École Nationale Supérieure des Télécommunications

Insensibilité dans les Réseaux de Files d'Attente et Applications au Partage de Ressources Informatiques

THÈSE

présentée et soutenue publiquement le 29 octobre 2007

pour l'obtention du

Doctorat de l'École Nationale Supérieure des Télécommunications
(spécialité: informatique et réseaux)

par

Minh Anh Trần

Composition du jury

Président :	Jean-Michel Fourneau
Rapporteurs :	Hans Daduna Jean Mairesse
Examinateur :	Laurent Decreusefond
Directeur de thèse :	François Baccelli
Codirecteur :	Thomas Bonald

Groupe de Recherche sur la Théorie des Réseaux Et Communications (TREC)
Laboratoire d'informatique de l'École Normale Supérieure - 45 rue d'Ulm

À mes parents,

Remerciements / Acknowledgments

Je tiens tout d'abord à remercier François Baccelli d'avoir accepté de diriger cette thèse. Ce fut pour moi un plaisir et un grand honneur de profiter de ses conseils et de son soutien.

Mille et un merci à Thomas Bonald. Sa créativité, sa rigueur scientifique et ses conseils m'ont guidé tout au long de ces années. Ce fut un réel plaisir de colaborer avec un codirecteur de thèse tellement rapide et enthousiaste.

Je souhaite remercier vivement Jean Mairesse d'avoir accepté de lire et de rapporter soigneusement ce travail. Ses remarques m'ont permis d'améliorer de nombreuses parties de cette thèse.

It is my great honour that Hans Daduna reviewed this PhD thesis. I would like to thank him for his valuable remarks. Ce fut aussi un grand honneur pour moi d'avoir dans le jury Laurent Decreusefond et Jean-Michel Fourneau, merci à eux.

J'aimerais remercier, pour l'ambiance très sympathique, tous les membres du groupe TREC : Charles Bordenave, Bartlomiej Blaszczyszyn, Pierre Brémaud, Giovanna Carofiglo, Augustin Chaintreau, Prassana Chaporkar, Yogeshwaran Dandhapani, Bruno Kauffmann, Frédéric Morlot, Alexandre Proutière, Bozidar Radunovic, Julien Reynier, Emmanuel Roy et Justin Salez. C'était à TREC que Bartek a dirigé mon premier stage de recherche il y a 4 ans et que Marc a corrigé ma rédation hier...

Ce travail académique n'aurait jamais vu le jour sans l'encouragement de mes parents, de mon frère et de mes amis. Je voudrais les remercier pour leur soutien sur les terrains non-académiques.

Le plus fort de mes remerciements est pour Hà. Ce sont sa présence, son amour qui alimentent mon inspiration.

Résumé

Nous abordons dans cette thèse le problème de l'insensibilité dans les réseaux de files d'attente et quelques applications au partage de ressources informatiques. Tout d'abord, nous montrons que les réseaux de files d'attente symétriques avec le routage de Jackson ou de Kelly sont tous insensibles à la distribution des demandes de service même si à l'arrivée, au départ ou au changement de files d'un client quelconque, les autres clients dans chaque file sont permutés au hasard selon certaine loi dépendante de l'état du réseau. Nous identifions également certaines disciplines de service non symétriques pour lesquellesla propriété d'insensibilité est satisfaite. Ensuite, nous proposons deux nouvelles métriques de débit pour les réseaux de données. Nous montrons quelques propriétés génériques satisfaites par ces deux métriques et nous illustrons leur différence à travers quelques exemples. Enfin, nous montrons que l'équilibrage de sources de trafic élastique détériore la performance en termes de débit, et en présence de contrôle d'admission, de probabilité de blocage.

Mots-clés: Files d'attente, Réseau, Insensibilité, Symétrique, Permutation, Débit, Métrique de Débit, Trafic élastique, Équilibrage de source, Équilibrage de trafic, Probabilité de blocage, Contrôle d'admission, Contrainte de capacité.

Abstract

In this thesis, we tackle the problem of insensitivity in queueing networks and consider some applications to computer resource sharing. First of all, we prove that networks of symmetric queues with Jackson or Kelly routing are both insensitive to the service requirement distribution even if at arrivals, departures or migration events, customers at each queue are randomly permuted according to some law that may depend on the network state. We also identify some non-symmetric service disciplines for which the insensitivity property holds. We then propose two new throughput metrics for data networks. We prove some generic properties satisfied by these two metrics and we illustrate their difference through some examples. Finally, we prove that balancing elastic traffic sources worsens performance in terms of throughput and, in the presence of admission control, of blocking probability.

Keywords: Queueing network, Insensitivity, Symmetric, Permutation, Throughput, Throughput measure, Elastic traffic, Traffic balancing, Source balancing, Blocking probability, Admission control, Capacity constraint.

Table des matières

Introduction

Cette thèse comporte de deux parties principales plus ou moins indépendantes, dont la première consiste en études de la propriété d'insensibilité dans les réseaux de files d'attente et la deuxième consiste en l'évaluation de la performance des réseaux informatiques.

Les files d'attente sont souvent rencontrées en pratique : à la poste, aux supermarchés, dans les gestions des avions au décollage ou à l'atterrissage,... La théorie des files d'attente utilise des outils probabilistes pour modéliser et étudier les solutions optimales de gestion de ces objets. Ce domaine de recherche, né en 1917, des travaux de l'ingénieur danois Erlang [Erl09] sur la gestion des réseaux téléphoniques, étudie aussi les réseaux de données et le réseau Internet.

En pratique, on s'intéresse typiquement aux nombres de clients dans le réseau, aux temps de séjour des clients et à leur débit. En général, l'expression de ces quantités dépend fortement de la façon dont arrivent les clients dans le système, des demandes de service des clients et de la discipline déterminant comment les clients sont servis. Dans la première partie, on s'intéresse aux réseaux de files d'attente où la distribution du nombre de clients ne dépend pas des distributions des demandes de service. Ces réseaux sont dits *insensibles* aux distributions des demandes de service ou *insensibles* tout court, et on dit aussi que la discipline correspondante est insensible. Cette propriété d'insensibilité est d'un grand intérêt car les nombres de clients, leur temps de séjour et leur débit peuvent être évalués sans savoir les statistiques précises des demandes de service.

En 1979, Kelly a introduit une large classe de disciplines insensibles, il s'agit des disciplines *symétriques* [Kel79]. Ces disciplines symétriques incluent la discipline processor-sharing (PS) et la discipline LIFO préemptive considérées par Baskett, Chandy, Muntz et Palacios [BCMP75].

Dans le premier chapitre, on rappellera des outils probabilistes fréquemment utilisés et des modèles connus de files d'attente insensibles. Dans le deuxième chapitre, on introduira les permutations aléatoires de clients [Dad01a, Yas80] dans les files d'attente symétriques et on montrera que ces files et leurs réseaux restent insensibles. Ensuite, on considérera l'insensibilité de quelques disciplines non symétriques avant de donner un résumé des disciplines insensibles/sensibles.

Pour la deuxième partie sur l'évaluation de la performance des réseaux informatiques, on définira deux nouvelles métriques de débit dans le troisième chapitre. Il s'agit du débit échantillonné en temps et du débit échantillonné par flot [KK02, LvdBB04, BT07b]. On montrera quelques propriétés génériques satisfaites par les deux métriques et on illustrera leur différence par quelques exemples.

Dans le quatrième chapitre, on considérera le modèle d'un lien partagé par des sources de circuit et de trafic élastique. On montrera que l'équilibrage de sources augmente la probabilité de blocage et diminue les débits. Alors le système peut être dimensionné sans risque en supposant des sources homogènes malgré la forte hétérogénéité observée en pratique.

Table des figures

Première partie

Insensibilité dans les réseaux de files d'attente

1

Préliminaires

Sommaire

L'objectif de ce chapitre est de rappeler des notions de modèles markoviens, de files d'attente et de leurs réseaux. Dans la première partie, on étudiera les chaînes et les processus de Markov, les notions de probabilités de transition, matrice de transition, mesures invariantes,... [Bre99] qui servent à étudier les modèles mathématiques dans la théorie de files d'attente. Dans la deuxième partie, on considérera quelques systèmes de files d'attente, les notions de disciplines, de distributions stationnaires, d'équations de balance, de l'insensibilité et quelques méthodes pour démontrer l'insensibilité d'un système de files d'attente.

1.1 Modèles markoviens

Dans cette partie, sont rappelés les définitions des chaînes de Markov, des processus markoviens, leurs propriétés et leur équilibre.

1.1.1 Chaînes de Markov

Considérons une suite de variables aléatoires $X_0, X_1, \ldots, X_n, \ldots$ où

– $X_0, X_1, \ldots$ sont des variables aléatoires à valeurs dans un espace d'états fini ou dénombrable $\mathcal{X}$. Par exemple, l'espace d'états $\mathcal{X}$ peut être $\{1, \ldots, N\}$, $\mathbb{N}$, ou $\mathbb{N}^d, \ldots$

– La variable X_n peut dépendre, a priori de $X_0, \ldots, X_{n-1}$, et même des valeurs futures $X_{n+1}, X_{n+2}, \ldots$

Définitions

Définition 1.1 *On appelle une matrice à termes positifs $P = \big(p(x,y), x, y \in \mathcal{X}\big)$ une matrice de trasition si*

$$\sum_{y \in \mathcal{X}} p(x,y) = 1, \qquad \forall x \in \mathcal{X}.$$

Définition 1.2 *On dit que la chaîne (X_n) a la propriété de Markov si*

$$\mathbb{P}(X_n = x_n \mid X_{n-1} = x_{n-1}, X_{n-2} = x_{n-2}, \ldots, X_0 = x_0) = \mathbb{P}(X_n = x_n \mid X_{n-1} = x_{n-1}), \quad (1.1)$$

pour tout $n \geq 1$. Et dans ce cas, on appelle (X_n) une chaîne de Markov.

Littéralement, la chaîne (X_n) a la propriété de Markov si sachant tout le passé jusqu'au rang $n-1$, le comportement ultérieur de la chaîne ne dépend que de la dernière variable aléatoire X_{n-1}.

Définition 1.3 *La chaîne de Markov (X_n) est homogène s'il existe une matrice de trasition P telle que*

$$\mathbb{P}(X_n = y \mid X_{n-1} = x) = p(x,y), \quad (1.2)$$

pour tout $n \geq 1$. Dans ce cas, on appelle P la matrice de transition de la chaîne de Markov (X_n).

Remarquons que l'homogénéité assure que la dynamique de la chaîne de Markov est *invariante* dans le temps.

Considérons par exemple une marche aléatoire (X_n) sur l'ensemble des entiers $\mathbb{Z}$ (voir la Figure 1.1) dont la probabilité d'avance est p et la probabilité de recul est $1-p$:

$$p(k, k+1) = p, \quad \text{et} \quad p(k, k-1) = 1 - p, \qquad \forall k \in \mathbb{Z}.$$

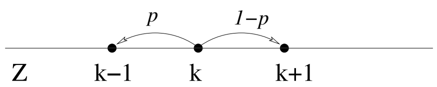

FIG. 1.1 – Une marche aléatoire sur $\mathbb{Z}$.

Cette marche aléatoire (X_n) est une chaîne de Markov homogène :

- Espace d'états $\mathbb{Z}$.
- Matrice de transition $P = \big(p(x,y), x, y \in \mathbb{Z}\big)$:

$$p(x,y) = \begin{cases} p & \text{si } y = x+1, \\ 1-p & \text{si } y = x-1, \\ 0 & \text{sinon.} \end{cases}$$

Considérons maintenant une autre marche aléatoire (X_n) sur l'ensemble $\mathbb{Z}$ avec les *probabilités de transition* hétérogènes :

$$\mathbb{P}(X_n = k+1 \mid X_{n-1} = k) = \frac{1}{n}, \quad \text{et} \quad \mathbb{P}(X_n = k-1 \mid X_{n-1} = k) = \frac{n-1}{n}, \qquad \forall k \in \mathbb{Z},$$

pour tout $n \geq 1$. Alors cette marche aléatoire est toujours une chaîne de Markov mais elle n'est plus homogène car les probabilités de transition dépendent du rang n de la variable aléatoire X_n.

Et si la loi de la variable aléatoire X_n dépend non seulement de X_{n-1} mais aussi de X_{n-2} par exemple, la chaîne (X_n) n'a plus la propriété de Markov.

À partir d'ici, une chaîne de Markov est sous-entendue homogène si l'on ne précise rien. Alors comme une marche aléatoire sur $\mathbb{Z}$, une chaîne de Markov peut être représentée sous forme d'un graphe, voir la Figure 1.2.

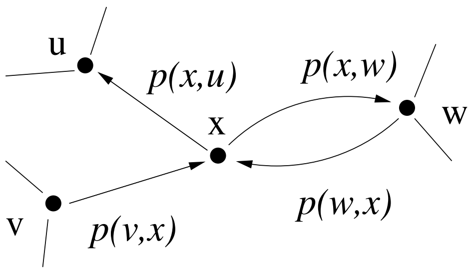

FIG. 1.2 – Représentation graphique d'une chaîne de Markov.

Loi d'une chaîne de Markov

Soit (X_n) une chaîne de Markov de matrice de transition $P = \big(p(x,y)\big)$.

Soit ν la *loi initiale* : $\mathbb{P}(X_0 = x_0) = \nu(x_0)$ pour tout $x \in \mathcal{X}$.

Alors la *loi de la chaîne de Markov* est déterminée par les *probabilités de trajectoires*

$$\mathbb{P}(X_0 = x_0, X_1 = x_1, \ldots, X_n = x_n) = \nu(x_0)p(x_0,x_1)p(x_1,x_2)\ldots p(x_{n-1},x_n), \qquad (1.3)$$

pour tous $n \geq 0$ et $x_0, \ldots, x_n \in \mathcal{X}$.

La loi de la variable aléatoire X_n est donnée par

$$\mathbb{P}(X_n = y) = \sum_{x_0,\ldots,x_{n-1} \in \mathcal{X}} \nu(x_0)p(x_0,x_1)p(x_1,x_2)\ldots p(x_{n-1},y), \qquad \forall y \in \mathcal{X},$$

pour tout $n \geq 0$.

En particulier, la probabilité de passer de x à y en n étapes est donnée par

$$\mathbb{P}(X_n = y \mid X_0 = x) = p^{(n)}(x, y) = P^{(n)}(x, y).$$

Notons $\mathbb{P}_\nu$ la loi de la chaîne de Markov sachant que la loi initiale soit ν :

$$\mathbb{P}_\nu(X_n \in A) = \sum_{x_0 \in \mathcal{X}, y \in A} \nu(x_0)\mathbb{P}(X_n = y \mid X_0 = x_0), \qquad \forall A \subset \mathcal{X}.$$

Périodicité

Définition 1.4 *On appelle période de l'état i et on note d_i le PGCD de l'ensemble M_i des $n \geq 1$ tels que $p^{(n)}(i, i) > 0$, avec la convention $d_i = \infty$ si cet ensemble est vide. Et on dit que i est périodique si $d_i > 1$.*

Irréductibilité

La chaîne de Markov (X_n) est *irréductible* si pour tous $x, y \in \mathcal{X}$, la probabilité d'atteindre l'état y est strictement possitive sachant que l'état initial est x :

$$\forall x, y \in \mathcal{X}, \exists n : \quad \mathbb{P}(X_n = y \mid X_0 = x) > 0.$$

Remarquons que l'irréductibilité est équivalente à l'existence d'un chemin de x à y sur le graphe de transition pour tous $x, y \in \mathcal{X}$. Si la chaîne est irréductible, ses états ont la même période ; de plus, si aucun état n'est périodique, la chaîne est dite *apériodique.*

Équilibre d'une chaîne de Markov

Définition 1.5 *Une mesure de probabilité $\pi(x)$ sur l'espace d'état $\mathcal{X}$ est invariante pour la matrice de transition $P = \big(p(x, y)\big)$ si*

$$\pi(x) = \sum_{y \in \mathcal{X}} \pi(y)p(y, x), \qquad \forall x \in \mathcal{X}. \tag{1.4}$$

Dans ce cas, π est appelée la *probabilité invariante* et aussi la *probabilité stationnaire* de la chaîne de Markov, et les équations (1.4) sont appelées les *équations de balance globale.*

Théorème 1.6 *Soient π une mesure de probabilité invariante pour la matrice P et (X_n) une chaîne de Markov de matrice de transition P telles que*

$$\mathbb{P}(X_0 = x) = \pi(x), \qquad \forall x \in \mathcal{X}$$

alors

$$\mathbb{P}(X_n = x) = \pi(x), \qquad \forall n \geq 0, \forall x \in \mathcal{X}.$$

On dit que la chaîne de Markov (X_n) est à l'équilibre.

Reprenons l'exemple de la marche alétoire (X_n) sur $\mathbb{Z}$ de matrice de transition P :

$$p(x,y) = \begin{cases} p & \text{si } y = x+1, \\ 1-p & \text{si } y = x-1, \\ 0 & \text{sinon.} \end{cases}$$

Alors les équations de balance globale sont

$$\pi(k) = p\pi(k-1) + (1-p)\pi(k+1), \qquad \forall k \in \mathbb{Z}. \tag{1.5}$$

et $\sum_{k\in\mathbb{Z}} \pi(k) = 1$.

Le système d'équations (1.5) admet des solutions

$$\pi(k) = C, \qquad \forall k \in \mathbb{Z},$$

pour un constance quelconque C. Comme ces mesures ne sont pas sommables, il n'existe pas de probabilité invariante, et par conséquence, il n'existe pas d'équilibre non plus.

Limitons maintenant l'espace d'états de la marche aléatoire (X_n) à un ensemble fini, par exemple, l'ensemble $\{0,1,\ldots,N\}, N \geq 1$:

$$p(x,y) = \begin{cases} p & \text{si } (x,y) = (x,x+1), 0 \leq x \leq N-1, \text{ ou } (x,y) = (N,N) \\ 1-p & \text{si } (x,y) = (x,x-1), 1 \leq x \leq N, \text{ ou } (x,y) = (0,0) \\ 0 & \text{sinon.} \end{cases}$$

Alors les équations de balance globale deviennent

$$\begin{cases} \pi(0) & = p\pi(0) + (1-p)\pi(1), \\ \pi(k) & = p\pi(k-1) + (1-p)\pi(k+1), \quad 1 \leq k \leq N-1, \\ \pi(N) & = p\pi(N-1) + (1-p)\pi(N), \end{cases}$$

et $\sum_{k\in\mathbb{Z}} \pi(k) = 1$.

Ce système d'équations admet comme solution une unique probabilité invariante

$$\pi(k) = \frac{1}{N+1}, \qquad 0 \leq k \leq N.$$

Unicité de probabilité invariante et Ergodicité

Théorème 1.7 *Une chaîne de Markov irréductible a au plus une probabilité invariante. Si l'espace d'états est fini, la chaîne a une unique probabilité invariante.*

Théorème 1.8 *Soit une chaîne de Markov irréductible apériodique (X_n) avec la probabilité invariante π. Alors $\mathbb{P}$-presque sûrement*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} f(X_i) = \mathbb{E}_\pi[f],$$

pour toute fonction f telle que $\mathbb{E}_\pi\big[|f|\big] < \infty$, où $\mathbb{E}_\pi[f] := \sum_{\mathcal{X}} \pi(x)f(x)$.

1.1.2 Processus de sauts

Loi exponentielle

Pour un réel strictement positif λ, on dit qu'une variable aléatoire X suit une distribution exponentielle de paramètre λ si

$$\mathbb{E}\big[f(X)\big] = \int_0^\infty f(x)\lambda e^{-\lambda x} dx. \tag{1.6}$$

Proposition 1.9 *Si une variable aléatoire X suit une loi exponentielle alors elle est sans mémoire, c.à.d. la variable aléatoire X satisfait pour tous x, y positifs :*

$$\mathbb{P}(X \geq x + y \mid X \geq y) = \mathbb{P}(X \geq x).$$

Si X modélise la durée de vie d'un individu A, la propriété que X est sans mémoire exprime que A ne vieillit pas : si A a vécu y années, la probabilité pour qu'il vive encore x années est la même que la probabilité pour qu'un individu similaire à A qui vient de naître vive lui aussi x années.

Processus de Poisson

Soit $\big(V_n, n \geq 1\big)$ une suite de variables aléatoires indépendantes, identiquement distribuées (i.i.d.), de même loi exponentielle de paramètre $\lambda, \lambda > 0$.

Alors le processus $(T_n = \sum_{i=1}^n V_i)$ est appelé un *processus de Poisson* d'*intensité* λ.

Définition des processus de sauts

Soient (X_n) une chaîne de Markov sur $\mathcal{X}$ de matrice de transition $P = \big(p(x,y)\big)$, et $q_x, x \in \mathcal{X}$, des réels strictement positifs. Considérons $(X(t))$ un processus satisfaisant :

- $X(0) = X_0 = x_0 \in \mathcal{X}$ et $X(t) = x_0$ pour tout $t < T_1 = V_1$, où V_1 est une variable aléatoire de loi exponentielle de paramètre q_{x_0}.
- $X(t) = X_1 = x_1 \in \mathcal{X}$ pour tout $t : T_1 \leq t < T_2 = T_1 + V_2$, où V_2 est une variable aléatoire de loi exponentielle de paramètre q_{x_1}.
- Pour tout $n \geq 1$, $X(t) = X_n = x_n \in \mathcal{X}$ pour tout $t : T_{n-1} \leq t < T_n = T_{n-1} + V_n$, où V_n est une variable aléatoire de loi exponentielle de paramètre q_{x_n}.

On appelle $(X(t))$ un *processus de sauts*, (X_n) la *chaîne incluse* du processus de sauts $(X(t))$. Remarquons que le processus de sauts $(X(t))$ suit la même trajectoire que la chaîne incluse (X_n), voir les Figures 1.3 et 1.4.

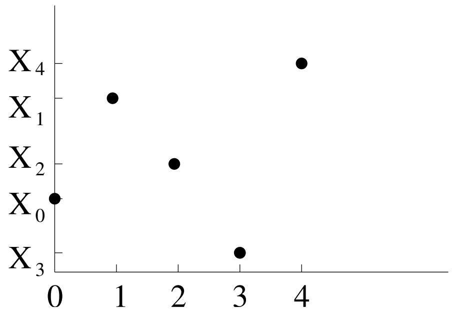

FIG. 1.3 – La chaîne incluse (X_n).

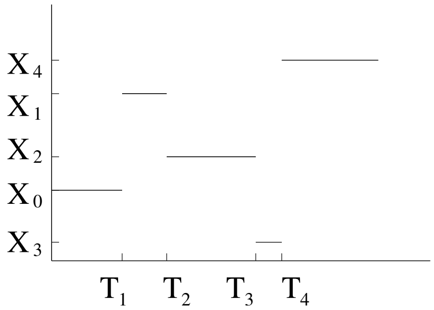

FIG. 1.4 – Le processus de sauts $\big(X(t)\big)$.

Exemples

Soient $(V_n, n \geq 1)$ une suite de variables aléatoires i.i.d. de loi exponentielle de paramètre λ et $(T_n = \sum_{i=1}^{n} V_i)$ le processus de Poisson correspondant. Alors ce processus de Poisson est équivalent à une fonction croissante $N(t)$ définie par le nombre de points T_n entre 0 et t :

$$N(t) := \sum_{n=1}^{\infty} \mathbb{I}(0 < T_n \leq t),$$

où $\mathbb{I}$ est la fonction indicatrice : $\mathbb{I}(A) = 1$ si l'événement A est vrai et $\mathbb{I}(A) = 0$ sinon.

Cette fonction $N(t)$ est appelée la *mesure de comptage* du processus de Poisson (T_n). Cette mesure de comptage définit un processus de sauts, illustré dans la Figure 1.5.

Générateur infinitésimal

Soit $\big(X(t)\big)$ un processus de sauts de chaîne incluse (X_n) et de matrice de transition $P = \big(p(x,y), x, y \in \mathcal{X}\big)$. Le générateur infinitésiaml $Q = \big(q(x,y), x, y \in \mathcal{X}\big)$ est défini par

$$q(x,y) = \begin{cases} q_x p(x,y) & \text{si } x \neq y, \\ -q_x & \text{si } x = y. \end{cases}$$

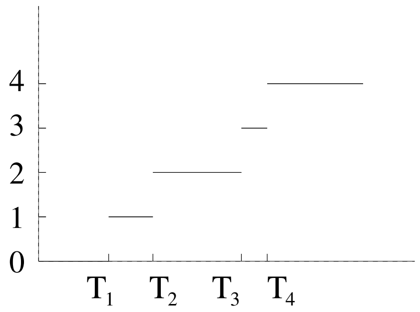

FIG. 1.5 – La mesure de comptage $(N(t))$.

Remarquons que ce générateur infinitésimal décrit complètement le processus de sauts original :

$$\begin{aligned} q_x &= \sum_{y \in \mathcal{X}} q(x,y), \\ p(x,y) &= \frac{q(x,y)}{q_x}, \quad x \neq y. \end{aligned} \tag{1.7}$$

On appelle $q(x,y)$ le *taux de transition* de l'état x à l'état y, pour tous $x, y \in \mathcal{X}$.

Définition 1.10 *Le processus de sauts $(X(t))$ est dit non-explosif si la suite des instants de sauts (T_n) satisfait :*

$$\lim_{n \to \infty} T_n = \infty, \quad \mathbb{P} - p.s.$$

Théorème 1.11 *Le processus de sauts non-explosif $(X(t))$ a la propriété de Markov :*

$$[X(s+t) \mid X(u), u \leq t, X(t) = x] \stackrel{dist.}{=} [X(s+t) \mid X(t) = x]$$

Dans ce cas, $(X(t))$ est appelé un processus markovien, et si de plus, la chaîne incluse est homogène, ce processus a la propriété de Markov homogène :

$$[X(s+t) \mid X(u), u \leq t, X(t) = x] \stackrel{dist.}{=} [X(s) \mid X(0) = x]$$

Notation markovienne

Si $(X(t))$ est un processus de sauts alors on note

$$\mathbb{P}_x(A) := \mathbb{P}(A \mid X(0) = x), \tag{1.8}$$
$$\mathbb{E}_x[V] := \mathbb{E}\big[V \mid X(0) = x\big], \tag{1.9}$$

où A est un événement et V est une variable aléatoire intégrable.

Ergodicité et Équilibre

Le processus markovien $(X(t))$ est dit *ergodique* si

$$\lim_{t\to\infty} \mathbb{P}_y(X(t) = x) = \pi(x), \qquad \forall x, y \in \mathcal{X},$$

où π est une probabilité sur l'espace d'états $\mathcal{X}$. Cette quantité $\pi(x)$ peut être interprétée comme la fréquence relative que le processus $(X(t))$ visite l'état x.

Supposons maintenant que la chaîne incluse admet l'unique probabilité invariante π_0. Comme le temps que le processus $(X(t))$ reste à un état x avant une transition à un autre état est de loi exponentielle de moyenne $1/q_x$, la fréquence qu'il reste à un état x est proportionnelle à la fréquence que la chaîne incluse visite cet état multipliée par $1/q_x$:

$$\pi(x) \propto \pi_0(x)\frac{1}{q_x}. \tag{1.10}$$

En remplaçant (1.10) et (1.7) dans (1.4), on obtient

$$\pi(x)\sum_{y\neq x} q(x,y) = \sum_{y\neq x} \pi(y)q(y,x), \qquad \forall x \in \mathcal{X}. \tag{1.11}$$

Ce sont les *équations de balance globale* pour le processus markovien $(X(t))$. Ces équations sont équivalentes à

$$\begin{aligned} & \sum_{y\in\mathcal{X}} \pi(y)q(y,x) = 0, \qquad \forall x \in \mathcal{X} \\ \iff \quad & \pi^t Q = 0, \end{aligned}$$

où π^t est le vecteur de ligne représentant la probabilité π.

Si π satisfait les équations de balance globale (1.11), π est appelée la *probabilité invariante* du processus de Markov $(X(t))$. Dans ce cas, si $X(0)$ suit la loi π, $X(t)$ suit aussi cette loi π pour tout $t \geq 0$, et de plus, pour tout $s \geq 0$,

$$(X(s+t), t \geq 0) \stackrel{dist.}{=} (X(t), t \geq 0).$$

Alors $(X(t))$ est un *processus de Markov stationnaire* et π est sa *distribution stationnaire.*

Dans la section suivante, on considérera un exemple fondamental d'un processus de Markov, il s'agit d'une file d'attente $M/M/1$.

1.2 File d'attente $M/M/1$ FIFO

En 1909, Erlang a publié le premier article dans la théorie de files d'attente [Erl09]. Dans ce travail, Erlang a utilisé la théorie des probabilités pour dimensionner les conversations téléphoniques, à savoir il a calculé la probabilité qu'un certain nombre d'appels soient engendrés dans un certain intervalle de temps et le délai de répondre à un appel.

Ensuite, en 1953, une notation a été introduite par Kendall pour classifier les files d'attente en différents types [Ken53]. Ainsi, une file d'attente est notée par $A/B/C/D$ avec

- A : Distribution des intervalles d'inter-arrivée.
- B : Distribution des demandes de service.
- C : Nombre de serveurs.
- D : Nombre maximal de clients dans le système à un instant quelconque.

A et B peuvent être un des types de distribution suivants :
- M : Distribution exponentielle - Markovien.
- D : Distribution déterministe.
- E_k : Distribution d'Erlang, c.à.d. une somme de k variables aléatoires exponentielles.
- G : Distribution générale - Distribution arbitraire.

Dans cette section, on considérera la file $M/M/1$ où les intervalles d'inter-arrivée sont exponentiels, les demandes de service le sont aussi et il n'y a qu'un serveur dans le système.

Modèle

On considère une file d'attente de type $M/M/1$:

Arrivées poisonniennes. Les clients arrivent à la file suivant un processus de Poisson d'intensité ν, c.à.d. les intervalles entre les instants d'arrivée sont i.i.d. de loi exponentielle de moyenne ν^{-1}, notée $exp(\nu)$. On appelle ν le *taux d'arrivée* de clients dans la file.

Service exponentiel. Les clients demandent des services aléatoires i.i.d. de loi exponentielle de moyenne μ^{-1}, notée $exp(\mu)$. On appelle μ le *taux de service.*

Discipline de service. La file dispose d'un seul serveur de capacité unitaire, dont la discipline de service est Permier Arrivé Premier Servi (First In First Out - FIFO) : ce serveur donne toute sa capacité au plus ancien client.

L'évolution du système peut être représentée par un processus de sauts dont les transitions possibles sont une arrivée d'un nouveau client à la fin de la file ou un départ du client en tête de la file s'il y a au moins un client dans la file (Figure 1.6).

Ici on ne s'intéresse qu'au nombre de clients présents dans la file. Comme les intervalles d'inter-arrivée et les demandes de service sont tous exponentiels, le processus représentant ce nombre de clients est un processus de Markov d'une chaîne incluse homogène, qui est une marche aléatoire sur l'ensemble des nombres naturels $\mathbb{N}$ (voir la Figure 1.7).

FIG. 1.6 – Arrivée et départ dans une file $M/M/1$ FIFO.

Notations

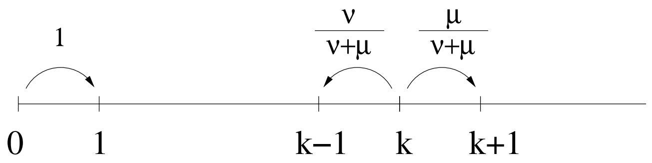

FIG. 1.7 – Graphe de la chaîne incluse du processus de sauts représentant une file $M/M/1$ FIFO.

État de la file : Le nombre de clients n. L'espace d'états est alors l'ensemble des nombres naturels $\mathbb{N}$.

Taux de charge : La proportion ρ entre le taux d'arrivée ν et le taux de service μ :

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

Distribution stationnaire : L'unique mesure de probabilité $\pi(n)$ (s'il existe) qui satisfait le système d'*équations de balance globale* :

$$\begin{aligned} \pi(0)\nu &= \pi(1)\mu, \\ \pi(n)(\nu + \mu) &= \pi(n-1)\nu + \pi(n+1)\mu, \qquad \forall n \geq 1 \end{aligned} \tag{1.12}$$

Dans cette file $M/M/1$, le système d'équations de balance globale est équivalent au système d'*équations de balance détaillée* suivant :

$$\pi(n)\nu = \pi(n+1)\mu, \qquad \forall n \geq 0, \tag{1.13}$$

dont l'unique solution est à forme géométrique et ne dépend que du taux de charge ρ :

$$\pi(n) = (1-\rho)\rho^n, \qquad \forall n \geq 0, \tag{1.14}$$

si la condition de stabilité est remplie :

$$\rho < 1. \tag{1.15}$$

Théorème 1.12 *Dans une file d'attente $M/M/1$ de taux de charge ρ,*
- *Si $\rho \geq 1$, la file est instable et il n'existe pas de mesure de probabilité invariante.*
- *Si $\rho < 1$, la file est stable et la distribution stationnaire du nombre de clients dans la file suit une loi géométrique de paramètre ρ :*

$$\pi(n) = (1-\rho)\rho^n, \qquad \forall n \in \mathbb{N}.$$

Par conséquence, il suffit de connaître le taux d'arrivée ν et le taux de service μ afin de pouvoir complètement mesurer un système $M/M/1$.

Le *nombre moyen de clients* dans la file ne dépend que du taux de charge ρ :

$$\mathbb{E}[n] = \sum_n n\pi(n) = \frac{\rho}{1-\rho}. \tag{1.16}$$

Ensuite, le résultat de Little [Lit61] dit que le nombre moyen de clients dans un système stable de files d'attente est donné par le produit de leur taux d'arivée ν et leur temps de séjour moyen $\mathbb{E}[T]$. Alors inversement, dans ce système $M/M/1$, le temps de séjour moyen est égal au nombre moyen de clients dans la file divisé par le taux d'arrivée ν :

$$\mathbb{E}[T] = \frac{\mathbb{E}[n]}{\nu} = \frac{1}{\mu - \nu}. \tag{1.17}$$

Remarquons que ce modèle requiert les arrivées poissonniennes et les demandes de service exponentielles afin d'avoir les expressions explicites ci-dessus. Une question très intéressante est naturellement posée : Y-a-t-il des disciplines de services, des modèles de files d'attente insensibles à la distribution des demandes de service, c.à.d. par exemple que l'expression de la distribution stationnaire du nombre de clients dans le système ne dépend pas de la loi des demandes de service ?

Une condition nécessaire de l'insensibilité a été donnée dans [FKAS82]. Cette condition, appelée la condition d'*attention instantanée*, dit qu'il faut allouer une proportion de service strictement positive au nouveau client à son arrivée. En particulier, dans le modèle d'une discipline de service FIFO, le serveur n'alloue aucun service au nouveau client s'il y a d'autres clients à son arrivée, alors cette discipline FIFO est sensible à la distribution des demandes de service : si les demandes de service ne sont pas exponentielles, on n'aura plus la forme géométrique de la distribution stationnaire du nombre de clients dans la file et dans ce cas, le système est très difficile à analyser.

Par contre, il existe des disciplines de service insensibles. Parmi ces disciplines, Processeur Partager (Processor Sharing - PS) et Dernier Arrivé Premier Servi (Last In First Out - LIFO) préemptive considérées par Baskett, Chandy, Muntz et Palacios [BCMP75] sont deux disciplines de service insensibles des plus connues et importantes dans la théorie des files d'attente. En 1979, Kelly a introduit dans son travail [Kel79] la discipline appelée *symétrique* définissant une large classe de disciplines de service insensibles qui contient les deux disciplines PS et LIFO mentionées ci-dessus. Dans la section suivante, on rappellera la définition de cette discipline de service symétrique et on montrera son insensibilité par deux méthodes différentes dont l'une est basée sur les demandes de service à phases [Kel79, CMP99] et l'autre est basée sur la théorie des Processus Semi-Markovien Généralisé avec Réallocations (*RGSMP*) [CMST98, Miy93, MSS95, Sch78c, Sch86, Sch78a].

1.3 Files d'attente symétriques

Dans ce modèle de file d'attente, la notion de positions (ou emplacements) dans la file d'attente est employée. Si un client trouve n clients dans la file à son arrivée, il est placé au hasard à une position l parmi $n+1$ positions $1, 2, \ldots, n+1$, les anciens clients aux positions

$l, l+1, \ldots, n$ doivent alors décaler vers la fin de la file, respectivement aux nouvelles positions $l+1, l+2, \ldots, n+1$. Lorsqu'un client en position l finit son service tandis qu'il y a n clients dans la file, il quitte la file et sa place est libérée, alors les clients aux positions $l+1, l+2, \ldots, n$, décalent vers la tête de la file, respectivement aux nouvelles positions $l, l+1, \ldots, n-1$.

Modèle

On considère une file d'attente symétrique $M/G/1$:

Arrivées poissoniennes. Les clients arrivent à la file suivant un processus de Poisson d'intensité ν, c.à.d. que les intervalles entre les instants d'arrivée sont i.i.d. de loi exponentielle de paramètre ν.

Positions dans la file. À l'arrivée, si $n-1$ autres clients se présentent dans la file, le nouveau client est placé à la position $l, l = 1, 2, \ldots, n$ avec la probabilité $\delta(l, n)$, les anciens clients aux positions $l, l+1, \ldots, n-1$ sont respectivement décalés vers la fin de la file, voir Figure 1.8. Ces *probabilités de positionnement* $\delta(l, n), l = 1, \ldots, n$, satisfont la condition suivante :

$$\sum_{l=1}^{n} \delta(l, n) = 1, \qquad \forall n \geq 1.$$

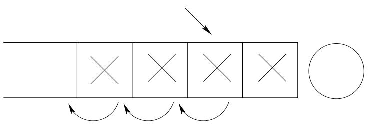

FIG. 1.8 – Décalages lors d'une arrivée.

Discipline de service. La file d'attente dispose d'un seul serveur de capacité constante 1. Si n clients se présentent dans la file, ce serveur alloue à chaque client une proportion de service en fonction de la position de ce client. La file est dite *symétrique* si cette proportion de service est exactement égale à la probabilité de positionnement $\delta(l, n)$ définie ci-dessus.

Lorsqu'un client en position $l, l = 1, \ldots, n$, finit son service et quitte la file, les clients derrière lui sont décalés vers la tête de la file, respectivement aux positions $l-1, \ldots, n-1$, prenant la place cédée par le client quittant, voir Figure 1.9.

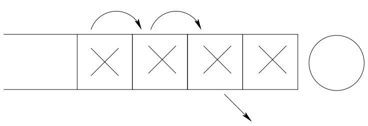

FIG. 1.9 – Décalages lors d'un départ.

Cette discipline de service symétrique en inclut la discipline PS, pour laquelle

$$\delta(l, n) = \frac{1}{n}, \qquad \forall n \geq 1, \forall l = 1, \ldots, n,$$

et la discipline LIFO préemptive, pour laquelle

$$\delta(1, n) = 1, \qquad \forall n \geq 1.$$

Remarquons tout d'abord que dans une file d'attente symétrique, la condition d'*attention instantanée* est satisfaite car à son arrivée, si le nouveau client est placé à une certaine position l avec la probabilité strictement positive $\delta(l, n)$, il est alloué une proportion strictement positive de service $\delta(l, n)$.

Maintenant, pour montrer l'insensibilité d'un système, il y a plusieurs méthodes possibles, par exemple, celle des phases utilisée par Kelly, Chao, Miyazawa et Pinedo [Kel79, CMP99], celle de Whittle utilisant l'équivalence entre la balance partielle sur un sous-ensemble $A \subset G$ et l'insensibilité à la distribution des demandes de service des clients dans A [Whi85, Whi86], et en fin, l'approche RGSMP (Processus Semi-Markovien Généralisé avec Réallocations) élaborée par Schassberger et Miyazawa [Sch86, Miy93]. Remarquons que la deuxième méthode est en fait une conséquence de la troisième. Alors on étudiera la première et la troisième en montrant l'insensibilité des files d'attente symétriques par la méthode des phases dans la Section 1.3.1 et puis par la méthode RGSMP dans la Section 1.3.2. Ensuite, afin de rester dans un cardre markovien, on empruntera la méthode des phases dans la plus part des preuves de l'insensibilité.

1.3.1 Méthode des phases

Demandes de service. Dans ce contexte, on suppose que les demandes de service sont i.i.d, mais elles suivent une loi générale arbitraire. Il est alors difficile de trouver la distribution stationnaire de la file. Afin de franchir cette difficulté, on suppose que ces demandes de service suivent un mélange de distributions d'Erlang [JT89, JT90a, JT90b, JT91], défini ci-dessous. Johnson et Taaffe ont montré qu'on peut approcher n'importe quelle distribution de support non-négatif par une suite de mélanges de distributions d'Erlang [JT89]. Alors comme dans les travaux de Kelly, Chao, Miyazawa et Pinedo [Kel79, CMP99], on peut conclure que le système est insensible à la distribution des demandes de service, au moins pour les mélanges de distributions d'Erlang. On peut appeler cette méthode *le raisonnement avec des services à phases.*

On suppose dans cette section que les clients requièrent des services aléatoires selon un mélange de distributions d'Erlang, c.à.d. que chaque demande de service est une somme d'un nombre aléatoire de phases exponentielles de moyenne μ^{-1}, $\mu > 0$. Pour tout $k \geq 1$, posons $s(k)$ la probabilité que le nombre de phases de service soit égal à k, $\bar{s}(k)$ la probabilité que le nombre de phases soit plus grand ou égal à k et $\bar{s}$ le nombre moyen de phases. On a :

$$\sum_{k \geq 1} s(k) = 1,$$

et

$$\bar{s}(k) = \sum_{j \geq k} s(j), \quad \bar{s} = \sum_{k \geq 1} k s(k) = \sum_{k \geq 1} \bar{s}(k).$$

Notons que la demande de service moyenne est égale à $\bar{s}/\mu$ et le taux de charge est donné par $\rho = \nu \bar{s}/\mu$. On suppose que la file est stable :

$$\rho = \frac{\nu \bar{s}}{\mu} < 1. \tag{1.18}$$

Macro-état. Soit n le nombre de clients présents dans la file. On s'aperçoit que ce macro-état ne suffit pas pour décrire l'évolution du système entier. Il faut considérer de plus les phases de service de tous les clients afin d'avoir un processus markovien décrivant le système.

Micro-état. Posons x_l la phase de service du client en position l, le micro-état de la file est défini par le vecteur de ligne $x = (x_1, \dots, x_n)$. Par exemple, $x_2 = 3$ veut dire que le client en position 2 est à sa troisième phase de service. On a : $n = |x|$, où $|x|$ signifie la longueur du vecteur x. Par convention, posons $x = \emptyset$ et $n = |\emptyset| = 0$ si la file est vide.

L'espace de micro-états est l'ensemble $\mathcal{X} = \{\emptyset\} \cup \{x = (x_1, \dots, x_n) : n \geq 1, x_i \geq 1, 1 \leq i \leq n\}$. Notons $\mathcal{X}(n) = \{x : |x| = n\}$ l'ensemble des micro-états correspondants au même macro-état n. Pour tous $x \in \mathcal{X}(n)$ et $l = 1, \dots, n+1$, notons $T^{k,l}x$ le micro-état obtenu à partir de x en ajoutant un client en $k^{\text{ème}}$ phase de service à la position l dans la file suivant la règle de décalage définie auparavant. De même, pour tous $x \in \mathcal{X}(n)$ et $l = 1, \dots, n$, notons $T_l x$ le micro-état obtenu à partir de x en effaçant le client en position l.

Théorème 1.13 *Dans une file d'attente symétrique stable (1.18), la distribution stationnaire du processus markovien décrivant le micro-état est donnée par l'expression explicite suivante :*

$$\pi(x) = (1-\rho)\left(\frac{\nu}{\mu}\right)^n \prod_{l=1}^{n} \bar{s}(x_l), \qquad \forall x \in \mathcal{X}.$$

Preuve. Les transitions de ce processus markovien sont nulles à part les transitions suivantes :

– Arrivée d'un nouveau client à la position l, menant la file de l'état x à l'état $T^{1,l}x$:

$$q(x, T^{1,l}x) = \nu\delta(l, n+1), \qquad l = 1, \dots, n+1.$$

– Changement de phase de service du client en position l, menant la file de l'état x à l'état $x + e_l$, où e_l désigne le vecteur unitaire de dimension n dont la $l^{\text{ème}}$ composante est égale à 1 et les autres sont nulles :

$$q(x, x+e_l) = \mu\delta(l, n)\frac{\bar{s}(x_l+1)}{\bar{s}(x_l)}, \qquad l = 1, \dots, n.$$

– Départ du client en position l, menant la file de l'état x à l'état $T_l x$:

$$q(x, T_l x) = \mu\delta(l, n)\frac{s(x_l)}{\bar{s}(x_l)}, \qquad l = 1, \dots, n.$$

À partir de ces probabilités de transition, on peut vérifier sans difficulté les équations de balance partielle pour la source :

$$\pi(x)q(x, T^{1,l}x) = \sum_{d \geq 1} \pi(T^{d,l}x)q(T^{d,l}x, x), \qquad \forall x \in \mathcal{X}, l = 1, \dots, n+1,$$

et celles pour la position l :

$$\pi(x)\Big(q(x,x+e_l)+q(x,T_lx)\Big) \quad = \quad \pi(T_lx)q(T_lx,x)+\pi(x-e_l)q(x-e_l,x),$$

pour tous $x \neq \emptyset$ et $l = 1, \dots, n$.

Ces équations de balance partielle assurent le système d'équations de balance globale :

$$\pi(x)\sum_{y\in\mathcal{X}} q(x,y) = \sum_{y\in\mathcal{X}} \pi(y)q(y,x).$$

Et ce système d'équations de balance globale montre que π est bel et bien la distribution stationnaire du processus markovien représentant le micro-état de la file symétrique.

□

Corollaire 1.14 *Dans une file d'attente symétrique stable, si les demandes de service sont i.i.d. de loi mélange de distributions d'Erlang, la distribution stationnaire du nombre de clients est à forme géométrique et donnée par :*

$$\pi'(n) = (1-\rho)\rho^n, \qquad n \geq 0.$$

Preuve. En sommant sur tout $x \in \mathcal{X}(n)$, on obtient l'expression de π' :

$$\begin{aligned}
\pi'(n) &= \sum_{x\in\mathcal{X}(n)} \pi(x) \\
&= \sum_{x_1,\dots,x_n} (1-\rho)\left(\frac{\nu}{\mu}\right)^n \prod_{l=1}^{n} \bar{s}(x_l) \\
&= (1-\rho)\left(\frac{\nu}{\mu}\right)^n \bar{s}^n \\
&= (1-\rho)\rho^n.
\end{aligned}$$

□

Remarquons que l'expression de la distribution stationnaire du nombre de clients ne dépend pas des paramètres $s, \bar{s}$ du mélange de distributions d'Erlang. Le système est donc insensible à la distribution des demandes de service, au moins pour les mélanges de distributions d'Erlang, la distribution stationnaire du nombre de clients est à forme géométrique et ne dépend que du taux de charge ρ :

$$\pi'(n) = (1-\rho)\rho^n, \qquad n \geq 0.$$

Le nombre moyen de clients dans la file est alors donné par

$$\mathbb{E}[n] = \frac{\rho}{1-\rho},$$

et le temps de séjour moyen d'un client est donné par

$$\mathbb{E}[T] = \frac{\bar{s}}{\mu(1-\rho)}.$$

Corollaire 1.15 ***(Temps de séjour conditionnel)*** *Le temps de séjour moyen d'un client est proportionnel au nombre de phases de service qu'il demande k et est donné par :*

$$\frac{k}{\mu(1-\rho)}.$$

Preuve. Notons N_l le nombre de clients en $l^{\text{ème}}$ phase de service dans la file. À partir de l'expression explicite de la distribution stationnaire π dans le Théorème 1.13, on peut déterminer la loi de cette v.a. N_l :

$$\begin{aligned}
\mathbb{P}(N_k = m) &= \sum_{n \geq m} \mathbb{P}(N_k = m \text{ et } |x| = n) \\
&= \sum_{n \geq m} (1-\rho)\Big(\frac{\nu}{\mu}\Big)^n \binom{n}{m} \bar{s}(l)^m \Big(\bar{s} - \bar{s}(l)\Big)^{n-m} \\
&= \Big(1 - \frac{\bar{s}(l)\nu/\mu}{1-\rho+\bar{s}(l)\nu/\mu}\Big)\Big(\frac{\bar{s}(l)\nu/\mu}{1-\rho+\bar{s}(l)\nu/\mu}\Big)^m, \quad \forall m \geq 0.
\end{aligned}$$

Le nombre moyen de clients en $l^{\text{ème}}$ phase dans la file est donc donné par

$$\mathbb{E}[N_l] = \frac{\bar{s}(l)\nu}{\mu(1-\rho)}.$$

D'après la loi de Little, le temps moyen qu'un client soit dans sa $l^{\text{ème}}$ phase de service sachant qu'il demande au moins l phases est donné par

$$\frac{\mathbb{E}[N_l]}{\nu \bar{s}(l)} = \frac{1}{\mu(1-\rho)}.$$

Remarquons que cette quantité ne dépend pas de la phase l. Alors sachant qu'un client demande k phases de service, sont temps de séjour moyen est donné par :

$$\frac{k}{\mu(1-\rho)}.$$

□

En considérant les phases infinitésimales, c.à.d. $\mu^{-1} \to 0$, on peut approcher n'importe quelle demande de service constante r. Et on peut déduire que le temps de séjour moyen d'un client est proportionnel à sa demande de service r et est donné par

$$\frac{r}{1-\rho}.$$

1.3.2 Méthode des processus semi-markoviens généralisés

Dans la section précédente, on a vu *la méthode des phases* utilisant l'approche d'une loi arbitraire de support non-négatif par des mélanges de distributions d'Erlang pour démontrer l'insensibilité des files d'attente simples et ultérieurement pour démontrer l'insensibilité des réseaux de files d'attente. Dans la théorie de l'insensibilité, il y une autre méthode, dite méthode des processus semi-markoviens généralisés (*GSMP*) [CMST98, Miy93, MSS95, Sch78c, Sch86, Sch78a] pour démontrer l'insensibilité de systèmes semi-markoviens généralisés avec des classes de clients.

La notion de GSMP a été introduite par K. Matthes en 1962 [Mat64] et a été ensuite développée par P. Franken, D. König et R. Schassbeger [FKAS82, Sch77, Sch78b, Sch78c, Sch86]. Il s'est avéré que même si cette notion permettait d'étudier l'insensibilité de plusieurs modèles, elle n'était en revanche pas capable de représenter les plus simples files d'attente dont l'insensibilité était connue. Même la file $M/GI/\infty$ échappait aux GSMP [Sch86].

Afin de modéliser les réseaux de files d'attente usuelles, une extention de GSMP a été proposée par Schassberger et Miyazawa [Sch86, Miy93]. Il s'agit de la notion de RGSMP (Processus Semi-Markovien Généralisé avec Réallocations). Dans cette section, on étudiera cette notion et son application à une file d'attente symétrique simple.

Notations de base

Soient G, S et D trois ensembles dénombrables. Les éléments de G, S et D sont appelés respectivement *macro-état*, *emplacement* et *type d'horloge*. Dans le modèle d'une file d'attente symétrique, on a respectivement :

Macro-états : n - le nombre de clients dans la file. L'espace des macro-états est $G = \mathbb{N}$.

Emplacements : une position $l > 0$, représentant le client à cette position dans la file, et la position virtuelle 0 représentant l'arrivée dans la file. L'ensemble des emplacements est $S = \mathbb{N}$.

À chaque macro-état $n \in G$ est associé un sous-ensemble fini $A(n) \in G$ appelé l'ensemble des emplacements *actifs*, c.à.d. des emplacements qui peuvent délencher une transition à l'état n. Dans la file symétrique, ce sous-ensemble $A(n)$ est $\{0, 1, \dots, n\}$ où l'emplacement 0 représente la prochaine arrivée dans la file et les emplacements $1, \dots, n$ représentent les n clients dans la file.

Types d'horloge : Chaque emplacement actif $l \in A(n)$ possède une horloge dont le type est noté par $\tau(l, n)$. Le type d'horloge détermine la distribution du temps de séjour nominal de l'horloge correspondant, ou encore la distribution de la demande de service du client correspondant dans la file symétrique.

Dans la file symétrique, la position 0 possède un type d'horloge indépendant du macro-état n, appelons a ce type d'horloge. De même, le type d'horloge des autres positions est indépendant de l'état n, appelons b ce type d'horloge :

$$\tau(0, n) = a, \ \forall n \geq 0, \quad \text{et} \quad \tau(l, n) = b, \ \forall n \geq 1, l = 1, \dots, n,$$

car les intervalles entre les arrivées successives sont i.i.d. et les demandes de service sont aussi i.i.d. dans notre exemple d'une file symétrique.

Transitions de macro-états

À chaque emplacement, il y a une horloge décomptant le temps de séjour nominal résiduel. Dans la file symétrique simple, le temps de séjour résiduel de la position virtuelle 0 est réduit avec la vitesse constante $c(0, n) = 1$ et celui de la position $l, l = 1, \dots, n$, est réduit avec la vitesse $c(l, n) = \delta(l, n)$. Pour simplifier les notations, notons $\delta(0, n)$ la vitesse constante 1, alors $c(l, n) = \delta(l, n)$ pour tous $n \geq 0$ et $l = 0, \dots, n$.

Dans le modèle d'une file d'attente symétrique, il n'y a que deux types de transitions, correspondant à deux types d'horloge :

Arrivée. Pour le type d'horloge a, si l'horloge à l'emplacement 0 s'épuise, une arrivée aura lieu : une nouvelle horloge sera créée à cet emplacement 0 déterminant le prochain instant d'arrivée et une autre horloge sera créée à un certain emplacement, décomptant le temps de séjour nominal du client qui vient d'arriver. Dans la file symétrique, cette dernière horloge est placée à l'emplacement l avec la probabilité $\delta(l, n+1)$, $l = 1, \dots, n+1$.

Le nouveau macro-état sera $n+1$ et l'ensemble des emplacements actifs sera $A(n+1) = \{0, 1, \dots, n+1\}$. Notons que le décalage des anciens clients correspond à une *réallocation bijective* d'horloges :

$$\Gamma_{(n,U,n+1,U')} : A(n) \setminus U \longrightarrow A(n+1) \setminus U', \tag{1.19}$$

où $U := \{0\}$ est l'ensemble des emplacements des horloges qui se sont épuisées et $U' := \{0, l\}$ est l'ensemble des emplacements des nouvelles horloges, de telle sorte que le type de chaque horloge reste *inchangé* :

$$\tau(l', n+1) = \tau(\Gamma^{-1}_{(n,U,n+1,U')}(l'), n), \qquad \forall l' \in A(n+1) \setminus U' \tag{1.20}$$

Dans notre modèle d'une file symétrique :

$$\Gamma_{(n,U,n+1,U')}(l') = \begin{cases} l' & \text{si } 1 \leq l' \leq l-1 \\ l'+1 & \text{si } l \leq l' \leq n \end{cases}$$

et le type de ces horloges reste égal à b.

Le taux de cette transition est donné par

$$p(n, U, n+1, U') = \delta(l, n+1).$$

Départ. Pour le type d'horloge b, si l'horloge à l'emplacement $l, l = 1, \dots, n$, s'épuise, un départ aura lieu : le macro-état sera $n-1$, l'ensemble des emplacements actifs deviendra $A(n-1) = \{0, \dots, n-1\}$. L'ensemble des emplacements des horloges qui s'épuisent sera $U = \{l\}$ et aucune nouvelle horloge ne sera créée dans cette transition : $U' = \emptyset$. Le décalage des autres clients correspond à une réallocation bijective d'horloges similaire :

$$\Gamma_{(n,U,n-1,U')} : A(n) \setminus U \longrightarrow A(n-1) \setminus U',$$

avec

$$\Gamma_{(n,U,n-1,U')}(l') = \begin{cases} l' & \text{si } 0 \leq l' \leq l-1 \\ l'-1 & \text{si } l+1 \leq l' \leq n \end{cases}$$

la condition (1.20) est toujours remplie car le type d'horloges réallouées reste toujours égal à b.

Dans ce cas, le taux de transition est égal à la constance 1 :

$$p(n, U, n-1, U') = 1.$$

On appelle le modèle formalisé ci-dessus un *Schéma Semi-Markovien Généralisé avec Réallocations* (RGSMS).

Notons $X(t)$ le macro-état du système à l'instant t, alors le processus stochastique $\{X(t); t \geq 0\}$ n'est markovien que dans peu de cas, à savoir, lorsque tous les temps de séjour nominaux sont indépendants et suivent des lois exponentielles. Ainsi, il faut en général considérer des informations supplémentaires afin de décrire complètement l'évolution du système.

Pour tout type d'horloge $d \in D$, notons F_d la fonction de distribution du temps de séjour nominal des horloges de type d. On suppose que

$$F_d(0+) = 0 \quad \text{et} \quad \mu_d^{-1} := \int_0^\infty (1 - F_d(u))du < \infty, \qquad \forall d \in D, \tag{1.21}$$

et que les temps de séjour nominaux sont indépendants. Dans notre modèle d'une file symétrique, on a les moyennes :

$$\mu_a^{-1} = \nu^{-1} \quad \text{et} \quad \mu_b^{-1} = \mu^{-1}.$$

Pour $X(t) = n$ et pour toute $l \in A(n)$, notons $R_l(t)$ le temps de séjour nominal résiduel de l'horloge à l'emplacement l à l'instant t et $Y(t) = (R_l(t); l \in A(X(t)))$. Posons $Z(t) = (X(t), Y(t))$ le micro-état du système à l'instant t.

Dans un premier temps, supposons que tous les temps de séjour nominaux suivent des lois exponentielles. Alors le processus $\{X(t); t \geq 0\}$ est markovien et dans ce cas, une mesure de probabilité $\pi(n)$ est la distribution stationnaire du processus markovien $\{X(t)\}$ si et seulement si π est solution du système d'équations de balance globale :

$$\Big(\sum_{l \in A(n)} c(l,n)\mu_{\tau(l,n)}\Big)\pi(n) = \sum_{n' \in G} \sum_{l' \in A(n')} \sum_{U \subset A(n)} c(l',n')\mu_{\tau(l',n')}\pi(n')p(n',U',n,U), \tag{1.22}$$

pour tout $n \in G$, où l'ensemble U' se compose d'un seul emplacement l' correspondant à l'horloge épuisée.

D'ailleurs, la condition :

$$\sum_{n \in G} \sum_{l \in A(n)} c(l,n)\mu_{\tau(l,n)}\pi(n) < \infty \tag{1.23}$$

est nécessaire et suffisante pour que la chaîne incluse du processus $\{X(t)\}$ considéré aux instants de saut soit récurrente positive.

Dans ce cas exponentiel, les équations de balance globale (1.22) pour une file symétrique sont équivalentes à celles pour une file d'attente $M/M/1$, elles admettent une unique solution :

$$\pi(n) = (1-\rho)\rho^n, \qquad \forall n \in G, \tag{1.24}$$

où $\rho = \nu/\mu$ désigne le taux de charge de la file.

Ensuite, retournons au cas général avec une famille de distributions arbitraires des demandes de service $\{F_d, d \in D\}$. On considérera de plus un système d'équations de balance, appelé le système d'*équations de balance locale.* Afin de définir ce système, notons $D' \subset D$ le sous-ensemble de types d'horloge tel que F_d soit exponentielle pour tout $d \in D'$, et notons $D'' = D \backslash D'$ l'ensemble complémentaire de types d'horloge tel que F_d soit une fonction de distribution arbitraire pour tout $d \in D''$. Et puis pour tout $n \in G$, notons $A'' = \{l : l \in A(n), \tau(l,n) \in D''\}$ l'ensemble des emplacements actifs dont le type d'horloge est dans D''. Alors le système d'équations de balance locale est donné par :

$$c(l,n)\mu_{\tau(l,n)}\pi(n) = \sum_{n' \in G} \sum_{l' \in A(n')} \sum_{U : l \in U} c(l',n')\mu_{\tau(l',n')}\pi(n')p(n',l',n,U), \tag{1.25}$$

pour tous $n \in G$ et $l \in A''(n)$.

Littéralement, ces équations de balance locale veulent dire que le flux de probabilité correspondant à la mort d'une horloge quelconque dans le macro-état n est égal au flux de probabilité correspondant à la naissance de cette horloge *menant* le système au même état n.

Jansen, König, Miyazawa et Schassberger [KJ77, Miy93, Sch78c] ont montré que les équations de balance globale, la récurrence positive et les équations de balance locale définies ci-dessus sont nécessaires et suffisantes pour la *décomposabilité en forme produit* de la distribution stationnaire du processus $\{Z(t)\}$.

Proposition 1.16 *Soit $\{\pi(n), n \in G\}$ une mesure de probabilité. Alors π satisfait les conditions (1.22), (1.23) et (1.25) si et seulement si la distribution stationnaire du processus $\{Z(t)\}$ est :*

$$P\Big(X(t) = n, R_l(t) \le x_l ; l \in A(n)\Big) = \pi(n) \prod_{l \in A(n)} F^{(r)}_{\tau(l,n)}(x_l) \tag{1.26}$$

et l'intensité du processus ponctuel considéré aux instants de sauts de $\{Z(t)\}$ est finie, où $F_d^{(r)}(x)$ est donnée par $\mu_d \int_0^x (1 - F_d(u))du$. Dans ce cas, on dit que la distribution stationnaire de $\{Z(t)\}$ est décomposable en forme produit.

Remarque 1.1 *Si le système est décomposable en forme produit, la mesure de probabilité π, solution des systèmes d'équations de balance (1.22) et (1.25), est la distribution stationnaire du nombre de clients dans le système. Cette distribution stationnaire ne dépend des fonctions de distribution $\{F_d, d \in D''\}$ qu'à travers leurs moyennes $\{\mu_d^{-1}, d \in D''\}$ et on dit que le système est insensible par rapport à l'ensemble D''.*

Cette remarque est simplement déduite de la Proposition 1.16 en faisant tendre les x_l vers ∞ dans l'expression (1.26). Notons que la décomposabilité implique l'insensibilité à la distribution des temps de séjour nominaux mais la réciproque est fausse. On peut trouver dans la Section 2.4.2 un exemple intéressant d'une file d'attente avec un nombre fini de places où l'on a l'insensibilité mais pas la décomposabilité.

Appliquons maintenant ces résultats pour notre modèle d'une file d'attente symétrique avec $D'' = \{1, \ldots, n\}$. Cet ensemble représente tous les clients dans la file, on suppose que la position virtuelle 0 possède une horloge exponentielle, c.à.d. que le processus d'arrivées est de Poisson. La mesure de probabilité $\pi(n) = (1-\rho)\rho^n$ définie en (1.24) satisfait la condition de récurrence positive (1.23) :

$$\nu\pi(0) + \sum_{n\geq 1}(\nu+\mu)\pi(n) < \infty.$$

On observe que la vitesse $c(l,n)$ de l'horloge en position l est égale à la probabilité de transition $p(n-1,\{0\},n,\{0,l\})$, grâce à la symétrie de la file. Alors les équations de balance locale (1.25) sont remplies :

$$\mu c(l,n)\pi(n) = \nu p(n-1,\{0\},n,\{0,l\})\pi(n-1), \qquad \forall n \in G, l \in A''(n).$$

D'après la Proposition 1.16, dans une file d'attente symétrique, la distribution stationnaire du processus markovien décrivant le micro-état de la file est *décomposable en forme produit*. Et par conséquence, d'après la Remarque 1.1, la distribution stationnaire du nombre de clients est la mesure de probabilité π, indépendamment de la distribution des demandes de service. Donc, la file symétrique est insensible à la distribution des demandes de service.

1.4 Réseaux de files d'attente

On a étudié les files d'attente simples commes les files $M/M/1$ et les files symétriques. Mais en pratique, on rencontre souvent un réseau de files d'attente et pas une file d'attente simple. Imaginons que les files d'attente sont connectées d'une certaine manière. Considérons par exemple une file PS et une file LIFO préemptive en tandem, voir Figure 1.10.

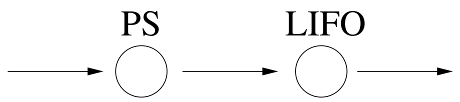

FIG. 1.10 – Réseau d'une file PS et une file LIFO préemptive en tandem.

Dans ce réseau simple de deux files d'attente, les clients arrivent à la première file, y demandent des services i.i.d. de loi arbitraire et partagent équitablement la capacité totale de cette file. Une fois qu'ils finissent leur service à la première file, ils continuent leur route à la deuxième file, et encore une fois, ils demandent des services de loi arbitraire, mais dans cette deuxième file, le dernier client est servi avec la capacité totale avant de quitter le réseau. Comme les disciplines PS et LIFO préemptive font partie des disciplines symétriques, elles sont insensibles. Ce fait implique-t-il l'insensibilité du réseau de deux files PS et LIFO préemptive en tandem ?

On considérera dans cette partie des réseaux de files d'attente, en particulier, des réseaux de Jackson et des réseaux de Whittle. On montrera l'insensibilité des réseaux de Whittle de files PS et puis la condition nécessaire et suffisante de l'insensibilité d'un réseau de Jackson de files PS avec capacités de service dépendantes de l'état.

1.4.1 Réseaux de Jackson

Dans cette partie, la définition des réseaux de Jackson de files d'attente $M/M/1$ sera donnée et quelques notations seront introduites. En particulier, on trouvera ici la définition du routage de Jackson, des équations de trafic et des capacités de service dépendantes de l'état du réseau.

Considérons un réseau ouvert de I files d'attente :

Arrivées. À chaque file i, les clients arrivent suivant un processus de Poisson d'intensité ν_i.

Demandes de service. Les clients dans la file i demandent des services exponentiels i.i.d. de moyenne $1/\mu_i$, $i = 1, \ldots, I$.

Routage de Jackson. Après l'accomplissement de service à la file i, un client est dirigé vers la file j avec la probabilité de routage p_{ij} et quitte le réseau avec la probabilité $p_i = 1 - \sum_{j=1}^{I} p_{ij}$. Un exemple de routage de Jackson est illustré dans la Figure 1.11.

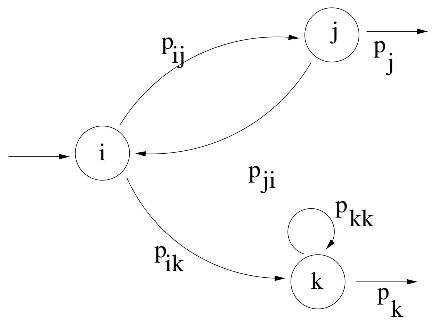

FIG. 1.11 – Routage de Jackson.

Équations de trafic. Le *taux d'arrivée effectif* λ_i à la file i est alors défini par les *équations de trafic* :

$$\lambda_i = \nu_i + \sum_{j=1}^{I} \lambda_j p_{ji}, \qquad i = 1, \ldots, I. \tag{1.27}$$

Dans la suite, on suppose que ces équations de trafic admettent une unique solution $(\lambda_1, \ldots, \lambda_I)$. Supposons de plus que tous les clients viennent d'une *source* unique. En sommant les équations de trafic (1.27), on obtient donc une nouvelle équation de trafic pour cette source :

$$\sum_{i=1}^{I} \nu_i = \sum_{i=1}^{I} \lambda_i p_i.$$

On parlera ultérieurement aussi des équations de balance partielle pour cette source.

Le modèle défini ci-dessus est appelé un *réseau de Jackson* de files d'attente $M/M/1$.

Notons $\rho_i = \lambda_i/\mu_i$ le *taux de charge* de la file i.

État du réseau. Notons n_i le nombre de clients dans la file i et $n = (n_1, \ldots, n_I)$ l'état du réseau. Alors l'espace d'états est $G = \mathbb{N}^I$.

Capacités unitaires. Dans ce modèle de réseau de Jackson, supposons que la capacité de service de chaque file est égale à 1.

Comme le processus d'arrivée est poissonnien et les demandes de service sont i.i.d. de loi exponentielle, le processus stochastique $\big(X(t), t \geq 0\big)$ décrivant l'évolution du nombre de clients à chaque file est un processus markovien. Soit e_i le vecteur unitaire dont la $i^{\text{ème}}$ composante est égale à 1 et les autres sont nulles. Alors lorsque le réseau est dans l'état n, les taux de transition sont nuls à part les trois transitions suivantes :

– une arrivée à la file i menant le réseau à l'état $n + e_i$ avec le taux de transition

$$q(n, n + e_i) = \nu_i, \qquad n \in G,$$

– un départ de la file i menant le réseau à l'état $n - e_i$ avec le taux de transition

$$q(n, n - e_i) = \phi_i(n)\mu_i p_i, \qquad n_i \geq 1,$$

– un mouvement d'un client à la file i vers la file j menant le réseau à l'état $n - e_i + e_j$ avec le taux de transition

$$q(n, n - e_i + e_j) = \phi_i(n)\mu_i p_{ij}, \qquad n_i \geq 1.$$

Alors π est la distribution stationnaire du nombre de clients à chaque file si et seulement si elle remplit les équations de balance globale :

$$\pi(n)\sum_{i=1}^{I}(\nu_i + \mu_i) = \sum_{i=1}^{I}\pi(n - e_i)\nu_i + \sum_{i=1}^{I}\pi(n + e_i)\mu_i p_i + \sum_{i,j}\pi(n - e_i + e_j)\mu_j p_{ji},$$

pour tout $n \in G$. À partir d'ici, on note par convention $\pi(n) = 0$ si n n'est pas dans l'espace d'états G.

Capacités de service dépendantes de l'état du réseau.

Les réseaux de files d'attente peuvent représenter des réseaux de données arbitraires. La capacité d'une file dans ce cas correspond à la bande passante allouée aux flots de données sur une route particulière dans le réseau de données. Alors ces flots doivent partager les liens de connexion communs avec les autres flots de la même route et aussi avec les flots d'autres routes. Par conséquence, la capacité de service d'un noeud dépend en général du nombre de clients à chaque noeud. D'où l'importance et l'intérêt d'étudier les réseaux de files d'attente avec capacités de service dépendantes du nombre de clients à chaque file.

Supposons maintenant que la capacité de chaque file d'attente i est une fonction $\phi_i(n)$ dépendante de l'état du système n. Supposons que $\phi_i(n) > 0$ si et seulement si $n_i > 0$.

Dans ce nouveau réseau avec capacités de service dépendantes de l'état, π est la distribution stationnaire du nombre de clients à chaque file si et seulement si elle remplit les équations de balance globale :

$$\begin{aligned}\pi(n)\sum_{i=1}^{I}\left(\nu_i+\phi_i(n)\mu_i\right) &= \sum_{i=1}^{I}\pi(n-e_i)\nu_i+\sum_{i=1}^{I}\pi(n+e_i)\phi_i(n+e_i)\mu_i p_i \\ &\quad +\sum_{i,j}\pi(n-e_i+e_j)\phi_j(n-e_i+e_j)\mu_j p_{ji}, \qquad (1.28)\end{aligned}$$

pour tout $n \in G$.

1.4.2 Réseaux de Whittle.

Une classe particulière de réseaux de files d'attente, connue sous le nom de réseaux de Whittle [Ser99], est caractérisée par la *propriété d'équilibre* suivante.

Définition 1.17 *Les capacités de service dans un réseau de files d'attente sont dites équilibrées si :*

$$\phi_i(n)\phi_j(n-e_i)=\phi_j(n)\phi_i(n-e_j), \qquad i,j=1,\dots,I : n_i>0, n_j>0.$$

Soit $< n, n-e_{i_1},\dots,n-e_{i_1}-\dots-e_{i_{m-1}},0 >$ un chemin direct de l'état n à l'état 0, de longueur m où $m=n_1+\dots+n_I$ donne le nombre total de clients dans l'état n. La propriété d'équilibre impose que l'expression

$$\Phi(n)=\frac{1}{\phi_{i_1}(n)\phi_{i_2}(n-e_{i_1})\dots\phi_{i_m}(n-e_{i_1}-\dots-e_{i_{m-1}})} \qquad (1.29)$$

est indépendante du chemin considéré. En particulier, les capacités $\phi_i(n)$ sont exclusivement caractérisées par la fonction d'équilibre Φ :

$$\phi_i(n)=\frac{\Phi(n-e_i)}{\Phi(n)}, \quad i=1,\dots,I, x_i>0. \qquad (1.30)$$

Réciproquement, s'il existe une fonction Φ telle que les capacités satisfont (1.30), ces capacités sont équilibrées. On dit que ces capacités sont équilibrées par la fonction Φ.

On a le résultat clef suivant.

Théorème 1.18 *Dans un réseau de Whittle, la distribution stationnaire du nombre de clients à chaque file est à forme produit et donnée par :*

$$\pi(n)=C\Phi(n)\prod_{i=1}^{I}\rho_i^{n_i}, \qquad n\in G,$$

si le réseau est stable, c'est à dire

$$\sum_{n\in G} \Phi(n) \prod_{i=1}^{I} \rho_i^{n_i} < \infty,$$

et C est la constance de normalisation :

$$C = \Big(\sum_{n\in G} \Phi(n) \prod_{i=1}^{I} \rho_i^{n_i} \Big)^{-1}.$$

Preuve. La démonstration consiste à vérifier les équations de balance partielle pour la source :

$$\pi(n) \sum_{i=1}^{I} \nu_i = \sum_{i=1}^{I} \pi(n+e_i)\phi_i(n+e_i)\mu_i p_i, \qquad \forall n \in G,$$

et celles pour chaque file $i, i = 1, \ldots, I$:

$$\pi(n)\phi_i(n)\mu_i = \pi(n-e_i)\nu_i + \sum_{j=1}^{I} \pi(n-e_i+e_j)\phi_j(n-e_i+e_j)\mu_j p_{ji}, \qquad \forall n \in G.$$

En tenant en compte que les capacités $\phi_i(n)$ sont équilibrées par la fonction $\Phi(n)$, ces équations de balance partielle sont équivalentes aux équations de trafic, respectivement :

$$\sum_{i=1}^{I} \nu_i = \sum_{i=1}^{I} \lambda_i p_i,$$

et

$$\lambda_i = \nu_i + \sum_{j=1}^{I} \lambda_j p_{ji}.$$

□

1.4.3 Insensibilité de la discipline PS

Les files d'attente PS ont été traditionnellement utilisées pour décrire des systèmes multi-accès d'ordinateurs [Kle75], ou encore pour décrire le processus de fonctionnement d'une unité centrale de traitement dans un ordinateur, où les tâches partagent équitablement la capacité de cette unité centrale. On peut voir ultérieurement dans le Chapitre 4 que ces files PS sont utilisées récemment dans les réseaux de donnée pour modéliser les liens de connexion dont les capacités sont partagées équitablement par les flots de données qui passent par ces liens [BFBP$^+$01, BM01, BP02a, BPRR01, MR00, BT07a].

Dans cette section, on considérera les réseaux de files PS avec routage de Jackson et avec capacités dépendantes de l'état. On montrera tout d'abord l'insensibilité des réseaux de Whittle de files PS et ensuite, on remontrera qu'un réseau de files PS est insensible si et seulement s'il s'agit d'un réseau de Whittle de files PS. Cette équivalence ont été établie par Bonald et Proutière en 2002 [BP02b].

D'après le Théorème 1.18, les réseaux de Whittle sont insensibles aux taux d'arrivée, aux taux de service et aux probabilités de routage, c.à.d. que la distribution stationnaire du processus markovien décrivant l'état du système n ne dépend de ces quantités qu'à travers les taux de charge $\rho_1, \ldots, \rho_I$.

Les réseaux de Whittle sont aussi connus pour être insensibles à la distribution des demandes de service. Pour la classe des distributions de Cox, qui forme un sous-ensemble dense de l'ensemble de toutes les distributions de variables aléatoires positives, cette propriété est une conséquence directe du résultat suivant.

Proposition 1.19 *Considérons un réseau de files PS tel que la capacité totale de deux files, par exemple, 1 et 2, ne dépend des nombres de clients présents à ces files qu'à travers leur somme et est partagée équitablement à leurs clients, c.à.d. que*

$$\frac{\phi_1(n)}{n_1} = \frac{\phi_2(n)}{n_2} = \frac{\phi_1(n) + \phi_2(n)}{n_1 + n_2}, \qquad n_1 > 0, n_2 > 0.$$

Alors les capacités $\phi_1, \ldots, \phi_I$ sont équilibrées par la fonction Φ si et seulement si les fonctions $\phi_1 + \phi_2, \phi_3, \ldots, \phi_I$ sont équilibrées par $\tilde{\Phi}$, avec

$$\Phi(n) = \binom{n_1 + n_2}{n_1} \tilde{\Phi}(\tilde{n}), \qquad \tilde{n} = (n_1 + n_2, n_3, \ldots, n_I).$$

Preuve. La démonstration est déduite directement de la caractérisation de la propriété d'équilibre (1.30).

□

Considérons un réseau de Whittle tel que la capacité totale de deux files, par exemple, 1 et 2, ne dépend des nombres de clients présents à ces files qu'à travers leur somme et est partagée équitablement à leurs clients. D'après la Proposition 1.19 et le Théorème 1.18, la distribution stationnaire du processus markovien décrivant l'état du système est donnée par

$$\pi(n) = C \binom{n_1 + n_2}{n_1} \tilde{\Phi}(\tilde{n}) \prod_{i=1}^{I} \rho_i^{n_i},$$

où C est la constance de normalisation.

En particulier, la distribution stationnaire $\tilde{\pi}$ des nombres de clients aux files $1 + 2, 3, \ldots, I$ est donnée par

$$\tilde{\pi}(\tilde{n}) = C \tilde{\Phi}(\tilde{n}) (\rho_1 + \rho_2)^{n_1 + n_2} \prod_{i=3}^{I} \rho_i^{n_i}, \tag{1.31}$$

où C est la constance de normalisation.

Considérons maintenant un réseau de Whittle où les distributions exponentielles des demandes de service sont remplacées par une distribution de Cox de k_i phases exponentielles i.i.d.

d'une même moyenne. Alors la file d'attente i avec les demandes de service de distribution de Cox de k_i phases est équivalente à un réseau linéaire de k_i files PS avec les demandes de service exponentielles. De plus, la capacité totale de ces k_i files ne dépend des nombres de clients présents à ces files quà travers leur somme et est partagé équitablement à tous les clients. D'après la Proposition 1.19, c'est encore un réseau de Whittle avec les capacités équilibrées. Alors d'après la propriété que l'on vient d'établir (1.31), la distribution stationnaire du nombre de clients à chaque file dans notre réseau initial reste inchangée.

En conclusion, les réseaux de Whittle sont insensibles aux distributions des demandes de service, au moins pour la classe des distributions de Cox.

Dans la suite, on établira la condition nécessaire et suffisante pour qu'un réseau de Jackson de files PS avec capacités dépendantes de l'état soit insensible à la distribution des demandes de service.

Théorème 1.20 *Un réseau de files PS est un réseau de Whittle si et seulement si la distribution stationnaire du nombre de clients à chaque file n reste inchangée lorsque les clients demandent des services de loi hyperexponentielle de paramètre θ_i et de moyenne $1/\mu_i$, c.à.d. un service exponentiel de moyenne $1/\theta_i \times 1/\mu_i$ avec la probabilité θ_i et un service nul avec la probabilité $1-\theta_i$, pour tous i et $0<\theta_i<1$.*

Preuve. On a montré auparavant qu'un réseau de Whittle est insensible à la distribution des demandes de service. Dans un réseau de Whittle, la distribution stationnaire de l'état n n'est donc pas changée lorsque les clients demandent des services hyperexponentiels au lieu des services exponentiels.

Il ne nous reste plus qu'à démontrer le sens inverse. Supposons que la distribution stationnaire du nombre de clients à chaque file est inchangée lors que les clients demandes des services hyperexponentiels. On montrera que les capacités $\phi_i(n)$ sont équilibrées afin de conclure que le réseau est de Whittle.

Lors que les demandes de service sont hyperexponentielles, le système est équivalent à un nouveau réseau de file PS avec les demandes de service exponentielles de moyenne $1/\theta_i\mu_i$ avec les nouveaux taux d'arrivée et avec le nouveau routage :

- Taux d'arrivée : $\bar{\nu}_i = \theta_i \nu_i$.
- Routage : $\bar{p}_{ij} = \theta_j p_{ij} + \sum_k p_{ik}(1-\theta_k)\bar{p}_{kj}$.
- Les nouveaux taux d'arrivée effectifs sont

$$\bar{\lambda}_i = \bar{\nu}_i + \sum_j \bar{\lambda}_j \bar{p}_{ji}.$$

Alors les nouveaux taux de charge sont donnés par

$$\bar{\rho}_i = \frac{\bar{\lambda}_i}{\theta_i \mu_i}.$$

Remarquons que la distribution stationnaire du nombre de clients à chaque file reste inchangée si et seulement si ces nouveaux taux de charge $\bar{\rho}$ sont égaux aux anciens taux de charge ρ, ce qui est équivalent à dire que le nouveau taux d'arrivée doit satisfaire : $\bar{\lambda}_i = \theta_i \lambda_i$ pour toute i. Cette condition nous suggère qu'en faisant θ_1 vers 0, le réseau asymptotique est équivalent à un réseau partiel de $I-1$ files $2, \ldots, I$. Dans la suite, on utilisera cette remarque pour montrer par récurrence sur le nombre de files dans le réseau que les capacités $\phi_i(n)$ sont équilibrées :

$$\phi_i(n) = \frac{\Phi(n-e_i)}{\Phi(n)}, \qquad i = 1, \ldots, I, \tag{1.32}$$

par la fonction

$$\Phi(n) = \frac{\pi(n)}{\prod_{i=1}^{I} \rho_i^{n_i}}.$$

Pour $I = 1$, le réseau consiste en une file PS simple de capacité $\phi_1(n)$ et une mesure stationnaire est donnée par

$$\pi(n) = \frac{1}{\prod_{l=1}^{n_1} \phi_1(l)} \rho_1^{n_1}.$$

Alors évidemment, la capacité $\phi_1(n)$ satisfait la propriété d'équilibre (1.32) avec :

$$\Phi(n) = \frac{\pi(n)}{\rho_1^{n_1}}.$$

Supposons que le résultat est valable pour $I-1$. Considérons un réseau de I files PS.

Remarquons tout d'arbord que pour les probabilités de routage, si $p_{ii} > 0$ pour une certaine file i, l'état du réseau est en fait le même que celui d'un réseau avec taux de service $\tilde{\mu_i} = \mu_i(1-p_{ii})$ et avec probabilités de routage $\tilde{p}_{ij} = p_{ij}/(1-p_{ij})$ si $j \neq i$, et $\tilde{p}_{ii} = 0$. Alors on peut supposer que les probabilités de routage $p_{ii} = 0$ pour toutes $i = 1, \ldots, I$.

Supposons que la distribution stationnaire du nombre de clients à chaque file est $\pi(n)$ lorsque les demandes de service à la file i sont i.i.d. de loi hyperexponentielle de paramètre θ_i et de moyenne $1/\mu_i$ pour toutes paramètres $\theta_1, \ldots, \theta_I$.

Faisons θ_1 tendre vers 0 alors le système asymptotique est le réseau partiel de $I-1$ files PS $2, 3, \ldots, I$ avec de nouveaux taux d'arrivée et routage :

$$\tilde{\nu}_i = \nu_i + \nu_1 p_{1i}, \quad \tilde{p}_{ij} = p_{ij} + p_{i1}p_{1j}, \quad \text{et} \quad \tilde{p}_i = 1 - \sum_{j \neq 1} \tilde{p}_{ij}, \qquad i, j \neq 1.$$

Les équation de trafic (1.27) deviennent donc

$$\begin{aligned} \tilde{\lambda}_i &= \tilde{\nu}_i + \sum_{j=2}^{I} \tilde{\lambda}_j \tilde{p}_{ji} \\ &= \nu_i + \sum_{j=2}^{I} \tilde{\lambda}_j p_{ji} + (\nu_1 + \sum_{j=2}^{I} \tilde{\lambda}_j p_{j1}) p_{1i}, \qquad i = 1, \ldots, I. \end{aligned}$$

Ces équations de trafic admettent une solution unique $(\lambda_2, \ldots, \lambda_I)$, où $\lambda_2, \ldots, \lambda_I$ sont les taux d'arrivée effectifs de notre réseau initial. Cela implique que les taux d'arrivée effectifs $\lambda_2, \ldots, \lambda_I$

du réseau partiel de $I-1$ files sont les mêmes du réseau initial. D'autre part, les moyennes de service $\mu_2^{-1}, \dots, \mu_I^{-1}$ sont les mêmes, alors les taux de charge $\rho_2, \dots, \rho_I$ le sont aussi.

Par conséquence, pour tout n_1 fixé, $\pi(n_1, \cdot)$ est une mesure stationnaire du nombre de clients à chaque file pour le réseau partiel de $I-1$ files PS $2, \dots, I$, de capacités $\phi_2(n_1, \cdot), \dots, \phi_I(n_1, \cdot)$. Cette mesure stationnaire est valide pour toutes demandes de service hyperexponentielles. Par l'hypothèse de récurrence, ces capacités sont équilibrées par la fonction

$$\Phi(n_1, \cdot) = \frac{\pi(n_1, \cdot)}{\prod_{i=2}^{I} \rho_i^{n_i}}.$$

En multipliant cette fonction par la constance $1/\rho_1^{n_1}$, on obtient à nouveau une fonction d'équilibre

$$\Phi(n) = \frac{\pi(n)}{\prod_{i=1}^{I} \rho_i^{n_i}},$$

qui satisfait

$$\phi_i(n) = \frac{\Phi(n - e_i)}{\Phi(n)}, \qquad \forall i \neq 1.$$

De plus, comme les files sont toutes équivalentes, on obtient enfin

$$\phi_i(n) = \frac{\Phi(n - e_i)}{\Phi(n)}, \qquad i = 1, \dots, I,$$

ce qui montre (1.32) pour les réseaux de I files PS.

En conclusion, un réseau de files PS est un réseau de Whittle si et seulement si la distribution stationnaire de l'état n reste inchagée lorsque les clients demandent des services hyperexponentiels de paramètre θ_i et de moyenne $1/\mu_i$.

$\square$

On déduit de l'insensibilité des réseaux de Whittle le résultat suivant

Corollaire 1.21 *Il y a équivalence entre*

1. *Un réseau de files PS est un réseau de Whittle de files PS, c.à.d. que les capacités de service sont équilibrées.*
2. *La distribution stationnaire du nombre de clients à chaque file est insensible à la distribution des demandes de service.*
3. *Cette distribution stationnaire ne dépend des taux d'arrivée, taux de service et probabilités de routage qu'à travers les taux de charge $\rho_1, \dots, \rho_I$.*

Preuve. On a montré qu'un réseau de Whittle de files d'attente PS est insensible à la distribution des demandes de service alors la première proposition implique la deuxième. Réciproquement, la deuxième proposition implique que la distribution stationnaire de l'état n reste inchangée lorsque les clients demandent des services hyperexponentiels, alors d'après le Théorème 1.20, ceci est un réseau de Whitlle de files d'attente PS. Par conséquence, 1 et 2 sont équivalentes.

D'ailleurs, on a montré qu'un réseau de Whittle de files PS ne dépend des taux d'arrivée, taux de service et probabilités de routage qu'à travers les taux de charge alors 1 implique 3. Il ne nous reste plus qu'à montrer que 3 implique 1 pour conclure la démonstration.

Supposons que 3 est satisfaite. Modifions les taux d'arrivée, taux de service et probabilités de routage comme suite :

- Taux d'arrivée $\tilde{\nu}_1 = \nu_1\theta_1$ et $\tilde{\nu}_i = \nu_i + \nu_1(1-\theta_1)p_{1i}$ pour $i \neq 1$.
- Taux de service $\tilde{\mu}_1 = \mu_1\theta_1$ et $\tilde{\mu}_i = \mu_i$ pour $i \neq 1$.
- Probabilités de routage $\tilde{p}_{1i} = p_{1i}$, $\tilde{p}_{i1} = p_{i1}$ et $\tilde{p}_{ij} = p_{ij} + p_{i1}(1-\theta_1)p_{1j}$ pour $i, j \neq 1$.

Alors on obtient un nouveau réseau qui est équivalent au réseau initial avec demandes de service hyperexponentielles à la file 1. La proposition 3 implique que ce réseau admet la même distribution stationnaire de l'état n que le réseau initial. De plus, toutes les files ont le même rôle, alors le réseau initial satisfait les conditions du Théorème 1.20, donc c'est un réseau de Whittle de files PS.

□

Bonald et Proutière [BP02b] ont montré que le temps de séjour moyen d'un client à la file i est proportionnel à sa demande de service r et est donné par :

$$r\frac{\mathbb{E}[n_i]}{\rho_i}.$$

1.4.4 Taux d'arrivée et routage dépendants de l'état

Supposons maintenant que les taux d'arrivée et les probabilités de routage sont aussi des fonctions de l'état du réseau : $\nu_i(n)$ et $p_{ij}(n), p_i(n)$. Les taux d'arrivée effectifs $\lambda_i(n)$ sont définis par les équations de trafic :

$$\lambda_i(n) = \nu_i(n) + \sum_{j=1}^{I} \lambda_j(n)p_{ji}(n+e_j).$$

On suppose que ces équations de trafic admettent une unique solution $(\lambda_1(n), \ldots, \lambda_I(n))$.

Les taux de charge dépendront aussi de l'état n :

$$\rho_i(n) = \frac{\lambda_i(n)}{\mu_i}, \qquad \forall n, \forall i.$$

Posons ensuite

$$\psi_i(n) = \frac{\rho_i(n-e_i)}{\phi_i(n)}, \qquad n_i > 0, \tag{1.33}$$

on obtient le résultat suivant

Théorème 1.22 *Si les fonctions $\psi_1(n), \ldots, \psi_I(n)$ sont équilibrées par une fonction positive $\Psi(n)$, la distribution stationnaire du processus markovien décrivant l'état du réseau est donnée par*

$$\pi(n) = C\Psi(n)^{-1}, \qquad n \in G,$$

où C est la constance de normalisation :

$$C = \Big(\sum_{n \in G} \Psi(n)^{-1} \Big)^{-1},$$

si la condition de stabilité est remplie : $0 < C < \infty$.

Preuve. Comme les taux d'arrivée et le routage sont en fonction de l'état, les équations de balance partielle deviennent, pour la *source*

$$\pi(n) \sum_{i=1}^{I} \nu_i(n) = \sum_{i=1}^{I} \pi(n+e_i)\phi_i(n+e_i)\mu_i p_i(n+e_i), \qquad \forall n \in G,$$

et pour chaque file $i, i = 1, \ldots, I$:

$$\pi(n)\phi_i(n)\mu_i = \pi(n-e_i)\nu_i(n-e_i) + \sum_{j=1}^{I} \pi(n-e_i+e_j)\phi_j(n-e_i+e_j)\mu_j p_{ji}(n-e_i+e_j), \qquad \forall n \in G.$$

Comme dans le Théorème 1.18, ces équations de balance partielle sont équivalentes aux équations de trafic tant que les fonctions $\psi_1(n), \ldots, \psi_I(n)$ sont équilibrées par la fonction $\Psi(n)$. $\square$

Remarque 1.2 *Si les taux de service $\phi_1(n), \ldots, \phi_I(n)$ sont équilibrés par $\Phi(n)$, les fonctions $\psi_1(n)$, …, $\psi_I(n)$ sont équilibrées par $\Psi(n)$ si et seulement si les taux de charge $\rho_1(n-e_1), \ldots,$ $\rho_I(n-e_I)$ sont équilibrés par $\Phi(n) \times \Psi(n)$.*

Proposition 1.23 *Supposons que la capacité totale de deux files, par exemple 1 et 2, ne dépend du nombre de clients présents à ces deux files qu'à travers leur somme et est équitablement partagée à leurs clients, c.à.d.*

$$\frac{\phi_1(n)}{n_1} = \frac{\phi_2(n)}{n_2} = \frac{\phi_1(n)+\phi_2(n)}{n_1+n_2}, \qquad n_1, n_2 > 0.$$

Supposons de plus que $\rho_1(n-e_1) + \rho_2(n-e_2)$ ne dépend de n_1, n_2 qu'à travers leur somme et que pour certaines constances w_1, w_2,

$$\frac{\rho_1(n-e_1)}{w_1} = \frac{\rho_2(n-e_2)}{w_2} = \frac{\rho_1(n-e_1)+\rho_2(n-e_2)}{w_1+w_2}, \qquad n_1, n_2 > 0.$$

Alors les fonctions $\psi_1(n), \ldots, \psi_I(n)$ sont équilibrées par $\Psi(n)$ si et seulement si $\frac{\rho_1(n-e_1)+\rho_2(n-e_2)}{\phi_1(n)+\phi_2(n)}$, $\psi_3(n)$, …, $\psi_I(n)$ sont équilibrées par $\tilde{\Psi}(n)$, avec

$$\Psi^{-1}(n) = \binom{n_1+n_2}{n_1} \frac{w_1^{n_1} w_2^{n_2}}{(w_1+w_2)^{n_1+n_2}} \tilde{\Psi}(n_1+n_2, n_3, \ldots, n_I)^{-1}.$$

Preuve. La démonstration est déduite du fait que

$$\frac{\rho_1(n-e_1)}{\phi_1(n)} = \frac{n_1+n_2}{n_1} \times \frac{w_1}{w_1+w_2} \times \frac{\rho_1(n-e_1)+\rho_2(n-e_2)}{\phi_1(n)+\phi_2(n)}, \qquad n_1, n_2 > 0.$$

□

Cette proposition suggère qu'en empruntant le raisonnement utilisé dans la Section 1.4.3, on peut montrer la condition suffisante de l'insensibilité d'un réseau de Jackson de files PS avec capacités, taux de service et routage dépendants de l'état. De plus, par récurrence sur le nombre de files dans le réseau, on peut montrer que cette condition suffisante est aussi nécessaire pour avoir l'insensibilité.

Théorème 1.24 *Dans un réseau de Jackson de files d'attente PS avec capacités, taux d'arrivée et routage dépendants de l'état, l'insensibilité est équivalente à la propriété d'équilibre des fonctions* $\psi_1, \ldots, \psi_I$ *définies en (1.33). Dans ce cas, la distribution stationnaire du nombre de clients à chaque file est insensible aux distributions des demandes de service, au moins pour la classe des distributions de Cox, et est donnée par*

$$\pi(n) = C\Psi(n)^{-1}.$$

2

Disciplines de service insensibles

Sommaire

Dans ce chapitre, on introduira des permutations aléatoires de clients dans les files d'attente symétriques et dans les réseaux de ces files. Lorsqu'un nouveau client arrive à ou quitte une file, les autres clients présents à chaque file sont permutés au hasard au lieu d'un simple décalage dans la discipline symétrique usuelle. Ces permutations aléatoires ont été considérées par Yashkov [Yas80], Daduna, Schassberger [DS83, Dad01b] et Yates [Yat90, Yat94]. Dans son livre [Dad01a], Daduna a considéré les files à temps discret avec des permutations aléatoires à chaque slot de temps, avant chaque arrivée et après chaque départ. Dans cette thèse, on considère les modèles à temps continu et on suppose qu'avant chaque arrivée, après chaque départ et lors d'un changement de files (pour les modèles de réseaux), les clients à chaque file peuvent être permutés au hasard. Remarquons que dans notre modèles, ces permutations aléatoires sont choisies suivant certaines mesures de probabilité. De plus, dans les modèles de réseaux, la mesure de probabilité de permutations à une file quelconque peut dépendre non seulement du nombre de clients à cette file mais aussi du nombre de clients à toute autre file dans le réseau. On montrera que la distribution stationnaire de l'état du système est la même que dans le cas sans permutation. En particuler, avec permutations aléatoires de clients à chaque file, les files symétriques simples, les réseaux de Jackson et les réseaux de Kelly de files symétriques sont tous insensibles.

Rappelons que dans le Chapitre 1, on a montré l'insensibilité d'une file symétrique par la méthode des phases et la méthode RGSMP. Similairement, on peut utiliser les deux méthodes pour démontrer l'insensibilité des modèles dans ce chapitre, mais on choisit ici la méthode des phases afin d'avoir des preuves directes à travers les équations de balance et d'être en accord avec les preuves pour la condition nécessaire dans certains modèles.

Enfin, on introduira quelques modèles insensibles non symétriques avant de récapituler les modèles de files d'attente insensibles/sensibles.

2.1 Disciplines symétriques

Les disciplines symétriques sont connus à être insensibles dans les modèles classiques sans permutation aléatoire. Dans cette partie, on définira les permutations aléatoires et on établira l'insensibilité d'une file symétrique simple munie de ces permutations aléatoires au lieu de décalages classiques.

2.1.1 Permutations aléatoires

On considère une file d'attente simple de macro-état n désignant le nombre de clients dans la file, et de micro-état x contenant les informations complémentaires afin de décrire l'évolution de système. Introduisons maintenant les permutations aléatoires dans cette file.

Pour tout $n \geq 1$, posons $\mathcal{P}(n)$ l'ensemble de permutations des éléments d'une séquence de longueur n et $\mathcal{P}(0)$ l'ensemble consistant seulement en l'application d'identité sur $\{\emptyset\}$. Pour tout micro-état $x \in \mathcal{X}(n)$, et pour toute permutation $\sigma \in \mathcal{P}(n)$, on note $\sigma(x)$ le micro-état obtenu à partir de x en permutant les clients selon σ.

Pour tous $n \geq 1$ et $l = 1, \ldots, n$, posons $\alpha(l, n, \cdot)$ une mesure de probabilité arbitraire sur $\mathcal{P}(n)$. On a :

$$\sum_{\sigma \in \mathcal{P}(n)} \alpha(l, n, \sigma) = 1.$$

Lorsqu'un client arrive à la position l de la file tandis que le nombre de clients avant son arrivée est $n, n \geq 1$, ces n clients sont permutés au hasard suivant la mesure de probabilité $\alpha(l, n, \cdot)$ avant cette arrivée. Si $l \leq n$, les clients aux positions $l, \ldots, n$ sont ensuite décalés aux positions $l+1, \ldots, n+1$, respectivement suivant la règle de décalage classique.

Supposons par exemple que :

$$\alpha(2, 4, \sigma) = 1, \quad \text{avec} \quad \sigma(1, 2, 3, 4) = (1, 4, 2, 3).$$

Alors lorsqu'un client arrive à la position 2 tandis qu'il y a 4 clients dans la file, l'ancien client en position 2 sera placé à la fin de la file, comme illustré dans la Figure 2.1.

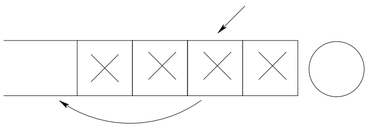

FIG. 2.1 – Une permutation simple : Le client en position 2 est placé à la fin de la file.

Supposons maintenant que :

$$\alpha(2,4,\sigma) = \frac{1}{2}, \quad \text{avec} \quad \sigma(1,2,3,4) = (1,4,2,3),$$

et

$$\alpha(2,4,\sigma') = \frac{1}{2}, \quad \text{avec} \quad \sigma'(1,2,3,4) = (3,1,4,2).$$

Alors lorsqu'un nouveau client arrive à la position 2 tandis qu'il y a 4 clients dans la file, ces quatre clients sont permutés selon la permutation de la Figure 2.1 avec la probabilité 1/2, selon la permutation plus complexe de la Figure 2.2 avec la probabilité 1/2.

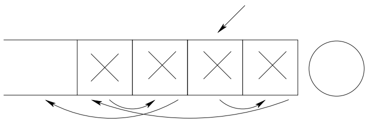

FIG. 2.2 – Une permutation plus complexe

De même, pour tous $n \geq 1$ et $l = 1, \dots, n+1$, posons $\beta(l, n, \cdot)$ une mesure de probabilité sur $\mathcal{P}(n)$. On a :

$$\sum_{\sigma \in \mathcal{P}(n)} \beta(l, n, \sigma) = 1.$$

Lorsque le client en position l quitte la file tandis que le nombre de clients avant son départ est $n+1, n \geq 1$, les n autres clients sont permutés au hasard selon la mesure de probabilité $\beta(l, n, \cdot)$ dès que la règle de décalage classique est appliquée.

Remarquons qu'une grande classe de disciplines de service peut être représentée par ces files symétriques avec permutations aléatoires. Considérons par exemple la discipline de service où à l'arrivée, un nouveau client reçoit immédiatement toute la capacité de la file :

$$\delta(1, n) = 1, \qquad \forall n \geq 1,$$

et après chaque accomplissement de service, un client choisi au hasard reçoit toute la capacité de service, c.à.d. que $\beta(1, n, \cdot)$ est la distribution uniforme sur l'ensemble des permutations $\mathcal{P}(n)$:

$$\beta(1, n, \sigma) = \frac{1}{n!}, \qquad \forall n \geq 1, \forall \sigma \in \mathcal{P}(n).$$

Cette discipline de service peut être vue comme un mélange de la discipline LIFO préemptive et la discipline *aléatoire* [Pal57].

2.1.2 Insensibilité

On suppose que les clients arrivent dans la file suivant un processus de Poisson d'intensité ν et la file est symétrique, de probabilités de positionnement $\delta(l,n)$: Si un client trouve $n-1$ autres clients à son arrivée, il choisit la position l avec la probabilité $\delta(l,n)$ et lorsqu'il y a n clients dans la file, le serveur alloue une proportion $\delta(l,n)$ de capacité de service au client en position $l, n \geq 1, l = 1, \ldots, n$.

Supposons que les demandes de service sont i.i.d. de loi mélange de distributions d'Erlang, c.à.d. que chaque demande de service est une somme d'un nombre aléatoire de phases exponentielles de moyenne $1/\mu, \mu > 0$.

Pour tout $k \geq 1$, posons $s(k)$ la probabilité que le nombre de phases de service soit égal à k, $\bar{s}(k)$ la probabilité que le nombre de phases soit plus grand ou égal à k et $\bar{s}$ le nombre moyen de phases. On a :

$$\sum_{k\geq 1} s(k) = 1, \quad \bar{s}(k) = \sum_{j\geq k} s(j), \quad \text{et} \quad \bar{s} = \sum_{k\geq 1} ks(k) = \sum_{k\geq 1} \bar{s}(k).$$

Notons que la demande de service moyenne est égale à $\bar{s}/\mu$ et appelons cette quantité le *taux de service* de la file. Le taux de charge est alors donné par $\rho = \nu\bar{s}/\mu$. On suppose que la file est stable :

$$\rho = \frac{\nu\bar{s}}{\mu} < 1. \tag{2.1}$$

Micro-état. Dans ce cas, le micro-état de la file est le vecteur de ligne x contenant les phases de service de chaque client. On a : $n = |x|$, où $|x|$ désigne la longueur du vecteur x. Par convention, $x = \emptyset$ et $n = |\emptyset| = 0$ si la file est vide.

Notons $\mathcal{X}$ l'ensemble de micro-états et $\mathcal{X}(n) = \{x : |x| = n\}$ l'ensemble de micro-états correspondants au même macro-état n. Pour tous $x \in \mathcal{X}(n)$ et $l = 1, \ldots, n+1$, notons $T^{k,l}x$ le micro-état obtenu à partir de x en ajoutant un client en k$^{\text{ème}}$ phase de service à la position l dans la file suivant la règle de décalage définie auparavant. De même, pour tous $x \in \mathcal{X}(n)$ et $l = 1, \ldots, n$, notons $T_l x$ le micro-état obtenu à partir de x en effaçant le client en position l.

Théorème 2.1 *La distribution stationnaire du processus markovien décrivant le micro-état x a une expression explicite :*

$$\pi(x) = (1-\rho)\left(\frac{\nu}{\mu}\right)^n \prod_{l=1}^{n} \bar{s}(x_l), \qquad \forall x \in \mathcal{X},$$

si la file est stable : $0 < \rho < 1$.

Preuve. Les taux de transition de ce processus markovien sont nuls à part les suivants :

- Arrivée d'un nouveau client à la position l, menant la file de l'état x à l'état $T^{1,l}\sigma(x)$, pour une certaine permutation $\sigma \in \mathcal{P}(n)$:

$$q\big(x, T^{1,l}\sigma(x)\big) = \nu\delta(l, n+1)\alpha(l, n, \sigma), \qquad x \in \mathcal{X}.$$

- Changement de phase de service du client en position l, menant la file de l'état x à l'état $x + e_l$, e_l étant le vecteur unitaire de dimension n dont la $l^{\text{ème}}$ composante est égale à 1 et les autres sont nulles :

$$q(x, x+e_l) = \mu\delta(l, n)\frac{\bar{s}(x_l+1)}{\bar{s}(x_l)}, \qquad x \neq \emptyset, l = 1, \ldots, n.$$

- Départ du client en position l, menant la file de l'état x à l'état $\sigma(T_l x)$, pour une certaine permutation $\sigma \in \mathcal{P}(n-1)$:

$$q\big(x, \sigma(T_l x)\big) = \mu\delta(l, n)\frac{s(x_l)}{\bar{s}(x_l)}\beta(l, n-1, \sigma), \qquad x \neq \emptyset, l = 1, \ldots, n.$$

En observant que $\pi(\sigma(x)) = \pi(x)$ pour toute permutation $\sigma \in \mathcal{P}(n)$, on peut vérifier que π satisfait les équations de balance partielle :

$$\pi(x) \sum_{\sigma \in \mathcal{P}(n)} q\big(x, T^{1,l}\sigma(x)\big) = \sum_{d \geq 1} \sum_{\sigma \in \mathcal{P}(n)} \pi\big(T^{d,l}\sigma^{-1}(x)\big) q\big(T^{d,l}\sigma^{-1}(x), x\big),$$

pour tous $x \in \mathcal{X}$ et $l = 1, \ldots, n+1$, et

$$\begin{aligned} \pi(x)\Big(q(x, x+e_l) + \sum_{\sigma \in \mathcal{P}(n-1)} q(x, \sigma(T_l x))\Big) \quad &= \quad \sum_{\sigma \in \mathcal{P}(n-1)} \pi\big(\sigma^{-1}(T_l x)\big) q\big(\sigma^{-1}(T_l x), x\big) \\ &\quad + \pi(x - e_l) q(x - e_l, x), \end{aligned}$$

pour tous $x \neq \emptyset$ et $l = 1, \ldots, n$. Par convention, $\pi(x) = 0$ si x n'est pas dans $\mathcal{X}$.

Alors la probabilité π est la distribution stationnaire du processus markovien décrivant le micro-état x.

□

Par conséquence, dans une file d'attente symétrique avec permutations aléatoires, la distribution stationnaire du nombre de clients dans la file est insensible à la distribution des demandes de service, au moins pour la classe des mélanges de distributions d'Erlang. Cette distribution stationnaire ne dépend que du taux de charge ρ :

$$\pi'(n) = (1-\rho)\rho^n, \qquad n \geq 0.$$

Corollaire 2.2 *La distribution stationnaire du nombre de clients dans une file d'attente symétrique stable avec permutations aléatoires est insensible à la distribution des demandes de service, au moins pour les mélanges de distributions d'Erlang, et ne dépend que du taux de charge $\rho = \nu\bar{s}/\mu$:*

$$\pi'(n) = (1-\rho)\rho^n, \qquad n \geq 0.$$

Le nombre moyen de clients dans l'état stationnaire ne dépend que du taux de charge ρ :

$$\mathbb{E}[n] = \sum_n n\pi(n) = \frac{\rho}{1-\rho}.$$

D'après la loi de Little, le temps de séjour moyen d'un client quelconque est insensible et donné par

$$\mathbb{E}[T] = \frac{\mathbb{E}[n]}{\nu} = \frac{\bar{s}}{\mu(1-\rho)}.$$

Comme dans la file symétrique classique (Section 1.3.1), on peut montrer que le temps de séjour moyen d'un client est proportionnel à sa demande de serice r et est donné par :

$$\frac{r}{1-\rho}.$$

2.2 Réseaux de Jackson avec permutations aléatoires

On considère maintenant un réseau de I files d'attente symétriques avec permutations aléatoires et avec le routage de Jackson [Jac57] et on l'appelle un réseau de Jackson de files symétriques avec permutations aléatoires. On décrit tout d'abord le modèle dans la première partie, et ensuite on étudie son insensibilité avant de considérer quelques extensions.

2.2.1 Modèle de réseaux de Jackson

Considérons un réseau de I files symétriques dont les arrivées externes à chaque file i forment un processus de Poisson d'intensité ν_i, avec

$$\sum_{i=1}^{I} \nu_i > 0.$$

Routage de Jackson. Après l'accomplissement de service à la file i, un client est dirigé vers la file j avec la probabilité p_{ij} et quitte le réseau avec la probabilité

$$p_i = 1 - \sum_{j=1}^{I} p_{ij}. \tag{2.2}$$

Les clients quittent éventuellement le réseau de telle sorte que les taux d'arrivée effectifs $\lambda_1, \ldots, \lambda_I$, en comptant les arrivées des autres files, soient uniquement définis par les *équations de trafic* :

$$\lambda_i = \nu_i + \sum_{j=1}^{I} \lambda_j p_{ji}, \qquad i = 1, \ldots, I. \tag{2.3}$$

Le macro-état du système est le vecteur $n = (n_1, \ldots, n_I)$ contenant le nombre de clients présents à chaque file.

Discipline de service symétrique : À la file i, un nouveau client choisit la position l avec la probabilité $\delta_i(l,n)$ s'il trouve le macro-état $n - e_i$ à son arrivée. Cette quantité $\delta_i(l,n)$ est aussi la proportion de service alouée à la position l lorsque le macro-état du système est n :

$$\sum_{l=1}^{n_i} \delta_i(l,n) = 1, \qquad \forall i = 1, \ldots, I, \forall n : n_i \geq 1.$$

e_i étant le vecteur unitaire de dimension I dont la $i^{\text{ème}}$ composante est égale à 1 et les autres sont nulles.

Remarquons que dans ce modèle, la discipline de service peut dépendre du macro-état n du réseau entier et non seulement de l'état n_i de la file i.

Permutations aléatoires : On appelle une permutation dans le macro-état n un élément de l'ensemble $\mathcal{S}(n) = \mathcal{P}(n_1) \times \ldots \times \mathcal{P}(n_I)$, où $\mathcal{P}(m)$ désigne l'ensemble ordinaire des permutations de m éléments, pour tout $m \geq 1$, et $\mathcal{P}(0)$ désigne l'ensemble consistant seulement en l'application d'identité sur $\{\emptyset\}$.

À partir d'ici, on suppose qu'un mouvement d'un client quelconque peut provoquer des permutations aléatoires de clients à chaque file d'attente. La distribution de permutations peut dépendre du macro-état et du client qui provoque la permutation. Pour tout macro-état n et pour tous $i, j = 1, \ldots, I$, posons $\alpha_i(n, \cdot)$, $\beta_i(n, \cdot)$ et $\gamma_{ij}(n, \cdot)$ des distributions arbitraires sur $\mathcal{S}(n)$.

Avant chaque arrivée externe à la file i lorsque le macro-état du système est n, les anciens clients du système sont permutés au hasard selon la distribution de permutations $\alpha_i(n, \cdot)$.

De même, lorsqu'un client la file i quitte le réseau tandis que le macro-état est n, les autres clients sont permutés au hasard selon la distribution de permutations $\beta_i(n - e_i, \cdot)$.

Et finalement, lorsqu'un client dans la file i termine son service dans le macro-état n et se dirige vers la file j avec la probabilité p_{ij}, les autres clients sont permutés selon la distribution de permutations $\gamma_{ij}(n - e_i, \cdot)$ après le départ de la file i et avant l'arrivée à la file j.

Remarquons qu'à l'arrivée d'un client, on ne le permute pas avec les autres clients et ce nouveau client est toujours alloué une proportion strictement positive de capacité de service à son arrivée. La condition d'attention instantanée est remplie.

Demandes de service : Supposons que les clients à la file i demandent des services i.i.d. de loi mélange de distributions d'Erlang, c.à.d. que chaque client à cette file demande une somme d'un nombre aléatoire de phases exponenetielles de moyenne $1/\mu_i$. Pour tout $k \geq 1$, notons $s_i(k)$ la probabilité que le nombre de phases soit égal à k, $\bar{s}_i(k)$ la probabilité que ce nombre dépasse k et $\bar{s}_i$ le nombre moyen de phases. Alors

$$\sum_{k \geq 1} s_i(k) = 1, \quad \bar{s}_i(k) = \sum_{j \geq k} s_i(j), \quad \text{et} \quad \bar{s}_i = \sum_{k \geq 1} k s_i(k) = \sum_{k \geq 1} \bar{s}_i(k).$$

On suppose que le réseau est stable :

$$\rho_i = \frac{\lambda_i \bar{s}_i}{\mu_i} < 1. \tag{2.4}$$

Le micro-état du réseau est $x = (x_{il}, 1 \leq i \leq I, 1 \leq l \leq n_i)$, où x_{il} désigne la phase de service du client en position l à la file i. Notons $|x| = n$ le macro-état du réseau. Par convention, $x_i = \emptyset$ et $n_i = |\emptyset| = 0$ si la file i est vide.

Notons $\mathcal{X}$ l'ensemble de micro-états et $\mathcal{X}(n) = \{x : |x| = n\}$ l'ensemble des micro-états correspondants au même macro-état n. Pour tous $x \in \mathcal{X}(n)$ et $l = 1, \ldots, n+1$, notons $T^{i,k,l}x$ le micro-état obtenu à partir de x en ajoutant un client en $k^{\text{ème}}$ phase de service à la position l dans la file i suivant la règle de décalage définie auparavant. De même, pour tous $x \in \mathcal{X}(n)$ et $l = 1, \ldots, n$, notons $T_{i,l}x$ le micro-état obtenu à partir de x en effaçant le client en position l à la file i.

Il y a quatre types de transition dans un réseau de Jackson

- *arrivée externe* d'un client à la position l de la file $i, l = 1, \ldots, n_i + 1$, menant le système de l'état x à l'état $T^{i,1,l}\sigma(x)$, pour une certaine permutation $\sigma \in \mathcal{S}(n)$. Le taux de cette transition est
$$q\big(x, T^{i,1,l}\sigma(x)\big) = \nu_i \delta_i(l, n + e_i)\alpha_i(n, \sigma), \qquad x \in \mathcal{X}.$$

- *changement de phase* du client en position l de la file i, menant le système de l'état x à l'état $x + e_{i,l}$, où $e_{i,l}$ désigne l'élément dont la composante (i, l) est égale à 1 et les autres sont nulles. Le taux de transition correspondant est
$$q(x, x + e_{i,l}) = \mu_i \delta_i(l, n)\frac{\bar{s}(x_{il} + 1)}{\bar{s}(x_{il})}, \qquad i = 1, \ldots, I, x_i \neq \emptyset, l = 1, \ldots, n_i.$$

- *changement de files :* mouvement du client en position l de la file i à la position l' de la file $j, l = 1, \ldots, n_i, l' = 1, \ldots, n_j + 1, n_i \geq 1$, menant le système de l'état x à l'état $T^{j,1,l'}\sigma(T_{i,l}x)$ pour une certain permutation $\sigma \in \mathcal{S}(n - e_i)$, où e_i désigne le vecteur unitaire de dimension I dont la $i^{\text{ème}}$ composante est égale à 1 et les autres sont nulles. On appelle aussi cette transition une *arrivée interne* à la position l' de la file j. Le taux de cette transition est donné par
$$q\big(x, T^{j,1,l'}\sigma(T_{i,l}x)\big) = \mu_i \delta_i(l, n)\frac{s(x_{il})}{\bar{s}(x_{il})}\gamma_{ij}(n - e_i, \sigma)p_{ij}\delta_j(l', n - e_i + e_j),$$
pour $i, j = 1, \ldots, I, x_i \neq \emptyset, l = 1, \ldots, n_i, l' = 1, \ldots, n_j + 1$.

- *départ définitif* du client en position l de la file $i, l = 1, \ldots, n_i$, menant le système de l'état x à l'état $\sigma(T_{i,l}x)$. Le taux de cette transition est
$$q\big(x, \sigma(T_{i,l}x)\big) = \mu_i \delta_i(l, n)\frac{s(x_{il})}{\bar{s}(x_{il})}\beta_i(n - e_i, \sigma)p_i, \qquad i = 1, \ldots, I, x_i \neq \emptyset, l = 1, \ldots, n_i,$$
où p_i est la probabilité qu'un client quitte le réseau après l'accomplissement de service à la file d'attente i.

2.2.2 Insensibilité des réseaux de Jackson

Théorème 2.3 *La distribution stationnaire du processus markovien décrivant le micro-état du système est à forme produit et donnée par*

$$\pi(x) = \prod_{i=1}^{I}(1-\rho_i)(\frac{\lambda_i}{\mu_i})^{n_i}\prod_{l=1}^{n_i}\bar{s}_i(x_{il}), \qquad \forall x \in \mathcal{X},$$

si le réseau est stable : $0 < \rho_i < 1$ *pour tout* $i = 1, \dots, I$ *(2.4).*

Preuve. Vérifions les équations de balance partielle pour la source et pour toute position l de la file i.

Pour la source, le flux de probabilité correspondant à une arrivée externe dans le micro-état x est donné par

$$\pi(x)\sum_{i=1}^{I}\nu_i,$$

égal à celui correspondant à un départ définitif d'un client menant le système à l'état x :

$$\begin{aligned}
& \sum_{i=1}^{I}\sum_{l=1}^{n_i+1}\sum_{\sigma\in\mathcal{S}(n)}\sum_{k\geq 1}\pi\big(T^{i,k,l}\sigma^{-1}(x)\big)q\big(T^{i,k,l}\sigma^{-1}(x),x\big) \\
= \quad & \sum_{i=1}^{I}\sum_{l=1}^{n_i+1}\sum_{\sigma\in\mathcal{S}(n)}\sum_{k\geq 1}\pi(T^{i,k,l}x)\mu_i\delta_i(l,n+e_i)\frac{s_i(k)}{\bar{s}_i(k)}\beta_i(n,\sigma)p_i \\
= \quad & \sum_{i=1}^{I}\sum_{l=1}^{n_i+1}\sum_{k\geq 1}\pi(x)\frac{\lambda_i}{\mu_i}\bar{s}_i(k)\mu_i\delta_i(l,n+e_i)\frac{s_i(k)}{\bar{s}_i(k)}p_i \\
= \quad & \pi(x)\sum_{i=1}^{I}\lambda_i p_i,
\end{aligned}$$

car $\sum_{k\geq 1}s_i(k) = 1$ pour toute i et d'après les équations de trafic (2.3),

$$\sum_{i=1}^{I}\nu_i = \sum_{i=1}^{I}\lambda_i p_i.$$

D'ailleurs, pour toute position l de la file i, le flux de probabilité correspondant à un changement de phase, un changement de files et un départ définitif du client en position l de la file i dans le micro-état x est égal à celui correspondant à l'accomplissement de phase de service de ce client :

$$\pi(x)\mu_i\delta_i(l,n). \tag{2.5}$$

De plus, le flux de probabilité correspondant à une arrivée externe à la position l de la file i menant le système à l'état x est donné par

$$
\begin{aligned}
& \sum_{\sigma\in\mathcal{S}(n-e_i)} \pi\big(T_{i,l}\sigma^{-1}(x)\big)q\big(T_{i,l}\sigma^{-1}(x),x\big) \\
= & \sum_{\sigma\in\mathcal{S}(n-e_i)} \pi(T_{i,l}x)\nu_i\delta_i(l,n)\alpha_i(n-e_i,\sigma)\mathbb{I}(x_{i,l=1}) \\
= & \ \pi(x)\frac{\mu_i}{\lambda_i}\delta_i(l,n)\nu_i\mathbb{I}(x_{il}=1), \qquad (2.6)
\end{aligned}
$$

où $\mathbb{I}$ est la fonction indicatrice : $\mathbb{I}(A)=1$ si l'événement A est vrai et $\mathbb{I}(A)=0$ sinon.

Le flux de probabilité correspondant à un changement de phase du client en position l de la file i menant le système à l'état x est donné par

$$
\begin{aligned}
& \pi(x-e_{i,l})q(x-e_{i,l},x) \\
= & \ \pi(x)\frac{\bar{s}_i(x_{il}-1)}{\bar{s}_i(x_{il})}\mu_i\delta_i(l,n)\frac{\bar{s}_i(x_{il})}{\bar{s}_i(x_{il}-1)}\mathbb{I}(x_{il}\geq 2) \\
= & \ \pi(x)\mu_i\delta_i(l,n)\mathbb{I}(x_{il}\geq 2), \qquad (2.7)
\end{aligned}
$$

et le flux de probabilité correspondant à une arrivée interne à la position l de la file i menant le système à l'état x est égal à

$$
\begin{aligned}
& \sum_{j=1}^{I}\sum_{l'=1}^{n_j+1}\sum_{k\geq 1}\sum_{\sigma\in\mathcal{S}(n-e_i)} \pi\big(T^{j,k,l'}\sigma^{-1}(T_{i,l}x)\big)q\big(T^{j,k,l'}\sigma^{-1}(T_{i,l}x),x\big) \\
= & \sum_{j=1}^{I}\sum_{l'=1}^{n_j+1}\sum_{k\geq 1}\sum_{\sigma\in\mathcal{S}(n-e_i)} \pi\big(T^{j,k,l'}T_{i,l}x\big)\mu_j\delta_j(l',n+e_j-e_i)\frac{s_j(k)}{\bar{s}_j(k)} \\
& \times\gamma_{ji}(n-e_i,\sigma)p_{ji}\delta_i(l,n)\mathbb{I}(x_{il}=1) \\
= & \sum_{j=1}^{I}\sum_{k\geq 1}\pi(x)\frac{\lambda_j}{\mu_j}\frac{\mu_i}{\lambda_i}\frac{\bar{s}_j(k)}{\bar{s}_i(1)}\mu_j\frac{s_j(k)}{\bar{s}_j(k)}p_{ji}\delta_i(l,n)\mathbb{I}(x_{il}=1) \\
= & \ \pi(x)\frac{\mu_i}{\lambda_i}\delta_i(l,n)\mathbb{I}(x_{il}=1)\sum_{j=1}^{I}\lambda_j p_{ji}, \qquad (2.8)
\end{aligned}
$$

car $\bar{s}_i(1)=1$ et $\sum_{k\geq 1}s_j(k)=1$.

On s'aperçoit que la somme des flux de probabilité (2.6), (2.7) et (2.8) donne le flux (2.5), d'après les équations de trafic (2.3).

Par conséquence, la probabilité π satisfait le système d'équations de balance partielle et donc est la distribution stationnaire du micro-état du système.

□

En sommant sur tout $x\in\mathcal{X}(n)$, on obtient la distribution stationnaire du nombre de clients dans chaque file pour les mélanges de distributions d'Erlang :

Théorème 2.4 *Le réseau stable de Jackson de files symétriques avec permutations aléatoires est insensible aux distributions de demandes de service, au moins pour les mélanges de distributions d'Erlang. La distribution stationnaire du nombre de clients à chaque file est à forme produit et ne dépend que des taux de charge $\rho_1, \ldots, \rho_I$:*

$$\pi'(n) = \prod_{i=1}^{I} (1 - \rho_i)\rho_i^{n_i}, \qquad n_i \geq 0 \ \forall i.$$

Corollaire 2.5 *À l'état stationnaire, le nombre moyen de clients à chaque file i ne dépend que de son taux de charge ρ_i :*

$$\mathbb{E}[n_i] = \sum_n n_i \pi'(n) = \frac{\rho_i}{1 - \rho_i}.$$

Le temps de séjour moyen d'un client à la file i est donné par

$$\mathbb{E}[T_i] = \frac{\mathbb{E}(n_i)}{\lambda_i} = \frac{\bar{s}_i}{\mu_i(1 - \rho_i)}.$$

Comme dans la file symétrique classique (Section 1.3.1), le temps de séjour moyen d'un client à la file i est proportionnel à sa demande de service r et est donné par :

$$\frac{r}{1 - \rho_i}.$$

Remarquons que les résultats restent valables dans les cas suivants :

- Les mesures de permutations dépendent de la position du client en mouvement : Les mesures $\alpha_i(n, \cdot)$ dépendent de la position choisie par le nouveau client. De même, les mesures $\beta_i(n, \cdot)$ dépendent de la position du client quittant le réseau, et les mesures $\gamma_{ij}(n, \cdot)$ dépendent à la fois de l'ancienne position et de la nouvelle position du client en mouvement.
- Les permutations aléatoires sont provoquées par un processus ponctuel externe à chaque file (par exemple, un processus de Poisson). Il suffit de considérer les files additionnelles dont le rôle est de provoquer les permutations (par exemple, à travers des arrivées externes à ces files).

Dans la suite, on donnera des résultats analogues dans les réseaux fermés de Jackson avec permutations aléatoires et dans les réseaux avec des capacités variables.

2.2.3 Réseaux fermés

Supposons que le réseau est fermé : il n'y a ni d'arrivée externe dans le réseau, ni de départ de clients du réseau, les clients ne circulent que dans le réseau. On a :

$$\nu_i = 0 \quad \text{et} \quad p_i = 0, \qquad \forall i.$$

Le nombre total de clients présents dans le réseau est fixé à une certaine valeur m et l'ensemble de macro-états est

$$G = \{n : \sum_{i=1}^{I} n_i = m\}.$$

Supposons que le routage est irréductible, c.à.d. que chaque file est visitée par les m clients. Les résultats similaires restent valables pour les réseaux appelés *mélangés* de files d'attente.

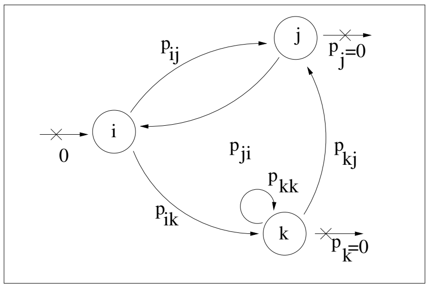

FIG. 2.3 – Un réseau fermé irréductible de Jackson.

La fréquence d'arrivée λ_i à chaque file i est uniquement définie par les équations de trafic (2.3) :

$$\lambda_i = \sum_{j=1}^{I} \lambda_j p_{ji},$$

avec $\sum_i \lambda_i = 1$.

Comme il n'y a plus d'arrivée externe et de départ définitif, les flux correspondants à ces transitions sont nuls et les équations de balance partielle deviennent : Le flux de probabilité correspondant à un changement de phases et à un changement de files du client en position l de la file i dans l'état x est égal au flux de probabilité correspondant à un changement de phases et à une arrivée interne à la position l de la file i menant le système à l'état x. On peut vérifier que ces équations de balance partielle admettent une unique solution donnée par

$$\pi(x) = C \prod_{i=1}^{I} (\frac{\lambda_i}{\mu_i})^{n_i} \prod_{l=1}^{n_i} \bar{s}_i(x_{il}), \qquad x \in \mathcal{X} : n \in G,$$

où C désigne la constance de normalisation.

Théorème 2.6 *Le réseau fermé de Jackson de files symétriques avec permutations aléatoires admet la distribution stationnaire à forme produit, donnée par*

$$\pi(x) = C \prod_{i=1}^{I} (\frac{\lambda_i}{\mu_i})^{n_i} \prod_{l=1}^{n_i} \bar{s}_i(x_{il}), \qquad x \in \mathcal{X} : n \in G,$$

où C désigne la constance de normalisation.

Corollaire 2.7 *Le réseau fermé de Jackson de files symétriques avec permutations aléatoires est insensible à la distribution des demandes de service, au moins pour les mélanges de distributions d'Erlang. La distribution stationnaire du nombre de clients à chaque file est à forme produit et ne dépend que des taux de charge :*

$$\pi(n) = C \prod_{i=1}^{I} \rho_i^{n_i}, \qquad \forall n \in G,$$

où $\rho_i = \lambda_i \bar{s}_i / \mu_i$ est le taux de charge à la file $i, i = 1, \ldots, I$, et C est la constance de normalisation :

$$C = \Big(\sum_{n \in G} \prod_{i=1}^{I} \rho_i^{n_i} \Big)^{-1}.$$

2.2.4 Capacités variables

Supposons maintenant qu'à chaque file i, la capacité de service n'est plus la constance 1 mais une fonction $\phi_i(n)$ dépendante du macro-état n. Supposons de plus que ces capacités de service possèdent la *propriété d'équilibre* (Section 1.4.2) des réseaux de Kelly-Whittle [Ser99], c.à.d. qu'il existe une fonction positive Φ sur $\mathbb{N}^I$ telle que

$$\phi_i(n) = \frac{\Phi(n - e_i)}{\Phi(n)}, \qquad \forall n : n_i \geq 1.$$

Dans la Section 1.4.3, on a vu que cette propriété d'équilibre est la condition nécessaire et suffisante afin d'avoir l'insensibilité dans un réseau de Jackson de files PS avec capacités dépendantes de l'état. Dans cette section, pour un réseau de Jackson de files d'attente symétriques avec permutations aléatoires et avec capacités et taux de service dépendants du macro-état du réseau, on montrera que la propriété d'équilibre des taux de service est la condition suffisante pour la propriéte d'insensibilié. Ensuite, la démonstration de la condition nécessaire est l'analogue de la Section 1.4.3.

Condition suffisante

On s'aperçoit que les équations de balance partielle restent les mêmes sauf que le taux de l'accomplissement de phase de service du client en position l de la file i devient $\phi_i(n)\mu_i\delta_i(l,n)$, au lieu de l'ancien terme $\mu_i\delta_i(l,n)$. Et on peut vérifier qu'en multipliant la distribution stationnaire $\pi(x)$ donnée dans le Théorème 2.3 par la fonction d'équilibre $\Phi(n)$ et par une constance de normalisation C, on obtient la distribution stationnaire du nouveau réseau de Jackson avec capacités dépendantes de l'état.

Théorème 2.8 *Le réseau de Jackson de files d'attente symétriques avec permutations aléatoires et avec capacités dépendantes de l'état admet la distribution stationnaire du micro-état x à forme explicite :*

$$\pi(x) = C\Phi(n) \prod_{i=1}^{I} (\frac{\lambda_i}{\mu_i})^{n_i} \prod_{l=1}^{n_i} \bar{s}_i(x_{il}), \qquad \forall x \in \mathcal{X},$$

si le réseau est stable : $0 < C < \infty$, où C est la constance de normalisation

$$C = \Big(\sum_{x \in \mathcal{X}} \Phi(n) \prod_{i=1}^{I} (\frac{\lambda_i}{\mu_i})^{n_i} \prod_{l=1}^{n_i} \bar{s}_i(x_{il}) \Big)^{-1}.$$

Corollaire 2.9 *Un réseau stable de Jackson de files d'attente symétriques avec permutations aléatoires et avec capacités dépendantes de l'état est insensible aux distributions des demandes de service, au moins pour les mélanges de distributions d'Erlang. La distribution stationnaire du nombre de clients à chaque file est à forme produit et donnée par :*

$$\pi(n) = C\Phi(n) \prod_{i=1}^{I} \rho_i^{n_i}.$$

Condition nécessaire

Par le raisonnement de récurrence sur le nombre de files utilisé pour le réseau de files d'attente PS (Section 1.4.3), on déduit que la condition nécessaire de l'insensibilité est la propriété d'équilibre des taux de service $\phi_i(n)$.

En conclusion, dans un réseau de Jackson de files d'attente symétriques avec permutations aléatoires et avec capacités de service dépendantes du macro-état n du réseau, on a l'insensibilité si et seulement si ces capacités de service $\phi_i(n)$ sont équilibrées par une certaine fonction positive $\Phi(n)$.

Similairement, si les taux d'arrivée et le routage sont aussi en fonction de l'état, l'insensibilité est équivalente à la propriété d'équilibre des fonctions $\psi_i(n)$ définies par

$$\psi_i(n) = \frac{\rho(n - e_i)}{\phi(n)},$$

où $\rho_i(n)$ désigne le taux de charge de la file i.

Remarque 2.1 *Comme la discipline symétrique en inclut la discipline PS, les réseaux de files PS (Section 1.4.3) appartiennent à la classe des réseaux de files symétriques. L'insensibilité des réseaux de files symétriques implique celle des réseaux de files PS.*

2.2.5 Exemples

Dans cette partie, on considérera quelques modèles de files d'attente avec permutations aléatoires. En particulier, on illustrera leur insensibilité/sensibilité par simulations.

Une file d'attente LIFO-préemptive avec permutations aléatoires

Considérons une file d'attente LIFO-préemptive avec :

$$\delta(1, n) = 1, \qquad n \geq 1.$$

Les clients arrivent dans la file suivant un processus de Poisson d'intensité $\nu = 0,5$. La distribution des demandes de service est hyperexponentielle de paramètre θ et de moyenne 1, pour un certain paramètre θ strictement positif. Avant chaque arrivée, les clients présents dans la file sont permutés au hasard.

Considérons deux types de permutations

Permutation uniforme. Les clients sont permutés au hasard selon une distribution uniforme, c.à.d.

$$\alpha(n, \sigma) = \frac{1}{n!}, \qquad \forall n \geq 1, l = 1, \ldots, n, \forall \sigma \in \mathcal{P}(n).$$

Permutation simple. Le client en service précédemment en tête de la file est placé à la fin de la file, c.à.d, $\forall n \geq 1, l = 1, \ldots, n,$

$$\alpha(n, \sigma) = 1$$

où σ désigne la permutation suivante :

$$\sigma(1, 2, 3, \ldots, n) = (2, 3, \ldots, n, 1).$$

D'après les Corollaires 2.2 et 2.5, le temps de séjour moyen est égal à 2, indépendamment de la distribution de demandes de service. Au contraire, la distribution du temps de séjour dépend fortement de la distribution de demandes de service, en particulier, l'écart type du temps de séjour en dépend aussi. Cela est illustré par des résultats de simulation dans la Figure 2.4 où le paramètre hyperexponentiel θ varie de 1 à 10. On peut observer que l'écart type est presque le même pour les deux types de permutations. Cette valeur se situe au dessous de celle du cas sans permutation, qui correspond à une file d'attente LIFO-préemptive ordinaire, et très près de celle d'une file d'attente PS.

Supposons maintenant que les permutations sont provoquées par un certain processus poissonnien indépendant d'intensité λ. Alors la file reste insensible dans ce cas. La Figure 2.5 nous donne la variation de l'écart type du temps de séjour en fonction du taux de permutation λ dans le cas de demandes de service exponentielles ($\theta = 1$). On peut observer que l'écart type diminue de celui d'une file LIFO-préemptive ordinaire à celui d'une file PS lorsque λ varie de 0 à 10. Lorsque le taux de permutation est infini, tous les clients reçoivent le même taux de service et la file est effectivement équivalent à une file PS.

Sensibilité de modèles avec permutations entre différentes files

Dans cette section, on considère quelques exemples de réseaux de files d'attente symétriques avec permutations de clients de différentes files d'attente. Plus précisément, on va considérer un réseau de deux files d'attente PS avec :

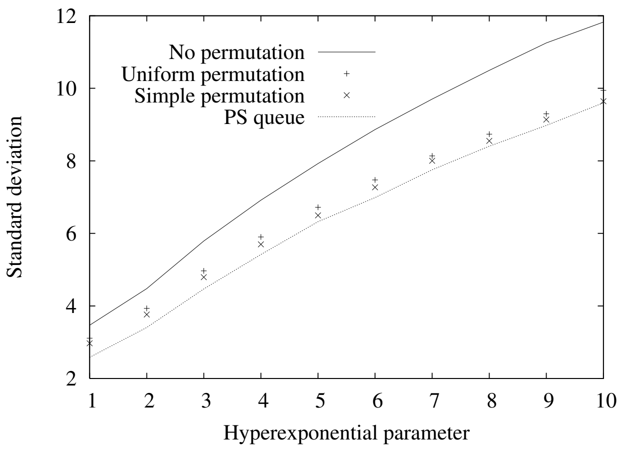

FIG. 2.4 – L'écart type du temps de séjour dans une file LIFO-préemptive avec permutations aléatoires aux arrivées.

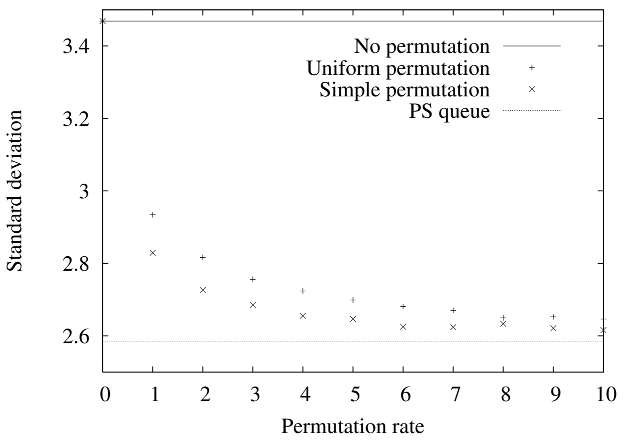

FIG. 2.5 – L'écart type du temps de séjour dans une file LIFO-préemptive avec permutations aléatoires aux arrivées d'un processus poissonien indépendant.

Taux d'arrivée : $\nu_1 = \nu_2 = 0.5$.

Demandes de service hyperexponentielles : Les clients arrivant à la file 1 demandent des services selon une loi hyperexponentielle de moyenne 1 et de paramètre θ_1.

De même, les clients arrivant à la file 2 demandent des services selon une loi hyperexponentielle de moyenne 1 et de paramètre θ_2.

Routage : Les clients quittent le réseau lors de l'accomplissement de service : $p_1 = p_2 = 1$, il n'y a pas de changement de files après l'accomplissement de service, les deux files sont en parallèle.

Permutations aléatoires : Supposons que les permutations sont provoquées par un certain processus poissonien indépendant d'intensité λ. On considère les permutations uniformes et simples définies auparavant. Une permutation simple consiste maintenant à permuter deux clients choisis au hasard, un de chaque file, et une permutation uniforme consiste à permuter tous les clients présents dans le réseau selon une distribution uniforme.

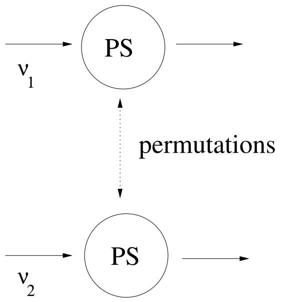

FIG. 2.6 – Réseau de deux files PS avec permutations des clients entre les deux files.

Cas 1. Les distributions des demandes de service sont différentes dans ces files d'attente.

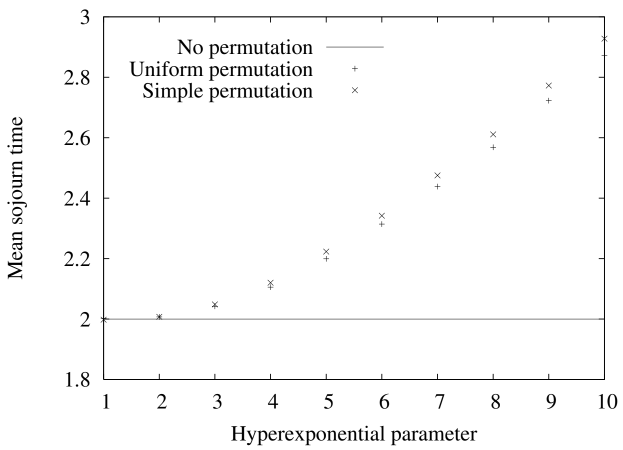

FIG. 2.7 – Sensibilité du temps de séjour moyen à la distribution des demandes de service dans un réseau de deux files avec permutations entre les deux files.

Les Figures 2.7 et 2.8 montrent la sensibilité du temps de séjour moyen aux distributions des demandes de service et au taux de permutation lorsque les distributions des demandes de service sont différentes dans les deux file d'attente PS.

La Figure 2.7 donne le temps de séjour moyen en fonction du paramètre hyperexponentiel θ_1 de la distribution des demandes de service à la file 1 lorsque la distribution des demandes de service à la file 2 est exponentielle : $\theta_2 = 1$, dans le cas où $\nu_1 = \nu_2 = 0,5$ et $\lambda = 1$.

La Figure 2.8 donne le temps de séjour moyen en fonction du taux de permutation λ lorsque $\theta_1 = 10, \theta_2 = 1$ et $\nu_1 = \nu_2 = 0,5$. Dans le cas ordinaire sans permutation, le réseau se compose

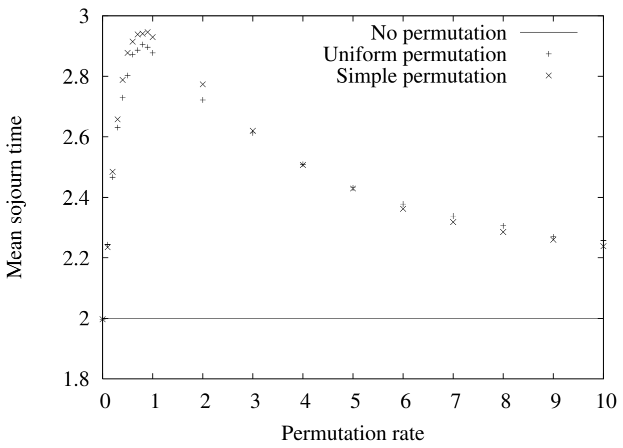

FIG. 2.8 – Sensibilité du temps de séjour moyen au taux de permutation dans un réseau de deux files avec permutations entre les deux files.

de deux files d'attente PS parallèles et le temps de séjour moyen est égal à :

$$\frac{1}{\mu_1 - \nu_1} = \frac{1}{\mu_2 - \nu_2} = 2.$$

Cas 2. Considérons maintenant le cas $\theta_1 = \theta_2$ où les distributions des demandes de service sont les mêmes dans les deux files.

D'une part, d'après le Corollaire 2.5, le temps de séjour moyen est indépendant de paramètres θ_1, θ_2 et λ.

D'autre part, le temps de séjour moyen des clients qui sont arrivés initialement à une file particulière, la file 1 ou la file 2, est sensible aux distributions des demandes de service (θ_1, θ_2) et au taux de permutation λ. Ce fait ne contredit pas l'insensibilité de la distribution stationnaire de l'état du réseau parce que l'état du réseau ne donne pas le nombre de clients arrivés par la file 1 ou 2. Ce phénomène est illustré par les Figures 2.9 et 2.10.

La Figure 2.9 donne le temps de séjour moyen en fonction du paramètre hyperexponentiel $\theta_1 = \theta_2$ lorsque $\nu_1 = 0,8, \nu_2 = 0,5$ et $\lambda = 1$.

La Figure 2.10 donne le temps de séjour moyen en fonction du taux de permutation λ lorsque $\mu_1 = 0,8, \nu_2 = 0,5$ et $\theta_1 = \theta_2 = 1$.

En absence de permutation, le réseau consiste en deux files d'attente PS parallèles et le temps de séjour moyen à la file 1 et file 2 sont respectivement donnés par :

$$\frac{1}{\mu_1 - \nu_1} = 5, \qquad \frac{1}{\mu_2 - \nu_2} = 2.$$

Le temps de séjour moyen dans le réseau est indépendant du taux de permutation et donné par :

$$\left(\frac{\nu_1}{1-\nu_1} + \frac{\nu_2}{1-\nu_2}\right)/(\nu_1 + \nu_2) = 50/13 \approx 3,85.$$

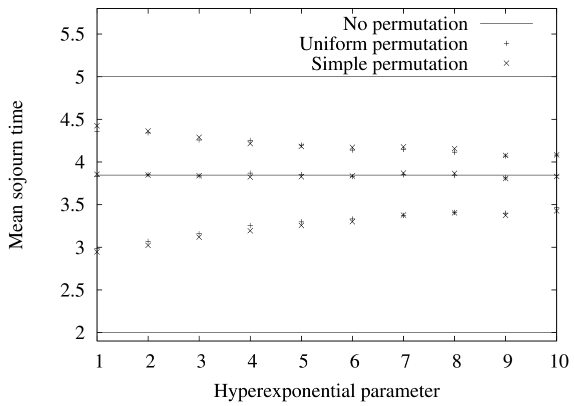

FIG. 2.9 – Temps de séjour moyen de clients qui entrent initialement à la file 1 (haut), file 2 (bas) et toute file (milieu) en fonction du paramètre hyperexponentiel.

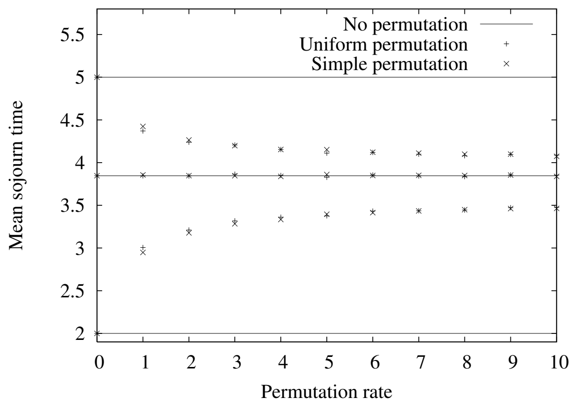

FIG. 2.10 – Temps de séjour moyen de clients qui entrent initialement à la file 1 (haut), file 2 (bas) et toute file (milieu) en fonction du taux de permutation.

2.3 Réseaux de Kelly avec permutations aléatoires

Dans cette section, la notion de *classe* de clients sera employée. Cette notion nous permet de représenter n'importe quel routage dans le réseau. Sans perte de généralité, on suppose que chaque classe est associée à une seule file d'attente. Les clients d'une même classe demandent des services i.i.d. de loi exponentielle à cette file et après l'accomplissement de service, rejoignent ensuite une autre classe ou quittent le réseau.

On considérera les réseaux de Kelly de files symétriques avec permutations aléatoires. Lorsqu'un client arrive à une file d'attente ou la quitte, les autres clients sont permutés à chaque file d'attente.

Afin de rester dans un cardre markovien, on emploiera dans cette section la méthode des phases. On montrera que les équations de balance partielle sont satisfaites pour chaque classe de client à chaque position de la file d'attente associée. Ces équations ne peuvent pas être déduites de celles des réseaux classiques sans permutation car dans notre modèle, les clients de différentes classes avec différentes demandes de service sont permutés dans chaque file d'attente. Ces équations de balance donnent la forme explicite de la distribution stationnaire du réseau et montrent la propriété d'insensibilité.

Remarquons que l'on peut utiliser la méthode des processus semi-markoviens généalisés, voir l'annexe A pour le modèle RGSMP de ces réseaux généraux de Kelly.

2.3.1 Modèle

Considérons un réseau ouvert de I files d'attente symétriques de taux de service unitaire 1.

Macro-état. Posons n_i le nombre de clients à la file i, $n = (n_1, \dots, n_I)$ le macro-état du réseau. Posons e_i le vecteur unitaire de dimension I dont la $i^{\text{ème}}$ composante est égale à 1 et les autres sont nulles.

Files symétriques. Les files d'attente sont symétriques de probabilités de positionnement $\delta_i(l, n)$ définies dans la Section 2.2.

Classes de clients. Posons $\mathcal{C}$ un ensemble dénombrable arbitraire de classes de clients. Sans perte de généralité, on suppose que chaque classe est associée à une seule file d'attente. Notons $\mathcal{C}_i$ l'ensemble de classes associées à la file d'attente i, alors les ensembles $\mathcal{C}_1, \dots, \mathcal{C}_I$ forment une répartition de l'ensemble $\mathcal{C}$.

Supposons que les clients de classe c arrivent suivant un processus de Poisson d'intensité ν_c. On appelle ces arrivées des arrivées *externes*. Ces clients demandent des services i.i.d. de loi exponentielle de moyenne $1/\mu_c$ à la file d'attente associée. Après l'accomplissement de service, un client de la classe c joint une autre classe c' avec la probabilité $p_{cc'}$, appelée la probabilité correspondante à une arrivée *interne*, ou quitte le réseau avec la probabilité :

$$p_c = 1 - \sum_{c' \in \mathcal{C}} p_{cc'}.$$

On suppose que le taux d'arrivée total est strictement positif : $\sum_{c \in \mathcal{C}} \nu_c > 0$ et que les clients finissent par quitter le réseau de telle sorte que le taux d'arrivée effectif λ_c des clients de la classe c soit la solution unique des équations de trafic :

$$\lambda_c = \nu_c + \sum_{c' \in \mathcal{C}} \lambda_{c'} p_{c'c}, \qquad \forall c \in \mathcal{C}. \tag{2.9}$$

En sommant sur toutes les classes c, on obtient :

$$\sum_{c' \in \mathcal{C}} \lambda_{c'} p_{c'} = \sum_{c \in \mathcal{C}} \nu_c. \tag{2.10}$$

Le taux de charge dû à la classe c est alors donné par :

$$\rho_c = \frac{\lambda_c}{\mu_c}.$$

Remarquons qu'en augmentant le nombre de classes, on peut décrire un ensemble arbitraire de routages dans un réseau.

Micro-état. Posons $x_i = (x_i(1), \ldots, x_i(n_i))$ l'état de la file i où $x_i(l)$ désigne la classe du client en position l de la file d'attente i. On appelle la donnée $x = (x_1, \ldots, x_I)$ le micro-état du système. Notons $\mathcal{X}$ l'ensemble de micro-états et $\mathcal{X}(n)$ l'ensemble des micro-états associés au même macro-état n.

Pour tout $l = 1, \ldots, n_i + 1$, on pose $T^{c,l}x$ le micro-état obtenu de x en ajoutant un client de classe c à la position l de la file associée i selon la règle de décalage classique.

De même, posons $T_{i,l}x$ le micro-état obtenu de x en effaçant le client en position l de la file i selon la règle de décalage classique, pour tous $i = 1, \ldots, I$ et $l = 1, \ldots, n_i$.

Permutations aléatoires. Soit $\mathcal{S}(n) = \mathcal{P}(n_1) \times \ldots \times \mathcal{P}(n_I)$, où $\mathcal{P}(m)$ désigne l'ensemble des permutations de m éléments, pour tout $m \geq 1$, et $\mathcal{P}(0)$ désigne l'ensemble consistant seulement en l'application d'identité sur $\{\emptyset\}$. Pour tout micro-état $x \in \mathcal{X}(n)$ et pour toute permutation $\sigma \in \mathcal{S}(n)$, notons $\sigma(x)$ le micro-état obtenu de x en permutant les clients suivant la permutation σ.

On suppose dans ce modèle que n'importe quel mouvement de client peut provoquer une permutation de clients à chaque file d'attente. La distribution des permutations peut dépendre du macro-état et du client qui provoque la permutation. Pour tout macro-état n, et toutes classes c, c', posons $\alpha_c(n, \cdot), \beta_c(n, \cdot), \gamma_{cc'}(n, \cdot)$ des distributions arbitraires sur $\mathcal{S}(n)$.

Lorsqu'un nouveau client de classe c arrive dans le macro-état n, les autres clients sont permutés selon la distribution $\alpha_c(n, \cdot)$ avant l'arrivée de ce client. En particulier, le micro-état devient $\sigma(x)$ avec la probabilité $\alpha_c(n, \sigma)$ avant cette arrivée.

De même, lorqu'un client de classe c quitte le réseau dans l'état n, avec $c \in \mathcal{C}_i$, les autres clients sont permutés selon la distribution $\beta_c(n - e_i, \cdot)$.

Enfin, lorsqu'un client rejoint une certaine classe c' depuis la classe c après l'accomplissement de service dans macro-état n, avec $c \in \mathcal{C}_i$ et $c' \in \mathcal{C}_j$, les autres clients sont permutés selon la distribution $\gamma_{cc'}(n - e_i, \cdot)$ après le départ de la file i et avant l'arrivée à la file j.

2.3.2 Forme produit et insensibilité

L'évolution du micro-état du système définit un processus markovien sur un espace d'états dénombrable.

Théorème 2.10 *Le micro-état du système possède la distribution stationnaire*

$$\pi(x) = C \prod_{i=1}^{I} \prod_{l=1}^{n_i} \rho_{x_i(l)},$$

si le réseau est stable (2.14) : $0 < C < \infty$, où C est la constante de normalisation que l'on déterminera plus tard dans l'équation (2.13).

Preuve. Le point clef de la démonstration est que la mesure π est invariante aux permutations.

Tout d'abord, on montre que la mesure de probabilité π satisfait les équations de balance partielle pour chaque classe de clients c à chaque position l de la file d'attente associée i.

Considérons un micro-état $x \in \mathcal{X}(n)$ tel que $x_i(l) = c$. Le flux de probabilité correspondant au départ du client en position l de la file i dans le micro-état x est donné par

$$\pi(x)\mu_c\delta_i(l,n). \tag{2.11}$$

D'autre part, le flux de probabilité correspondant à une arrivée externe à la position l de la file i *menant* le système au micro-état x est donné par

$$\sum_{\sigma \in \mathcal{S}(n-e_i)} \pi(x')\nu_c\alpha_c(n-e_i,\sigma)\delta_i(l,n),$$

où x' désigne l'unique micro-état satisfaisant $x = T^{c,l}\sigma(x')$. En utilisant la relation $\pi(x) = \pi(x')\rho_c$, on déduit que ce flux de probabilité est égal à

$$\pi(x)\frac{\nu_c}{\rho_c}\delta_i(l,n) \sum_{\sigma \in \mathcal{S}(n-e_i)} \alpha_c(n-e_i,\sigma) = \pi(x)\delta_i(l,n)\frac{\mu_c}{\lambda_c}\nu_x.$$

De même, le flux de probabilité correspondant à une arrivée interne à la position l de la file i *menant* le système au état x est donné par

$$\sum_{i=1}^{I} \sum_{c' \in \mathcal{C}_j} \sum_{l'=1}^{n_j+1} \sum_{\sigma \in \mathcal{S}(n-e_i)} \pi(x')\mu_{c'}\delta_j(l', n+e_j-e_i)p_{c'c}\gamma_{c'c}(n-e_i,\sigma)\delta_i(l,n),$$

où x' désigne l'unique micro-état satisfaisant $x = T^{c,l}\sigma(T_{j,l'}x')$. En utilisant la relation $\pi(x) = \pi(x')\rho_c/\rho_{c'}$, on déduit que ce flux de probabilité est égal à

$$\begin{aligned}&\sum_{i=1}^{I} \sum_{c' \in \mathcal{C}_j} \pi(x)\frac{\rho_c'}{\rho_c}\mu_{c'}p_{c'c}\delta_i(l,n) \sum_{l'=1}^{n_j+1} \delta_j(l', n+e_j-e_i) \sum_{\sigma \in \mathcal{S}(n-e_i)} \gamma_{c'c}(n-e_i,\sigma)\\ &= \pi(x)\delta_i(l,n)\frac{\mu_c}{\lambda_c}\sum_{c' \in \mathcal{C}} \lambda_{c'}p_{c'c}.\end{aligned}$$

En sommant ces deux flux, on obtient

$$\pi(x)\delta_i(l,n)\frac{\mu_c}{\lambda_c}\left(\nu_c + \sum_{c'\in\mathcal{C}} \lambda_{c'} p_{c'c}\right)$$

qui est égal au flux de probabilité (2.11) selon les équations de trafic (2.9).

Afin de conclure la démonstration, il reste à vérifier les équations de balance partielle pour la source. Soit $x \in \mathcal{X}(n)$, le flux de probabilité correspondant aux arrivées externes dans le micro-état x est donné par

$$\pi(x)\sum_{c\in\mathcal{C}} \nu_c. \tag{2.12}$$

D'ailleurs, le flux de probabilité correspondant au départ d'un client *menant* le système au micro-état x est donné par

$$\sum_{j=1}^{I}\sum_{c'\in\mathcal{C}_j}\sum_{l'=1}^{n_j+1}\sum_{\sigma\in\mathcal{S}(n)} \pi(x')\mu_{c'}\delta_j(l',n+e_j)p_{c'}\beta_{c'}(n,\sigma),$$

où x' désigne l'unique micro-état satisfaisant $x = \sigma(T_{j,l'}x')$.

En utilisant la relation $\pi(x) = \pi(x')/\rho_{x'}$, on obtient le flux de probabilité

$$\sum_{j=1}^{I}\sum_{c'\in\mathcal{C}_j} \pi(x)\rho_{c'}\mu_{c'}p_{c'}\sum_{l'=1}^{n_j+1}\delta_j(l',n+e_j)\sum_{\sigma\in\mathcal{S}(n)}\beta_{c'}(n,\sigma) = \pi(x)\sum_{c'\in\mathcal{C}}\lambda_{c'}p_{c'},$$

qui est égal au flux de probabilité (2.12) grâce aux équations de trafic (2.10).

□

Posons ρ_i le taux de charge de la file i :

$$\rho_i = \sum_{c\in\mathcal{C}_i} \rho_c.$$

En conséquence du Théorème 2.10, le macro-état possède la mesure de probabilité stationnaire

$$\pi'(n) = \sum_{x\in\mathcal{X}(n)} \pi(x) = C\sum_{x\in\mathcal{X}(n)}\prod_{i=1}^{I}\prod_{l=1}^{n_i}\rho_{x_i(l)} = C\prod_{i=1}^{I}\rho_i^{n_i},$$

où la dernière équation se déduit de l'invariance à permutations de π. On obtient aussi l'expression de la constante C :

$$C = \prod_{i=1}^{I}(1-\rho_i). \tag{2.13}$$

En particulier, le réseau est stable si et seulement si

$$\rho_i < 1, \qquad \forall i = 1,\dots,I. \tag{2.14}$$

Dans ce cas là, la distribution stationnaire du macro-état est à forme produit et est donnée par

$$\pi'(n) = \prod_{i=1}^{I} (1 - \rho_i) \rho_i^{n_i}.$$

Alors la distribution stationnaire du macro-état ne dépend des classes de clients qu'à travers les taux de charge de chaque file d'attente. En utilisant des classes de clients pour représenter des phases de service exponentielles, on peut déduire la propriété d'insensibilité, au moins pour les distributions de services à phases.

Théorème 2.11 *Si le réseau est stable (2.14), il est insensible aux distributions des demandes de service, au moins pour les distributions à phases, la distribution stationnaire du nombre de clients à chaque file est à forme produit et ne dépend que des taux de charge $\rho_1, \dots, \rho_I$:*

$$\pi'(n) = \prod_{i=1}^{I} (1 - \rho_i) \rho_i^{n_i},$$

À l'état stationnaire, le nombre moyen de clients à la file i est donné par

$$\mathbb{E}[n_i] = \frac{\rho_i}{1 - \rho_i},$$

le temps de séjour moyen d'un client à la file i est donné par

$$\mathbb{E}[T] = \frac{\rho_i}{\lambda_i (1 - \rho_i)},$$

où $\lambda_i = \sum_{c \in \mathcal{C}_i} \lambda_c$.

Plus généralement, la propriété d'insensibilité reste valable pour la distribution stationnaire du nombre de clients à chaque classe. Posons y_c le nombre de clients de la classe c dans micro-état $x \in \mathcal{X}(n)$:

$$y_c = \sum_{i=1}^{n_i} \mathbb{I}(x_i(l) = c), \qquad c \in \mathcal{C}_i.$$

Posons $y = (y_c, c \in \mathcal{C})$, on a le résultat suivant.

Théorème 2.12 *Sous les conditions de stabilité (2.14), la distribution stationnaire du nombre de clients de chaque classe est insensible, ne dépend des distributions des demandes de service qu'à travers leur moyenne et est à forme produit :*

$$\pi''(y) = \prod_{i=1}^{I} (1 - \rho_i) n_i! \times \prod_{c \in \mathcal{C}} \frac{\rho_c^{y_c}}{y_c!}.$$

2.3.3 Extensions

Comme pour les réseaux de Jackson (Section 2.2), les résultats restent valables dans les cas suivants

- les permutations aléatoires sont provoquées par un processus ponctuel externe, par exemple, un processus de Poisson indépendant.

- réseaux fermés de Kelly avec permutations aléatoires.

Considérons maintenant le cas où les capacités de service dépendent du nombre de clients à chaque file : $\phi_i(n)$; et les taux d'arrivée et les probabilités de routage dépendent aussi de ces nombres de clients de telle sorte que les taux de charge de chaque file i soient en fonction de n : $\rho_i(n)$. Alors, on peut montrer que le réseau est insensible aux distributions des demandes de service si et seulement si les fonctions $\psi_i(n) = \rho_i(n - e_i)/\phi_i(n), i = 1, \dots, I$, sont équilibrées. La condition suffisante se déduit des équations de balance partielle, et la condition nécessaire se montre par récurrence sur le nombre de files d'attente dans le réseau.

2.4 Disciplines non symétriques

Jusqu'à ici, on a considéré l'insensibilité des files d'attente symétriques et leurs réseaux avec permutations aléatoires. Dans ce chapitre, on introduira quelques nouvelles disciplines de service insensibles non symétriques.

2.4.1 Disciplines avec labels

Considérons la discipline de service pour laquelle on associe à chaque client un certain label parmi L labels donnés. Les clients de label l demandent des services i.i.d. de moyenne μ_l^{-1}.

Posons n_l le nombre de clients de label l dans la file alors l'état de la file d'attente est le vecteur $n = (n_1, \dots, n_L)$.

Les clients arrivent dans la file suivant un processus de Poisson d'intensité ν et à un nouveau client est associé un label l avec la probabilité $\delta(l, n)$ dans l'état n :

$$\sum_{l=1}^{L} \delta(l, n) = 1.$$

Dans l'état n, les clients de label l partagent équitablement une proportion $\delta'(l, n)$ de capacité de service :

$$\sum_{l=1}^{L} \delta'(l, n) = 1.$$

L'évolution de la file peut être décrite par une chaîne de Markov sur un sous-ensemble de $\mathbb{N}^L$. On suppose que cette chaîne de Markov est irréductible.

Cette discipline peut être vue comme un réseau de files d'attente PS (Section 1.4.3) avec taux d'arrivée et capacités de service dépendants du nombre de clients à chaque file :

Capacités de sevice : $\delta'(l,n), l = 1,\ldots,L$.

Arrivées poissoniennes d'intensité $\nu\delta(l,n), l = 1,\ldots,L$.

D'après la Section 1.4.3, ce réseau est insensible si et seulement si les fonctions

$$\psi_l(n) = \frac{\nu\delta(l,n-e_l)}{\mu_l\delta'(l,n)}$$

sont équilibrées, e_l étant le vecteur unitaire de dimension L dont la $l^{\text{ème}}$ composante est égale à 1 et les autres sont nulles.

En particulier, cette nouvelle discipline donne la file d'attente LIFO préemptive avec un nombre limité de L places si :

$$\left\{\begin{array}{lll}
\delta(1,0) & = \delta'(1,e_1) & = 1 \\
\delta(2,e_1) & = \delta'(2,e_1+e_2) & = 1 \\
\delta(3,e_1+e_2) & = \delta'(3,e_1+e_2+e_3) & = 1 \\
 & \vdots & \\
\delta(L,e_1+\ldots+e_{L-1}) & = \delta'(L,e_1+\ldots+e_L) & = 1.
\end{array}\right.$$

Dans cet exemple, les labels correspondent aux positions de clients dans la file d'attente.

La discipline LIFO préemptive considérée ci-dessus est symétrique mais en général, notre discipline avec labels ne l'est pas. Par exemple, considérons un cas simple où $L = 2$ et $\delta(1,n) = \delta'(1,n) = 0$ sauf si $n_2 = 0$. Dans cette discipline, la priorité est donnée au label 2, c.à.d. que si les clients de label 2 sont présents dans la file, toute la capacité est donnée à ces clients et un nouveau client est toujours associé le label 2. Cette discipline contient alors deux classes de priorité de telle sorte que la file alloue toute sa capacité aux clients de la classe 2, la plus favorisée. Les labels correspondent alors aux classes de priorité.

Dans ce cas, l'état de la file est de la forme (n_1,n_2), où n_i désigne le nombre de clients de label i. Dans la Figure 2.11, sont illustrées deux transitions possibles à partir de l'état $(2,3)$: Une arrivée a lieu avec le taux de transition ν, cet arrivé est associé le label 2, menant la file à l'état $(2,4)$, et un départ de l'un des trois clients de classe 2 a lieu avec le taux de transition μ, menant la file à l'état $(2,2)$.

On peut former une autre discipline de service de priorité en considérant $L = 2$ et pour tout $n \geq 1$:

$$\left\{\begin{array}{l}
\delta(2,ne_1) = p_n \\
\delta'(2,ne_1+e_2) = 1.
\end{array}\right.$$

Ainsi, si n clients sont présents dans la file à une arrivée, le nouveau client se sert immédiatement de toute la capacité de la file avec la probabilité p_n. Dans ce cas là, aucune arrivée

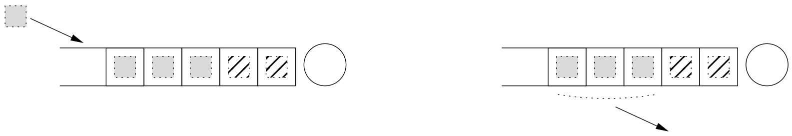

FIG. 2.11 – Une file avec 2 classes de priorité : Arrivée d'un nouveau client de la classe 2 et départ de l'un des trois clients de cette classe.

ne sera acceptée pendant le service de ce client favorisé. Et sinon, ce nouveau client partage équitablement la capacité avec les autres clients avec la probabilité $1 - p_n$.

Par convention, $p_0 = 0$, c.à.d. que si un client arrive lorsque la file est vide, il est associé à la première classe, l'état de la file devient $(1, 0)$. L'espace d'états est alors l'ensemble $G = \{(0,0)\} \cup \{(n_1, n_2) : n_1 > 0, n_2 = 0, 1\}$.

L'insensibilité de cette file est montrée comme suit : On s'aperçoit que pour tout état $n = (n_1, n_2) \in G$, il y a un seul chemin direct de cet état à l'état vide $0 = (0, 0)$, alors le produit d'une fonction $\psi_l(n)$ quelconque prise sur un chemin direct de n à 0 ne dépend pas de ce chemin. En particulier, notre fonction ψ est équilibrée, alors cette discipline de service est insensible.

Enfin, considérons la discipline de *transfert* avec $L = 2$ labels, telle que s'il y a autant de clients de label 1 que ceux de label 2, toute la capacité de la file est allouée équitablement aux clients de label 1 et la file n'accepte que des arrivées de clients de label 2, sinon, si les clients de label 2 sont plus nombreux, la capacité est alors transférée à ces clients, c.à.d. que la capacité de la file est allouée équitablement aux clients de label 2, et la file n'accepte que des arrivées de clients de label 1.

À partir de l'état vide, la file ne peut atteindre que les états de formes (n, n) ou $(n, n+1), n \in \mathbb{N}$. Le serveur sert successivement les deux classes de clients.

FIG. 2.12 – Discipline de transfert : Arrivées et départs dans les états (n, n) et $(n, n + 1)$.

2.4.2 Nombre fini de places

Dans la Section 1.3.2, on a montré que la décomposabilité d'un modèle RGSMP assure son insensibilité et on a remarqué qu'il n'y a pas d'équivalence entre ces deux propriétés. Le modèle que l'on introduira dans cette partie illustrera cette propriété de non-équivalence. En effet, notre modèle sera insensibile mais non-décomposable en forme produit.

Considérons une file d'attente dont le nombre de places est limité à une certaine valeur N. Supposons que les clients arrivent dans la file suivant un processus de Poisson d'intensité ν. S'il y a déjà N clients dans la file, aucune arrivée n'est permise, dans le cas contraire où il n'y a que n clients dans la file, $n < N$, un nouveau client choisit la position l avec la probabilité $\delta(l, n+1)$.

Une proportion $\delta'(l,n)$ de capacité de service est allouée au client en position l lorsqu'il y a n clients dans la file, $n \leq N$.

Commençons par les deux cas les plus simples avec $N = 1$ et $N = 2$.

N égal à 1. Dans ce cas, à son arrivée, un client prend forcément la place unique dans la file et le serveur donne toute sa capacité à ce client :

$$\delta'(1,1) = \delta(1,1) = 1.$$

Ce cas est trivial car cette file est à la fois symétrique et insensible.

N égal à 2. On a aussi $\delta'(1,1) = \delta(1,1) = 1$. Le problème est de trouver les valeurs de $\delta'(1,2), \delta'(2,2)$ et $\delta(1,2), \delta(2,2)$ qui assurent l'insensibilité de la file.

Pour la condition nécessaire, on suppose que les clients demandent les services aléatoires de 2 phases exponentielles de même moyenne $1/\mu$. Si les clients arrivent dans la file selon un processus de Poisson d'intensité ν, le taux de charge est donné par

$$\rho = 2\frac{\nu}{\mu}.$$

Le micro-état de la file se compose de la phase de service de chaque client et l'espace des micro-états contient 7 éléments :

$$\mathcal{X} = \{(0), (1), (2), (1,1), (1,2), (2,1), (2,2)\}.$$

La distribution stationnaire π doit satisfaire un système de 7 équations de balance globale. Si la distribution stationnaire du nombre de clients est insensible, π doit satisfaire de plus les conditions suivantes

$$\begin{cases} \pi(1) + \pi(2) & = \rho\pi(0) \\ \pi(1,1) + \pi(1,2) + \pi(2,1) + \pi(2,2) & = \rho^2\pi(0) \end{cases}$$

En étudiant le système d'équations de balance globale et ces deux conditions d'insensibilité, on peut trouver la condition nécessaire de l'insensibilité :

$$\big(\delta'(1,2) - \delta'(2,2)\big)\big(\delta'(1,2) - \delta(1,2)\big) = 0$$

Cette condition est équivalente à la symétrie de la file, qui est aussi une condition suffisante [Kel79].

Alors dans les deux premiers cas où $N = 1$ et $N = 2$, l'insensibilité est équivalente à la symétrie. Au contraire, pour à partir de $N = 3$, le système d'équations de balance globale devient trop compliqué pour étudier afin de trouver la condition nécessaire et suffisante de l'insensibilité. Dans la suite, on donnera une condition suffisante qui est moins restreinte que la symétrie. D'une part, cela montre que la symétrie n'est plus nécessaire pour l'insensibilité si $N \geq 3$. D'autre part, comme les disciplines avec labels considérées dans la section précédente, cette nouvelle discipline non symétrique peut donner une perspective pour identifier la condition nécessaire et suffisante de l'insensibilité.

N plus grand ou égal à 3.

On étudiera ici une discipline non symétrique avec

$$\begin{cases} \delta'(l,n) = \delta(l,n), & \forall n < N, l = 1, \dots, n \\ \delta'(l,N) = \frac{1}{N}, & \forall l = 1, \dots, N. \end{cases} \tag{2.15}$$

Tout d'abord, remarquons qu'en faisant le nombre de places tendre vers ∞, on obtient une file symétrique sans limite de place, qui est insensible.

Retournons maintenant aux cas où N est fini. Si $\delta(l,N) \neq 1/N$, cette file d'attente est non symétrique et en général, sensible à la distribution des demandes de service, mais on montrera qu'elle reste insensible si les probabilités de positionnement satisfont les conditions suivantes

$$\begin{cases} (N-1)\delta(1,N) & + & \delta(2,N) & = 1 \\ (N-2)\delta(2,N) & + & 2\delta(3,N) & = 1 \\ & \vdots & & \\ \delta(N-1,N) & + & (N-1)\delta(N,N) & = 1 \end{cases} \tag{2.16}$$

Remarquons qu'en sommant ces égalités, on retrouve la condition :

$$\sum_{l=1}^{N} \delta(l,N) = 1,$$

de telle sorte que ces équations admettent une infinité de solutions dont l'une est la discipline symétrique $\delta(l,N) = 1/N$.

Distribution stationnaire.

On suppose que les demandents de services sont i.i.d. de loi mélange de distributions d'Erlang, c.à.d. une somme d'un nombre aléatoire de phases exponentielles de moyenne μ^{-1}. Posons $s(k)$ la probabilité que le nombre de phases soit égal à k, $\bar{s}(k)$ la probabilité que le nombre de phases dépasse k et $\bar{s}$ le nombre moyen de phases. Alors

$$\sum_{k=1}^{\infty} s(k) = 1, \quad \bar{s}(k) = \sum_{j \geq k} s(j), \quad \text{et} \quad \bar{s} = \sum_{k=1}^{\infty} k s(k) = \sum_{k=1}^{\infty} \bar{s}(k).$$

Le taux de charge de la file est alors

$$\rho = \frac{\nu \bar{s}}{\mu}.$$

Posons x_l la phase de service du client en position l, $x = (x_1, \dots, x_n)$ le micro-état de la file, où $|x| = n$ désigne le nombre de clients dans la file, $n \leq N$.

Théorème 2.13 *Une mesure stationnaire de l'état de la file est de la forme :*

$$\pi(x) = \begin{cases} (\frac{\nu}{\mu})^n \prod_{l=1}^{n} \bar{s}(x_l) & si\ |x| = n < N \\ (\frac{\nu}{\mu})^N \prod_{l=1}^{N} \bar{s}(x_l) f(x) & si\ |x| = N, \end{cases} \tag{2.17}$$

où $f(x)$ est une fonction définie ultérieurement par (2.24).

Preuve. La démonstration est basée sur les équations de balance partielle.

Posons $T^{k,l}x$ (resp. $T_l x$) le micro-état obtenu de x en ajoutant un client en phase de service k à la position l (resp. le micro-état obtenu de x en effaçant le client en position l).

- Si $0 \leq n < N-1$, le flux de probabilité d'un départ *menant* la file à l'état x est donné par

$$\sum_{l=1}^{n+1} \sum_{k \geq 1} \pi(T^{k,l}x)\mu\delta'(l,n+1)\frac{s(k)}{\bar{s}(k)} = \pi(x)\nu \sum_{l=1}^{n+1} \sum_{k \geq 1} \delta'(l,n+1)s(k),$$

  car $\pi(T^{k,l}x) = \pi(x)\bar{s}(k+1)/\bar{s}(k)$.

  Ce flux de probabilité est égal au flux de probabilité d'une arrivée dans l'état x :

$$\pi(x)\nu.$$

- Si $1 \leq n \leq N-1$, le flux de probabilité correspondant à une arrivée *menant* la file à l'état x est donné par

$$\sum_{l=1}^{n} \pi(T_l x)\nu\delta(l,n)\mathbb{I}(x_l = 1) = \pi(x)\mu \sum_{l=1}^{n} \delta(l,n)\mathbb{I}(x_l = 1),$$

  car $\bar{s}(x_l) = \bar{s}(1) = 1$ et $\pi(T_l x) = \pi(x)\mu/\nu$ si $x_l = 1$.

  Par ailleurs, le flux de probabilité correspondant à une transition de phases *menant* la file à l'état x est donné par

$$\sum_{l=1}^{n} \pi(x - e_l)\mu\delta'(l,n)\frac{\bar{s}(x_l)}{\bar{s}(x_l - 1)}\mathbb{I}(x_l > 1) = \pi(x)\mu \sum_{l=1}^{n} \delta(l,n)\mathbb{I}(x_l > 1),$$

  car $\delta'(l,n) = \delta(l,n)$ si $n \leq N-1$, et $\pi(x - e_l) = \pi(x)\bar{s}(x_l - 1)/\bar{s}(x_l)$ si $x_l > 1$, e_l étant le vecteur unitaire de dimension n dont la $l^{\text{ème}}$ composante est égale à 1 et les autres sont nulles.

  En sommant ces deux flux de probabilité, on obtient le flux de probabilité correspondant à un accomplissement de phase de service d'un client lorsque la file est dans l'état x :

$$\pi(x)\mu.$$

Alors π est une mesure stationnaire du micro-état de la file si et seulement si elle satisfait les deux équations de balance partielle suivantes

$$\pi(x)\nu = \sum_{l=1}^{N} \sum_{k \geq 1} \pi(T^{k,l}x)\mu\delta'(l,N)\frac{s(k)}{\bar{s}(k)}, \quad \text{si } |x| = N-1, \tag{2.18}$$

et

$$\begin{aligned}\pi(x)\mu &= \sum_{l=1}^{N-1} \pi(T_l x)\nu\delta(l,N)\mathbb{I}(x_l = 1) \\ &\quad + \sum_{l=1}^{N} \pi(x - e_l)\mu\delta'(l,N)\frac{\bar{s}(x_l)}{\bar{s}(x_l - 1)}\mathbb{I}(x_l > 1), \quad \text{si } |x| = N. \end{aligned} \tag{2.19}$$

L'équation (2.18) est équivalente à la condition suivante sur la fonction $f(x)$:

$$\sum_{l=1}^{N}\sum_{k\geq 1} s(k)f(T^{k,l}x) = N, \qquad \forall x : |x| = N-1. \tag{2.20}$$

Et l'équation (2.19) est équivalente à la formule récursive de la fonction $f(x)$:

$$f(x) = \sum_{l=1}^{N}\Big(\delta(l,N)\mathbb{I}(x_l = 1) + \frac{1}{N}f(x - e_l)\mathbb{I}(x_l > 1)\Big), \qquad \forall x : |x| = N. \tag{2.21}$$

Alors s'il existe une fonction f satisfaisant (2.20) et (2.21), la mesure π donnée dans le Théorème 2.13 forme une mesure stationnaire du micro-état x. Dans la suite, on montrera l'existence d'une telle fonction.

En appliquant l'équation (2.21) pour $x = e_1 + e_2 + \ldots + e_N$, on obtient la condition initiale de la fonction $f(x)$:

$$f(e_1 + e_2 + \ldots + e_N) = \sum_{l=1}^{N}\delta(l,N) = 1. \tag{2.22}$$

Alors la fonction $f(x)$ est complètement déterminée par cette condition initiale (2.22) et la formule récursive (2.21). Il ne nous reste qu'à montrer que la fonction définie récursivement par (2.22) et (2.21) remplit la condition (2.20) pour tout x tel que $|x| = N-1$. En particulier, on montrera par récurrence une condition plus forte :

$$\sum_{l=1}^{N} f(T^{k,l}x) = N, \qquad \forall k \geq 1, \forall x : |x| = N-1. \tag{2.23}$$

Cette condition signifie que pour tout état $x, |x| = N$, la somme de f sur tous les états obtenus à partir de x en déplaçant un client particulier, est égal à N, indépendamment de sa phase de service k.

Commençons par $x = e_1+e_2+\ldots+e_{N-1}$ et $k = 1$, alors $T^{k,l}x = e_1+e_2+\ldots+e_N, \forall l = 1,\ldots,N$, et cette condition (2.23) devient équivalent à la condition initiale (2.22) :

$$\sum_{l=1}^{N} f(e_1 + e_2 + \ldots + e_N) = N.$$

La condition (2.23) est ensuite jusitifiée à être remplie pour le couple (k, x), $x = e_1 + e_2 + \ldots + e_{N-1}$, $\forall k \geq 1$, par récurrence sur k :

$$\begin{aligned} f(T^{k+1,l}x) &= \sum_{l'=1}^{l-1} \delta(l', N) + \frac{1}{N} f(T^{k,l}x) + \sum_{l'=l+1}^{N} \delta(l', N) \\ &= 1 - \delta(l, N) + \frac{1}{N} f(T^{k,l}x). \end{aligned}$$

En sommant sur l, on obtient l'égalité (2.23) pour le couple $(k+1, x)$ par récurrence :

$$\sum_{l=1}^{N} f(T^{k+1,l}x) = N - \sum_{l=1}^{N} \delta(l, N) + \frac{1}{N} \sum_{l=1}^{N} f(T^{k,l}x) = N.$$

Supposons maintenant que la condition (2.23) est remplie pour (k, x'), pour tout $k \geq 1$ et pour tout x' tel que $x' \prec x$, c.à.d. $x'_l \leq x_l \; \forall l = 1, \ldots, N$ et qu'il existe un indice l tel que $x'_l < x_l$. On montrera par récurrence sur k que cette condition est aussi remplie pour $(k, x), \forall k \geq 1$.

Pour $k = 1$, on a :

$$\begin{aligned} f(T^{1,l}x) &= \sum_{l'=1}^{l-1} \Big(\delta(l', N) \mathbb{I}(x_{l'} = 1) + \frac{1}{N} f(T^{1,l}(x - e_{l'})) \mathbb{I}(x_{l'} > 1) \Big) + \delta(l, N) \\ &\quad + \sum_{l'=l}^{N-1} \Big(\delta(l'+1, N) \mathbb{I}(x_{l'} = 1) + \frac{1}{N} f(T^{1,l}(x - e_{l'})) \mathbb{I}(x_{l'} > 1) \Big). \end{aligned}$$

En utilisant (2.23) pour le couple $(1, x - e_{l'}), l' = 1, \ldots, N$, on vérifie que cette condition est aussi satisfaite pour le couple $(1, x)$:

$$\sum_{l=1}^{N} f(T^{1,l}x) = N.$$

Supposons ensuite que la condition (2.23) est rempie pour le couple (k, x), on vérifiera cette condition pour $(k+1, x)$. On a :

$$\begin{aligned} f(T^{k+1,l}x) &= \sum_{l'=1}^{l-1} \Big(\delta(l', N) \mathbb{I}(x_{l'} = 1) + \frac{1}{N} f(T^{k+1,l}(x - e_{l'})) \mathbb{I}(x_{l'} > 1) \Big) \\ &\quad + \frac{1}{N} f(T^{k,l}x) \\ &\quad + \sum_{l'=l}^{N-1} \Big(\delta(l'+1, N) \mathbb{I}(x_{l'} = 1) + \frac{1}{N} f(T^{k+1,l}(x - e_{l'})) \mathbb{I}(x_{l'} > 1) \Big). \end{aligned}$$

En utilisant (2.23) pour $(k+1, x - e_{l'}), l' = 1, \ldots, N$, et pour (k, x), on obtient pour le couple $(k+1, x)$:

$$\sum_{l=1}^{N} f(T^{k+1,l}x) = N.$$

Alors on a (2.23) pour $(k,x), \forall k \geq 1, \forall x : |x| = N-1$, si (2.23) est vraie pour $(k,x'), \forall k \geq 1, \forall x' : |x'| = N-1, x' \prec x$. Or initialement, cette condition se vérifie pour (k,x), $x = e_1 + e_2 + \ldots + e_{N-1}$, $\forall k \geq 1$, alors on déduit que (2.23) est remplie pour tout couple $(k,x), \forall k \geq 1, \forall x : |x| = N-1$.

La fonction $f(x)$ déterminée récursivement par :

$$\begin{cases} f(x) &= 1, \quad \text{pour } x = e_1 + e_2 + \ldots + e_N \\ f(x) &= \sum_{l=1}^{N} \Big(\delta(l,N) \mathbb{I}(x_l = 1) + \frac{1}{N} f(x - e_l) \mathbb{I}(x_l > 1) \Big), \qquad \forall x : |x| = N \end{cases} \tag{2.24}$$

satisfait la condition (2.23) :

$$\sum_{l=1}^{N} f(T^{k,l}x) = N, \qquad \forall k \geq 1, \forall x : |x| = N-1.$$

Par conséquence, la mesure correspondante π définie dans le Théorème 2.13 donne une mesure stationnaire du micro-état x.

□

Impact des probabilités de positionnement $\delta(l,N)$ sur la fonction f

Remarquons que la fonction f dépend fortement des probabilités $\delta(1,N), \ldots, \delta(N,N)$. Même si on est dans le cas $N \geq 3$, nos études conviennent aussi avec le cas $N = 2$, et même dans ce cas simple, on peut voir le grand impact des probabilités de positionnement sur la fonction f. Afin de simplifier les notations, posons $\delta_1 = \delta(1,2)$ et $\delta_2 = \delta(2,2)$. Dans ce cas, la file est insensible pour tous δ_1, δ_2 tels que

$$\delta_1 + \delta_2 = 1,$$

et la fonction f est définie par :

$$\begin{cases} f(1,1) &= 1 \\ f(u,v) &= \delta_1 \mathbb{I}(u=1) + \frac{1}{2} f(u-1,v) \mathbb{I}(u>1) \\ & \quad + \delta_2 \mathbb{I}(v=1) + \frac{1}{2} f(u,v-1) \mathbb{I}(v>1) \qquad u,v \geq 1. \end{cases}$$

La forme explicite de $f(u,v)$ est très compliquée alors on se contente à regarder l'impact de δ_1 et δ_2 sur les valeurs de $f(k,1)$ et $f(1,k)$:

$$\begin{cases} f(k,1) &= \frac{1}{2^{k-1}} + 2\delta_2 (1 - \frac{1}{2^{k-1}}) \\ f(1,k) &= \frac{1}{2^{k-1}} + 2\delta_1 (1 - \frac{1}{2^{k-1}}). \end{cases}$$

Alors on peut vérifier que la condition (2.23) est remplie pour $x = (1)$: $f(k,1) + f(1,k) = 2$.

Maintenant, si chaque nouveau client choisit sa position suivant une distribution uniforme, c.à.d. que $\delta_1 = \delta_2 = 1/2$, $f(k,1) = f(1,k) = 1$ qui coïncide avec le cas d'une file symétrique et en fait la file est symétrique car $\delta(l,n) = \delta'(l,n)$ pour tous n, l.

Par contre, si un nouveau client choisit toujours de se placer à la tête de la file, c.à.d. que $\delta_1 = 1$ et $\delta_2 = 0$, la file est non symétrique et $f(k,1) = 1/2^{k-1} << f(1,k)$. Dans ce cas, le plus nouveau des deux clients dans la file (celui en tête de la file) a une forte probabilité d'être à sa première phase de service, sachant que l'un des deux clients l'est.

Insensibilité

La file d'attente est insensible à la distribution des demandes de service, au moins pour les mélanges de distributions d'Erlang :

Théorème 2.14 *La file d'attente est insensible, une mesure stationnaire du nombre de clients dans la file est à forme géométrique et ne dépend que de la charge ρ :*

$$\pi'(n) = \rho^n, \qquad \forall n \leq N.$$

Preuve. Pour $n < N$, on a :

$$\pi'(n) = \sum_{x:|x|=n} (\frac{\nu}{\mu})^n \prod_{l=1}^{n} \bar{s}(x_l) = \rho^n.$$

Pour $n = N$, on a :

$$\pi'(N) = \sum_{x:|x|=N} (\frac{\nu}{\mu})^N \prod_{l=1}^{N} \bar{s}(x_l) f(x)$$

Remarquons que pour toute permutation $\sigma \in \mathcal{P}(N)$, on a :

$$\prod_{l=1}^{N} \bar{s}(\sigma(x)_l) = \prod_{l=1}^{N} \bar{s}(x_l),$$

et grâce à la condition (2.23), il se déduit que :

$$\sum_{\sigma \in \mathcal{P}(N)} f(\sigma(x)) = \sum_{\sigma \in \mathcal{P}(N)} 1.$$

Par conséquence, le terme $\pi'(N)$ reste inchangé si on remplace $f(x)$ par sa moyenne sur toutes les permutations $\sigma \in \mathcal{P}(N)$:

$$\begin{aligned}
\pi'(N) &= \sum_{x:|x|=N} (\frac{\nu}{\mu})^N \prod_{l=1}^{N} \bar{s}(x_l) f(x) \\
&= \sum_{x:|x|=N} (\frac{\nu}{\mu})^N \prod_{l=1}^{N} \bar{s}(x_l) \frac{\sum_{\sigma \in \mathcal{P}(N)} f(\sigma(x))}{\sum_{\sigma \in \mathcal{P}(N)} 1} \\
&= \sum_{x:|x|=N} (\frac{\nu}{\mu})^N \prod_{l=1}^{N} \bar{s}(x_l) \\
&= \rho^N.
\end{aligned}$$

Alors la file est insensible à la distribution des demandes de service, au moins pour les mélanges de distributions d'Erlang, et la distribution stationnaire du nombre de clients est à forme produit $\pi'(n) = C\rho^n, \forall n \leq N$, où C désigne la constance de normalisation. Par conséquence, le temps de séjour moyen d'un client dans la file est aussi insensible à la distribution des demandes de service.

$\square$

Décomposabilité

On a introduit une nouvelle discipline insensible non symétrique. Remarquons que cette discipline n'est pas décomposable en forme produit (Proposition 1.16) parce que dans le cas exponentiel, les équations de balance locale (1.25) ne sont pas satisfaites pour le macro-état N :

Le taux de mort du client en position l est :

$$\pi'(N)\frac{\mu}{N},$$

tandis que son taux de naissance est :

$$\pi'(N-1)\nu\delta(l,N) = \pi'(N)\mu\delta(l,N).$$

Si $\delta(l,N) \neq 1/N$, ces deux taux sont différents, les équations de balance locale ne sont pas satisfaites et la file n'est pas décomposable en forme produit. Pourtant, cette file d'attente est insensible, ce qui montre que la propriété de décomposabilité est plus forte que l'insensibilité.

Sensibilité du contre exemple

Si les conditions (2.16) sur les probabilités de positionnement $\delta(l,N)$ ne sont pas satisfaites, la file d'attente est sensible en général. Considérons l'exemple simple d'une file d'attente limitée à $N = 3$ places, avec le processus d'arrivées poissonnien d'intensité $\nu = 1$ et un nouveau client est toujours placé à la fin de la file. Cette file d'attente dispose d'un serveur de taux de service 2 donné en totalité au client à la fin de la file s'il y a moins de trois clients dans la file et sinon, partagé équitablement aux trois clients dans la file. On vérifie bien que les conditions (2.15) sont remplies :

$$\delta'(n,n) = \delta(n,n) = 1, \qquad \forall n < N,$$

et

$$\delta'(l,N) = \frac{1}{N}, \qquad \forall l = 1,\ldots,N.$$

Par contre, les probabilités de positionnement $\delta(l,N)$ ne satisfont pas les conditions (2.16) comme $\delta(3,3) = 1$ et $\delta(1,3) = \delta(2,3) = 0$:

$$\begin{cases} 2\delta(1,3)+ & \delta(2,3) & = 0 \neq 1 \\ \delta(2,3)+ & 2\delta(3,3) & = 2 \neq 1. \end{cases}$$

Supposons que les demandes de service sont i.i.d. de loi hyperexponentielle de paramètre θ et de moyenne 1. La Figure 2.13 montre la sensibilité du temps de séjour moyen au paramètre de la loi hyperexponentielle θ.

Si $\theta = 1$, la loi hyperexponentielle devient exponentielle et le temps de séjour moyen est égal à celui d'une file d'attente symétrique insensible de 3 places. Par contre, le temps de séjour moyen devient plus important lorsque la loi de demandes de service varie.

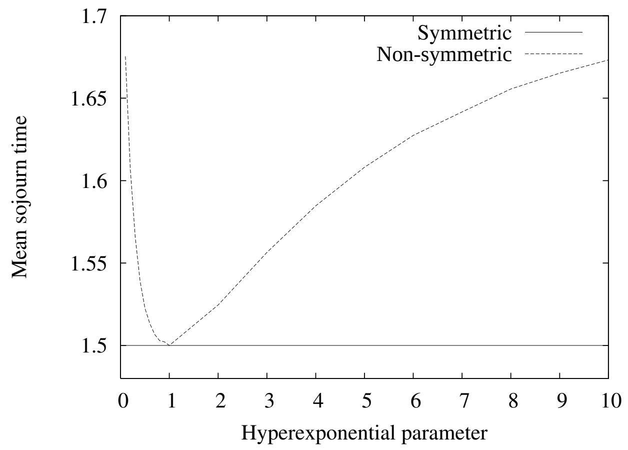

FIG. 2.13 – Temps de séjour moyen dans une file d'attente de 3 places.

2.5 Bilan des disciplines insensibles / sensibles

Dans cette partie, on récapitulera des disciplines insensibles / sensibles.

1. Les disciplines symétriques sont insensibles. Les disciplines PS et LIFO-préemptive font partie de ces disciplines symétriques.

   Les réseaux de Jackson et de Kelly de ces files d'attente sont aussi insensibles aux distributions de demandes de service. De plus, si les capacités de service, les taux d'arrivée et le routage dépendent du nombre de clients à chaque file, ces réseaux sont insensibles si et seulement s'ils satisfont une certaine propriété d'équilibre.

   À partir de ces réseaux, on peut construire de nombreuses nouvelles disciplines insensibles. Par exemple, la discipline de transfert avec 2 classes de clients dont le serveur sert alternativement l'une des deux classe est une des disciplines insensibles avec labels introduites dans la Section 2.4.1.

   Les réseaux de files symétriques restent insensibles même si l'on y introduit des permutations de clients à chaque file. Par contre, s'il y a permutations de clients de deux files de différentes distributions des demandes de service, ces réseaux sont sensibles, voir la Section 2.2.5. À la fin de la même section, on a aussi considéré des permutations entre les clients de deux files PS parallèles de même distribution des demandes de service. Par simulation, on a montré que le temps de séjour moyen de tous clients dans le système est insensible mais au contraire, le temps de séjour moyen échantillonné sur les clients entrants initialement à une file particulière est sensible à la distribution des demandes de service et aussi sensible au taux de permutation.

2. Dans la Section 2.4.2, une classe des disciplines nonsymétriques avec un nombre limite de places a été introduite et prouvée d'être insensible, mais non-décomposable, qui donne au même temps une illustration du fait qu'en général, l'insensibilité n'est pas équivalente à la décomposabilité.

3. En revanche, on rencontre aussi des disciplines sensibles, par exemple, la discipline FIFO, la discipline à priorité, la discipline limitée à un nombre fini de places que l'on a considérée dans la section précédente,...

Deuxième partie

Applications au partage de ressources informatiques

3

Métriques de débit

Sommaire

Le débit est un indicateur clef de la performance pour l'accès Internet. Il est souvent le résultat du multiplexage statistique d'un nombre aléatoire de flots partageant un lien commun. Les exemples typiques incluent les liens d'accès dans un réseau sans fil, cablé ou d'Ethernet haut-débit. Dans certains autres cas comme l'accès DSL, les utilisateurs disposent de leur propre lien. À condition que le réseau n'est pas limitant, le débit est alors presque constant et déterminé par la vitesse du lien d'accès, variant de 512 kbit/s à 20 Mbit/s pour un accès DSL. Le développement d'accès optique, de très haute vitesse, de 40 Mbit/s à 1 Gbit/s, tend à augmenter les constraintes du débit au réseau collecte. Le débit sera alors le résultat d'un multiplexage statistique.

Le choix d'une "bonne" métrique de débit, qui est critique pour dimmensionner les réseaux, n'est pas un problème facile sachant la nature aléatoire du trafic. Considérons par exemple deux tranferts indépendants d'un même fichier dont le premier est complété au taux 1 Mbit/s et le deuxième au taux 3 Mbit/s. D'une part, il est naturel à dire que le débit moyen est égal à 2 Mbit/s. D'autre part, le calcul basé sur la durée moyenne de transfert donne 1.5 Mbit/s, qui correspond au débit moyen *harmonique* ; ceci est aussi le débit d'un flot virtuel qui comprendrait le transfert consécutif des deux fichiers.

Dans cette partie, on définira formellement deux métriques de débit, à savoir le débit *échantillonné par flot* et celui *échantillonné en temps*. Les définitions ne sont pas restreintes au débit moyen mais appliquées à la distribution du débit. On donnera quelques propriétés génériques, par exemple les expressions asymptotiques dans les cas limites de flots de taille infiniment petite et infiniment large. On montrera que le débit échantillonné en temps peut être interprété comme le débit instantané *pondéré* par le nombre de flots, ce qui fournit un moyen pratique pour l'évaluer et le mesurer. Ceci explique pourquoi le débit moyen échantillonné en temps est plus souvent utilisé, voir [BFBP$^+$01, BMPV06, DPR04, HLN97, LB03].

Les notions de débits moyens échantillonnés par flot et en temps sont considérées dans [KK02, LvdBB04]. Il est notamment observé que le premier est dure à évaluer même dans le cas le plus simple d'un lien unique partagé équitablement, mais peut être approximé par le débit instantané moyen qui dispose de l'expression explicite [BK00, KK02, LvdBB04]. On verra que le débit instantané souffre en fait un biais d'échantillonnage : il dépend de la proportion p du trafic échantillonné. Par exemple, pour un lien partagé équitablement, le débit instantané moyen coïncide avec le débit moyen échantillonné par flot pour $p = 1$ et avec le débit moyen échantillonné en temps pour $p \to 0$. D'ailleurs, le débit instantané moyen pondéré par le nombre de flots coïncide avec le débit moyen échantillonné en temps pour toute proportion du trafic échantillonné.

La contribution clef de cette partie est d'étendre les notions d'échantillonnage par flot et en temps à la distribution du débit. Il est particulièrement approprié pour les flux adaptatifs qui requièrent un débit minimal. Dans cette section, on se focalise sur le trafic élastique où le taux de service de flots s'ajuste pour remplir la bande passante disponible. Mais les résultats peuvent être utilisés comme une approximation conservative pour un scénario plus réaliste où les trafics élastiques et adaptatifs sont multiplexés [BP04]. Il est clair que le débit moyen n'est pas une bonne métrique pour évaluer la qualité de flots adaptatifs. Des résultats asymptotiques sur la durée et le débit de flot, voir [BBNnQ05, BvOZ05, GRZ04], ne suffisent pas non plus à ce but. Ce qui est vraiment nécessaire est la distribution du débit, comme décrit dans la suite.

3.1 Modèle de trafic

Considérons un flot arbitraie de taille potentiellement infinie initié à l'instant 0. On considère ce flot comme un fluide et on note $\varphi(t)$ son débit instantané à l'instant t. Ce processus stochastique n'est pas stationnaire à cause de l'impact du flot étiqueté sur l'état du système. Supposons que ce processus atteint son état stationnaire caractérisé par une certaine variable aléatoire φ^{perm} correspondante à la distribution du débit d'un flot *permanent*. Supposons aussi que le processus $\varphi(t)$ prend des valeurs positives dans un certain ensemble discret borné $\mathcal{X}$ et qu'il est régulier, continu à droite avec une limite à gauche et ergodique, c.à.d. que

$$\lim_{T\to\infty} \frac{1}{T}\int_0^T \mathbb{I}(\varphi(t) = x) = \mathbb{P}(\varphi^{perm} = x) \quad \text{p.s.}, \qquad \forall x \in \mathcal{X}. \tag{3.1}$$

Notons $\varphi^{virt} = \varphi(0)$ le débit d'un flot virtuel de taille nulle.

Si le flot étiqueté est de taille s, sa durée $D(s)$ est une variable aléatoire définie par l'égalité

$$s = \int_0^{D(s)} \varphi(t)dt.$$

Soit T_0 le premier instant t tel que $\varphi(t) \neq \varphi(0)$. On a évidemment $\varphi(t) = \varphi(0)$ pour tout $t < T_0$ et

$$D(s) = \frac{s}{\varphi^{virt}} \quad \text{si } s < \varphi^{virt} T_0.$$

Comme $\varphi^{virt} T_0 > 0$ presque sûrement, on obtient

$$\lim_{s \to 0} \frac{s}{D(s)} = \varphi^{virt} \quad \text{p.s.} \tag{3.2}$$

Ensuite, en utilisant le fait que l'ensemle $\mathcal{X}$ est borné, on a

$$\lim_{s \to \infty} D(s) = \infty \quad \text{p.s.} \tag{3.3}$$

Alors par l'ergodicité,

$$\lim_{s \to \infty} \frac{s}{D(s)} = \mathbb{E}[\varphi^{virt}] \quad \text{p.s.} \tag{3.4}$$

Dans la suite, on pose S une taille aléatoire de flot et $D = D(S)$ la durée correspondante :

$$S = \int_0^D \varphi(t) dt.$$

3.2 Débit échantillonné par flot

3.2.1 Définition

Supposons maintenant que le flot étiqueté est de taille aléatoire S. Pour tout $x \in \mathcal{X}$, considérons la fraction moyenne de temps que ce flot ait débit x

$$\mathbb{E}\Big[\frac{1}{D} \int_0^D \mathbb{I}(\varphi(t) = x) dt\Big] \tag{3.5}$$

Cette fraction moyenne définit une distribution sur l'ensemble $\mathcal{X}$. On appelle la variable aléatoire correspondante X^{flow} le *débit échantillonné par flot*.

La définition ci-dessus peut s'étendre à des flots de toute taille fixée $s > 0$. La variable aléatoire correspondante $X^{flow}(s)$ est alors définie par

$$\mathbb{P}\Big(X^{flow}(s) = x\Big) = \mathbb{E}\Big[\frac{1}{D(s)} \int_0^{D(s)} \mathbb{I}(\varphi(t) = x) dt\Big]. \tag{3.6}$$

3.2.2 Propriété

Le débit moyen échantillonné par flot, noté γ^{flow}, est donné par

$$\begin{aligned}
\gamma^{flow} &= \mathbb{E}\Big[\mathcal{X}^{flow}\Big] \\
&= \sum_{x\in\mathcal{X}} x\mathbb{E}\Big[\frac{1}{D}\int_0^D \mathbb{I}(\varphi(t)=x)dt\Big] \\
&= \mathbb{E}\Bigg[\frac{1}{D}\int_0^D \big(\sum_{x\in\mathcal{X}} x\mathbb{I}(\varphi(t)=x\big)dt\Bigg] \\
&= \mathbb{E}\Big[\frac{1}{D}\int_0^D \varphi(t)dt\Big] \\
&= \mathbb{E}\Big[\frac{S}{D}\Big].
\end{aligned} \tag{3.7}$$

De même, on obtient le débit moyene échantilloné par flot d'un flot de taille s

$$\gamma^{flow}(s) = \mathbb{E}\big[X^{flow}(s)\big] = \mathbb{E}\Big[\frac{s}{D(s)}\Big]. \tag{3.8}$$

On énonce ci-dessous des résultats pour les cas limites de flots de taille infiniment petite et infiniment large.

Théorème 3.1 *Le débit échantillonné par flot $X^{flow}(s)$ tend en distribution vers φ^{virt} lorsque s tend vers 0 et vers φ^{perm} lorsque s tend vers ∞.*

Preuve. Rappelons que

$$\mathbb{P}\Big(X^{flow}(s)=x\Big) = \mathbb{E}\Bigg[\frac{1}{D(s)}\int_0^{D(s)} \mathbb{I}(\varphi(t)=x)dt\Bigg].$$

Comme dans la Section 3.2.1, posons T_0 le premier instant t tel que $\varphi(t)\neq\varphi(0)$. On a

$$\frac{1}{D(s)}\int_0^{D(s)} \mathbb{I}(\varphi(t)=x)dt = \mathbb{I}(\varphi^{virt}=x) \quad \text{si } s<\varphi^{virt}T_0.$$

Comme $\varphi^{virt}T_0>0$ presque sûrement, on obtient

$$\lim_{s\to 0}\frac{1}{D(s)}\int_0^{D(s)} \mathbb{I}(\varphi(t)=x)dt = \mathbb{I}\Big(\varphi^{virt}=x\Big) \quad \text{p.s.}$$

et

$$\lim_{s\to 0}\mathbb{P}\Big(X^{flow}(s)=x\Big) = \mathbb{P}\Big(\varphi^{virt}=x\Big).$$

L'autre limite est déduite par l'ergodicité en utilisant (3.1) et (3.3) :

$$\lim_{s\to\infty} \frac{1}{D(s)} \int_0^{D(s)} \mathbb{I}(\varphi(t) = x)dt = \mathbb{P}\Big(\varphi^{perm} = x\Big) \quad \text{p.s.}$$

□

3.3 Débit échantillonné en temps

3.3.1 Définition

Supposons que le flot étiqueté est de taille aléatoire S. Pour tout $x \in \mathcal{X}$, considérons le rapport de la durée moyenne au débit x à la durée moyenne de flot :

$$\frac{\mathbb{E}\Bigg[\int_0^D \mathbb{I}(\varphi(t) = x)dt\Bigg]}{\mathbb{E}[D]}. \tag{3.9}$$

Ce rapport définit une distribution sur l'ensemble $\mathcal{X}$ sachant que la durée moyenne de flot est finie. On appelle la variable aléatoire correspondante X^{time} le *débit échantillonné en temps*.

Cette définition s'étend à toute taille fixée de flot $s > 0$. La variable aléatoire correspondante $X^{time}(s)$ satisfait pour tout $x \in \mathcal{X}$

$$\mathbb{P}\Big(X^{time}(s) = x\Big) = \frac{\mathbb{E}\Bigg[\int_0^{D(s)} \mathbb{I}(\varphi(t) = x)dt\Bigg]}{\mathbb{E}\big[D(s)\big]}. \tag{3.10}$$

Afin d'illustrer la différence avec l'autre métrique de débit, considérons I flows de taille s, de durées respectives $d_1, \ldots, d_I$ et de proportions de temps au débit x : $f_1, \ldots, f_I$. Pour la métrique échantillonnée par flot, la probabilité qu'un flot de taille s ait le débit x est estimée par la moyenne empirique

$$\frac{1}{I}\sum_{i=1}^{I} f_i.$$

Pour la métrique échantillonnée en temps, elle est estimée par la moyenne empirique *pondérée* par les durées de flots

$$\frac{\sum_{i=1}^I d_i f_i}{\sum_{i=1}^I d_i}.$$

Cette quantité correspond au rapport du *temps* que les flots ont débit x au temps total de durées de flots, d'où le terme *échantillonné en temps*. Remarquons que, sachant que les I échantillonages soient i.i.d., les deux expressions ci-dessus tendent vers (3.6) et (3.10) respectivement, lorsque I tend vers l'infini.

3.3.2 Propriétés

Le débit moyen échantillonné en temps γ^{time} est donné par

$$\begin{aligned}
\gamma^{time} &= \mathbb{E}\Big[\mathcal{X}^{time}\Big] \\
&= \sum_{x\in\mathcal{X}} x\frac{1}{\mathbb{E}[D]}\mathbb{E}\Big[\int_0^D \mathbb{I}(\varphi(t)=x)dt\Big] \\
&= \frac{1}{\mathbb{E}[D]}\mathbb{E}\Bigg[\int_0^D \big(\sum_{x\in\mathcal{X}} x\mathbb{I}(\varphi(t)=x\big)dt\Bigg] \\
&= \frac{1}{\mathbb{E}[D]}\mathbb{E}\Big[\int_0^D \varphi(t)dt\Big] \\
&= \frac{\mathbb{E}[S]}{\mathbb{E}[D]}. \qquad (3.11)
\end{aligned}$$

De même, on a le débit moyen échantillonné en temps d'un flot de taille s

$$\gamma^{time}(s) = \mathbb{E}\Big[X^{time}(s)\Big] = \frac{s}{\mathbb{E}[D(s)]}. \qquad (3.12)$$

Maintenant, afin de montrer la différence avec l'autre métrique, reprenons I flots de taille s et de durées respectives $d_1,\ldots,d_I$. Pour la métrique échantillonnée par flot, le débit moyen est estimé par la moyenne *arithmétique* empirique

$$\frac{1}{I}\sum_{i=1}^{I}\frac{s}{d_i},$$

tandis que pour la métrique échantillonnée en temps, il est estimé par la moyenne *harmonique* empirique

$$\left(\frac{1}{I}\sum_{i=1}^{I}\frac{d_i}{s}\right)^{-1}.$$

Sachant que les I échantillonnages soient i.i.d., ces expressions tendent vers (3.8) et (3.12) respectivement, lorsque I tend vers l'infini.

On a le résultat suivant.

Théorème 3.2 *Pour toute taille de flot, le débit moyen échantillonné en temps est inférieur au débit moyen échantillonné par flot :*

$$\gamma^{time}(s) \leq \gamma^{flow}(s), \qquad \forall s>0.$$

Preuve. La démonstraition se déduit de l'inégalité de convexité

$$\mathbb{E}[D(s)]\mathbb{E}\left[\frac{1}{D(s)}\right] \geq 1.$$

□

Pour les cas limites de flots de taille infiniment petite et infiniment large, on a les résultats suivants.

Théorème 3.3 *Le débit échantillonné en temps $X^{time}(s)$ tend en distribution vers φ^{perm} lorsque s tend vers l'infini. De plus,*

$$\lim_{s\to 0} \mathbb{P}\Big(X^{time}(s) = x\Big) = \frac{\mathbb{P}\Big(\varphi^{virt} = x\Big)}{x\mathbb{E}\Big[1/\varphi^{virt}\Big]}.$$

Preuve. Rappelons que

$$\mathbb{P}\Big(X^{time}(s) = x\Big) = \frac{\mathbb{E}\Bigg[\int_0^{D(s)} \mathbb{I}(\varphi(t) = x)dt\Bigg]}{\mathbb{E}[D(s)]}.$$

Par l'ergodicité, il est déduit de (3.1), (3.3) et (3.4) que

$$\lim_{s\to\infty} \frac{1}{s}\int_0^{D(s)} \mathbb{I}(\varphi(t) = x)dt = \frac{\mathbb{P}\Big(\varphi^{perm} = x\Big)}{\mathbb{E}\Big[\varphi^{perm}\Big]} \quad \text{p.s.}$$

et

$$\lim_{s\to\infty} \frac{\mathbb{E}[D(s)]}{s} = \frac{1}{\mathbb{E}[\varphi^{perm}]},$$

alors

$$\lim_{s\to\infty} \frac{1}{\mathbb{E}[D(s)]}\int_0^{D(s)} \mathbb{I}(\varphi(t) = x)dt = \mathbb{P}\Big(\varphi^{perm} = x\Big) \quad \text{p.s.}$$

Pour l'autre limite, on utilise encore une fois le fait que $\varphi(t) = \varphi^{virt}$ pour tout $t < T_0$ de sorte que

$$\frac{1}{D(s)}\int_0^{D(s)} \mathbb{I}(\varphi(t) = x)dt = \mathbb{I}(\varphi(t) = x) \quad \text{si } s < \phi^{virt}T_0.$$

En utilisant (3.2), on obtient

$$\lim_{s\to 0} \frac{1}{s}\int_0^{D(s)} \mathbb{I}(\varphi(t) = x)dt = \frac{1}{x}\mathbb{I}\Big(\varphi^{virt} = x\Big) \quad \text{p.s.}$$

et

$$\lim_{s\to 0} \frac{D(s)}{s} = \frac{1}{\varphi^{virt}} \quad \text{p.s.}$$

alors

$$\lim_{s\to 0} \frac{\mathbb{E}\Bigg[\int_0^{D(s)} \mathbb{I}(\varphi(t) = x)dt\Bigg]}{\mathbb{E}[D(s)]} = \frac{\mathbb{P}\Big(\varphi^{virt} = x\Big)}{x\mathbb{E}\Big[1/\varphi^{virt}\Big]}.$$

□

3.4 Interprétation du débit instantané

En pratique, les statistiques de débit sont beaucoup plus faciles à estimer à travers le débit instantané mesuré sur un certain intervalle de temps choisi. Considérons un flux avec taux d'arrivée λ de flots du même débit instantané. Comme dans la seciton précédente, on note S leur taille aléatoire et D leur durée aléatoire. On suppose que le système est dans son état stationnaire et on note n le nombre de flots actifs dans le flux considéré et φ leur débit instantané, bien défini si $n \geq 1$. Par la loi de Little [BB03], on a

$$\mathbb{E}[n] = \lambda \mathbb{E}[D] \quad \text{et} \quad \mathbb{E}\Big[n \mathbb{I}(\varphi = x)\Big] = \lambda \mathbb{E}\left[\int_0^{D(s)} \mathbb{I}(\varphi(t) = x)\right], \tag{3.13}$$

alors

$$\mathbb{P}\Big(X^{time} = x\Big) = \frac{\mathbb{E}\big[n \mathbb{I}(\varphi = x)\big]}{\mathbb{E}[n]}. \tag{3.14}$$

Par conséquence, le débit échantillonné en temps peut être interprété comme le débit instantané pour une nouvelle mesure de probabilité où chaque événement est pondéré par le nombre de flots. Cela fournit un moyen utile pour évaluer et mesurer ce débit échantillonné en temps.

D'après (3.14), on a

$$\gamma^{time} = \mathbb{E}\Big[X^{time}\Big] = \frac{\mathbb{E}[n\varphi]}{\mathbb{E}[n]}. \tag{3.15}$$

En utilisant (3.13), on obtient

$$\begin{aligned} \mathbb{E}[n\varphi] &= \mathbb{E}\Big[n \sum_{x \in \mathcal{X}} x \mathbb{I}(\varphi = x)\Big] \\ &= \sum_{x \in \mathcal{X}} x \lambda \mathbb{E}\Big[\int_0^D \mathbb{I}(\varphi(t) = x) dt\Big] \\ &= \lambda \mathbb{E}\left[\int_0^D \varphi(t) dt\right] \\ &= \lambda \mathbb{E}[S], \end{aligned}$$

qui correspond à l'intensité du trafic, définie par le produit du taux d'arrivée et de la taille moyenne de flot. On déduit donc l'expression suivante pour le débit moyen échantillonné en temps

$$\gamma^{time} = \frac{\lambda \mathbb{E}[S]}{\mathbb{E}[n]}. \tag{3.16}$$

Notons qu'en fait, cette expression est déduite directement de (3.11) et de la loi de Little.

En général, le débit instantané est utilisé comme une approximation pour le débit échantillonné par flot [KK02, LvdBB04]. Comme il est défini si et seulement si $n \geq 1$, la distribution correspondante est

$$\mathbb{P}\Big(X^{inst} = x\Big) = \mathbb{P}(\varphi = x \mid n \geq 1), \qquad \forall x \in \mathcal{X}.$$

Dans la suite, on verra que cette métrique de débit est biaisée : elle dépend de la proportion de trafic représentée par la classe de flots considérée.

3.5 Exemples

3.5.1 Partage équitable

Considérons un lien de capacité unitaire dont les flots arrivent suivant un processus de Poisson d'intensité λ et sont de tailles i.i.d. caractérisées par la variable aléatoire S. Le partage est supposé à être équitable de telle sorte que le débit instantané φ de chaque flot est égal à $1/n$ lorsqu'il y a n flots, avec $n \geq 1$. En particulier, le débit prend des valeurs dans l'ensemble discret borné $\mathcal{X} = \{1, 1/2, 1/3, \ldots\}$. Le système correspondant est une file d'attente PS simple. Sachant que le taux de charge $\rho = \lambda \mathbb{E}[S]$ est inférieur à 1, la file est stable, insensible à la distribution de tailles de flot et possède la distribution stationnaire

$$\pi(n) = (1-\rho)\rho^n.$$

En présence d'un flot permanent, la distribution staionnaire du nombre n d'autres flots devient

$$\pi'(n) = (n+1)(1-\rho)^2\rho^n.$$

On déduit les expressions suivantes pour les distributions du débit instantané, du débit d'un flot virtuel de taille nulle et du débit d'un flot permanent, où $x = 1/k, k \geq 1$, désigne un élément quelconque de l'ensemble $\mathcal{X}$:

$$\begin{aligned}
\mathbb{P}(\varphi = x \mid k \geq 1) &= \frac{\pi(k)}{1-\pi(0)} = (1-\rho)\rho^k, \\
\mathbb{P}\Big(\varphi^{virt} = x\Big) &= \pi(k-1) = (1-\rho)\rho^{k-1}, \\
\mathbb{P}\Big(\varphi^{perm} = x\Big) &= \pi'(k-1) = k(1-\rho)^2\rho^{k-1}.
\end{aligned}$$

En particulier, on a

$$\mathbb{E}\Big[\varphi^{virt}\Big] = -(1-\rho)\frac{ln(1-\rho)}{\rho}, \tag{3.17}$$

et

$$\mathbb{E}\Big[\varphi^{perm}\Big] = \mathbb{E}\Big[1/\varphi^{virt}\Big]^{-1} = 1-\rho \tag{3.18}$$

Débits moyens

Il s'avère que le débit moyen échantillonné par flot n'a pas d'expression simple, même dans le cas le plus simple d'une distribution exponentielle de taille de flot. On n'a des expressions explicites que dans les cas limites de flots de taille infiniment petite et infiniment large. Il se déduit du Théoème 3.1, (3.17) et (3.18) que

$$\lim_{s\to 0} \gamma^{flow}(s) = -(1-\rho)\frac{ln(1-\rho)}{\rho},$$

et

$$\lim_{s\to\infty} \gamma^{flow}(s) = 1-\rho.$$

Le débit moyen échantillonné en temps est déduit directement de (3.16)

$$\gamma^{time} = 1-\rho.$$

Ensuite, il se déduit du Théorème 3.3, de (3.17) et de (3.18) que cela est aussi le débit moyen échantillonné en temps pour les flots de taille infiniment petite et infiniment large. Cette expression est en fait valable pour toute taille de flot. Ceci est déduit de (3.12) et du fait que le temps de séjour moyen d'un client dans un file d'attente PS est proportionnel à sa demande de service [Kle75].

Figure 3.1 compare les deux métriques de débit pour flots de taille exponentielle de moyenne unitaire avec le taux de charge $\rho = 0.5$. Le débit moyen échantillonné en temps est donné par $1-\rho = 0.5$, le débit moyen échantillonné par flot est simulé sur 10000000 poins.

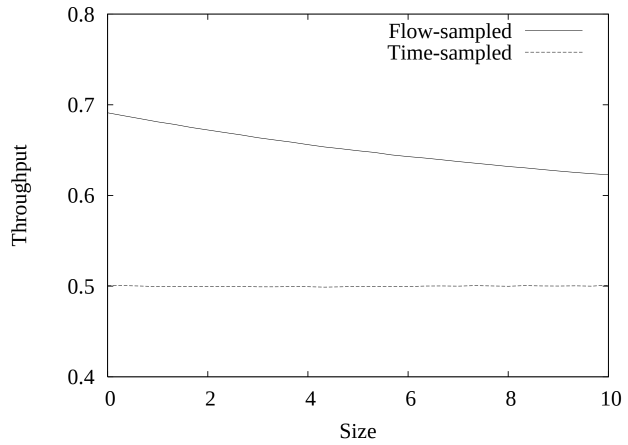

FIG. 3.1 – Débit moyen échantillonné en temps et par flot (taux de charge $\rho = 0.5$).

Distribution de débit

Comme dans les autres modèles, il n'y a pas d'expression explicite pour la distribution du débit échantillonné par flot sauf dans les cas limites avec flots de taille infiniment petite et infiniment large. D'après le Théorème 3.1, si les flots sont de taille infiniment petite, cette distribution est celle de φ^{virt} :

$$\mathbb{P}\Big(\varphi^{virt} = x\Big) = (1-\rho)\rho^{k-1}, \qquad \forall x = 1/k, k \geq 1,$$

et si les flots sont de taille infiniment large, il s'agit de la distribution de φ^{perm} :

$$\mathbb{P}\Big(\varphi^{perm} = x\Big) = k(1-\rho)^2\rho^{k-1}, \qquad \forall x = 1/k, k \geq 1.$$

D'ailleurs, pour le débit échantillonné en temps, on a, d'après (3.14), que pour tout $x = 1/k, k \geq 1$,

$$\mathbb{P}\Big(X^{time} = x\Big) = \frac{k\pi(k)}{\mathbb{E}[n]} = k(1-\rho)^2\rho^{k-1}.$$

D'après le Théorème 3.3, cette expression est aussi la distribution de débit échantillonné en temps pour les flots de taille infiniment petite et infiniment large. Et cette expression est en fait valable pour toute taille de flot.

Les résultats sont illustrés par la Figure 3.2 pour les flots de taille infiniment petite, appelés *des flots courts* dans la suite. Remarquons que la différence entre les deux métriques est significative. En particulier, la probabilité que le débit soit maximal est égal à $1 - \rho = 0.5$ pour le débit échantillonné par flot et seulement $(1-\rho)^2 = 0.25$ pour le débit échantillonné en temps.

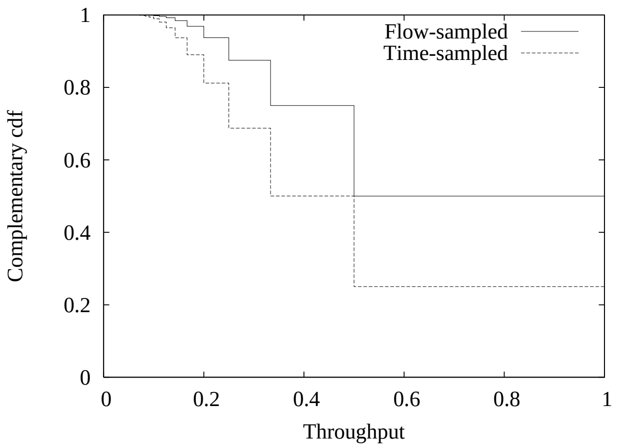

FIG. 3.2 – Distribution de débit échantillonné en temps et par flot (flots courts, taux de charge $\rho = 0.5$).

Biais d'échantillonnage

Maintenant, on illustrera le biais d'échantillonnage associé à la métrique de débit instantané. Supposons que les calculs sont basés sur un sous-ensemble de flots échantillonné au hasard avec la probabilité p. On appelle ce sous-ensemble la classe 1 et le sous-ensemble complémentaire la classe 2. Notons $\rho_1 = p\rho$ et $\rho_2 = (1-p)\rho$ les taux de charge correspondants. Alors la distribution stationnaire du nombre de flots de chaque classe est donnée par

$$\pi(n_1, n_2) = \binom{n_1 + n_2}{n_1}(1-\rho)\rho_1^{n_1}\rho_2^{n_2}.$$

Le débit instantané moyen évalué par la classe 1 est donné par

$$\gamma_1^{inst} = \mathbb{E}[\varphi \mid n_1 \geq 1] = \frac{\sum_{n_1 \geq 1, n_2} \frac{\pi(n_1,n_2)}{n_1+n_2}}{\sum_{n_1 \geq 1, n_2} \pi(n_1, n_2)} = -\frac{ln\left(1 - \frac{\rho_1}{1-\rho_2}\right)}{\frac{\rho_1}{1-\rho_2}}(1-\rho).$$

Cette quantité donne le débit moyen échantillonné par flot lorsque $p = 1$ et celui échantillonné en temps lorsque p tend vers 0.

Le biais disparaît lorsque la mesure de probabilité est pondérée par le nombre de flots. D'après (3.15), le débit moyen échantillonné en temps pour la classe 1 est donné par

$$\gamma_1^{time} = \frac{\mathbb{E}[n_1\varphi]}{\mathbb{E}[n_1]} = \frac{\sum_{n_1 \geq 1, n_2} \frac{n_1\pi(n_1,n_2)}{n_1+n_2}}{\sum_{n_1 \geq 1, n_2} n_1\pi(n_1, n_2)} = 1 - \rho.$$

De même, pour tout $x = 1/k, k \geq 1$,

$$\mathbb{P}\left(X_1^{time} = x\right) = \frac{\mathbb{E}\left[n_1\mathbb{I}(\varphi = x)\right]}{\mathbb{E}[n_1]} = \frac{\sum_{n_1+n_2=k} n_1\pi(n_1, n_2)}{\sum_{n_1,n_2} n_1\pi(n_1, n_2)} = k(1-\rho)^2\rho^{k-1}.$$

Cette expression est indépendante du choix de la classe 1. En choisissant des flots de taille s ($\pm ds$), on déduit que la distribution de débit échantillonné en temps est indépendante de la taille de flot.

3.5.2 Contraintes de capacité

En pratique, on rencontre souvent des cas où les flots n'ont pas de droit de se servir de la capacité totale du lien mais sont imposés par une limite de capacité. Dans le cas simple d'une capacité limite commune c, le débit instantané φ de chaque flot est égal à $min(c, 1/n)$ lorsqu'il y a n flots, avec $n \geq 1$. La distribution stationnaire devient

$$\pi(n) = \frac{\rho}{nc}\pi(n-1) \quad \text{si } nc \leq 1, \qquad \pi(n) = \rho\pi(n-1) \quad \text{sinon.}$$

En présence d'un flot permanent, on obtient

$$\pi'(n) = \frac{\rho}{nc}\pi'(n-1) \quad \text{si } (n+1)c \leq 1, \qquad \pi'(n) = \frac{n+1}{n}\rho\pi'(n-1) \quad \text{sinon.}$$

On en déduit les distributions de φ, φ^{virt} et φ^{perm} comme auparavant. Considérons par exemple $\rho = 0.5$ et $c = 0.5$. Dans ce cas, s'il y a un seul flot dans le système, ce flot ne reçoit que le débit $c = 0.5$ au lieu de la capacité totale 1 ; et si le nombre des flots est plus grand ou égal à 2, ces flots partagent équitablement la capacité totale du lien. Les distributions stationnaires π et π' peuvent être calculées explicitement :

n	0	1	2	3	4	...
$\pi(n)$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}\rho$	$\frac{1}{3}\rho^2$	$\frac{1}{3}\rho^3$	...
$\pi'(n)$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{3}{8}\rho$	$\frac{4}{8}\rho^2$	$\frac{5}{8}\rho^3$	...

La probabilité qu'un flot obtienne le débit instantané maximal est la probabilité qu'il y ait moins de deux autres flots dans le système :

$$\big(\pi(1)+\pi(2)\big)/\big(\pi(1)+\pi(2)+\pi(3)+\dots\big) = (\frac{1}{3}+\frac{1}{6})/(1-\frac{1}{3}) = 0.75,$$

La probabilité qu'un flot obtienne le débit maximal échantillonné en temps est

$$\pi'(0)+\pi'(1) = 0.5,$$

la probabilité qu'un flot court obtienne le débit maximal échantillonné par flot est

$$\pi(0)+\pi(1) = 0.67.$$

Bien que la différence entre les deux métriques est plus faible que sans contrainte de capacité, elle est toujours importante pour les flots courts. Figures 3.3 et 3.4 sont les analogues de Figures 3.1 et 3.2 pour ce cas où $c = 0.5$.

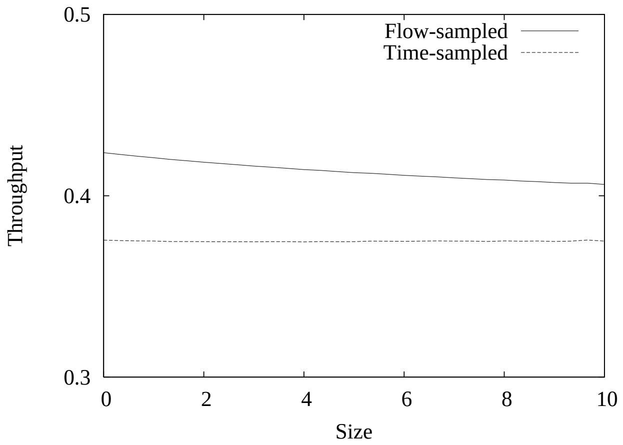

FIG. 3.3 – Débit moyen échantillonné en temps et par flot (taux de charge $\rho = 0.5$, capacité limite $c = 0.5$).

3.5.3 Partage inéquitable

Considérons enfin un lien de capacité unitaire sous un partage inéquitable. En particulier, on considère deux classes de flots dont le débit d'un flot de la classe 1 est égal à w_1/w_2 fois celui d'un flot concurrent de la classe 2, pour certains poids w_1 et w_2. Le modèle correspondant est la file PS discriminatoire [FMI80].

On n'a pas d'expression simple pour la distribution stationnaire, même pour les flots de tailles exponentielles. Figures 3.5 et 3.6 sont les analogues de Figures 3.1 et 3.2 pour une fraction de

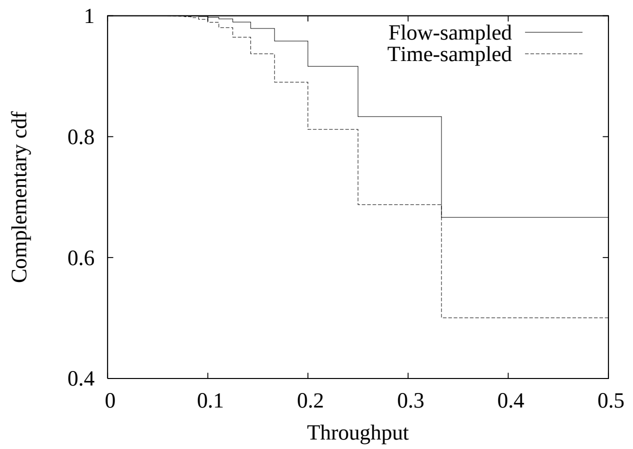

FIG. 3.4 – Distribution de débit échantillonné en temps et par flot (flots courts, taux de charge $\rho = 0.5$ et capacité limite $c = 0.5$).

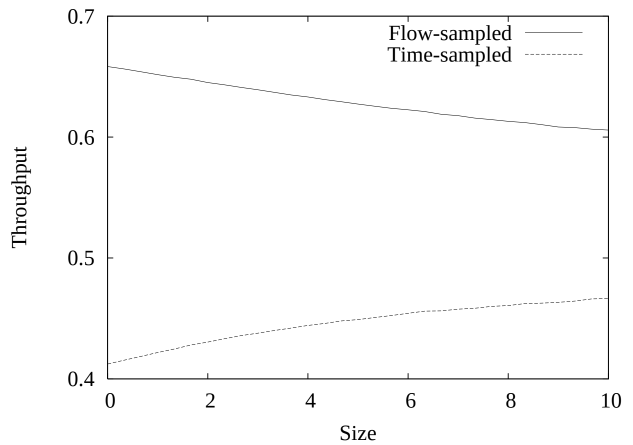

FIG. 3.5 – Débit moyen échantillonné en temps et par flot (flots de faible priorité, taux de charge $\rho = 0.5$).

poids $w_1/w_2 = 2$. Remarquons que les débits échantillonnés par flot et en temps dépendent tous les deux de la taille du flot étiqueté et la différence entre les deux métriques est significative.

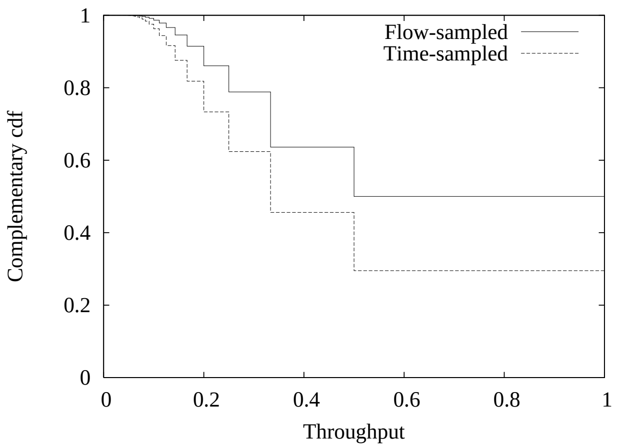

FIG. 3.6 – Distribution de débit échantillonné en temps et par flot (flots courts de faible priorité, taux de charge $\rho = 0.5$).

4

Équilibrage de sources

Sommaire

On considère dans ce chapitre un ensemble de sources hétérogènes partageant un lien commun. Ce modèle s'avère très utile en dimensionnement de réseaux [HLN97, BK00, BOR03]. Ceci est l'analogue du modèle d'Engset introduit en 1915 pour le trafic téléphonique [Eng]. La formule d'Engset, qui relie la probabilité de blocage des appels à l'intensité du trafic et au nombre de lignes téléphoniques, a été ensuite prolongée pour les sources de trafic hétérogènes [Coh79]. Il s'est avéré que les résultats correspondants sont trop compliqués pour tout but pratique. Ce résultat a motivé le travail de Dartois qui a montré en 1970 que l'équilibrage du trafic augmente la probabilité de blocage [Dar70]. Par conséquence, les réseaux téléphoniques peuvent être dimensionnés en supposant que l'on est dans le pire cas avec des sources de trafic homogènes.

En effet, on montrera que l'équilibrage de sources augmente les probabilités de blocage et diminue les débits pour les trafics élastiques.

Dans la première section, on montrera l'augmentation de probabilités de blocage et de la probabilité de saturation pour les modèles de circuit, y compris le modèle d'Engset. Et dans la deuxième section, on considérera les trafics élastiques. On montrera que l'équilibrage de sources diminue les débits et en présence d'un contrôle d'admission, l'équilibrage de sources augmente la probabilité de saturation ainsi que les probabilités de blocage dans certains cas.

4.1 Trafic de circuits

Dans cette partie, on considérera I sources partageant un lien commun dont chaque source demande un nombre fixe de circuits. S'il n'y a pas assez de circuits au moment où une source devient active, cette source est bloquée ; elle redevient alors inactive, et s'il n'y a assez de circuits libres, la source reçoit le nombre de circuits demandé et devient active.

Dans la première partie, on considère le modèle le plus simple où chaque source ne demande qu'un circuit. Ensuite, on considérera le modèle multidébit où chaque source demande un nombre différent de circuits.

4.1.1 Modèle d'Engset

Modèle

Considérons I sources partageant un lien commun de C circuits avec $C < I$. Chaque source demande 1 circuit si elle est active. La source i devient active après des intervalles inactifs exponentiels de durée moyenne $1/\nu_i$. À l'instant où la source i s'active, s'il n'y a pas de circuit libre, cette source est alors bloquée, sinon elle prend 1 circuit, devient active et quitte le système après un temps de séjour exponentiel de moyenne $1/\mu_i$. Notons $\rho_i = \nu_i/\mu_i$ le taux de charge correspondant.

Distribution stationnaire

Notons x_i l'état d'activité de la source i et x le vecteur $(x_1, \ldots, x_L)$:

$$x_i = \begin{cases} 1 & \text{si la source } i \text{ est active} \\ 0 & \text{sinon} \end{cases}$$

Notons $n = \sum_i x_i$ le nombre de sources actives.

L'espace d'état est alors l'ensemble $\mathcal{X} = \{x : n \leq C\}$.

Théorème 4.1 *Une mesure stationnaire de l'état du système x est donnée par*

$$\alpha(x) = \prod_{i=1}^{I} \rho_i^{x_i}, \qquad x \in \mathcal{X}.$$

La démonstration de ce théorème consiste à vérifier les équations de balance détaillée :

$$\alpha(x - e_i)\nu_i = \alpha(x)\mu_i, \qquad \forall x \in \mathcal{X} : x_i = 1,$$

où e_i est le vecteur unitaire de dimension I dont la $i^{\text{ème}}$ composante est égale à 1 et les autres sont nulles.

Corollaire 4.2 *Soit $\mathcal{X}(n)$ l'ensemble des états dans lesquels il y a n sources actives, $n \leq C$, alors une mesure stationnaire du nombre de sources actives est donnée par*

$$\pi(n) = \sum_{x \in \mathcal{X}(n)} \prod_{i=1}^{I} \rho_i^{x_i}, \qquad n \leq C,$$

Probabilité de blocage

Soit $\mathcal{X}_i \subset \mathcal{X}$ l'ensemble des états où la source i est inactive, c.à.d. $x_i = 0$, et $\mathcal{X}_i(n) \subset \mathcal{X}_i$ l'ensemble des états où la source i est inactive et il y a n sources actives. La probabilité de blocage des flots de la source i est calculée par la probabilité que C sources soient actives sachant que la source i est inactive :

$$B_i = \frac{\sum_{x \in \mathcal{X}_i(C)} \alpha(x)}{\sum_{x \in \mathcal{X}_i} \alpha(x)}.$$

La probabilité de blocage moyenne est donnée par

$$B = \frac{\sum_{i=1}^{I} \nu_i \sum_{x \in \mathcal{X}_i(C)} \alpha(x)}{\sum_{i=1}^{I} \nu_i \sum_{x \in \mathcal{X}_i} \alpha(x)}.$$

La probabilité de blocage d'une source virtuelle, appelée également la probabilité de saturation, est donnée par

$$B^{virt} = \frac{\pi(C)}{\sum \pi(n)}.$$

Équilibrage de sources

Dans cette section, on évaluera l'impact de l'équilibrage de sources. Supposons que les sources ont les mêmes statistiques de telle sorte que chaque source a le même taux de charge :

$$\rho = \frac{1}{I} \sum_{i=1}^{I} \rho_i.$$

Alors les mesures stationnaires deviennent :

$$\alpha'(x) = \rho^n, \qquad x \in \mathcal{X},$$

et

$$\pi'(n) = \binom{I}{n} \rho^n, \qquad n \leq C.$$

Théorème 4.3 *On a une comparaison entre la mesure originale π et la mesure à l'équilibrage π' :*

$$\frac{\pi(n)}{\pi'(n)} \geq \frac{\pi(n+1)}{\pi'(n+1)}.$$

pour tout $n = 0, 1, \ldots, I-1$.

Preuve. En remplaçant π et π' par leur expression, l'inégalité en question est équivalente à la suivante :

$$\frac{\sum_{x\in\mathcal{X}(n)}\prod_{i=1}^{I}(\frac{\rho_i}{\rho})^{x_i}}{\binom{I}{n}} \geq \frac{\sum_{x\in\mathcal{X}(n+1)}\prod_{i=1}^{I}(\frac{\rho_i}{\rho})^{x_i}}{\binom{I}{n+1}}, \quad \text{pour tout } n = 0, 1, \dots, I-1. \tag{4.1}$$

Posons

$$a_i = \frac{\rho_i}{\rho}$$

et

$$S(n) = \sum_{x\in\mathcal{X}(n)}\prod_{i=1}^{I} a_i^{x_i},$$

alors

$$S(1) = \sum_{i=1}^{I} a_i = I,$$

et l'inégalité (4.1) devient :

$$\frac{S(n)}{\binom{I}{n}} \geq \frac{S(n+1)}{\binom{I}{n+1}}, \quad \text{pour tout } n = 0, 1, \dots, I-1, \tag{4.2}$$

D'une part, on a :

$$S(1)S(n) = (n+1)S(n+1) + S'(n), \tag{4.3}$$

où

$$S'(n) = \sum_{i=1}^{I} a_i^2 \sum_{x\in\mathcal{X}(n-1),x_i=0}\prod_{j=1}^{I} a_j^{x_j}.$$

D'autre part, on a :

$$\begin{aligned}
(I-n)S'(n) &= \sum_{i\neq j}(a_i^2+a_j^2)\sum_{x\in\mathcal{X}(n-1),x_i=x_j=0}\prod_{k=1}^{I} a_k^{x_k}, \\
&\geq \sum_{i\neq j} 2a_ia_j \sum_{x\in\mathcal{X}(n-1),x_i=x_j=0}\prod_{k=1}^{I} a_k^{x_k}, \\
&= 2\binom{n+1}{2}S(n+1). \qquad (4.4)
\end{aligned}$$

Alors en remplaçant l'inégalité (4.4) dans l'expression (4.3), on obtient que :

$$\begin{aligned}
S(1)S(n) &\geq (n+1)S(n+1) + \frac{2}{I-n}\binom{n+1}{2}S(n+1), \\
&= I\frac{n+1}{I-n}S(n+1).
\end{aligned}$$

Comme $S(1) = I$, on en déduit l'inégalité (4.2).

□

On a ensuite un lemme très simple :

Lemme 4.4 *Soient $a_1, \ldots, a_k, b_1, \ldots, b_k$ et $t_1, \ldots, t_k$ des nombres positifs satisfaisant :*

$$\frac{a_1}{b_1} \geq \frac{a_2}{b_2} \geq \ldots \geq \frac{a_k}{b_k}, \quad \textit{et} \quad t_1 \leq t_2 \leq \ldots \leq t_k,$$

alors

$$\frac{\sum_{i=1}^k t_i a_i}{\sum_{i=1}^k t_i b_i} \leq \frac{\sum_{i=1}^k a_i}{\sum_{i=1}^k b_i}.$$

Preuve. L'inégalité est équivalente à :

$$\begin{aligned}
&\sum_{i=1}^k t_i a_i \sum_{i=1}^k b_i \leq \sum_{i=1}^k t_i b_i \sum_{i=1}^k a_i \\
&\Longleftrightarrow \sum_{i \neq j} (t_i - t_j) a_i b_j \leq 0 \\
&\Longleftrightarrow \sum_{i > j} (t_i - t_j)(a_i b_j - a_j b_i) \leq 0.
\end{aligned}$$

□

Théorème 4.5 *Le nombre de sources actives est plus grand en distribution lorsque les sources sont équilibrées :*

$$\mathbb{P}(n \geq m) \geq \mathbb{P}'(n \geq m), \qquad m = 0, \ldots, C,$$

où $\mathbb{P}$ est la probabilité correspondante à la distribution π et $\mathbb{P}'$ est celle correspondante à π'.

Preuve. En appliquant le Lemme 4.4 pour les nombres suivants :

$$\frac{\pi(0)}{\pi'(0)} \geq \frac{\pi(1)}{\pi'(1)} \geq \ldots \geq \frac{\pi(C)}{\pi'(C)} \quad \text{et} \quad t_0 = \ldots = t_{m-1} = 0, t_m = \ldots = t_C = 1,$$

on obtient que :

$$\frac{\sum_{n=m}^C \pi(n)}{\sum_{n=m}^C \pi'(n)} \leq \frac{\sum_{n=0}^C \pi(n)}{\sum_{n=0}^C \pi'(n)}.$$

□

Par l'équilibrage des sources, les probabilités de blocage deviennent :

$$B' = B'_i = \frac{\sum_{x \in \mathcal{X}_i(C)} \alpha'(x)}{\sum_{x \in \mathcal{X}_i} \alpha'(x)}, \qquad i = 1, \ldots, I,$$

et

$$B'^{virt} = \frac{\pi'(C)}{\sum \pi'(n)}.$$

Théorème 4.6 *La probabilité de blocage d'une source virtuelle, c.à.d. la probabilité de saturation, est augmentée lorsque les sources sont équilibrées.*

Preuve. D'après la Théorème 4.3, on a

$$\frac{\pi(0)}{\pi'(0)} \geq \frac{\pi(1)}{\pi'(1)} \geq \ldots \geq \frac{\pi(C)}{\pi'(C)}.$$

Alors en appliquant le Lemme 4.4, on obtient

$$\frac{\pi(0) + \pi(1) + \ldots + \pi(C)}{\pi'(0) + \pi'(1) + \ldots + \pi'(C)} \geq \frac{\pi(C)}{\pi'(C)}.$$

D'où, $\pi'(C) \geq \pi(C)$ et donc $B'^{virt} \geq B^{virt}$.

□

Corollaire 4.7 *La probabilité de blocage d'une source est augmentée lorsque les autres sources sont équilibrées.*

Preuve. La probabilité de blocage d'une source particulière est en fait égale à la probabilité de saturation du sous-système des $I-1$ autres sources. Alors comme l'équilibrage de sources de ce sous-système augmente sa probabilité de saturation, ceci augmente aussi la probabilité de blocage de la source considérée.

□

On a montré que l'équilibrage de toutes les sources augmente la probabilité de saturation. Dans la suite, on montrera que l'équilibrage d'un sous-ensemble de sources augmente aussi la probabilité de saturation. Tout d'abord, on suppose que $\rho_1 < (\rho_1+\rho_2)/2 = \rho < \rho_2$ et on montrera que si la moyenne ρ et les autres taux de charge $\rho_3, \ldots, \rho_I$ sont constants, la probabilité de saturation B^{virt} est est une fonction croissante de ρ_1. Une fois que cette monotonie est établie, on en déduit que l'équilibrage d'un sous-ensemble quelconque de sources augmente la probabilité de saturation.

Théorème 4.8 *Si les taux $\rho = (\rho_1 + \rho_2)/2$ et $\rho_3, \ldots, \rho_I$ sont constants, et si $\rho_1 < \rho < \rho_2$, la proportion $\pi(n-1)/\pi(n)$ est une fonction décroissante de ρ_1 pour tout $n \geq 1$.*

Preuve. On montrera ce théorème par l'étude de la dérivée de cette proportion $\pi(n-1)/\pi(n)$ par rapport à ρ_1. Posons $S(k)$ la somme de tous les produits de k termes d'entre $\rho_3, \ldots, \rho_I$:

$$S(k) = \sum_{3 \leq j_1 < j_2 < \ldots < j_k \leq I} \rho_{j_1} \ldots \rho_{j_k}, \qquad k \geq 1,$$

et par convention, $S(0) = 1$. Nous obtenons

$$\pi(n) = \Big(\rho_1\rho_2 S(n-2) + 2\rho S(n-1) + S(n)\Big), \qquad n \geq 2.$$

Et donc
$$\frac{d}{d\rho_1}\pi(n) = 2(\rho - \rho_1)S(n-2), \qquad n \geq 2.$$

Alors similairement au raisonnement dans la preuve du Théorème 4.3, on peut montrer que
$$\Big(\frac{d}{d\rho_1}\pi(n-1)\Big)\pi(n) \leq \Big(\frac{d}{d\rho_1}\pi(n)\Big)\pi(n-1), \qquad n \geq 3,$$
tant que la condition $\rho_1 < \rho < \rho_2$ est remplie. Donc la proportion $\pi(n-1)/\pi(n)$ est décroissante pour tout $n \geq 3$.

De plus, $\pi(0) = 1$ et $\pi(1) = 2\rho + S(1)$ sont toutes constantes et $\pi(2)$ est croissante alors évidemment la proportion $\pi(n-1)/\pi(n)$ est décroissante pour $n = 1, 2$ et donc pour tout $n \geq 1$.

□

Remarquons que ce résultat est plus fort que le Théorème 4.3.

Corollaire 4.9 *Si les taux $\rho = (\rho_1 + \rho_2)/2$ et $\rho_3, \ldots, \rho_I$ sont constants, et si $\rho_1 < \rho < \rho_2$, la probabilité de saturation B^{virt} est une fonction croissante de ρ_1.*

Preuve. Rappelons que
$$B^{virt} = \frac{\pi(C)}{\pi(0) + \pi(1) + \ldots + \pi(C)}.$$
Comme la proportion $\pi(n-1)/\pi(n)$ est décroissante pour tout n entre 1 et C, il en déduit que $\pi(n)/\pi(C)$ l'est aussi pour tout n entre 0 et C, et donc B^{virt} est croissante. Par conséquence, la probabilité de saturation B^{virt} est croissante.

□

Si $\rho_1 < \rho_2$, on peut remplacer le couple (ρ_1, ρ_2) par n'importe quel couple (ρ_1', ρ_2') tel que $\rho_1 < \rho_1' < \rho_2' < \rho_2$ et $\rho_1' + \rho_2' = \rho_1 + \rho_2 = 2\rho$ afin d'obtenir une probabilité de saturation plus importante. Autrement dit, la probabilité de saturation est augmentée si on remplace les taux de charge ρ_1, ρ_2 par deux nouveaux taux de la même somme mais d'une différence moins importante. Ceci implique que l'équilibrage d'un sous-ensemble quelconque de sources augmente la probabilité de saturation.

Théorème 4.10 *L'équilibrage d'un sous-ensemble quelconque de sources augmente la probabilité de saturation B^{virt}. Et par conséquence, l'équilibre d'un sous-ensemble de sources augmente la probabilité de blocage d'une source particulière en dehors de ce sous-ensemble.*

Remarquons que l'équilibrage de sources augmente la probabilité de saturation et la probabilité de blocage d'une source particulière mais il peut augmenter et aussi diminuer la probabilité de blocage moyen B. Considérons le cas le plus simple où $C = 1$ et $I = 2$:
$$B = \frac{\nu_1\rho_1 + \nu_2\rho_2}{\nu_1 + \nu_2 + \nu_1\rho_1 + \nu_2\rho_2},$$
et
$$B' = B_i' = \frac{\rho}{1+\rho}, \qquad i = 1, 2.$$

Alors $B < B'$ si et seulement si $(\nu_1 - \nu_2)(\rho_2 - \rho_1) < 0$, ce qui n'est pas toujours vrai. En particulier, si dans le système original, tous les flots demandent des temps de séjour d'une même moyenne $1/\mu$, l'équilibrage des sources augmente la probabilité de blocage moyenne.

On s'aperçoit que l'équilibrage d'un sous-ensemble de sources augmente la probabilité de blocage de toutes les sources en dehors de ce sous-ensemble mais peut diminuer la probabilité de blocage moyenne. Ce phénomène ne peut être expliqué que par la diminution de la probabilité de blocage d'une ou plusieurs sources dans le sous-ensemble équilibré. En effet, l'équilibrage de deux sources 1 et 2 dans l'exemple précédent diminue la probabilité de blocage de l'une des deux sources car la probabilité de blocage $B_1' = B_2' = \rho/(1+\rho)$ du système équilibré se situe toujours entre les deux probabilités de blocage $B_1 = \rho_1/(1+\rho_1)$ et $B_2 = \rho_2/(1+\rho_2)$ du système original.

4.1.2 Modèle multidébit

Dans ce modèle, les sources partagent un lien commun de C circuits, la source i demande un nombre c_i de circuits, $c_i \geq 1, i = 1, \ldots, I$. Posons ρ_i le taux de charge virtuel correspondant à la source i.

L'espace d'états du système devient compliqué

$$\mathcal{X} = \{x : C(x) \leq C\},$$

avec $C(x) = \sum_{i=1}^{I} x_i c_i$ le nombre de circuits occupés.

Théorème 4.11 *Une mesure stationnaire de l'état du système est à forme produit et est donnée par*

$$\alpha(x) = \prod_{i=1}^{I} \rho_i^{x_i}, \qquad x \in \mathcal{X}.$$

Corollaire 4.12 *Une mesure stationnaire du nombre de sources actives est à forme produit et est donnée par*

$$\pi(n) = \sum_{x \in \mathcal{X}(n)} \prod_{i=1}^{I} \rho_i^{x_i}, \qquad n \leq I,$$

où $\mathcal{X}(n)$ désigne l'ensemble des états où le nombre de sources actives est égal à n.

La probabilité de blocage d'une source virtuelle qui demande c circuits est donnée par

$$B_c^{virt} = \frac{\sum_{x : C - c < C(x) \leq C} \alpha(x)}{\sum_{x \in \mathcal{X}} \alpha(x)}.$$

Considérons un exemple simple de $I = 2$ sources avec $c_1 < c_2 < C < c_1 + c_2$.

L'espace d'états est $\mathcal{X} = \{(0,0),(1,0),(0,1)\}$ et la distribution stationnaire α devient

$$\alpha(0,0) = 1, \quad \alpha(1,0) = \rho_1, \quad \text{et} \quad \alpha(0,1) = \rho_2.$$

Considérons une source virtuelle qui demande c circuits avec $c_1+c < C < c_2+c$. La probabilité de blocage de cette source est

$$B_c^{virt} = \frac{\rho_2}{1+\rho_1+\rho_2}.$$

On s'aperçoit que l'équilibrage de sources augmente cette probabilité de blocage si $\rho_2 < \rho_1$ et la diminue sinon. Alors en général, l'équilibrage de sources peut augmenter et aussi diminuer la probabilité de blocage d'une source virtuelle et de même pour la probabilité de blocage d'une source particulière.

Et il suffit de reprendre l'exemple considéré dans ce modèle d'Engset afin de conclure que l'équilibrage de sources peut augmenter et aussi diminuer la probabilité de blocage moyenne.

Considérons maintenant le cas particulier où les sources forment des classes de sources de même demande de circuits. L'équilibrage d'une classe augmente-il les probabilités de blocage ? Pour la probabilité de blocage moyenne, la réponse est non parce que même dans le modèle d'Engset avec une seule classe de sources, l'équilibrage de sources peut augmenter et aussi diminuer cette probabilité de blocage moyenne. Dans la suite, on montrera que la réponse est aussi non pour la probabilité de blocage d'une source particulière et d'une source virtuelle.

Remarquons qu'il suffit de travailler avec l'équilibrage d'une classe de 2 sources. Alors considérons l'ensemble de I sources qui demandent respectivement $c_1, \ldots, c_I$. Ces sources partagent un lien commun de C circuits :

$$C(x) = \sum_{i=1}^{I} c_i x_i \leq C.$$

Supposons que $c_1 = c_2$, on considérera l'impact de l'équilibrage des deux sources correspondantes 1 et 2.

La probabilité de blocage d'une source virtuelle qui demande c circuits est donnée par

$$B_c^{virt} = \frac{f(C)}{g(C)},$$

où les fonctions f et g sont définies par

$$f(t) = \sum_{t-c<C(x)\leq t} \alpha(x)$$

et

$$g(t) = \sum_{C(x)\leq t} \alpha(x).$$

Cette expression est réécrite comme suit

$$B_c^{virt} = \frac{\tilde{f}(C) + (\rho_1+\rho_2)\tilde{f}(C-c_1) + \rho_1\rho_2\tilde{f}(C-2c_1)}{\tilde{g}(C) + (\rho_1+\rho_2)\tilde{g}(C-c_1) + \rho_1\rho_2\tilde{g}(C-2c_1)},$$

où $\tilde{f}$ et $\tilde{g}$ sont deux fonctions définies de la même manière que f et g mais pour le sous-système de $I-2$ sources : $3, \ldots, I$.

À l'équilibrage des deux sources 1 et 2, cette probabilité de blocage devient

$$B_c'^{virt} = \frac{\tilde{f}(C) + 2\rho\tilde{f}(C-c_1) + \rho^2\tilde{f}(C-2c_1)}{\tilde{g}(C) + 2\rho\tilde{g}(C-c_1) + \rho^2\tilde{g}(C-2c_1)},$$

avec $\rho = (\rho_1 + \rho_2)/2$.

En utilisant le fait que pour tous nombres réels positifs a, a', b, b', la fraction $(a+a')/(b+b')$ se situe toujours entre les deux fractions a/b et a'/b', on peut montrer que $B_c^{virt} < B_c'^{virt}$ si et seulement si

$$\begin{aligned}
& \frac{\tilde{f}(C) + 2\rho\tilde{f}(C-c_1) + \rho_1\rho_2\tilde{f}(C-2c_1)}{\tilde{g}(C) + 2\rho\tilde{g}(C-c_1) + \rho_1\rho_2\tilde{g}(C-2c_1)} < \frac{\tilde{f}(C-2c_1)}{\tilde{g}(C-2c_1)} \\
\iff \quad & \frac{\tilde{f}(C) + 2\rho\tilde{f}(C-c_1)}{\tilde{g}(C) + 2\rho\tilde{g}(C-c_1)} < \frac{\tilde{f}(C-2c_1)}{\tilde{g}(C-2c_1)}, \qquad \forall\rho \\
\iff \quad & \begin{cases} \frac{\tilde{f}(C)}{\tilde{g}(C)} < \frac{\tilde{f}(C-2c_1)}{\tilde{g}(C-2c_1)} \\ \frac{\tilde{f}(C-c_1)}{\tilde{g}(C-c_1)} < \frac{\tilde{f}(C-2c_1)}{\tilde{g}(C-2c_1)} \end{cases}
\end{aligned}$$

Cette inégalité est équivalente à dire que pour le sous-système de $I-2$ sources, la probabilité de blocage d'une source virtuelle est augmentée si l'on diminue la capacité du lien. Considérons un sous-système particulier de 20 sources de taux de charge ρ_1 dont chacune demande 1 circuit et 2 sources de taux de charge ρ_2 dont chacune demande 2 circuits alors pour $c = 1$, on peut calculer par exemple :

$$\frac{\tilde{f}(20)}{\tilde{g}(20)} = \frac{\rho_2^2 + \binom{20}{10}\rho_1^{10}\rho_2 + \rho_1^{20}}{\rho_2^2 + \binom{2}{1}\rho_2\sum_{k=0}^{10}\binom{20}{k}\rho_1^k + \sum_{k=0}^{20}\binom{20}{k}\rho_1^k},$$

et

$$\frac{\tilde{f}(19)}{\tilde{g}(19)} = \frac{\binom{20}{9}\binom{2}{1}\rho_2\rho_1^9 + \binom{20}{19}\rho_1^{19}}{\binom{2}{1}\rho_2\sum_{k=0}^{9}\binom{20}{k}\rho_1^k + \sum_{k=0}^{19}\binom{20}{k}\rho_1^k}.$$

Faisons ρ_2 tendre vers ∞, on obtient

$$\frac{\tilde{f}(20)}{\tilde{g}(20)} = 1 > \frac{\tilde{f}(19)}{\tilde{g}(19)},$$

ce qui montre que diminuer la capacité du lien partagé n'augmente pas la probabilité de blocage d'une source virtuelle dans ce cas. En conclusion, dans le modèle de circuits avec des classes de sources de même demande de circuits, l'équilibrage de sources n'augmente pas toujours la probabilité de blocage d'une source virtuelle. De même, l'équilibrage de sources n'augmente pas toujours la probabilité de blocage d'une source particulière dans le système.

D'ailleurs, il existe des cas où l'équilibrage de sources augmente ces probabilités de blocage, par exemple dans le modèle d'Engset, alors on n'a pas de résultat similaire pour le modèle multi-débit.

4.1.3 Réseaux de circuits

Dans cette partie, on considérera l'impact de l'équilibrage de sources sur la probabilité de blocage d'une source particulière dans un réseau de circuits. On montrera que l'équilibrage de sources n'augmente pas toujours cette probabilité de blocage.

Considérons le modèle le plus simple d'un réseau linéaire de deux liens $1, 2$ de capacipités $C_1 = 2, C_2 = 1$ dont les sources $1, 2$ demandent les services du premier lien, la source 4 demande le deuxième lien et la source 3 demande un circuit de chaque lien (voir la Figure 4.1).

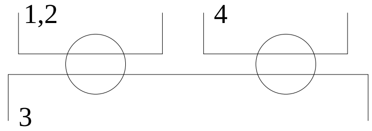

Fig. 4.1 – Un réseau linéaire.

L'espace d'état est alors l'ensemble $\mathcal{X} = \{x = (x_1, x_2, x_3, x_4) : x_1 + x_2 + x_3 \leq 2, x_3 + x_4 \leq 2\}$. La probabilité de blocage de la source 4 est donnée par

$$\begin{aligned} B_4 &= \frac{\rho_3(1 + \rho_1 + \rho_2)}{\rho_3(1 + \rho_1 + \rho_2) + 1 + \rho_1 + \rho_2 + \rho_1\rho_2} \\ &= \frac{\rho_3(1 + 2\rho)}{\rho_3(1 + 2\rho) + 1 + 2\rho + \rho_1\rho_2}, \end{aligned}$$

où $\rho = (\rho_1 + \rho_2)/2$. On s'aperçoit que dans ce cas, l'équilibrage de deux sources 1 et 2 diminue la probabilité de blocage de la source 4. Dans un réseau de circuits, l'équilibrage de sources peut augmenter et aussi diminuer la probabilité de blocage.

4.2 Trafic élastique

On considère un ensemble fini de sources qui partagent dynamiquement un lien commun. Ce lien peut être une partie d'un réseau de câble, DSL ou d'un réseau optique par exemple, ou encore le lien d'accès à un serveur de Web. Chaque source engendre des flots de données de taille aléatoire séparés par les intervalles d'inactivité de durée aléatoire. L'élasticité du trafic signifie que les intervalles de congestion où il y a un grand nombre de sources actives augmentent la durée des flots de données. D'autre part, la durée des intervalles d'inactivité est supposée être indépendante du niveau de congestion du lien : à la fin de chaque flot, la source entre dans un intervalle inactif de durée aléatoire avant d'engendrer un nouveau flot.

Dans cette section, on montre pour le trafic élastique l'analogue du résultat dans la section précédente. En particulier, on montre que l'équilibrage du trafic diminue les débits moyens et augemente la probabilité de blocage s'il y a un contrôle d'admission, où un nombre maximal de sources actives est imposé. Ainsi les réseaux de donnée peuvent être dimensionnés en prenant des sources de trafic homogènes, malgré l'hétérogénité observée en réalité.

4.2.1 Modèle

Considérons I sources partageant un lien commun de capacité C bit/s. La source i engendre des flots de taille moyenne σ_i séparés par les intervalles d'inactivité de durée moyenne $1/\nu_i$. Les tailles des flots et les durées des intervalles d'inactivité d'une source donnée possèdent des distributions arbitraires et peuvent être corrélées.

On note $\mu_i = C/\sigma_i$ le taux virtuel d'accomplissement des flots de la source i et $\rho_i = \nu_i/\mu_i$ la charge virtuelle correspondante.

On suppose qule partage du lien est parfaitement équitable, c.à.d. que chaque source reçoit exactement un débit C/n s'il y a n sources actives.

Distribution stationnaire

On note x_i l'état d'activité de la source i et x le vecteur $(x_1, \ldots, x_I)$:

$$x_i = \begin{cases} 1 & \text{si la source } i \text{ est active} \\ 0 & \text{sinon} \end{cases}$$

L'état du système x appartient à l'espace d'états $\mathcal{X} = \{0,1\}^I$.

Notons $n = \sum_{i=1}^{I} x_i$ le nombre de sources actives.

Théorème 4.13 *Le système est insensible, la distribution stationnaire de l'état du système x est la même que celle dans le cas où les tailles des flots et les durées des intervalles d'inactivité sont indépendantes et suivent les distributions exponentielles. Le processus décrivant l'état du système x est alors markovien, réversible sur l'espace d'états $\mathcal{X}$, de distribution stationnaire*

$$\alpha(x) = \alpha(0) n! \prod_{i=1}^{I} \rho_i^{x_i}, \quad x \in \mathcal{X}. \tag{4.5}$$

Dans ce modèle, l'insensibilité veut dire que la distribution stationnaire $\alpha(x)$ n'est pas changée même si les tailles des flots et les durées des intervalles d'inactivité d'une source donnée possèdent des distributions arbitraires et sont corrélées, *pourvu que* les charges ρ_i restent toujours inchangées. Remarquons qu'ultérieurement, l'équilibrage de sources changera les valeurs de ces charges et donc la distribution stationnaire α sera aussi changée par cet équilibrage.

Afin d'illustrer la corrélation entre les flots et les intervalles d'inactivité successifs, imaginons un routage simple de Kelly où la loi de la longueur de chaque intervalle peut dépendre de la phase de route. Par exemple, ce routage peut imposer à une certaine source un flot long, un intervalle d'inactivité de courte durée, un flot court et enfin un intervalle d'inactivité de durée longue avant de recommencer une nouvelle session. Alors on peut voir la corrélation entre les flots et les intervalles d'inactivité successifs : en particulier, après un long intervalle d'inactivité, il y aura un flot long avec une forte probabilité, et par contre, après un court intervalle d'inactivité, il y aura un flot court avec une forte probabilité.

Cette propriété d'insensibilité a été montrée dans les références [Ser99, BFBP+01] et l'expression de la distribution stationnaire est déduite des équations de balance détaillée

$$\alpha(x)\nu_i = \alpha(x+e_i)\frac{\mu_i}{n+1}, \quad \forall x \in \mathcal{X} : x_i = 0, \tag{4.6}$$

où e_i est le vecteur unitaire de dimension I dont la $i^{\text{ème}}$ composante est égale à 1 et les autres sont nulles.

Corollaire 4.14 *Soit $\mathcal{X}(n)$ l'ensemble des états dans lesquels il y a n sources actives, alors la distribution stationnaire du nombre de sources actives est donnée par :*

$$\pi(n) = \pi(0)n! \sum_{x\in\mathcal{X}(n)} \prod_{i=1}^{I} \rho_i^{x_i}, \quad n = 0, \ldots, I. \tag{4.7}$$

Débit échantillonné en temps

Dans cette partie, on est intéressé par le débit moyen échantillonné en temps (Section 3.3), défini par la fraction entre la taille moyenne et la durée moyenne des flots, et la distribution de ce débit échantillonné en temps.

Soit τ_i la durée moyenne des flots de la source i. Le débit moyen échantillonné en temps de cette source est alors

$$\gamma_i^{time} = \frac{\sigma_i}{\tau_i}. \tag{4.8}$$

Notons p_i la probabilité que la source i soit active. Par la loi de Little, le taux d'arrivée λ_i des flots de la source i satisfait

$$\tau_i = \frac{p_i}{\lambda_i}, \quad \frac{1}{\nu_i} = \frac{1-p_i}{\lambda_i}.$$

On en déduit que

$$\gamma_i^{time} = C\frac{\rho_i(1-p_i)}{p_i}. \tag{4.9}$$

Considérons maintenant le débit moyen échantillonné en temps de toutes les sources

$$\gamma^{time} = \frac{\sigma}{\tau},$$

où σ et τ sont respectivement la taille moyenne et la durée moyenne des flots :

$$\sigma = \frac{1}{\lambda}\sum_{i=1}^{I} \lambda_i\sigma_i, \quad \text{et} \quad \tau = \frac{1}{\lambda}\sum_{i=1}^{I} \lambda_i\tau_i \quad \text{avec } \lambda = \sum_{i=1}^{I} \lambda_i.$$

Soit p la probabilité qu'au moins une source soit active. En utilisant les équations de balance détaillée (4.6), on obtient

$$\begin{aligned}
\sum_{i=1}^{I} \lambda_i \sigma_i &= \sum_{i=1}^{I} (1-p_i)\nu_i \sigma_i, \\
&= \sum_{i=1}^{I} \sum_{x:x_i=0} \alpha(x)\nu_i\sigma_i, \\
&= \sum_{i=1}^{I} \sum_{x:x_i=0} \alpha(x+e_i)\frac{\mu_i}{n+1}\sigma_i, \\
&= C\sum_{i=1}^{I} \sum_{x:x_i=1} \frac{\alpha(x)}{n}, \\
&= Cp.
\end{aligned}$$

Or

$$\sum_{i=1}^{I} \lambda_i \tau_i = \sum_{i=1}^{I} p_i = \mathbb{E}[n].$$

On en déduit une expression simple pour le débit moyen échantillonné en temps de toutes les sources :

$$\gamma^{time} = C\frac{p}{\mathbb{E}[n]}.$$

D'ailleurs, d'après (3.14), la distribution du débit échantilloné en temps est donnée par

$$\mathbb{P}(X^{time} = d) = \frac{\mathbb{E}\big[n\mathbb{I}(\varphi = d)\big]}{\mathbb{E}[n]} = \frac{k\pi(k)}{\mathbb{E}[n]}, \qquad d = \frac{C}{k},$$

où φ est le débit instantané des flots. Cette expression est valable pour toute taille de flot.

Débit échantillonné par flot

On s'intéresse maintenant au débit échantillonné par flot dans ce modèle. Même si cette quantité n'admet pas d'expression simple, on peut étudier ses limites au cas de flots de taille infiniment petite et infiniment large, appelés respectivement des *flots courts* et des *flots longs*. D'après le Théorème 3.1, on a

$$\lim_{s\to 0} \gamma^{flow}(s) = E\big[\varphi^{virt}\big],$$

et

$$\lim_{s\to \infty} \gamma^{flow}(s) = E\big[\varphi^{perm}\big] = \gamma^{time} = C\frac{p}{E[n]},$$

où s désigne la taille d'un flot arbitraire, φ^{virt} et φ^{perm} désignent respectivement le débit d'un flot virtuel de taille nulle et celui d'un flot permanent.

Le débit échantillonné par flot des flots longs est égal au débit échantillonné en temps. Pour les flots courts, l'espérance $E[\varphi^{virt}]$ est donnée par

$$\begin{aligned} E[\varphi^{virt}] &= \frac{\sum_{n=0}^{I} \varphi^{virt}\pi(n)}{\sum_n \pi(n)} \\ &= \sum_{n=0}^{I} \frac{C}{n+1}\pi(n), \end{aligned}$$

alors les flots couts reçoivent le débit moyen échantillonné par flot :

$$\lim_{s\to 0} \gamma^{flow}(s) = C \sum_{n=0}^{I} \frac{1}{n+1}\pi(n). \tag{4.10}$$

De plus, pour la distribution de débit échantillonné par flot des flots courts, on a

$$\lim_{s\to 0} \mathbb{P}\Big(X^{flow}(s) = d\Big) = \mathbb{P}(\varphi^{virt} = d) = \pi(k-1), \qquad d = \frac{C}{k}.$$

Et la distribution pour les flots longs est celle du débit échantillonné en temps :

$$\lim_{s\to \infty} \mathbb{P}\Big(X^{flow}(s) = d\Big) = \lim_{s\to\infty} \mathbb{P}\Big(X^{time}(s) = d\Big) = \frac{k\pi(k)}{\mathbb{E}[n]}, \qquad d = \frac{C}{k}.$$

4.2.2 Équilibrage de sources

Dans cette section, on évaluera l'impact de l'équilibrage de sources. En particulier, on suppose que toute source engendre des flots de même taille moyenne σ séparés par les intervalles d'inactivité de même durée moyene $1/\nu$.

Notons $\mu = C/\sigma$ le taux virtuel d'accomplissement des flots de chaque source et $\rho = \nu/\mu$ la charge virtuelle correspondante. On suppose que la charge moyenne de toutes les sources est conservée :

$$\rho = \frac{1}{I}\sum_{i=1}^{I} \rho_i. \tag{4.11}$$

Distribution stationnaire

À partir du Corrolaire 4.14, on obtient la distribution stationnaire du nombre de sources actives :

$$\begin{aligned} \pi'(n) &= \pi'(0) n! \sum_{x\in\mathcal{X}(n)} \rho^n, \\ &= \pi'(0) n! \binom{I}{n} \rho^n. \end{aligned} \tag{4.12}$$

On a les analogues des Théorèmes 4.3, 4.8 et 4.5

Théorème 4.15 *On a pour tout $n \leq I-1$,*

$$\frac{\pi(n)}{\pi'(n)} \geq \frac{\pi(n+1)}{\pi'(n+1)}.$$

Plus généralement, la proportion $\pi(n)/\pi(n+1)$ est diminuée lorsque l'on remplace les taux de charge $\rho_i, \rho_j, i \neq j$, par deux nouveaux taux de la même somme mais d'une différence moins importante.

Théorème 4.16 *Le nombre de sources actives est plus grand en distribution lorsque les sources sont équilibrées :*

$$\sum_{n \geq m} \pi'(n) \geq \sum_{n \geq m} \pi(n), \quad \textit{pour tout } m = 0, 1, \ldots, I.$$

Débits moyens

Un nombre plus grand de sources actives entraîne un débit échantillonné en temps plus faible :

Théorème 4.17 *Le débit moyen échantillonné en temps est plus faible lorsque les sources sont équilibrées.*

Preuve. Il suffit d'appliquer le Lemme 4.4 pour les nombres suivants :

$$\frac{\pi(1)}{\pi'(1)} \geq \frac{\pi(2)}{\pi'(2)} \geq \ldots \geq \frac{\pi(I)}{\pi'(I)} \quad \text{et} \quad t_n = n, n = 1, \ldots, I,$$

on déduit :

$$\frac{\sum_{n=1}^{I} n\pi(n)}{\sum_{n=1}^{I} n\pi'(n)} \leq \frac{\sum_{n=1}^{I} \pi(n)}{\sum_{n=1}^{I} \pi'(n)},$$

et donc

$$\Longrightarrow \frac{p}{E[n]} \geq \frac{p'}{E'[n]}.$$

$\square$

Notons que ce résultat n'est vrai que pour le débit moyen de toutes les sources, et pas pour le débit moyen de chaque source. Par exemple, dans le cas de $I = 2$ sources, les débits moyens de chaque source sont :

$$\gamma_1^{time} = C\frac{1+\rho_2}{1+2\rho_2} \quad \text{et} \quad \gamma_2^{time} = C\frac{1+\rho_1}{1+2\rho_1}.$$

Ainsi, une source bénéficie de l'équilibrage du trafic si et seulement si l'autre source contribue à la plus grande partie de la charge totale.

De même, on a une comparaison sur le débit moyen échantillonné par flot des flots courts :

Théorème 4.18 *Les flots courts reçoivent un débit moyen échantillonné par flot plus faible lorsque les sources sont équilibrées.*

Preuve. En appliquant le Lemme 4.4 pour

$$\frac{\pi'(I)}{\pi(I)} \geq \frac{\pi'(I-1)}{\pi(I-1)} \geq \ldots \geq \frac{\pi'(0)}{\pi(0)},$$

et les constantes en ordre décroissant

$$\frac{1}{I+1} \leq \frac{1}{I} \leq \ldots \leq \frac{1}{1},$$

on obtient

$$\frac{\sum_{n=0}^{I} \frac{1}{n+1}\pi'(n)}{\sum_{n=0}^{I} \frac{1}{n+1}\pi(n)} \leq \frac{\sum_{n=0}^{I} \pi'(n)}{\sum_{n=0}^{I} \pi(n)} = 1.$$

D'où,

$$C\sum_{n=0}^{I} \frac{1}{n+1}\pi'(n) \leq C\sum_{n=0}^{I} \frac{1}{n+1}\pi(n).$$

□

Remarquons qu'en utilisant la deuxième partie du Théorème 4.15, on peut montrer un résultat plus fort que l'équilibrage d'un sous-ensemble de sources diminue aussi ces débits moyens.

Distributions des débits

Dans cette partie, on montrera que l'équilibrage de sources diminue les débits en distribution.

Théorème 4.19 *Le débit échantillonné en temps est plus faible en distribution lorsque les sources sont équilibrées :*

$$\mathbb{P}'\big(X^{time} \geq d\big) \leq \mathbb{P}\big(X^{time} \geq d\big),$$

où $\mathbb{P}$ est la probabilité correspondante à la distribution π du système original et $\mathbb{P}'$ est celle correspondante à la distribtuion π'.

Preuve. Rappelons que la distribution du débit échantillonné en temps est donnée par

$$\mathbb{P}(X^{time} = d) = \frac{k\pi(k)}{\mathbb{E}[n]}, \qquad d = \frac{C}{k}.$$

Alors

$$\mathbb{P}(X^{time} \geq d) = \frac{\sum_{n\leq k} n\pi(n)}{\sum n\pi(n)}, \qquad d = \frac{C}{k}.$$

Ensuite, en utilisant le Théorème 4.15 et le Lemme 4.4, on obtient

$$\frac{\sum_{n\leq k} n\pi'(n)}{\sum n\pi'(n)} \leq \frac{\sum_{n\leq k} n\pi(n)}{\sum n\pi(n)}$$

□

Pour les flots courts, la distribution du débit échantillonné par flot est donnée par

$$\lim_{s\to 0} \mathbb{P}(X^{flow}(s) = d) = \pi(k-1), \qquad d = \frac{C}{k}.$$

Alors

$$\lim_{s\to 0} \mathbb{P}(X^{flow}(s) \geq d) = \sum_{n\leq k} \pi(n-1), \qquad d = \frac{C}{k},$$

qui est reliée à la distribution du nombre de sources actives (Théorème 4.16). On a le résultat suivant.

Théorème 4.20 *Le débit échantillonné par flot des flots courts est plus faible en distribution lorsque les sources sont équilibrées.*

Remarquons aussi qu'en utilisant la deuxième partie du Théorème 4.15, on peut montrer que les débits sont plus faibles en distribution lorsqu'un sous-ensemble de sources sont équilibrées.

4.2.3 Contrôle d'admission

Dans cette section, le nombre de sources actives est limité à une valeur fixe $J \leq I$, qui assure un débit minimal de C/J pour tout transfert de donnée. Si une source essaie d'engendrer un nouveau flot lorsque J sources sont actives, ce flot est bloqué et la source revient dans l'état inactif. On s'intéresse à la probabilité de blocage et aux débits échantillonnés en temps et par flot.

Distribution stationnaire

Par les propriétés d'insensibilité et de réversibilité, la distribution stationnaire π de l'état du système avec le contrôle d'admission est la restriction de la distribution stationnaire (4.7) aux états $n \leq J$. De même, la distribution stationnaire π' du nombre de sources actives lorsque les sources sont équilibrées est la restriction de la distribution stationnaire (4.12) aux états $n \leq J$. On a les analogues des Théorèmes 4.15 et 4.16 :

Théorème 4.21 *Pour tout $n \leq J-1$, on a*

$$\frac{\pi(n)}{\pi'(n)} \geq \frac{\pi(n+1)}{\pi'(n+1)}.$$

De plus, la proportion $\pi(n)/\pi(n+1)$ est diminuée lorsque l'on remplace les taux de charge $\rho_i, \rho_j, i \neq j$ par deux nouveaux taux de même somme mais d'une différence plus faible.

Théorème 4.22 *Le nombre de sources actives est plus grand en distribution lorsque les sources sont équilibrées :*

$$\sum_{n \geq m} \pi'(n) \geq \sum_{n \geq m} \pi(n), \quad \textit{pour tout } m = 0, 1, \ldots, J.$$

Débits

Comme les distributions stationnaires ne sont que les restrictions des distributions du cas sans contrôle d'admission, on s'aperçoit que les débits eux-mêmes sont plus faibles en distribution et leurs moyennes sont plus faibles lorsque toutes les sources ou un sous-ensemble de sources sont équilibrées.

Probabilité de blocage

Soit $\mathcal{X}_i \subset \mathcal{X}$ l'ensemble des états où la source i est inactive et $\mathcal{X}_i(n) \subset \mathcal{X}_i$ l'ensemble des états où la source i est inactive et n sources sont actives. La probabilité de blocage des flots de la source i est égale à la probabilité que J sources soient actives sachant que la source i est inactive :

$$B_i = \frac{\sum_{x \in \mathcal{X}_i(J)} \alpha(x)}{\sum_{x \in \mathcal{X}_i} \alpha(x)}.$$

La probabilité de blocage moyenne est donnée par :

$$B = \frac{\sum_{i=1}^{I} \nu_i \sum_{x \in \mathcal{X}_i(J)} \alpha(x)}{\sum_{i=1}^{I} \nu_i \sum_{x \in \mathcal{X}_i} \alpha(x)}. \tag{4.13}$$

Il n'est pas vrai en général que l'équilibrage des sources augmente la probabilité de blocage. Par exemple pour $I = 2$ sources et au plus $J = 1$ source active, on a :

$$B_1 = \frac{\rho_2}{1 + \rho_2} \quad \text{et} \quad B_2 = \frac{\rho_1}{1 + \rho_1}.$$

Ainsi, une source bénéficie de l'équilibrage du trafic si et seulement si l'autre source génère la plus grande partie de la charge totale. Pour la probabilité de blocage moyenne, on obtient :

$$B = \frac{\nu_1 \rho_2 + \nu_2 \rho_1}{\nu_1(1 + \rho_2) + \nu_2(1 + \rho_1)},$$

qui n'est pas maximal pour $\rho_1 = \rho_2$ en général.

Une condition suffisante pour que cette propriété soit vraie est que les flots de toute source aient la même moyenne de demande de service $1/\mu$ dans le système original.

Au contraire, on a les analogues des résultats sur la probabilité de saturation du modèle d'Engset : l'équilibrage de toutes les sources ou d'un sous-ensemble de sources augmente la probabilité de saturation :

$$B^{virt} = \pi(J).$$

Et de plus, l'équilibrage d'un sous-ensemble de sources augmente la probabilité de blocage d'une source en dehors de ce sous-ensemble.

4.2.4 Contraintes de capacité

Introduisons maintenant des contraintes de capacité dans le modèle. On considère le cas simple où la capacité est limitée à une constante commune $c < C$ pour tous les flots.

Notons n_0 la partie entière de C/c. Lorsque le nombre de flots n est strictement supérieur à n_0, $C/n < c$, donc chaque flot reçoit une capacité C/n bit/s ; chaque flot reçoit une capacité c sinon.

Distribution stationaire

Les équations de balance détaillée sont légèrement modifiées :

$$\alpha(x)\nu_i = \alpha(x+e_i)\mu_i \frac{c}{C}, \quad \forall x \in \mathcal{X} : x_i = 0, n \leq n_0,$$

et

$$\alpha(x)\nu_i = \alpha(x+e_i)\frac{\mu_i}{n+1}, \quad \forall x \in \mathcal{X} : x_i = 0, n > n_0.$$

On a l'analogue du Théorème 4.13 :

Théorème 4.23 *Le système est insensible et la distribution stationnaire de l'état du système est donnée par*

$$\alpha(x) = \alpha(0)\Big(\frac{C}{c}\Big)^n \prod_{i=1}^{I} \rho_i^{x_i}, \quad si\ n \leq n_0,$$

et

$$\alpha(x) = \alpha(0)\Big(\frac{C}{c}\Big)^{n_0} \frac{n!}{n_0!} \prod_{i=1}^{I} \rho_i^{x_i}, \quad si\ n > n_0.$$

Corollaire 4.24 *La distribution stationnaire du nombre de sources actives est donnée par*

$$\pi(n) = \pi(0)\Big(\frac{C}{c}\Big)^n \sum_{x \in \mathcal{X}(n)} \prod_{i=1}^{I} \rho_i^{x_i}, \quad si\ n \leq n_0,$$

et

$$\pi(n) = \pi(0)\Big(\frac{C}{c}\Big)^{n_0} \frac{n!}{n_0!} \sum_{x \in \mathcal{X}(n)} \prod_{i=1}^{I} \rho_i^{x_i}, \quad si\ n > n_0.$$

On s'aperçoit que pour chaque nombre de flots fixé n, ces distributions stationnaires sont égales à celles obtenues dans le cas sans contrainte de capacité multipliées par certaines constantes. Alors on a les analogues des Théorèmes 4.15 et 4.16 :

Théorème 4.25 *On a pour tout $n \leq I-1$,*

$$\frac{\pi(n)}{\pi'(n)} \geq \frac{\pi(n+1)}{\pi'(n+1)}.$$

De plus, la proportion $\pi(n)/\pi(n+1)$ est diminuée lorsque l'on remplace les taux de charge $\rho_i, \rho_j, i \neq j$ par deux nouveaux taux de même somme mais d'une différence plus faible.

Théorème 4.26 *Le nombre de sources actives est plus grand en distribution lorsque les sources sont équilibrées :*

$$\sum_{n \geq m} \pi'(n) \geq \sum_{n \geq m} \pi(n), \qquad \forall m = 0, \ldots, I.$$

Débits

Lorsque toutes les sources ou un sous-ensemble de sources sont équilibrées, les débits sont plus faibles en distribution et leurs moyennes sont plus faibles.

Probabilité de blocage

Si l'on introduit un contrôle d'admission imposant que le nombre de sources actives soit inférieur à un seuil J, on peut montrer comme dans la Section 4.2.3 que l'équilibrage d'un sous-ensemble de sources augmente la probabilité de saturation et la probabilité de blocage d'une source en dehors de ce sous-ensemble.

4.2.5 Réseaux

On montrera dans cette section qu'en général, les résultats obtenus auparavant sont faux pour les réseaux. En particulier, l'équilibrage de sources peut augmenter et aussi diminuer les débits instantanés moyens, les probabilités de blocage, en fonction des paramètres du réseau.

Reprenons l'exemple d'un réseau linéaire de deux liens mais cette fois ci partagés par 4 sources de trafic élastique, voir la Figure 4.2.

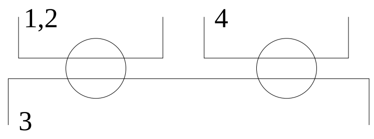

FIG. 4.2 – Un réseau linéaire.

L'espace d'états de ce résau est l'ensemble $\mathcal{X} = \{0,1\}^4$. Le réseau est équivalent à celui de quatre files PS avec taux de service dépendants de l'état. Alors une mesure stationnaire est donnée par

$$\alpha(x) = \Phi(x) \prod_{i=1}^{4} \rho_i^{x_i}, \qquad \forall x \in \mathcal{X},$$

où Φ désigne la fonction de balance qui peut être calculée explicitement dans ce cas.

Pour la source i, si active, c.à.d. $x_i = 1$, son débit instantané est donné par

$$\varphi_i(x) = \frac{\Phi(x - e_i)}{\Phi(x)}.$$

Alors son débit instantané moyen est donné par

$$\begin{aligned}
\gamma_i &= \mathbb{E}\Big[\frac{\Phi(x - e_i)}{\Phi(x)} \mid x_i = 1\Big] \\
&= \frac{\sum_{x:x_i=1} \frac{\Phi(x-e_i)}{\Phi(x)}\alpha(x)}{\sum_{x:x_i=1}\alpha(x)} \\
&= \frac{\sum_{x:x_i=0} \rho_i\alpha(x)}{\sum_{x:x_i=1}\alpha(x)} \\
&= \rho_i \frac{\mathbb{P}(x_i = 0)}{\mathbb{P}(x_i = 1)},
\end{aligned}$$

on retrouve l'expression (4.8) du débit moyen échantillonné en temps γ_i^{time}.

En particulier, pour la source 4, le débit instantané moyen est donné par

$$\gamma_4 = \frac{1 + \rho_1 + \rho_2 + \rho_3 + 2(\rho_1\rho_2 + \rho_1\rho_3 + \rho_2\rho_3) + 6\rho_1\rho_2\rho_3}{1 + \rho + 1 + \rho_2 + 2\rho_3 + 2\rho_1\rho_2 + 6\rho_1\rho_3 + 6\rho_2\rho_3 + 8\rho_1\rho_2\rho_3}.$$

Lorsque $\rho_3 \to \infty$, ce débit moyen tend à la fraction

$$\frac{1 + \rho_1 + \rho_2 + 6\rho_1\rho_2}{2 + 3\rho_1 + 3\rho_2 + 8\rho_1\rho_2},$$

qui est augmenté lorsque les sources 1 et 2 sont équilibrées. Donc l'équilibrage de sources peut augmenter le débit instantané moyen de la source 4.

Au contraire, lorsque $\rho_4 \to 0$, la source 4 est presque sûrement inactive et le système asymptotique est le modèle d'un seul lien partagé par les sources 1, 2 et 3. Par conséquence, l'équilibrage des sources 1 et 2 peut diminuer le débit instantané moyen des trois sources 1, 2 et 3.

Introduisons maintenant un contrôle d'admission dans ce réseau tel que le premier lien accepte au plus $C_1 = 2$ sources actives et le deuxième n'accepte qu'une source active en même temps. Dans ce cas, on observe que l'équilibrage des sources 1 et 2 diminue la probabilité de blocage de la source 4 car cette probabilité est donnée par

$$B_4 = \frac{\rho_3 + 2\rho_1\rho_3 + 2\rho_2\rho_3}{1 + \rho_1 + \rho_2 + \rho_3 + 2\rho_1\rho_2 + 2\rho_1\rho_3 + 2\rho_2\rho_3},$$

et l'équilibrage augmente le produit $\rho_1\rho_2$, il diminue la probabilité de blocage B_4.

Par contre, si $C_1 = 3$ et $C_2 = 1$, la probabilité de blocage de la source 4 est donnée par

$$B_4 = \frac{\rho_3 + 2\rho_1\rho_3 + 2\rho_2\rho_3 + 6\rho_1\rho_2\rho_3}{1 + \rho_1 + \rho_2 + \rho_3 + 2\rho_1\rho_2 + 2\rho_1\rho_3 + 2\rho_2\rho_3 + 6\rho_1\rho_2\rho_3},$$

qui est augmentée lorsque les sources 1 et 2 sont équilibrées.

En conclusion, les résultats précédents sont faux pour les réseaux, c.à.d. que l'équilibrage de sources peut diminuer et aussi augmenter les débits, les probabilités de blocage, en fonction des paramètres précis du système.

Conclusion

En conclusion, nous présentons les principales contributions de ce travail et mentionnons également de nombreux problèmes ouverts qui constitueraient des perspectives de recherche intéressantes.

Insensibilité dans les réseaux de files d'attente

Dans la première partie, nous avons montré que la file symétrique est insensible même si les permutations aléatoires de clients sont permis à chaque arrivée et à chaque départ, voir la section 2.1. Ensuite, l'insensibilité des réseau de Jackson et des réseaux de Kelly avec permutations aléatoires ont été aussi conclue, voir les sections 2.2 et 2.3 respectivement. De plus, nous avons montré que si les capacités de service, les taux de service et le routage dépendent du nombre de clients à chaque file dans ces réseaux, l'insensibilité est équivalente à la propriété de balance (Section 2.2.4).

En utilisant ce résultat, on peut construire de nombreuses nouvelles disciplines de service insensibles (Section 2.4.1) ; et donc étend la classe de disciplines insensibles connues afin d'approcher la classe de toutes les disciplines de service insensibles.

Il est très intéressant à étudier l'insensibilité des modèles avec des transitions en batch et avec des clients négatifs. Ces modèles sont élaborés par plusieurs auteurs, voir [BvD91, YM98, Miy97, MT97, MW96, MY93, Tay00, CHPT97, HNT94, HT90, HPTvD90, vDS90] et [CMP99, FGS96, Gel93, FG92, Gel91, GGS91] respectivement.

Métriques de débit

Dans le chapitre 3, nous avons introduit deux nouvelles métriques de débit dans les modèles de trafic élastique. Il s'agit des débits échantillonnés par flots et en temps. Ces deux débits sont très utiles pour évaluer la performance d'un modèle.

Équilibrage de sources de trafic élastique

Dans le dernier chapitre, nous avons considéré un lien unique partagé par un nombre de sources de trafic élastique hétérogènes. Nous avons montré que l'équilibrage des sources diminue le débit moyen et diminue également les débits échantillonnés par flots et en temps. En présence d'un contrôle d'admission, l'équilibrage d'un sous-ensemble de sources augmente la probabilité de blocage de toute source en dehors de ce sous-ensemble. Si les flots de toute source ont la même moyenne de demande de service dans le système original, l'équilibrage des sources augmente la probabilité de blocage du système. Ce résultat assure que l'on peut dimmensionner le modèle en supposant les souces homogènes malgré la forte hétérogénéité observée en pratique.

En général, le résultat est en revanche faux dans le modèle multidébit et dans un réseau de liens, de trafic élastique ou de trafic de circuits. L'équilibrage d'une classe de sources peut augmenter ou diminuer le débit et la probabilité de blocage des autres sources. Existe-il des conditions particulières qui assurent dans un réseau que l'équilibrage des sources augmente la probabilité de blocage et diminue le débit ? Y-a-t-il un autre moyen pour simplifier le dimensionnement, similaire à l'équilibrage des sources ? Ces questions ne sont pas encore traitées et ouvrent un perspective de recherche intéressant.

A

Réseaux de Kelly avec permutations décrits par RGSMP

Dans cette partie, on décrira les réseaux de Kelly avec permutations aléatoires (Section 2.3) par un modèle RGSMP et on montrera leur propriété d'insensibilité.

Modèle

Dans notre réseau, il y a des classes de clients et chaque client demande un service à une et seulement une file d'attente fixée correspondante à sa classe. Les clients de la classe c arrivent dans le réseau selon un processus de Poisson d'intensité ν_c et chacun demande un service de moyenne $1/\mu_c$. Ensuite, à l'achèvement de son service, ces clients rejoignent la classe c' avec la probabilité $p_{cc'}$, ou quittent le réseau avec la probabilité

$$p_c = 1 - \sum_{c' \in \mathcal{C}} p_{cc'},$$

où $\mathcal{C}$ désigne l'ensemble des classes.

Posons $n = \{n_c, c \in \mathcal{C}\}$ le macro-état du système, où n_c donne le nombre de clients de la classe c, et G l'ensemble de ces macro-états.

Ce réseau peut être décrit par un processus semi-markovien généralisé dont les emplacements actifs sont les couples (c, l). Chaque couple (c, l) correspond soit à un client dans le réseau, où c désigne la classe du client et l désigne sa position dans la file correspondante, soit à un arrivé, où c désigne la classe du nouvel arrivé et $l = 0$ par convention.

Pour chaque classe c, il y a deux types d'horloges dont

- a_c représente des arrivées avec la moyenne du temps de séjour nominal $1/\mu_{a_c} = 1/\nu_c$.
- b_c représente des clients de cette classe avec la moyenne du temps de séjour nominal $\mu_{b_c} = \mu_c$.

Il y a trois types de transitions dans le réseau :

Arrivée d'un client de la classe c à la file correspondante i. Ce client est placé à la position l et les anciens clients sont permutés à chaque file selon une certaine permutation σ avec la probabilité $\alpha_c(n,\sigma)$. Dans ce cas, l'horloge à l'emplacement $(c,0)$ s'épuise, une nouvelle horloge est créée à cet emplacement et une aure horloge est créée à l'emplacement (c,l).

Alors l'ensemble des emplacements des horloges qui se sont épuisées est $U = \{(c,0)\}$ et l'ensemble des nouvelles horloges est $U' = \{(c,l),(c,0)\}$. Notons $n' = n + e_c$ le nouveau macro-état alors la permutation des anciens clients correspond à une réallocation bijective d'horloges :

$$\Gamma_{(n,U,n',U')} : A(n) \setminus U \longrightarrow A(n') \setminus U'$$

de telle sorte que le type de chaque horloge reste inchangé :

$$\tau(l',n') = \tau(\Gamma^{-1}_{(n,U,n',U')}(l'),n), \qquad \forall l' \in A(n') \setminus U'.$$

Le taux de cette transition est donné par

$$p(n,U,n',U') = \sum_{\sigma} \delta_i(l,n')\alpha_c(n,\sigma) = \delta_i(l,n'),$$

où i désigne la file correspondante à la classe c du nouvel arrivé.

Départ du client de la classe en position l de la file correspondante et les autres clients sont permutés à chaque file selon une certaine permutation σ avec la probabilité $\beta_c(n',\sigma)$, où $n' = n - e_c$ désigne le macro-état obtenu après le départ. Dans ce cas, l'horloge à l'emplacement (c,l) s'épuise et aucune horloge n'est créée :

$$U = \{(c,l)\} \text{ et } U = \emptyset.$$

Le taux de cette transition est donné par

$$p(n,U,n',U') = \sum_{\sigma} \beta_c(n-e_c,\sigma)p_c = p_c.$$

Changement de classes. Après l'achèvement de son service, le client de la classe c à la position l rejoint la classe c' et prend la position l' de la file correspondante à sa nouvelle classe. Les autres clients sont permutés à chaque file selon une certaine permutation σ avec la probabilité $\gamma_{cc'}(n-e_c,\sigma)$, où $n - e_c$ désigne le macro-état obtenu du macro-état n en supprimant un client de la classe c. Dans ce cas, l'horloge à l'emplacement (c,l) s'épuise et une nouvelle horloge est créée à l'emplacement (c',l') :

$$U = \{(c,l)\} \text{ et } U = \{c',l'\}.$$

Posons $n' = n - e_c + e_{c'}$ le macro-état du réseau après ce changement de classes, le taux de cette transition est donné par

$$p(n,U,n',U') = \sum_{\sigma} \gamma_{cc'}(n-e_c,\sigma)p_{cc'} = p_{cc'}.$$

Remarquons que dans ce modèle RGSMP, le macro-état considéré se compose du nombre de clients de chaque classe. De plus, la fonction δ et les permutations aléatoires peuvent dépendre de ce macro-état. On démontrera directement le Théorème 2.12 sur la distribution stationnaire du nombre de clients de chaque classe. Le Théorème 2.11 sur le nombre de clients à chaque file est ensuite une conséquence de ce résultat.

Cas exponentiel

Théorème A.27 *Si les demandes de services sont toutes exponentielles, la distribution stationnaire du macro-état du réseau est à forme produit et est donnée par*

$$\pi(n) = \prod_{c \in \mathcal{C}} (1-\rho_c)\rho_c^{n_c}, \qquad n \in G,$$

où λ_c désigne le taux d'arrivée effectif de la classe c défini par les équations de traffic pour le réseau de Kelly (2.9) et $\rho_c = \lambda_c/\mu_c$ désigne le taux de charge dû à la classe c.

Preuve. Remarquons que les permutations se simplifient dans les expressions des taux de transition. Alors comme dans le cas de réseau de Kelly sans permutation, la mesure de probabilité π donnée dans le théorème satifait les équations de balance globale (1.22) :

$$\pi(n) \sum_{c \in \mathcal{C}} \left(\nu_c + \sum_l \mu_c \delta_i(l,n) \right) = \sum_{n' \in G} \sum_{(c',l') \in A(n')} \sum_{U \subset A(n)} \delta_j(l',n') \mu_{\tau_{c'}(l',n')} \pi(n') p(n',U',n,U),$$

où i,j désignent la file correspondante à la classe c, c' respectivement, et $\tau_{c'}(l',n')$ désigne le type du client à l'emplacement (c',l').

Alors π est la distribution stationnaire du macro-état n. $\square$

Insensibilité

Théorème A.28 *Le réseau de Kelly avec des permutations aléatoires est insensible à la distribution des demandes de service des clients de chaque classe. La distribution stationnaire du nombre de clients de chaque classe est donnée par*

$$\pi(n) = \prod_{c \in \mathcal{C}} (1-\rho_c)\rho_c^{n_c},$$

Preuve. Tout d'abord, la chaîne incluse du processus semi-markovien considéré est récurrente positive (1.23) :

$$\sum_{n \in G} \pi(n) \sum_{c \in \mathcal{C}} \left(\nu_c + \sum_l \mu_c \delta_i(l,n) \right) < \infty.$$

Ensuite, les processus d'arrivées sont tous poisonniens dans notre modèle, alors les temps de séjour des horloges aux emplacements $(c, 0)$ sont exponentiels pour toutes classes c. Les équations de balance locale pour les horloges non-exponentielles (1.25) deviennent

$$\mu_c \delta_i(l,n)\pi(n) = \pi(n - e_c)\nu_c \delta_i(l,n) + \pi(n - e_c + e_{c'})\mu_{c'}\delta_i(l,n)p_{c'c}, \qquad \forall (c,l) \in A(n), l > 0.$$

Remarquons que ces équations de balance locale ressemblent aux équations de balance partielle dans la preuve du Théorème 2.10. En remplaçant π par sa forme explicite, on peut vérifier que la mesure de probabilité π satisfait ces équations de balance locale. D'après la Proposition 1.16 et la Remarque (1.1), le modèle est insensible à la distribuion des demandes de service des clients de chaque classe. $\square$

Bibliographie

[BB03] F. Baccelli and P. Brémaud. *Elements of queueing theory*, volume 26 of *Applications of Mathematics (New York)*. Springer-Verlag, Berlin, second edition, 2003. Palm martingale calculus and stochastic recurrences, Stochastic Modelling and Applied Probability.

[BBNnQ05] R. Bekker, S. Borst, and R. Núñez Queija. Performance of tcp-friendly streaming sessions in the presence of heavy-tailed elastic flows. *Perform. Eval.*, 61(2-3) :143–162, 2005.

[BCMP75] F. Baskett, K.M. Chandy, R.R. Muntz, and F.G. Palacios. Open, closed, and mixed networks of queues with different classes of customers. *J. Assoc. Comput. Mach.*, 22 :248–260, 1975.

[BFBP$^+$01] S. Ben Fred, T. Bonald, A. Proutiere, G. Régnié, and J.W. Roberts. Statistical bandwidth sharing : a study of congestion at flow level. *SIGCOMM Comput. Commun. Rev.*, 31(4) :111–122, 2001.

[BK00] A.W. Berger and Y. Kogan. Dimensioning bandwidth for elastic traffic in high-speed data networks. *IEEE/ACM Trans. Netw.*, 8(5) :643–654, 2000.

[BM01] T. Bonald and L. Massoulié. Impact of fairness on internet performance. *SIGMETRICS Perform. Eval. Rev.*, 29(1) :82–91, 2001.

[BMPV06] T. Bonald, L. Massoulié, A. Proutière, and J. Virtamo. A queueing analysis of max-min fairness, proportional fairness and balanced fairness. *Queueing Syst.*, 53(1-2) :65–84, 2006.

[BOR03] T. Bonald, P. Olivier, and J.W. Roberts. Dimensioning high-speed IP access networks. In *Proceedings of the ITC 18 Conference*, pages 241–251, 2003.

[BP02a] T. Bonald and A. Proutière. Insensitive bandwidth sharing. In *Proceedings of the IEEE GLOBECOM Conference*, 2002.

[BP02b] T. Bonald and A. Proutière. Insensitivity in processor-sharing networks. *Perform. Eval.*, 49(1-4) :193–209, 2002.

[BP04] T. Bonald and A. Proutière. On performance bounds for the integration of elastic and adaptive streaming flows. *SIGMETRICS Perform. Eval. Rev.*, 32(1) :235–245, 2004.

[BPRR01] T. Bonald, A. Proutière, G. Régnié, and J.W. Roberts. Insensitivity results in statistical bandwidth sharing. In *Proceedings of the ITC 17 Conference*, 2001.

[Bre99] P. Bremaud. *Markov chains, Gibbs fields, Monte-Carlo simulation and queues.* Springer-Verlag, New York, 1999.

[BT07a] T. Bonald and M.A. Tran. Balancing elastic traffic sources. *IEEE Communications Letters*, 11(8) :692–694, 2007.

[BT07b] T. Bonald and M.A. Tran. Flow vs. time sampling for throughput performance evaluation. *Perform. Eval.*, 64(9-12) :1181–1193, 2007.

[BvD91] R.J. Boucherie and N.M. van Dijk. Product forms for queueing networks with state-dependent multiple job transitions. *Adv. in Appl. Probab.*, 23(1) :152–187, 1991.

[BvOZ05] S. Borst, D. van Ooteghem, and B. Zwart. Tail asymptotics for discriminatory processor-sharing queues with heavy-tailed service requirements. *Perform. Eval.*, 61(2-3) :281–298, 2005.

[CHPT97] J.L. Coleman, W. Henderson, C.E.M. Pearce, and P.G. Taylor. A correspondence between product-form batch-movement queueing networks and single-movement networks. *J. Appl. Probab.*, 34(1) :160–175, 1997.

[CMP99] X. Chao, M. Miyazawa, and M. Pinedo. *Queueing networks. Customers, signals, and product form solutions.* Wiley, 1999.

[CMST98] X. Chao, M. Miyazawa, R.F. Serfozo, and H. Takada. Markov network processes with product form stationary distributions. *Queueing Systems Theory Appl.*, 28(4) :377–401, 1998.

[Coh79] J.W. Cohen. The multiple phase service network with generalized processor sharing. *Acta Inform.*, 12(3) :245–284, 1979.

[Dad01a] H. Daduna. *Queueing networks with discret time scale : Explicit expressions for the steady behavior of discrte time stochastic networks*, volume 2046 of *Lecture Notes in Computer Science.* Springer, Berlin, 2001.

[Dad01b] H. Daduna. Stochastic networks with product form equilibrium. In *Stochastic processes : theory and methods*, volume 19 of *Handbook of Statist.*, pages 309–364. North-Holland, Amsterdam, 2001.

[Dar70] J.P. Dartois. Lost call cleared systems with unbalanced traffic sources. In *Proceedings of the ITC 6 Conference*, 1970.

[DPR04] F. Delcoigne, A. Proutière, and G. Régnié. Modeling integration of streaming and data traffic. *Performance Evaluation*, 55 :185–209, 2004.

[DS83] H. Daduna and R. Schassberger. Networks of queues in discrete time. *Z. Oper. Res. Ser. A-B*, 27(5) :A159–A175, 1983.

[Eng] T.O. Engset. On the calculation of switches in an automatic telephone system. In : *Tore Olaus Engset : The man behind the formula*, Eds : A. Myskja, O. Espvik, 1998.

[Erl09] A.K. Erlang. The Theory of Probabilities and Telephone Conversations. *Nyt Tidsskrift for Matematik*, 20 :33, 1909.

[FG92] J.-M. Fourneau and E. Gelenbe. Multiple class G-networks. In *Computer science and operations research (Williamsburg, VA, 1992)*, pages 149–157. Pergamon, Oxford, 1992.

[FGS96] J.-M. Fourneau, E. Gelenbe, and R. Suros. G-networks with multiple classes of negative and positive customers. *Theoret. Comput. Sci.*, 155(1) :141–156, 1996.

[FKAS82] P. Franken, D. König, U. Arndt, and V. Schmidt. *Queues and point processes.* Wiley Series in Probability and Mathematical Statistics : Applied Probability and Statistics. John Wiley & Sons Ltd., Chichester, 1982.

[FMI80] G. Fayolle, I. Mitrani, and R. Iasnogorodski. Sharing a processor among many job classes. *J. Assoc. Comput. Mach.*, 27(3) :519–532, 1980.

[Gel91] E. Gelenbe. Product-form queueing networks with negative and positive customers. *J. Appl. Probab.*, 28(3) :656–663, 1991.

[Gel93] E. Gelenbe. G-networks with triggered customer movement. *J. Appl. Probab.*, 30(3) :742–748, 1993.

[GGS91] E. Gelenbe, P. Glynn, and K. Sigman. Queues with negative arrivals. *J. Appl. Probab.*, 28(1) :245–250, 1991.

[GRZ04] F. Guillemin, P. Robert, and B. Zwart. Tail asymptotics for processor-sharing queues. *Adv. in Appl. Probab.*, 36(2) :525–543, 2004.

[HLN97] D.P. Heyman, T.V. Lakshman, and Arnold L. Neidhardt. A new method for analysing feedback-based protocols with applications to engineering web traffic over the internet. *SIGMETRICS Perform. Eval. Rev.*, 25(1) :24–38, 1997.

[HNT94] W. Henderson, B.S. Northcote, and P.G. Taylor. Geometric equilibrium distributions for queues with interactive batch departures. *Ann. Oper. Res.*, 48(1-4) :493–511, 1994.

[HPTvD90] W. Henderson, C.E.M. Pearce, P.G. Taylor, and N.M. van Dijk. Closed queueing networks with batch services. *Queueing Systems Theory Appl.*, 6(1) :59–70, 1990.

[HT90] W. Henderson and P.G. Taylor. Product form in networks of queues with batch arrivals and batch services. *Queueing Systems Theory Appl.*, 6(1) :71–87, 1990.

[Jac57] J.R. Jackson. Networks of waiting lines. *Operations Res.*, 5 :518–521, 1957.

[JT89] M.A. Johnson and M.R. Taaffe. Matching moments to phase distributions : mixtures of Erlang distributions of common order. *Comm. Statist. Stochastic Models*, 5(4) :711–743, 1989.

[JT90a] M.A. Johnson and M.R. Taaffe. Matching moments to phase distributions : density function shapes. *Comm. Statist. Stochastic Models*, 6(2) :283–306, 1990.

[JT90b] M.A. Johnson and M.R. Taaffe. Matching moments to phase distributions : nonlinear programming approaches. *Comm. Statist. Stochastic Models*, 6(2) :259–281, 1990.

[JT91] M.A. Johnson and M.R. Taaffe. An investigation of phase-distribution moment-matching algorithms for use in queueing models. *Queueing Systems Theory Appl.*, 8(2) :129–147, 1991.

[Kel79] F.P. Kelly. *Reversibility and stochastic networks*. John Wiley & Sons Ltd., Chichester, 1979. Wiley Series in Probability and Mathematical Statistics.

[Ken53] D.G. Kendall. Stochastic processes occuring in the theory of queues and their analysis by the method of the imbedded Markov chain. *Ann. Math. Statistics*, 24 :338–354, 1953.

[KJ77] D. König and U. Jansen. Stochastic processes and properties of invariance for queueing systems with speeds and temporary interruptions. In *Transactions of the Seventh Prague Conference on Information Theory, Statistical Decision Functions, Random Processes and of the Eighth European Meeting of Statisticians (Tech. Univ. Prague, Prague, 1974), Vol. A*, pages 335–343. Reidel, Dordrecht, 1977.

[KK02] A.A. Kherani and A. Kumar. Stochastic models for throughput analysis of randomly arriving elastic flows in the internet. In *Proceedings of INFOCOM*, 2002.

[Kle75] L. Kleinrock. *Queueing Systems*, volume 2. Wiley, 1975.

[LB03] R. Litjens and R.J. Boucherie. Elastic calls in an integrated services network : the greater the call size variability the better the qos. *Perform. Eval.*, 52(4) :193–220, 2003.

[Lit61] J.D.C. Little. A proof for the queuing formula : $L = \lambda W$. *Operations Res.*, 9 :383–387, 1961.

[LvdBB04] R. Litjens, J.L. van den Berg, and R.J. Boucherie. Throughputs in processor sharing models for integrated stream and elastic traffic. Memorandum 1708, Enschede, 2004.

[Mat64] K. Matthes. Zur Theorie der Bedienungsprozesse. In *Trans. Third Prague Conf. Information Theory, Statist. Decision Functions, Random Processes (Liblice, 1962)*, pages 513–528. Publ. House Czech. Acad. Sci., Prague, 1964.

[Miy93] M. Miyazawa. Insensitivity and product-form decomposability of reallocatable GSMP. *Adv. in Appl. Probab.*, 25(2) :415–437, 1993.

[Miy97] M. Miyazawa. Structure-reversibility and departure functions of queueing networks with batch movements and state dependent routing. *Queueing Systems Theory Appl.*, 25(1-4) :45–75, 1997.

[MR00] L. Massoulié and J.W. Roberts. Bandwidth sharing and admission control for elastic traffic. *Telecommun. Syst.*, 15 :185–201, 2000.

[MSS95] M. Miyazawa, R. Schassberger, and V. Schmidt. On the structure of an insensitive generalized semi-Markov process with reallocation and point-process input. *Adv. in Appl. Probab.*, 27(1) :203–225, 1995.

[MT97] M. Miyazawa and P.G. Taylor. A geometric product-form distribution for a queueing network with non-standard batch arrivals and batch transfers. *Adv. in Appl. Probab.*, 29(2) :523–544, 1997.

[MW96] M. Miyazawa and R.W. Wolff. Symmetric queues with batch departures and their networks. *Adv. in Appl. Probab.*, 28(1) :308–326, 1996.

[MY93] M. Miyazawa and G. Yamazaki. Note on batch arrival LCFS and related symmetric queues. *Oper. Res. Lett.*, 14(1) :35–41, 1993.

[Pal57] C. Palm. Waiting times with random served queue. *Tele*, 1, 1957.

[Sch77] R. Schassberger. Insensitivity of steady-state distributions of generalized semi-Markov processes. I. *Ann. Probability*, 5(1) :87–99, 1977.

[Sch78a] R. Schassberger. The insensitivity of stationary probabilities in networks of queues. *Adv. in Appl. Probab.*, 10(4) :906–912, 1978.

[Sch78b] R. Schassberger. Insensitivity of steady-state distributions of generalized semi-Markov processes. II. *Ann. Probability*, 6(1) :85–93, 1978.

[Sch78c] R. Schassberger. Insensitivity of steady-state distributions of generalized semi-Markov processes with speeds. *Adv. in Appl. Probab.*, 10(4) :836–851, 1978.

[Sch86] R. Schassberger. Two remarks on insensitive stochastic models. *Adv. in Appl. Probab.*, 18(3) :791–814, 1986.

[Ser99] R.F. Serfozo. *Introduction to stochastic networks*, volume 44 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, 1999.

[Tay00] P.G. Taylor. Quasi-reversibility and networks of queues with nonstandard batch movements. *Math. Comput. Modelling*, 31(10-12) :335–341, 2000. Stochastic models in engineering, technology, and management (Gold Coast, 1996).

[vDS90] N.M. van Dijk and E. Smeitink. A nonexponential queueing system with independent arrivals and batch servicing. *J. Appl. Probab.*, 27(2) :401–408, 1990.

[Whi85] P. Whittle. Partial balance and insensitivity. *J. Appl. Probab.*, 22(1) :168–176, 1985.

[Whi86] P. Whittle. Partial balance, insensitivity and weak coupling. *Adv. in Appl. Probab.*, 18(3) :706–723, 1986.

[Yas80] S.F. Yashkov. Properties of invariance of probabilistic models of adaptive scheduling in shared-use systems. *Automat. Control Comput. Sci.*, 14(6) :46–51, 1980.

[Yat90] R.D. Yates. *High Speed Round Robin Queueing Networks*. PhD thesis, MIT Dept. of Electrical Engineering and Computer Science, 1990.

[Yat94] R.D. Yates. Analysis of discrete time queues via the reversed process. *Queueing Systems Theory Appl.*, 18(1-2) :107–116, 1994.

[YM98] H. Yamashita and M. Miyazawa. Geometric product form queueing networks with concurrent batch movements. *Adv. in Appl. Probab.*, 30(4) :1111–1129, 1998.