Cahier de Mathématiques Appliquées n^o 14

Files d'attente

B. Ycart

La théorie des files d'attente a de nombreuses applications, en particulier dans les réseaux de communication et les réseaux informatiques. Nous insisterons surtout sur les modèles markoviens, en supposant acquises les notions de base sur les chaînes de Markov et les processus markoviens de saut, qui ont fait l'objet de cahiers dans la même collection. L'objectif est de donner une compréhension concrète des phénomènes, tout en présentant les résultats mathématiques de base de la théorie, sans insister sur les détails techniques. Pour compléter ce qui suit, on pourra se reporter aux ouvrages suivants.

E. Gelenbe et G. Pujolle Introduction aux réseaux de files d'attente.
Eyrolles, Paris, 1985.

L. Kleinrock Queuing systems, vol. 1: theory.
Wiley, New York, 1975.

L. Kleinrock Queuing systems, vol. 2: computer applications.
Wiley, New York, 1976.

P. Robert Réseaux et files d'attente : méthodes probabilistes.
Springer-Verlag, Berlin, 2000.

Ce "cahier de mathématiques appliquées" doit beaucoup aux relectures scrupuleuses de Jean-Stéphane Dhersin et Dominique Seret, au dynamisme de Sylvie Sevestre-Ghalila, au soutien de l'Ecole Supérieure de la Statistique et de l'Analyse de l'Information de Tunisie, par son directeur Makki Ksouri, ainsi qu'à la compétence de Habib Bouchriha, directeur du Centre des Publications Universitaires de la Tunisie.

Table des matières

1 Processus de naissance et de mort **71**
1.1 Définitions et exemples 71
1.2 Comportement asymptotique 74
1.3 Equations de Chapman-Kolmogorov 77

2 Files markoviennes **80**
2.1 Modèles d'attente 80
2.2 File M/M/1 83
2.3 File M/M/s 87
2.4 Files à capacité limitée 90
2.5 Files à arrivées ou services groupés 92

3 Files quasi-markoviennes **97**
3.1 File M/GI/1 97
3.2 File GI/M/1 102

4 Réseaux de files d'attente **105**
4.1 Réseaux de Jackson ouverts 105
4.2 Réseaux de Jackson fermés 110
4.3 Réseaux de Petri 113

5 Exercices **117**

1 Processus de naissance et de mort

1.1 Définitions et exemples

Un processus est une collection de variables aléatoires $\{Z_t\,,\, t \geq 0\}$, indicée par le temps. Ici, il nous servira à décrire l'évolution aléatoire au cours du temps d'un nombre d'individus, dans une population ou un système d'attente. Les variables aléatoires Z_t prennent donc leurs valeurs dans l'ensemble des entiers $\mathbb{N}$. Le processus évolue comme un processus markovien de saut : le nombre d'individus reste constant pendant une certaine durée exponentielle, puis il saute vers une autre valeur. S'agissant d'une population ou d'une file d'attente, nous ne considérerons que des sauts vers les deux valeurs voisines : la taille de la population peut soit augmenter de 1 (naissance ou arrivée) soit diminuer de 1 (mort ou départ). L'intensité de ces sauts est gouvernée par deux suites de réels positifs, $(\lambda_n)_{n\in\mathbb{N}}$ (taux de naissance) et $(\mu_n)_{n\in\mathbb{N}^*}$ (taux de mort). Pour éviter les cas particuliers, nous supposerons dans un premier temps que ces taux sont tous strictement positifs.

Définition 1.1 *Soient $(\lambda_n)_{n\in\mathbb{N}}$ et $(\mu_n)_{n\in\mathbb{N}^*}$ deux suites de réels strictement positifs. Un* processus de naissance et de mort *de taux de naissance (λ_n) et taux de mort (μ_n) est un processus markovien de saut $\{Z_t\,,\, t \geq 0\}$ à valeurs dans $\mathbb{N}$ tel que pour tout $n \geq 0$ le taux de transition de n vers $n+1$ est λ_n, et pour tout $n \geq 1$, le taux de transition de n vers $n-1$ est μ_n. Aucune autre transition n'est possible.*

Le diagramme de transition de la figure 1 illustre cette définition.

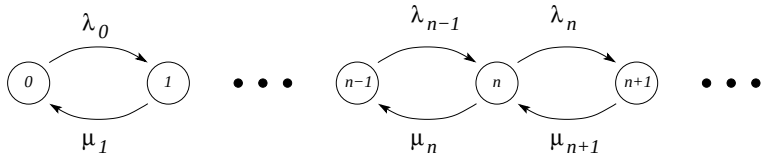

FIGURE 1 – Diagramme de transition d'un processus de naissance et de mort.

Pour comprendre la dynamique d'un processus de naissance et de mort, on peut revenir à la construction d'un processus de saut par sa chaîne incluse. Ici c'est une chaîne de Markov à valeurs dans $\mathbb{N}$, qui saute de n vers $n+1$ avec probabilité $\frac{\lambda_n}{\lambda_n+\mu_n}$ et de n vers $n-1$ avec probabilité $\frac{\mu_n}{\lambda_n+\mu_n}$. Le passage de la chaîne au processus se fait en rajoutant des temps de séjour aléatoires, indépendants entre eux et indépendants de la chaîne, dont la loi dépend de l'état courant : le temps de séjour dans l'état n suit la loi exponentielle $\mathcal{E}(\lambda_n + \mu_n)$. En d'autres termes, on peut simuler un processus de naissance et mort par l'algorithme suivant.

$t \longleftarrow 0$
Initialiser Z dans $\mathbb{N}$

Répéter
  $n \longleftarrow Z$ *(état présent)*
  Si $n = 0$ alors $Z \longleftarrow 1$; $t \longleftarrow t - \log(\text{Random})/\lambda_0$
    sinon
      Si Random $< \frac{\lambda_n}{\lambda_n+\mu_n}$
        alors $Z \longleftarrow n+1$
        sinon $Z \longleftarrow n-1$
      finSi
      $t \longleftarrow t - \log(\text{Random})/(\lambda_n + \mu_n)$
  finSi
Jusqu'à (arrêt de la simulation)

La famille des lois exponentielles possède une propriété de stabilité qui sera souvent invoquée dans ce qui suit.

Proposition 1.2 *Considérons k variables aléatoires indépendantes $X_1, \ldots, X_k$, de lois respectives $\mathcal{E}(\lambda_1), \ldots, \mathcal{E}(\lambda_k)$. Posons :*

$$Y \;=\; \min\{X_1, \ldots, X_k\}\,.$$

Alors Y suit la loi $\mathcal{E}(\lambda_1 + \cdots + \lambda_k)$ et pour tout $i = 1, \ldots, n$:

$$I\!P[Y = X_i] \;=\; \frac{\lambda_i}{\lambda_1 + \cdots + \lambda_k}\,.$$

Considérons alors un processus de naissance et de mort arrivant dans un état $n > 0$. Nous pouvons envisager deux variables aléatoires indépendantes A et D, de lois $\mathcal{E}(\lambda_n)$ et $\mathcal{E}(\mu_n)$, que nous interpréterons respectivement comme le temps d'attente avant une prochaine naissance (arrivée) et le temps d'attente avant une prochaine mort (départ). Si $A < D$, alors le prochain événement sera une naissance, et le processus passera de n à $n+1$. Si $A > D$, le prochain événement sera une mort et le processus passera de n à $n-1$. La proposition 1.2 montre que le plus petit des deux temps d'attente suit la loi $\mathcal{E}(\lambda_n + \mu_n)$, et que le prochain saut ira de n à $n+1$ avec probabilité $\frac{\lambda_n}{\lambda_n+\mu_n}$. En d'autres termes, on peut remplacer l'algorithme de simulation par la chaîne incluse par le suivant, qui est strictement équivalent du point de vue probabiliste.

$t \longleftarrow 0$
Initialiser Z dans $I\!N$
Répéter
  $n \longleftarrow Z$ *(état présent)*
  Si $n = 0$ alors $Z \longleftarrow 1$; $t \longleftarrow -\log(\text{Random})/\lambda_0$
    sinon
      $A \longleftarrow -\log(\text{Random})/\lambda_n$
      $D \longleftarrow -\log(\text{Random})/\mu_n$
      Si $A < D$

alors $Z \longleftarrow n+1$; $t \longleftarrow t+A$
sinon $Z \longleftarrow n-1$; $t \longleftarrow t+D$
finSi
finSi
Jusqu'à (arrêt de la simulation)

Une autre propriété des lois exponentielles joue un rôle fondamental dans la modélisation par les processus de naissance et de mort, la propriété d'*absence de mémoire*.

Proposition 1.3 *Une variable aléatoire X à valeurs dans $\mathbb{R}^+$, de fonction de répartition continue suit une loi exponentielle si et seulement si pour tout $t, h \geq 0$,*

$$\mathbb{P}[X > t+h \,|\, X > t] \;=\; \mathbb{P}[X > h] \;.$$

Son interprétation est la suivante. Si l'attente d'un événement a débuté dans le passé, et que celui-ci ne s'est pas encore produit à l'instant t, alors la durée qui reste à attendre après t suit la même loi que la durée totale, si celle-ci est exponentielle. Nous retrouverons cet argument plusieurs fois dans les exemples qui suivent.

Exemple 1 : *Processus de naissance pure*

L'hypothèse que les taux de naissance et les taux de mort sont tous strictement positifs assure que le processus est irréductible : la probabilité de passer d'un état n à un état m entre deux instants distincts est toujours strictement positive. Si un taux de naissance est nul, le processus ne peut pas dépasser l'état correspondant. Ce sera le cas pour des files d'attente à capacité limitée. Si un taux de mort est nul, le processus ne peut pas revenir en arrière. Dans le cas particulier où les taux de mort sont tous nuls, on parle de "processus de naissance pure". Un tel processus reste dans l'état n un temps exponentiel de paramètre λ_n, puis passe à l'état $n+1$. Si de plus tous les taux de naissance sont égaux à λ, c'est un processus de Poisson d'intensité λ.

Exemple 2 : *La file* M/M/1

La notation M/M/1 sera justifiée plus loin. Il s'agit du modèle le plus simple de file d'attente. On suppose que des clients arrivent dans une file à un seul serveur, et reçoivent chacun leur tour un service d'une certaine durée. Si un client arrivant trouve le serveur libre, il passe immédiatement au service. Si le serveur est occupé, il attend son tour. Les hypothèses probabilistes sont les suivantes.

- Les clients arrivent un par un selon un processus de Poisson d'intensité λ : le temps séparant deux arrivées successives suit la loi $\mathcal{E}(\lambda)$.
- Le temps de service de chaque client suit la loi $\mathcal{E}(\mu)$.
- Les temps séparant deux arrivées successives et les temps de service sont des variables aléatoires indépendantes dans leur ensemble.

Nous examinons le nombre Z_t de clients dans le système (qu'ils soient en train d'attendre ou d'être servis). Supposons qu'il y en ait $n > 0$ à l'instant t. Par la propriété d'absence de mémoire, le temps qui sépare t de la prochaine arrivée suit la loi $\mathcal{E}(\lambda)$, et le temps résiduel du client en train d'être servi suit la loi $\mathcal{E}(\mu)$. Par la proposition 1.2, le prochain événement modifiant la composition de la file surviendra au bout d'un temps exponentiel de paramètre $\lambda + \mu$. Ce sera une arrivée avec probabilité $\frac{\lambda}{\lambda+\mu}$, ou un départ avec probabilité $\frac{\mu}{\lambda+\mu}$. En d'autres termes, $\{Z_t\,,\, t \geq 0\}$ est un processus de naissance et de mort de taux de naissance constants $\lambda_n = \lambda\ \forall n \geq 0$, et taux de mort également constants $\mu_n = \mu\ \forall n \geq 1$.

Exemple 3 : *Croissance de population*

Considérons une population de bactéries sur laquelle on fait les hypothèses suivantes.

- La durée de vie de chaque bactérie est exponentielle de paramètre μ.
- Chaque bactérie vivante donne naissance à une autre au bout d'un temps exponentiel de paramètre λ.
- Des bactéries arrivent de l'extérieur selon un processus de Poisson d'intensité ν.
- Toutes les variables aléatoires considérées sont indépendantes.

Nous examinons le nombre Z_t de bactéries vivantes à l'instant t. Supposons qu'il y en ait n. Chacune d'entre elles a deux horloges : sa durée de vie et son temps de reproduction. Par la propriété sans mémoire, la durée de vie au-delà de t suit la loi $\mathcal{E}(\lambda)$ et le temps au bout duquel elle donnera naissance à une autre (si elle n'est pas morte) suit la loi $\mathcal{E}(\mu)$. Le temps qui sépare t de la prochaine immigration suit la loi $\mathcal{E}(\nu)$. Par la proposition 1.2, le prochain événement modifiant la population surviendra au bout d'un temps exponentiel de paramètre $n\lambda + \nu + n\mu$. Cet événement sera une naissance ou une arrivée avec probabilité $\frac{n\lambda+\nu}{n\lambda+\nu+n\mu}$, ce sera une mort avec probabilité $\frac{n\mu}{n\lambda+\nu+n\mu}$. L'évolution du nombre de bactéries peut être décrite par un processus de naissance et de mort de taux de naissance $\lambda_n = n\lambda + \nu$, et de taux de mort $\mu_n = n\mu$.

1.2 Comportement asymptotique

Etudier le comportement asymptotique des processus de naissance et de mort en toute généralité est assez compliqué. Nous nous contenterons de dégager quelques idées générales, les plus importantes pour les applications.

Nous commençons par écarter un cas pathologique que l'on ne rencontre pas en pratique, celui de l'*explosion à temps fini*. Pour comprendre ce phénomène, considérons un processus de naissance pure, dont les taux de naissance λ_n sont tous strictement positifs. On suppose qu'il part de 0 à l'instant 0. On peut construire ce processus à partir d'une suite de variables aléatoires indépendantes (X_n), où pour tout $n \geq 0$, X_n suit la loi exponentielle $\mathcal{E}(\lambda_n)$. La variable X_n est le temps de séjour dans l'état n, à la suite duquel le processus saute vers l'état $n + 1$. Le temps que le processus met à parcourir les

états $0, \ldots, n$ est donc $X_0 + \cdots + X_n$. La somme de la série $\sum X_i$ est le temps total que le processus passe dans $\mathbb{N}$. En tant que série de variables indépendantes, $\sum X_i$ converge ou diverge presque sûrement (conséquence de la loi du zéro-un de Kolmogorov). Si elle diverge, le processus reste un temps infini dans $\mathbb{N}$, et la valeur de Z_t reste finie pour tout t. Mais si la série converge, le temps de séjour dans $\mathbb{N}$ est presque sûrement fini, et le processus $\{Z_t\,,\, t \geq 0\}$ n'est défini que pour $t \leq \sum X_i$. L'espérance de X_n est $\frac{1}{\lambda_n}$. Si la série $\sum \frac{1}{\lambda_n}$ converge, alors le temps moyen de séjour du processus dans $\mathbb{N}$ est fini. Il se trouve que c'est aussi la condition nécessaire et suffisante de convergence pour la série $\sum X_i$: la proposition suivante se déduit du théorème de convergence des séries aléatoires, conséquence de la théorie des martingales.

Proposition 1.4 *Soit $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires indépendantes, telles que pour tout $n \geq 0$, X_n suit la loi $\mathcal{E}(\lambda_n)$. La série $\sum X_n$ converge presque sûrement si et seulement si $\sum \frac{1}{\lambda_n} < +\infty$.*

La figure 2 illustre une situation d'explosion à temps fini, pour un processus de naissance pure, de taux de naissance $\lambda_n = \frac{1}{(n+1)^2}$.

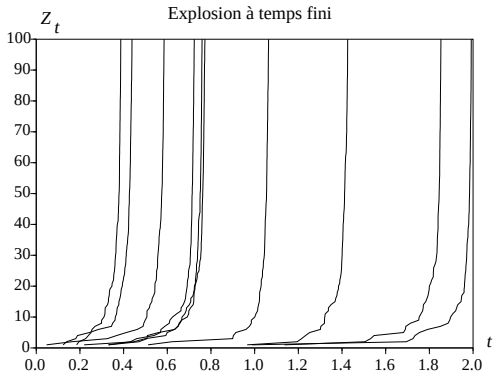

FIGURE 2 – Trajectoires indépendantes d'un processus de naissance pure, de taux de naissance $\frac{1}{(n+1)^2}$: les trajectoires explosent à temps fini.

L'explosion à temps fini peut se produire pour un processus de naissance et de mort dans le cas général, mais il n'est pas aussi facile de donner une condition nécessaire et suffisante. Nous nous contenterons d'admettre que si la série $\sum \frac{1}{\lambda_n}$ diverge, alors le temps total passé dans $\mathbb{N}$ est nécessairement infini. Ceci fournit une condition suffisante d'existence dans le cas général.

Proposition 1.5 *Soit $\{Z_t\,,\, t \geq 0\}$ un processus de naissance et de mort de taux de naissance $\lambda_n > 0$. Si la série $\sum \frac{1}{\lambda_n}$ diverge, alors le processus est défini pour tout $t \geq 0$.*

La condition $\sum \frac{1}{\lambda_n} = +\infty$ est satisfaite dans tous les modèles d'intérêt pratique. Nous la supposons vérifiée désormais.

Une fois le processus $\{Z_t\}$ défini pour tout t, se pose la question de son comportement quand t tend vers l'infini. Pour un processus irréductible, comme pour une chaîne irréductible apériodique, trois cas sont possibles en fonction du retour dans un état donné.

- Si le processus, partant d'un état donné, a une probabilité non nulle de ne jamais y retourner, il est dit *transient*. En pratique cela signifie que le nombre que l'on étudie (taille de population ou nombre de clients) tend vers l'infini. Nous interprétons cette situation comme une *saturation* du système.
- Si le processus, partant d'un état donné, y revient nécessairement au bout d'un temps fini en moyenne, il est dit *récurrent positif*. Dans ce cas, il converge en loi vers sa *mesure stationnaire*, interprétée comme une situation d'équilibre du système étudié.
- Si le processus, partant d'un état donné, y revient avec probabilité 1 mais que l'espérance du temps de retour est infinie, il est dit *récurrent nul*. Dans les applications, ce cas, frontière entre l'équilibre et la saturation, ne sera pas distingué du cas transient.

Nous allons montrer que le comportement asymptotique d'un processus de naissance et de mort dépend de rapports des taux de naissance aux taux de mort. Pour cela, nous commençons par examiner la probabilité de retour en 0. Un processus de naissance et de mort partant de 0 à l'instant 0 atteint l'état 1 à son premier saut. La question est de savoir s'il peut revenir de 1 vers 0. Pour tout $n \geq 0$, notons p_n la probabilité de retour en 0 à partir de l'état n. Pour tout $n > 0$, la dynamique de la chaîne incluse permet d'écrire l'équation de récurrence suivante.

$$p_n = \frac{\lambda_n}{\lambda_n + \mu_n} p_{n+1} + \frac{\mu_n}{\lambda_n + \mu_n} p_{n-1} \ .$$

On en déduit l'expression de p_n en fonction de p_1.

$$p_n = 1 - (1 - p_1) \left(1 + \frac{\mu_1}{\lambda_1} + \cdots + \frac{\mu_1 \dots \mu_{n-1}}{\lambda_1 \dots \lambda_{n-1}} \right) \ .$$

Si la série de terme général $\frac{\mu_1 \dots \mu_{n-1}}{\lambda_1 \dots \lambda_{n-1}}$ diverge, alors nécessairement $p_1 = 1$, le retour en 0 est certain, et le processus est donc récurrent.

Examinons maintenant le temps moyen de passage de n à $n-1$, que nous notons e_n. L'équation de récurrence est la suivante :

$$e_n = \frac{1}{\lambda_n + \mu_n} + \frac{\lambda_n}{\lambda_n + \mu_n} (e_{n+1} + e_n) \ .$$

On en déduit la relation suivante entre e_{n+1} et e_1.

$$e_{n+1} \frac{\lambda_1 \dots \lambda_n}{\mu_1 \dots \mu_n} = e_1 - \frac{1}{\lambda_0} \left(\frac{\lambda_0}{\mu_1} + \cdots + \frac{\lambda_0 \dots \lambda_{n-1}}{\mu_1 \dots \mu_n} \right) \ .$$

Si la série de terme général $\frac{\lambda_0 \ldots \lambda_{n-1}}{\mu_1 \ldots \mu_n}$ diverge, alors nécessairement e_1 doit être infini : le processus est soit récurrent nul, soit transient. Nous admettrons réciproquement que si la série de terme général $\frac{\lambda_0 \ldots \lambda_{n-1}}{\mu_1 \ldots \mu_n}$ converge, alors le processus est récurrent positif. Nous retrouverons cette série dans le calcul de la mesure stationnaire au paragraphe suivant.

Nous résumons les conditions obtenues sur le cas particulier de la file M/M/1.

Proposition 1.6 *Soit Z_t le nombre de clients à l'instant t dans une file* M/M/1 *pour laquelle les temps séparant deux arrivées sont exponentiels de paramètre λ et les temps de service sont exponentiels de paramètre μ. Le processus $\{Z_t\,,\, t \geq 0\}$ est :*

- *transient* *si* $\lambda > \mu$
- *récurrent positif* *si* $\lambda < \mu$
- *récurrent nul* *si* $\lambda = \mu$.

Démonstration : Pour une file M/M/1, $\{Z_t\,,\, t \geq 0\}$ est un processus de naissance et de mort, pour lequel les taux de naissance et les taux de mort sont constants : $\lambda_n = \lambda\ \forall n \geq 0$ et $\mu_n = \mu\ \forall n \geq 1$. Les rapports $\frac{\mu_1 \ldots \mu_n}{\lambda_1 \ldots \lambda_n}$ et $\frac{\lambda_0 \ldots \lambda_{n-1}}{\mu_1 \ldots \mu_n}$ valent respectivement $(\frac{\mu}{\lambda})^n$ et $(\frac{\lambda}{\mu})^n$. Si $\lambda > \mu$, la série $\sum(\frac{\mu}{\lambda})^n$ converge, la série $\sum(\frac{\lambda}{\mu})^n$ diverge. Si $\lambda < \mu$ c'est le contraire. Si $\lambda = \mu$, les deux séries divergent. □

Revenons à l'interprétation des paramètres : $\frac{1}{\lambda}$ est le temps moyen qui sépare deux arrivées consécutives, et λ est le nombre moyen de clients qui arrivent dans la file par unité de temps. C'est l'intensité du processus de Poisson des arrivées, nous parlerons de *taux d'arrivée*. De même $\frac{1}{\mu}$ est la durée moyenne d'un service. Donc μ est le nombre moyen de clients que le serveur traiterait s'il fonctionnait sans interruption, autrement dit la capacité moyenne maximale du service. Nous parlerons aussi de *taux de service*. Si le nombre de clients arrivant par unité de temps est strictement supérieur à la capacité maximale, le serveur ne peut pas faire face : les clients s'accumulent en attente, et la file sature : en moyenne, on y trouvera à pau près $(\lambda - \mu)t$ clients au temps t. Si le nombre de clients arrivant par unité de temps est inférieur à la capacité du serveur, celui-ci peut faire face à toutes les demandes, et la file sera équilibrée. La figure 3 montre une simulation d'une file M/M/1 dans les trois cas possibles.

Le raisonnement ci-dessus est tout à fait général et nous le retrouverons souvent dans la suite : l'équilibre d'une file d'attente dépend du rapport ρ du taux d'arrivée au taux maximal de service. Une file d'attente est équilibrée si et seulement si ρ est strictement inférieur à 1.

1.3 Equations de Chapman-Kolmogorov

Nous commençons par une description des mouvements possibles d'un processus de naissance et de mort sur un petit intervalle de temps. On consi-

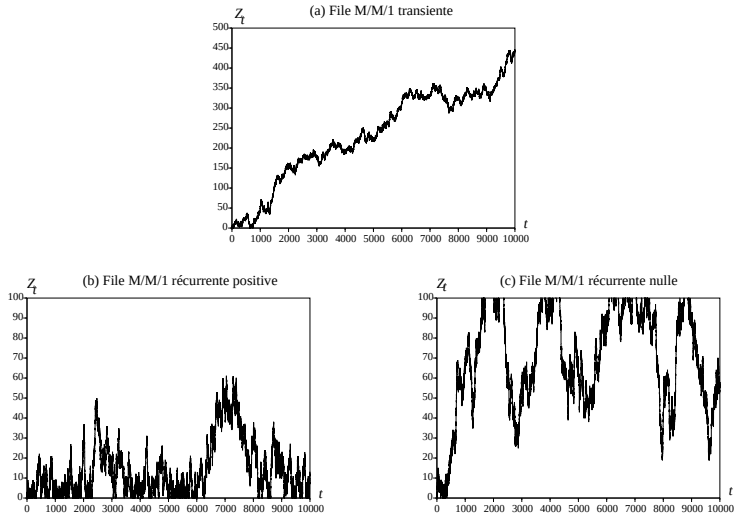

FIGURE 3 – Simulation d'une file M/M/1 dans les trois cas transient (a), récurrent positif (b) et récurrent nul (c). Le taux de service vaut 1 dans les trois cas, les taux d'arrivée valent respectivement 1.05, 0.95 et 1.

dère un processus de naissance et de mort $\{Z_t\,,\, t \geq 0\}$, de taux de naissance $(\lambda_n)_{n\in I\!N}$ et taux de mort $(\mu_n)_{n\in I\!N^*}$. Supposons que le processus soit dans l'état $n > 0$ à l'instant t. Le prochain saut l'amènera soit en $n+1$, soit en $n-1$. Les différentes probabilités de localisation de Z_{t+h}, pour h petit, sont données ci-dessous.

$$\begin{array}{ll} I\!P[Z_{t+h} = n+1 \,|\, Z_t = n] & = \lambda_n h + o(h) \\ I\!P[Z_{t+h} = n-1 \,|\, Z_t = n] & = \mu_n h + o(h) \\ I\!P[Z_{t+h} = n \,|\, Z_t = n] & = 1 - (\lambda_n + \mu_n)h + o(h)\ . \end{array}$$

On en déduit la relation suivante entre la loi de Z_{t+h} et la loi de Z_t.

$$\begin{aligned} I\!P[Z_{t+h} = n] = {} & \lambda_{n-1} h I\!P[Z_t = n-1] \\ & + \mu_{n+1} h I\!P[Z_t = n+1] + (1 - h(\lambda_n + \mu_n)) I\!P[Z_t = n] + o(h)\ . \end{aligned}$$

Soit en réarrangeant les différents termes :

$$\begin{aligned} \frac{1}{h}\Big(I\!P[Z_{t+h} = n] - I\!P[Z_t = n]\Big) = {} & \lambda_{n-1} I\!P[Z_t = n-1] \\ & - (\lambda_n + \mu_n) I\!P[Z_t = n] + \mu_{n+1} I\!P[Z_t = n+1] + \frac{o(h)}{h}\ . \end{aligned}$$

Posons $p_n(t) = I\!P[Z_t = n]$. En passant à la limite quand h tend vers 0, on arrive à l'équation différentielle suivante.

$$p_n'(t) = \lambda_{n-1}\, p_{n-1}(t) - (\lambda_n + \mu_n)\, p_n(t) + \mu_{n+1}\, p_{n+1}(t)\ .$$

Pour l'état $n = 0$, l'équation est légèrement différente.

$$p_0'(t) = -\lambda_0\, p_0(t) + \mu_1\, p_1(t)\ .$$

Le système d'équations différentielles ainsi obtenu est le *système de Chapman-Kolmogorov.*

$$\begin{cases} p_0'(t) = -\lambda_0\, p_0(t) + \mu_1\, p_1(t) \\ p_n'(t) = \lambda_{n-1}\, p_{n-1}(t) - (\lambda_n + \mu_n)\, p_n(t) + \mu_{n+1}\, p_{n+1}(t)\ , \quad \forall n \geq 1\ . \end{cases} \tag{1.1}$$

Sauf dans quelques cas très particuliers, il est hors de question de le résoudre explicitement. La théorie dit que pour une loi de probabilité initiale $(p_n(0))_{n\in\mathbb{N}}$ donnée, le système (1.1) admet une solution $(p_n(t))_{n\in\mathbb{N}}$ unique, et que dans le cas où il n'y a pas explosion à temps fini, cette solution est une loi de probabilité pour tout t.

Une *mesure stationnaire* est une loi de probabilité $\pi = (\pi_n)_{n\in\mathbb{N}}$ telle que si la loi de Z_0 est π, alors la loi de Z_t reste π pour tout $t > 0$. Une telle mesure est donc nécessairement une solution constante du système de Chapman-Kolmogorov :

$$\begin{cases} 0 = -\lambda_0\, \pi_0 + \mu_1\, \pi_1 \\ 0 = \lambda_{n-1}\, \pi_{n-1} - (\lambda_n + \mu_n)\, \pi_n + \mu_{n+1}\, \pi_{n+1}\ , \quad \forall n \geq 1\ . \end{cases} \tag{1.2}$$

Il est facile de résoudre ce système, en vérifiant par récurrence que pour tout $n \geq 1$:

$$\mu_n \pi_n - \lambda_{n-1} \pi_{n-1} = 0\ .$$

On en déduit immédiatement l'expression de π_n en fonction de π_0.

$$\pi_n = \pi_0 \frac{\lambda_0 \dots \lambda_{n-1}}{\mu_1 \dots \mu_n}\ . \tag{1.3}$$

Le système (1.2) admet donc toujours une solution, unique à un coefficient multiplicatif près. Mais pour que ce soit une loi de probabilité, il est nécessaire que la somme des coefficients soit 1. C'est le cas si et seulement si la série de terme général $\frac{\lambda_0 \dots \lambda_{n-1}}{\mu_1 \dots \mu_n}$ converge. Nous avons déjà vu que c'est aussi la condition pour que le processus soit récurrent positif. Nous résumons dans le théorème suivant, que nous admettons pour l'essentiel, les différents comportements d'un processus de naissance et de mort qui nous intéressent en pratique.

Théorème 1.7 *Soit $\{Z_t\,,\, t \geq 0\}$ un processus de naissance et de mort, de taux de naissance $(\lambda_n)_{n\in\mathbb{N}}$ et taux de mort $(\mu_n)_{n\in\mathbb{N}^*}$, supposés strictement positifs. Notons $(p_n(t))_{n\in\mathbb{N}}$ la loi de Z_t, posons $\alpha_0 = 1$ et pour $n \geq 1$:*

$$\alpha_n = \frac{\lambda_0 \dots \lambda_{n-1}}{\mu_1 \dots \mu_n}\ .$$

1. *Si la série* $\sum \alpha_n$ *diverge, alors le processus est récurrent nul ou transient. Pour tout* $n \in \mathbb{N}$, $p_n(t)$ *tend vers* 0 *quand* t *tend vers l'infini, et donc* Z_t *converge en probabilité vers* $+\infty$.
2. *Si la série* $\sum \alpha_n$ *converge, alors le processus est récurrent positif. Posons :*

$$\pi_n = \left(\sum_{m=0}^{\infty} \alpha_m \right)^{-1} \alpha_n \ . \tag{1.4}$$

La loi de probabilité $\pi = (\pi_n)_{n \in \mathbb{N}}$ *est la* mesure stationnaire *du processus. Pour tout* $n \in \mathbb{N}$, $p_n(t)$ *tend vers* π_n *et donc* Z_t *converge en loi vers* π. *De plus si* f *est une fonction quelconque de* $\mathbb{N}$ *dans* $\mathbb{R}$, *on a :*

$$\lim_{\tau \to \infty} \frac{1}{\tau} \int_0^{\tau} f(Z_t)\, dt = \sum_{i \in \mathbb{N}} f(i)\, \pi_i \ , \quad p.s. \tag{1.5}$$

Pour comprendre la condition d'équilibre d'un processus de naissance et de mort, utilisons le critère de d'Alembert. Le rapport de deux termes consécutifs de la série $\sum \alpha_n$ est :

$$\frac{\alpha_{n+1}}{\alpha_n} = \frac{\lambda_n}{\mu_{n+1}} \ .$$

Si ce rapport reste inférieur à $c < 1$ pour n assez grand, la série converge. S'il est supérieur à 1, elle diverge. Nous retrouvons donc l'interprétation déjà mentionnée pour le cas particulier de la file M/M/1. Si les taux d'arrivée λ_n l'emportent sur les taux de départ μ_n, le nombre de clients tend à augmenter, et le processus tend vers l'infini. Dans le cas contraire, un régime stationnaire peut s'établir, et ce régime stationnaire est toujours atteint au bout d'un temps suffisamment long, quelle que soit la loi de Z_0 (l'état initial du système). La dernière assertion du théorème s'interprète comme suit. Comprenons f comme le coût du séjour dans n : le concepteur du système verse $f(n)$ par unité de temps passée dans n. L'intégrale $\int_0^{\tau} f(Z_t)\, dt$ est le coût total de la trajectoire jusqu'à l'instant τ. En divisant par τ, on obtient le coût moyen par unité de temps de cette trajectoire. La formule (1.5) dit que ce coût moyen, qui varie a priori d'une trajectoire à l'autre, est en fait assez stable sur le long terme, puisqu'il converge vers la valeur moyenne de f pour la mesure stationnaire π. Si on prend comme cas particulier pour f la fonction indicatrice de l'état n, alors le membre de gauche de (1.5) est la proportion de temps passé dans l'état n par une trajectoire entre 0 et τ. Quand τ tend vers l'infini, cette proportion tend vers π_n.

2 Files markoviennes

2.1 Modèles d'attente

Dans un système d'attente, des clients arrivent de manière aléatoire dans un système d'où ils partiront après avoir reçu un service, dont la durée est elle-

même aléatoire. Historiquement, la recherche sur les files d'attente a débuté avec les travaux d'Erlang (1909) sur les réseaux téléphoniques. Les clients dans ce cas sont des appels téléphoniques arrivant à un central, et les temps de service sont les durées des communications. Mais beaucoup d'autres situations pratiques relèvent des modèles d'attente. Dans un aéroport, les avions sont les clients, le temps de service est la durée nécessaire à un avion pour atterrir et dégager la piste. L'attente pour un avion consiste à tourner au-dessus de l'aéroport avant le feu vert de la tour de contrôle. La même situation se retrouve dans un port de commerce où les cargos attendent un quai libre pour décharger. Dans une usine, les clients peuvent être des machines, immobilisées pendant le temps nécessaire à leur maintenance ou leur réparation. Avec l'essor de l'informatique, on a assisté à un regain d'intérêt pour les files d'attente, dû à leurs applications dans la gestion des réseaux. Là les clients sont des programmes à traiter ou des paquets de données à acheminer. Evidemment les modèles à une seule file sont trop simplistes pour rendre compte de toute la complexité d'un réseau informatique.

La figure 4 est le schéma classique d'un modèle d'attente. On peut le caractériser par les éléments suivants.

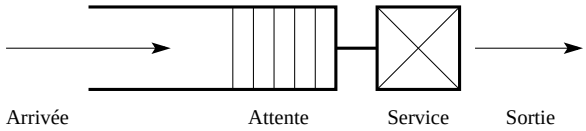

FIGURE 4 – Schéma d'une file d'attente.

1. *Le processus d'arrivée* : il est déterminé en général par la loi conjointe des intervalles de temps séparant deux arrivées consécutives, que l'on appelle des *temps d'interarrivée*. En général, on suppose que leur loi ne dépend pas des clients. Souvent, on est conduit pour des raisons de simplification à supposer que les temps d'interarrivée sont indépendants entre eux. Le cas le plus facile à traiter mathématiquement est celui où ils suivent tous la même loi exponentielle. Dans ce cas, les clients arrivent selon un processus de Poisson.
2. *Les durées de service* : ce sont des variables aléatoires que l'on considère en général comme indépendantes et de même loi. Là encore, c'est la loi exponentielle qui donne le traitement mathématique le plus simple. D'autres familles de lois sont parfois utilisées (masse de Dirac pour des temps de service constants, loi d'Erlang pour des services en série, loi de Weibull, etc...).
3. *Le nombre de serveurs* : dans un premier temps, on considérera toujours une file unique, ayant éventuellement plusieurs serveurs. Mais tous les serveurs seront identiques, et chaque client n'aura affaire qu'à un seul

d'entre eux. Nous aborderons dans la quatrième partie des modèles plus complexes, où plusieurs files d'attente sont connectées en réseau.

4. *La capacité maximale du système* : même si nous considérerons souvent pour simplifier que la file peut contenir un nombre arbitraire de clients, ce n'est pas réaliste. Certains modèles prennent en compte le nombre maximal de clients qui peuvent séjourner simultanément dans le système.

5. *La population des usagers* : dans certains modèles (comme par exemple la maintenance des machines d'une usine), le nombre total de clients est fixé.

6. *La discipline de service* : c'est l'ordre dans lequel les clients en attente sont traités. Par défaut, on considérera que les clients sont servis dans leur ordre d'arrivée : c'est la dicipline PAPS (premier arrivé premier servi) ou FIFO en anglais (first in first out). D'autres disciplines sont couramment envisagées selon les applications :
   - Dernier arrivé premier servi (ou LIFO : last in first out). C'est ce qui est pratiqué en informatique pour les piles de données.
   - Aléatoire (ou FIRO : first in random out). Le serveur choisit au hasard parmi les clients présents le prochain qui sera servi.
   - Temps partagé (ou PS : processor sharing). Quand plusieurs programmes sont présents simultanément dans une unité centrale, ils reçoivent chacun leur tour une partie de leur traitement, jusqu'à achèvement de toutes les tâches.

D'autres paramètres peuvent également entrer en ligne de compte, par exemple dans les cas où les clients arrivent ou sont servis par paquets, nous en verrons plus loin des exemples. La *notation de Kendall* résume par des symboles, séparés par des barres obliques les différents paramètres de définition du modèle.

Arrivées / Services / Serveurs / Capacité / Population / Discipline

Ainsi la file M/M/1 a des temps d'interarrivée et de service exponentiels et indépendants (le M signifie "markovien"), un seul serveur, et les autres paramètres sont fixés par défaut : $+\infty$ pour la capacité et la population, PAPS pour la discipline. Nous étudierons aussi la file M/M/s (s serveurs), la file M/M/1/N (capacité limitée à N), la file M/GI/1 (les services sont indépendants mais leur loi est générale), etc...

Par exemple, la file E_2/D/5/100/FIRO aurait des temps d'interarrivée suivant une loi d'Erlang de paramètre 2 (somme de deux exponentielles indépendantes), des services constants (D pour Dirac), 5 serveurs, une capacité limitée à 100 clients et fonctionnerait selon la discipline aléatoire.

2.2 File M/M/1

Nous avons déjà vu plusieurs caractéristiques du modèle le plus simple. C'est un processus de naissance et de mort dont les taux sont constants : $\lambda_n = \lambda \ \forall n \geq 0$ et $\mu_n = \mu \ \forall n \geq 1$. Son comportement asymptotique dépend du rapport $\frac{\lambda}{\mu}$ (voir figure 3). Si ce rapport est supérieur ou égal à 1, la file sature, s'il est strictement inférieur à 1, elle admet un régime stationnaire. Dans ce cas, la mesure stationnaire $\pi = (\pi_n)$ est donnée par :

$$\pi_n = \left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^n \ .$$

La loi π est la loi du nombre de clients présents dans le système en régime stationnaire. Nous noterons Z la variable aléatoire correspondante. Son espérance est le nombre moyen de clients présents dans le système en régime stationnaire :

$$I\!E[Z] = \frac{\frac{\lambda}{\mu}}{1 - \frac{\lambda}{\mu}} = \frac{\lambda}{\mu - \lambda} \ . \tag{2.6}$$

Comme on pouvait s'y attendre, le nombre moyen de clients dans le système augmente quand λ se rapproche de μ.

Sous la loi π, la probabilité que le serveur soit occupé est $I\!P[Z > 0] = 1 - \pi_0 = \frac{\lambda}{\mu}$. On peut donc interpréter le rapport $\frac{\lambda}{\mu}$ comme la probabilité que le serveur soit occupé à l'équilibre, ou comme la proportion moyenne du temps qu'il passe à servir les clients. On parle de *coefficient d'occupation* du serveur.

Notre but ici est d'une part d'étudier le temps passé par un client dans le système en régime stationnaire, d'autre part d'étudier le régime transitoire en présentant deux techniques de résolution du système de Chapman-Kolmogorov (1.1). Plus que le traitement de la file M/M/1, qui est un modèle trop simple pour être réaliste, l'intérêt de ce qui suit est de mettre en place des méthodes mathématiques utiles dans beaucoup d'autres contextes.

Nous commençons par étudier le régime stationnaire. Nous supposons donc que λ est strictement inférieur à μ et que la loi de Z_0 est π. Nous dirons que la file est "à l'équilibre", ce qui ne se produit en pratique qu'approximativement, au bout d'un temps suffisamment long. Considérons alors un client à l'instant de son arrivée et notons T le temps total qu'il passera dans le système.

Proposition 2.1 *La loi du temps de séjour T d'un client arrivant en régime stationnaire est la loi exponentielle de paramètre $\mu - \lambda$.*

Démonstration : Si le système est vide quand le client arrive, son temps de séjour est égal à son temps de service, et il suit la loi $\mathcal{E}(\mu)$. S'il y a $n \geq 1$ clients quand il arrive, son temps de séjour est la somme :

- du temps de service résiduel du client en train d'être servi, qui suit la loi $\mathcal{E}(\mu)$ par la propriété d'absence de mémoire.
- des $n-1$ temps de service des clients en attente devant lui (chacun suit la loi $\mathcal{E}(\mu)$).
- de son propre temps de service, toujours de loi $\mathcal{E}(\mu)$.

Au total, la loi conditionnelle de T sachant que n clients sont présents est la loi de la somme de $n+1$ exponentielles indépendantes, autrement dit la loi gamma $\mathcal{G}(n+1,\mu)$, appelée aussi loi d'Erlang. Elle a pour densité :

$$\frac{\mu^{n+1}t^n}{n!}e^{-\mu t}\,\mathbb{1}_{I\!R^+}(t)\ .$$

Connaissant la probabilité π_n que n clients soient présents en régime stationnaire, on peut donc écrire la densité de T sous la forme suivante.

$$\begin{aligned} f_T(t) &= \sum_{n=0}^{\infty} \pi_n \,\frac{\mu^{n+1}t^n}{n!}e^{-\mu t}\,\mathbb{1}_{I\!R^+}(t) \\ &= \sum_{n=0}^{\infty} \left(1-\frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^n \frac{\mu^{n+1}t^n}{n!}e^{-\mu t}\,\mathbb{1}_{I\!R^+}(t) \\ &= (\mu-\lambda)e^{-(\mu-\lambda)t}\,\mathbb{1}_{I\!R^+}(t)\ . \end{aligned}$$

La loi du temps de séjour d'un client arrivant en régime stationnaire est donc bien la loi exponentielle $\mathcal{E}(\mu-\lambda)$. Son espérance vaut :

$$I\!E[T] = \frac{1}{\mu-\lambda}\ . \tag{2.7}$$

□

La comparaison de (2.6) et (2.7) fait apparaître une relation qui est valable dans des situations beaucoup plus générales que celle de la file M/M/1. C'est la *formule de Little*.

Proposition 2.2 *Dans une file d'attente en régime stationnaire, le nombre moyen de clients est le produit du taux d'arrivée par le temps moyen de séjour d'un client dans le système.*

$$I\!E[Z] = \lambda\, I\!E[T]\ . \tag{2.8}$$

Une démonstration rigoureuse de ce résultat sous des hypothèses générales dépasse le cadre restreint de ce cours. Nous nous contenterons de le vérifier sur des cas particuliers dans ce qui suit.

Nous nous intéressons maintenant au régime transitoire de la file M/M/1, c'est-à-dire à la résolution du système différentiel de Chapman-Kolmogorov

(2.9).

$$\begin{cases} p_0'(t) = -\lambda\, p_0(t) + \mu\, p_1(t) \\ p_n'(t) = \lambda\, p_{n-1}(t) - (\lambda+\mu)\, p_n(t) + \mu\, p_{n+1}(t)\ , \quad \forall n \geq 1\ . \end{cases} \tag{2.9}$$

Ce sera l'occasion de mettre en œuvre deux outils mathématiques importants pour l'étude des modèles d'attente markoviens, les transformées de Laplace et les fonctions génératrices.

Commençons par les transformées de Laplace. Si φ est une fonction de $\mathbb{R}^+$ dans $\mathbb{R}$, sa transformée de Laplace, qui sera notée φ^* est la fonction définie pour tout $s \geq 0$ par :

$$\varphi^*(s) = \int_0^{+\infty} e^{-st}\varphi(t)\, dt\ .$$

Si X est une variable aléatoire, sa transformée de Laplace (ou plutôt celle de sa loi) est la fonction qui à s associe :

$$\mathbb{E}[\exp(-sX)]\ .$$

Dans le cas où X admet pour densité f, sa transformée de Laplace est celle de sa densité :

$$\mathbb{E}[\exp(-sX)] = \int_0^{+\infty} e^{-st} f(t)\, dt = f^*(s)\ .$$

L'usage que nous ferons des transformées de Laplace est justifié par les propriétés suivantes.

Proposition 2.3

1. *La transformée de Laplace est linéaire : si φ et ψ sont deux fonctions de $\mathbb{R}^+$ dans $\mathbb{R}$, α et β sont deux constantes, alors la transformée de Laplace de $\alpha\varphi + \beta\psi$ est $\alpha\varphi^* + \beta\psi^*$.*
2. *La transformée de Laplace de $\varphi'(t)$ est $s\varphi^*(s) - \varphi(0)$.*
3. *La fonction $\varphi(t)$ admet une limite en $+\infty$ si et seulement si $s\,\varphi^*(s)$ admet une limite en 0, et si elles existent ces limites sont égales.*

$$\lim_{t\to\infty} \varphi(t) = \lim_{s\to 0} s\,\varphi^*(s)\ .$$

4. *Soit X une variable aléatoire, f^* sa transformée de Laplace et k un entier tel que $\mathbb{E}[X^k] < \infty$, alors :*

$$E[X^k] = (-1)^k \frac{d^k f^*}{ds^k}(0)\ .$$

Nous ne démontrerons pas cette proposition, pas plus que nous ne justifierons par la suite les nombreux points techniques liés à l'utilisation des transformées de Laplace et des fonctions génératrices : dans tout ce qui suit, les convergences de séries et d'intégrales, les dérivations sous le signe somme, les interversions de sommes ou d'intégrales seront affirmées sans démonstration.

Pour résoudre le système de Chapman-Kolmogorov (2.9), l'idée est de remplacer la fonction inconnue $p_n(t)$ par sa transformée de Laplace $p_n^*(s)$. Grâce au point *2* de la proposition 2.3, remplacer les $p_n(t)$ par leurs transformées de Laplace permet de supprimer le caractère différentiel du système.

$$\begin{cases} s\,p_0^*(s) - p_0(0) = -\lambda\,p_0^*(s) + \mu\,p_1^*(s) \\ s\,p_n^*(s) - p_n(0) = \lambda\,p_{n-1}^*(s) - (\lambda+\mu)\,p_n^*(s) + \mu\,p_{n+1}^*(s)\ , \quad \forall n \geq 1\ . \end{cases}$$

Comme condition initiale, nous supposerons que la file est vide à l'instant 0, soit $p_0(0) = 1$ et $p_n(0) = 0$ pour $n \geq 1$. Les $p_n^*(s)$ sont donc solution de l'équation de récurrence :

$$sp_n^*(s) = \lambda p_{n-1}^*(s) - (\lambda+\mu)p_n^*(s) + \mu p_{n+1}^*(s)\ .$$

La solution générale de cette équation est :

$$p_n^*(s) = A(s)\,\alpha(s)^n + B(s)\,\beta(s)^n\ ,$$

où $\alpha(s)$ et $\beta(s)$ sont les racines de l'équation caractéristique associée, à savoir :

$$\alpha(s) = \frac{1}{2\mu}\Big(s+\lambda+\mu+\sqrt{(s+\lambda+\mu)^2-4\lambda\mu}\Big)\ ,$$

et

$$\beta(s) = \frac{1}{2\mu}\Big(s+\lambda+\mu-\sqrt{(s+\lambda+\mu)^2-4\lambda\mu}\Big)\ .$$

On vérifie facilement que $0 < \beta(s) < 1 < \alpha(s)$. Par linéarité de la transformée de Laplace, $\sum p_n(t) = 1$ entraîne $\sum p_n^*(s) = 1/s$. Comme $\sum p_n^*(s)$ est une série convergente, nécessairement $A(s)$ doit être nul. La condition $\sum p_n^*(s) = 1/s$ donne :

$$B(s) = \frac{1}{s}(1-\beta(s))\ ,$$

d'où l'expression de $p_n^*(s)$:

$$p_n^*(s) = \frac{1}{s}(1-\beta(s))(\beta(s))^n\ .$$

Le retour de $p_n^*(s)$ à $p_n(t)$ est plus difficile. Il existe des méthodes numériques de calcul des transformées de Laplace inverses, ainsi que des tables formelles. On peut étudier le comportement asymptotique de $p_n(t)$ à l'aide du point *3* de la proposition 2.3.

$$\lim_{t\to\infty} p_n(t) = \lim_{s\to 0} sp_n^*(s) = (1-\beta)\beta^n\ ,$$

où

$$\beta = \lim_{s\to 0} \beta(s) = \frac{1}{2\mu}(\lambda + \mu - |\lambda - \mu|) \ .$$

Si $\lambda \geq \mu$ (saturation) alors $\beta = 1$, donc pour tout n, $\lim p_n(t) = 0$. Si $\lambda < \mu$ (équilibre) alors $\beta = \lambda/\mu$, donc $\lim p_n(t) = \pi_n$. On retrouve ainsi, pour la file M/M/1, les deux cas du théorème 1.7.

Présentons maintenant la technique des fonctions génératrices. L'idée est de remplacer les fonctions $p_n(t)$ par leur série entière $G(z,t)$, définie pour $|z| < 1$ et $t \geq 0$ par :

$$G(z,t) = \sum_{n=0}^{\infty} z^n \, p_n(t) \ .$$

En multipliant la n-ième équation du système (2.9) par z^n et en sommant, on obtient l'équation aux dérivées partielles suivante.

$$z\frac{\partial G(z,t)}{\partial t} = G(z,t)(\lambda z^2 - (\lambda + \mu)z + \mu) + p_0(t)\mu(z-1) \ .$$

Dans ce cas l'équation obtenue dépend de $p_0(t) = G(0,t)$, et elle n'a pas de solution explicite. On peut, en utilisant à nouveau la transformée de Laplace, en déduire aussi le comportement asymptotique de la file. Nous verrons dans la suite plusieurs exemples d'application de cette technique.

2.3 File M/M/s

Le modèle que nous étudions ici est l'extension la plus simple de la file M/M/1 : les hypothèses probabilistes sont les mêmes, seul change le nombre de serveurs.

- Les clients arrivent un par un selon un processus de Poisson d'intensité λ : le temps séparant deux arrivées successives suit la loi $\mathcal{E}(\lambda)$.
- Les s serveurs sont identiques. Si un client arrivant dans la file trouve un serveur libre, il l'occupe. Si les s serveurs sont occupés, le nouvel arrivant attend. Le temps de service de chaque client suit la loi $\mathcal{E}(\mu)$.
- Toutes les variables aléatoires considérées (temps séparant deux arrivées successives et temps de service) sont indépendantes dans leur ensemble.

Nous notons encore Z_t le nombre de clients présents dans le système (en attente ou au service) à l'instant t. Supposons que ce nombre soit n. Le prochain événement peut être une arrivée ou bien la fin de l'un des services. Le temps qui sépare t de la prochaine arrivée est exponentiel de paramètre λ, par la propriété d'absence de mémoire. Si $n \leq s$, alors n serveurs sont occupés (personne n'est en train d'attendre), et les temps résiduels de ces n services sont exponentiels de paramètre μ. Le temps qui sépare t du prochain saut est exponentiel de paramètre $\lambda + n\mu$, ce prochain saut sera une arrivée avec probabilité $\frac{\lambda}{\lambda+n\mu}$, ou un départ avec probabilité $\frac{n\mu}{\lambda+n\mu}$ (proposition 1.2). Le cas où n est plus grand que s (tous les serveurs sont occupés et des clients

sont en attente), est différent, puisqu'alors le prochain départ ne peut être que celui d'un des s clients en train d'être servi. Dans ce cas le temps qui sépare t du prochain saut est exponentiel de paramètre $\lambda + s\mu$, ce prochain saut sera une arrivée avec probabilité $\frac{\lambda}{\lambda+s\mu}$, ou un départ avec probabilité $\frac{s\mu}{\lambda+s\mu}$. En résumé, $\{Z_t\,,\, t \geq 0\}$ est un processus de naissance et de mort, de taux de naissance constants $\lambda_n = \lambda\ \forall n \geq 0$, et de taux de mort μ_n tels que :

$$\mu_n = \begin{cases} n\mu\ \forall n = 1, \ldots, s \\ s\mu\ \ \forall n \geq s\ . \end{cases}$$

D'après le théorème 1.7, la condition d'équilibre du système est la convergence de la série de terme général $\alpha_n = \frac{\lambda_0 \ldots \lambda_{n-1}}{\mu_1 \ldots \mu_n}$, avec :

$$\alpha_n = \begin{cases} \dfrac{\lambda^n}{n!\mu^n} & \forall n = 0, \ldots, s \\ \dfrac{\lambda^n}{s!s^{n-s}\mu^n}\ \forall n \geq s\ . \end{cases}$$

La série $\sum \alpha_n$ converge si et seulement si $\lambda < s\mu$, comme on pouvait le prévoir. Si chacun des s serveurs fonctionnait en continu, ils traiteraient $s\mu$ clients par unité de temps, donc $s\mu$ est la capacité maximale du service, qui doit être supérieure au nombre λ de clients arrivant par unité de temps. Si la condition d'équilibre est réalisée, alors la mesure stationnaire (π_n) est donnée par (1.4), à savoir $\pi_n = (\sum \alpha_n)^{-1}\alpha_n$, avec :

$$\sum_{n=0}^{\infty} \alpha_n = \left(\sum_{n=0}^{s-1} \frac{\lambda^n}{n!\mu^n} \right) + \frac{\lambda^s}{s!\mu^s} \frac{1}{1 - \frac{\lambda}{s\mu}}\ .$$

Ayant l'expression explicite des π_n, on peut calculer le nombre moyen de clients dans le système en régime stationnaire, soit $I\!E[Z] = \sum n\,\pi_n$. Il est plus simple de l'exprimer en fonction de la probabilité π_s. On trouve :

$$I\!E[Z] = \frac{\lambda}{\mu} + \frac{\pi_s \lambda s\mu}{(s\mu - \lambda)^2}\ . \tag{2.10}$$

Nous allons maintenant étudier la loi du temps d'attente d'un client arrivant dans le système en régime stationnaire. Nous supposons donc que $\lambda < s\mu$ et que la loi de Z_t est π, pour tout t.

Proposition 2.4 *La loi du temps d'attente d'un client en régime stationnaire est un mélange*

1. *de la masse de Dirac en* 0*, avec probabilité* $1 - \frac{\pi_s}{1-\frac{\lambda}{s\mu}}$ *,*
2. *de la loi exponentielle* $\mathcal{E}(s\mu - \lambda)$*, avec probabilité* $\frac{\pi_s}{1-\frac{\lambda}{s\mu}}$ *.*

Démonstration : Tout d'abord, un client n'attend que si tous les serveurs sont occupés, donc si au moins s clients sont déjà présents quand il arrive. Ceci se produit avec probabilité :

$$\pi_s + \pi_{s+1} + \cdots = \pi_s \sum_{k=0}^{\infty} \frac{\lambda^k}{s^k \mu^k} = \frac{\pi_s}{1 - \frac{\lambda}{s\mu}} \ .$$

S'il y a $n \geq s$ clients déjà présents dans le système, celui qui arrive trouve $n-s$ clients en attente avant lui. Il devra attendre $(n-s+1)$ libérations de serveurs, chaque libération survenant au bout d'un temps exponentiel de paramètre $s\mu$. La loi conditionnelle du temps d'attente sachant qu'il y a $n \geq s$ clients déjà présents dans le système est donc la loi gamma $\mathcal{G}(n-s+1, s\mu)$, de densité :

$$\frac{(s\mu)^{n-s+1} t^{n-s}}{(n-s)!} e^{-s\mu t} \, \mathbb{1}_{I\!R^+}(t) \ .$$

La loi du temps d'attente est un mélange de la masse de Dirac en 0 (s'il n'y a pas attente) et d'une loi sur $I\!R^+$, dont la densité est :

$$\frac{1 - \frac{\lambda}{s\mu}}{\pi_s} \sum_{n=s}^{\infty} \pi_n \frac{(s\mu)^{n-s+1} t^{n-s}}{(n-s)!} e^{-s\mu t} \, \mathbb{1}_{I\!R^+}(t) \ .$$

Après simplifications, on vérifie que cette densité est bien celle de la loi exponentielle de paramètre $s\mu - \lambda$. □

L'espérance du temps d'attente est donc :

$$\frac{\pi_s}{1 - \frac{\lambda}{s\mu}} \frac{1}{s\mu - \lambda} = \frac{\pi_s s\mu}{(s\mu - \lambda)^2} \ .$$

Pour obtenir le temps moyen de séjour d'un client arrivant en régime stationnaire, il faut ajouter le temps de service au temps d'attente :

$$I\!E[T] = \frac{1}{\mu} + \frac{\pi_s s\mu}{(s\mu - \lambda)^2} \ . \tag{2.11}$$

En comparant (2.10) et (2.11), on retrouve la formule de Little, à savoir $I\!E[Z] = \lambda I\!E[T]$. On peut aussi obtenir d'autres quantités (toujours pour le régime stationnaire), comme :

- la longueur moyenne de la file : $I\!E[Z] - \frac{\lambda}{\mu}$,
- le nombre moyen de serveurs occupés : $I\!E[\mathrm{O}] = \frac{\lambda}{\mu} = s\frac{\lambda}{s\mu}$.

Dans cette dernière formule, on retrouve l'interprétation du rapport $\frac{\lambda}{s\mu}$ (rapport du taux d'arrivée à la capacité de service) comme un coefficient d'occupation.

Comme nouvel exemple d'utilisation des fonctions génératrices, nous traitons maintenant le cas de la file M/M/∞. Même si supposer qu'il y a une

infinité de serveurs n'est pas vraiment réaliste en pratique, cela peut constituer une bonne approximation de la file M/M/s dans le cas où le coefficient d'occupation $\frac{\lambda}{s\mu}$ est faible. Une file M/M/∞ est un processus de naissance et de mort de taux de naissance constants $\lambda_n = \lambda \ \forall n \geq 0$, et de taux de mort linéaires $\mu_n = n\mu \ \forall n \geq 1$. Il ne peut pas y avoir de saturation dans une telle file. Elle est toujours récurrente positive, avec pour mesure stationnaire la loi de Poisson $\mathcal{P}(\frac{\lambda}{\mu})$.

$$\pi_n = e^{-\frac{\lambda}{\mu}} \frac{(\frac{\lambda}{\mu})^n}{n!} \ .$$

Le système de Chapman-Kolmogorov est le suivant.

$$\begin{cases} p_0'(t) = -\lambda\, p_0(t) + \mu\, p_1(t) \\ p_n'(t) = \lambda\, p_{n-1}(t) - (\lambda + n\mu)\, p_n(t) + (n+1)\mu\, p_{n+1}(t) \ , \quad \forall n \geq 1 \ . \end{cases} \tag{2.12}$$

Posons $G(z,t) = \sum z^n\, p_n(t)$, multiplions la n-ième équation par z^n et sommons. En remarquant que :

$$\frac{\partial G(z,t)}{\partial t} = \sum_{n=0}^{\infty} z^n p_n'(t) \quad \text{et} \quad \frac{\partial G(z,t)}{\partial z} = \sum_{n=1}^{\infty} n z^{n-1} p_n(t) \ ,$$

on arrive à l'équation aux dérivées partielles suivante.

$$\frac{\partial G(z,t)}{\partial t} = \lambda(z-1)G(z,t) + \mu(1-z)\frac{\partial G(z,t)}{\partial z} \ . \tag{2.13}$$

Cette équation admet pour solution particulière la fonction génératrice de la loi de Poisson de paramètre $\rho(t)$:

$$G(z,t) = e^{\rho(t)(z-1)} \ ,$$

avec

$$\rho(t) = \frac{\lambda}{\mu} + \left(\rho(0) - \frac{\lambda}{\mu}\right) e^{-\mu t} \ .$$

Cela signifie que si la loi du nombre de clients dans la file à l'instant 0 est une loi de Poisson, alors il en sera de même à chaque instant : seul le paramètre $\rho(t)$ dépend du temps. Quand t tend vers l'infini, $\rho(t)$ converge vers $\frac{\lambda}{\mu}$, qui est le paramètre de la mesure stationnaire π. On retrouve donc la convergence en loi vers la mesure stationnaire. De plus ici, cette convergence a lieu à vitesse exponentielle (en $e^{-\mu t}$), ce qui signifie en pratique que l'équilibre sera atteint rapidement.

2.4 Files à capacité limitée

Nous allons examiner trois modèles particuliers, pour lesquels le nombre de clients dans le système est un processus de naissance et de mort à nombre d'états fini.

Le premier est la file M/M/1/N, qui fonctionne comme la file M/M/1, sauf que le nombre maximal de clients dans le système est limité à N : un client qui arrive lorsque la file est pleine est rejeté. Ceci revient à annuler le taux de transition de N vers $N+1$. On obtient un processus de naissance et de mort sur $\{0,\dots,N\}$, dont les taux de naissance valent $\lambda_n = \lambda\ \forall n = 0,\dots,N-1$, et les taux de mort $\mu_n = \mu\ \forall n = 1,\dots,N$. Tout processus de saut sur un ensemble fini est récurrent positif, il n'est évidemment pas question de saturation ici. La mesure stationnaire π sur $\{0,\dots,N\}$ est donnée par la formule (1.3), à savoir :

$$\forall n = 1,\dots,N\ , \quad \pi_n = \pi_0 \left(\frac{\lambda}{\mu}\right)^n\ .$$

On calcule π_0 par la condition $\sum_{n=0}^{N} \pi_n = 1$.

$$\pi_0 = \begin{cases} \frac{1-\frac{\lambda}{\mu}}{1-(\frac{\lambda}{\mu})^{N+1}} & \text{si } \frac{\lambda}{\mu} \neq 1 \\ \frac{1}{N+1} & \text{si } \frac{\lambda}{\mu} = 1\ . \end{cases}$$

On peut en déduire le nombre moyen de clients dans le système en régime stationnaire.

$$I\!E[Z] = \sum_{n=0}^{N} n\,\pi_n = \begin{cases} \frac{\frac{\lambda}{\mu}}{1-\frac{\lambda}{\mu}} - \frac{(N+1)(\frac{\lambda}{\mu})^{N+1}}{1-(\frac{\lambda}{\mu})^{N+1}} & \text{si } \frac{\lambda}{\mu} \neq 1 \\ \frac{N}{2} & \text{si } \frac{\lambda}{\mu} = 1\ . \end{cases}$$

Comme on pouvait s'y attendre, lorsque le rapport $\frac{\lambda}{\mu}$ tend vers 0, la file est le plus souvent vide, et se comporte comme une file M/M/1. Par contre si $\frac{\lambda}{\mu}$ tend vers l'infini, π_N tend vers 1 et $I\!E[Z]$ vers N : la file est pleine la plus grande partie du temps.

Examinons maintenant la file M/M/s/s qui a une capacité limitée au nombre de ses serveurs : c'est une file sans attente possible, puisqu'un client arrivant lorque tous les serveurs sont occupés est rejeté. C'est le premier modèle introduit par Erlang pour les réseaux téléphoniques : les clients sont des appels et les serveurs des lignes. Si un usager appelle lorsque toutes les lignes sont occupées, il n'est pas mis en attente. Le nombre de clients dans le système est un processus de naissance et de mort sur $\{0,\dots,s\}$, dont les taux de naissance valent $\lambda_n = \lambda\ \forall n = 0,\dots,s-1$, et les taux de mort $\mu_n = n\mu\ \forall n = 1,\dots,s$. Ici encore il n'y a pas de staturation possible. La mesure stationnaire est donnée par :

$$\pi_n = \pi_0 \frac{\lambda^n}{n!\mu^n}\ ,$$

avec

$$\pi_0 = \left(\sum_{n=0}^{s} \frac{\lambda^n}{n!\mu^n}\right)^{-1}\ .$$

La valeur de π_s est la "probabilité de perte", à savoir la probabilité qu'un client arrivant en régime stationnaire soit rejeté. Son expression en fonction de λ, μ et s est la "formule d'Erlang" ou "formule des téléphonistes". Avant les ordinateurs, les "abaques d'Erlang" traduisaient graphiquement cette formule afin de faciliter le calcul approché de π_s.

$$\pi_s = \frac{\lambda^s}{s!\mu^s} \left(\sum_{n=0}^{s} \frac{\lambda^n}{n!\mu^n} \right)^{-1} .$$

On peut exprimer le nombre moyen de lignes occupées en fonction de π_s :

$$I\!E[Z] = \frac{\lambda}{\mu}(1 - \pi_s) .$$

Si le rapport $\frac{\lambda}{\mu}$ tend vers 0, tout appel trouve une ligne et la probabilité de perte tend vers 0. S'il tend vers l'infini, la probabilité de perte tend vers 1 et $I\!E[Z]$ tend vers s.

Nous examinons pour finir un modèle de maintenance où la population des usagers est finie. Supposons qu'il y ait dans une usine m machines du même type. Quand l'une d'entre elles tombe en panne, elle entre dans l'un des s ateliers de réparation, s'il y en a un de libre. On suppose que les temps de fonctionnement des machines avant une panne sont exponentiels de paramètre λ, et que les temps de réparation sont exponentiels de paramètre μ, tous ces temps étant indépendants dans leur ensemble. Notons Z_t le nombre total de machines immobilisées à l'instant t. Des raisonnements analogues à ceux que nous avons déjà menés pour les modèles précédents montrent que $\{Z_t\,,\, t \geq 0\}$ est un processus de naissance et de mort sur $\{0, \ldots, m\}$, avec $\lambda_n = (m - n)\lambda \; \forall n = 0, \ldots, m-1$, les taux de mort μ_n étant ceux de la file M/M/s :

$$\mu_n = \begin{cases} n\mu \; \forall n = 0, \ldots, s \\ s\mu \; \forall n = s, \ldots, m \; . \end{cases}$$

On obtient encore la mesure stationnaire π par la formule (1.3). Le responsable de l'usine connaît les coûts associés aux différents états : il sait combien lui rapporte une machine en fonctionnement, combien lui coûte une réparation, quels sont ses coûts fixes. Connaissant le coût $f(n)$ associé à chaque état n ainsi que sa probabilité stationnaire il peut donc évaluer le coût moyen $\sum f(n)\,\pi_n$. D'après le théorème 1.7, cette quantité est une prédiction de ce que son usine lui rapportera en moyenne par unité de temps. Il peut utiliser cette prédiction, par exemple pour optimiser le nombre d'ateliers de réparation.

2.5 Files à arrivées ou services groupés

La caractéristique des processus de naissance et de mort est d'évoluer par sauts de $+1$ ou -1. Pour un tel modèle, les arrivées et les départs sont

donc individuels. Or dans de nombreuses situations pratiques, cette hypothèse n'est pas réaliste. Pour les arrivées, pensons à l'entrée d'un théâtre, où les spectateurs arrivent en couple ou en famille. Pour les services, prenons un téléphérique, qui monte d'un coup tout un groupe de touristes. Nous allons examiner un modèle d'arrivées groupées, puis un modèle de services groupés.

Le premier est la file $\mathrm{M}^{(X)}/\mathrm{M}/1$. Comme pour la file M/M/1, les temps d'interarrivée sont exponentiels de paramètre λ. Mais à chaque instant d'arrivée, c'est non pas 1, mais un nombre aléatoire X de clients qui arrivent. Les autres hypothèses sont celles de la file M/M/1 : il n'y a qu'un seul serveur, qui sert les clients un par un. Les temps de service sont exponentiels de paramètre μ. Toutes les variables aléatoires considérées (temps d'interarrivée, temps de service et nombres de clients à chaque arrivée) sont indépendantes dans leur ensemble.

Les tailles aléatoires des paquets de clients sont de même loi, et cette loi est supposée connue. On peut supposer sans perte de généralité que les paquets contiennent tous au moins un client ($\mathbb{P}[X = 0] = 0$), et on note q_k la probabilité qu'un paquet contienne k clients. Nous utiliserons aussi la fonction génératrice de la loi de X, notée g.

$$g(z) = \sum_{k=1}^{\infty} z^k \, q_k \ .$$

Le nombre moyen de clients d'un paquet est la dérivée en 1 de la fonction g :

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} k \, q_k = g'(1) \ .$$

Puisqu'il y a en moyenne λ arrivées de paquets par unité de temps et que chaque paquet contient en moyenne $g'(1)$ clients, il y a $\lambda g'(1)$ clients entrant dans le système par unité de temps. On peut donc s'attendre à ce que la condition d'équilibre soit $\lambda g'(1) < \mu$. C'est ce que nous allons vérifier par le calcul.

La variable d'intérêt est encore le nombre total de clients présents dans le système à l'instant t, noté Z_t. Comme pour la file M/M/1, si n clients sont présents dans le système, le prochain événement surviendra au bout d'un temps exponentiel de paramètre $\lambda + \mu$. Ce sera une arrivée avec probabilité $\frac{\lambda}{\lambda+\mu}$, ou un départ avec probabilité $\frac{\mu}{\lambda+\mu}$. Mais si c'est une arrivée, le processus ne saute pas nécessairement vers $n + 1$. Il saute vers $n + k$ avec probabilité $\frac{\lambda q_k}{\lambda+\mu}$. Le processus $\{Z_t \, , \, t \geq 0\}$ est un processus markovien de saut sur $\mathbb{N}$. La figure 5 représente son diagramme de transition.

Posons $p_n(t) = \mathbb{P}[Z_t = n]$. On écrit le système de Chapman-Kolmogorov

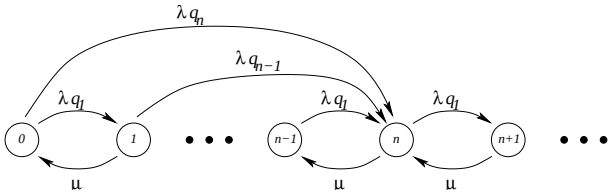

FIGURE 5 – Modèle à arrivées groupées : diagramme de transition de la file $M^{(X)}/M/1$.

à partir du diagramme de transitions.

$$\begin{cases} p_0'(t) = -\lambda\, p_0(t) + \mu\, p_1(t) \\ p_n'(t) = \displaystyle\sum_{j=1}^{n} \lambda q_j\, p_{n-j}(t) - (\lambda+\mu)\, p_n(t) + \mu\, p_{n+1}(t)\ , \quad \forall n \geq 1\ . \end{cases} \tag{2.14}$$

Nous cherchons seulement à déterminer la solution constante, notée $\pi = (\pi_n)_{n\in I\!N}$.

$$\begin{cases} 0 = -\lambda\, \pi_0 + \mu\, \pi_1 \\ 0 = \displaystyle\sum_{j=1}^{n} \lambda q_j\, \pi_{n-j} - (\lambda+\mu)\, \pi_n(t) + \mu\, \pi_{n+1}\ , \quad \forall n \geq 1\ . \end{cases}$$

Pour résoudre ce système, on utilise la technique des fonctions génératrices. Posons donc :

$$f(z) = \sum_{n=0}^{\infty} z^n\, \pi_n\ .$$

On multiplie la n-ième équation par z^n, et on somme pour obtenir :

$$0 = \lambda g(z) f(z) - \lambda f(z) - \mu(f(z) - \pi_0) + \frac{\mu}{z}(f(z) - \pi_0)\ ,$$

soit,

$$f(z) = \pi_0 \frac{\mu(1-z)}{\mu(1-z) - \lambda z(1-g(z))}\ .$$

Pour déterminer π_0, il reste à écrire que la somme des π_n vaut 1, soit $f(1) = 1$. Un développement limité à l'ordre 1 du dénominateur permet de lever l'indétermination : $g(z) = 1 - (1-z)g'(1) + o(1)$. On trouve :

$$\pi_0 = 1 - \frac{\lambda g'(1)}{\mu}\ .$$

Evidemment cette solution n'est acceptable que si $\lambda g'(1) < \mu$, comme nous l'avions prévu. La file $M^{(X)}/M/1$ n'admet de mesure stationnaire que si le

nombre moyen de clients entrant par unité de temps est inférieur à la capacité de service.

Comme cas particulier, si un seul client arrive à chaque instant, alors $g(z) = z$, et on retrouve bien sûr la mesure stationnaire de la file M/M/1 :

$$f(z) = \frac{1 - \frac{\lambda}{\mu}}{1 - \frac{\lambda z}{\mu}} = \sum_{n=0}^{\infty} \left(1 - \frac{\lambda}{\mu}\right) \left(\frac{\lambda}{\mu}\right)^n z^n \, .$$

Comme modèle à services groupés, nous aurions pu choisir des services par paquets de taille aléatoire, et traiter l'analogue du modèle précédent. Le modèle ci-après nous paraît plus proche des applications ; il s'agit de la file $\mathrm{M}/\mathrm{M}^{(a,b)}/1$. Les clients arrivent un par un, selon un processus de Poisson d'intensité λ. Les temps de service sont exponentiels de paramètre μ. Les temps d'interarrivée et les temps de service sont indépendants dans leur ensemble. La discipline de service est la suivante. Si immédiatement après un service, le serveur trouve moins de a clients dans la file, il attend jusqu'à ce qu'il y en ait a, et les prend tous ; s'il en trouve plus de a, mais moins de b, il les prend tous ; s'il en trouve plus de b, il en prend b, pendant que les autres continuent d'attendre. En d'autres termes, chaque service est une "fournée" qui doit contenir au moins a clients, mais ne peut pas en contenir plus de b. La capacité maximale du service est de $b\mu$ clients. La condition d'équilibre sera donc $\lambda < b\mu$, ce que nous allons vérifier mathématiquement.

Ce modèle est sensiblement plus compliqué que les précédents, dans la mesure où le nombre total de clients dans le système n'est pas un processus markovien de saut. Pour comprendre pourquoi, il suffit de considérer un cas particulier. Prenons $a = 2$ et $b = 3$, et supposons qu'à l'instant t il y ait 3 clients dans le système. Cela peut se produire dans deux cas :

– soit deux clients sont au service et un en attente,
– soit trois clients sont au service et personne n'attend.

Mais si le prochain événement est une fin de service, alors après le saut, il restera un client dans le premier cas, personne dans le second.

Pour obtenir un processus markovien de saut, nous allons devoir prendre en compte l'état du serveur. Les états possibles du système seront donc des couples $(0, n)$ ou $(1, n)$.

- $(0, n)$: le serveur est libre et n clients attendent $(n = 0, \ldots, a-1)$
- $(1, n)$: le serveur est occupé et n clients attendent $(n \geq 0)$.

La figure 6 représente le diagramme de transition du processus entre ces états.

On note :

- $p_{0,n}(t)$ la probabilité qu'à l'instant t le serveur soit libre et n clients attendent $(n = 0, \ldots, a-1)$
- $p_{1,n}(t)$: la probabilité qu'à l'instant t le serveur soit occupé et n clients attendent $(n \geq 0)$.

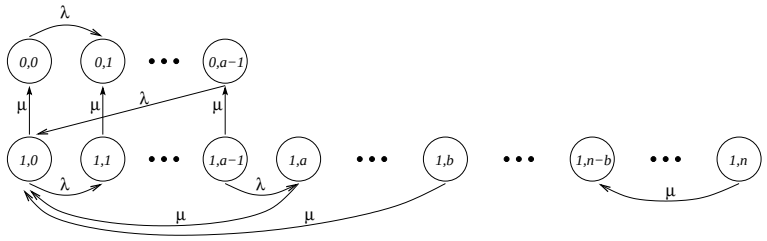

FIGURE 6 – Modèle à services groupés : diagramme de transition de la file $\mathrm{M}/\mathrm{M}^{(a,b)}/1$.

Le système de Chapman-Kolmogorov correspondant au diagramme de transition de la figure 6 est le suivant.

$$\begin{cases} p'_{0,0}(t) = -\lambda\, p_{0,0}(t) + \mu\, p_{1,0}(t) \\ p'_{0,n}(t) = \lambda\, p_{0,n-1}(t) - \lambda\, p_{0,n}(t) + \mu\, p_{1,n}(t)\ , \quad n = 1, \ldots, a-1 \\ p'_{1,0}(t) = \lambda\, p_{0,a-1}(t) - (\lambda+\mu)\, p_{1,0}(t) + \mu \sum_{n=a}^{b} p_{1,n}(t) \\ p'_{1,n}(t) = \lambda\, p_{1,n-1}(t) - (\lambda+\mu)\, p_{1,n}(t) + \mu\, p_{1,n+b}(t)\ , \quad n \geq 1\ . \end{cases} \tag{2.15}$$

Bien entendu, nous ne calculerons que la solution constante de ce système, en notant $\pi_{0,n}$ et $\pi_{1,n}$ les probabilités stationnaires des différents états.

$$\begin{cases} \text{(a)} \quad 0 = -\lambda\, \pi_{0,0} + \mu\, \pi_{1,0} \\ \text{(b)} \quad 0 = \lambda\, \pi_{0,n-1} - \lambda\, \pi_{0,n} + \mu\, \pi_{1,n}\ , \quad n = 1, \ldots, a-1 \\ \text{(c)} \quad 0 = \lambda\, \pi_{0,a-1} - (\lambda+\mu)\, \pi_{1,0} + \mu \sum_{n=a}^{b} \pi_{1,n} \\ \text{(d)} \quad 0 = \lambda\, \pi_{1,n-1} - (\lambda+\mu)\, \pi_{1,n} + \mu\, \pi_{1,n+b}\ , \quad n \geq 1\ . \end{cases} \tag{2.16}$$

Considérons l'équation de récurrence linéaire (d). L'équation caractéristique associée est la suivante.

$$0 = \lambda - (\lambda+\mu)z + \mu z^{b+1}\ . \tag{2.17}$$

Le lemme suivant se démontre par des techniques d'analyse complexe (théorème de Rouché).

Lemme 2.5

1. *Si $\lambda \geq b\mu$, l'équation* (2.17) *n'admet aucune racine de module strictement inférieur à* 1.
2. *Si $\lambda < b\mu$, l'équation* (2.17) *admet une unique racine de module strictement inférieur à* 1. *Cette racine est réelle et strictement comprise entre* 0 *et* 1.

Il est facile de vérifier, en étudiant le graphe du polynôme $\lambda-(\lambda+\mu)z+\mu z^{b+1}$, que (2.17) admet une racine réelle comprise entre 0 et 1 si $\lambda < b\mu$. Nous la noterons r. Remarquons que r peut se calculer numériquement avec une excellente précision (par exemple avec la méthode de Newton).

Rappelons qu'une mesure stationnaire est une loi de probabilité. La solution de l'équation (d) que nous cherchons doit être une série convergente. Or toute solution de (d) est une combinaison linéaire des puissances de racines de l'équation caractéristique associée (2.17). Dans le cas $\lambda \geq b\mu$, les puissances de racines ne peuvent pas être les termes généraux de séries convergentes. Le processus n'admet pas de mesure stationnaire, le système sature. Dans le cas $\lambda < b\mu$ en revanche, la solution que nous cherchons vérifie nécessairement, pour tout $n \geq 0$:

$$\pi_{1,n} = \pi_{1,0}\, r^n \ .$$

Reste à calculer les autres probabilités. Nous les exprimons toutes en fonction de $\pi_{0,0}$ et r. En combinant les équations (a), (b) et (c) de (2.16), on trouve :

$$\pi_{0,n} = \pi_{0,0}\, \frac{1-r^{n+1}}{1-r} \ , \quad n = 1, \ldots, a-1 \ ,$$

$$\pi_{1,n} = \pi_{0,0}\, \frac{(1-r^b)r^{n+1}}{1-r} \ , \quad n \geq 0 \ .$$

On calcule ensuite $\pi_{0,0}$, en écrivant que la somme de toutes les probabilités vaut 1. On trouve :

$$\pi_{0,0} = \left(\frac{a}{1-r} + \frac{r(r^a - r^b)}{(1-r)^2} \right)^{-1} \ .$$

3 Files quasi-markoviennes

Les traitements mathématiques des modèles étudiés jusqu'ici étaient tous basés sur les processus markoviens de saut, et donc sur l'hypothèse de durées exponentielles. Or supposer que des temps d'interarrivée ou des temps de service sont exponentiels n'est pas toujours réaliste dans les applications. L'intérêt des modèles markoviens est que beaucoup des résultats que l'on y démontre sont robustes aux changements de loi, au sens où ils restent vrais quand on remplace les lois exponentielles des services ou des interarrivées par d'autres lois. Pour donner une base concrète à cette intuition, nous allons reprendre le cadre le plus simple de la file M/M/1, en généralisant d'abord la loi des temps de service, ensuite celle des temps d'interarrivée.

3.1 File M/GI/1

Nous commençons par une description d'un modèle moins précis que celui de la file M/M/1. Considérons une file d'attente à un seul serveur. Les clients sont servis un par un, et le temps de service de chaque client est aléatoire, mais de loi inconnue. Le nombre de clients arrivant dans la file pendant le n-ième service est aussi une variable aléatoire, que nous noterons A_n. L'hypothèse essentielle est que les A_n sont indépendantes et de même loi sur $\mathbb{N}$. Nous

notons X_n le nombre de clients présents dans le système immédiatement après le n-ième service. Si X_n est strictement positif, alors $X_{n+1} = X_n - 1 + A_{n+1}$ (un client est parti, et A_{n+1} sont arrivés). Si X_n est nul, alors $X_{n+1} = A_{n+1}$. On peut donc écrire :

$$X_{n+1} = X_n - \mathbb{1}_{I\!N^*}(X_n) + A_{n+1} \ , \tag{3.18}$$

ce qui montre que (X_n) est une chaîne de Markov, à valeurs dans $I\!N$.

Nous noterons $q = (q_k)_{k \in I\!N}$ la loi commune des nombres de clients arrivant pendant un service, g sa fonction génératrice et ρ son espérance.

$$g(z) = \sum_{k=0}^{\infty} z^k \, q_k \ ; \ \rho = \sum_{k=0}^{\infty} k \, q_k = g'(1) \ .$$

Le paramètre ρ est le nombre moyen de clients qui arrivent pendant un temps de service. C'est le coefficient d'occupation de la file. Le comportement asymptotique de la chaîne (X_n) est facile à deviner intuitivement. Si $\rho < 1$, le serveur peut faire face à toutes les demandes : les clients ne s'accumulent pas et un régime d'équilibre peut s'établir ; la chaîne (X_n) est récurrente positive. Si $\rho > 1$, les clients sont trop nombreux et la file sature : X_n croît en moyenne comme $n(\rho - 1)$; la chaîne (X_n) tend presque sûrement vers $+\infty$ et elle est donc transiente. On démontre que la chaîne est récurrente nulle pour $\rho = 1$. Nous donnons ci-après les justifications les plus faciles.

Proposition 3.1 *Si $\rho > 1$, la chaîne (X_n) tend vers l'infini presque sûrement, elle est donc transiente. Si $\rho < 1$ la chaîne est récurrente.*

Démonstration : A partir de la définition (formule (3.18)), on peut écrire immédiatement :

$$X_n \geq -n + \sum_{m=1}^{n} A_m = n\Big(-1 + \frac{1}{n} \sum_{m=1}^{n} A_m \Big) \ .$$

D'après la loi forte des grands nombres, $\frac{1}{n} \sum_{m=1}^{n} A_m$ converge presque sûrement vers ρ, d'où le résultat dans le cas $\rho > 1$.

Si $X_0 = i$, on voit aisément à partir de la même formule (3.18), que $\sum_{m=1}^{n} A_m < n$ entraîne que parmi $X_1, \ldots, X_n$, au moins une des valeurs est égale à i. La probabilité de retour en i est donc minorée par $I\!P[\sum_{m=1}^{n} A_m < n]$. Pour $\rho < 1$, cette probabilité tend vers 1 quand n tend vers l'infini. □

Dans le cas où un régime d'équilibre s'établit, il est possible de calculer explicitement la fonction génératrice de la mesure stationnaire. Le traitement est très proche de celui de la file $\mathrm{M}^{(X)}/\mathrm{M}/1$.

Proposition 3.2 *La chaîne (X_n) admet une mesure stationnaire si et seulement si le coefficient d'occupation ρ est strictement inférieur à 1. Dans ce cas, la fonction génératrice de la mesure stationnaire est :*

$$f(z) = \frac{(1-\rho)(1-z)g(z)}{g(z) - z} \ . \tag{3.19}$$

Démonstration : A partir de la formule de définition (3.18), il est facile d'écrire les probabilités de transition de la chaîne (X_n). La probabilité de transition de 0 à j est q_j pour tout $j \geq 0$, et pour $i > 0$, la probabilité de transition de i à j est q_{j-i+1} si $j \geq i-1$, 0 sinon. Notons $\pi = (\pi_n)$ la mesure stationnaire. Si elle existe, elle vérifie le système d'équations suivant :

$$\begin{cases} \pi_0 = \pi_0\, q_0 + \pi_1\, q_0 \\ \pi_n = \pi_0\, q_n + \pi_1\, q_n + \cdots + \pi_{n+1}\, q_0\ , \quad \forall n \geq 1\ . \end{cases}$$

La fonction génératrice de π est définie par $f(z) = \sum z^n\, \pi_n$. Pour la faire apparaître, on multiplie par z^n la n-ième équation du système et on somme :

$$\begin{aligned} f(z) &= \pi_0\, g(z) + \pi_1\, g(z) + \cdots + \pi_{n+1} z^n\, g(z) + \cdots \\ &= \frac{g(z)}{z}\Big(\pi_0 z - \pi_0 + f(z)\Big)\ . \end{aligned}$$

On en déduit l'expression de $f(z)$ en fonction de π_0 et $g(z)$:

$$f(z) = \pi_0 \frac{(1-z)g(z)}{g(z)-z}\ .$$

Pour déterminer la valeur de π_0, il faut utiliser le fait que π doit être une mesure de probabilité et que donc $f(1)$ doit valoir 1. Or $z = 1$ annule le numérateur et le dénominateur de l'expression ci-dessus. Pour lever l'indétermination, on écrit :

$$g(z) = 1 + (z-1)\rho + o(z-1)\ .$$

On en déduit facilement que $\pi_0 = 1-\rho$. Donc la mesure stationnaire ne peut être une loi de probabilité que si $\rho < 1$. □

De manière plus spécifique, supposons que les temps d'interarrivée soient exponentiels de paramètre λ. Les temps de service ont une loi commune notée G sur $\mathbb{R}^+$. Les temps d'interarrivée et de service sont indépendants dans leur ensemble. La notation de Kendall correspondant à ce modèle est M/GI/1. Nous allons voir que les quantités intéressantes relatives au régime stationnaire (loi du nombre de clients mais aussi du temps de séjour) s'expriment de manière explicite à l'aide de la transformée de Laplace de la loi du temps de service, notée G^*, et définie pour $s > 0$ par :

$$G^*(s) = \int_0^\infty e^{-st}\, dG(t)\ .$$

Par exemple, pour la file M/M/1, la durée de service est exponentielle de paramètre μ, la transformée de Laplace de cette loi est $G^*(s) = \frac{\mu}{\mu+s}$.

Nous utiliserons les deux premiers moments de la loi G, qui s'expriment à l'aide des dérivées de G^* en 0. L'espérance d'un temps de service est :

$$\int_0^\infty t\, dG(t) = -G^{*\prime}(0)\ .$$

Pour conserver des notations homogènes avec la file M/M/1, nous noterons $\frac{1}{\mu}$ cette espérance. La dérivée seconde de G^* en 0 est l'espérance du carré du temps de service :

$$\int t^2 \, dG(t) = G^{*\prime\prime}(0) \; .$$

Nous noterons v la variance d'un temps de service :

$$v = G^{*\prime\prime}(0) - \frac{1}{\mu^2} \; .$$

Les résultats qui suivent sont dus à Pollaček et Khintchine.

Proposition 3.3 *Considérons la file d'attente M/GI/1, et supposons-la en régime stationnaire. Soit f la fonction génératrice du nombre de clients après un service, W la loi du temps de séjour d'un client, et W^* sa transformée de Laplace. Alors :*

1.
$$f(z) = \frac{(1-\rho)(1-z)G^*(\lambda - \lambda z)}{G^*(\lambda - \lambda z) - z} \; . \tag{3.20}$$

2.
$$W^*(s) = \frac{s(1-\rho)G^*(s)}{s - \lambda(1 - G^*(s))} \; . \tag{3.21}$$

3.
$$I\!E[W] = \frac{1}{\mu} + \frac{\lambda}{2(1-\rho)} \left(v + \frac{1}{\mu^2} \right) \; . \tag{3.22}$$

Démonstration : Le processus d'arrivée est un processus de Poisson d'intensité λ. Le nombre d'arrivées pendant un intervalle de temps d'amplitude t suit donc la loi de Poisson de paramètre λt. La probabilité que k clients arrivent pendant cet intervalle est donc $e^{-\lambda t} \frac{(\lambda t)^k}{k!}$. Pour obtenir la probabilité que k clients arrivent pendant un temps de service, il faut intégrer cette expression par rapport à la loi G :

$$q_k = \int_0^\infty e^{-\lambda t} \frac{(\lambda t)^k}{k!} \, dG(t) \; .$$

La fonction génératrice de la loi q s'écrit donc :

$$\begin{aligned} g(z) &= \sum_{k=0}^{\infty} z^k \, q_k \\ &= \int_0^\infty \left(\sum_{k=0}^{\infty} e^{-\lambda t} \frac{(\lambda t)^k z^k}{k!} \right) dG(t) \\ &= \int_0^\infty e^{-\lambda t + \lambda z t} \, dG(t) \\ &= G^*(\lambda - \lambda z) \; . \end{aligned}$$

On déduit donc (3.20) de (3.19).

Par la même technique, on peut relier la transformée de Laplace du temps de séjour d'un client arrivant en régime stationnaire à la fonction génératrice de la mesure stationnaire f. Considérons la probabilité π_n. Par définition, c'est la probabilité de l'événement "à la fin d'un service en régime stationnaire, il reste n clients dans le système". Mais comme la discipline de service est FIFO, ces n clients sont exactement ceux qui sont arrivés pendant le temps de séjour du client qui vient de partir. Notons W la loi de ce temps de séjour. On a :

$$\pi_n = \int_0^\infty e^{-\lambda t} \frac{(\lambda t)^n}{n!}\, dW(t) \ .$$

Et donc par le même calcul que précédemment :

$$f(z) = \sum_{n=0}^{\infty} z^n\, \pi_n = W^*(\lambda - \lambda z) \ ,$$

en notant W^* la transformée de Laplace de la loi W. En revenant à l'expression de $f(z)$ (formule (3.20)), on obtient l'expression suivante :

$$W^*(\lambda - \lambda z) = \frac{(1-\rho)(1-z)G^*(\lambda - \lambda z)}{G^*(\lambda - \lambda z) - z} \ .$$

Soit (3.21), en remplaçant formellement $\lambda - \lambda z$ par s.

Le coefficient d'occupation ρ s'exprime simplement. Nous l'avons défini comme le nombre moyen de clients arrivant pendant la durée d'un service.

$$\rho = \sum_{k=0}^{\infty} k\, q_k = g'(1) = -\lambda G^{*\prime}(0) \ .$$

Or $-G^{*\prime}(0) = \frac{1}{\mu}$ est l'espérance d'un temps de service. On a alors, comme pour la file M/M/1 :

$$\rho = \frac{\lambda}{\mu} \ .$$

Le nombre moyen de clients présents à la fin d'un service en régime stationnaire est l'espérance de la loi π, à savoir $f'(1)$. De même le temps moyen de séjour d'un client arrivant en régime stationnaire est l'espérance de la loi W, soit $-W^{*\prime}(0)$. Comme $f(z) = W^*(\lambda - \lambda z)$, on a $f'(1) = -\lambda W^{*\prime}(0)$. On retouve donc la formule de Little : le nombre moyen de clients dans le système est le produit du nombre moyen de clients arrivant par unité de temps par le temps moyen de séjour. L'expression du temps moyen de séjour (3.22) s'obtient en dérivant (3.21).

On constate que le temps moyen de séjour est minimal lorsque la variance des temps de service est nulle, à savoir lorsque ceux-ci sont constants : pour optimiser une file d'attente, on a donc intérêt à standardiser les temps de service. □

Comme cas particulier, si G est la loi exponentielle de paramètre μ, sa transformée de Laplace est $G^*(s) = \frac{\mu}{s+\mu}$. La fonction génératrice de la mesure stationnaire est :

$$f(z) = \frac{\mu - \lambda}{\mu - \lambda z} \ .$$

La transformée de Laplace du temps de séjour est :

$$W^*(s) = \frac{\mu - \lambda}{\mu - \lambda + s} \ .$$

Ce sont bien les résultats que nous avions trouvé pour la file M/M/1.

3.2 File GI/M/1

Nous commençons par une description assez peu réaliste du modèle. On considère une file d'attente à un seul serveur. Les clients arrivent un par un. Lors de la n-ième arrivée, on donne au serveur un quota aléatoire D_n de clients à servir. Si moins de D_n clients sont présents, ils sont tous servis, sinon D_n sont servis. Les quotas D_n sont des variables aléatoires indépendantes et de même loi $q = (q_k)$ sur $\mathbb{N}$. La fonction génératrice de cette loi est notée g. Le nombre de clients présents dans le système immédiatement avant la n-ième arrivée est noté X_n. Il est défini par :

$$X_{n+1} = \max\{0, X_n + 1 - D_n\} \ , \tag{3.23}$$

donc (X_n) est une chaîne de Markov.

Notons ρ le coefficient d'occupation de la file, à savoir le rapport du nombre d'arrivées au nombre de services par unité de temps. Comme D_n est un nombre de services pour une arrivée, son espérance doit être $1/\rho$.

$$\sum_{k=1}^{\infty} k\, q_k = g'(1) = \frac{1}{\rho} \ .$$

Le comportement asymptotique de la chaîne est toujours le même : si $\rho \geq 1$ la file sature, si $\rho < 1$ un régime d'équilibre peut s'établir. Ici encore, nous ne proposons que les démonstrations les plus faciles.

Proposition 3.4 *Si $\rho > 1$, la chaîne (X_n) tend vers l'infini presque sûrement, elle est donc transiente. Si $\rho < 1$ la chaîne est récurrente.*

Démonstration : A partir de la définition (formule (3.23)), on vérifie immédiatement que pour tout $n \geq 1$, $X_n \geq X_0 + n - (D_1 + \cdots + D_n)$. Par la loi des grands nombres $(D_1 + \cdots + D_n)/n$ converge vers $1/\rho$. Si $\rho > 1$, $n - (D_1 + \cdots + D_n)$ tend vers $+\infty$ presque sûrement, donc la chaîne est transiente. Inversement, supposons $\rho < 1$. Le même raisonnement indique que

l'événement "$D_1 + \cdots + D_n > n$" a une probabilité qui tend vers 1 quand n tend vers l'infini. Or si $X_0 = 0$, cet événement implique qu'il existe $m \leq n$ tel que $X_m = 0$. La probabilité de retour en 0 tend donc vers 1 et la chaîne est récurrente. □

Dans le cas où un régime d'équilibre s'établit, il est possible de calculer explicitement la mesure stationnaire.

Proposition 3.5 *La chaîne (X_n) admet une mesure stationnaire si et seulement si le coefficient d'occupation ρ est strictement inférieur à 1. Dans ce cas, l'équation $g(z) = z$ admet une unique solution de module inférieur à 1. Cette solution, notée r, est réelle et strictement comprise entre 0 et 1. La mesure stationnaire $\pi = (\pi_n)_{n\in\mathbb{N}}$ est définie par :*

$$\pi_n = (1 - r)r^n \ . \tag{3.24}$$

Démonstration : A partir de la formule de définition (3.18), il est facile d'écrire les probabilités de transition de la chaîne. La probabilité de transition de i vers $j = 1, \ldots, i + 1$ est q_{i-j+1}. La probabilité de transition de i vers 0 est $\sum_{k=i+1}^{\infty} q_k$. Si la mesure stationnaire existe, elle vérifie pour tout $n > 0$ l'équation suivante.

$$\pi_n = \sum_{k=0}^{\infty} q_k \, \pi_{n+k-1} \ .$$

Cette équation de récurrence linéaire ne peut admettre comme solution une loi de probabilité que si l'équation caractéristique associée admet une racine de module strictement inférieur à 1. Or l'équation caractéristique associée est $g(z) = z$. Il est facile de vérifier, en étudiant le graphe de la fonction, qu'elle admet une racine réelle strictement comprise entre 0 et 1 (notée r) si et seulement si $g'(1) > 1$, soit $\rho < 1$. En utilisant le théorème de Rouché, on démontre qu'il n'y a pas de racine de module < 1 si $\rho \geq 1$, et que r est la seule racine de module < 1 si $\rho < 1$. Donc il n'existe pas de mesure stationaire si $\rho \geq 1$ (la file sature), et si $\rho < 1$, la mesure stationnaire est donnée par (3.24). □

Voici une définition plus précise du modèle GI/M/1. Les clients arrivent un par un et leurs arrivées sont séparées par des durées indépendantes et de même loi, notée G. Ils sont servis un par un et les services sont exponentiels de paramètre μ. Les temps d'interarrivée et de service sont indépendants dans leur ensemble. Le "quota" de clients que le serveur peut traiter pendant un temps d'interarrivée est le nombre de clients qu'il traiterait s'il fonctionnait en continu. Si c'était le cas, le nombre de clients servis au cours du temps serait un processus de Poisson d'intensité μ. Sur un intervalle de temps d'amplitude t, la probabilité de traiter k clients serait $e^{-\mu t}\frac{(\mu t)^k}{k!}$. Pour obtenir la probabilité d'un quota de k clients, il faut intégrer cette expression par rapport à la

loi G :

$$q_k = \int_0^\infty e^{-\mu t} \frac{(\mu t)^k}{k!}\, dG(t) \ .$$

La fonction génératrice de la loi q s'écrit donc :

$$\begin{aligned} g(z) &= \sum_{k=0}^{\infty} z^k\, q_k \\ &= \int_0^\infty \sum_{k=0}^{\infty} e^{-\mu t} \frac{(\mu t)^k z^k}{k!}\, dG(t) \\ &= \int_0^\infty e^{-\mu t + \mu z t}\, dG(t) \\ &= G^*(\mu - \mu z) \ , \end{aligned}$$

en notant G^* la transformée de Laplace de la loi G. Remarquons que $g'(1) = -\mu G^{*\prime}(0)$. Si l'espérance d'un temps d'interarrivée est $-G^{*\prime}(0) = \frac{1}{\lambda}$, alors la taille moyenne d'un quota de clients est $\frac{1}{\rho} = \frac{\mu}{\lambda}$.

La détermination de la durée de séjour d'un client dans le système en régime stationnaire se fait exactement de la même façon que pour la file M/M/1.

Proposition 3.6 *Supposons que la chaîne (X_n) soit en régime stationnaire. Notons T la durée de séjour d'un client. La loi de T est la loi exponentielle de paramètre $\mu(1-r)$, où r est la racine de $z = g(z)$ strictement comprise entre 0 et 1.*

Démonstration : elle est identique à celle de la proposition 2.1. La probabilité que le client qui arrive trouve n clients arrivés avant lui dans le système est $\pi_n = (1-r)r^n$. Si c'est le cas, son temps de séjour sera la somme de $n+1$ temps de service indépendants, exponentiels de paramètre μ. La loi conditionnelle de T sachant qu'il y a n clients dans le système est donc la loi d'Erlang ou loi gamma $\mathcal{G}(n+1, \mu)$. La densité de T peut donc s'écrire :

$$\begin{aligned} f_T(t) &= \sum_{n=0}^{\infty} \pi_n \frac{\mu^{n+1} t^n}{n!} e^{-\mu t}\, \mathbb{1}_{I\!R^+}(t) \\ &= \sum_{n=0}^{\infty} (1-r)\,(r)^n\, \frac{\mu^{n+1} t^n}{n!} e^{-\mu t}\, \mathbb{1}_{I\!R^+}(t) \\ &= \mu(1-r) e^{-\mu(1-r)t}\, \mathbb{1}_{I\!R^+}(t) \ . \end{aligned}$$

□

Remarquons que l'espérance du temps de séjour est $\frac{1}{\mu(1-r)}$, alors que le nombre moyen de clients dans le système est $\sum n\pi_n = \frac{r}{1-r}$: la formule de Little ne s'applique pas dans ce cas. Ceci est dû au fait que l'on n'observe la file qu'à certains instants particuliers, juste avant une arrivée.

4 Réseaux de files d'attente

Pour donner une petite idée de la complexité des problèmes d'attente liés à l'informatique, examinons le trajet d'une tâche dans une unité centrale. Son traitement peut être interrompu soit par un défaut de pagination, auquel cas elle sera envoyée dans le tambour de pagination, soit par une entrée-sortie, auquel cas elle se placera dans une mémoire tampon. Elle reviendra plus tard dans l'unité centrale pour poursuivre son traitement. Si on voit cette tâche comme un client qui doit recevoir un certain service, ce sont trois files d'attente (l'unité centrale, le tambour de pagination et la mémoire tampon) qu'elle visite peut-être un grand nombre de fois chacune avant de terminer son exécution. Les réseaux de file d'attente sont des modèles décrivant le parcours d'un client dans un système où il doit visiter successivement plusieurs files avant de sortir. Nous traiterons d'abord les modèles markoviens les plus simples qui sont les réseaux de Jackson, avant de présenter le cadre plus général des réseaux de Petri.

4.1 Réseaux de Jackson ouverts

Un réseau de Jackson ouvert se compose de K files d'attente, que les clients parcourent de manière aléatoire. La description du modèle est la suivante. Elle est illustrée par la figure 7.

1. Les clients arrivent un par un dans le système, selon un processus de Poisson d'intensité λ.
2. Un client qui arrive se dirige vers la file numéro i avec probabilité $p_{0,i}$. Ces probabilités vérifient :

$$\sum_{i=1}^{K} p_{0,i} = 1 \ .$$

3. Les clients sont servis un par un dans chaque file. Le i-ième serveur fonctionne de manière markovienne, avec un taux de service $\mu_i(n_i)$ qui dépend du nombre n_i de clients présents dans la file. Autrement dit, lorsque n_i clients sont présents dans la file, le temps de service d'un client est exponentiel de paramètre $\mu_i(n_i)$. Ceci permet en particulier de prendre en compte des serveurs multiples du type M/M/s ou M/M/∞.

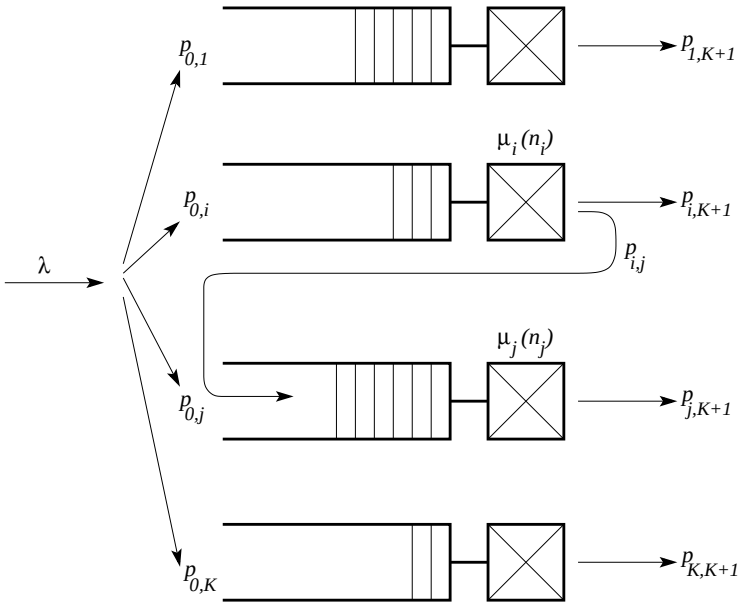

FIGURE 7 – Schéma d'un réseau de Jackson ouvert.

4. Un client sortant de la i-ième file se dirige soit vers la file numéro j avec probabilité $p_{i,j}$, soit vers l'extérieur (il quitte définitivement le système), avec probabilité $p_{i,K+1}$. Pour tout $i = 1, \ldots, K$, ces probabilités, dites de "routage", vérifient :
$$\sum_{j=1}^{K+1} p_{i,j} = 1 \, .$$

5. Les temps d'interarrivée, les temps de service et les choix aléatoires des clients en entrée et en sortie de chaque file, sont des variables aléatoires indépendantes dans leur ensemble.

Les files étant numérotées de 1 à K, un client arrive de l'extérieur, noté comme une 0-ième file, parcourt un certain nombre de files de manière aléatoire, puis se dirige vers l'extérieur, noté comme une $(K+1)$-ième file. La suite des files parcourues par un client est une chaîne de Markov sur $\{0, \ldots, K+1\}$, partant de 0 et telle que $K+1$ est un état absorbant. Nous supposons que cette file n'a pas d'autres composantes irréductibles récurrentes, c'est-à-dire qu'un client qui rentre dans le système ne peut pas y rester bloqué : il en sortira forcément.

L'état du système sera codé par un vecteur d'entiers $\mathbf{n} = (n_1, \ldots, n_k)$, dont la i-ième coordonnée représente le nombre de clients présents dans la file numéro i. Les hypothèses du modèle pemettent de décrire l'évolution de

l'état du système comme un processus markovien de saut, à valeurs dans $\mathbb{N}^K$. Comme pour un processus de naissance et de mort, les sauts vont d'un état vers un état voisin. Ils sont de trois types.

1. *Arrivée* : un client est arrivé de l'extérieur dans la file numéro i ; la i-ième coordonnée du vecteur d'état est augmentée de 1. Nous notons $a(\mathbf{n}, i)$ le nouveau vecteur d'état :

$$a(\mathbf{n}, i) = (n_1, \ldots, n_i + 1, \ldots, n_k) \; .$$

   La transition de $\mathbf{n}$ vers $a(\mathbf{n}, i)$ se produit avec un taux égal à $\lambda p_{0,i}$.

2. *Départ* : un client a quitté la file numéro i pour se diriger vers l'extérieur ; la i-ième coordonnée du vecteur d'état est diminuée de 1. Nous notons $b(\mathbf{n}, i)$ le nouveau vecteur d'état :

$$b(\mathbf{n}, i) = (n_1, \ldots, n_i - 1, \ldots, n_k) \; .$$

   La transition de $\mathbf{n}$ vers $b(\mathbf{n}, i)$ se produit avec un taux égal à $\mu_i(n_i) p_{i,K+1}$.

3. *Changement de file* : un client a quitté la file numéro i pour se diriger vers la file numéro j ; la i-ième coordonnée du vecteur d'état est diminuée de 1, la j-ième est augmentée de 1. Nous notons $c(\mathbf{n}, i, j)$ le nouveau vecteur d'état :

$$c(\mathbf{n}, i, j) = (n_1, \ldots, n_i - 1, \ldots, n_j + 1, \ldots, n_k) \; .$$

   La transition de $\mathbf{n}$ vers $c(\mathbf{n}, i, j)$ se produit avec un taux égal à $\mu_i(n_i) p_{i,j}$.

Notons $p_{\mathbf{n}}(t)$ la probabilité que le système soit dans l'état $\mathbf{n}$ à l'instant t. En examinant les transitions qui conduisent à l'état $\mathbf{n}$, on écrit l'équation de Chapman-Kolmogorov relative à cet état.

$$\begin{aligned} p'_{\mathbf{n}}(t) = & \sum_{i=1}^{K} \mu_i(n_i + 1) p_{i,K+1} \, p_{a(\mathbf{n},i)}(t) \\ & + \sum_{i=1}^{K} \lambda p_{0,i} \, p_{b(\mathbf{n},i)}(t) \\ & + \sum_{i \neq j = 1}^{K} \mu_j(n_j + 1) p_{j,i} \, p_{c(\mathbf{n},i,j)}(t) \\ & - \left(\lambda + \sum_{i=1}^{K} \mu_i(n_i)(1 - p_{i,i}) \right) p_{\mathbf{n}}(t) \; . \end{aligned} \tag{4.25}$$

Cette équation est valable pour tout état $\mathbf{n}$, à condition de poser $\mu_i(0) = 0$, et $p_{\mathbf{n}}(t) = 0$ dès que $\mathbf{n}$ a une coordonnée négative.

Il n'est évidemment pas question de chercher une solution générale à l'équation (4.25). Il est déjà assez étonnant que l'on puisse déterminer explicitement la mesure stationnaire. Elle est donnée par le *théorème de Jackson*.

Théorème 4.1 *Soit $(e_i)_{i=1,\dots,K}$ la solution du système linéaire suivant :*

$$\forall i = 1, \dots, K \ , \quad e_i = p_{0,i} + \sum_{j=1}^{K} e_j p_{j,i} \ . \tag{4.26}$$

Pour tout $i = 1, \dots, K$, posons $\alpha_{i,0} = 1$ et pour tout $n \geq 1$, notons $\alpha_{i,n}$ le produit :

$$\alpha_{i,n} = \prod_{m=1}^{n} \frac{\lambda e_i}{\mu_i(m)} \ . \tag{4.27}$$

Le système est récurrent positif si et seulement si pour tout $i = 1, \dots, K$, $\alpha_{i,n}$ est le terme général d'une série convergente. Si c'est le cas, notons $\pi_{\mathbf{n}}$ la mesure stationnaire de l'état $\mathbf{n} = (n_i)$. Elle s'exprime en fonction de la mesure stationnaire de l'état $\mathbf{0}$ où toutes les files sont vides, de la façon suivante.

$$\pi_{\mathbf{n}} = \pi_{\mathbf{0}} \prod_{i=1}^{K} \alpha_{i,n_i} \ , \tag{4.28}$$

avec :

$$\pi_{\mathbf{0}} = \left(\sum_{\mathbf{n} \in I\!N^K} \prod_{i=1}^{K} \alpha_{i,n_i} \right)^{-1} \ .$$

Nous allons donner une interprétation concrète de ce théorème. Si le système est en équilibre, alors chacune des K files doit également être en équilibre, au sens où le nombre de clients qui y rentrent par unité de temps doit être égal en moyenne au nombre de clients qui en sortent. Les équations (4.26) doivent être comprises comme des équations d'équilibre de flux : λe_i est le nombre moyen de clients qui passent dans la file numéro i par unité de temps. Multiplions la i-ième équation par λ :

$$\lambda e_i = \lambda p_{0,i} + \sum_{j=1}^{K} \lambda e_j p_{j,i} \ .$$

Le nombre de clients qui entrent dans la file numéro i est la somme du nombre de clients venant de l'extérieur, et des nombres de clients sortant des autres files et qui se dirigent vers i. Or il entre λ clients dans le système par unité de temps. Parmi ceux-là une proportion $p_{0,i}$ se dirige vers la file numéro i. Il arrive donc dans cette file en moyenne $\lambda p_{0,i}$ clients venant de l'extérieur. Si les files sont en équilibre, il sort λe_j clients de la file numéro j, parmi lesquels une proportion $p_{j,i}$ se dirige vers la file numéro i. Il entre donc $\lambda e_j p_{j,i}$ clients venant de la file j. La somme de tous ces nombres doit être le flux de la file numéro i, soit λe_i. L'expression de la mesure stationnaire sous forme produit est très particulière : la probabilité d'un état est le produit des probabilités de ses coordonnées. Tout se passe donc comme si à l'équilibre les K files

fonctionnaient de manière indépendante. De plus la probabilité qu'il y ait n clients dans la i-ième file, proportionnelle à $\alpha_{i,n}$, est la même que pour un processus de naissance et de mort, de taux de naissance λe_i, et taux de mort $\mu_i(n)$. Donc la mesure stationnaire donnée par le théorème 4.1 est la loi d'un K-uplet de files d'attentes indépendantes, telles que les clients arrivent dans la i-ième file selon un processus de Poisson de paramètre λe_i, et y sont servis avec un taux $\mu_i(n)$.

Démonstration : Nous nous contenterons de vérifier que la mesure stationnaire proposée dans le théorème 4.1 est bien solution constante de l'équation de Chapman-Kolmogorov (4.25). Commençons par écrire l'équation que doit vérifier la mesure stationnaire, relative à l'état $\mathbf{n}$, en la divisant par $\pi_{\mathbf{n}}$.

$$\begin{aligned}\left(\lambda + \sum_{i=1}^{K} \mu_i(n_i)(1-p_{i,i})\right) &= \sum_{i=1}^{K} \mu_i(n_i+1)p_{i,K+1}\,\frac{\pi_{a(\mathbf{n},i)}}{\pi_{\mathbf{n}}} \\ &+ \sum_{i=1}^{K} \lambda p_{0,i}\,\frac{\pi_{b(\mathbf{n},i)}}{\pi_{\mathbf{n}}} \\ &+ \sum_{i\neq j=1}^{K} \mu_j(n_j+1)p_{j,i}\,\frac{\pi_{c(\mathbf{n},i,j)}}{\pi_{\mathbf{n}}}\ .\end{aligned} \tag{4.29}$$

Etant donnée la forme produit de la mesure stationnaire proposée, les rapports de probabilités se simplifient notablement. A partir des expressions (4.27) et (4.28), on trouve :

$$\frac{\pi_{a(\mathbf{n},i)}}{\pi_{\mathbf{n}}} = \frac{\lambda e_i}{\mu_i(n_i+1)}\ , \quad \frac{\pi_{b(\mathbf{n},i)}}{\pi_{\mathbf{n}}} = \frac{\mu_i(n_i)}{\lambda e_i}\ , \quad \frac{\pi_{c(\mathbf{n},i,j)}}{\pi_{\mathbf{n}}} = \frac{\lambda e_j}{\mu_j(n_j+1)}\frac{\mu_i(n_i)}{\lambda e_i}\ .$$

Donc :

$$\begin{aligned}\left(\lambda + \sum_{i=1}^{K} \mu_i(n_i)(1-p_{i,i})\right) &= \sum_{i=1}^{K} \lambda e_i\, p_{i,K+1} \\ &+ \sum_{i=1}^{K} \frac{\mu_i(n_i)}{e_i} p_{0,i} \\ &+ \sum_{i\neq j=1}^{K} \mu_i(n_i)\frac{e_j}{e_i} p_{j,i}\ .\end{aligned}$$

En reportant les termes en $\mu_i(n_i)p_{i,i}$ dans le membre de droite on obtient :

$$\lambda + \sum_{i=1}^{K} \mu_i(n_i) = \sum_{i=1}^{K} \lambda e_i p_{i,K+1} + \sum_{i=1}^{K} \frac{\mu_i(n_i)}{e_i}\left(p_{0,i} + \sum_{j=1}^{K} p_{j,i}e_j\right)\ .$$

Dans la dernière parenthèse, on reconnaît le membre de droite de l'équation d'équilibre des flux (4.26). Les termes en $\mu_i(n_i)$ se simplifient donc et il reste :

$$\lambda = \sum_{i=1}^{K} \lambda e_i\, p_{i,K+1}\ . \tag{4.30}$$

Cette dernière équation exprime l'équilibre global du système : λ est le nombre moyen de clients entrant par unité de temps, λe_i est le nombre moyen de clients qui sortent de la file numéro i, parmi lesquels une proportion $p_{i,K+1}$ sort définitivement du système. Pour vérifier (4.30), il suffit de sommer toutes les équations de (4.26). On obtient :

$$\sum_{i=1}^{K} p_{0,i} = \sum_{i=1}^{K} \left(e_i - \sum_{j=1}^{K} e_j p_{j,i} \right) ,$$

soit en intervertissant les deux sommes dans le membre de droite :

$$1 = \sum_{i=1}^{K} e_i \left(1 - \sum_{j=1}^{K} p_{i,j} \right) = \sum_{i=1}^{K} e_i \, p_{i,K+1} .$$

□

4.2 Réseaux de Jackson fermés

On simplifie le modèle précédent en supprimant les échanges avec l'extérieur : aucun client ne rentre dans le réseau ($\lambda = 0$) ni n'en sort ($p_{i,K+1} = 0$). Le nombre total de clients reste donc fixe, et nous le notons N. Nous justifierons intuitivement plus loin par un exemple qu'un réseau ouvert à capacité limitée peut être remplacé par un réseau fermé, ce qui légitime le modèle des réseaux de Jackson fermés. Par rapport au paragraphe précédent, on définit maintenant un processus markovien de saut sur un ensemble *fini* d'états.

$$E = \left\{ \mathbf{n} = (n_i) \in \mathbb{N}^K \, ; \, \sum_{i=1}^{K} n_i = N \right\} .$$

Même s'il est fini, le nombre d'états peut être très grand. Le cardinal de E vaut en effet :

$$|E| = \binom{N+K-1}{K-1} ,$$

qui est de l'ordre de N^{K-1} quand N tend vers l'infini. Nous examinerons plus loin les conséquences algorithmiques de cette explosion combinatoire.

La suite des files visitées par un client est une chaîne de Markov, dont nous supposons qu'elle est irréductible : chaque file peut être visitée à partir de n'importe quelle autre. Le théorème suivant, également dû à Jackson, donne la mesure stationnaire du réseau.

Théorème 4.2 *Soit* $(e_i)_{i=1,\dots,K}$ *une solution du système suivant.*

$$\forall i = 1, \dots, K , \quad e_i = \sum_{j=1}^{K} e_j \, p_{j,i} . \tag{4.31}$$

Pour $i = 1, \ldots, K$, posons $\alpha_{i,0} = 1$ et pour $n \geq 1$, notons $\alpha_{i,n}$ le produit :

$$\alpha_{i,n} = \prod_{m=1}^{n} \frac{e_i}{\mu_i(m)} \ . \tag{4.32}$$

Le réseau de Jackson fermé admet pour mesure stationnaire la mesure $\pi = (\pi_{\mathbf{n}})_{\mathbf{n} \in E}$, définie pour $\mathbf{n} = (n_i) \in E$ par :

$$\pi_{\mathbf{n}} = \gamma(N, K)^{-1} \prod_{i=1}^{K} \alpha_{i,n_i} \ , \tag{4.33}$$

avec :

$$\gamma(N, K) = \sum_{\mathbf{n} \in E} \prod_{i=1}^{K} \alpha_{i,n_i} \ . \tag{4.34}$$

On vérifie que la mesure stationnaire π ainsi définie est solution constante des équations de Chapman-Kolmogorov, tout comme dans la démonstration du théorème 4.1.

L'interprétation est légèrement différente. Le système linéaire (4.31) est celui que vérifie la mesure stationnaire d'une chaîne de Markov de matrice de transition $(p_{i,j})$. Ses solutions sont déterminées à une constante près, qui s'élimine dans l'expression de $\pi_{\mathbf{n}}$ avec la constante de normalisation $\gamma(N, K)$. Nous pouvons donc arbitrairement décider que (e_i) est une loi de probabilité sur $\{1, \ldots, K\}$: c'est la mesure stationnaire de la chaîne de Markov des visites d'un client. A l'équilibre, e_i représente la proportion des N clients qui visitent la file numéro i. Comme chacune des files doit être en équilibre, le nombre de clients qui entrent dans la file i par unité de temps et le nombre de clients qui en sortent doivent être égaux, et proportionnels à e_i. L'équation relative à la file i est encore un équilibre de flux : les clients entrant viennent d'une autre file j, et se dirigent vers i avec probabilité $p_{j,i}$. Une fois que les e_i ont été fixés, la mesure stationnaire de chaque file est proportionnelle aux $(\alpha_{i,n})$: c'est la mesure stationnaire d'un processus de naissance et de mort de taux de naissance constants égaux à e_i et de taux de mort $\mu_i(n)$. A l'équilibre, tout se passe comme si les K files étaient indépendantes. Chacune des files a une capacité limitée à N. Pour la i-ième file, les arrivées se font selon un processus de Poisson d'intensité e_i, le taux de service est $\mu_i(n)$.

A titre d'exemple, considérons un processus à deux files, avec bouclage : les clients sortant de la file 1 se dirigent forcément vers la file 2, et réciproquement. Les probabilités de routage $p_{i,j}$ sont donc les suivantes.

$$p_{1,1} = 0 \ , \quad p_{1,2} = 1 \ , \quad p_{2,1} = 1 \ , \quad p_{2,2} = 0 \ .$$

Nous supposons que les taux de service des deux files sont constants, μ pour la première, et λ pour la seconde. Les états du système sont les couples

$(n, N-n)$, pour $n = 0, \ldots, N$. Nous noterons π_n la probabilité stationnaire de l'état $(n, N-n)$. En appliquant le théorème 4.2, on trouve :

$$\pi_n = \gamma(N,2)^{-1}\mu^{-n}\lambda^{n-N} = \frac{1-\frac{\lambda}{\mu}}{1-(\frac{\lambda}{\mu})^{N+1}}\left(\frac{\lambda}{\mu}\right)^n \ .$$

C'est exactement la mesure stationnaire d'une file M/M/1/N (voir paragraphe 2.4). On peut généraliser cet exemple de la façon suivante. Considérons un réseau de Jackson ouvert, de capacité limitée à N clients au total. Ajoutons-lui une $(K+1)$-ième file, de taux de service λ constant. Les clients sortant de cette $(K+1)$-ième file se dirigent vers la i-ième avec probabilité $p_{0,i}$. A un état $\mathbf{n} = (n_1, \ldots, n_K)$ du système initial, correspond l'état $(n_1, \ldots, n_K, N - \sum n_i)$ du nouveau système. Les probabilités stationnaires de ces états pour leurs systèmes respectifs sont les mêmes. Dans les applications, les capacités infinies n'existent pas. Ceci justifie donc l'étude des réseaux fermés, comme alternative aux réseaux ouverts à capacité limitée.

Nous revenons maintenant sur le calcul algorithmique de la constante de normalisation $\gamma(N, K)$. Calculer directement les $|E|$ termes et les sommer est impensable, pour peu que N et K soient un peu grands. Une astuce à base de fonctions génératrices conduit à un algorithme dont le nombre d'opérations est d'ordre N^2, au lieu de N^{K-1} pour une application naïve de la formule (4.34). L'idée consiste à introduire une fonction génératrice g_i, pour la mesure stationnaire de la i-ième file.

$$g_i(z) = \sum_{n=0}^{N} z^n \, \alpha_{i,n} \ .$$

Définissons ensuite les h_i comme les produits cumulés des g_i :

$$h_i(z) = g_1(z) \cdots g_i(z) \ .$$

Observons que la constante $\gamma(N, K)$ est le coefficient de z^N dans le développement de h_K en série entière. Plus généralement, pour tout $n = 1, \ldots, N$ et pour tout $j = 1, \ldots, K$, $\gamma(n, j)$ est le terme en z^n dans le développement de h_j en série entière. Mais la relation $h_j = h_{j-1}g_j$ se traduit sur les développements en série entière par la relation de récurrence suivante.

$$\gamma(n,j) = \sum_{m=0}^{n} \gamma(m, j-1)\alpha_{j,n-m} \ . \tag{4.35}$$

Cette récurrence est initialisée par :

$$\gamma(n,1) = \alpha_{1,n} \quad \text{et} \quad \gamma(0,j) = 1 \ .$$

L'algorithme itératif qui implémente cette récurrence pour le calcul de $\gamma(N, K)$ nécessite $N(N+1)K$ opérations en virgule flottante (flops) pour le calcul des $\alpha_{i,n}$, et autant pour la relation (4.35) elle-même.

On peut utiliser le même type d'astuce pour calculer d'autres quantités relatives à la mesure stationnaire, comme par exemple le coefficient d'occupation du serveur numéro i, noté $\rho_i(N,K)$.

$$\rho_i(N,K) = \sum_{\substack{\mathbf{n}\in E \\ n_i>0}} \pi_{\mathbf{n}} \ .$$

Avec les notations précédentes, considérons la fonction $\widetilde{g_i}$ définie par :

$$\widetilde{g_i} = h_K(z)\frac{g_i(z)-1}{g_i(z)} = g_1(z)\cdots(g_i(z)-1)\cdots g_K(z) \ .$$

Le coefficient de z^N dans le développement en série de $\widetilde{g_i}(z)$ est le produit $\gamma(N,K)\rho_i(N,K)$. Pour calculer le développement en série de $\widetilde{g_i}(z)$, on l'écrit sous la forme :

$$\widetilde{g_i}(z) = h_K(z) - h_K(z)(g_i(z)^{-1} \ .$$

On a donc :

$$\gamma(N,K)\rho_i(N,K) = \gamma(N,K) - \sum_{n=0}^{N}\gamma(n,K)\theta_i(N-n) \ ,$$

où $\theta_i(m)$ est le coefficient de z^m dans le développement en série de $(g_i(z))^{-1}$. On calcule d'abord les $\theta_i(m)$ de manière itérative en initialisant par $\theta_i(0)=1$, puis en écrivant $g_i(z)(g_i(z))^{-1}=1$, soit :

$$\sum_{k=0}^{m}\theta_i(k)\alpha_{i,m-k} = 0 \ .$$

4.3 Réseaux de Petri

Les réseaux de Petri peuvent être vus comme le modèle le plus général pour des files d'attente synchronisées. Le langage établi s'écarte quelque peu de celui des files d'attente classiques. Un réseau de Petri se compose d'un nombre fini de K *places* (ou sites : ce sont les files) et d'un ensemble fini de T *transitions*. Chaque place i peut contenir un nombre entier n_i de *marques* (aussi appelées jetons ou charges : ce sont les clients). Le *marquage* du réseau est le vecteur à composantes entières :

$$\mathbf{n} \ = \ (n_1,\ldots,n_K) \ .$$

L'évolution de ce marquage est déterminée par le *déclenchement* des transitions. Lorsque la transition τ déclenche, elle apporte des marques à certaines places et en enlève à d'autres. Nous notons $c_{i,\tau}$ le nombre (positif, nul ou négatif) de marques ajoutées (ou retranchées) à la place i par le déclenchement de la transition τ. La matrice $C = (c_{i,\tau}) \ , \ i = 1,\ldots,K \ , \ \tau = 1,\ldots,T$ est

dite *matrice d'incidence* du réseau. Nous notons C_τ sa τ-ième colonne, qui décrit l'effet de la transition τ sur l'ensemble des places. Si le marquage est $\mathbf{n}$ avant que la transition τ déclenche, il devient $\mathbf{n}+C_\tau$ ensuite. Les réseaux de Jackson sont un cas particulier de réseau de Petri : les transitions ajoutent une marque (arrivée), en suppriment une (départ), ou transfèrent une marque d'une place à une autre (changement de file).

Un réseau de Petri est *temporisé* dès que l'on précise le processus de déclenchement des transitions. Pour cela, il faut décrire la loi des temps séparant deux déclenchements successifs. Nous nous plaçons uniquement dans le cadre des réseaux de Petri markoviens. Notons $Z_t \in \mathbb{N}^K$ le marquage à l'instant t. Le processus $\{Z_t\,,\, t \geq 0\}$ est un processus markovien de sauts à valeurs dans $\mathbb{N}^K$. Ses sauts correspondent au déclenchement des transitions du réseau de Petri. Si c'est la transition τ qui déclenche, le marquage courant passe de $\mathbf{n}$ à $\mathbf{n}+C_\tau$. Le taux de ce saut, noté $\lambda_\tau(\mathbf{n})$ dépend de l'ensemble du marquage $\mathbf{n}$ (et pas seulement des places concernées par la transition). C'est par l'intermédiaire de ces taux de transition que l'on traduit les synchronisations. Dans la pratique, on peut souvent se limiter au cas où les fonctions $\lambda_\tau(\mathbf{n})$ ne prennent que deux valeurs, 0 et $\lambda_\tau > 0$: pour certains marquages la transition τ est possible (et déclenche avec taux λ_τ), pour les autres elle est impossible. L'ensemble des valeurs possibles du vecteur $(\lambda_\tau(\mathbf{n}))_{\tau=1,\ldots,T}$ des taux de transition est alors fini et le processus est harmonisable. Nous noterons ν son horloge interne.

$$\nu \;=\; \sum_{\tau=1}^{T} \lambda_\tau \;.$$

L'algorithme général de simulation par la chaîne harmonisée est le suivant.

$t \longleftarrow 0$
Initialiser $Z = (Z_1, \ldots, Z_N)$
Répéter
    choisir $\tau \in \{1, \ldots, T\}$ avec probabilité λ_τ/ν
    Si $(Z \in E_\tau)$ alors $Z \longleftarrow Z + C_\tau$
    finSi
    $t \longleftarrow t - \log(\mathsf{Random})/\nu$
Jusqu'à (arrêt de la simulation)

Selon les cas, on sera amené à simuler le choix d'une transition par décomposition, eventuellement en plusieurs étapes.

Exemple : Files M/M/1 à sorties synchronisées.

Une représentation graphique standardisée des réseaux de Petri s'est imposée. Elle consiste à représenter les places par des cercles, les transitions par des barres, et à les relier par des flèches symbolisant l'effet des transitions. Dans l'exemple que nous traitons ici (voir figure 8), il y a 6 places et 7 transitions.

Les 4 premières transitions amènent des marques une par une dans les 4 premières places. Les deux premières places ont des sorties synchronisées. Lorsque la transition β_1 déclenche, une marque disparaît de chacune des places 1 et 2, et deux marques apparaissent dans la place 5. Les sorties des places 3 et 4, puis 5 et 6 sont synchronisées de façon analogue. La matrice d'incidence du réseau est la suivante.

	α_1	α_2	α_3	α_4	β_1	β_2	γ
1	1	0	0	0	−1	0	0
2	0	1	0	0	−1	0	0
3	0	0	1	0	0	−1	0
4	0	0	0	1	0	−1	0
5	0	0	0	0	2	0	−1
6	0	0	0	0	0	2	−1

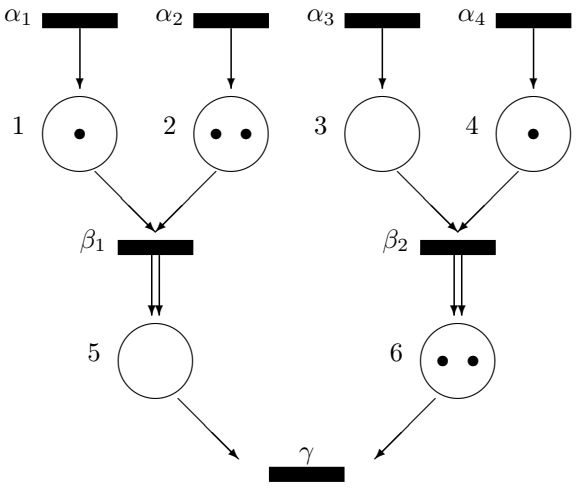

FIGURE 8 – *Exemple de réseau de Petri : files M/M/1 synchronisées.*

Nous supposerons que les taux des 4 premières transitions sont constants, et égaux à $\lambda_\alpha > 0$. Les taux des transitions β_1 et β_2 valent $\lambda_\beta > 0$ si les places en amont de ces transitions ont des marquages strictement positifs, 0 sinon. Le taux de la transition γ est $\lambda_\gamma > 0$ si les places 5 et 6 ont des marquages strictement positifs, 0 sinon. L'horloge interne vaut $\nu = 4\lambda_\alpha + 2\lambda_\beta + \lambda_\gamma$. Voici un algorithme possible pour la simulation de cet exemple.

$t \longleftarrow 0$
Initialiser $(Z_1, \ldots, Z_6)$

```
Répéter
    type ⟵ α, β ou γ avec probabilités 4λ_α/ν, 2λ_β/ν, λ_γ/ν
    Selon type
        type = α faire
            k ⟵ Random({1, 2, 3, 4})
            Z_k ⟵ Z_k + 1
        type = β faire
            k ⟵ Random({1, 2})
            Selon k
                k = 1 faire
                    Si (Z_1 > 0 et Z_2 > 0) alors
                        Z_1 ⟵ Z_1 − 1
                        Z_2 ⟵ Z_2 − 1
                        Z_5 ⟵ Z_5 + 2
                    finSi
                k = 2 faire
                    Si (Z_3 > 0 et Z_4 > 0) alors
                        Z_3 ⟵ Z_3 − 1
                        Z_4 ⟵ Z_4 − 1
                        Z_6 ⟵ Z_6 + 2
                    finSi
            finSelon
        type = γ faire
            Si (Z_5 > 0 et Z_6 > 0) alors
                Z_5 ⟵ Z_5 − 1
                Z_6 ⟵ Z_6 − 1
            finSi
    finSelon
    t ⟵ t − log(Random)/ν
Jusqu'à (arrêt de la simulation)
```

5 Exercices

NB : Les exercices qui suivent proposent surtout des développements mathématiques à base de techniques markoviennes. L'hypothèse de temps exponentiels sera donc omniprésente. Il est recommandé, parallèlement au traitement mathématique, d'effectuer des simulations du modèle proposé, et dans la mesure du possible, de vérifier expérimentalement les résultats théoriques. Il sera également intéressant de remplacer dans le modèle simulé, les lois exponentielles par d'autres lois de même espérance, afin de vérifier quels sont ceux parmi les résultats théoriques qui restent valides.

Exercice 1 On doit connecter n terminaux à une unité centrale. Le temps de traitement d'une tâche en unité centrale est exponentiel, de moyenne 500 millisecondes. L'intervalle de temps entre deux requêtes provenant d'un même terminal est exponentiel, de moyenne 20 secondes.

1. Quel est le nombre maximal de terminaux que l'on peut connecter sans qu'il y ait saturation ?
2. Pour 36 terminaux connectés, quel sera le temps de réponse moyen pour une requête provenant d'un terminal donné ?

Exercice 2 Des malades arrivent à l'hôpital selon un processus de Poisson d'intensité 12 par heure. La loi des durées de consultation est exponentielle, de moyenne 10 minutes.

1. Combien de médecins au minimum faut-il pour éviter l'engorgement ?
2. On décide d'affecter 3 médecins aux consultations. Quel sera le temps d'attente moyen d'un malade en régime stationnaire ?
3. Combien y aura-t-il de malades en attente en moyenne ?
4. Combien de médecins seront inoccupés en moyenne ?

Exercice 3

1. Il y a dans une poste deux files d'attente identiques et totalement séparées : ce sont deux files du type M/M/1. Pour chacune d'elles les arrivées sont séparées par des temps exponentiels de paramètre λ, les temps de service sont exponentiels de paramètre μ (on suppose $\lambda < \mu$). Toutes les variables aléatoires du système sont indépendantes.
   (a) Combien de clients en moyenne y a-t-il dans la poste en régime stationnaire ?
   (b) Quel est le temps d'attente moyen d'un client arrivant en régime stationnaire ?
2. On décide de fusionner les deux files pour n'en faire qu'une seule file à deux serveurs. Les clients continuent à arriver à la poste au même rythme (2λ en moyenne par unité de temps). Le temps de service de chaque client est encore exponentiel de paramètre μ.

(a) Combien de clients en moyenne y a-t-il dans la poste en régime stationnaire ?

(b) Quel est le temps d'attente moyen d'un client arrivant en régime stationnaire ?

3. Le fonctionnement a-t-il été amélioré, du point de vue de la poste, et du point de vue des clients, et pourquoi ?

Exercice 4 On considère un système d'attente du type M/M/1. Les clients se présentent devant la file selon un processus de Poisson d'intensité $\lambda > 0$. Les temps de service sont exponentiels de paramètre $\mu > 0$. On fait l'hypothèse supplémentaire suivante : un client qui arrive devant la file choisit de rester ou non en fonction du nombre de clients déjà présents dans le système. Si ce nombre est n, le client choisit de rester avec la probabilité q_n ou de quitter instantanément le système avec la probabilité $1 - q_n$. On suppose que $(q_n)_{n\in I\!N}$ est une suite décroissante de nombres positifs, et on note q sa limite. Les temps d'interarrivée, les temps de service et les choix des clients sont des variables aléatoires indépendantes dans leur ensemble. On note Z_t le nombre de clients présents dans le système (attente et service) à l'instant t.

1. Montrer que $\{Z_t\,,\, t \geq 0\}$ est un processus de naissance et de mort et donner ses taux de transition.
2. A quelle condition ce processus est-il récurrent positif ? Interpréter.
3. On suppose que q_n est égal à $(1 - n/N)/(n + 1)$, pour n de 0 à $N > 0$ et à 0 pour $n \geq N$ (N est un entier fixé).
   (a) Déterminer la mesure stationnaire.
   (b) Quel est le nombre moyen de clients dans le système en régime stationnaire ?
   (c) Que devient ce nombre quand λ tend vers 0 ? vers l'infini ? Interpréter.
4. On suppose désormais que q_n est égal à $1/(n + 1)$, pour tout $n \in I\!N$.
   (a) Déterminer la solution stationnaire.
   (b) Quel est le nombre moyen de clients dans le système en régime stationnaire ?
   (c) Quelle est la probabilité qu'un client arrivant en régime stationnaire reste dans le système ?
   (d) Quel est le temps moyen passé par ce client dans le système, s'il a décidé de rester et si n clients sont déjà présents quand il arrive ?
   (e) En déduire le temps moyen passé par un client dans le système, en régime stationnaire.
   (f) Soit $G(z, t)$ la fonction génératrice de la loi de Z_t. Déduire du système de Chapman-Kolmogorov une équation en G, $\partial G/\partial t$ et $\int_0^z G(u, t)\, du$.

(g) Soit $F(z,s)$ la transformée de Laplace de $G(z,t)$. Montrer que $F(z,s)$ est solution de l'équation différentielle suivante.

$$\frac{\partial F}{\partial z}(z,s)\left(\frac{sz}{z-1}+\mu\right)-F(z,s)\left(\frac{s}{(z-1)^2}+\lambda\right)+\frac{1}{(z-1)^2}=0\ .$$

Exercice 5 On considère une file M/M/1 dans laquelle peuvent passer deux types de clients. Les clients du premier type arrivent selon un processus de Poisson de paramètre λ_1, ceux du second type selon un processus de Poisson de paramètre λ_2. Les temps de service sont exponentiels de paramètres respectifs μ_1 et μ_2 pour les clients du premier et du second type.

Les clients du premier type ont une priorité absolue sur ceux du second. Aucun client de type 2 ne peut être servi si au moins un client de type 1 est présent dans le système. Si un client de type 2 est au service lorsqu'un client de type 1 arrive, celui-ci passe immédiatement au service et le client de type 2 est remis en attente.

On note :

- $p_{n,m}(t)$ la probabilité qu'il y ait n clients de type 1 et m clients de type 2 dans le système à l'instant t,
- $p^1_n(t)$ la probabilité qu'il y ait n clients de type 1 dans le système à l'instant t,
- $p^2_m(t)$ la probabilité qu'il y ait m clients de type 2 dans le système à l'instant t.

On note $\pi_{n,m}$, π^1_n, π^2_m les probabilités stationnaires correspondantes, quand elles existent.

1. Ecrire le système de Chapman-Kolmogorov dont les $(p_{n,m}(t))$ sont solution.
2. En déduire les équations dont les $(\pi_{n,m})$ sont solution.
3. Exprimer les $p^1_n(t)$ et $p^2_m(t)$ en fonction des $p_{n,m}(t)$.
4. Déterminer la condition d'équilibre du système pour les clients du type 1. On suppose désormais que cette condition est vérifiée.
5. Calculer les (π^1_n). Quel est le temps de séjour moyen d'un client du type 1 dans le système en régime stationnaire ? Quel est le nombre moyen de clients du type 1 présents dans le système en régime stationnaire ?
6. Quelle proportion du temps le serveur peut-il consacrer aux clients du type 2 quand l'équilibre est atteint pour ceux du type 1 ? En déduire le nombre moyen de clients du type 2 que le serveur peut traiter par unité de temps, puis la condition d'équilibre du système pour les clients du type 2.
7. En sommant par rapport à n les équations d'équilibre, écrire les équations reliant π^2_0 et $\pi_{0,1}$ puis π^2_m, π^2_{m-1}, $\pi_{0,m}$ et $\pi_{0,m+1}$. A partir de ces équations, calculer $\pi_{0,0}$ et retrouver la condition de la question précédente.

Exercice 6 On considère un système d'attente à une seule file et un seul serveur, fonctionnant sous les hypothèses suivantes. Les clients arrivent deux par deux. Chaque arrivée de deux clients est séparée de la précédente par une durée exponentielle de paramètre $\lambda > 0$. Les clients sont servis un par un. Chaque temps de service est exponentiel de paramètre $\mu > 0$. Toutes les variables aléatoires sont indépendantes. Pour tout $t \geq 0$ on note Z_t le nombre total de clients présents dans le système à l'instant t.

1. Montrer que $\{Z_t\,,\, t \geq 0\}$ est un processus markovien de saut sur $\mathbb{N}$ et représenter son diagramme de transition.
2. Pour tout $n \in \mathbb{N}$ on note $p_n(t) = P[Z_t = n]$. Ecrire le système de Chapman-Kolmogorov dont les $p_n(t)$ sont solution.
3. Si le processus admet une mesure stationnaire, on la notera $(\pi_n)_{n\in\mathbb{N}}$. De quel système linéaire la mesure (π_n) est-elle solution ?
4. On note g la fonction génératrice de la loi (π_n) :

$$g(z) = \sum_{n=0}^{\infty} z^n\, \pi_n\ .$$

   Exprimer $g(z)$ en fonction de λ, μ et π_0.
5. En déduire l'expression de π_0 en fonction de λ et μ puis la condition nécessaire et suffisante d'existence de la mesure stationnaire. Interpréter.
6. En supposant la condition précédente satisfaite et le régime stationnaire atteint, quel est le nombre moyen de clients dans le système en fonction de λ et μ ?
7. Ce système fonctionne-t-il mieux ou moins bien qu'une file M/M/1 dont le taux d'arrivée serait 2λ et le taux de service μ ? Expliquer pourquoi.

Exercice 7 On considère un système d'attente à une seule file et un seul serveur, fonctionnant sous les hypothèses suivantes. Les clients arrivent un par un. Chaque arrivée de client est séparée de la précédente par une durée exponentielle de paramètre $\lambda > 0$. Les clients sont servis deux par deux. S'il y a moins de deux clients dans la file quand un service se termine, le serveur attend qu'il y en ait deux. S'il y en a au moins deux, il prend les deux premiers et les autres restent en attente. La durée de chaque service (pour deux clients à la fois) est exponentielle de paramètre $\mu > 0$. Toutes les variables aléatoires sont indépendantes. Pour tout $t \geq 0$ on note Z_t le nombre total de clients présents dans le système à l'instant t.

1. Montrer que $\{Z_t\,,\, t \geq 0\}$ est un processus de Markov sur $\mathbb{N}$ et représenter son diagramme de transition.
2. Pour tout $n \in \mathbb{N}$ on note $p_n(t) = P[Z_t = n]$. Ecrire le système de Chapman-Kolmogorov dont les $p_n(t)$ sont solution.

3. Si le processus admet une mesure stationnaire, on la notera $(\pi_n)_{n\in I\!N}$. De quel système linéaire (π_n) est-elle solution ?
4. Toute solution du système précédent peut s'écrire sous la forme :

$$\pi_n = A\alpha^n + B\beta^n + C\gamma^n \ ,$$

où A, B et C sont des constantes quelconques et α, β et γ sont les racines de l'équation caractéristique associée (ECA) :

$$\lambda - (\lambda + \mu)z + \mu z^3 = 0 \ .$$

Montrer que pour $\lambda \geq 2\mu$ cette équation n'a aucune racine de module strictement inférieur à 1. En déduire que sous cette condition le processus n'a pas de mesure stationnaire. Interpréter.

5. Pour $\lambda < 2\mu$, montrer que l'équation (ECA) a une racine r dans $]0,1[$, les autres de module ≥ 1. En déduire l'expression de la mesure stationnaire en fonction de r.
6. En supposant la condition $\lambda < 2\mu$ satisfaite et le régime stationnaire atteint, quel est le nombre moyen de clients dans le système en fonction de λ et μ ?
7. Ce système fonctionne-t-il mieux ou moins bien qu'une file M/M/1 dont le taux d'arrivée serait λ et le taux de service 2μ ? Expliquer pourquoi.

Exercice 8 On considère un système d'attente à une seule file fonctionnant selon les hypothèses suivantes. Les clients arrivent dans la file selon un processus de Poisson de paramètre λ. Chaque client doit passer successivement par deux stations dont la première est dite "service" et la deuxième "contrôle". La durée de séjour d'un client dans le contrôle est exponentielle de paramètre ν. Le service ne peut fonctionner que si le contrôle est libre. Quand c'est le cas, le temps de service est exponentiel de paramètre μ. Quand le service d'un client se termine, celui-ci passe au contrôle, et le service est désactivé jusqu'à ce qu'il en sorte. Les temps d'interarrivée, les temps de service et les temps de contrôle sont indépendants dans leur ensemble.

On considère les états suivants du système :

– Etat $(0,n)$: le contrôle est libre et il y a n clients dans le système.
– Etat $(1,n)$: le contrôle est occupé et il y a n clients en attente.

On note $p_{0,n}(t)$ (respectivement $p_{1,n}(t)$) la probabilité qu'à l'instant t le système soit dans l'état $(0,n)$ (respectivement $(1,n)$). Si le système admet une mesure stationnaire, on notera $\pi_{0,n}$ et $\pi_{1,n}$ les probabilités stationnaires des états $(0,n)$ et $(1,n)$.

1. Décrire les taux de transition du système et écrire le système de Chapman-Kolmogorov.
2. En déduire le système d'équations linéaires que doit vérifier la solution stationnaire si elle existe.

3. On pose
$$g_0(z) = \sum_{n=0}^{\infty} z^n \, \pi_{0,n} \quad \text{et} \quad g_1(z) = \sum_{n=0}^{\infty} z^n \, \pi_{1,n} \; .$$
Déduire de la question précédente un système d'équations linéaires en $g_0(z)$, $g_1(z)$ et $\pi_{0,0}$.
4. Calculer $g_0(z)$ et $g_1(z)$ en fonction de $\pi_{0,0}$.
5. Si la solution stationnaire existe, que vaut $g_0(1) + g_1(1)$? En déduire l'expression de $\pi_{0,0}$ en fonction de λ, μ et ν.
6. Quelle est la condition nécessaire et suffisante d'existence de la mesure stationnaire ? On supposera désormais que cette condition est vérifiée.
7. Calculer la limite quand ν tend vers l'infini de $g_0(z)$ et $g_1(z)$. Interpréter.
8. Calculer la probabilité pour que le service soit actif en régime stationnaire.
9. Calculer la probabilité pour que le contrôle soit occupé en régime stationnaire.
10. Calculer le nombre moyen de clients dans le système en régime stationnaire.
11. En déduire le temps moyen passé par un client dans le système en régime stationnaire.
12. Le système proposé fonctionne en fait comme une file M/GI/1 pour laquelle le temps de service de chaque client serait la somme de deux exponentielles indépendantes, une de paramètre μ, l'autre de paramètre ν. Vérifier les résultats des questions précédentes en les comparant au théorème 3.3 (théorème de Pollaček-Khintchine).

Exercice 9 On considère un système M/M/1, muni, à l'entrée de la file, d'une porte qui ne laisse passer qu'un client sur deux. Les temps séparant deux arrivées successives dans la file sont exponentiels de paramètre λ, les temps de service sont exponentiels de paramètre μ, et toutes ces variables aléatoires sont indépendantes.

Quand la porte est ouverte et qu'un client se présente devant le système, il rentre, et la porte se referme immédiatement derrière lui. Quand la porte est fermée et qu'un client se présente devant le système, il est rejeté, mais ce rejet provoque l'ouverture de la porte pour le client suivant.

1. On considère deux types d'états.
   - $(0, n)$: la porte est fermée, il y a n clients au total dans le système.
   - $(1, n)$: la porte est ouverte, il y a n clients au total dans le système.

   Décrire les taux de transitions du système entre ces différents états.
2. On note

- $p_{0,n}(t)$ la probabilité que le système soit dans l'état $(0,n)$ à l'instant t.
- $p_{1,n}(t)$ la probabilité que le système soit dans l'état $(1,n)$ à l'instant t.

Ecrire le système de Chapman-Kolmogorov dont sont solution les $p_{0,n}(t)$ et les $p_{1,n}(t)$.

3. On note $\{\pi_{0,n}\ ,\ \pi_{1,n}\ ,\ n \in I\!N\}$ la mesure stationnaire, si elle existe. Montrer que nécessairement les suites $(\pi_{0,n})_{n\in I\!N}$ et $(\pi_{1,n})_{n\in I\!N}$ sont solution de la même équation de récurrence linéaire :

$$0 = \lambda^2 U_n - (\lambda+\mu)^2 U_{n+1} + 2\mu(\lambda+\mu)U_{n+2} - \mu^2 U_{n+3}\ .$$

On désigne par (ECA) l'équation caractéristique associée.

4. Montrer que (ECA) a trois racines réelles, dont une égale à 1 , et une ≥ 1. A quelle condition (ECA) a-t-elle une racine dans $]0,1[$? Interpréter cette condition. On la suppose désormais réalisée et on note r l'unique racine de (ECA) dans $]0,1[$.
5. Montrer que la mesure stationnaire est :

$$\pi_{0,0} = \frac{1-r}{2}\left(1-\frac{\mu r}{\lambda}\right) = \frac{3}{4} - \frac{1}{4}\sqrt{1+\frac{4\lambda}{\mu}}\ ,$$
$$\pi_{0,n} = \pi_{0,0}\,\frac{\lambda}{\mu}\,r^{n-1}\ ,\ \forall n \geq 1\ ,$$
$$\pi_{1,n} = \pi_{0,0}\,\frac{1-r}{2}\,r^n\ ,\ \forall n \geq 0\ .$$

6. En déduire la probabilité pour qu'il y ait n clients dans le système en régime stationnaire, en fonction de λ, μ, et r.
7. Quel est le nombre moyen de clients dans le système en régime stationnaire, en fonction de λ, μ, et r ?
8. Déterminer les limites des quantités calculées aux deux questions précédentes, quand λ/μ tend vers 0 et quand λ/μ tend vers 2. Interpréter.
9. Déterminer la loi du temps de séjour dans le système d'un client entrant en régime stationnaire. En déduire le temps de séjour moyen d'un client dans le système en régime stationnaire. Quelles sont les limites de ces quantités quand λ/μ tend vers 0 et quand λ/μ tend vers 2 ?
10. On note

$$G_0(z,t) = \sum_{n=0}^{\infty} z^n\, p_{0,n}(t) \quad \text{et} \quad G_1(z,t) = \sum_{n=0}^{\infty} z^n\, p_{1,n}(t)\ .$$

Déduire du système de Chapman-Kolmogorov, un système de deux équations différentielles en G_0, G_1, $\partial G_0/\partial t$, $\partial G_1/\partial t$, $p_{0,0}(t)$ et $p_{1,0}(t)$.

11. On note $f_{0,n}(s)$, $f_{1,n}(s)$, $F_0(z,s)$ et $F_1(z,s)$ les transformées de Laplace respectives des fonctions $p_{0,n}(t)$, $p_{1,n}(t)$, $G_0(z,t)$ et $G_1(z,t)$. En supposant qu'à l'instant initial le système est vide et la porte est ouverte, calculer $F_0(z,s)$ et $F_1(z,s)$ en fonction de $f_{0,0}(s)$ et $f_{1,0}(s)$. En déduire les valeurs de $f_{0,0}(s)$ et $f_{1,0}(s)$.
12. Le système proposé évolue en fait comme une file $E_2/M/1$, où E_2 (loi d'Erlang) désigne la somme de deux exponentielles indépendantes de paramètre λ. Comparer les résultats des questions précédentes avec ceux du paragraphe 3.2.

Exercice 10 Le but de l'exercice est d'étudier et de comparer deux politiques de gestion d'une file d'attente ayant deux serveurs différents.

Modèle A.

On considère un système à une seule file et deux serveurs sous les hypothèses suivantes. Les temps d'interarrivée sont exponentiels de paramètre $\lambda > 0$. Les temps de service du premier serveur sont exponentiels de paramètre $\mu_1 > 0$. Les temps de service du second serveur sont exponentiels de paramètre $\mu_2 > 0$. Toutes ces variables aléatoires sont indépendantes dans leur ensemble. Quand le système est vide, le prochain client qui arrive est dirigé vers le premier serveur. Quand la file d'attente est vide et le premier serveur occupé, le prochain client est dirigé vers le second serveur. Quand la file d'attente n'est pas vide, les deux serveurs sont occupés. Dès que l'un d'eux se libère, il prend le premier client de la file.

On code l'état du système de la façon suivante.

- $(0,0)$: le système est vide.
- $(1,0)$: un client au premier service, aucun au second.
- $(0,1)$: un client au second service, aucun au premier.
- $(n,1), n > 0$: les deux services sont occupés, $n-1$ clients attendent.

1. Représenter le diagramme de transition du système.
2. Les probabilités respectives pour que le système soit dans les états $(0,0)$, $(1,0)$, $(0,1)$, $(n,1)$ à l'instant t seront notées $p_{0,0}(t)$, $p_{1,0}(t)$, $p_{0,1}(t)$, $p_{n,1}(t)$. Les probabilités stationnaires correspondantes seront notées $\pi_{0,0}$, $\pi_{1,0}$, $\pi_{0,1}$, $\pi_{n,1}$. Ecrire le système de Chapman-Kolmogorov.
3. Quelle est la condition d'équilibre du système ?
4. Sous cette condition d'équilibre, déterminer la mesure stationnaire du système.
5. En déduire le nombre moyen de clients dans le système en régime stationnaire.
6. Pendant quelle proportion du temps le premier serveur fonctionne-t-il en régime stationnaire ?
7. Quel est le temps moyen passé dans le système par un client arrivant en régime stationnaire ?

Modèle B.

On considère maintenant un système composé de deux files d'attente de type $M/M/1$. Les clients arrivent dans le système selon un processus de Poisson de paramètre λ. Ils sont dirigés vers la première file avec probabilité p, vers la seconde avec probabilité $(1–p)$, ce routage aléatoire étant indépendant des arrivées et des services. La première file a un taux de service μ_1, la seconde un taux de service μ_2.

8. Quelles sont les conditions d'équilibre du système ?
9. Sous ces conditions, quel est le nombre moyen de clients dans le système en régime stationnaire ?
10. Quel est le temps d'attente moyen d'un client arrivant en régime stationnaire ?
11. On suppose que λ, μ_1 et μ_2 sont donnés, et vérifient $\lambda < (\mu_1 + \mu_2)$, $\mu_1 < \lambda$ et $\mu_2 < \lambda$. Montrer que la valeur suivante de p est optimale pour le fonctionnement du système.

$$p \;=\; \Big(\frac{1}{y_1} - \frac{1}{y_2} + y_2\Big)\Big(\frac{1}{y_1 + y_2}\Big)\,,$$

où

$$y_1 = \sqrt{\frac{\lambda}{\mu_1}} \quad \text{et} \quad y_2 = \sqrt{\frac{\lambda}{\mu_2}}\,.$$

Application.

Dans un certain réseau local, chacune des 500 stations de travail envoie en moyenne une tâche à exécuter toutes les 5 minutes. Deux serveurs sont disponibles. Le premier serveur exécute une tâche en moyenne en 7.26 millisecondes, le second serveur en 8.64 millisecondes. Le gestionnaire du réseau envisage plusieurs options, correspondant aux modèles A et B étudiés ci-dessus.

- option 1 : Installer une file d'attente commune (modèle A). Quand le système est vide, la prochaine tâche est envoyée au premier serveur.
- option 2 : Même chose, mais quand le système est vide, la prochaine tâche est envoyée au second serveur.
- option 3 : Chaque serveur conserve sa file d'attente et les tâches sont routées au hasard avec probabilité p, selon le modèle B.

12. Pour chacune des trois options, calculer le temps moyen de réponse d'une tâche. Laquelle des trois options le gestionnaire devra-t-il choisir ?
13. Que devrait-il décider si les deux serveurs étaient identiques ?
14. Que devrait-il décider si le premier serveur traitait ses tâches en moyenne en 0.726 millisecondes ?

Exercice 11 On considère le modèle de croissance d'une population animale défini par les hypothèses suivantes. Chaque femelle vivante engendre un descendant au bout d'un temps exponentiel de paramètre λ. Tout nouvel individu est soit une femelle avec probabilité p, soit un mâle avec probabilité $1-p$. Les durées de vie des femelles sont exponentielles de paramètre μ. Les durées de vie des mâles sont exponentielles de paramètre ν. Toutes ces variables aléatoires sont indépendantes dans leur ensemble. Les paramètres λ, μ, ν sont strictement positifs, p est strictement compris entre 0 et 1.

On notera :

- X_t le nombre de femelles vivant à l'instant t,
- Y_t le nombre de mâles vivant à l'instant t,
- $(q_n(t))_{n\in I\!N}$ la loi de probabilité de X_t,
- $G(z,t)$ sa fonction génératrice,

$$G(z,t) = \sum_{n=0}^{\infty} z^n \, q_n(t) \; .$$

- $(p_{n,m}(t))_{n,m\in I\!N}$ la loi de probabilité du couple (X_t, Y_t),
- $F(z,u,t)$ sa fonction génératrice,

$$F(z,u,t) = \sum_{n=0}^{\infty} \sum_{m=0}^{\infty} z^n u^m \, p_{n,m}(t) \; .$$

I. Etude de la population des femelles.

1. Montrer que le processus $\{X_t \, , \, t \geq 0\}$ est un processus de naissance et de mort et déterminer ses taux de transition.
2. Ecrire le système de Chapman-Kolmogorov en $q_n(t)$.
3. Déduire de ce système une équation aux dérivées partielles en $\partial G/\partial t$ et $\partial G/\partial z$.
4. Vérifier que si ϕ est une application dérivable de $I\!R$ dans $I\!R$ quelconque, alors :
$$G(z,t) = \phi\left(\frac{\mu - \lambda p z}{1-z} e^{(\mu-\lambda p)t}\right)$$
est solution de l'équation précédente. On admettra que dans le cas où $\lambda p \neq \mu$, toute solution peut s'écrire ainsi.
5. On suppose que à l'instant 0 , la population comprend une seule femelle. Que vaut $G(z,0)$? Montrer que nécessairement :
$$\phi(x) = \frac{x-\mu}{x-\lambda p} \; .$$
6. En déduire $q_0(t)$, puis $\lim\limits_{t\to\infty} q_0(t)$. Interpréter.
7. Calculer $I\!E[X_t]$, puis $\lim\limits_{t\to\infty} I\!E[X_t]$. Interpréter.

II. Etude de la population globale.

1. Ecrire le système de Chapman-Kolmogorov en $p_{n,m}(t)$.
2. En déduire une équation aux dérivées partielles en $\partial F/\partial t$, $\partial F/\partial z$ et $\partial F/\partial u$.
3. Sous quelles conditions, portant sur les paramètres du modèle la population s'éteindra-t-elle avec probabilité 1 ?

Exercice 12 Par suite de son environnement biochimique dans la cellule, un gène donné alterne des périodes d'activité au cours desquelles il s'exprime en produisant des molécules d'une certaine protéine P, et des périodes d'inactivité au cours desquelles il ne produit rien. Les molécules de P ont une durée de vie limitée au bout de laquelle leur structure stéréochimique se dégrade et leurs composants sont recyclés par la cellule. Le but du problème est d'étudier un modèle markovien permettant de prédire la loi du nombre de molécules de P présentes dans la cellule. On fait les hypothèses suivantes. Les périodes d'inactivité du gène suivent la loi $\mathcal{E}(\lambda)$. Les périodes d'activité du gène suivent la loi $\mathcal{E}(\mu)$. Quand le gène est actif la durée séparant deux naissances de molécules de P suit la loi $\mathcal{E}(\nu)$. La durée de vie de chaque molécule de P suit la loi $\mathcal{E}(\delta)$. Toutes ces variables aléatoires sont indépendantes dans leur ensemble.

On considère deux types d'états.

- $(0, n)$: le gène est inactif et n molécules de P sont présentes.
- $(1, n)$: le gène est actif et n molécules de P sont présentes.

On notera

- $p_{0,n}(t)$ la probabilité que le système soit dans l'état $(0, n)$ à l'instant t.
- $p_{1,n}(t)$ la probabilité que le système soit dans l'état $(1, n)$ à l'instant t.

1. Décrire les transitions élémentaires entre ces différents états.
2. Si le gène était actif en permanence ($\mu = 0$), à quel modèle de files d'attente correspondrait le nombre de molécules de P présentes dans la cellule à l'instant t ? Dans le cas $\mu = 0$ et $\delta > 0$, quel serait le nombre moyen de molécules de P présentes dans la cellule en régime stationnaire ? Dans le cas $\mu = \delta = 0$, quel serait le nombre moyen de molécules de P présentes dans la cellule à l'instant t ?
3. Ecrire le système de Chapman-Kolmogorov dont sont solution les $p_{0,n}(t)$ et les $p_{1,n}(t)$.
4. On note

$$G_0(z,t) = \sum_{n=0}^{\infty} p_{0,n}(t)\, z^n \quad \text{et} \quad G_1(z,t) = \sum_{n=0}^{\infty} p_{1,n}(t)\, z^n \ .$$

Déduire du système de Chapman-Kolmogorov, un système de deux équations différentielles en G_0, G_1, $\partial G_0/\partial z$, $\partial G_1/\partial z$, $\partial G_0/\partial t$, $\partial G_1/\partial t$.

5. Exprimer la probabilité que le gène soit actif à l'instant t à l'aide de G_1. On suppose qu'à l'instant 0 le gène est inactif et aucune molécule de P n'est présente dans la cellule. Que valent $G_0(z,0)$ et $G_1(z,0)$? Calculer $G_1(1,t)$ pour tout t (remarquer que l'état du gène (0 ou 1) est un processus de Markov à deux états). En déduire

$$\lim_{t\to\infty} G_0(1,t) \quad \text{et} \quad \lim_{t\to\infty} G_1(1,t) \; .$$

6. On suppose $\delta > 0$. Expliquer intuitivement pourquoi le processus considéré admet un régime stationnaire. On note :

$$g_0(z) = \lim_{t\to\infty} G_0(z,t) \quad \text{et} \quad g_1(z) = \lim_{t\to\infty} G_1(z,t) \; .$$

De quel système d'équations différentielles g_0 et g_1 sont-elles solution ?

7. On pose :

$$a = \frac{\lambda+\mu+\delta}{\delta} \; , \quad b = \frac{\nu}{\delta} \text{ et } \quad c = \frac{\lambda\,\nu}{\delta^2} \; .$$

Montrer que g_0 est solution de l'équation différentielle suivante.

$$(1-z)\,g_0''(z) - (a+b(1-z))\,g_0'(z) + c\,g_0(z) = 0 \; .$$

En déduire l'expression de la dérivée n-ième de g_0 en $z=1$:

$$g_0^{(n)}(1) = \frac{c(c+b)\dots(c+b(n-1))}{a(a+1)\dots(a+n-1)}\,\frac{\mu}{\lambda+\mu} \; .$$

8. On note $g = g_0 + g_1$ la fonction génératrice de la mesure stationnaire du nombre de molécules présentes dans la cellule. Montrer que :

$$g(z) = 1 + \sum_{n=1}^{\infty} \frac{1}{n!}\,\frac{c(c+b)\dots(c+b(n-1))}{(a-1)a\dots(a+n-2)}\,(z-1)^n \; .$$

9. Soit m l'espérance du nombre de molécules de P présentes dans la cellule, en régime stationnaire. Montrer que :

$$m = \frac{\lambda}{\lambda+\mu}\,\nu\delta \; .$$

Interpréter cette relation en utilisant les questions 2 et 5.

10. On note :

$$m_0(t) = \frac{\partial G_0}{\partial z}(1,t) \quad \text{et} \quad m_1(t) = \frac{\partial G_1}{\partial z}(1,t) \; .$$

Déduire de 4 et 5 un système d'équations différentielles dont m_0 et m_1 sont solution. Soit $m(t)$ le nombre moyen de molécules de P présentes dans la cellule à l'instant t. Quelle relation y a-t-il entre $m(t)$, $m_0(t)$ et $m_1(t)$?

11. Dans le cas $\delta > 0$, on note :

$$m_0 = \lim_{t\to\infty} m_0(t) \quad \text{et} \quad m_1 = \lim_{t\to\infty} m_1(t) \; .$$

Déduire de la question précédente un système linéaire de deux équations en m_0 et m_1. En déduire m_0 et m_1 et retrouver l'expression de m obtenue précédemment.

12. Dans le cas $\delta = 0$, montrer la relation :

$$m'(t) = \nu G_1(1,t).$$

En déduire l'expression suivante de $m(t)$:

$$m(t) = \frac{\lambda}{\lambda+\mu}\,\nu t + \frac{\lambda}{\lambda+\mu}\,\frac{\nu}{\lambda+\mu}\,(e^{-(\lambda+\mu)t} - 1) \; .$$

Interpréter le comportement asymptotique de $m(t)$ en utilisant les questions 2 et 5.

Exercice 13 On considère K files d'attente du type M/M/1 en tandem : les clients arrivent dans la première file suivant un processus de Poisson d'intensité λ, ils en sortent pour se diriger vers la seconde, puis vers la troisième, et sortent du système par la K-ième. Les temps de service de chaque client dans la i-ième file sont exponentiels de paramètre μ_i.

1. Quelle est la condition d'équilibre du système ?
2. En supposant cette condition satisfaite, déterminer la mesure stationnaire.
3. En déduire l'espérance du nombre total de clients dans le système en régime stationnaire.

Exercice 14 On considère un réseau formé de K files identiques, de type M/M/∞, disposées en tandem. Les clients arrivent selon un processus de Poisson de paramètre λ. Chaque client qui arrive dans la i-ième file reçoit sans attendre, un service exponentiel de paramètre μ_i. Pour $i = 1, \ldots, K-1$, les clients qui sortent de la i-ième file se dirigent vers la i+1-ième. Les clients qui sortent de la K-ième file sortent définitivement du système. Toutes les variables aléatoires (temps d'interarrivée et temps de service) sont indépendantes dans leur ensemble.

1. Quelle est la loi du temps de séjour d'un client dans le système ?
2. Quelle est la mesure stationnaire ?
3. Quelle est la loi du nombre total de clients dans le système en régime stationnaire ?

Exercice 15 On considère un réseau de files d'attentes fermé, de K files, contenant en tout N clients. Chacune des K files est de type M/M/∞ : le taux de sortie de la file i quand elle contient n_i clients, est $n_i\mu_i$, où $\mu_i > 0$. La probabilité qu'un client sortant de la file i se dirige vers la file j est $p_{i,j}$. On suppose que la matrice de transition $(p_{i,j})$ est irréductible. Toutes les variables aléatoires (temps de services et choix des clients) sont indépendantes dans leur ensemble. L'état du système à l'instant t est le N uplet d'entiers $(n_1, \ldots, n_K)$, où n_i est le nombre de clients présents dans la file i ($n_1 + \cdots + n_K = N$).

1. On suppose pour commencer qu'il n'y a qu'un seul client dans le système ($N = 1$). Montrer que le temps de séjour du client dans la file numéro i est exponentiel, de paramètre $\mu_i(1 - p_{ii})$.
2. Pour $i = 1, \ldots, N$, on notera c_i l'état du système où le client est dans la file numéro i. Soit $C_t \in \{c_1, \ldots, c_K\}$ l'état du système à l'instant t. Montrer que $\{C_t\,,\, t \geq 0\}$ est un processus markovien de saut. Montrer que le taux de transition de c_i vers c_j est $\lambda_{ij} = \mu_i p_{ij}$.
3. On note $(\rho_i)_{i=1,\ldots,K}$ la mesure stationnaire du processus de Markov de taux de transition (λ_{ij}), et $(e_i)_{i=1,\ldots,K}$ la mesure stationnaire de la chaîne de matrice de transition (p_{ij}). Montrer que les vecteurs $(\rho_i)_{i=1,\ldots,K}$ et $(e_i/\mu_i)_{i=1,\ldots,K}$ sont proportionnels.
4. Cas général : il y a N clients dans le système. Montrer que la mesure stationnaire de l'état du système est la loi multinomiale de paramètres $K, \rho_1, \ldots, \rho_K$.
5. Montrer que si la loi de l'état du système à l'instant 0 est la loi multinomiale de paramètres $N, \rho_1(0), \ldots, \rho_K(0)$, alors pour tout $t > 0$ la loi à l'instant t est multinomiale de paramètres $N, \rho_1(t), \ldots, \rho_K(t)$. Donner l'expression du vecteur $(\rho_i(t))_{i=1,\ldots,K}$ en fonction du vecteur $(\rho_i(0))_{i=1,\ldots,K}$ et du générateur correspondant aux taux de transition (λ_{ij}).

Exercice 16 On considère un système d'attente, constitué de deux files et un serveur. Les arrivées dans les deux files sont poissonniennes, d'intensités respectives λ_1 et λ_2. Les services sont exponentiels de paramètre μ. Le serveur sert en même temps deux clients, un dans chaque file. Si immédiatement après une fin de service l'une des files au moins est vide, le serveur attend qu'il y ait un client dans chaque file. Toutes les variables aléatoires (temps d'interarrivées et temps de services) sont indépendantes dans leur ensemble. Ecrire et implémenter un algorithme de simulation de ce système. Comme variables d'entrée, le programme comportera les paramètres λ_1, λ_2 et μ. En sortie, il devra afficher les compteurs des nombres de clients entrés et des nombres de clients servis dans chacune des deux files, ainsi que le compteur de temps.

1. Décrire le système comme un réseau de Petri à 2 places et 3 transitions.

Décrire la matrice d'incidence ainsi que les taux de déclenchement des transitions.

2. Vérifier expérimentalement que ce système ci-dessus n'a pas de régime stationnaire. Le démontrer rigoureusement.
3. Considérer ensuite le cas où la capacité de la première file est limitée à N_1 clients, celle de la deuxième file à N_2 clients. Déterminer la solution stationnaire. Dans le cas $\lambda_1 < \lambda_2 < \mu$ comparer le régime stationnaire de la première file avec celui d'une file M/M/1/N_1, de taux d'arrivée λ_1, taux de service λ_2.

Exercice 17 On considère un système d'attente, constitué de deux files et trois serveurs : un serveur pour chacune des files, et un troisième serveur jouant le rôle de contrôle pour chaque sortie. Les arrivées dans les deux files sont poissonniennes, d'intensités respectives λ_1 et λ_2. Les services sont exponentiels de paramètre μ_1 et μ_2 respectivement. Les temps de contrôle sont exponentiels de paramètre ν. Chaque client, en sortant du service de sa propre file, doit passer par le contrôle. La présence d'un client au contrôle bloque les deux services pendant toute la durée du contrôle. Toutes les variables aléatoires (temps d'interarrivée, temps de service et temps de contrôle) sont indépendantes dans leur ensemble.

1. Décrire le système comme un réseau de Petri à 3 places et 4 transitions. Décrire la matrice d'incidence ainsi que les taux de déclenchement des transitions.
2. Ecrire et implémenter un algorithme de simulation de ce système. Comme variables d'entrée, le programme comportera les paramètres λ_1, λ_2, μ_1, μ_2 et ν. En sortie, il devra afficher les compteurs des nombres de clients entrés et des nombres de clients servis dans chacune des deux files, ainsi que le compteur de temps.
3. Déterminer les conditions, portant sur les 5 paramètres λ_1, λ_2, μ_1, μ_2 et ν sous lesquelles le système admet un régime stationnaire. Plusieurs moyens pourront être envisagés pour répondre à cette question : expériences de simulation, système de Chapman-Kolmogorov, raisonnements de bon sens.
4. Généraliser la situation au cas de K files d'attente de type M/M/1, dont les sorties seraient couplées par un même contrôle, exponentiel de paramètre ν. Dans ce cas on supposera que les taux d'arrivée et de service dans chaque file sont les mêmes (λ et μ). On cherchera à déterminer la condition d'équilibre en fonction de K, λ, μ et ν.

Index

chaîne de Markov, 98, 102, 106, 110
Chapman-Kolmogorov, 78, 85, 90, 94, 96, 107
coefficient d'occupation, 83, 89, 98, 102, 113
condition d'équilibre, 77, 80, 88, 93, 95, 98, 102, 108

déclenchement, 113
diagramme de transition, 71, 93, 95
discipline de service, 82

équilibre, 76
explosion à temps fini, 74

file
  GI/M/1, 102
  $M^{(X)}$/M/1, 93
  M/GI/1, 97
  $M/M^{(a,b)}/1$, 95
  M/M/∞, 90
  M/M/s, 87
  M/M/1, 73, 77, 83, 95, 102
  M/M/1/N, 91, 112
  M/M/1/s/s, 91
fonction génératrice, 87, 89, 93, 94, 98, 112
formule
  d'Erlang, 92
  de Little, 84, 89, 105

loi
  d'Erlang, 82, 84, 104
  de Poisson, 90
  exponentielle, 72, 81, 83, 89, 99
  gamma, 84, 89, 104

marquage, 113
mesure stationnaire, 76, 79, 83, 88, 94, 96, 98, 103, 108, 111
modèle
  d'attente, 80
  de population, 74

notation de Kendall, 82

probabilité de perte, 92
processus
  de naissance et de mort, 71, 88, 90, 92, 109, 111
  de naissance pure, 73, 74
  de Poisson, 73, 100, 105
  markovien de saut, 71, 95, 107, 114
  récurrent nul, 76, 80
  récurrent positif, 76, 80
  transient, 76, 80

réseau de Jackson fermé, 110
réseau de Jackson ouvert, 105
réseau de Petri, 113

saturation, 76, 102
simulation
  d'un processus de naissance et de mort, 71
  d'un réseau de Petri, 114
système de Chapman-Kolmogorov, 78, 85, 90, 94, 96, 107

taux
  d'arrivée, 77
  de mort, 71
  de naissance, 71
  de service, 77, 105
temps
  d'attente, 88
  d'interarrivée, 81
  de séjour, 83, 100, 104
  de service, 81
théorème
  de Jackson, 107, 110
  de Pollaček-Khintchine, 100
transformée de Laplace, 85, 99, 104