{\mu_2}\right)^{n_2}. \tag{8.37}$$

8.4.2 Théorème de Jackson pour les réseaux fermés

Le réseau de la section 8.4.1 n'est qu'un exemple très simple de réseau fermé, mais pour lequel, de manière similaire aux réseaux ouverts, la solution a une forme produit. Plus généralement, considérons un réseau fermé,

- formé de M files d'attentes de type ./M/1, dont le vecteur des taux de service est $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_M)^T$,
- décrit par une matrice de routage R, dont chaque composante r_{ij} est la probabilité qu'un client sortant de la file i aille dans la file j,
- et dans lequel circulent N clients.

Les taux moyens des arrivées dans chaque file $1 \leq i \leq M$ sont tels que

$$\lambda_i = \sum_{j=1}^{M} r_{ji}\lambda_j,$$

et sont donc une solution non nulle de l'équation vectorielle

$$\boldsymbol{\lambda} = R^T\boldsymbol{\lambda}. \tag{8.38}$$

Cette équation admettant une infinité de solutions, les flux λ_i sont définis à une constante multiplicative λ_1 près : on peut donc écrire, si $\lambda_1 \neq 0$, que

$$\lambda_i = \alpha_i\lambda_1, \tag{8.39}$$

les α_i étant les solutions uniques du système d'équations

$$\alpha_1 = 1 \tag{8.40}$$

$$\alpha_i = r_{1i} + \sum_{j=2}^{M} r_{ji}\alpha_j \quad \text{pour } 2 \leq i \leq M. \tag{8.41}$$

Gordon et Newell ont alors étendu le théorème de Jackson aux réseaux fermés, en montrant que la probabilité que l'état du système soit $(n_1, n_2, \ldots, n_M)$ est, à l'équilibre,

$$\pi^\star_{n_1,n_2,\ldots,n_M} = \frac{1}{G(N,M)} \prod_{i=1}^{M} f(n_i) \tag{8.42}$$

où $f(n_i)$ est une fonction ne dépendant que de n_i (Par exemple, dans le cas de serveurs ./M/1, $f(n_i) = (\lambda_i/\mu_i)^{n_i}$ tandis que dans le cas de serveurs ./M/N/./N [1], $f(n_i) = (\lambda_i/\mu_i)^{n_i}/n_i!$) et $G(N, M)$ est une constante de normalisation telle que

$$G(N, M) = \sum_{n_1+...+n_M=N} \prod_{i=1}^{M} f(n_i). \tag{8.43}$$

Le calcul de cette constante peut être très long, car il faut évaluer l'expression $\prod_{i=1}^{M} f(n_i)$ pour toutes les combinaisons possibles de n_i telles que $\sum_{i=1}^{M} n_i = N$. Un algorithme récursif, l'algorithme de convolution de Buzen, permet de la calculer assez facilement pour des réseaux de taille raisonnable.

8.4.3 Analyse de la Valeur Moyenne

Il arrive souvent qu'on se contente, dans une analyse de performance, de connaître les valeurs moyennes des caractéristiques du réseau. Dans ce cas, une méthode très élégante permet de calculer ces quantités sans recourir à la constante de normalisation $G(N, M)$: c'est l'*analyse de la valeur moyenne* ("Mean Value Analysis" (MVA)), qui se base sur le théorème des arrivées :

Théorème 20 *Dans un réseau fermé de N clients à l'état stationnaire, le nombre de clients présents dans la file i au moment de l'arrivée d'un nouveau client dans cette file suit la même distribution que la solution stationnaire du réseau comportant $N-1$ clients au lieu de N.*

On déduit de ce théorème que

$$E[R_i(N)] = \frac{1}{\mu_i}\,[1 + E[N_i(N-1)]], \tag{8.44}$$

où $E[N_i(N-1)]$ est le nombre moyen de clients dans la file i du type ./M/1, lorsque $N-1$ clients circulent dans le réseau, et où $E[R_i(N)]$ est le temps de réponse moyen de la $i$ème file lorsque N clients circulent dans le réseau.

Si la file i est du type ./M/N/./N au lieu d'être du type ./M/1, l'équation (8.44) doit être remplacée par

$$E[R_i(N)] = \frac{1}{\mu_i}. \tag{8.45}$$

En faisant usage de la loi de Little, on trouve

$$N = \sum_{i=1}^{M} E[N_i(N)] = \sum_{i=1}^{M} \lambda_i(N) E[R_i(N)] = \lambda_1(N) \sum_{i=1}^{M} \alpha_i E[R_i(N)]$$

1. Comme la population des usagers est limitée à N, une file ./M/N/./N est équivalent à une file ./M/∞ puisqu'un maximum de N serveurs peuvent être utilisés simultanément.

et en combinant cette relation avec le système d'équations (8.39) on détermine les flux internes

$$\lambda_i(N) = \frac{\alpha_i N}{\sum_{j=1}^{M} \alpha_j E[R_j(N)]}. \tag{8.46}$$

En utilisant à nouveau la loi de Little, on trouve le nombre moyen de clients séjournant dans la file i,

$$E[N_i(N)] = \lambda_i(N) E[R_i(N)], \tag{8.47}$$

et on peut recommencer une nouvelle itération.

Par conséquent, on peut calculer récursivement les valeurs de $E[R_i(N)]$ et $E[N_i(N)]$ à partir des valeurs initiales obtenues lorsque $N = 0$: $E[N_i(0)] = 0$. L'algorithme est donc défini par les équations (8.44) ou (8.45), (8.46) et (8.47).

8.4.4 Application 3 : Modèle à serveur central d'un système à temps partagé

Le modèle à serveur central de l'ordinateur étudié à la section 8.3.4 est à présent inséré dans un réseau fermé comportant N terminaux, comme représenté à la figure 8.6. L'ordinateur fonctionne en mode interactif ; le nombre maximum de tâches qu'il doit traiter est N. Le groupe des N terminaux sont modélisés comme une file ./M/N/./N, avec un temps de service moyen qui est le temps moyen de réflexion des utilisateurs $1/\mu_T = 1$ sec.

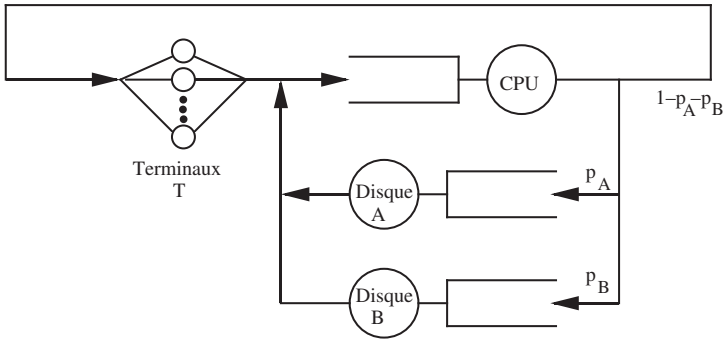


FIGURE 8.6 – Modèle à serveur central d'un système à temps partagé.

Les autres données restent inchangées par rapport au module précédent : le CPU et les deux disques sont des files ./M/1, dont les temps de service sont respectivement $1/\mu_{CPU} = 0.005$ sec, $1/\mu_A = 0.02$ sec et $1/\mu_B = 0.03$ sec. Après avoir reçu son service dans le CPU, une tâche a une probabilité $p_A = 0.4$ d'être envoyée vers le disque A et $p_B = 0.5$ vers le disque B.

La matrice de routage du réseau fermé est

$$R = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 - p_A - p_B & 0 & p_A & p_B \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix},$$

et on obtient les coefficients $\alpha_i = \lambda_i/\lambda_T$ suivants, en prenant

$$\alpha_T = 1,$$

et en écrivant (8.41) sous forme vectorielle,

$$\begin{bmatrix} \alpha_{CPU} \\ \alpha_A \\ \alpha_B \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 & 1 & 1 \\ p_A & 0 & 0 \\ p_B & 0 & 0 \end{bmatrix} \begin{bmatrix} \alpha_{CPU} \\ \alpha_A \\ \alpha_B \end{bmatrix}$$

qui a comme solution

$$\begin{bmatrix} \alpha_{CPU} \\ \alpha_A \\ \alpha_B \end{bmatrix} = \begin{bmatrix} 1 & -1 & -1 \\ -p_A & 1 & 0 \\ -p_B & 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 1/(1-p_A-p_B) \\ p_A/(1-p_A-p_B) \\ p_B/(1-p_A-p_B) \end{bmatrix} = \begin{bmatrix} 10 \\ 4 \\ 5 \end{bmatrix}.$$

On peut maintenant mettre en oeuvre l'algorithme MVA pour déterminer les performances moyennes de ce système en fonction du nombre d'utilisateurs. On prendra comme temps de réponse moyen du système le temps moyen que passe une tâche dans l'ordinateur (CPU + disques A et B), entre son départ et son retour au terminal. A partir de la formule de Little, on trouve

$$E[R] = \frac{1}{\lambda_T} \left(\lambda_{CPU} E[R_{CPU}] + \lambda_A E[R_A] + \lambda_B E[R_B]\right)$$

Remarquons que cette formule est la même que celle donnant le temps de réponse du réseau ouvert modélisant cet ordinateur au module précédent, mais où le flux d'arrivées externes Λ_s est remplacé par le flux d'arrivées provenant des terminaux, λ_T, qui n'est plus une donnée du problème.

L'algorithme est initialisé aux valeurs $E[N_T(0)] = E[N_{CPU}(0)] = E[N_A(0)] = E[N_B(0)] = 0$.

A la première itération ($N = 1$), on trouve

$$\begin{aligned}
E[R_T(1)] &= 1/\mu_T = 1 \text{ sec} \\
E[R_{CPU}(1)] &= (1 + E[N_{CPU}(0)])/\mu_{CPU} = 0.005 \text{ sec} \\
E[R_A(1)] &= (1 + E[N_A(0)])/\mu_A = 0.02 \text{ sec} \\
E[R_B(1)] &= (1 + E[N_B(0)])/\mu_B = 0.03 \text{ sec} \\
\lambda_T(1) &= \alpha_T \times 1/(\alpha_T E[r_T] + \alpha_{CPU} E[R_{CPU}] + \alpha_A E[R_A] + \alpha_B E[R_B]) = 0.781 \text{ tâches/sec} \\
\lambda_{CPU}(1) &= \alpha_{CPU} \lambda_T(1) = 7.813 \text{ tâches/sec} \\
\lambda_A(1) &= \alpha_A \lambda_T(1) = 3.125 \text{ tâches/sec} \\
\lambda_B(1) &= \alpha_B \lambda_T(1) = 3.906 \text{ tâches/sec} \\
E[N_T(1)] &= \lambda_T(1) E[R_T(1)] = 0.781 \text{ tâches} \\
E[N_{CPU}(1)] &= \lambda_{CPU}(1) E[R_{CPU}(1)] = 0.039 \text{ tâches} \\
E[N_A(1)] &= \lambda_A(1) E[R_A(1)] = 0.063 \text{ tâches} \\
E[N_B(1)] &= \lambda_B(1) E[R_B(1)] = 0.117 \text{ tâches} \\
E[R(1)] &= (\lambda_{CPU} E[R_{CPU}] + \lambda_A E[R_A] + \lambda_B E[R_B])/\lambda_T = 0.28 \text{ sec}
\end{aligned}$$

On peut alors commencer la seconde itération, et ainsi de suite. On trouve les résultats suivants :

N	2	3	4	5	10	20	30	50	100
$E[N_T(N)]$	1.533	2.251	2.928	3.560	5.821	6.661	6.667	6.667	6.667
$E[N_{CPU}(N)]$	0.080	0.122	0.164	0.207	0.398	0.499	0.500	0.500	0.500
$E[N_A(N)]$	0.130	0.204	0.282	0.365	0.800	1.135	1.143	1.143	1.143
$E[N_B(N)]$	0.257	0.424	0.626	0.868	2.981	11.705	21.690	41.690	91.690
$E[R(N)]$	0.305	0.333	0.366	0.405	0.716	2.003	3.500	6.500	14.000

Remarquons que lorsque N devient supérieur à 20, le nombre de tâches présentes en moyenne dans les terminaux, le CPU et le disque A a atteint une valeur constante quel que soit N, au contraire du disque B, qui est donc confirmé comme goulot d'étranglement du système.

Pour de grands réseaux, le calcul récursif des valeurs moyennes par l'algorithme MVA exact de Reiser peut devenir très long. On peut alors recourir à des méthodes approximatives, qui évitent un calcul récursif, et qui donnent une bonne estimation des temps de réponse des appareils. La *méthode MVA approximative de Schweitzer* est l'une des meilleures techniques.

8.4.5 Application 4 : Régulation de flux par fenêtre coulissante

Si la demande de trafic sur un réseau de communication augmente, celui-ci peut être soumis à des phénomènes de congestion, qui peuvent se traduire par un effondrement général de ses performances. Une méthode permettant de contrôler le trafic et d'éviter cette congestion (utilisée par exemple par le protocole TCP/IP) est la méthode dite *à fenêtre coulissante* ("Sliding window flow control") limitant le nombre de paquets pouvant circuler entre la source et la destination à un nombre N, qui est la taille de la fenêtre coulissante. Normalement, cette taille est adatpée par un mécanisme "Additive Increase, multiplicative decrease", ici nous la supposerons constante (ce n'est d'ailleurs pas la pire des hypothèses simplificatrices que nous allons faire). Chaque fois qu'un paquet émis par la source arrive à destination, celle-ci envoie un acquittement ("acknowledgment") à la source. Cette dernière ne peut émettre plus de N paquets consécutifs avant d'avoir reçu l'acquittement d'au moins le premier des N paquets qu'elle a envoyés. Une fois cet acquittement enregistré par la source, la fenêtre coulissante est décalée d'une unité, autorisant le départ d'un nouveau paquet si tous les autres $N-1$ paquets que peut contenir la fenêtre sont déjà sur le réseau.

Une connection est modélisée par une cascade de $M-1$ files ./M/1, de taux de service $\mu'_i C_i$ où $1/\mu'_i$ est la longueur moyenne des paquets sur la $i$ème ligne, de capacité C_i, que parcourent les paquets transitant sur le trajet entre source et destination, qu'on suppose identique pour tous les paquets. Les acquittements, qui sont des paquets de taille très petite, seront supposés être transmis sur le réseau avec la plus haute priorité. Ceci nous amène à faire l'hypothèse simplificatrice que les acquittements parviennent sans délai à la source dès qu'un paquet arrive à destination, et à modéliser le mécanisme de fenêtre coulissante, par le réseau fermé de la figure 8.7. La source et la destination sont reliées par une file artificielle numérotée M, dont le taux de service est λ_{source}, le taux auquel la source émettrait des paquets sans contrôle du flux dans la connection (qui serait alors modélisé par un réseau "cascade" ouvert). Si le nombre

de paquets en transit (c'est-à-dire dans les $M-1$ premières files) est égal à N, la taille de la fenêtre, la Mième file est vide, et aucun paquet ne peut être émis. Par contre, au moment où un paquet arrive à destination, un nouveau paquet le remplace dans la file M, et est émis en déans un temps distribué exponentiellement, avec une moyenne de $1/\lambda_{source}$ secondes.

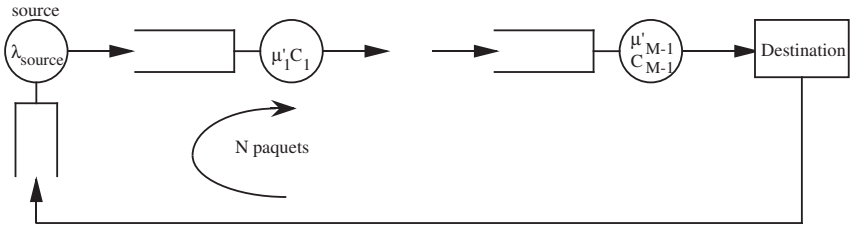


FIGURE 8.7 – Modèle de régulation de flux par fenêtre coulissante.

Le temps moyen de transit de paquet de bout en bout est simplement

$$E[R_{transit}] = \sum_{i=1}^{M-1} E[R_i] = E[R] - E[R_M] \tag{8.48}$$

et peut être calculé par l'algorithme MVA :

$$\begin{aligned}
E[R_i(N)] &= (1 + E[N_i(N-1)])/\mu'_i C_i \qquad \text{pour } 1 \leq i \leq M-1 \\
E[R_M(N)] &= (1 + E[N_M(N-1)])/\lambda_{source} \\
E[R(N)] &= \sum_{i=1}^{M} E[R_i(N)] \\
\lambda(N) &= N/E[R(N)] \\
E[N_i(N)] &= \lambda(N) E[R_i(N)] \qquad \text{pour } 1 \leq i \leq M.
\end{aligned}$$

On peut alors déterminer la taille optimale de la fenêtre N qui maximise la puissance du réseau $\lambda/E[R]$ (cfr exercice 20).

Il est important de noter que les hypothèses de tailles de paquets exponentielles pour des paquets IP, de même que l'hypothèse de temps d'une source émettant à un taux constant λ_{source} et suivant un processus de Poisson en boucle ouverte, est irréaliste pour un réseau TCP/IP. Une modélisation plus réaliste est nécssaire, c'est un sujet de recherches actuel.

8.5 Exercices

1. Un concentrateur a un tampon pouvant contenir 5 paquets, et est modélisé par une file M/M/1/5. La ligne de sortie a une capacité de 2400 bps, et la taille moyenne des paquets est de 1000 bits. Un paquet arrive au concentrateur toutes les secondes.

   (a) Quelle est la probabilité qu'un paquet soit rejeté ?

(b) Quel est le nombre moyen de paquets présents dans le concentrateur ?

(c) Quel est le temps de réponse moyen du système ? (Attention à l'application de la formule de Little !)

2. A présent, on suppose que le concentrateur a une mémoire de taille infinie, mais que tous les paquets ont la même longueur de 1000 bits. Les arrivées sont toujours poissonniennes, avec une moyenne de 1 sec^{-1}.

   (a) Quel est le nombre moyen de paquets présents dans le concentrateur ?

   (b) Quel est le temps de réponse moyen du système ?

3. Calculer la valeur de $E[N]$ pour une file M/M/1 en partant directement des équations de balance, sans les résoudre mais en utilisant la fonction génératrice de probabilités.

4. On considère une file M/M/1 à l'état stationnaire. Calculer

   (a) la loi de probabilité de la v.a. Q comptant le nombre de clients dans le buffer (Hint : $Q = 0$ si $N = 0$ et $Q = N - 1$ si $N \geq 1$).

   (b) le nombre moyen de clients en attente, $E[Q]$.

   (c) la loi de probabilité de la v.a. W mesurant le temps d'attente d'un client (Hint : W est une v.a. mixte).

   (d) le temps d'attente moyen $E[W]$.

   Vous pouvez retrouver le résultat des questions b et d de plusieurs manières différentes.

5. Deux lignes de capacité C chacune relient deux noeuds d'un réseau de communication. Les paquets de données, dont la longueur suit une loi exponentielle de moyenne μ', empruntent l'une de ces deux lignes aléatoirement. Ce système a-t-il un temps de réponse moyen plus élevé, plus faible ou égal à celui d'un système n'utilisant qu'une seule ligne, mais de capacité $2C$? On suppose les arrivées de paquets poissonniennes.

6. Des clients arrivant à un restaurant "fast-food" à un taux moyen de 5 par minute et attendent en moyenne 5 minutes pour recevoir leur commande. Ils mangent dans le restaurant avec une probabilité 0.5, et quittent le restaurant avec leur repas avec une probabilité 0.5. La durée moyenne d'un repas est 15 minutes. Quelle est le nombre moyen de clients dans le restaurant ?

7. Considérons la file M/G/1.

   (a) Calculer $E[A^2(k)]$.

   (b) En élevant les deux membres de (8.16) au carré et en prenant leur espérance, retrouver la formule de Pollaczek-Khintchine (8.21).

8. (a) Montrer que la fonction génératrice d'un système M/D/1 à l'état stationnaire est

$$G_N(z) = \frac{(1-\rho)(1-z)}{1 - ze^{\rho(1-z)}}.$$

(b) Que vaut la probabilité $\pi_0^\star$ que le système soit vide ?

(c) Que vaut la probabilité $\pi_1^\star$ qu'il ait un (seul) "client" dans le système ?

9. Des paquets arrivent suivant un processus de Poisson, à un taux moyen de 200 paquets par seconde, à une station qui les stocke dans une mémoire-tampon (suivant l'ordre FIFO) avant de les transmettre sur un canal bruité, en utilisant un protocole "Stop and wait" (ou Alternating Bit Protocol), dans lequel chaque paquet transmis doit être acquitté avant que le suivant puisse être transmis. Autrement, en l'absence d'acquittement, le même paquet est retransmis au bout d'un temps mort (Time out), jusqu'à ce que la transmission soit fructueuse et qu'un acquittement soit reçu. Ce protocole est la forme la plus élémentaire d'un protocole ARQ (Automatic Repeat reQuest).

   Le temps de transmission (y compris le temps de propagation) d'un paquet est d'une milliseconde, la probabilité qu'il parvienne correctement à destination est 0.9. Le temps de transmission des acquittements (acknowledgments) est négligé, de plus on suppose que ceux-ci sont toujours transmis sans erreur. La valeur du time-out est également fixée à 1 msec (normalement, ce serait un peu plus). Par conséquent, 1 msec après avoir émis un paquet, la station envoie soit le paquet suivant si elle vient de recevoir l'acquittement précédent, soit à nouveau le même paquet dans le cas contraire. Quel est le temps moyen qui s'écoule entre l'arrivée d'un paquet dans la file et le moment où il arrive sans erreurs à destination ?

10. On considère un système de télṕhonie mobile micro-cellulaire le long d'une autoroute, et on cherche à déterminer le taux des appels issus dans ce tronçon (“Originating Calls (OC)") λ_{OC} (appels/minute), et le taux des appels transférés d'autres cellules (“Handoff Calls (HC)"), λ_{HC} (appels/minute). On suppose ici que tous les appels sont acceptés (aucun appel bloqué), contrairememt à l'exercice 10 du module précédent (en fait cet exercice pourrait servir de préparation à l'exercice 10 du module précédent).

    Les cellules sont ici rectangulaires, et couvrent un tronçon de longueur L de l'autoroute, comme celui représenté à la figure 8.8. La longueur totale de l'autoroute supposée infinie (pour simplifier), et on est en régime stationnaire.

    Le nombre d'appels $A(t, x)$ générés pendant un intervalle de temps de longueur t et sur une portion d'autoroute de longueur x est un processus de Poisson sur $\mathbb{R}^+ \times \mathbb{R}^+$, de taux λ appels par minute et km d'autoroute. Le processus $A(t, x)$ est plus précisémment défini comme suit : (i) le nombre d'appels générés pendant un intervalle de temps de longueur t et sur une portion d'autoroute de longueur x suit une loi de Poisson de moyenne λxt, et (ii) les nombres d'appels générés dans des intervalles de temps disjoints sont indépendants, de même que les nombres d'appels générés sur des portions d'autoroute ne se recouvrant pas.

    (a) Quelle est la probabilité que n appels aient été générés sur le tronçon de longueur L de la microcellule considérée pendant un intervalle de temps de longueur t ? Déduisez-en le taux λ_{OC}.

    (b) La durée des appels forme une suite de v.a. i.i.d, dont la moyenne est $1/\mu$ minutes, et tous les véhicules roulent à la même vitesse v (en km/minute). Soit $N_{HC}(t)$ le nombre

de “Handoff Calls" arrivant pendant un intervalle de temps de longueur t. Calculer $P(N_{HC}(t) = n)$ et déduisez-en le taux λ_{HC}. Hint : transformez les temps d'appels en longueurs de trajet parcouru pendant la durée d'un appel. Soit $1/\mu'$ l'espérance de ces v.a.. $N_{HC}(t)$ est alors est le nombre de clients “en service” dans un système M/GI/∞, soumis à un taux de λt appels par km et dont le taux de service par serveur est de μ' appels par km.

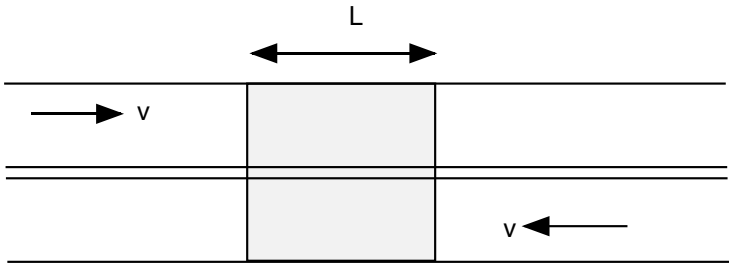


FIGURE 8.8 – Microcellule le long d'une autoraoute.

11. On considère un système $M/D/\infty$, avec un taux d'arrivée λ et une durée constante de service $1/\mu$ à l'état stationnaire, et on considère ce système pour tout $t \in \mathbb{R}$ (on suppose que le système est en place depuis toujours). Soit $N(t)$ le nombre de clients dans ce système au temps t. Calculez la moyenne et la fonction d'auto-corrélation de $N(t)$. Hint : quel lien y a-t-il avec le Poisson shot noise ?

12. On considère un système identique à une file M/M/1 à l'état stationnaire, à la différence près que quand le système se vide, le service ne reprend que lorsqu'il y a J clients présents dans le système (Une file M/M/1 est donc le cas particulier $J = 1$). Une fois que le service a repris, il procède normalement jusqu'à ce que le système devienne à nouveau vide.

    (a) Dessiner le graphe des transitions entre états de ce système. Hint : il est utile de considérer chaque état comme ayant deux variables : la première n est, comme d'habitude, le nombre de clients dans le système, la seconde indique lequel des deux états 0 ou J a été le plus récemment atteint. On a donc, suivant la valeur de n, les états possibles suivants :

$$\begin{aligned} n = 0 &\rightarrow \text{état} = (n, 0) = (0, 0) \\ 1 \leq n \leq J-1 &\rightarrow \text{état} = (n, 0) \text{ ou } (n, J) \\ n \geq J &\rightarrow \text{état} = (n, J) \end{aligned}$$

    En effet, si $n \geq J$, le dernier état atteint devait être J, tandis que si $1 \leq n \leq J-1$, le dernier état atteint pouvait être soit 0, si le système était dernièrement vide (et donc le service n'a pas encore repris) ou J, dans le cas contraire (et donc un client est en train d'être servi).

    (b) Ecrire un ensemble d'équations de balance (les équations de balance globales, ou tout autre ensemble d'équations équivalentes, mais plus simples).

(c) Résoudre les équations de balance (i.e., calculer les probabilités d'avoir n clients dans le système).

(d) Calculer le nombre moyen de clients dans le système. Montrer que ce nombre est égal au nombre moyen de clients dans une file M/M/1 augmenté de $(J-1)/2$.

(e) Calculer le temps moyen de séjour d'un client dans le système.

13. Considérons le système de la figure 8.9(a), pour laquelle $\Lambda_s/\mu = 0.25$ (le serveur est exponentiel, les arrivées poissonniennes).

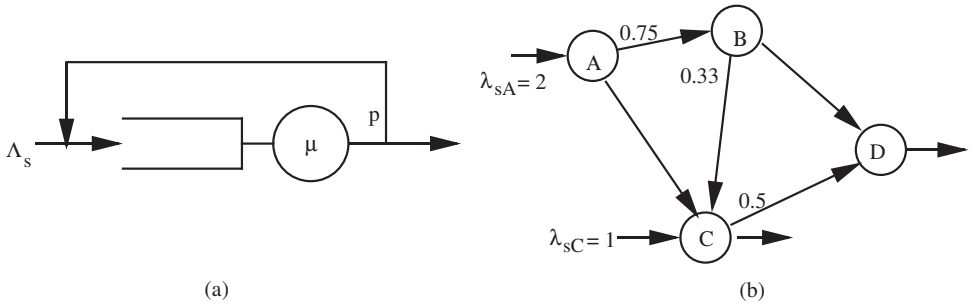


FIGURE 8.9 – File unique avec rétroaction (a) et réseau de communication (b).

(a) Quelles sont les valeurs de p, $0 \leq p < 1$, pour lesquelles ce système est stable ?

(b) Quelle est la probabilité d'avoir 10 clients ou plus dans le système, si $p = 0.5$?

14. Quel est le temps de transmission moyen d'un paquet transmis sur le réseau représenté à la figure 8.9(b) ? Les paquets ont une longueur (distribuée exponentiellement) moyenne de 1000 bits sur tout le réseau, et les capacités des lignes sont $C_{AB} = 9.6$ kbps, $C_{AC} = C_{BC} = C_{BD} = 2.4$ kbps, et $C_{CD} = 4.8$ kbps.

15. Résoudre le probème dual de l'optimisation d'un réseau de transit : calculer les capacités minimisant le temps de réponse moyen du réseau tout en donnant un coût total inférieur à une valeur maximale $W_{\max}$. Les coûts sont des fonctions linéaires des capacités : $W = \sum_{i=1}^{M}(d_i C_i + f_i)$. Ce problème peut-il ne pas admettre de solution ?

16. On considère trois files ./M/1 représentées à la figure 8.10, ayant chacune un taux moyen de service égal à μ. Le taux moyen d'arrivée externe est noté λ, les probabilités de routage sont données à la figure 8.10. Que vaut la probabilité qu'il y ait exactement un client dans chacune des trois files ?

17. Les temps de service des serveurs du réseau fermé de la figure 8.5(a) sont $1/\mu_1 = 1$ sec et $1/\mu_2 = 2$ sec, tandis que le nombre de clients circulant dans le système complet est $N = 5$. Calculer le nombre moyen de clients occupant chaque file,

(a) en utilisant la formule classique $E[N_1] = \sum_{n_1=1}^{5} n_1 \pi^{\star}_{n_1, N-n_1}$ ou à partir de la fonction génératrice,

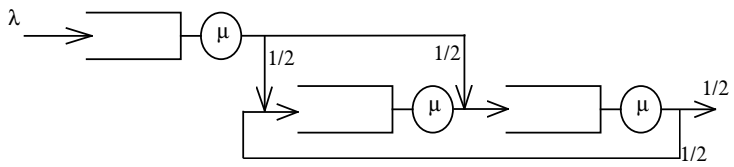


FIGURE 8.10 – Trois files ./M/1.

(b) en utilisant l'algorithme MVA.

18. On considère à nouveau le réseau de la figure 8.5, où cette fois $\mu_1 = \mu_2$.

    (a) Que vaut $\pi^{\star}_{n_1,n_2}$?

    (b) Quel est le nombre moyen de clients résidant dans chaque file ?

19. Le disque B de l'ordinateur de l'application 3 est supprimé, de sorte que toutes les opérations E/S sont à présent dirigées vers le disque A. Que deviennent le temps de réponse et les nombres moyens de tâches à chaque appareil pour $N = 5$, 10 et 20 utilisateurs ?

20. Si toutes les lignes ont la même capacité C et les messages la même longueur moyenne μ' tout le long du circuit virtuel de la figure 8.7, tandis que $\lambda_{source} = \mu' C$, quelle est la taille de la fenêtre N qui maximise la puissance du réseau de la figure 8.7 ?

21. Un ordinateur, modélisé par une file ./M/1, est relié à N terminaux, modélisés par une file ./M/N/./N ("delay center"). Ce système, qui fonctionne en mode interactif, est représenté la figure 8.11. Le temps de service moyen de l'ordinateur est $1/\mu_{\text{ord}} = 0.1$ seconde tandis que le temps de réflexion moyen des utilisateurs est $1/\mu_T = 1$ seconde.

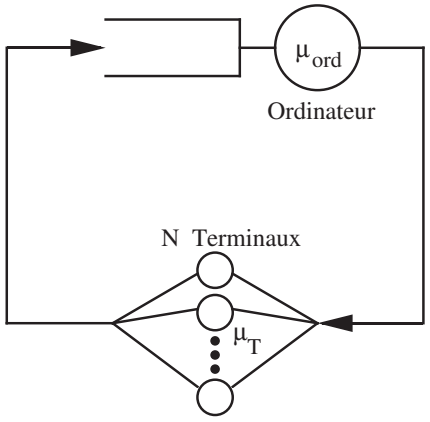


FIGURE 8.11 – Ordinateur relié à N terminaux.

(a) S'il y a $N = 2$ utilisateurs, quelle est la probabilité qu'aucune des deux tâches ("jobs") ne soit présente dans l'ordinateur, lorsque le système est à l'état stationnaire ?

(b) Supposons à présent qu'il y ait $N = 14$ utilisateurs. On a calculé que le nombre moyen de tâches présentes dans l'ordinateur est alors égal à 4.568. Que devient ce nombre lorsqu'il y a 15 utilisateurs au lieu de 14 ?