HABILITATION À DIRIGER DES RECHERCHES

présentée devant

L'Université de Rennes 1
Institut de Formation Supérieure
en Informatique et en Communication

par

Bruno Tuffin

Modélisation mathématique pour la conception, l'analyse quantitative et le contrôle de la qualité de service des systèmes

soutenue le 10 Avril 2006 devant le jury composé de

M.	Pierre L'Ecuyer	Président
M.	Eitan Altman	Rapporteur
M.	Jean-Yves Le Boudec	Rapporteur
M.	Jean Walrand	Rapporteur
M.	François Le Gland	Examinateur
M.	Gerardo Rubino	Examinateur

Table des matières

Avant-propos

Dans ce rapport, nous discutons des méthodes de modélisation, des techniques d'évaluation quantitative des performances et de la sûreté de fonctionnement de systèmes, principalement orientées autour des méthodes de simulation. Le domaine privilégié d'application est l'étude des réseaux de communication. Un sujet plus récent que nous abordons également est la détermination des paramètres optimisant les performances des systèmes, principalement dans le cas de la tarification des réseaux de communication où il s'agit de déterminer le schéma de tarification ainsi que les prix maximisant le revenu du fournisseur d'accès ou le bien-être social.

Une partie non négligeable des résultats décrits dans ce travail doit être attribuée à des collaborations actives avec des collègues que je me dois d'associer à ce rapport. Chronologiquement, tout d'abord Gerardo Rubino, mon directeur de thèse et depuis chef du projet INRIA ARMOR au sein duquel je développe mes recherches, avec qui j'ai pu travailler sur les méthodes de simulation d'événement rare de type Monte Carlo. Ces travaux, toujours en cours de développement, ont également été obtenu en collaboration avec Hector Cancela, ancien thésard de l'équipe, et maintenant Professeur à l'Université de Montevideo, collaboration se poursuivant toujours. Juste après ma thèse, j'ai pu apprécié les collaborations avec Louis-Marie Le Ny sur les méthodes quasi-Monte Carlo et avec Bruno Sericola sur l'analyse des files fluides. Mon séjour à Duke University en 1999 m'a permis de travailler sur la modélisation par réseaux de Petri avec Kishor Trivedi et son équipe (Dong Chen, Christophe Hirel étant mes autres co-auteurs), ainsi que sur leur simulation de type Monte Carlo, employant des techniques de réduction de la variance, et l'extension d'un logiciel largement diffusé dans l'industrie et le monde académique, SPNP. A mon retour, j'ai mis à profit ces connaissances pour modéliser et analyser les files à seuils avec hystérésis avec Louis-Marie Le Ny. De plus, suite à des contacts initiés lors de conférences, j'ai eu depuis de nombreuses collaborations sur les méthodes de quasi-Monte Carlo, avec Christian Lécot (Université de Savoie), Giray Ökten (Ball State University puis Florida State University) et Pierre L'Ecuyer (Université de Montréal), ces collaborations étant toujours actives et constituant une part non négligeable de mon travail actuel. Une activité initiée depuis mon retour de Duke University (Janvier 2000) et relativement dissociée de mes travaux précédents est la mise au point de méthodes de tarification des réseaux de communication. Tous ces travaux sont réalisés au sein d'un sous groupe de l'équipe ARMOR, avec David Ros, Yezekael Hayel, Patrick Maillé et Ricardo Orozco ; nous y analysons les méthodes de tarification tant sur le plan théorique (via des outils tels que la théorie des jeux, les techniques d'optimisation...) que sur le plan pratique. Par l'intermédiaire de l'ARC (*Action de*

Recherche Coopérative) PRIXNET, j'ai aussi pu obtenir quelques résultats avec Eitan Altman et Rachid El Azouzi de l'INRIA Sophia-Antipolis ; cette collaboration nous a aussi conduits à travailler au sein de l'ARC TCP sur l'analyse de flux TCP/IP en compétition, en collaboration également avec Milan Vojnović de Microsoft Research. Parallèlement, nous continuons à travailler sur la simulation d'événements rares pour l'analyse de la fiabilité, dans le cadre de l'ACI sécurité Sure-Paths.

Pour clore cet avant-propos, c'est avec sincérité que je remercie les membres de mon jury : Pierre L'Ecuyer, qui s'est donné la peine de venir de Montréal pour présider la soutenance, Eitan Altman, Jean-Yves Le Boudec et Jean Walrand qui ont pris sur leur emploi du temps pourtant chargé pour rédiger un rapport, et François Le Gland et Gerardo Rubino pour leur participation.

Chapitre 1

Introduction

Les concepteurs de systèmes informatiques et de télécommunication –cadre dans lequel notre travail s'est situé, mais pouvant être étendu à bien d'autres domaines– ont besoin de méthodes de modélisation et de quantification de certains paramètres tels que la performance ou la fiabilité. En effet, les systèmes informatiques sont de plus en plus complexes, évoluent de plus en plus rapidement et sont de plus en plus essentiels dans la vie de tous les jours. Il y a donc un besoin de plus en plus important d'outils pour comprendre leur comportement afin de répondre aux questions de coût et de performances qui se posent au cours de la vie de ces systèmes :

- Au cours de leur création, implémentation et de leur dimensionnement : il faut déterminer les « bonnes » architectures et les « bons » protocoles permettant d'obtenir les performances souhaitées, ceci sous des contraintes de coût.
- Durant l'évolution de la configuration et de la charge de travail : une installation pour l'Internet par exemple doit être réalisée de manière à satisfaire la demande accrue et prévisible de ressources (un surdimensionnement de la capacité des câbles est par exemple moins coûteux qu'une modification des installations) ainsi qu'à répondre aux exigences de la future génération de l'Internet.

Dans ces conditions, représenter le système par un modèle abstrait qui sera analysé mathématiquement par la suite est une solution judicieuse.

1.1 Modélisation

La modélisation consiste donc en une représentation abstraite d'un système réel, dans le but de l'analyser mathématiquement. Le processus de modélisation doit être basé sur des hypothèses motivées par deux considérations contradictoires :

- Simplicité : il faut veiller à éliminer les détails « sans intérêt ». L'expression « sans intérêt » reste volontairement vague car tout détail a son importance. Une simplicité accrue permet d'utiliser des méthodes d'analyse beaucoup plus simples et plus rapides.
- Adéquation des résultats avec le système réel : les valeurs réelles doivent être très proches des résultats obtenus avec le modèle utilisé.

L'objectif d'un modélisateur est d'obtenir un compromis entre ces deux notions. Les modèles mathématiques de base qui sont utilisés sont des systèmes à événements discrets, des processus stochastiques, et bien souvent des chaînes de Markov en raison de leurs bonnes propriétés (exploitables).

Nous avons travaillé sur ces abstractions. Cependant, il est souhaitable pour certains utilisateurs de disposer d'une description de plus haut niveau des systèmes, à partir de laquelle le modèle mathématique peut être algorithmiquement construit. Ainsi, pour certaines personnes méconnaîssant les outils mathématiques, une traduction automatique et transparente peut néanmoins leur permettre de les utiliser. De même, le niveau de complexité de certains systèmes est tel qu'une construction « manuelle » du modèle est extrêmement difficile et une description de plus haut niveau, traduite ensuite algorithmiquement, peut s'avérer extrêmement utile. Les files d'attente sont un exemple très utilisé, mais nous nous sommes aussi particulièrement intéressés aux réseaux de Petri, qui sont des graphes formels bien adaptés à la représentation des systèmes comprenant des problèmes de concurrence et de synchronisation.

1.2 Méthodes d'évaluation de performance et de sûreté de fonctionnement

Le modèle mathématique étant obtenu, se pose alors la question de son évaluation. Les mesures qui nous intéressent sont typiquement la fiabilité d'un système, sa disponibilité (dans le cas de la sûreté de fonctionnement), le taux de perte, le temps de réponse, ou le débit moyen dans le cadre de l'évaluation de performance.

Dans le cas idéal, l'évaluation peut être réalisée *analytiquement*, c'est à dire résolue exactement. Cependant ceci ne peut être obtenu que dans le cas de modèles relativement simples, avec des contraintes fortes (Markov...). De plus, il est important de noter que même quand ces hypothèses sont respectées, l'évaluation n'est pas toujours possible. C'est par exemple le cas des chaînes de Markov pour lesquelles la détermination de la mesure invariante nécessite la résolution d'un système d'équations linéaires de taille donnée par le nombre d'états, en général très grand, demandant un temps de calcul extrêmement important. Aussi, les mesures d'intérêt peuvent parfois s'exprimer sous forme d'une intégrale calculable directement.

Les techniques d'analyse numérique sont alors une alternative exploitable, comme l'utilisation de règles de quadratures pour le calcul d'intégrales, ou, encore à titre d'exemple la méthode de Gauss-Seidel dans le cas de la résolution de systèmes linéaires. Dans ces cas cependant, des hypothèses et contraintes fortes, bien que réduites par rapport au cas analytique, restent en général nécessaires.

La technique nécessitant le moins d'hypothèses est la *simulation*. Simulation est un mot utilisé, probablement excessivement, dans beaucoup de domaines. Nous la définissons ici en tant que technique d'évaluation utilisant des nombres aléatoires (souvent définie alors par *simulation Monte Carlo*). Cette technique d'évaluation a constitué un élément moteur de mon travail et sera donc mise en évidence dans ce document.

1.3 Une méthode d'évaluation particulière : la simulation

La simulation de type Monte Carlo est utilisée dans deux cadres différents :
- la simulation statique où on simule un modèle à un instant donné ;
- la simulation dynamique où il est nécessaire de simuler l'évolution du système au cours du temps.

Dans la plupart des cas, la simulation dynamique est dite à événements discrets, c'est à dire que seuls les instants de changements d'états sont simulés, le système restant non modifié entre ces instants. Il est à noter que la simulation de modèles fluides peut elle aussi être souvent réalisée par une simulation à événements discrets [187].

L'analyse des résultats de la simulation, correspondant à la donnée d'un intervalle de confiance pour les quantités à estimer, est basée sur le théorème de la limite centrale et ses dérivés. L'efficacité d'un estimateur est alors fonction de deux paramètres : la variance de l'estimateur, et le temps de calcul. Réduire ces paramètres permet d'améliorer l'efficacité de l'estimateur, et donc *d'accélérer la simulation*.

Jouer sur l'efficacité n'est pas seulement un confort, mais peut représenter une nécessité comme dans le cas de la simulation d'événements rares. En effet, si on cherche par exemple à valider une probabilité de perte de 10^{-9} pour un système, il faudra un échantillon de taille 10^9 pour obtenir en moyenne une fois l'événement, et bien plus pour obtenir un intervalle de confiance. Plusieurs méthodes d'accélération existent dans la littérature. Nous avons principalement travaillé sur les techniques d'échantillonnage préférentiel (*importance sampling*) qui modifient la loi de probabilité de l'échantillon, introduisant un biais qui doit être contrebalancé par le calcul de la fonction de vraisemblance, et sur les techniques de ramification de trajectoires (*importance splitting*) qui, d'une certaine manière, décomposent l'événement rare en une succession d'événements non rares : dès qu'un événement intermédiaire est atteint, la trajectoire est scindée en plusieurs nouvelles trajectoires tentant d'atteindre les événements suivants.

Une autre technique permettant d'aller plus vite que la méthode de Monte Carlo est celle dite de quasi-Monte Carlo. Cette technique ne consiste pas à utiliser des nombres aléatoires mais à utiliser une suite de points se répartissant « au mieux » sur l'espace d'états considéré et, grâce à ce meilleur balayage des possibilités, permettre une meilleure estimation des quantités recherchées. Très attractive a priori, cette technique se heurte à quelques difficultés : la détermination pratique de l'erreur et un champ d'application plus limité (car nécessitant quelques hypothèses non présentes dans les méthodes de Monte Carlo). Nous avons travaillé sur ces deux limitations afin d'en réduire la portée.

1.4 Tarification pour contrôler les systèmes de communication

Modéliser et évaluer les systèmes sont des problèmes cruciaux. Les contrôler apparaît tout aussi vital. Une manière de contrôler les services rendus par un système est de gérer la demande par notamment l'introduction d'un mécanisme de tarification. Considérons par exemple le cadre des réseaux de communication, comme le réseau Internet. Il est fort probable que la future génération de l'Internet sera encore plus consommatrice en ressources

que l'actuelle, conséquence entre autres de l'intégration des réseaux télévisés et téléphoniques. Dès lors, le système de tarification actuel basé sur un abonnement fixe, indépendant de l'utilisation, est une stimulation de la consommation qui, bien qu'ayant été très utile au démarrage du réseau, devient ingérable en cas de congestion si l'on souhaite faire de ce dernier un réseau multiservice efficace. De nombreuses théories mathématiques de tarification basées sur l'utilisation ont été récemment développées afin de satisfaire différents critères de qualité de service et de répondre à des règles d'utilisation équitables, elles aussi mathématiquement définies. Les outils mathématiques que nous utilisons pour développer nos modèles sont les techniques d'optimisation (pour déterminer les prix maximisant le revenu du vendeur ou le bien-être social), la théorie des files d'attente (pour déterminer les mesures de qualité de service perçues par les utilisateurs, régissant la demande) et la théorie des jeux non coopératifs (pour modéliser le comportement individualiste des utilisateurs, et dont les qualités de service sont interdépendantes, conduisant ainsi à un *équilibre de Nash*, s'il existe).

1.5 Logiciels/implémentations

Développer des méthodes théoriques n'a de sens réel que si elles sont implémentables et utilisables par nos pairs. Pour cela, nous avons implémenté les méthodes de modélisation et de simulation dans des logiciels et bibliothèques. Nous avons tout d'abord créé une bibliothèque C (QMCLIB) pour l'utilisation des suites à discrépance faible qui se retrouvent dans les méthodes de quasi-Monte Carlo. Nous avons aussi participé au développement du logiciel SPNP (pour *Stochastic Petri Net Package*) permettant la modélisation et l'évaluation de systèmes à l'aide de réseaux de Petri. Ce logiciel est distribué à la fois dans le monde académique et le monde industriel.

Il est aussi à noter que, dans notre activité sur la tarification des réseaux, nous commençons à étudier l'implémentation des mécanismes théoriques que nous avons développés et continuons de développer, via l'élaboration de protocoles et leur test sur une plate-forme dédiée à la tarification.

1.6 Structure du document

Le document est organisé de la manière suivante. Le chapitre 2 concerne nos travaux liés à la modélisation, principalement la modélisation par réseaux de Petri (suite aux travaux réalisés à Duke University) et à leur application, entre autres pour l'analyse des files à seuils. Nous discuterons aussi de la modélisation du protocole TCP/IP, et l'étude de l'influence de la stratégie de perte sur le débit moyen, sur sa variabilité et sur l'équité du mécanisme. D'autres travaux annexes seront brièvement décrits en note. Le chapitre 3 traitera d'une technique particulière d'analyse, la simulation, celle-ci ayant constitué le sujet principal de ma thèse [215] et étant toujours un de mes domaines privilégiés de recherche. Différents cas seront étudiés : tout d'abord nous verrons les différents types de techniques d'accélération ; ensuite nous nous intéresserons au cas particulier de la simulation d'événements rares ; enfin nous étudierons la simulation de type quasi-Monte Carlo. Dans le

chapitre 4, nous nous intéressons à la tarification comme moyen de contrôle dans les réseaux de communication. Différentes approches sont abordées : celle où plusieurs classes de services sont disponibles de manière à ce que les utilisateurs demandant une qualité plus importante payent plus cher, et celle où aucune classe n'est imposée mais où la différenciation se fait naturellement par l'intermédiaire d'un jeu. Le chapitre 5 traite des logiciels, bibliothèques et implémentations de nos résultats. Enfin, le chapitre 6 résume mes contributions et discute des voies de recherche futures qu'il me semble intéressant d'explorer.

Chapitre 2

Contributions à la modélisation pour l'évaluation de performances

La modélisation d'un système constitue la phase initiale, et essentielle, de son évaluation. Nous nous concentrons dans ce chapitre principalement sur la modélisation par réseau de Petri, description à partir de laquelle un modèle mathématique peut être automatiquement construit. Cependant, nous décrirons en fin de chapitre quelques travaux pour lesquels le modèle mathématique a été directement construit, avec une attention spéciale pour la modélisation du protocole TCP/IP, permettant d'évaluer l'influence de différentes stratégies de perte.

2.1 Réseaux de Petri : outil de modélisation

2.1.1 Définition et dynamique

Les réseaux de Petri stochastiques de différents types sont un paradigme puissant pour analyser les performances et la sûreté de fonctionnement de systèmes complexes (voir par exemple [3, 179]), notamment dans le cas de systèmes comportant des problèmes de synchronisation. Plus récemment, les réseaux de Petri stochastiques fluides [50] ont généralisé les réseaux de Petri stochastiques pour contourner le problème de l'explosion d'états, mais aussi pour gérer plus facilement des systèmes où un grand nombre d'événements intervient au cours de périodes très courtes (comme par exemple à un routeur d'un réseau).

Un réseau de Petri stochastique fluide est défini par un 13-uplet

$$(\mathcal{P}, \mathcal{T}, a, f, g, >, d, r, F, \omega, b, m^0, x^0)$$

où

– $\mathcal{P}$, l'ensemble des places, est partitionné en $\mathcal{P}_D$, l'ensemble de D places discrètes, et $\mathcal{P}_C$, l'ensemble de C places continues (ou fluides). Soit (x, m) désignant le mar-

quage du réseau de Petri, avec le vecteur x (resp. vecteur m) donnant le contenu des places fluides (resp. discrètes, qui seront dites contenir des jetons). Soit $\mathcal{S}$ l'ensemble des états possibles du réseau de Petri. Nous supposerons que le nombre total de jetons dans les places discrètes est borné.

- $\mathcal{T}$ est l'ensemble des transitions qui peut être partitionné en deux sous ensembles : $\mathcal{T}_T$, l'ensemble des transitions temporisées et $\mathcal{T}_I$, l'ensemble des transitions immédiates.
- a décrit la cardinalité (dépendante du marquage) ou l'impulsion fluide des arcs d'entrée et de sortie connectant les arcs et les places. La fonction a est définie de $((\mathcal{P}_D \times \mathcal{T}) \cup (\mathcal{T} \times \mathcal{P}_D)) \times \mathcal{S}$ dans $\mathbb{N}$ pour les places discrètes et de $((\mathcal{P}_C \times \mathcal{T}) \cup (\mathcal{T} \times \mathcal{P}_C)) \times \mathcal{S}$ dans $\mathbb{R}^+$ pour les places fluides.
- $f : ((\mathcal{P}_C \times \mathcal{T}) \cup (\mathcal{T} \times \mathcal{P}_C)) \times \mathcal{S} \rightarrow \mathbb{R}^+$ est le taux fluide (dépendant du marquage) pour les arcs connectant les transitions aux places fluides.
- $g : \mathcal{T} \times \mathcal{S} \rightarrow \{0, 1\}$ est la fonction de garde de chaque transition (mise par défaut à 1).
- $>$ définit la priorité pour le déclenchement des transitions.
- d définit la distribution du temps avant déclenchement de chaque transition. Elle peut dépendre du marquage. En pratique, chaque transition a sa propre horloge $c(t)$ activée quand la transition devient elle-même activée, et son temps de déclenchement est déterminé par la distribution donnée par d.
- r est la politique de ré-échantillonage statique pour chaque transition quand elle est désactivée par le déclenchement d'une transition concurrente et redevient activée plus tard : cela peut être par exemple PRI (répétition identique préemptive, où le temps de déclenchement est réinitialisé mais reste identique) PRD (répétition différente préemptive, où le temps est ré-échantillonné) ou PRS (reprise en cours, où le temps passé est pris en compte).
- F définit la politique de ré-échantillonage de chaque transition quand une autre transition se déclenche mais la transition considérée reste activée. PRI, RPS et PRD sont encore trois possibilités.
- $\omega : \mathcal{T} \times \mathcal{S} \rightarrow \mathbb{R}_*^+ = \mathbb{R}^+ - \{0\}$ est la fonction de poids sur les transitions.
- $b : \mathcal{P}_C \times \mathcal{S} \rightarrow \mathbb{R}_*^+$ est une borne pour chaque place continue.
- (x^0, m^0) est le marquage initial.

Une transition $t \in \mathcal{T}$ est active quand le marquage est (x, m) si

$$\forall p \in \mathcal{P}_D, a_{p,t}(x, m) \leq m_p \text{ et } g_t(x, m) = 1.$$

Décrivons maintenant la dynamique d'un réseau de Petri stochastique fluide quand le marquage est (x, m). Toutes les transitions actives ont un temps de déclenchement échantillonné selon la distribution d quand elles sont activées.

Du fluide coule de manière continue à travers les arcs des transitions actives dans ou à partir des places fluides. Le taux *potentiel* de changement de niveau dans le marquage (x, m) pour la place i est donné par

$$\delta_i^{pot}(x, m) = \sum_{t \in \mathcal{E}(x,m)} f_{t,i}(m, x) - f_{i,t}(x, m),$$

$f_{t,i}$ donnant le flux d'entrée (de la transition vers la place), et $f_{i,t}$ le taux de sortie. Cepen-

dant le taux de remplissage/vidage peut être différent si la place est pleine/vide ; le taux *réel* est alors

$$\frac{dx_i}{dt} = \begin{cases} 0 & \text{si } (x_i = 0 \text{ et } \delta_i^{pot}(x,m) \leq 0) \text{ ou} \\ & (x_i = b_i(m) \text{ et } \delta_i^{pot}(x,m) \geq 0) \\ \delta_i^{pot}(x,m) & \text{sinon,} \end{cases}$$

où $b_i(m)$ est la borne fluide pour la $i$ème place continue quand le marquage discret est m.

Après un temps donné par le minimum des temps de déclenchements des transitions actives, et selon le niveau de priorité $>$ si plusieurs sont déclenchables en même temps, la transition correspondante $t \in \mathcal{E}(x,m)$ est déclenchée, menant à un nouveau marquage (x', m') donné par

$$\begin{aligned} \forall p \in \mathcal{P}_D,\ m'_p &= m_p + a_{t,p}(x,m) - a_{p,t}(x,m) \\ \forall i \in \mathcal{P}_C,\ x'_i &= \min(b_i(m'), \max(0, x_i + a_{t,i}(x,m) - a_{i,t}(x,m))). \end{aligned} \quad (2.1)$$

Immédiatement après avoir atteint ce nouveau marquage, l'ensemble des transitions actives et les temps de déclenchements sont remis à jour selon $>$, d, r et F. Les transitions immédiates actives sont alors déclenchées. Quand au moins deux transitions immédiates sont actives en même temps (ou si deux transitions sont supposées se déclencher en même temps et ont même priorité), la transition u est déclenchée en premier avec la probabilité

$$\frac{\omega_u(x,m)}{\sum_{u'} \omega_{u'}(x,m)}.$$

Alors, le processus recommence : du fluide coule dans les places fluides, jusqu'au prochain déclenchement...

Dans le cas de transitions exponentielles et immédiates, le modèle sous-jacent est alors une chaîne de Markov. L'avantage est que la transformation de la représentation sous forme de réseau de Petri à celle de chaîne de Markov, peut être facilement réalisée automatiquement, alors qu'une détermination directe de la chaîne de Markov peut s'avérer très difficile. Une utilisation de tous les outils standards de résolution des chaînes de Markov est alors possible. Dans le cas où le système ne peut être représenté par une chaîne de Markov, des outils d'analyse existent encore dans certains cas particuliers, mais il faut la plupart du temps avoir recours aux méthodes de simulation. Nous décrirons en section 5.2 un logiciel d'analyse des réseaux de Petri, SPNP (*Stochastic Petri Net Package*) au développement duquel nous avons participé.

Un autre formalisme couramment utilisé est celui des système dits hybrides (voir par exemple [12, 13, 102, 181, 233]), aussi appelés automates hybrides, qui sont des modèles constitués d'«objets» digitaux ou continus interagissant entre eux, tels que les variables continues d'un objet sont contrôlées par un contrôleur digital. Ces modèles ont reçu beaucoup d'intérêt pour, notamment, le contrôle aérien, les réseaux de communication, les systèmes intelligents... Les questions auxquelles s'adressent habituellement ces modèles sont la stabilité, la contrôlabilité et l'observabilité. Par définition, un FSPN est un système hybride, puiqu'il intègre leur formalisme. Dès lors, on peut se demander quels types de systèmes hybrides peuvent être représentés par des FSPNs. Les réseaux de Petri ont certains avantages par rapport aux systèmes hybrides usuels de la théorie du contrôle, comme sa

disposition à gérer la concurrence, à réduire la complexité du modèle, et à générer automatiquement le processus stochastique sous-jacent à partir de la description concise de haut niveau. Comparer les deux types de modèles, comme réalisé dans [226], peut permettre d'utiliser les méthodes d'analyse d'un formalisme sur l'autre, avec les outils associés. La complémentarité des deux notions est mise en évidence. On peut noter que l'idée de comparer les réseaux de Petri et modèles d'automates d'états (non temporisés) a été étudiée dans [45, page 104] ; la conclusion était là aussi que le meilleur modèle dépend de l'application particulière considérée, et que les deux approches sont complémentaires.

2.1.2 Méthodes de résolution

Dans ce cadre des réseaux de Petri, il nous est apparu intéressant, afin d'aider un utilisateur dans le choix de la méthode de résolution, de comparer les différentes méthodes possibles. Dans [227], nous avons considéré l'exemple d'un système client/serveur, c'est à dire un système où un serveur reçoit des requêtes de stations clientes, sert les requêtes, et répond aux stations clientes [106]. Nous avons comparé les méthodes en faisant varier le nombre de stations connectées au serveur. Nous avons donc considéré un système distribué constitué de N stations de travail et d'un serveur de fichiers interconnectés par un réseau local. Nous avons comparé les méthodes analytiques-numériques avec les méthodes de simulation pour l'analyse stationnaire, ainsi que dans le cas transitoire (cumulé ou à un instant donné). Nous avons ainsi étudié la précision des résultats, mis en balance avec le temps de calcul nécessaire pour les obtenir. Ceci nous a permis d'illustrer les limites de chaque méthode. Afin, pour un utilisateur, de choisir entre les méthodes de simulation et les méthodes analytiques-numériques, nous avons proposé les recommandations suivantes :

- Quand le système n'est pas Markovien (particulièrement s'il n'a pas de structure régénérative), très peu de méthodes analytiques-numériques sont disponibles (on peut néanmoins citer [47, 48] par exemple où si pas plus d'une transition non exponentielle n'est possible dans chaque marquage, on obtient alors un processus de Markov régénératif pour lequel des outils d'analyse existent). La simulation est alors souvent la seule possibilité.
- Quand le système est Markovien
  - si l'espace d'états est grand (et qu'aucun modèle approché réduisant l'espace d'état n'est disponible), le graphe d'accessibilité ne peut être généré. La simulation est alors encore la seule possibilité (dans nos exemples, sur une station de travail Sun SparcStation Ultra 60 avec 640Mb de mémoire réelle et 982Mb de mémoire swap, nous sommes allés jusqu'à 6 millions d'états et 17 millions d'entrées non nulles dans le générateur infinitésimal).
  - Dans les autres cas, le choix dépend des spécifications de l'application. Néanmoins, de notre expérience et de l'exemple traité dans [227], nous pouvons montrer que pour de « petits » espaces d'états, les méthodes analytiques-numériques sont performantes. Quand l'espace d'états devient plus grand, il existe toujours un seuil à partir duquel le temps de simulation devient bien plus efficace (le cas extrême est celui où le graphe d'accessibilité est trop grand pour être généré). Nous pouvons souligné que le seuil naturel entre méthodes analytiques-numériques et simulation est plus rapide pour l'étude du comportement stationnaire, puis le com-

portement transitoire cumulatif, et enfin le comportement transitoire instantané (excepté pour des horizons petits). Habituellement, la simulation requiert un temps très long pour obtenir de bons résultats dans le cas d'horizons éloignés.
Notons que même si le temps de calcul pour les méthodes analytiques-numériques est un peu plus grand que celui de la simulation, il est toujours pertinent de les utiliser car elles donnent un résultats précis au lieu d'un intervalle de confiance. La différence pour passer de l'une à l'autre est subjective et dépend de l'utilisateur. De plus, il est important de noter que l'utilisation de modèles approchés s'avère très intéressante même quand on utilise la simulation, ce qui est pourtant rarement fait en pratique. En effet, même si l'espace d'états n'est pas généré comme dans les méthodes analytiques-numériques, un modèle approché peut considérablement réduire le temps de calcul par itération en réduisant le nombre d'événements à traiter.

2.2 Modélisation des files à seuils

Nous présentons dans cette section des résultats obtenus dans [120, 119, 229] et qui traitent de files d'attente dont le comportement est régulé grâce à l'introduction de seuils avec hystérésis.

2.2.1 Files monoclasses à seuils et hystérésis

Dans [120], nous avons étudié une file d'attente à K serveurs ($K \geq 2$) permettant de contrôler la congestion à l'aide de seuils avec hystérésis. Un ensemble de $K-1$ seuils ascendants $(n_1, n_2, .., n_{K-1})$ et un autre ensemble de seuils descendants $(m_1, m_2, .., m_{K-1})$ sont définis comme suit : pour $i = 1, ..., K-1$, quand le nombre de clients du système dépasse un seuil ascendant n_i, un serveur supplémentaire est immédiatement ajouté. De même, quand le nombre de clients devient inférieur au seuil descendant m_i ($m_i < n_i$), un serveur est aussitôt enlevé. Ce type de file d'attente a des applications potentielles dans les problèmes de routages dynamiques dans les réseaux de communication pour le contrôle de congestion. Notre modèle suppose des arrivées Poissonniennes et des serveurs hétérogènes dont les temps de service sont exponentiels. Pour indication la figure 2.1 présente l'espace d'états où (n, i) signifie que n clients sont présents dans la file, et i serveurs sont actifs. Nous avons proposé dans [120] une méthode basée sur la notion de coupe pour le calcul sous forme close des probabilités stationnaires d'état. La formule obtenue avait déjà été trouvée dans le cas de serveurs homogènes mais seulement pour 2 serveurs dans le cas hétérogène [107]. De plus, la méthode proposée présente l'avantage de demander peu de développements mathématiques par rapport à [107].

Dans [229] nous avons considéré une extension de ces travaux en considérant des lois générales pour les inter-arrivées et les services. De même les temps d'attente d'un nouveau serveur ainsi que de sa libération ont des lois de probabilité non nécessairement exponentielles. Nous avons choisi de modéliser et d'analyser ce système par les réseaux de Petri stochastiques (SPNs) précédemment décrits, qui donnent une représentation fidèle et néanmoins performante des systèmes complexes. Nous avons montré que la représentation est

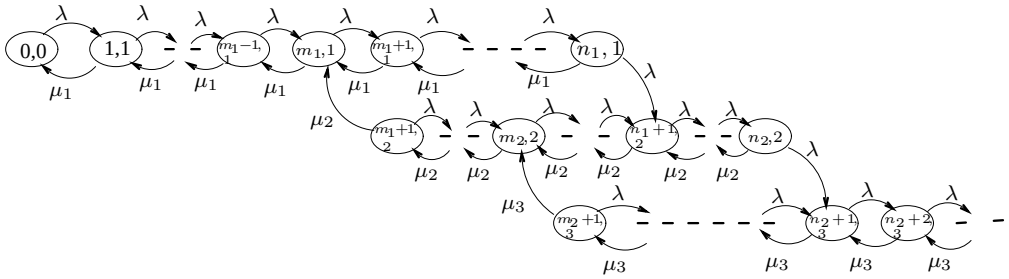


FIG. 2.1 – File à seuil et hystéresis pour $K = 3$ seuils

alors très simple et permet d'étudier et de configurer la file afin de la conformer à des spécifications données. De plus, les réseaux de Petri stochastiques fluides (FSPNs) constituent une approximation efficace pour l'analyse des réseaux haut débit. Plusieurs illustrations numériques sont données, attestant de l'intérêt des files à seuils avec hystérésis et de l'utilité de leur représentation par des SPNs et FSPNs.

2.2.2 Files multiclasses à seuils et hystérésis

Dans [119] nous avons généralisé l'étude précédente au cas multiclasse. Ce cadre a suscité beaucoup d'intérêt dans la litérature. Ainsi, dans [128], différentes politiques d'ordonnancement dans un commutateur ATM sont étudiées. Dans [15], des politiques à seuils sont utilisées dans des réseaux de files d'attente à un serveur dans le but d'implémenter différentes disciplines de service. Les paquets temps réel tels que la voix ont une priorité supérieure aux autres tels que les données. Des politiques de pure priorité étant très pénalisantes en terme de qualité de service pour les paquets de priorité inférieure, l'introduction de seuils permet de satisfaire les exigences du trafic prioritaire tout en assurant des performances acceptables aux trafics non prioritaires. Le même type de processus peut aussi modéliser le protocole de signalisation SS $n^o 7$ en incluant une politique de priorité fournissant une meilleure qualité de service à certaines classes [46]. Une étude quantitative similaire pour des services différenciés dans l'Internet (DiffServ) est proposée dans [191]. Tous ces travaux montrent que les méthodes analytiques deviennent rapidement complexes ; nous avons donc choisi dans [119] une autre voie consistant à utiliser les réseaux de Petri stochastiques. Ici encore, la description du modèle est particulièrement simple et la traduction mathématique automatisée. Les conclusions sont prometteuses quant à l'utilisation de cette technique, puisque nous avons pu analyser les cas markovien, non markovien, avec délai pour l'activation des serveurs, et fluide pour différentes politiques de service entre les classes (priorités strictes, différents types de seuils...). Grâce à l'utilisation d'une fonction coût, il a été possible de sélectionner les meilleures politiques en fonction de différents paramètres tels que le débit la proportion des clients de chaque classe, l'intensité du trafic, etc...

2.3 Modélisation des flux TCP/IP par des processus AIMD

Un domaine d'étude important dans les télécommunications est la modélisation, l'analyse, voire l'amélioration du protocole de transmission TCP/IP qui dirige la majeure partie du trafic Internet. En résumé, toute session envoie les paquets de plus en plus rapidement (selon une augmentation linéraire) tant qu'aucune perte n'est perçue ; dès qu'une perte est détectée, le débit est diminué par un facteur multiplicatif. Certains modèles s'intéressent à une session soumise à un processus de perte exogène [174], mais cette approche suppose une grande quantité de trafic. Une autre approche étudie plusieurs connections partageant un goulot d'étranglement mais suppose l'hypothèse simplificatrice que toutes les connections réduisent simultanément leur fenêtre quand il y a congestion [2, 35, 115], cependant cette hypothèse n'est pas valide pour des connections asymétriques.

Nous avons étudié dans [9, 10] pour deux connections, et généralisé dans [8] pour N connections l'influence de différentes stratégies de perte à un goulot d'étranglement sur des connections TCP modélisées par un processus AIMD (*Additive Increase Multiplicative Decrease*). Notre modèle, basé sur celui présenté dans [21], est le suivant. On considère N sessions TCP en compétition pour la bande passante à un routeur de capacité C. On note $\eta^{(i)}$ le taux de croissance additif pour la session i et $\beta^{(i)}$ son taux de décroissance multiplicatif. Habituellement, $\beta^{(i)} = 1/2\ \forall i$ et pour TCP $\eta^{(i)}$ est l'inverse du carré du temps d'aller-retour (RTT) pour la session i. Soient $Y^{(i)}(t)$ et $Y_n^{(i)}$ les débits de la session i à l'instant t et au n-ème instant de congestion T_n respectivement, et $p_n^{(i)}$ la probabilité qu'elle expérimente une perte au n-ème instant de congestion, de sorte que

$$Y^{(i)}(t) = Y_n^{(i)} + \eta^{(i)}(t - T_n)$$

pour $T_n \leq t < T_{n+1}$. Nous supposons de plus qu'une perte est observée dès que la capacité du routeur est atteinte (c'est à dire qu'on suppose que le routeur n'a pas de tampon).

Nous avons montré le résultat surprenant suivant dans le cas de connections symétriques ($\eta^{(i)} = \eta$ et $\beta^{(i)} = \beta\ \forall i$) :

Théorème 1 *Pour N connections symétriques en compétition telles que exactement une connection expérimente une perte aux instants de connection, le débit moyen des sessions est indépendant de la stratégie de perte adoptée (c'est à dire des $p_n^{(i)}$) :*

$$\forall i\ \ \bar{Y}^{(i)} = \frac{1+\beta}{(N+1)+(N-1)\beta} C,$$

de sorte que l'utilisation moyenne du routeur est $N\frac{1+\beta}{(N+1)+(N-1)\beta}C$.

Ce résultat est démontré dans [8] pour N connections et dans [9, 10] pour deux connections. De plus, alors que la preuve dans [9, 10] est pour le cas où les probabilités de pertes dépendent des flux aux instants de congestion uniquement, aucune hypothèse similaire n'est faite pour la généralisation obtenue dans [8].

Cependant nous avons pu montrer que les stratégies de perte ont une influence sur la variabilité du débit (le moment d'ordre 2). Avoir un débit stable permet une constance dans le service. Nous nous sommes intéressés à quatre stratégies de perte particulières :

– *sbd* : un flux précis est systématiquement sélectionné jusqu'à ce que les autres flux occupent toute la capacité ; un de ces autres flux est ensuite systématiquement sélectionné, et ainsi de suite ;
– *cst* : le cas où la probabilité de perte est fixe pour chaque session (indépendante du débit de la session aux instants de congestion) ;
– *prop* : le cas où la probabilité de perte est proportionnelle au débit aux instants de congestion ($p_n^{(i)} = Y_n^{(i)}/C$) ;
– *ltl* : le cas où la session ayant le plus grand débit aux instants de congestion expérimente nécessairement la perte.

Nous avons alors calculé la variabilité du débit comme l'espérance de la somme cumulée du carré du débit entre deux instants de congestion $\mathbb{E}[S_2]$ divisée par le temps moyen entre deux instants de congestion $\mathbb{E}[\tau]$, et obtenu le résultat suivant, dans le cas de deux sessions symétriques en compétition :

Proposition 1 *Considérons 2 sessions symétriques en compétition et la valeur standard $\beta = 1/2$. Si on utilise la stratégie qui sélectionne un flux jusqu'à ce que l'autre atteigne la totalité de la capacité, alors*

$$Q_{sbd} = \frac{\mathbb{E}[S_2]}{\mathbb{E}[\tau]} = \frac{1}{3}C^2 \approx 0.3333C^2;$$

si on utilise la stratégie de probabilité de perte constante, alors

$$Q_{cst} = \frac{\mathbb{E}[S_2]}{\mathbb{E}[\tau]} = \frac{5}{24}C^2 \approx 0.20833C^2;$$

si on utilise la stratégie proportionnelle

$$Q_{pro} = \frac{\mathbb{E}[S_2]}{\mathbb{E}[\tau]} = \frac{679}{3396}C^2 \approx 0.19994C^2;$$

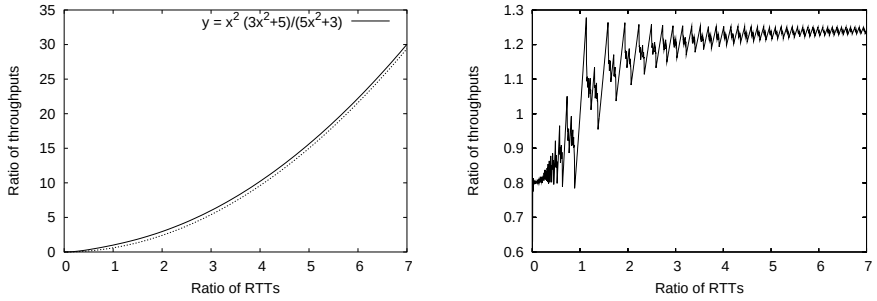
alors que si on utilise la stratégie de perte au plus grand débit

$$Q_{ltl} = \frac{\mathbb{E}[S_2]}{\mathbb{E}[\tau]} = \frac{4}{21}C^2 \approx 0.1905C^2.$$

Pour ces exemples, la stratégie *ltl* donne donc ici les résultats les plus probants, et la stratégie *sbd* les plus mauvais. Ces résultats vérifient l'intuition car plus les politiques cherchent à équilibrer les flux (en pénalisant les plus gros), plus la variance est petite.

Nous avons aussi pu mettre en évidence que, sous l'hypothèse qu'exactement une connection subit une perte aux instants de congestion, le débit moyen aggrégé converge rapidement vers 1 (comme en témoigne la formule du théorème précédent). Ainsi, l'option de décomposer le trafic d'un utilisateur en plusieurs flux TCP parallèles est valide, mais il est ici possible de déterminer un seuil pour le nombre de flux à utiliser au delà duquel le gain sera dérisoire (voir [8]). Implémenter une telle stratégie d'unique perte semble raisonnable dans de nombreux contextes, comme en home-networking par exemple.

Nous nous sommes aussi intéressés au cas assymétrique plus courant en pratique. Nous avons ainsi pu obtenir une forme close du débit moyen dans le cas de la stratégie de probabilité de perte constante (avec une probabilité différente par flux) dans le cas de deux



(a) Stratégie de pertes constantes (b) Stratégie *ltl*

FIG. 2.2 – Quotient des débits moyens $\bar{Y}^{(1)}/\bar{Y}^{(2)}$ en fonction du quotient des temps d'aller retour $R^{(2)}/R^{(1)}$

sessions [10]. Etudier ce cas était initialement motivé par l'idée de fixer les probabilités de perte différente par session en fonction d'une méthode de tarification. Nous avons aussi étudié l'équité entre deux sessions hétérogènes caractérisées par des temps d'aller-retour différents $RTT^{(1)}$ et $RTT^{(2)}$ (mesure de distance entre origine et destination) donnant des taux additifs $\eta^{(1)}$ et $\eta^{(2)}$ différents. Les figures 2.2(a) et 2.2(b) donnent le quotient des débits pour deux connections en fonctions du quotient des temps d'aller-retour pour les stratégies *cst* et *ltl*. Conserver ce quotient le plus proche possible de 1 est souhaitable/juste. On peut ainsi observer que le partage du débit est bien plus équitable pour la stratégie *ltl* que la stratégie *cst*, car bien moins sensible aux différences en termes de RTT (noter la différence d'échelle entre les deux courbes). Par exemple, une connection avec un RTT 3 fois plus petit aura seulement 1,21 fois plus de débit sous la stratégie *ltl* alors qu'il en aura 6 fois plus avec la stratégie *cst*. Le cas de la stratégie *prop* avait été étudié dans [5] : pour le même exemple la connection avec un RTT 3 fois plus petit obtient un débit 2,75 fois plus grand. La stratégie *ltl* s'avère donc aussi la plus efficace en terme d'équité. Nous avons aussi pu prouver que si le quotient de RTTs tend vers l'infini, le quotient des débits tend vers 5/4, d'où une grande efficacité.

En conclusion, même si le débit moyen (dans le cas symétrique) est indépendant de la stratégie de perte adoptée, la variabilité de ce débit, ainsi que l'équité, en dépendent. La stratégie *ltl* est prouvée être la plus efficace sur ces deux métriques. L'implémentation de cette stratégie en pratique (plus complexe que la stratégie proportionnelle) est donc à étudier.

2.4 Notes sur la modélisation

Soulignons que nous avons également travaillé sur la modélisation et l'analyse de files d'attentes fluides dans [199], où nous avons développé un algorithme approché, contrôlable à une erreur spécifiable à l'avance, de calcul de la distribution stationnaire d'un tampon fluide dirigé par une chaîne de Markov.

Enfin, nous avons, au cours de travaux de valorisation industrielle, collaboré avec le

CELAR (direction générale de l'armement) sur la modélisation de briques de sécurité pour les réseaux haut débit (gigabit ou térabit), ainsi qu'avec Cril Telecom Software sur l'analyse de la couverture et le positionnement des stations de base dans les réseaux UMTS. Nous avons également décrit dans [54] un formalisme qui, étant à la fois basé sur l'analyse fonctionnelle et sur un langage dynamique de type State Chart / State Flow, permet d'obtenir un modèle très réaliste, reproduisant en tout point le comportement du système réel et qui laisse à l'utilisateur la possibilité de paramétrer entièrement les transitions entres les états (fonctionnels ou de pannes) des composants du système. Ceci permet de retrouver au sein du même modèle des comportements statiques ou dynamiques, déterministes ou aléatoires.

Chapitre 3

Contributions à la simulation de systèmes

3.1 Généralités

Le chapitre précédent traitait de la modélisation de systèmes et de la recherche, via des formules explicites voire via des méthodes d'analyse numérique déterministes, de mesures de performance de ces systèmes. Le champ d'application de ces méthodes analytiques-numériques est cependant relativement limité : ainsi, si le modèle est représenté par un processus stochastique (qui constitue d'ailleurs le cas général de nos études), des hypothèses de processus Markovien, semi-Markovien ou au minimum de régénération sont la plupart du temps nécessaires (c'est à dire des propriétés d'indépendance stochastique). Le cas où les techniques sont les plus développées est celui des processus de Markov. L'argument le plus fort pour leur utilisation est qu'ils peuvent approcher n'importe quel processus, en augmentant l'espace d'états (via l'utilisation de lois de type PH). Ceci conduit à la deuxième limitation majeure des méthodes précédemment citées : leurs performances sont limitées par la taille de l'espace d'états étudié et/ou la dimension mathématique du problème. Par exemple, les chaînes de Markov nécessitent de travailler sur des matrices dont l'ordre est la taille de l'espace d'états, donc très rapidement irréalisable en un temps de calcul « raisonnable ». Dans certains cas, la décomposition modulaire et la représentation tensorielle compacte permettent cependant de repousser un peu plus les limites des méthodes.

Dès lors que l'une des limitations n'est plus vérifiée, les méthodes de simulation s'avèrent souvent les seules utilisables. La simulation standard est en général relativement simple à utiliser, car elle consiste à mimer le comportement dynamique des systèmes. Par simulation, nous voulons dire « simulation Monte Carlo », c'est à dire simulation utilisant des nombres aléatoires. Cette définition est plus restrictive que le sens commun du terme simulation qui est « action de faire paraître comme réelle une chose qui ne l'est pas » ou, plus techniquement, « méthode de mesure et d'étude consistant à remplacer un système à étudier par un modèle ayant un comportement analogue». En effet, la technique de mesure utilise ici des nombres aléatoires. Notre travail a consisté à « accélérer » la simulation, c'est à dire à obtenir une meilleure précision pour un même temps de simulation.

On peut regrouper les modèles à simuler en différentes classes, pour lesquelles des méthodes spécifiques existent. Tout d'abord, il faudra distinguer les modèles à événements discrets, pour lesquels les dates de changement d'état forment une suite finie ou dénombrable, aux modèles continus, pour lesquels l'état peut évoluer de manière continue dans le temps. De même, on peut opposer les modèles statiques, où le temps ne joue pas de rôle, aux modèles dynamiques, où l'on considère explicitement une évolution au cours du temps.

En règle générale, modèles statiques et dynamiques peuvent être regroupés dans le cadre général du calcul d'une intégrale de la forme

$$\mathcal{I} = \int_{[0,1]^s} f(x)dx.$$

Par exemple, toute espérance mathématique qui peut être estimée par simulation peut être écrite sous cette forme, s étant potentiellement infini comme par exemple dans le cas de simulation de chemins sur un processus stochastique jusqu'à un temps d'arrêt non borné. L'intérêt de cette formulation est que la source (générateur) d'aléatoire est souvent une suite de nombres visant à imiter des variables aléatoires indépendantes uniformément distribuées sur $[0,1)$. Ces nombres sont alors transformés pour obtenir l'estimateur. La dimension s est donc le nombre d'appels au générateur. Le choix d'un tel générateur est donc capital et constitue une branche importante de la recherche en simulation (voir par exemple [71, 123]). Cependant notre travail a porté sur l'utilisation efficace, et non la création, de tels générateurs.

La méthode de Monte Carlo standard est la suivante. Soit $(X^{(i)})_{1\leq i\leq I}$ une suite finie de I vecteurs aléatoires indépendants uniformément distribués sur $[0,1)^s$. On sait, par la loi des grands nombres, qu'un estimateur sans biais de $\mathcal{I}$ est

$$\bar{f}_I = \frac{1}{I}\sum_{i=1}^{I} f(X^{(i)}).$$

La variance de l'estimateur $\bar{f}_I$ est alors σ^2/I ; σ^2 étant la variance de $f(X)$ avec X uniformément distribué sur $[0,1)^s$. Le théorème de la limite centrale nous donne une idée de la distribution de cet estimateur en fonction de I : on sait que

$$\sqrt{I}\frac{\bar{f}_I - \mathcal{I}}{\sigma}$$

a approximativement pour distribution la loi normale centrée réduite. Ceci nous permet d'obtenir un intervalle de confiance pour $\mathcal{I}$:

$$\mathcal{I} \in \left[\frac{1}{I}\sum_{i=1}^{I} f(X^{(i)}) - \frac{c_\alpha\sigma}{\sqrt{I}}, \frac{1}{I}\sum_{i=1}^{I} f(X^{(i)}) + \frac{c_\alpha\sigma}{\sqrt{I}}\right]$$

au risque α, où $c_\alpha = \Phi^{-1}(1-\frac{\alpha}{2})$ et Φ est la fonction de répartition de la loi normale centrée réduite. La vitesse de convergence d'une telle méthode est donc, en moyenne, en $O(I^{-1/2})$, indépendamment de la dimension s du problème. L'efficacité de l'estimateur $\bar{f}_I$ est alors $I/(t\sigma^2)$ où t est le temps de calcul pour obtenir $\bar{f}_I$.

3.2 Simulation Monte Carlo : réduction de la variance

Améliorer l'efficacité d'un estimateur signifie obtenir la même précision (i.e., largeur de l'intervalle de confiance) en un temps plus court. Il y a deux possibilités pour cela.

- On peut tout d'abord chercher à diminuer le temps de calcul d'une réplication, de manière à générer plus de variables en un temps donné.
- On peut aussi chercher à réduire la variance de l'estimateur.

Différentes techniques de réduction de la variance existent dans la littérature [71, 88]. Les principales techniques sont :

- l'utilisation de variables de contrôle. La variance est la moyenne du carré des écarts de la variable $f(X)$ par rapport à son espérance. Si on dispose d'une fonction g suffisamment simple pour être intégrée théoriquement et qui reproduit et absorbe la plupart des variations de f, l'intégrale de f peut être exprimée comme celle de g plus celle de la différence entre f et g, mais la variance devient alors celle, plus petite, de $(f - g)(X)$.
- L'utilisation de variables antithétiques. Ici il s'agit de chercher un second estimateur ayant la même espérance mais ayant une forte corrélation négative avec l'estimateur standard. En considérant la moyenne des deux estimateurs, on obtient ainsi une estimation non biaisée, et avec une très faible variance. Nous avons pu appliquer cette technique dans [216] à la simulation des files d'attente multi classes à forme produit et démontrer que son utilisation permet de systématiquement obtenir une réduction de la variance.
- L'échantillonnage stratifié. Il s'agit de partitionner le domaine d'intégration en sous domaines, et d'appliquer un échantillonnage Monte Carlo standard à chacun des sous domaines. Si la proportion de l'échantillonnage total réservé à chaque sous domaine est égale à la proportion du volume de ce sous domaine par rapport au volume total, nous sommes sûrs d'obtenir une réduction de la variance.
- L'échantillonnage préférentiel. Cette méthode consiste à modifier la loi des variables aléatoires échantillonnées. Dès lors, il ne s'agit plus d'estimer la fonction f de la nouvelle variable, mais f multipliée par une fonction appelée quotient de vraisemblance introduite pour éliminer le biais (plus exactement la distorsion) résultant du changement de variable. L'objectif de l'échantillonnage préférentiel est de concentrer la distribution de l'échantillon aux endroits « d'importance » pour l'évaluation de l'intégrale plutôt qu'uniformément.
  De manière générale, nous avons appliqué cette technique à la simulation des réseaux de Petri [232].

D'autres méthodes, adaptées à des problèmes très spécifiques, sont difficiles à expliquer dans un cadre général. La technique de ramification des trajectoires par exemple est spécifiquement adaptée à l'estimation d'événements rares. Nous la présenterons dans la section 3.3.3.

3.3 Simulation d'événements rares

Les événements rares interviennent dans de nombreux domaines. Si on considère par exemple un réseau de communication, on peut être intéressé par l'estimation de la proba-

bilité de perte à un routeur, ou la probabilité que deux sites terminaux d'un réseau soient connectés. Habituellement, ces probabilités sont très petites (souvent inférieures à 10^{-9}). Dans ces situations, les méthodes de simulation directes rencontrent des difficultés majeurs, puisque la faible probabilité de l'événement considéré rendent improbable son observation en pratique, conduisant à une mauvaise précision de l'estimation. Le développement de méthodes alternatives est donc un défi majeur. Cette section se décompose donc en trois parties : dans la première, nous étudions les différentes propriétés qu'un tel estimateur doit vérifier. La seconde se focalise sur un type de problème particulier : la simulation des systèmes Markoviens hautement fiables par échantillonnage préférentiel. Enfin dans la troisième, nous étudions les méthodes dites de ramification des trajectoires.

Rappelons qu'une fonction f est dite $o(\varepsilon^d)$ si $f(\varepsilon)/\varepsilon^d \to 0$ quand $\varepsilon \to 0$. Similairement $f(\varepsilon) = O(\varepsilon^d)$ si $|f(\varepsilon)| \leq c_1\varepsilon^d$ pour une constante $c_1 > 0$ et pour tout ε suffisamment petit. Elle est dite $\underline{O}(\varepsilon^d)$ si $|f(\varepsilon)| \geq c_2\varepsilon^d$ pour une constante $c_2 > 0$ et pour tout ε suffisamment petit. Finalement, $f(\varepsilon) = \Theta(\varepsilon^d)$ si $f(\varepsilon) = \underline{O}(\varepsilon^d)$ et $f(\varepsilon) = O(\varepsilon^d)$.

3.3.1 Propriétés des estimateurs d'événements rares

Considérons en toute généralité l'estimation de la probabilité γ d'un événement rare. La rareté est charactérisée par l'introduction d'un paramètre ε tel que $\gamma \to 0$ quand $\varepsilon \to 0$. Pour les modèles de fiabilité par exemple, ε peut représenter le taux de panne maximum d'un composant [200]. Pour les modèles de files d'attente, on peut considérer $\varepsilon = 1/B$ où B est la taille du tampon, de sorte que la probabilité de perte $\gamma \to 0$ quand $\varepsilon \to 0$ [99].

Considérons un estimateur sans biais $\hat{\gamma}$ de γ obtenu pour un échantillon de taille n. L'erreur relative bornée (ErrRB) a été définie par P. Shahabuddin dans [200].

Définition 1 *Soit σ_n^2 la variance de $\hat{\gamma}$ pour un échantillon de taille$n$ et soit z_δ le quantile d'ordre $1 - \delta/2$ de la loi normale centrée réduite. L'erreur relative ErrR est définie par*

$$ErrR = z_\delta \frac{\sqrt{\sigma_n^2}}{\gamma}. \tag{3.1}$$

On dit dit qu'il y a erreur relative bornée (ErrRB) si ErrB reste borné quand $\varepsilon \to 0$.

Ce concept est important car il assure que, quelle que soit la fiabilité du système, la largeur de l'intervalle de confiance restera de taille relative mesurée.

Une autre notion de robustesse, appelée *optimalité asymptotique*, a été très utilisée pour l'échantillonnage préférentiel appliqué aux files d'attente. Rappelons que l'échantillonnage préférentiel consiste à modifier la mesure de probabilité du système : si $\gamma = E_f[g(X)] = \int g(x)f(x)dx$, alors $\gamma = \int g(x)L(x)f^*(x)dx = E_{f^*}[g(X)L(X)]$ avec $L(x) = f(x)/f^*(x)$ quotient de vraisemblance (en supposant $f^* > 0$ si $fg > 0$). Il s'agit donc d'échantillonner selon f^* et de moyenner les valeurs obtenues à la fonction gL.

Définition 2 *Un estimateur utilisant l'échantillonnage préférentiel $\hat{\gamma}_{IS}$ est dit asymptotiquement optimal si*

$$\lim_{\varepsilon \to 0} \frac{\ln E_{f^*}[g(X)^2 L(X)^2]}{\ln \gamma} = 2.$$

On peut noter que cette quantité est toujours inférieure ou égale à 2.

Les relations entre optimalité asymptotique et ErrRB ont été étudiées par Sandmann [193], montrant que l'ErrRB est une notion plus forte que l'optimalité asymptotique. Elle est malheureusement dans certains cas bien plus difficile à obtenir.

Cependant, ces mesures de robustesse ne s'intéressent qu'à la taille relative de l'intervalle de confiance, mais pas à la validité de ce même intervalle, via l'évolution de son niveau de confiance quand $\varepsilon \to 0$. Nous avons défini l'approximation normale bornée (ANB) [218] qui assure que l'approximation de la loi normale, et donc le niveau de confiance résultant quand on applique le théorème central limite, reste valide quand ε tend vers 0. Cette définition se base sur la borne d'approximation de la loi normale :

Théorème 2 (Berry-Esseen) *[24] Soit $\rho = E(|X - E(X)|^3)$, $\sigma^2 = E((X - E(X))^2)$ et $\mathcal{N}(x)$ la distribution normale standard (i.e., espérance 0 et variance 1). Pour $X_1, \cdots, X_I$ I copies i.i.d. de X, posons $\bar{X}_I = I^{-1}\sum_{i=1}^{I} X_i$, $\hat{\sigma}_I^2 = I^{-1}\sum_{i=1}^{I}(X_i - \bar{X}_I)^2$ et F_I la distribution de la somme centrée et normalisée $(X_1 + \cdots + X_I)/(\hat{\sigma}_I\sqrt{I}) - E(X)\sqrt{I}/\hat{\sigma}_I$. Alors il existe une constante absolue $c > 0$ telle que, pour tout x et I,*

$$|F_I(x) - \mathcal{N}(x)| \leq \frac{c\rho}{\sigma^3\sqrt{I}}.$$

On définit alors l'approximation normale bornée comme suit.

Définition 3 *On dit que $\hat{\gamma}$ vérifie l'approximation normale bornée (ANB) si ϱ/σ^3 reste borné quand $\varepsilon \to 0$.*

Cette propriété garantit que l'approximation de la loi normale, et donc la validité de l'intervalle de confiance, reste bornée quand la fiabilité augmente. Il faut cependant noter que cette condition est *suffisante* mais pas nécessaire (d'autant que la distance est toujours bornée par 1). Dans [218], nous avons discuté à partir du développement d'Edgeworth de l'aspect nécessaire de la condition.

Dans [42], nous avons mis en évidence que l'ErrRB et l'ANB pouvaient s'avérer insuffisantes pour caractériser la robustesse d'un estimateur. En effet, ces propriétés s'intéressent à des tailles d'échantillon fixées, mais il existe des estimateurs pour lesquels ce n'est pas la variance qui diminue avec ε, mais le temps moyen d'exécution pour une réplication. C'est pourquoi nous avons généralisé les propriétés d'erreur relative bornée et approximation normale bornée comme suit.

L'efficacité relative bornée (EffRB) généralise l'ErrRB en étudiant la variance moyenne pour un temps de simulation donné plutôt que pour un numbre de réplications donné.

Définition 4 *Soit $\hat{\gamma}$ un estimateur de γ, σ_n^2 sa variance pour n réplications (pouvant être dépendantes). Soit t_n le temps moyen pour obtenir ces n réplications. L'efficacité relative de $\hat{\gamma}$ est*

$$\mathit{EffR} = \frac{\gamma^2}{\sigma_n^2 t_n}.$$

On dit que $\hat{\gamma}$ vérifie l'efficacité relative bornée (EffRB) s'il existe $d > 0$ tel que EffR soit minoré par d pour tout ε.

De manière similaire à l'ErrRB, on peut généraliser l'ANB de manière à prendre en compte le temps moyen de simulation par réplication [43] :

Définition 5 *Définissons $n(T)$ comme le nombre moyen de réplications obtenus en un temps de simulation T. On dit que $\hat{\gamma}$ vérifie l'approximation normale bornée généralisée (ANBG) si $\varrho/(\sigma^3\sqrt{n(T)})$ reste borné quand $\varepsilon \to 0$, ou, de manière équivalente, si $\varrho\sqrt{t_1}/\sigma^3$ reste borné quand $\varepsilon \to 0$ puisque $T = n(T)t_1$.*

En d'autres termes, si l'ANBG est vérifiée, le niveau de confiance de l'intervalle est robuste quand $\varepsilon \to 0$, *en considérant un temps de simulation donné*.

L'ANBG considère *une borne supérieure* de la distance entre la loi empirique et la loi normale (et donc une condition suffisante, mais non nécessaire, de couverture). Une manière de contrôler la validité réelle de l'intervalle est d'utiliser la fonction de couverture, introduite dans l'article de L.W. Schruben [196]. Supposons en toute généralité qu'on construit un intervalle de confiance $R(\eta, \boldsymbol{X})$ ipour l'estimation de γ, au niveau de confiance η et avec les données (aléatoires) $\boldsymbol{X}$. Si l'intervalle est bien construit, on doit avoir $\Pr[\gamma \in R(\eta, \boldsymbol{X})] = \eta$ et si on définit $\eta^* = \inf\{\eta \in [0,1] : \gamma \in R(\eta, \boldsymbol{X})\}$, η^* devrait être uniformément distribué sur [0,1] : $F_{\eta^*}(\eta) = \Pr[\eta^* \leq \eta] = \eta$. Pour chaque niveau de confiance *souhaité* η, notons $F_{\eta^*}(\eta)$ la couverture (ou niveau de confiance) *réelle*. Une évaluation pratique de cette fonction de couverture pour $n(T)$ replications peut être intéressante. Ceci peut être obtenu en considérant I données indépendantes $\boldsymbol{X}_i$ ($1 \leq i \leq I$) et en calculant les valeurs correspondantes η_i^*. À partir de ces valeurs, une estimation empirique de la distribution de η^* est construite.

Une discussion autour de ces différentes notions de robustesse a été réalisée dans [43] sur un modèle statique pour l'évaluation de la fiabilité d'un réseau de communication.

3.3.2 Echantillonnage préférentiel pour la simulation des systèmes Markoviens hautement fiables

Nous nous intéressons dans cette section à un type de problème particulier, la simulation des systèmes Markoviens hautement fiables par échantillonnage préférentiel. Les systèmes multi composants tolérant les fautes et hautement fiables apparaissent fréquemment dans les technologies modernes comme, par exemple, les réseaux de communication, les systèmes informatiques ou la recherche spatiale. Il est donc important de déterminer des mesures telles que la fiabilité, le temps moyen d'atteinte d'un état de défaillance ou la disponibilité pour ces systèmes. Pour cela on construit un modèle stochastique permettant de les analyser. Nous modéliserons le système par une chaîne de Markov à temps continu (CMTC). Cependant, l'espace d'états s'avère bien souvent trop grand en pratique pour espérer calculer les mesures directement [100]. De même, les techniques d'approximation numériques [84, 160] demandent beaucoup de temps de calcul ou/et une trop grande capacité mémoire. On utilise donc des méthodes de simulation. Une simulation naïve étant inefficace à cause de la rareté des défaillances, des méthodes d'échantillonnage préférentiel ont été développées.

Modèle

On considère un système multi composants constitué de C différents types de composants et n_i composants de type i, sujet à des défaillances et des réparations aléatoires, suivant toutes des lois exponentielles. Le modèle est alors donné par une CMTC $(Y_t)_{t\geq 0}$

définie sur l'espace fini d'états S où tout $x \in S$ donne pour chaque type i de composants, $i = 1, \ldots, C$, le nombre de composants opérationnels de type i, noté $n_i(x)$. Nous noterons $\mathbf{1}$ l'état ayant la totalité des composants opérationnels. L'espace S est partitionné en deux sous-ensembles U et F où U dénote l'ensemble des états opérationnels et F l'ensemble des états de défaillance. Nous supposerons que si $x \in U$ et $y \in S$ sont tels que $n_i(y) \geq n_i(x)$ pour tous les types de composants i, alors $y \in U$.

Sachant que nous souhaitons étudier des systèmes hautement fiables, on introduit un paramètre $\varepsilon > 0$, tel que $\varepsilon \ll 1$, mettant en évidence la rareté des défaillances, de sorte que le taux de défaillance d'un composant de type i quand on est dans l'état x est de la forme $a_i(x)\varepsilon^{b_i(x)}$, tels que $a_i(x) \geq 0$ et $b_i(x) \geq 1$ de dépendent pas de ε. Les réparations à partir de l'état x ramène à l'état y (où plusieurs composants pouvant être réparés simultanément) avec un taux ne dépendant pas de ε. La propagation des défaillances est autorisée. On note $p(y; x, i)$ (pouvant dépendre de ε) la probabilité que, si le système est dans l'état x et un composant de type i tombe en panne, le système aille instantanément dans l'état y par propagation de la défaillance à d'autres composants. Généralement, une transition (x, y) d'un état x à un état y sera dite une transition de défaillance si $\forall 1 \leq i \leq C$, $n_i(y) \leq n_i(x)$, avec une inégalité stricte pour au moins un type de composants et est notée $y \succ x$. On définit de même les transitions (x, y) de réparation par $y \prec x$ si $\forall 1 \leq i \leq C$, $n_i(y) \geq n_i(x)$, avec une inégalité stricte pour au moins un type de composants. Soit Γ l'ensemble de toutes les transitions possibles.

Soit X la chaîne de Markov à temps discret (CMTD) incluse et $\boldsymbol{P}$ sa matrice de transition. Les transitions de défaillance sont aussi rares (c'est à dire $O(\varepsilon)$) pour X, contrairement aux transition de réparation. On peut montrer (voir [162]) qu'il existe une fonction entière $b(x, y)$ et un entier $b_0 = \min_{1 \leq i \leq C} b_i(\mathbf{1})$ tels que pour tout $(x, y) \in \Gamma$

$$\boldsymbol{P}(x, y) = \begin{cases} \Theta(\varepsilon^{b(x,y)}) & \text{si } x \neq \mathbf{1} \\ \Theta(\varepsilon^{b(x,y)-b_0}) & \text{si } x = \mathbf{1}. \end{cases}$$

Notons que $b(x, y) \geq 1$ si $y \succ x$ et $b(x, y) = 0$ sinon. Définissons aussi Φ comme la mesure correspondante sur l'ensemble des chemins possibles de la CMTD.

Un type de système particulier auquel nous serons amenés à nous intéresser est celui des systèmes dits équilibrés.

Définition 6 *Nous dirons que le système est équilibré si toutes les transitions de défaillance sont du même ordre de grandeur (i.e. pour chaque état x, les $b(x, y)$ sont identiques).*

Nous nous intéressons à partir de maintenant à une mesure de sûreté de fonctionnement particulière, la MTTF (*Mean Time To Failure*) représentant le temps moyen d'atteinte de la défaillance, mais nous pourrions considérer d'autres mesures (fiabilité, disponibilité...). La MTTF peut être exprimée (en considérant les temps de séjour dans les états non pas aléatoires, mais constants et égaux aux moyennes ($1/q(i)$ des temps de séjour dans les états $i \in S$)) par le quotient (voir [85])

$$MTTF = \frac{E_\Phi \left(\sum_{k=0}^{\min(\tau_{\mathbf{1}}, \tau_F)} 1/q(X_k) \right)}{E_\Phi \left(1_{(\tau_F < \tau_{\mathbf{1}})} \right)}, \tag{3.2}$$

τ_F étant le temps d'atteinte de F à partir de $\mathbf{1}$ et $\tau_{\mathbf{1}}$ le temps de retour en $\mathbf{1}$.

Simulation et échantillonnage préférentiel

La $MTTF$ est estimée par simulation régénérative, en utilisant des cycles $(C_i)_{1\leq i\leq I}$ de la chaîne de Markov X, où C_i décrit X entre les $i-1$-ème et i-ème retour en $\mathbf{1}$, et en appliquant le théorème de la limite centrale à ces cycles. Un estimateur classique de la $MTTF$ est

$$\widehat{MTTF} = \frac{\sum_{i=1}^{I} G(C_i)}{\sum_{i=1}^{I} H(C_i)} \tag{3.3}$$

où $G(C_i)$ est la somme, pour tous les états parcourus pendant le cycle C_i, des espérances des temps de séjour dans ces états et $H(C_i) = 1_{(\tau_F<\tau_{\mathbf{1}})}(C_i)$.

Les défaillances constituant des événements rares, une estimation standard de la $MTTF$ s'avérera inefficace à cause du dénominateur. En effet, l'événement $1_{(\tau_F<\tau_{\mathbf{1}})}$ n'interviendra en moyenne pour la première fois qu'au bout $1/E_\Phi(1_{(\tau_F<\tau_{\mathbf{1}})})$ itérations. Cette espérance étant très petite, le nombre d'observations nécessaires est donc très important (et bien souvent trop) pour obtenir un intervalle de confiance précis. On peut utiliser alors des techniques de réduction de la variance de type échantillonnage préférentiel. Il est possible de simuler de manière indépendante les numérateurs et dénominateurs de l'expression (3.2) de la $MTTF$. Les fonctions étant différentes, on choisit deux mesures d'échantillonnage préférentiel différentes pour le numérateur et le dénominateur, ce qui permet ainsi de simuler chacune des fonctions de manière optimale. Sur les I cycles régénératifs, ξI sont réservés à l'estimation du numérateur et $(1-\xi)I$ à celle du dénominateur. Ainsi le numérateur peut efficacement être simulé par une simulation standard, et nous pouvons nous concentrer sur l'estimation du dénominateur

$$\gamma = E_\Phi[1_{[\tau_F<\tau_{\mathbf{1}}]}].$$

Pour estimer γ, nous utilisons l'échantillonnage préférentiel en choisissant une nouvelle matrice $\boldsymbol{P}'$ conduisant à une nouvelle mesure Φ' et telle que

$$\gamma = E_{\Phi'}[1_{[\tau_F<\tau_{\mathbf{1}}]}L]$$

où $L(x_0,\cdots x_n) = \dfrac{\Phi\{(X_0,\cdots,X_{\tau_F}) = (x_0,\cdots,x_n)\}}{\Phi'\{(X_0,\cdots,X_{\tau_F}) = (x_0,\cdots,x_n)\}}$ est défini pour tout chemin de la CMTD $(x_0,\cdots,x_n)$.

Différents choix ont étés proposés pour $\boldsymbol{P}'$ [39, 40, 41] :

- *Failure Biasing (FB)* [131] : l'idée est, pour chaque état x d'affecter (de forcer) une probabilité α (en général α sera pris entre 0,5 et 0,9) à l'ensemble des transitions de défaillance, et $1-\alpha$ aux réparations. Les probabilités individuelles des transitions dans chaque sous-ensemble sont prises proportionnelles à celles d'origine.
- *Selective Failure Biasing (SFB)* [85] : on procède ici comme dans le cas précédent, mais l'ensemble des transitions de défaillances est lui aussi partitionné en deux sous-ensembles, celui des transitions provoquées par une *première* défaillance d'un type de composant (probabilité totale $\alpha(1-\beta)$) et celui des transitions provoquées par des types de composants ayant déjà subi des défaillances (probabilité totale $\alpha\beta$). L'idée est d'augmenter les défaillances des types de composants déjà atteints, pour aller plus vite vers la panne.

- *Selective failure biasing for "series-like" systems (SFBS)* [41] : ici encore on sépare pour chaque état les transitions de défaillance en deux sous-groupes, celles qui provoquent les défaillances des types de composants les plus critiques, c'est à dire pour lesquels le nombre de composants devant tomber en pannes pour atteindre F est le plus petit (probabilité $\alpha\beta$) et les autres (probabilité $\alpha(1-\beta)$). Pour les systèmes en série, cette méthode doit être performante.
- *Selective failure biasing for "parallel-like" systems (SFBP)* [41] : L'idée, de la même manière que précédemment, est d'accélérer les transitions critiques d'abord, c'est à dire ici ceux dont le nombre de composants restant à tomber en panne est le plus grand, puis les non critiques. Cette méthode doit être performante pour les systèmes parallèles.
- *Inverse Failure Biasing (IFB) [176]* : inspirée par les techniques d'échantillonnage préférentiel pour la file M/M/1, cette technique consiste à échanger les probabilités des ensembles transition des défaillances et réparation (les probabilités individuelles dans chaque sous groupe restant proportionnelles à celles d'origine).
- *Distance-based selected failure biasing (DSFB) [44]* : cette méthode calcule (via les coupes minimales) la distance (en nombre minimal de composants restant à tomber en panne) de chaque état à la défaillance. À partir de tout état x, l'ensemble des transitions de défaillance (x, y) est partitionné en sous-ensembles selon la distance à la panne de y, auxquels on associe une probabilité donnée.
- *Méthodes équilibrées* [41, 162] : la méthode equilibréé de base, construite à partir de FB, consiste non plus à prendre les probabilités individuelles des transitions dans chaque sous-ensemble proportionnelles à celles d'origine, mais selon une répartition uniforme. Les méthodes sont dites équilibrées car toutes les transitions sont alors $\Theta(1)$, ce qui n'était pas nécessairement le cas, puisque deux transitions d'ordre différent (pour Φ) dans un même sous-ensemble restaient d'ordre différent. Dans [41], nous l'avons généralisée à toutes les autres methodes précédemment citées.

Une comparaison de ces propositions de mesures a été réalisée [41] sur différents modèles, pour déterminer leurs mérites respectifs.

Propriétés des estimateurs

Les temps de simulation moyens par cycle étant ici $\Theta(1)$, nous n'avons pas ici à nous intéresser à l'efficacité bornée où l'approximation normale bornée *généralisée*. Le critère de robustesse habituellement utilisé pour la simulation des systèmes markoviens est celui d'erreur relative bornée, que l'on peut spécifiquement redéfinir comme suit :

Définition 7 *[200] Définissons $\sigma^2_{\Phi'}$ comme la variance de la variable aléatoire $1_{[\tau_F<\tau_{\mathbf{1}}]}L$ sous la mesure Φ' (qui a pour espérance γ) et z_δ comme le $1-\delta/2$ quantile de la loi normale centrée réduite. Alors l'erreur relative pour un échantillon de taille I est définie par*

$$RE = z_\delta \frac{\sqrt{\sigma^2_{\Phi'}/I}}{\gamma}.$$

On dit qu'il y a erreur relative bornée si RE reste bornée quand $\varepsilon \to 0$.

Une condition nécessaire et suffisante sur la mesure d'échantillonnage Φ' a été obtenue dans [162] pour vérifier cette propriété.

Nous avons introduit dans [218, 224] le concept d'*approximation normale bornée* basé sur la définition 3 :

Définition 8 *Si $\rho_{\Phi'}$ dénote le moment absolu d'ordre trois et $\sigma_{\Phi'}$ la déviation standard de la variable aléatoire $1_{[\tau_F<\tau_{\mathbf{1}}]}L$ sous la mesure de probabilité Φ', on dit qu'il y a approximation normale bornée si $\rho_{\Phi'}/\sigma^3_{\Phi'}$ est borné quand $\varepsilon \to 0$.*

Nous avons obtenu une condition nécessaire et suffisante sur la mesure d'échantillonnage préférentiel afin de vérifier cette propriété. Soit r et s les entiers tels que $\gamma = \Theta(\varepsilon^r)$ et $\sigma^2_\Phi = \Theta(\varepsilon^s)$. Soit

$$\Delta_m = \{(x_0, \cdots, x_n) : n \geq 1, x_0 = \mathbf{1}, x_n \in F, x_i \notin \{\mathbf{1}, F\}\, \forall 1 \leq i \leq n-1, (x_i, x_i+1) \in \Gamma$$
$$\text{et } \Phi\{(X_0, \cdots, X_{\tau_F}) = (x_0, \cdots, x_n)\} = \Theta(\varepsilon^m)\}.$$

$$\Delta = \bigcup_{m=r}^{\infty} \Delta_m,$$

et

$$\begin{aligned}\Delta_{m,k} = \{ \quad & (x_0, \cdots, x_n) \in \Delta \ : \\ & \Phi\{(X_0, \cdots, X_{\tau_F}) = (x_0, \cdots, x_n)\} = \Theta(\varepsilon^m) \text{ et} \\ & \Phi'\{(X_0, \cdots, X_{\tau_F}) = (x_0, \cdots, x_n)\} = \Theta(\varepsilon^k)\},\end{aligned}$$

$$\Delta'_t = \bigcup_{m,k\,:\,m-k=t} \Delta_{m,k}.$$

Théorème 3 *Soit $\mathcal{I}$ la classes de mesures Φ' correspondants aux matrices $\boldsymbol{P}'$ telles que : pour tout $(\omega, y) \in \Gamma$, $\omega \neq \mathbf{1}$ et $y \succ \omega$,*

$$\textit{si } \boldsymbol{P}(\omega, y) = \Theta(\varepsilon^d), \textit{ alors } \boldsymbol{P}'(\omega, y) = \underline{O}(\varepsilon^{d-1})$$

et pour tout (ω, y) avec soit $y \prec \omega$ ou $y \succ \omega$ et $\omega = \mathbf{1}$,

$$\textit{si } \boldsymbol{P}(\omega, y) = \Theta(\varepsilon^d), \textit{ alors } \boldsymbol{P}'(\omega, y) = \underline{O}(\varepsilon^{d}).$$

L'approximation normale est bornée pour un nombre fixé d'observations et une mesure $\Phi' \in \mathcal{I}$ si et seulement si $\forall k, m$ tels que $m - k < r$, $(x_0, \cdots, x_n) \in \Delta_{m,k}$,

$$\Phi'\{(X_0, \cdots, X_{\tau_F}) = (x_0, \cdots, x_n)\} = \underline{O}(\varepsilon^{3m/2-3s/4})$$

(i.e. $k \leq 3m/2 - 3s/4$).

Nous pouvons dès lors établir les relations entre les propriétés d'erreur relative et d'approximation normale bornées avec les «bonnes estimations» asymptotiques de γ et de la variance. Définissons plus formellement cette notion de bonne estimation.

Définition 9 *Soit f une fonction définie sur Δ (et $f = 0$ sinon) et soit $t \geq 0$ tel que*

$$E_\Phi[f(X_0, \cdots, X_{\tau_F})] = \Theta(\varepsilon^t).$$

On dira que $E_\Phi[f(X_0, \cdots, X_{\tau_F})]$ est bien estimé asymptotiquement quand $\varepsilon \to 0$ sous Φ' si pour tout $(x_0, \cdots x_n) \in \Delta$ tel que $f(x_0, \cdots x_n)\Phi\{(X_0, \cdots, X_{\tau_F}) = (x_0, \cdots, x_n)\} = \Theta(\varepsilon^t)$, alors

$$\Phi'\{(X_0, \cdots, X_{\tau_F}) = (x_0, \cdots, x_n)\} = \Theta(1).$$

Autrement dit tout chemin contribuant au premier terme du développement de Taylor de $E_\Phi[f(X_0, \cdots, X_{\tau_F})]$ est non rare sous Φ', permettant de correctement l'estimer lorsque $\varepsilon \to 0$.

Nous avons alors la chaîne suivante de propriétés (voir [224]), mettant en évidence l'importance de l'approximation normale bornée.

Proposition 2 *Soit $\Phi' \in \mathcal{I}$ (comme défini dans le théorème 3). L'approximation normale bornée implique que $\sigma^2_{\Phi'}$ est bien estimé asymptotiquement.*
Par contre, la réciproque est fausse en général.

Proposition 3 *Si $\sigma^2_{\Phi'}$ est bien estimé asymptotiquement en utilisant une mesure de probabilité Φ', on a alors erreur relative bornée.*
Par contre, la réciproque est fausse en général.

Proposition 4 *Si on a erreur relative bornée, γ est bien estimé asymptotiquement.*
Par contre, la réciproque est fausse en général.

Il est intéressant de noter que toutes les méthode équilibrées, par exemple, vérifient la propriété d'approximation normale bornée.

3.3.3 Réplication des trajectoires pour la simulation d'événements rares

Il existe une autre grande famille pour la simulation des événements rares : les techniques basées sur la réplication (ou ramification) des trajectoires. Ces techniques ont été employées depuis les années 50 en physique mais n'ont été que récemment appliquées à des problèmes de télécommunication et de fiabilité. Il existe deux « écoles » pour l'application de ces méthodes : la première chronologiquement a été développée dans [235, 236, 237, 238] et a été appelée RESTART ; la deuxième dans [80, 81, 82, 83] où elle est appelée « splitting ». D'autres travaux peuvent être encore cités, comme [75, 89, 110, 195].

Supposons qu'on souhaite estimer la probabilité d'un événement rare A. La simulation standard étant inefficace on l'améliore en définissant des seuils où chaque trajectoire sera séparée en sous trajectoires. Considérons $k+1$ ensembles B_i tels quet $A = B_{k+1} \subset \cdots \subset B_1$ et utilisons [195]

$$P(A) = P(A|B_k)P(B_k|B_{k-1}) \cdots P(B_2|B_1)P(B_1) \tag{3.4}$$

tel que chaque événement conditionnel n'est plus rare. L'idée est d'échantillonner selon une loi de Bernoulli pour regarder si l'événement (non rare) B_1 est atteint. Dans ce cas l'essai est ramifié/séparé en R_1 essais et on regarde alors (toujours par des loi de Bernoulli) pour

chaque essai si B_2 est atteint, et ainsi de suite. Si un seuil n'est pas atteint, A ne l'est pas non plus, donc on arrête l'essai en cours. On a ainsi considéré $R_1 \cdots R_k$ essais dépendants. En utilisant R_0 réplications de la procédure, un estimateur de $P(A)$ est

$$\hat{p} = \frac{1}{R_0 \cdots R_k} \sum_{i_0=1}^{R_0} \cdots \sum_{i_k=1}^{R_k} \mathbf{1}_{i_0} \mathbf{1}_{i_0 i_1} \cdots \mathbf{1}_{i_0 i_1 \cdots i_k} \tag{3.5}$$

où $\mathbf{1}_{i_0 i_1 \cdots i_j}$ est le résultat du i-ème essai de Bernoulli à l'étape j. Il est démontré dans [238] qu'il est optimal de choisir un nombre de seuils $k = -1/2 \ln(P(A)) - 1$, et de prendre les seuils tels que $P(B_i|B_{i-1}) = e^{-2}$ et $R_i \approx 1/P(B_i|B_{i-1}) = e^2$.

Le cas de l'estimation d'un événement rare pour un processus stochastique est cependant plus complexe car la distribution stationnaire pour atteindre le seuil suivant est en général inconnue. Dans ce cas, on utilise des chemins, et nous ne sommes donc pas en stationnaire. Nous avons implémenté ces techniques pour la simulation des réseaux de Petri dans le logiciel SPNP, voir le chapitre 5 et [231]. Une combinaison avec les méthodes quasi-Monte Carlo de la section suivante est aussi depuis très récemment en cours d'étude [61].

3.4 Méthode quasi-Monte Carlo

3.4.1 Résultats généraux

Les méthodes de Monte Carlo ont comme grand avantage, par rapport aux méthodes d'analyse numérique classiques, d'avoir une vitesse de convergence en $O(1/\sqrt{N})$ (pour un échantillon de N points), donc indépendante de la dimension du problème. Néanmoins, il doit exister des suites de nombres telles que la convergence soit plus rapide, en supprimant l'aspect aléatoire. Ceci conduit à s'intéresser aux méthodes dites de Quasi Monte Carlo.

Ces méthodes approchent $\int_{[0,1]^s} f(x)dx$ par $\frac{1}{N}\sum_{n=1}^{N} f(\xi^{(n)})$ où $(\xi^{(n)})_{n\epsilon\mathbb{N}}$ est une suite déterministe. Comme nous voulons qu'il y ait asymptotiquement convergence quand $N \to +\infty$, la suite $(\xi^{(n)})_{n\in\mathbb{N}}$ doit être équirépartie, d'où la notion de discrépance [163, 65] :

Définition 10 *Soit $\mathcal{P} = (\xi^{(n)})_{n\in\mathbb{N}}$, $A_N(B,\mathcal{P})$ le nombre d'éléments appartenant à B parmi les N premiers de la suite $\mathcal{P}$, c'est-à-dire $A_N(B,\mathcal{P}) = \sum_{n=1}^{N} 1_B(\xi^{(n)})$, et soit $\mathcal{B}$ une famille de sous-ensembles des boréliens de $[0,1]^s$. La discrépance des N premiers termes de $\mathcal{P}$ est alors définie par $D_N(\mathcal{B},\mathcal{P}) = \sup_{B\epsilon\mathcal{B}} |\frac{A_N(B,\mathcal{P})}{N} - \lambda_s(B)|$ où λ_s est la mesure uniforme sur $[0,1)^s$.*

Le type de discrépance la plus utilisée est la discrépance étoile :

$$D_N^*(\mathcal{P}) = D_N(\mathcal{I}^*,\mathcal{P}) \text{ où } \mathcal{I}^* = \left\{ \prod_{i=1}^{s} [0,u_i) : u_i \in [0,1] \right\}.$$

Un autre type de discrépance facilement manipulable est la discrépance moyenne carré

$$T_N^{(2)*}(\mathcal{P}) = \left(\int_{[0,1]^s} \left(\frac{A_N([0,x),\mathcal{P})}{N} - \lambda_s([0,x)) \right)^2 dx \right)^{1/2}.$$

Une suite $\mathcal{P}$ est équirépartie si et seulement si $D_N^*(\mathcal{P}) \to 0$ quand $N \to \infty$. Il est possible de borner l'erreur d'approximation à l'aide de la discrépance. Soit $V(f)$ la variation au sens de Hardy et Krause de la fonction f. On a alors la borne de Koksma-Hlawka [163]

$$\left| \frac{1}{N} \sum_{n=1}^{N} f(\xi^{(n)}) - \int_{[0,1]^s} f(u)du \right| \leq V(f) D_N^*(\mathcal{P}).$$

Pour une suite infinie $\mathcal{S}$, il a été montré qu'on ne peut avoir mieux que $D_N^*(\mathcal{S}) = O(N^{-1} \log N)$ pour $s = 1$, et $O(N^{-1}(\log N)^{\alpha(s)})$ pour s quelconque (voir [163, pages 23-25], [31, page 10]). La fonction $\alpha(s)$ est encore indéterminée. On sait que $s/2 \leq \alpha(s) \leq s$ [31] et on conjecture que $\alpha(s) = s$.

Définition 11 *Une suite $\mathcal{S}$ de $[0,1]^s$ vérifiant $D_N^*(\mathcal{S}) = O(N^{-1}(\log N)^s)$ est appelée une suite à discrépance faible et les méthodes d'approximation utilisant ces suites sont appelées les méthodes de quasi-Monte Carlo.*

Asymptotiquement, la convergence est donc plus rapide que pour la méthode de Monte Carlo standard qui, on le rappelle, est en $O(1/\sqrt{N})$.

Dans [214], nous nous sommes intéressés à des techniques dites de réduction de la variation analogues à celles de réduction de la variance dans Monte Carlo, où, par transformation de l'intégrale, il est possible de réduire la variation. ceci nécessite une reformulation des différentes bornes. Cependant, comme nous le verrons plus tard, ceci n'a qu'un intérêt limité, car les bornes de l'erreur du type Koksma-Hlawka n'ont pas de réelle utilisation pratique.

3.4.2 Exemples de suites à discrépance faible

Présentons certaines suites à discrépance faible que nous utiliserons par la suite. Pour une description plus précise, voir [163, chapitre 3] ou [32, chapitre 2].

Suites de Halton

Ces suites forment une généralisation multi dimensionnelle des suites en dimension 1 de Van der Corput. Soit $n \in \mathbb{N}$ et $\sum_{j=0}^{\infty} a_j(n) b^j$ son développement en base b. On définit la fonction radicale inverse en base b par $\phi_b(n) = \sum_{j=0}^{\infty} a_j(n) b^{-j-1} \in [0,1)$. On considère s entiers p_i ($i = 1, \cdots, s$) premiers deux à deux et on pose $\mathcal{P} = (\xi^{(n)})_{n \in \mathbb{N}}$ la suite de Halton avec

$$\xi^{(n)} = (\phi_{p_1}(n), \cdots, \phi_{p_s}(n)) \in [0,1)^s.$$

Alors

$$D_N^*(\mathcal{P}) < \frac{s}{N} + \frac{1}{N} \prod_{i=1}^{s} \left(\frac{p_i - 1}{2 \log p_i} \log N + \frac{p_i + 1}{2} \right) = O(N^{-1}(\log N)^s). \tag{3.6}$$

En pratique, on considère habituellement p_i le i^{eme} nombre premier. Malgré la discrépance de ces suites très faible asymptotiquement, on peut observer que, en grande dimension, la propriété de bonne distribution peut n'apparaître qu'au bout d'un très grand nombre

d'itérations (i.e. on obtient une mauvaise répartition pour un petit nombre d'itérations). Par exemple, considérant $p_{15} = 47$ et $p_{16} = 53$, les 15^{eme} et 16^{eme} coordonnées seront constituées de sous suites croissantes de longueur respectives 47 et 53. Si on regroupe par paire ces coordonnées successives, on obtient pour un nombre modéré de points, comme illustré figure 3.1, des problèmes de régularité créés par les cycles monotones.

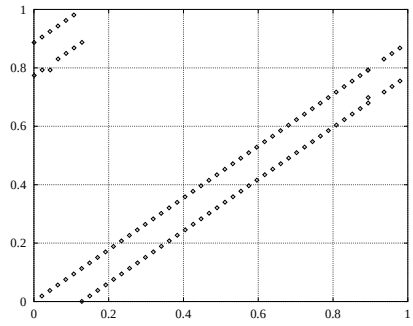


FIG. 3.1 – Projection de la suite de Halton sur les 15^{eme} et 16^{eme} coordonnées.

Un grand effort de recherche a eu lieu ces dernières années pour améliorer ce phénomène. Morokoff et Caflish [159] permutent aléatoirement (suivant une loi uniforme) les suites sur chaque coordonnée : pour les s suites que l'on considère alors, les N points sont rangés du plus petit au plus grand et sont alors permutés aléatoirement. On peut facilement montrer que cette opération ne change pas la borne de la discrépance donnée par (3.6). Cette technique est cependant moins élaborée que celle qui suit. Braaten et Weller [33] ont amélioré la distribution uniforme de la suite par l'introduction, pour chaque nombre premier p_i ($1 \leq i \leq s$), d'une permutation π_{p_i} sur $\{0, \cdots, p_i - 1\}$ telle que $\pi_{p_i}(0) = 0$, laissant apparaître un comportement plus chaotique. Ainsi, à partir du développement de n en base p_i, on pose

$$S_{p_i}(n) = \pi_{p_i}(a_0)/p_i + \pi_{p_i}(a_1)/p_i^2 + \cdots + \pi_{p_i}(a_j)/p_i^{j+1}$$

et la nouvelle suite de Halton est $\mathcal{P} = (\xi^{(n)})_{n \in \mathbb{N}}$ avec $\xi^{(n)} = (S_{p_1}(n), \cdots, S_{p_s}(n))$. Quelles que soient les permutations π_{p_i}, avec la restriction $\pi_{p_i}(0) = 0$, la borne de la discrépance de $\mathcal{P}$ donnée par (3.6) reste inchangée (i.e. $\mathcal{P}$ est toujours une suite à discrépance faible). Le problème est alors de choisir les bonnes permutations. Braaten et Weller [33] ont construit les permutations π_{p_i} unidimensionnellement telles que si on connaît $\pi_{p_i}(1), \cdots, \pi_{p_i}(j)$, on choisit $\pi_{p_i}(j+1)$ comme l'élément minimisant la discrépance moyenne carrée $T_{j+1}^{(2)*}$ des $j+1$ points $\{\pi_{p_i}(1)/p_i, \cdots, \pi_{p_i}(j)/p_i, \pi_{p_i}(j+1)/p_i\}$. Ces suites sont connues parmi les plus efficaces en pratique, et même les plus efficaces pour certaines dimensions et certaines applications [175]. Récemment, Faure [70] a amélioré ce choix, de manière lui-aussi uni-dimensionnelle. Bien que le choix de toutes ces permutations donne de bons résultats, on peut se demander pourquoi une concaténation de suites uni-dimensionnelles conduit à une bonne suite multi-dimensionnelle. C'est pourquoi nous avons construit de nouveaux algorithmes dans [210, 213].

Une première méthode de choix de permutations donne une table unique, disponible pour toute dimension. L'algorithme crée la table ligne par ligne. Il est différent de celui de Braaten et Weller qui génère la table élément par élément. Une fois les permutations $\pi_{p_1}, \cdots, \pi_{p_j}$ $1 \leq j < s$ déterminées, la permutation $\pi_{p_{j+1}}$ pour la $(j+1)^{eme}$ coordonnée est choisie parmi toutes les permutations de $\{0, \cdots, p_{j+1}-1\}$ telles que $\pi_{p_{j+1}}(0) = 0$, et $\pi_{p_{j+1}}$ minimise la discrépance à l'origine moyenne carrée $T^{(2)*}_{p_{j+1}-1}(\mathcal{P})$ de la suite en dimension $j+1$ définie par $\mathcal{P} = (S_{p_1}(n), \cdots, S_{p_{j+1}}(n))_{1 \leq n \leq p_{j+1}-1}$. Pour un j fixé, le nombre de permutations possible pour la $j+1^{eme}$ coordonnée est $(p_{j+1}-1)!$. Pour éviter cette explosion combinatoire, on choisit les permutations au moyen d'une simulation de type Monte Carlo.

De la même manière, on peut trouver un algorithme créant une table entière pour chaque dimension : pour une dimension s fixée, on cherche une table $(\pi_{p_1}, \cdots, \pi_{p_s})$, avec π_{p_i} permutation de $\{0, \cdots, p_i - 1\}$ telle que $\pi_{p_i}(0) = 0$ et π_{p_i} minimise $T^{(2)*}_{p_s-1}(\mathcal{P}_{p_1,\cdots,p_s})$ où $\mathcal{P}_{p_1,\cdots,p_s}$ est la suite en dimension s associée aux permutations $(\pi_{p_1}, \cdots, \pi_{p_s})$. Comme le nombre de permutations possibles est plus grand que celui pour l'algorithme créant la table ligne par ligne, on utilise encore une approche de type Monte Carlo.

Les permutations obtenues sont décrites dans [213]. En dimension 16 par exemple, elles permettent une réduction de la discrépance d'un facteur 10^4 par rapport à la suite de Halton standard, et d'un facteur 10 par rapport aux permutations proposées par Braaten et Weller.

(t, s)-suites

Définition 12 *Soit s fixé et b nombre premier tel que $b \geq s$. Un intervalle*

$$E = \prod_{i=1}^{s} [a_i b^{-d_i}, (a_i + 1)b^{-d_i}],\ a_i, b_i \in \mathbb{Z},\ d_i \geq 0,\ 0 \leq a_i < b$$

est appelé un intervalle élémentaire en base b.

Définition 13 *Soient t,m entiers tels que $t \leq m$. Un (t,m,s)-treillis en base b est un ensemble $\mathcal{P} = (\xi^{(n)})_{1 \leq n \leq b^m}$ comprenant b^m points de $[0,1)^s$ tel que pour tout intervalle élémentaire E en base b vérifiant $\lambda_s(E) = b^{t-m}$ on ait $A(E, \mathcal{P}) = \sum_{\xi^{(n)} \in \mathcal{P}} 1_E(\xi^{(n)}) = b^t$.*

Définition 14 *Soit $t \geq 0$. La suite $\mathcal{P} = (\xi^{(n)})_{n \in \mathbb{N}}$ est une (t,s)-suite de $[0,1)^s$ en base b si $\forall k \geq 0$ et $m > t$, $\{\xi^{(n)} : kb^m \leq n < (k+1)b^m\}$ est un (t,m,s)-treillis en base b.*

Les (t, s)-suites sont donc très bien réparties sur tout l'intervalle $[0,1)^s$ et prouvées être à discrépance faible.

Pour construire une (t,s)-suite, on choisit

- (S1) un anneau $\mathcal{R}$ commutatif tel que $card(\mathcal{R}) = b$;
- (S2) des bijections ψ_r et $\eta_{ij} : \mathbb{Z}/b\mathbb{Z} \to \mathbb{Z}/b\mathbb{Z}$ pour $r \geq 0$, $1 \leq i \leq s$ et $j \geq 1$;
- (S3) des éléments $c_{jr}^{(i)} \in \mathcal{R}$ pour $1 \leq i \leq s$, $r \geq 0$ et $j \geq 1$.

Soit $n \in \mathbb{N}$ et $\sum_{r=0}^{\infty} a_r(n) b^r$ son développement en base b. On pose
$\xi_i^{(n)} = \sum_{j=1}^{\infty} y_{ij}^{(n)} b^{-j}$ pour $n \geq 0$, $1 \leq i \leq s$ avec $y_{ij}^{(n)} = \eta_{ij}(\sum_{r=0}^{\infty} c_{jr}^{(i)} \psi_r(a_r(n))) \in \mathbb{Z}/b\mathbb{Z}$, puis

$$\xi^{(n)} = (\xi_1^{(n)}, \cdots, \xi_s^{(n)}). \tag{3.7}$$

Nb. de points	10^3	10^4	10^6	10^9	10^{12}	10^{16}
Borne	1.50	$3.34\ 10^{-1}$	$2.14\ 10^{-2}$	$1.21\ 10^{-4}$	$5.83\ 10^{-7}$	$1.92\ 10^{-10}$

TAB. 3.1 – Bornes de la discrépance en dimension 5

Pour garantir que $\xi^{(n)}$ appartienne à $[0,1)^s$, et non à $[0,1]^s$, nous prenons la condition suffisante : pour $1 \leq i \leq s$, $r \geq 0$ et tout j suffisamment grand, $\eta_{ij}(0) = 0$ et $c_{jr}^{(i)} = 0$.

Pour une construction précise, voir [163].

Règles de quadrillage

Les règles de quadrillage construisent des suites finies. Elles consistent à utiliser un vecteur $g \in \mathbb{Z}^s$ et considèrent $\mathcal{P} = (\xi^{(n)})_{1 \leq n \leq N}$ avec $\xi^{(n)} = \{\frac{n}{N} g\}$ où $\{\cdot\}$ rend les parties fractionnaires de chaque coordonnée du vecteur. Un bon choix de couple (g, N) a été un thème actif de recherche [130, 163, 203].

3.4.3 Inconvénients de quasi-Monte Carlo

Un des principaux avantages des méthodes Monte Carlo est qu'elles peuvent être facilement appliquées à des problèmes très généraux. Quasi-Monte Carlo au contraire, même si asymptotiquement plus rapide, exige un « échantillon » de plus en plus grand pour être performant quand la dimension augmente. Ceci limite fortement son efficacité pour des problèmes de grande dimension.

Un deuxième point concerne l'estimation de l'erreur : Monte Carlo d'un côté donne aisément un intervalle de confiance par application du théorème de la limite centrale. De l'autre côté quasi-Monte Carlo, par l'intermédiaire de la borne de Koksma-Hlawka par exemple est supposé fournir une borne plus forte (car non probabiliste) pour majorer l'erreur. Malheureusement cette borne, composée de la variation de la fonction et de la discrépance de la suite, est la plupart du temps inutile en pratique. En effet, il existe des fonctions à variation infinie comme par exemple en dimension $s = 3$

$$f(x_1, x_2, x_3) = \min(x_1 + x_2 + x_3, 1)$$

pour lesquelles quasi-Monte Carlo converge (rapidement) [32]. De plus, même si la variation est finie, elle est composée de $2^s - 1$ variations au sens de Vitali, chacune au moins aussi difficile à estimer que l'intégrale de départ elle-même, leur somme devenant (excessivement) grande quand s augmente. De même, les bornes de la discrépance sont elles aussi inutiles en pratique car, bien que ce soient elles qui donnent l'ordre de convergence asymptotique connu $O(N^{-1}(\ln N)^s)$, la valeur de N pour obtenir cet ordre est souvent irréalisable en pratique. Les Tables 3.1 et 3.2 donnent par exemple les bornes pour des $(0,5)$- suite de Niederreiter en base $b = 5$ et $(0,25)$-suite de Niederreiter en base $b = 27$ et différentes valeurs de N. Les nombres de points nécessaires pour obtenir des valeurs modérées sont irréalisables en dimension 25. Pourtant, les méthodes de quasi-Monte Carlo ont montré leur vitesse en pratique dans ces cas.

Les deux sous-sections suivantes montrent comment ces problèmes peuvent être contournés.

Nombre de points	10^3	10^9	10^{15}	10^{21}	10^{35}
Borne	$2.88\ 10^5$	$1.11\ 10^6$	$2.46\ 10^5$	$5.22\ 10^3$	$1.62\ 10^{-3}$

TAB. 3.2 – Bornes de la discrépance en dimension 25

3.4.4 Quasi-Monte Carlo randomisé

Méthodes et ordre de convergence

Les techniques de randomisation permettent d'utiliser le théorème de la limite centrale pour déterminer/estimer l'erreur dans les méthodes QMC. Elles peuvent être aussi vues comme une technique de réduction de la variance dans les méthodes Monte Carlo, en tant que généralisation des variables antithétiques. Parmi les méthodes de randomisation possibles, nous nous intéresserons plus particulièrement aux méthodes dites shiftées sur lesquelles nous avons apporté plusieurs contributions.

Suites à discrépance faible shiftées (aussi dites translatées).

Soit $(\xi^{(n)})_{n\in\mathbb{N}}$ une suite à discrépance faible. La technique de randomisation [57] consiste à ajouter une même variable aléatoire X uniformément distribuée sur $[0,1)^s$ à chaque élément de la suite afin de conserver la bonne répartition de la suite. Si on s'intéresse à

$$Z = \frac{1}{N}\sum_{n=1}^{N} f(\{X+\xi^{(n)}\}) \tag{3.8}$$

au lieu de $\frac{1}{N}\sum_{n=1}^{N} f(\xi^{(n)})$ dans QMC, où $\{x\}$ est le vecteur des parties fractionnaires des coordonnées de x, l'idée est de calculer la valeur moyenne obtenue sur I copies indépendantes Z_i de Z,

$$\frac{1}{I}\sum_{i=1}^{I} Z_i \tag{3.9}$$

et d'obtenir un intervalle de confiance en utilisant le théorème de la limite centrale sur ces I variables. La bonne répartition de la suite à dicrépance faible doit permettre de réduire la variance de l'estimateur (sans biais).

Pour appeler la fonction f le même nombre de fois pour chaque méthode, nous devons comparer la variance de l'estimateur (3.9) avec celle de l'estimateur Monte Carlo standard pour NI variables aléatoires. Nous obtiendrons une réduction de la variance si et seulement si

$$\sigma^2\left(\frac{1}{N}\sum_{n=1}^{N} f(\{X+\xi^{(n)}\})\right) < \frac{1}{N}\sigma^2(f(X)). \tag{3.10}$$

Dans [215, 216, 223], nous avons montré qu'il est possible d'obtenir une réduction conséquente de la variance par comparaison avec celle de la somme de N variables aléatoires indépendantes et identiquement distribuées $f(X)$. Citons certains résultats de [215, 216, 223] :

Théorème 4 *Soit $\mathcal{P} = (\xi_n)_{n\in\mathbb{N}}$ une suite à discrépance faible sur $[0,1)^s$. Si f est une fonction à variation finie et X une v.a. uniformément répartie sur $[0,1)^s$, on a*

$$\sigma^2\left(\frac{1}{N}\sum_{n=1}^{N} f(\{X+\xi_n\})\right) = O(N^{-2}(\ln N)^{2s}). \tag{3.11}$$

Une parallélisation de la méthode a été proposée dans [228].

De plus, on peut montrer que cette vitesse de convergence est aussi valable pour d'autres fonctions que celles à variation finie. En effet, considérons l'ensemble $\mathcal{F}$ des fonctions continues, muni de la mesure de Wiener qui est concentrée sur des fonctions à variation infinie [159]. Rappelons que la mesure de Wiener μ_W est une mesure gaussienne d'espérance nulle et de noyau de covariance

$$R(x,y) = \int_{\mathcal{F}} f(x)f(y)d\mu_W(f) = \prod_{i=1}^{s}\min(x_i,y_i).$$

On peut alors obtenir le théorème suivant :

Théorème 5 *Si $\mathcal{P} = (\xi_n)_{n\in\mathbb{N}}$ est suite à discrépance faible sur $[0,1)^s$, la variance moyenne $\sigma^2_{N,avg}$ de la variable aléatoire $\frac{1}{N}\sum_{n=1}^{N} f(\{\xi_n+X\})$, moyenne prise sur l'ensemble $\mathcal{F}$ des fonctions f continues sur $[0,1]^s$ que l'on munit de la mesure de Wiener μ_W, est en $O(N^{-2}(\ln N)^{2s})$.*

La vitesse de convergence peut même être plus rapide pour certaines classes de fonctions [217]. Soit $\alpha > 1$, $C > 0$ et $\forall h \in \mathbb{Z}^s$, $r(h) = \prod_{i=1}^{s}\max(1,|h_i|)$. Soit $E^s_\alpha(C)$ l'ensemble des fonctions périodiques $f : \mathbb{R}^s \to \mathbb{R}$, de période 1 sur chaque coordonnée, telles que le coefficient de Fourier de rang h de la fonction f, $\hat{f}(h) = \int_{[0,1]^s} f(x)e(-h.x)dx$ ($x.y$ étant le produit scalaire standard de $x, y \in \mathbb{R}^s$), vérifie

$$|\hat{f}(h)| \le Cr(h)^{-\alpha} \text{ pour tout } h \in \mathbb{Z}^s.$$

On considère comme suite à discrépance faible les règles de quadrillage $(\{ng/N\})_{1\le n\le N}$ avec g vecteur judicieusement choisi.

Théorème 6 *Soit X variable uniforme sur $[0,1]^s$. Alors, pour tout $\alpha > 1$, $C > 0$ et $N \ge 1$, il existe un $g \in \mathbb{Z}^s$ tel que*

$$\max_{f\in E^s_\alpha(C)} \sigma^2\left(\frac{1}{N}\sum_{n=0}^{N-1} f\left(\left\{X+\frac{n}{N}g\right\}\right)\right) = O(N^{-2\alpha}(\ln N)^{2\alpha s}).$$

Ce travail sur les règles de quadrillage a été étendu dans [127, 129], où le choix d'un bon g est cherché.

Autres randomisations : pour chaque type de randomisation, l'idée est d'utiliser I copies indépendantes et d'appliquer le théorème de la limite centrale à ces I copies.

– (t, s)-**suites brouillées**. Les (t, s)-suites brouillées ont été introduites par A. Owen dans [169, 170]. L'idée est de perturber les digits des (t, s)-suites en base b en utilisant des permutation aléatoires de digits, tout en préservant la propriété de discrépance faible. Toute permutation est choisie parmi les $b!$ permutations possibles de $\{0, \ldots, b-1\}$. Pour un point $\xi = \sum_{k=1}^{\infty} a_k b^{-k}$ en base b, la permutation utilisée pour a_1 est $\pi_.$, la permutation utilisée pour a_2 est $\pi_{.,a_1}$ (dépendante de a_1 mais indépendante de $\pi_.$), et généralement la permutation utilisée pour a_k est $\pi_{.,a_{k-1},\ldots,a_1}$. D'autres choix de permutation, moins coûteux en termes de stockage et en temps de génération, ont plus récemment été développées. Pour certaines classes de fonction, il est montré que la variance peut décroître aussi rapidement que $O\left(N^{-3}(\log N)^{s-1}\right)$ [171].
– **Départ aléatoire des suites de Halton**. Cette méthode [209, 241] voit les suites de Halton comme une application multidimensionnelle de la transformation de von Neuman-Kakutani [116] avec vecteur orbite $(0, \cdots, 0)$. En choisissant un vecteur d'orbite aléatoire, on obtient une suite de Halton randomisée.

Des comparaisons numériques des différentes techniques de randomisation et des différentes suites ont été réalisées. Dans [211] notamment, nous incitons à l'utilisation pour la méthode du shift des suite de Sobol (des (t, s)-suites) utilisant le code de Gray [16].

Nous avons appliqué efficacement ces techniques à la simulation de divers systèmes importants dans le domaine de l'évaluation des performances et des télécommunication. Dans [216], nous avons ainsi amélioré la simulation des files d'attente multi-classes à forme produit et obtenu sur certains exemples une amélioration de l'efficacité de 626 pour $N = 10^4$ par rapport à la meilleure méthode Monte Carlo connue (sachant que cette amélioration grandira avec N en raison de la vitesse de convergence plus rapide). Dans [212], nous l'avons appliqué à l'évaluation d'un système cellulaire avec partage dynamique des ressources, et dans [219] aux réseaux à perte.

3.4.5 Quasi-Monte Carlo pour les problèmes de grande dimension

Comme nous l'avons dit, les méthodes de quasi-Monte Carlo sont souvent peu efficaces pour un nombre de points modéré en grande dimension (plus exactement quand la dimension *effective* est importante, c'est à dire quand il n'est pas possible de concentrer la majorité des variations de la fonction sur un petit nombre de composantes). Or les problèmes de grande dimension sont très présents en pratique (comme par exemple pour la simulation de chaînes de Markov). Il existe différentes manières d'adapter les méthodes de quasi-Monte Carlo au cas de problème à grande dimension. Nous nous concentrons ici sur celles où nous avons contribué. D'autres peuvent être trouvées en [173, 207].

Méthode mixte

La méthode dite mixte a été initialement développée par G. Ökten et J. Spanier [166, 207, 206], puis renommée « padding » par A. Owen. Elle consiste, quand on cherche à intégrer (ou simuler un problème) en dimension s très grande, à considérer une suite à discrépance faible en dimension $d < s$ de sorte que les d premières coordonnées des points de l'échantillon seront générées via la suite à discrépance faible, et les suivantes en utilisant des variables aléatoires uniformes.

Formellement, on approche l'intégrale $\mathcal{I} = \int_{[0,1]^s} f(x)dx$, par $\frac{1}{N}\sum_{k=1}^{N} f(x^{(k)})$, où

$$x^{(k)} = (\xi^{(k)}, X^{(k)})$$

est une suite s-dimensionnelle, avec $(\xi^{(k)})_{k\geq 1}$, suite à discrépance faible d-dimensionnelle, et $X^{(k)}, k \geq 1$, variables aléatoires uniformes sur $(0,1)^{s-d}$ indépendantes. Il est espéré que la bonne répartition de la suite à discrépance faible va améliorer la convergence, particulièrement pour les problèmes à dimension effective faible, où la variance se concentre sur les premières coordonnées.

Posons les variables aléatoires

$$Y_k = f\left(\xi_1^{(k)}, \cdots, \xi_d^{(k)}, X_{d+1}^{(k)}, \cdots, , X_s^{(k)}\right),$$

d'espérences $\mu_k = E[Y_k]$ et variances $\sigma_k^2 = Var(Y_k)$. Posons aussi $s_N^2 = \sigma_1^2 + ... + \sigma_N^2$. Nous avons obtenu des bornes de la discrépance pour cette méthode [167, 168]. Nous avons aussi obtenu les résultats suivants :

Théorème 7 *Supposons f bornée sur $[0,1]^s$ et que les fonctions*

$$g(x_1, ..., x_d) = \int_{[0,1]^{s-d}} f(x_1, ..., x_d, X_{d+1}, ..., X_s)^2 dX_{d+1}...dX_s$$
$$h(x_1, ..., x_d) = \left(\int_{[0,1]^{s-d}} f(x_1, ..., x_d, X_{d+1}, ..., X_s)^2 dX_{d+1}...dX_s\right)^2$$

sont à variation au sens de Hardy et Krause bornée (condition suffisante de convergence, mais les convergences

$$\frac{1}{N}\sum_{k=1}^{N} g(\xi_1^{(k)}, \cdots, \xi_d^{(k)}) \to \int_{[0,1]^d} f(x)^2 dx$$

et

$$\frac{1}{N}\sum_{k=1}^{N} h(\xi_1^{(k)}, \cdots, \xi_d^{(k)}) \to \int_{[0,1]^d} h(y)dy = \int_{[0,1]^d} \left(\int_{[0,1]^{s-d}} f(y,x)dx\right)^2 dy$$

suffisent). Alors

1. *La somme normalisée*
$$\frac{\sum_{k=1}^{N} Y_k - \sum_{k=1}^{N} \mu_k}{s_N}$$
converge en loi vers une loi normale centrée réduite.
2. *On a*
$$s_N^2/N \to L = \int_{[0,1]^s} f(x)^2 dx - \int_{[0,1]^d} \left(\int_{[0,1]^{s-d}} f(y,x)dx\right)^2 dy;$$

3. *La méthode mixte donne asymptotiquement une réduction de la variance par rapport à Monte Carlo, avec un facteur de réduction*

$$\frac{\int_{[0,1]^s} f(x)^2 dx - \int_{[0,1]^d} \left(\int_{[0,1]^{s-d}} f(y,x) dx \right)^2 dy}{\int_{[0,1]^s} f(x)^2 dx - \left(\int_{[0,1]^s} f(x) dx \right)^2} \leq 1. \tag{3.12}$$

Il est important de noter que l'utilisation de suites à discrépance faible n'est pas requise dans ce théorème, l'hypothèse clé étant la convergence vers la distribution uniforme de la suite déterministe. Cependant les suites à discrépance faible sont utiles pour deux raisons :

– elles réduisent autant que possible le biais de l'estimateur ;
– La variance de l'estimateur converge plus rapidement vers sa valeur asymptotique.

La méthode présente cependant deux problèmes : l'estimateur est biaisé comme nous venons de le dire et la variance est difficile à estimer. Cependant, en utilisant les techniques de randomisations précédemment citées, l'estimateur devient sans biais et la variance est facilement estimable. On utilise donc un estimateur

$$\frac{1}{I} \sum_{i=1}^{I} \left(\frac{1}{N} \sum_{k=1}^{N} f(u^{(k,i)}) \right)$$

où les suites randomisées que nous considérons sont en fait des suites de quasi-Monte Carlo randomisées des section précédentes pour les d premières coordonnées, alors que les coordonnées restantes $u_n^{(k,i)}$ ($n = d+1, \ldots, s$, $1 \leq k \leq N$, $1 \leq i \leq I$) sont des variables aléatoires uniformes sur $[0,1)$. L'efficacité de la méthode (ainsi que sa version randomisée) a été démontrée dans [168], et le gain par rapport à Monte Carlo (3.12) vérifié.

Quasi-Monte Carlo et simulation des chaînes de Markov

Un type de problème où les méthodes de quasi-Monte Carlo peuvent s'avèrer inefficaces en raison de la grande dimension des suites nécessaires est la simulation des chaînes de Markov. Nous avons proposé dans [122] une adaptation particulière des méthodes QMC à la simulation transitoire des chaînes de Markov à temps discret. Nous avons proposé d'utiliser une suite à discrépance faible en dimension 2 et de simuler les chaînes en parallèle, mais de les réordonner à chaque étape.

Le problème général est modélisé comme suit. Soit E un ensemble fini ou dénombrable ($E = \mathbb{N}$ ou $E = \mathbb{Z}$) représentant l'espace d'états. Considérons une chaîne de Markov à temps discret $(Y_n)_{n \in \mathbb{N}}$ de loi initiale $\mu = (\mu\{i\} : i \in E)$ et matrice de transition $P = (p(i,j) : i,j \in E)$. Nous cherchons à approcher la loi de probabilité après n étapes de la chaîne de Markov. Noter que dans cette section, n désigne le nombre d'étapes considérées de la chaîne, et L est le nombre de chaînes.

L'algorithme est le suivant. Soit $X = \{\mathbf{x}_p\}_{p \in \mathbb{N}}$ une $(t,2)$-suite en base b et X^n l'ensemble de points $\{\mathbf{x}_p = (x_{p,1}, x_{p,2}) \ : \ nL \leq p < (n+1)L\}$ avec $L = b^q$. Si proj_1 et proj_2 dénotent les projections définies par $\mathrm{proj}_i(x_1, x_2) = x_i$, pour $i = 1, 2$, on suppose que $\forall n \geq 0$

$$\mathrm{proj}_1 X^n \text{ is a } (0, q, 1)\text{-net in base } b, \tag{3.13}$$

et que, si $E = \mathbb{Z}$,

$$0 \notin \mathrm{proj}_2 X. \tag{3.14}$$

Considérons L chaînes distinctes $Y^{(i)} = \{Y_n^{(i)}\}_{n\in\mathbb{N}}$, $0 \le i \le L-1$.

1. A partir de l'ensemble de points X^0, on échantillonne l'état initial (au temps $j = 0$) de chaque chaîne selon μ : l'état initial $i_0^{\lfloor Lx_{p,1}\rfloor}$ de la chaîne $\lfloor Lx_{p,1}\rfloor$ est échantillonné en utilisant $x_{p,2}$ (comme avec un nombre pseudo aléatoire). Notons que grâce à (3.13), chaque chaîne sera échantillonnée une unique fois.
2. On réordonne les chaînes selon leur état : $i_j^0 \le i_j^1 \le \cdots \le i_j^{L-1}$.
3. $j = j + 1$. Considérons X^j. On échantillone le nouvel état de chaque chaîne selon la matriceS P : le nouvel état de la chaîne $\lfloor Lx_{p,1}\rfloor$ est choisi en utilisant $x_{p,2}$, selon la distribution discrète $P_{i_{j-1}^{\lfloor Lx_{p,1}\rfloor},\cdot}$.
4. si $j = n$ stop sinon retourner en 2)
5. Considérer la moyenne de la fonction de récompense f sur les L chaînes :

$$\frac{1}{L}\sum_{l=0}^{L-1} f(i_N^l).$$

Pour prouver la convergence de la méthode, il est nécessaire de reformuler les diverse notions pour les adapter au contexte (comme réalisé dans [214]). La *λ-discrépance étoile* d'un ensemble de points I pour une distribution λ est (re)définie par

$$D^\star(I,\lambda) := \sup_{k\in E}\left|\frac{1}{L}\sum_{0\le\ell<L} 1_{F_k}(i_\ell) - \sum_{i\in F_k}\lambda\{i\}\right|,$$

où

$$F_k := \{i \in E : i < k\}.$$

Définissons la *variation* d'une suite u comme

$$V(u) := \sum_{i\in E'} |u(i+1) - u(i)|,$$

avec $E' := \{i \in E : i + 1 \in E\}$. Le théorème de Koksma-Hlawka peut être reformulé :

Lemme 1 *Soit λ une distribution sur E. Si u est une suite à variation bornée et si I est un ensemble de points $i_0, \ldots, i_{L-1} \in E$, alors*

$$\left|\frac{1}{L}\sum_{0\le\ell<L} u(i_\ell) - \sum_{i\in E}\lambda\{i\}u(i)\right| \le V(u)D^\star(I,\lambda).$$

Soit $I^n = \{i_j^0, i_j^1, \cdots, i_j^{L-1}\}$ l'ensemble d'états échantillonné à partir de I^{n-1} en utilisant l'ensemble de points X^n. La proposition suivante est prouvée dans [122].

Proposition 5 *Si la matrice P vérifie*

$$\forall k \in E \quad \sum_{i \in E'} \left| \sum_{j<k} p(i+1,j) - \sum_{j<k} p(i,j) \right| \leq 1,$$

alors, pour $q \geq t$,

$$D^\star(I^n, \mu P^n) \leq D^\star(I^0, \mu) + 2nb^{-\lfloor \frac{q-t}{2} \rfloor}. \tag{3.15}$$

Corollaire 1 *Si on cherche à estimer la moyenne $\mathbb{E}(u(Y_n))$ d'une suite u telle que $V(u) < \infty$ à la n-ème étape d'une chaîne de Markov à temps discret $(Y_n)_{n \in \mathbb{N}}$, l'estimation précédente converge quand $L \to \infty$.*

De même, si on estime la récompense cumulée $\sum_{m=1}^{n} \mathbb{E}(u(Y_m))/n$ jusqu'à l'instant n, la méthode converge sous les mêmes hypothèses.

La borne (3.15) prouve la convergence, mais est relativement pessimiste , car elle donne une convergence d'ordre $O(L^{-1/2})$ comme Monte Carlo (notons quand même qu'il s'agit d'une borne au pire cas et non en moyenne comme pour Monte Carlo). Cependant, dans [121], nous avons comparé numériquement sur plusieurs exemples les ordres de convergence pour cette méthode avec ceux de Monte Carlo et de diverses variantes de quasi-Monte Carlo. L'efficacité de la méthode y est vérifiée.

En collaboration avec l'Université de Savoie et l'Université de Montréal, nous avons travaillé sur une version randomisée de cette méthode (afin d'estimer l'erreur), avec une extension aux chaînes de Markov dont le temps de simulation est aléatoire et peut être non borné, au cas de n'importe quelle suite à discrépance faible, ainsi qu'au cas où d variables aléatoires uniformes sont nécessaires pour simuler une transition. Nous utilisons aussi ici le fait qu'un même ensemble de points à discrépance faible peut être utilisé à chaque étape [124, 126, 125]. L'algorithme est similaire à celui déterministe aux point suivants près :

- à chaque étape, on utilise une randomisation différente de l'ensemble de points utilisé, de manière à obtenir un estimateur sans biais ;
- seules les chaînes pour lesquelles le temps d'arrêt n'est pas atteint sont rééchatillonnées, les autres étant considérées dans l'état $+\infty$.

Nous avons pu ici aussi obtenir des résultats de convergence dans le pire cas, similaires à ceux pour la méthode déterministe. Nous avons aussi obtenu des résultats sur la variance dans certaines situations [124, 126] : une borne au pire cas en $O(L^{-1/2})$ est encore prouvée, ainsi qu'une borne en moyenne $O(L^{-3/2})$ sous certaines (fortes) conditions.

Du fait des majorations fortes des bornes obtenues, il parait intéressant d'insister sur une illustration numérique de l'efficacité réelle en pratique. La table 3.4.5 donne l'ordre de réduction de variance obtenu sur la simulation du temps moyen d'attente des 100 premiers clients d'une file M/M/1 initiallement vide en utilisant les équations de Lindsley et différentes valeurs d'intensité du trafic ρ. Les différentes méthodes utilisées ici sont les méthodes QMC randomisées classiques (en dimension 100) avec différentes suites, et la méthode présentée ci-dessus (dénotée par Array) avec ces même suites. Le gain observé est important, par rapport aux méthodes QMC randomisées classiques, et plus encore par rapport à Monte Carlo. Ce gain relatif augmente avec le nombre de points utilisés.

ρ		$k=10$	$k=12$	$k=14$	$k=16$	$k=18$	$k=20$
0.2	Classique-Korobov-Baker	5	8	15	16	59	117
	Classique-Sobol	1	1	3	1	13	28
	Array-Korobov	18	55	49	292	850	2169
	Array-Sobol	87	282	836	3705	10640	47850
	Array-Sobol-NoGray	46	112	276	874	2914	7429
0.5	Classique-Korobov-Baker	10	7	13	6	14	14
	Classique-Sobol	2	1	4	5	9	10
	Array-Korobov	14	46	33	231	686	2034
	Array-Sobol	123	504	1083	5651	13830	55160
	Array-Sobol-NoGray	55	130	302	1188	3507	11260
0.8	Classique-Korobov-Baker	11	2	15	17	21	26
	Classique-Sobol	3	2	4	6	10	11
	Array-Korobov	15	85	33	337	727	5119
	Array-Sobol	370	1281	3240	19730	57290	233100
	Array-Sobol-NoGray	117	288	996	4580	13210	48660

TAB. 3.3 – Facteurs de réduction de variance pour les méthodes randomisées par rapport à Monte Carlo pour l'estimation du temps moyen d'attente de 100 clients en utilisant $L \approx 2^k$ points. Classique est pour la méthode QMC randomisée classique. On utilise les suites de Sobol, avec ou sans code de Gray [16], un cas particulier des (t, s)-suites, celles de Korobov [163], un cas particulier des règles de quadrillage. Dans certains cas, une transformation dite de baker (boulanger) est utilisée [103].

3.5 Notes

Nous travaillons actuellement sur l'étude du niveau de confiance de l'intervalle obtenu par les méthodes QMC randomisées. Nous cherchons à déterminer si garder une taille d'échatillon fixe et augmenter la taille de la suite à discrépance faible ne détériore pas la qualité de l'intervalle de confiance [225]. De même, nous souhaitons appliquer la méthode pour les chaînes de Markov à des contextes spécifiques, tels que MCMC (Markov Chain Monte Carlo) ou la simulation parfaite. Une combinaison de la méthode avec la technique de ramification des trajectoires pour la simulation d'événements rares est aussi en cours d'élaboration [61].

Chapitre 4

Contributions à la tarification des réseaux de communication

Ce chapitre traite d'un sujet ayant récemment pris de l'importance dans la communauté scientifique, et auquel nous avons accordé beaucoup d'attention depuis quelques années : la tarification des réseaux pour contrôler la congestion et mieux partager les ressources.

4.1 Généralités

Le développement très rapide des réseaux de communication en général et de l'Internet en particulier nécessite une grande attention de la communauté scientifique, tant sur le plan qualitatif (pour le développement d'architectures, de protocoles, de procédures de contrôle et de test) que sur le plan quantitatif (indispensable pour dimensionner correctement ces architectures et ces services : l'évaluation des performances, de la sûreté de fonctionnement, de la qualité de service). L'un des problèmes majeurs est la difficulté à gérer convenablement la « qualité de service » (QoS) pour des applications comme le multimédia ou la téléphonie sur IP en raison du phénomène de congestion. Le protocole TCP avec ses mécanismes de contrôle de flux n'est pas adapté à ces nouvelles utilisations. Si la modification du comportement de TCP est délicate et ne pourra se faire que lentement, différentes approches pour améliorer le comportement du réseau sont suivies par de nombreux groupes de recherche. Ces tentatives passent par la modification de la gestion des files d'attente dans les équipements d'interconnexion. Même si certains pensent actuellement qu'il sera toujours possible de sur-dimensionner le réseau, cette approche semble dangereuse car le sur-dimensionnement incite à l'emploi d'applications de plus en plus consommatrices, et, surtout, il semble peu probable que certains réseaux d'accès puissent être suffisamment dimensionnés. Il faut noter que même en cas de surdimensionnement, les fournisseurs d'accès semblent désireux de différencier les services (quitte à réduire volontairement la qualité des classes non prioritaires dans le but d'offrir un meilleur service à ceux qui peuvent payer le plus). L'approche d'architecture privilégiée actuellement dans le monde de l'Internet est celle de la différenciation de services (DiffServ) définie par l'IETF.

Elle vise à offrir à certaines classes de trafic une proportion importante de la bande passante pour leur fournir une meilleure QoS. Un cadre où la pénurie de ressources est et restera incontestable est celui des réseaux sans fils. La troisième génération de réseaux cellulaires, l'UMTS en Europe, utilise la technologie CDMA (pour *Code Division Multiple Access*), où la congestion se traduit non pas en terme de délais accrus et de pertes de paquets, mais en termes d'interférences et de puissance d'émission à augmenter.

Parmi les domaines jusqu'à présent sans grand intérêt pour la communauté scientifique, celui des procédures de tarification des services réseaux fait maintenant l'objet de recherches intenses, conséquence du fait qu'un meilleur contrôle d'un réseau tel que l'Internet peut, en principe, être obtenu via des mécanismes de tarification sophistiqués. Le système de tarification actuel basé sur un abonnement fixe, indépendant de l'utilisation, est une stimulation à la consommation qui, bien qu'ayant été très utile au démarrage du réseau, devient donc impossible à gérer si l'on souhaite faire de ce dernier un réseau multiservice efficace. Ainsi, toute méthode de différenciation de services doit nécessairement être accompagnée d'une méthode de tarification spécifique sans quoi l'utilisateur demandera toujours la « classe de service prioritaire ». De nombreux modèles de tarification basés sur l'utilisation ont été récemment développés afin de satisfaire différents critères de qualité de service et de répondre à des règles d'utilisation équitable.

Les modèles utilisés sont issus de l'économie. L'une des principales notions est celle d'*utilité* qui permet de quantifier les préférences des utilisateurs vis à vis du service proposé. Un utilisateur acceptera un service si et seulement si son utilité est supérieure au prix à payer. La différence entre ces deux quantités est appelée *surplus* et représente le gain de l'utilisateur, et le *surplus social* ou *bien-être social* est la somme des surplus pour tous les utilisateurs. Dans la suite nous confondrons parfois (abusivement) utilité et surplus. Dans la littérature, les prix sont déterminés de manière à maximiser soit le revenu du réseau, soit le surplus social. Une propriété importante à vérifier est la *compatibilité d'incitation* qui incite les utilisateurs à déclarer au préalable leur vrais besoins, sans quoi leur surplus pourrait diminuer.

Pour des articles introductifs ou généraux sur la tarification, se référer à [56, 59, 63, 67, 101, 186, 208, 91, 222, 220]. Les différents types de tarification proposés dans la littérature sont les suivants.

1. Un premier groupe de chercheurs (voir par exemple [14, 165]) estime que même si le nombre d'utilisateurs (et leur demande) augmente rapidement, la capacité du réseau s'adaptera à la demande, et que si le réseau a survécu jusqu'ici, avec le succès qu'on a connu, pourquoi devrait-on introduire un modèle de tarification coûteux à mettre en place ?
2. Pour un second groupe de personnes, un modèle de tarification incitatif sera nécessaire pour réguler les différentes qualité de service requises, et certains services devront être *garantis*. Comme dans les réseaux ATM, les modèles de tarification pour la garantie de service devront se baser sur le contrôle d'admission [68, 77, 78, 112, 205], et la réservation de ressources. Dans l'Internet, la réservation de ressources peut être réalisée par RSVP [242]. Dans [240] et [79], la réservation est utilisée et seulement les services non garantis utilisent le best effort, adaptée à la volonté de payer des utilisateurs. Dans [178], le contrôle d'admission et la réservation de ressources sont appliqués aux réseaux à pertes ; la programmation dynamique est utilisée pour

calculer les prix optimaux et il est montré que la tarification par créneaux horaires approche efficacement la tarification dynamique optimale (voir aussi les extensions dans [132, 177]).

3. Une autre alternative a été suggérée par A. Odlyzko dans [164]. La proposition appelée *Paris Metro Pricing* (PMP) s'inspire de système de tarification du métro parisien il y a une vingtaine d'années. Le réseau est décomposé en plusieurs réseaux séparés et chaque réseau fonctionne comme l'Internet actuel, mais avec un coût différent, de sorte que les plus chers devraient être moins congestionnés. Il n'y a pas de garantie de qualité de service mais le modèle présente l'avantage de pouvoir être implémenté très facilement et à moindre côut.
4. La méthode de tarification dite du Cumulus (CPS) [183, 184] est aussi une possibilité relativement simple. Un contrat est négocié entre un ISP et l'utilisateur. L'utilisation est mesurée sur des intervalles de temps et des points cumulus (positifs ou négatifs) sont donnés selon le respect ou non du contrat. En fonction des points, un coût supplémentaire peut être imposé et le contrat peut être renégocié.
5. Un autre groupe suggère d'utiliser la tarification basée sur le niveau de priorité, sans réservation de ressource [29, 52, 53, 86, 105, 155]. Chaque classe est servie selon cette priorité à chaque nœud du réseau. La tarification basée sur la priorité peut se décomposer en deux sous-classes :
   (a) la tarification prioritaire statique où le prix par classe est fixé à l'avance. Dans [29], à chaque client est alloué un quota de paquets de haute priorité (selon un contrat), et en cas de sur-consommation, une pénalité est appliquée le mois suivant. Dans [52], un marquage de priorité de traitement est assigné à tout paquet selon le type de service, mais aussi un marquage de priorité de rejet pour les services pouvant supporter des pertes. Dans [148, 147, 146], un modèle à événement discret est décrit où le temps est discrétisé par petits intervalles. Les prix optimisant le bénéfice du réseau sont calculés. Dans [87, 155], une tarification optimale possédant la propriété de compatibilité d'incitation est mise en place pour la file M/M/1 multi-classe.
   (b) La deuxième sous-classe est la tarification prioritaire dynamique où la prix par paquet dépend du niveau de trafic dans le réseau [86].
6. Les enchères pour la priorité ont été proposées dans [153, 154]. Les utilisateurs font une offre pour chaque paquet et seulement les plus élevées sont acceptées. Dans [118, 197, 198], les enchères par paquet sont remplacées par des enchères pour la bande passante afin de réduire la complexité de la méthode. Les propriétés d'efficacité (en terme de surplus social), de stabilité et de compatibilité d'incitation sont démontrées dans le cas d'un nœud, mais aussi de réseaux interconnectés.
7. Un dernier groupe propose la tarification pour le trafic élastique basée sur le taux de transfert. Dans le travail de Kelly et al. [111, 113], les utilisateurs décident du prix qu'ils souhaitent payer et recoivent alors un taux de transmission (équitable) calculé de manière distribuée par le réseau. Dans le travail de Low et al., [19, 18, 135, 134, 136], le problème est inversé : les utilisateurs demandent un taux de transmission et le réseau calcule le prix à payer.

Une vaste littérature s'est développée ces dernières années sur la future évolution de l'Internet et l'intégration de différents services ainsi que sur les problèmes d'équité dans la partage des ressources [30, 150, 151, 157]. Dire quel groupe de méthode sera effectivement implanté est pour le moment un pari. Cependant, comme dans [86], nous pensons que les arguments en faveur d'un simple sur-dimensionnement de la capacité sont dangereux (surtout dans un contexte sans fils). De plus la réservation de ressources et la garantie de qualité de service est difficile à mettre en place. Nous sommes donc enclins à penser que la future tarification sera un schéma de tarification sans réservation de ressources.

Nous pouvons classifier nos contributions dans le domaine de la tarification en deux classes principales : la tarification multi-classes, avec différents types de politique d'ordonnancement possibles, ou la tarification basée sur la théorie des jeux, avec comme sous-thème moteur les enchères pour la bande passante.

On peut cependant aussi noter que nous avons porté une attention particulière au modèle dit du cumulus, qui est un modèle intermédiaire entre le modèle au forfait et le modèle basé sur le volume de consommation. Dans [96, 95], nos contributions on été les suivantes :

- nous avons étudié le modèle économique, tant du point de vue utilisateur que du point de vue réseau, en nous focalisant sur l'influence des seuils de consommation pour l'attribution des points de cumulus sur les bénéfices respectifs des protagonistes. Dans notre modèle, à chaque point cumulus est associé une valeur financière, de même qu'un coût est associé aux mesures de la consommation des utilisateurs pour le réseau. Toutes ces fonctions de coût interviennent dans les fonction d'utilité des utilisateurs et le bénéfice du réseau.
- Par une analyse mathématique, nous montrons que le modèle original du cumulus n'a pas la propriété de compatibilité d'incitation, dans le sens où les utilisateurs sont incités à mentir sur la quantité de ressources qu'ils pensent consommer.
- Par une modification mineure du modèle, nous donnons une condition suffisante sur les valeurs des seuils pour l'attribution des points de cumulus afin que la propriété de compatibilité d'incitation soit vérifiée (ce qui permet de mieux gérer le réseau). De manière très intéressante, cette condition est indépendante de la forme des fonctions d'utilité des utilisateurs (dont la forme est difficile à connaître en pratique).
- Sous cette contrainte, nous établissons comment les valeurs optimales (pour maximiser le revenu) peuvent être déterminées. De manière générale, obtenir des résultats théoriques a été impossible ; nous utilisons alors un algorithme de type recuit simulé. Néanmoins, dans des cas particuliers, tels que des seuils symétriques ou linéaires, nous avons été capable d'obtenir des valeurs explicites.

4.2 Tarification multi-classes et politiques d'ordonnancement

Une des manières de réaliser une différenciation réside dans la séparation des services en plusieurs classes, traitées et tarifées différemment. Différentes politiques d'ordonnancement sont envisageables, et certaines ont déjà fait l'objet d'attention (comme nous avons pu le voir dans la section précédente). Nous avons traité la politique de séparation totale entre classes, la politique de priorité stricte, et celles du processeur partagé généralisé ou

discriminatoire.

4.2.1 Traitement par séparation logique des classes (PMP)

Ce modèle, basé sur le principe (révolu) de tarification du métro parisien, sépare le réseau en sous réseaux logiquement indépendants [164]. La propositon initiale du PMP [164] consiste à partitionner le réseau en sous réseaux logiquement séparés, chacun ayant une fraction fixe du réseaux d'origine. Chaque sous réseau serait géré avec les protocoles de l'actuel Internet. Il n'y a pas de garantie formelle de QoS, mais on peut imaginer que si les réseaux sont tarifés à différents prix, les moins chers seront moins congestionnés, produisant donc une meilleure QoS. Le nom du modèle, *Paris Metro Pricing*, provient des règles du métro parisien il y a une vingtaine d'années, où les trains étaient composés de wagons rigoureusement identiques, mais séparés en deux classes à l'accès différents, de sorte que moins d'usagers étaient en première classe : il en résultait un meilleure perception de QoS.

Nous avons étudié mathématiquement un modèle de PMP dans le but de déterminer les prix optimisant le revenu d'un fournisseur [189, 188]. Ce problème de déterminer le « bon » ensemble de prix pour une partition donnée peut être résumé ainsi : pour être efficace, un réseau PMP doit fixer les prix des meilleures classes suffisamment hauts pour qu'elles soient peu chargées, mais pas trop haut, car sinon ces classes resteront vides. Dans notre modèle, les sous réseaux sont représentés par un unique goulot d'étranglement et les clients (des paquets de données) choisissent leur réseau selon le prix, mais aussi le délai moyen qui est supposé avoir un impact économique. Formellement, il y a N classes/sous réseaux, et le prix d'entrée pour un paquet à la classe i est p_i ($1 \leq i \leq N$). Une utilité est associée à chaque paquet, représentée par une variable aléatoire U. Le taux d'arrivée total *potentiel* est $\tilde{\lambda}$, correspondant au cas où l'accès est gratuit. Pour l'accès à la classe i, le coût total est $p_i + \gamma d_i$, où d_i est le délai moyen au sous réseau i et γ est une constante. Un paquet n'entre au sous réseau i que si $i = \text{argmin}_j p_j + \gamma d_j$ et $U \geq p_i + \gamma d_i$ (il choisit donc le moins coûteux). Nous avons aussi considéré le cas de plusieurs types d'applications, telles que U et γ soient indexés par j. Nous avons obtenu des conditions nécessaires et suffisantes pour la stabilité du système et analysé le problème de la maximisation du revenu du réseau. Nous avons comparé le revenu optimal avec celui qui serait obtenu dans le cas d'un unique réseau et, de manière surprenante, nous avons pu voir que le revenu sera toujours supérieur si le réseau ne partitionne pas son réseau en sous réseaux. Une première explication de ce résultat proviendrait de la non utilisation des ressources quand une classe est vide et pas l'autre car alors seulement une partie de la bande passante est utilisée (phénomène connu en théorie des files d'attente). Nos travaux ont donc une conclusion négative similaire à celle obtenue dans [76], où il a été montré que le modèle n'est pas viable dans un contexte de concurrence entre fournisseurs.

4.2.2 Traitement par priorités strictes

Dans le cadre de le tarification par classes, l'ordonnancement par priorités strictes est certainement le cas qui a reçu le plus d'attention dans la littérature. L'idée de base des méthodes de tarification est d'allouer de manière efficace les ressources entre des utilisateurs

égoïstes. L'utilisation de la ressource par un utilisateur impose un coût aux autres, connu sous le nom d'*externalité négative*, ou *coût de congestion*. En général, ce coût est basé sur le délai qu'impose l'utilisateur à ceux déjà présents. La plupart des modèles utilisent le coût dit *marginal*, une des notions fondamentales en économie. Cependant ce coût marginal présente le désavantage de ne pas partager « équitablement » les coûts entre utilisateurs. Il nous a donc paru intéressant d'étudier/appliquer un modèle alternatif de partage de coûts, connu en économie sous le nom de mécanisme de Aumann-Shapley [98].

Le modèle de base est une file d'attente multi-classes. Nous considérons N classes. Les clients de classe i arrivent selon un processus de Poisson de taux λ_i, $1 \leq i \leq N$. Le temps de service pour la classe i suite une loi exponentielle de moyenne c_i. Nous supposons tous les temps d'inter-arrivée et de service indépendants. Chaque classe i est caractérisée par un coût (unitaire) de délai v_i. Nous considérons une règle de priorité non-préemptive où les classes sont rangées de sorte que $\frac{v_1}{c_1} \geq \frac{v_2}{c_2} \geq \ldots \geq \frac{v_N}{c_N}$, assignant les plus haute priorité à la classe 1, puis la classe 2, et ainsi de suite. Soit $W_i(\underline{\lambda})$ le délai moyen (temps de réponse) pour un client/paquet de classe i job quand le vecteur des taux d'arrivées est $\underline{\lambda} = (\lambda_1, \ldots, \lambda_N)$. Selon un résultat classique des files d'attente,

$$W_i(\underline{\lambda}) = \frac{\sum_{j=1}^{N} c_j^2 \lambda_j}{(1 - \sum_{j=1}^{i-1} c_j \lambda_j)(1 - \sum_{j=1}^{i} c_j \lambda_j)} + c_i.$$

L'idée que nous avons cherché à développer est d'appliquer le mécanisme d'Aumann-Shapley à la fonction de coût globale

$$L(\underline{\lambda}) = \sum_{i=1}^{N} v_i \lambda_i W_i(\underline{\lambda}) \tag{4.1}$$

représentant le coût global induit par les clients dans le système. Les coûts de Aumann-Shapley $C_i = \int_0^1 \frac{\partial L}{\partial \lambda_i}(t\underline{\lambda})dt$ associés à chaque classe i sont les seuls à vérifier simultanément les cinq axiomes suivants :

- le coût total est partagé entre les utilisateurs $\sum_{i=1}^{N} \lambda_i C_i = L(\underline{\lambda})$;
- les coûts sont additifs ;
- les coûts sont positifs ou nuls ;
- les coûts sont « consistants » : séparer la ressource n'a pas d'effet sur la répartition des coûts ;
- résistance au changement d'échelle : tout changement d'échelle dans la ressource produit un changement équivalent dans le partage.

Notre objectif est ici de maximiser le surplus social. Soit $V_i(\lambda_i)$ la valuation agrégée des clients de classe i dans le système. Soit $V(\underline{\lambda}) = \sum_{i=1}^{N} V_i(\lambda_i)$. On cherche donc à maximiser

$$\max_{\underline{\lambda}} \left(V(\underline{\lambda}) - L(\underline{\lambda}) \right). \tag{4.2}$$

Soit le vecteur de prix pour l'accès à chaque classe $\underline{p} = (p_1, \ldots, p_N)$. Demande et prix sont liés par une relation de demande. Les prix sont choisis, ou plus exactement calculés, par le mécanisme de Aumann-Shapley appliqué au surplus marginal :

$$p_i = \int_0^1 \left(V_i'(\lambda_i t) - v_i W_i(\underline{\lambda} t) \right) dt, \tag{4.3}$$

donnant la relation de demande pour la classe i

$$\frac{1}{\lambda_i}\int_0^{\lambda_i} V_i'(u)du = \frac{V_i(\lambda_i)}{\lambda_i} = p_i + v_i \int_0^1 W_i(\underline{\lambda}t)dt.$$

Ceci signifie que la valuation moyenne $\frac{V_i(\lambda_i)}{\lambda_i}$ est égale au coût composé un prix direct et du coût du délai $v_i \int_0^1 W_i(\underline{\lambda}t)dt$ ressenti par un client de classe i. Connaissant les V_i et v_i, la demande est donc contrôlée par (4.3).

Les prix optimaux sont calculés dans [98], pour le cas de services homogènes et hétérogènes (c'est à dire des temps de service moyens différents) et possèdent la propriété de partager le coût de congestion (4.1) selon le mécanisme de Aumann-Shapley. Ces prix sont donc obtenus par le calcul de $\underline{\lambda}$ maximisant le surplus social en $\underline{\lambda}$, et utilisation de la relation de demande (nous n'explicitons ici pas les prix, car la formule n'est pas très informative). Les propriétés de compatibilité d'incitation sont prouvées être vérifiées dans les deux cas. Des extensions à la tarification dynamique et à tout un réseau sont aussi obtenues [98].

4.2.3 Traitement par processeur partagé généralisé (GPS)

Les implémentations de l'architecture DiffServ telle qu'elle est définie, sont réalisées en utilisant l'un des deux types d'ordonnancement suivants : les priorités strictes et le « weighted fair queuing » qui peut être approché par un processeur généralisé par classe (et une gestion FIFO au sein de chaque classe). Dans [94], nous avons repris un modèle pour l'optimisation du revenu dans le cas des priorités strictes [145], et déterminé les prix maximisant le revenu dans le cas d'un processeur partagé généralisé (ou plus exactement une approximation de ce cas) afin de comparer les deux politiques d'ordonnancement sur le plan économique. Dans ce but, on considère deux types de trafic, que l'on appelle (abusivement) *voix*, qui a les exigences les plus fortes en terme de délai, et *données*. On indexe par v (respectivement d) les notions relatives à trafic de type voix (respectivement données). L'utilité (ou plutôt ici surplus) par paquet dépend du délai moyen $I\!E D$ d'attente au goulot d'étranglement, mais aussi du prix p par paquet selon les relations quasi-linéaires

$$U_d(I\!E D) = \frac{1}{(I\!E D)^{\alpha_d}} - p, \tag{4.4}$$

$$U_v(I\!E D) = \frac{1}{(I\!E D)^{\alpha_v}} - p, \tag{4.5}$$

avec $0 < \alpha_d < \alpha_v$ de sorte que le type de service indexé par v a une préférence plus forte pour les petits délais. On suppose qu'une source envoie du trafic dès que son utilité est positive ou quitte le système si son utilité est négative. Ainsi, le nombre de sources de chaque type (supposé infiniment divisible) à l'équilibre est tel que l'utilité de ce type de trafic est nulle. On considère que le goulot d'étranglement est représenté par une file M/M/1 de taux de service μ et qu'une source de type-v envoie des paquets avec un taux λ_v alors qu'une source de type-d envoie des paquets avec un taux λ_d.

On suppose que le réseau sépare le trafic en deux classes (appelées 1 et 2) traitées différemment. Le cas où le traitement se fait par priorités strictes a été étudié dans [145]. Nous

considérons ici le cas d'une politique GPS où une proportion γ ($0 \leq \gamma \leq 1$) est réservée à la file 2, et donc $1-\gamma$ est alloué à la file 1. Comme malheureusement il n'existe pas de forme close connue pour exprimer le délai dans une telle file d'attente, nous l'approchons par un modèle de deux files séparées pour chaque classe (de sorte que la capacité d'une classe n'est pas allouée à l'autre quand elle est inutilisée), l'une avec un taux de service $(1-\gamma)\mu$ et l'autre avec un taux $\gamma\mu$. Cette approximation est justifiée dans le cas d'une intensité de trafic importante, constituant en fait la motivation pour appliquer une différenciation de service.

Soit p_i le prix d'accès à la file i ($i = 1, 2$). Il y a trois variables à considérer pour l'optimisation du revenu $p_1 N_1(p_1) + p_2 N_2(p_2)$: p_1, p_2 et γ. Nous considérons deux cas de gestion des files :

- celui des files *dédiées* : la file 1 est allouée au trafic de type-v et la file 2 au trafic de type-v. Le nombre de sources de chaque type (annulant les utilités (4.4) et (4.5)) est donné par

$$N_2(p_2) = N_d(p_2) = \frac{\gamma\mu - \sqrt[\alpha_d]{p_2}}{\lambda_d}, \tag{4.6}$$

  pour $p_2 < (\gamma\mu)^{\alpha_d}$, 0 sinon, pour le trafic de type données et

$$N_1(p_1) = N_v(p_1) = \frac{(1-\gamma)\mu - \sqrt[\alpha_v]{p_1}}{\lambda_v}, \tag{4.7}$$

  pour $p_1 < ((1-\gamma)\mu)^{\alpha_v}$, 0 sinon. Les valeurs optimales de p_1, p_2 et γ sont alors déterminées [94].
- Le cas des files *ouvertes* : les sources sont libres de choisir la file de leur choix. Nous avons pu montrer que, si $p_1 < 1$, seuls les paquets de type-d entrent dans la file 1. De même, si $p_2 < 1$, seuls les paquets de type-d entrent dans la file 2. Inversement, si $p_1 > 1$ (resp. $p_2 > 1$), seuls les paquets de type-v entrent dans la file 1 (resp. la file 2). Le nombre de sources de chaque type dans chaque file peut alors être déterminé en fonction des valeurs de p_1 et p_2 selon des règles équivalentes à (4.6) et (4.7), puis les prix p_1, p_2 et la répartition de bande passante γ permettant d'optimiser le revenu.

Le résultat le plus marquant que nous avons prouvé est cependant que, afin d'optimiser le revenu, utiliser l'ordonnancement par priorités strictes produira toujours un revenu (optimal) supérieur au cas GPS (ou tout du moins l'approximation en classes séparées) avec une répartition $0 < \gamma < 1$ pour une des deux classes. Ce résultat est vérifié à la fois pour le cas dédié et pour le cas ouvert.

4.2.4 Traitement par processeur partagé discriminatoire (DPS)

La politique de traitement par processeur partagé discriminatoire consiste à assigner un poids γ_i à chaque client de classe de service i, et à lui assigner une proportion du serveur $\frac{\gamma_i}{\sum_j N_j \gamma_j}$ avec N_j le nombre de clients de classe j. Cette politique est intéressante à plusieurs égards. Tout d'abord, il existe contrairement à GPS une forme explicite du délai moyen. Ensuite, cette politique est réputée modéliser fidèlement le comportement des flux TCP à un routeur [11, 36], résolvant la limitation de la modélisation au niveau paquet par une file M/M/1 d'un réseau tel que l'Internet.

Dans [97], nous nous sommes limités au cas de files dédiées. Nous avons encore considéré le cas de deux classes de trafic appelées voix et données avec fonctions d'utilité U_d et U_v données par les équations (4.4) et (4.5). Les files sont pour simplifier indexées par v et d (au lieu de 1 et 2). L'analyse au niveau flux pour la modélisation de TCP au lieu du niveau paquet pour la sous section précédente nécessite une légère modification du modèle. Les sessions de chaque type s'ouvrent ici selon des processus de Poisson. N_d et N_v représentent les nombres moyens de sessions données et voix dans le réseau en stationnaire (au lieu du nombre exact précédemment). De plus, λ_d et λ_v sont ici les tailles moyennes des flux données et voix (suivants des lois aléatoires indépendantes). On suppose que les connections ont le même RTT (Round Trip Time). Les autres paramètres sont identiques au cas GPS. Soit γ le poids du serveur assigné à une connection de type données et $1-\gamma$ celui assigné aux connections de type voix (nous avons sans perte de généralité normalisé les poids de manière à n'avoir qu'un paramètre). Une forme close du délai pour une telle file est donnée dans [90, page 86] par

$$I\!E D_v = \frac{\left(1 + \frac{\lambda_d N_d (2\gamma - 1)}{\mu - (1-\gamma)\lambda_v N_v - \gamma \lambda_d N_d}\right)}{\mu - \lambda_v N_v - \lambda_d N_d}, \tag{4.8}$$

et

$$I\!E D_d = \frac{\left(1 - \frac{\lambda_v N_v (2\gamma - 1)}{\mu - (1-\gamma)\lambda_v N_v - \gamma \lambda_d N_d}\right)}{\mu - \lambda_v N_v - \lambda_d N_d}. \tag{4.9}$$

L'analyse du nombre (moyen) d'utilisateurs de chaque type est ici bien plus complexe que dans le cas priorité ou l'approximation de GPS par files séparées en raison de la forte interaction entre les classes (car le délai de chaque classe dépend du nombre de sessions dans l'autre classe). Pour des prix fixés, nous avons donc un jeu (voir la section suivante) entre les types de trafic pour déterminer le nombre (moyen) de sources, annulant l'utilité. Nous avons pu prouver l'existence et la convergence vers un équilibre unique entre les types de trafic, dont la valeur dépend des paramètres. Ce point d'équilibre (N_d^*, N_v^*) est spécifié Figure 4.1 et est facilement exprimable sous forme close [97]. On peut observer Figure 4.1 que si les courbes $U_d = 0$ et $U_v = 0$ ne se coupent pas, il n'y a qu'un seul type de trafic dans la file. Le cas désirable est donc le cas (c) de la figure, où les deux courbes se croisent car les deux types de trafic sont présents. Appelons ce cas l'équilibre non trivial. Nous avons pu donner une condition nécessaire et suffisante sur les prix pour obtenir cet équilibre. Pour simplifier les notations, posons $q_d = (p_d)^{1/\alpha_d}$ et $q_v = (p_v)^{1/\alpha_v}$. L'équilibre sera non trivial si et seulement si p_d et p_v vérifient

$$0 < q_v < \quad \mu, \tag{4.10}$$

$$\sqrt{\mu q_v + (\frac{\mu - q_v}{2}\frac{1-\gamma}{\gamma})^2} - \frac{\mu - q_v}{2}\frac{1-\gamma}{\gamma} < \quad q_d, \tag{4.11}$$

$$\frac{\mu\gamma q_v + (1-\gamma)q_v^2}{\gamma q_v + \mu(1-\gamma)} > \quad q_d. \tag{4.12}$$

L'objectif pour le réseau est de déterminer p_d^*, p_v^* et γ^* maximisant le revenu

$$R^* = \max_{p_d, p_v, \gamma} \lambda_d N_d(p_d, p_v, \gamma) p_d + \lambda_v N_v(p_d, p_v, \gamma) p_v$$

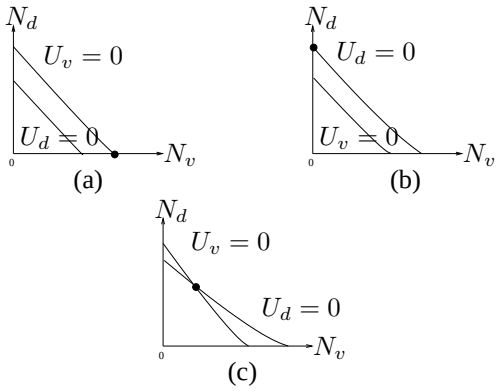


FIG. 4.1 – Courbes $U_d = 0$ et $U_v = 0$ et point d'équilibre donnant les nombres de sessions de chaque type en stationnaire. Le point d'équilibre est spécifié par le point épais.

sous contraintes $p_v, p_d \geq 0$ et $\gamma \in [0, 1]$. Nous avons pu réaliser cette optimisation numériquement. Une importante conjecture est que, comme pour le cas de GPS, le revenu est maximisé pour $\gamma^* = 0$ ou 1, qui signifie qu'une priorité stricte est encore préférable.

4.3 Tarification et enchères

En toute généralité, la théorie des jeux se propose d'étudier toute situation dans laquelle des agents rationnels interagissent. Son champ d'application est donc très vaste, englobant en particulier toute la micro-économie traditionnelle. Dans les télécommunications, la théorie des jeux est apparue récemment, en raison de la dérégulation des réseaux, mais surtout en raison de l'apparition du mode datagram (gestion par paquet) qui réduit les coûts de gestion, mais introduit des phénomènes forts de concurrence/compétition. La tarification des réseaux est un cadre naturel d'application de la théorie des jeux où les décisions et service obtenus de chacun sont déterminés par les prix et les choix des autres participants. Une notion importante est celle d'équilibre de Nash, qui est un état dans lequel aucun joueur n'a d'intérêt à unilatéralement changer de décision (sans quoi son surplus diminuerait). C'est donc un état dans lequel, étant donné les choix des autres participants, un joueur maximise son surplus. Les sous-sections précédentes présentaient un équilibre de Nash particulier où les joueurs entraient ou sortaient (le jeu se situait donc à leur niveau) mais où l'équilibre se situait au niveau du nombre de sources de chaque type. Nous regardons ici des types de jeux plus directs où les joueurs misent un prix et/ou une quantité de ressource, donc de de type enchères.

4.3.1 Différenciation de services par tarification d'un routeur RED

Dans la littérature sur la gestion active des files d'attente pour contrôler la qualité de service dans le réseau Internet, les tampons RED occupent une place prépondérante. Schématiquement, ces tampons rejettent des paquets aléatoirement, selon une probabilité croissante avec l'occupation moyenne du tampon de manière à prévenir au mieux l'engorge-

ment. Plus formellement, il y a deux seuils $q_{\min}$ et $q_{\max}$ tels que la probabilité de perte est 0 si la longueur moyenne de la file q est en dessous de $q_{\min}$, 1 si elle est au-dessus de $q_{\max}$, et $P_{max}(x - q_{\min})/(q_{\max} - q_{\min})$ si c'est x avec $q_{\min} < x < q_{\max}$, P_{max} étant la probabilité de rejet au seuil q_{max}. Le mécanisme est illustré Figure 4.2.

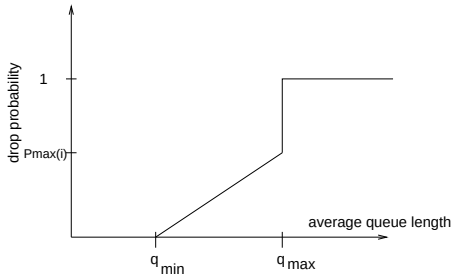


FIG. 4.2 – Probabilité de rejet en fonction de l'occupation moyenne q pour un tampon RED

L'idée de base que nous avons développé dans [6, 7] consiste à considérer des valeurs de P_{max} différentes par flux, $0 < P_{max}(i) \leq 1$ pour le flux i, où la valeur $P_{max}(i)$ dépendrait du prix pour le flux i : plus le prix est élevé, plus $P_{max}(i)$ est petit, et donc moins les paquets du flux sont rejetés.

Nous avons considéré un ensemble $\mathcal{N}$ de N flux TCP et un ensemble $\mathcal{I}$ de I flux temps réel utilisants UDP. Tous ces flux utilisent un tampon RED, qui les traite donc différemment. Nous supposons $q_{\min}$ et $q_{\max}$ fixes, et par conséquent seul $P_{max}(i)$ varie avec i. Soit t_i la pente de la partie linéaire in Figure 4.2 pour tout flux i,

$$t_i = \frac{P_{max}(i)}{q_{\max} - q_{\min}}$$

et $\mathbf{t} = (t_i, i \in \mathcal{I} \cup \mathcal{N})$ le vecteur décrivant ces pentes pour tous les flux. Dénotons par μ le taux de service du goulot d'étranglement (RED) du réseau. Le débit des flux TCP est dirigé par la formule classique

$$\lambda_i = \frac{1}{R_i}\sqrt{\frac{\alpha}{p_i}}, \quad i \in \mathcal{N}, \tag{4.13}$$

où R_i et p_i sont les temps d'aller-retour et probabilité de perte du flux TCP i. En pratique, α est $3/2$ ou $3/4$. Les débits des flux temps réels λ_i, pour $i \in \mathcal{I}$, quant à eux, ne sont pas contrôlés. En général, comme le goulot d'étranglement est vu comme un file fluide,

$$\sum_{j \in \mathcal{I} \cup \mathcal{N}} \lambda_j (1 - p_j) = \mu.$$

Dans la partie linéaire de RED, on aboutit au système d'équations

$$\left\{ \begin{array}{rcll} \sum_{j \in \mathcal{I} \cup \mathcal{N}} \lambda_j (1 - p_j) & = & \mu & \\ p_i & = & t_i (q - q_{\min}), & \forall i \in \mathcal{I} \cup \mathcal{N} \end{array} \right.$$

avec $(N + I + 1)$ inconnues : q (occupation moyenne de la file), et $p_i,\ i \in \mathcal{I} \cup \mathcal{N}$, avec λ_i, $i \in \mathcal{N}$ donné par (4.13). Nous avons prouvé que ce système admet une unique solution. De plus, dans le cas où il n'a pas que des flux TCP ou que des flux temps réel, nous avons obtenu une forme explicite pour les inconnues (la solution peut être trouvée numériquement sinon).

Tout utilisateur i doit décider du prix qu'il souhaite payer (et par conséquent la pente t_i de la partie linéaire de RED). Il en résulte un jeu. **t** est le vecteur de stratégie pour tous les flux et $(t_i, [\mathbf{t}]_{-i})$, le vecteur où la stratégie du flux i est t_i (mise en évidence) alors que les stratégies des autres flux sont donnés par le vecteur $[\mathbf{t}]_{-i}$. L'utilité (ou surplus ici) pour le flux i, $U_i(t_i, [\mathbf{t}]_{-i})$ est donnée par

$$a_i\lambda_i(1 - p(t_i, [\mathbf{t}]_{-i})) - b_i p(t_i, [\mathbf{t}]_{-i}) - d(t_i)$$

où le premier terme représente le débit réel obtenu, le deuxième la désutilité provenant du taux de perte, et le dernier le prix $d(t_i)$ à payer. Pour les flux TCP, on prend $b_i = 0$ car l'impact des pertes est déjà pris en compte dans le débit λ_i. Nous avons alors pu obtenir des conditions suffisantes afin d'être dans la région linéaire de RED et d'obtenir un équilibre de Nash.

Un objectif pour le réseau est alors de déterminer la fonction de prix d qui maximisera ses bénéfices :

$$c(\mathbf{t}^*) = \arg\max_d \sum_{i=1}^{I} d(t_i^*),$$

où $\mathbf{t}^*$ est l'équilibre de Nash pour la fonction d. Numériquement, pour certaines classes de fonctions (comme par exemple $d(t) = \delta/e^t$), et distributions d'utilisateurs (avec paramètres aléatoires a_i, b_i...), il est possible de déterminer les paramètres (δ pour l'exemple précédent) permettant d'optimiser le revenu.

4.3.2 Enchères pour la bande passante

Enchérir pour la bande passante est une alternative qui semble intéressante depuis la proposition initiale de McKie Mason et Varian [154] où des enchères par paquet (définissant la priorité de traitement) sont définies, mais sont coûteuses en terme de complexité. Nous nous intéressons ici aux enchères pour la bande passante, moins complexes à mettre en place. D'autres techniques d'enchères existent dans la littérature, mais ne seront pas développées dans ce document (voir [55, 185, 201]).

Enchères progressives au second prix (PSP)

Les enchères progressives au second prix ont été initialement développées à l'Université de Columbia [118, 197] et peuvent être résumées comme suit. Considérons une ressource de capacité Q. Supposons que I joueurs sont en compétition pour l'accès à la ressource, et utilisent pour le partage un mécanisme d'enchères où ils soumettent leurs enchères séquentiellement. Tout joueur i soumet une enchère $s_i = (q_i, p_i)$ où q_i est la capacité que le joueur i demande et p_i le prix *unitaire* qu'il propose. Posons $s = (s_1, \cdots, s_I)$ le profil des enchères et $s_{-i} = (s_1, \cdots, s_{i-1}, s_{i+1}, \cdots, s_I)$ le profil quand le joueur i est

exclu du jeu. On écrit parfois $s = (s_i; s_{-i})$ pour mettre en valeur l'enchère du joueur i. Pour $y \geq 0$, on définit

$$\underline{Q}_i(y; s_{-i}) = \left[Q - \sum_{k \neq i \ : p_k \geq y} q_k \right]^+$$

la quantité restante à un prix inférieur à y. Le mécanisme d'enchères progressives au second prix alloue au joueur i une bande passante

$$a_i(s) = \min(q_i, \underline{Q}_i(p_i; s_{-i})), \tag{4.14}$$

de sorte que la ressource est allouée en priorité aux enchères les plus hautes et le coût imposé au joueur i est calculé selon le principe du second prix, c'est à dire par les enchères des joueurs exclus par la présence du joueur i, i.e.,

$$c_i(s) = \sum_{j \neq i} p_j [a_j(0; s_{-i}) - a_j(s_i; s_{-i})]. \tag{4.15}$$

Il est supposé qu'un coût ε est imposé chaque fois qu'un joueur soumet une nouvelle enchère, et que chaque joueur i a une contrainte de budget b_i : $c_i(s_i, s_{-i}) \leq b_i$. Soit $\mathcal{S}_i(s_{-i})$ l'ensemble des enchères des joueurs vérifiant cette contrainte.

On suppose que le joueur i cherche à maximiser sa fonction d'utilité $u_i(s) = \theta_i(a_i(s)) - c_i(s)$, où θ_i est la valuation du joueur i pour son allocation, i.e. $\theta_i(a_i(s))$ est le prix que le joueur i est prêt à payer pour obtenir $a_i(s)$.

Enfin, une enchère $s_0 = (Q, p_0)$ est introduite, signifiant que le vendeur n'allouera pas la bande passante en dessous d'un prix minimum p_0, appelé *prix de réserve*. Le vendeur est donc vu comme un (nouveau) joueur avec fonction de valuation $\theta_i(q) = p_0 q$.

Sous de hypothèses de régularité des fonctions θ_i, les propriétés suivantes avaient été démontrées :

- **Compatibilité d'incitation**. Définissons une enchère $s_i = (q_i, p_i)$ *sincère* par une enchère vérifiant $p_i = \theta_i'(q_i)$, c'est à dire la valuation marginale de la quantité demandée. Soit

  $$Q_i(y; s_{-i}) = \left[Q - \sum_{k \neq i \ : p_k > y} q_k \right]^+$$

  et $G_i(s_{-i}) = \sup \{z : z \leq Q_i(\theta_i'(z), s_{-i}) \text{ et } c_i(z) \leq b_i\}$. $\forall 1 \leq i \leq I, \forall s_{-i}$ tel que $Q_i(0, s_{-i}) = 0$, $\forall \varepsilon > 0$, il existe une ϵ-meilleure réponse sincère, c'est à dire une enchère sincère $t_i = (v_i, w_i)$ qui garantit que i va à ϵ près obtenir la meilleure utilité possible :

  $$t_i(s_{-i}) = (v_i = [G_i(s_{-i}) - \varepsilon/\theta_i'(0)]^+, \omega_i = \theta_i'(v_i))$$

- **Convergence**. Si tous les joueurs soumettent comme nous venons de décrire et de manière séquentielle, le jeu converge vers un ϵ-équilibre de Nash, où un ϵ-équilibre de Nash est un profil d'enchères s tel que $\forall i \in \mathcal{I}$,

  $$\begin{cases} s_i \in \mathcal{S}_i(s_{-i}) \\ u_i(s_i\ ; s_{-i}) \geq u_i(s_i'\ ; s_{-i}) - \epsilon, \forall s_i' \in \mathcal{S}_i(s_{-i}) \end{cases}$$

– **Optimalité**. Pour le ϵ-équilibre de Nash précédent, l'utilité totale résiduelle des joueurs (vendeur inclus) $\sum_{i\in\mathcal{I}\cup\{0\}} \theta_i(a_i)$ est maximisée.

Dans [118, 197], ce mécanisme est étendu à des fournisseurs en interaction (représentés par un goulot d'étranglement), et en compétition pour l'achat de la bande passante à des vendeurs de ressources.

Nos contributions sur les enchères au second prix ont été les suivantes.

Dans [221], nous avons modifié la règle d'allocation (4.14) par

$$a_i(s) = \min\left(q_i, \frac{q_i}{\sum_{k\,:p_k=p_i} q_k} Q_i(p_i; s_{-i})\right) \tag{4.16}$$

tout en gardant la fonction de coût (4.15). Cette modification a deux motivations :

– La règle d'allocation (4.14) n'allouait pas optimallement la bande passante quand des joueurs misait le même prix unitaire. Ainsi, par exemple si $Q = 100$, $I = 2$, $s_1 = (60, 4)$, $s_2 = (70, 4)$, l'allocation était $a_1(s) = 30$ et $a_2(s) = 40$, conduisant donc à une sous utilisation potentielle de la bande passante. Notre nouvel algorithme corrige ce défaut, en allouant la bande passante proportionnellement entre les utilisateurs ayant soumis le même prix unitaire.
– La règle d'allocation (4.14) entraîne une erreur dans les preuves des propriétés du mécanisme. Nous avons pu montrer que, par notre modification, les propriétés s'avèrent formellement vérifiées.

Dans [137, 139], nous avons montré que le PSP présente encore deux problèmes majeurs. Tout d'abord, le premier joueur a tout intérêt à surestimer le prix unitaire de sa bande passante (en déclarant vouloir toute la bande passante à un prix très élevé). Ainsi les joueurs suivants ne chercherons pas à entrer dans le jeu, car le prix unitaire du premier joueur est trop élevé, et au final, le premier joueur récupérera toute la bande passante à un coût faible car selon la règle du second prix, il paiera le prix de réserve. Deuxièmement, nous avons pu observé que, selon l'ordre dans lequel les joueurs soumettent leurs enchères, l'équilibre de Nash résultant peut être différent, ainsi que (par conséquent) le revenu du vendeur ; il serait pourtant souhaitable de prédire ce revenu. Pour remédier à ces problèmes, nous avons proposé dans [139] que les joueurs qui n'obtiennent pas du tout de bande passante soumettent, sans frais, une enchère correspondant à ce qu'il auraient soumis s'ils étaient seuls de manière à maximiser leur utilité. En utilisant cette stratégie, le problème du premier joueur est résolu. En effet, si le premier joueur a soumis (Q, p_{max}), le second, exclu du jeu, soumettra $(q_2, \theta'_2(q_2))$, de sorte que le joueur 1 devra payer $p_0 Q + q_2(\theta'_2(q_2) - p_0)$, qui peut être plus grand que $\theta'_1(q_1)$, démontrant ainsi qu'il n'a pas d'intérêt à surestimer son enchère. De plus l'unicité (à ε près) de l'équilibre est alors aussi démontrée.

Dans [60], nous avons travaillé sur le seul degré de liberté restant pour le vendeur de la ressource dans le mécanisme PSP : le prix de réserve p_0. Déterminer le bon prix de réserve pour optimiser les bénéfices du vendeur est un problème de compromis : augmenter le prix de réserve augmente le prix de vente, mais réduit aussi la probabilité que le ressource soit vendue. Nous avons supposé qu'il y a T types de clients/joueurs possibles, et que le nombre de clients suite un loi connue. En supposant sans perte de généralité que $\theta'_{(1)}(0) \geq \theta'_{(2)}(0) \geq ... \geq \theta'_{(T)}(0)$, où $\theta'_{(i)}$ est la fonction de valuation marginale pour tout joueur de type i, et sous l'hypothèse que les fonctions de demandes résultant des fonction de valuation

sont concaves, l'espérance du revenu est prouvée être concave en p_0 sur chaque segment $[\theta'_{(t+1)}(0), \theta'_{(t)}(0)], 1 \leq t \leq T$ (avec $\theta'_{(T+1)}(0) = 0$). Optimiser le revenu est donc aisé.

Enfin dans [138], nous avons étudié le comportement du PSP dans un cadre aléatoire, avec des joueurs entrant et quittant le jeu. Tout changement de conditions entraîne alors une nouvelle phase de convergence du PSP, qui implique une perte d'efficacité du mécanisme en moyenne. Ceci nous a conduit à imaginer un mécanisme d'enchères multiples aux propriétés équivalentes à PSP, mais plus facilement adaptable aux variations du réseau, et moins lourd en termes de complexité (d'autant qu'il ne nécessite pas de communiquer le profil des enchères à tous les joueurs).

Enchères multiples

La définition du mécanisme d'enchères multiples [140] se base sur celle du PSP, en utilisant les mêmes hypothèses sur les fonctions de valuation des joueurs. Le processus de soumission est le suivant :

– Soit $\mathcal{I}$ l'ensemble des joueurs. Quand un joueur i entre dans le jeu, il soumet un ensemble de M_i enchères à deux dimensions $s_i = \{s_i^1, ..., s_i^{M_i}\}$, où $\forall i\ m, 1 \leq m \leq M_i, s_i^m = (q_i^m, p_i^m)$ est défini comme pour le PSP : q_i^m est la quantité de ressource et p_i^m le prix unitaire pour cette quantité. On suppose les enchères triées afin que $p_i^1 \leq p_i^2 \leq ... \leq p_i^{M_i}$. L'ensemble des enchères multiples possible est défini par

$$S = \bigcup_{M \geq 0} \left(\mathbb{R}^+ \times \mathbb{R}^+\right)^M, \quad \text{avec } \left(\mathbb{R}^+ \times \mathbb{R}^+\right)^0 = \emptyset.$$

Le vendeur (noté 0) définit ici encore un prix de réserve, de sorte qu'il soumet une enchère avec $M_0 = 1$ et $s_0^1 = (Q, p_0)$.

– Le vendeur rassemble alors le profil des enchères $s = (s_i)_{i \in \mathcal{I}}$. À partir de ce profil des enchères, il peut calculer la *fonction de pseudo demande agrégée* en fonction de prix p :

$$\bar{d}(p) = \sum_{i \in \mathcal{I} \cup \{0\}} \bar{d}_i(p) \tag{4.17}$$

avec $\bar{d}_i$ pseudo fonction de demande agrégée pour le joueur i définie par

$$\bar{d}_i(p) = \begin{cases} 0 & \text{si } s_i = \emptyset \text{ ou } p_i^{M_i} < p \\ \max\limits_{1 \leq m \leq M_i} \{q_i^m : p_i^m \geq p\} & \text{sinon.} \end{cases} \tag{4.18}$$

Le pseudo prix du marché (tel que pseudo demande égale capacité) est alors défini par

$$\bar{u} = \sup \left\{p : \bar{d}(p) > Q\right\}. \tag{4.19}$$

Afin de donner la règle d'allocation, notons pour toute fonction réelle f, $f(x^+) = \lim_{z \to x, z > x} f(z)$. Nous suggérons que soit alloué au joueur i

$$a_i(s_i, s_{-i}) = \bar{d}_i(\bar{u}^+) + \frac{\bar{d}_i(\bar{u}) - \bar{d}_i(\bar{u}^+)}{\bar{d}(\bar{u}) - \bar{d}(\bar{u}^+)}(Q - \bar{d}(\bar{u}^+)). \tag{4.20}$$

En d'autres termes, chaque joueur obtient ce qu'il demande au plus petit prix $\bar{u}^+$ qui excède la pseudo demande. Le surplus non alloué $Q - \bar{d}(\bar{u}^+)$ est alors partagé entre le joueurs qui ont soumis un prix $\bar{u}$, proportionnellement aux sauts des fonctions de pseudo demande $\bar{d}_i(\bar{u}) - \bar{d}_i(\bar{u}^+)$ pour conserver un allocation complète. Chaque joueur $i \in \mathcal{I}$ doit payer le prix total $c_i(s)$, avec

$$c_i(s_i, s_{-i}) = \sum_{j \in \mathcal{I} \cup \{0\}, j \neq i} \int_{a_j(s)}^{a_j(s_{-i})} \bar{\theta}'_j(q) dq. \tag{4.21}$$

Le coût total obéit donc au principe du second prix, mais en utilisant les fonctions de pseudo valuation $\bar{\theta}'_i$, définies par

$$\bar{\theta}'_i(q) = \begin{cases} 0 & \text{si } s_i = \emptyset \text{ ou } q_i^1 < q \\ \max\limits_{1 \leq m \leq M_i} \{p_i^m : q_i^m \geq q\} & \text{sinon,} \end{cases}$$

au lieu des fonction réelles θ'_i inconnues du vendeur.

Ce mécanisme possède des propriétés similaires au PSP. Définissons une enchère sincère comme une enchère telle que

$$s_i = \emptyset \text{ ou } \forall m, 1 \leq m \leq M_i,\ p_i^m = \theta'_i(q_i^m).$$

Nous avons pu démontrer dans [140] que, comme pour le PSP, la compatibilité d'incitation (chaque joueur i a intérêt à soumettre une enchère sincère) et l'efficacité sont prouvées, mais à une constante contrôlée près. Il est aussi suggéré que, pour être le plus efficace possible, soumettre un ensemble d'enchères selon une répartition uniforme des quantiles des enchères est préférable de manière à réduire la constante en question : $(q_i^m, p_i^m = \theta'_{t_i}(q_i^m))\ \forall 1 \leq m \leq M$ tel que

$$\int_{d_i(p_i^{m+1})}^{d_i(p_i^m)} (\theta'_{t_i}(q) - p_i^m) dq = C\ \forall m, \text{ où } \begin{cases} p_i^{M+1} = \theta'_{t_i}(0) \\ p_i^0 = p_0, \end{cases} \tag{4.22}$$

d_i étant la fonction de demande du joueur i.

Les enchères multiples présentent les avantages suivants par rapport au PSP :

- comme les enchères sont soumises une unique fois, aucune phase de convergence n'est nécessaire où les joueurs re-soumettent des enchères jusqu'à ce qu'un équilibre soit atteint (situation d'autant plus intéressante dans le cas de joueurs entrant puis quittant le jeu).
- Avec le PSP, chaque joueur doit connaître le profil avant de re-soumettre une enchère, et donc le profil dot être communiqué à tous les joueurs. Ceci n'est pas utile pour les enchères multiples, provoquant un gain important en signalisation.

Dans [141] nous avons illustré le gain obtenu par les enchères multiples par rapport au PSP. À titre d'exemple, la figure 4.3 présente le gain obtenu quand le nombre de joueur est fixé et jusqu'à l'équilibre pour le PSP pour deux types de joueurs, 3 de joueurs de type 1 et 2 de type 2 (avec fonction de valuation indiquée figure 4.3). La Figure 4.4 illustre quant à elle l'évolution quand les joueurs entrent et quittent le jeu (avec $M_i = 3\ \forall i$). Dans les deux cas, le gain sur le revenu et le surplus collectif moyen semble important.

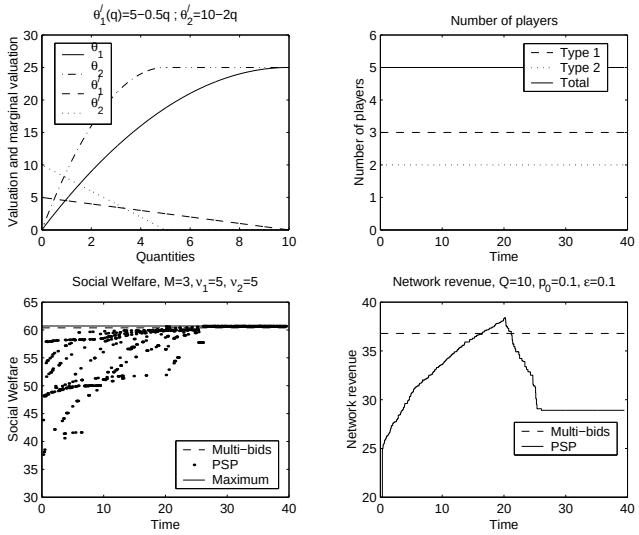


FIG. 4.3 – Comparaison de PSP et des enchères multiples pour un nombre fixé de joueurs, jusqu'à convergence pour le PSP

Nous avons étendu le mécanisme des enchères multiples au cas d'un type de réseau particulier, le réseau en forme d'arbre [143]. En effet, il est raisonnable de penser que le coeur de réseau n'est pas engorgé, et que la congestion n'est obtenue qu'aux réseaux d'accès, où changer de technologie pour surdimensionner les liens est trop coûteux [25]. De même, il est raisonnable de modéliser les réseaux d'accès par un arbre [28] : les clients ne sont pas connectés directement au réseau haut-débit, mais à des réseaux « locaux», eux-même connectés à des réseaux régionaux... L'algorithme 1 décrit l'adaptation du mécanisme à un arbre, et procède des feuilles vers la racine (point d'entrée du coeur de réseau). Il rend un vecteur $a = (a_1, ..., a_{|\mathcal{I}|}) \in \mathbb{R}_+^{\mathcal{I}}$ et se termine dès que l'allocation au lien directement connecté à la racine est calculée. L'objectif de l'algorithme est de judicieusement modifier les enchères multiples des joueurs, dans le but que la règle d'allocation garde les propriétés obtenues sur un lien unique, en s'assurant que la demande (et donc l'allocation) sur les liens en amont n'excède pas l'allocation du lien courant. Ceci est réalisé à l'étape 2b de l'algorithme en éliminant les enchères qui excède l'allocation actuelle. Nous avons prouvé que toutes les propriétés (compatibilité d'incitation, efficacité) restent vérifiées comme dans le cas d'un lien unique.

Enfin, nous avons étendu le mécanisme au cas de fonctions d'utilité ne dépendant pas uniquement de l'allocation instantannée comme dans ce qui précède et la litérature, mais de l'histoire de l'allocation jusqu'à l'instant présent [142]. Ici encore, les propriétés compatibilité d'incitation, efficacité et rationalité individuelle sont prouvées, tout du moins à chaque instant, conditionellement au passé.

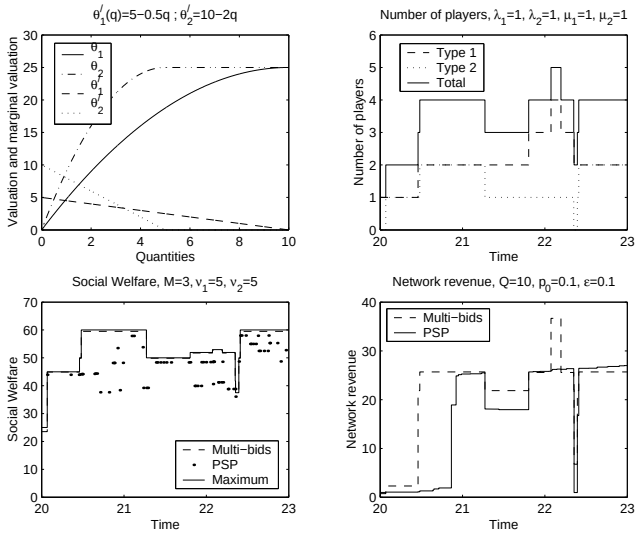


FIG. 4.4 – Comportement de PSP et des enchères multiples avec joueurs entrant et quittant le jeu

4.4 Notes

Nous travaillons (et finalisons) actuellement des extensions des travaux sur la tarification des réseaux à priorités strictes, où nous incluons une composante liée à l'incertitude sur la mesure du trafic. En effet, mesurer la consommation (en nombre de paquets) est un problème difficile ; il faudra avoir recours à des outils de métrologie passive pour mesurer le débit, et active pour mesurer le délai (si le prix par paquet en dépend). L'aversion au risque lié à cette incertitude/erreur sur les mesures est inclue dans les modèles. Les paramètres optimaux peuvent alors être déterminés. Nous avons obtenu des résultats préliminaires en ce sens dans [92].

Une autre contribution, combine les travaux sur la tarification et l'étude quantitative de la sûreté de fonctionnement : dans [230], nous avons étudié une tarification alternative, pour un réseau sur-dimensionné, où la prix dépendra de la disponibilité des connections. Par des algorithmes génétiques, nous avons alors déterminé comment étendre ou fiabiliser de manière optimale le réseau en fonction de la demande.

Enfin, nous finalisons actuellement un travail sur une tarification multiclasse statique dans le cadre des réseaux sans fil CDMA. L'analyse est similaire à celle de la file M/M/1 avec politique DPS de la section 4.2.4, mais utilise les spécificités liées à la technologie CDMA [93].

Alg. 1 Allocation sur un arbre

Entrées :
- L'arbre défini par l'ensemble $\mathcal{L}$ de liens, et les capacités $Q^l, l \in \mathcal{L}$
- l'ensemble $\mathcal{I}$ des joueurs et leurs routes $\{r_i, i \in \mathcal{I}\}$; $\mathcal{I}^l$ est alors l'ensemble des joueurs dont la route passe le lien l.
- le profil des enchères multiples s.

1. Pour tous les joueurs $i \in \mathcal{I}$, définissons l'*enchère multiple révisée* $\underline{s}_i$ comme $\underline{s}_i = s_i$.
2. Choisir un lien feuille de l'arbre $l \in \mathcal{L}$, et posons $\underline{s}_{\mathcal{I}^l} = (\underline{s}_i)_{i \in \mathcal{I}^l}$.
   (a) Calculer $\bar{u}^l = \bar{u}(\underline{s}_{\mathcal{I}^l}, Q^l))$ et $a_i^l = a_i(\underline{s}_{\mathcal{I}^l}, Q^l)$ pour tout $i \in \mathcal{I}^l$, en appliquant les formules des enchères multiples pour un lien aux enchères multiples révisées $\underline{s}_{\mathcal{I}^l}$.
   (b) $\forall i \in \mathcal{I}^l$, modifier l'enchère révisée comme suit :
   - Posons $\underline{s}_i = \underline{s}_i \setminus \{s_i^m : q_i^m > a_i^l\}$.
   - Si $\bar{u}^l > \max\{p_i^m : (q_i^m, p_i^m) \in \underline{s}_i\}$ alors posons $\underline{s}_i = \underline{s}_i \cup \{(a_i^l, \bar{u}^l)\}$.

   (c) Posons $\mathcal{L} = \mathcal{L} \setminus \{l\}$, i.e. enlevons le lien l de l'arbre
3. **si** $\mathcal{L} \neq \emptyset$ **aller à** 2
   sinon retourner $a = \left(a_1^{l^{root}}, a_2^{l^{root}}, ..., a_{|\mathcal{I}|}^{l^{root}}\right)$

Chapitre 5

Logiciels

La plupart de nos travaux ont nécessité des applications numériques pour lesquelles il a fallu coder les méthodes. Il apparaît cependant intéressant de mettre en valeur deux implantations : la première est une bibliothèque C pour la simulation générant des suites quasi-aléatoires, ainsi que leurs versions randomisées créée au cours de ma thèse. La deuxième est une participation active au code du logiciel SPNP distribué par Duke University.

5.1 Librairie Quasi-Monte Carlo

La création d'une bibliothèque C a été importante et permet de ne pas recoder pour chaque application les suites utilisées. Cette bibliothèque contient la génération des suites suivantes :

- suites de Halton classiques [163], ou avec permutations de Braaten et Weller [33], de Faure [70] ou celles que nous avons proposés dans [213] ;
- suite SQRT [163] ;
- suite de Lapeyre-Pagès [117] ;
- suites de Niederrreiter [163] en base p supérieure à la dimension, ou en base 2, en utilisant le code de Gray [73] ;
- Les suites de Faure [69] ;
- les suites de Sobol [16] ;
- Les randomisations des (t, m, s)-nets de Owen [172], Matousek [152] (la randomisation par translation ne nécessite pas d'implémentation particulière car elle utilise directement n'importe quelle suite).

Il faut noter que, depuis, de nombreuses bibliothèques mises à jour régulièrement sont devenues disponibles sur Internet. On peut notamment recenser La bibliothèque Java SSJ de P. L'Ecuyer à Montréal (voir `ttp://www.iro.umontreal.ca/~simardr/ssj/indexf.`tml), la bibliothèque C++ de A. Keller et I. Friedel (`ttp://www.multires.caltec`.edu/software/libseq/index.html), ainsi que la bibliothèque C de C. Lemieux pour les suites QMC randomisées (`ttp://www.mat`.ucalgary.ca/ lemieux/randqmc/).

5.2 SPNP

SPNP (*Stochastic Petri Net Package*) [104] est un outil de modélisation et d'analyse par réseau de Petri stochastique. Les modèles sont implantés dans un langage appelé CSPL (*C-based SPN Language*), une extension du langage de programmation C. Une interface graphique est aussi disponible. Ce logiciel permet de générer automatiquement et de résoudre des modèles de récompense Markoviens à partir d'un réseau de Petri. Nous avons contribué sur l'extension au cas de réseaux de Petri non-Markoviens et fluides, et sur l'implantation des méthodes de simulation. Les modèles pouvant être décrits par le logiciel son ceux présentés Section 2.1. La figure 5.1 montre les méthodes d'analyse actuellement implantées dans SPNP. La notation SOR est pour la méthode *Successive Overrelaxation*,

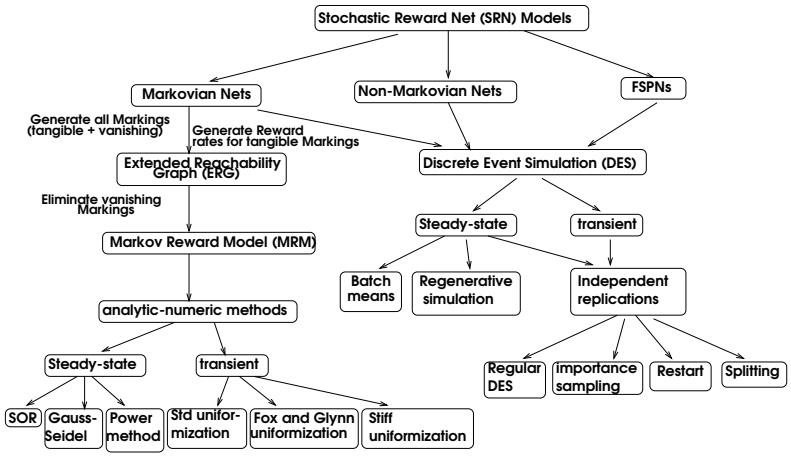


FIG. 5.1 – Méthodes disponibles dans SPNP

en général la plus rapide pour l'étude du stationnaire, mais dont la convergence n'est pas garantie. Dans ce cas, la méthode de Gauss-Seidel peut converger. En dernier ressort, la méthodes des puissances est démontrée converger.

Dans le cas transitoire, les techniques sont basées sur l'uniformisation, une détection de stationnaire pouvant être utilisée pour les modèles raides [144].

Dans le cas où les méthodes décrites ci-dessus ne peuvent être utilisées (en raison d'un espace d'états trop grand où parce qu'il n'y a pas de restrictions assez forte sur le modèle), on utilise la simulation. Les méthodes suivantes de simulation sont implantées :

- simulation utilisant des réplications indépendantes ;
- simulation utilisant des "batches" où une itération est considérée et coupée en blocs pour obtenir un intervalle de confiance ;
- simulation par ramification des trajectoires pour l'évaluation d'événements rares [231] ;
- simulation par échantillonnage préférentiel [232] ;
- simulation régénérative pour des estimations stationnaires non biaisées.

5.3 Notes

Parmi les autres implantations notables, nos travaux sur les files d'attente à forme produit nous ont conduit à créer une version modifiée du logiciel MonteQueue 2.0 de Keith Ross [190]. Nos travaux sur les chaînes de Markov hautement fiables ont été quant à eux implantés dans *BB* (*Balls and Baskets*) [38].

Chapitre 6

Conclusions

6.1 Importance de la thématique

La modélisation et l'évaluation de performances constituent un enjeu majeur pour la conception et le contrôle de systèmes, qui sont de plus en plus complexes et évoluent de plus en plus rapidement. Il y a donc un besoin de plus en plus important d'outils pour comprendre leur comportement, pour minimiser les coûts de déploiement et de gestion et pour optimiser les performances, au cours de la vie de ces systèmes.

Pour répondre à ces questions, l'utilisation d'une démarche intuitive peut s'avérer très dangereuse, et une évaluation expérimentale extrêmement coûteuse. Une méthode intermédiaire entre ces deux extrêmes est la modélisation abstraite et l'évaluation théorique. Le processus de modélisation est basé sur des hypothèses qui doivent être motivées par deux considérations contradictoires : simplicité suffisante du modèle afin de pouvoir résoudre le problème, et adéquation des résultats avec le système réel.

Le choix du modèle est donc intimement lié aux méthodes de résolution existantes. Le développement de méthodes d'analyse est donc un point capital, car il permet d'utiliser un modèle plus proche de la réalité. Contrôler un système, par un bon choix des paramètres ajustables, est de la même manière très important afin de maximiser ses performances.

6.2 Contributions

Nous pouvons décomposer nos contributions en plusieurs sous domaines. La thématique générale est la modélisation, l'analyse et le contrôle des performances et de la sûreté de performance des réseaux de communications, avec une implication majeure dans deux domaines particuliers : les méthodes de simulation et les modèles de tarification.

– Dans le cadre général de la modélisation et de l'analyse de la qualité de service et de la sûreté de fonctionnement, nous avons travaillé sur la modélisation par réseau de Petri. Nous avons notamment participé (et continuons de participer) au développement du logiciel SPNP distribué par Duke University. Nous avons principalement travaillé sur l'implémentation de diverses techniques de simulation. Ce formalisme

nous a aussi été utile pour nos études de files à seuils avec hystérésis dans les cas il n'a pas été possible de déterminer une forme explicite pour les probabilités stationnaires. Un autre cadre de modélisation mathématique important auquel nous nous sommes intéressés (concernant le réseau Internet) est celui des processus AIMD, modélisant bien le protocole TCP. Dans le but de mieux contrôler la qualité de service, nous avons cherché si une politique de perte particulière aux instants de congestion devait être privilégiée. De manière surprenante, nous avons montré que, dans le cas de connections symétriques, le débit moyen est indépendant de cette politique de perte. Cependant, ceci n'est pas vrai pour la variabilité du débit, ainsi que pour le cas non symétrique (en terme d'équité). Dans ces cas, une analyse de quatre politiques naturelles nous a permis de prouver qu'infliger la perte à la connection de plus haut débit était préférable.

– Plus particulièrement, les techniques de simulation Monte Carlo et quasi-Monte Carlo ont été un thème moteur de nos recherches. Ainsi, nous les avons développées dans deux directions. Tout d'abord sur la simulation des systèmes Markoviens hautement fiables qui constitue un domaine important de l'analyse de la sûreté de fonctionnement. Nous avons proposé de nouvelles mesures d'échantillonnage, et surtout défini la notion d'approximation normale bornée qui assure la validité de l'intervalle de confiance quand la fiabilité augmente ; nous avons donné une condition nécessaire et suffisante pour obtenir cette propriété et donné les relations entre approximation normale bornée, erreur relative bornée et bonne approximation asymptotique de la valeur cherchée et la variance de l'estimateur. La seconde contribution majeure porte sur l'analogue déterministe des méthodes de Monte Carlo, les méthodes dite de Quasi-Monte Carlo, convergeant asymptotiquement plus rapidement. Nos activités principales ont porté sur les méthodes dites « randomisées » qui permettent d'obtenir une estimation de l'erreur (ce qui n'était pas le cas en pratique pour les méthodes de quasi-Monte Carlo). Nous avons aussi pu définir une méthode quasi-Monte Carlo (et sa version randomisée) pour la simulation des chaînes de Markov. Nous avons appliqué avec succès ces méthodes à l'évaluation des performances de réseaux de files d'attente à forme produit par exemple. Il est à noter le troisième type de méthode de simulation auquel nous nous sommes intéressés : la technique de réplications des trajectoires.
– Les modèles de tarification des réseaux ont constitué ces dernières années une part importante de nos travaux. Ces modèles doivent permettre de contrôler le niveau de congestion dans les réseaux, et par conséquent la qualité de service. On peut ici encore (non exhautivement) décomposer les contributions en trois sous classes. La première porte sur la tarification multi classes : nous avons déterminé les prix optimisant le revenu du fournisseur ou le surplus social pour différentes politiques d'ordonnancement, et également comparé leur performances respectives. En outre, nous avons travaillé sur les techniques d'enchères pour la bande passante et défini une nouvelle technique basée sur des enchères multiples. Enfin, nous avons travaillé sur une méthode de tarification où la probabilité de perte dans un routeur RED varie selon la volonté de payer des utilisateurs.

6.3 Futures directions de recherche

Citons maintenant, parmi les directions de recherche que nous envisageons, celles qui nous paraissent les plus importantes. Sans être exhaustif, orientons cette section vers les travaux qui peuvent découler de ce qui a été présenté dans le manuscrit.

Tout d'abord, nos travaux sur les files d'attente à seuils et hystérésis peuvent avoir des applications et extensions dans différents domaines. Un premier axe est l'utilisation de ces modèles dans le cadre de la tarification. On peut ainsi imaginer d'établir des seuils sur l'occupation du tampon à partir desquels les prix augmenteraient, et donc la demande diminuerait, de manière à mieux contrôler la congestion. L'hystérésis permettrait d'éviter des oscillations autour des seuils, et donc des oscillations trop fréquentes de prix. Les prix optimaux devraient être alors légèrement différents par rapport aux modèles précédemment étudiés. Les files à seuils (avec ou sans hystérésis) semblent également être un outil privilégié pour modéliser les mécanismes de gestion active de files d'attente dans les résaeux, comme les algorithmes RED (*Random Early Detection*) [72] ou par exemple RIO (*RED In and Out*) [51], une version multi dimensionnelle avec plusieurs classes de trafic. Ceci pourrait permettre de déterminer les seuils à partir desquels les probabilités de perte deviennent non nulles, ainsi que ceux à partir desquels tout paquet d'une classe donnée sera rejeté.

Les techniques de simulation ont elles aussi un potentiel de perfectionnement non négligeable. Un axe très prometteur est celui relativement récent de la *simulation parfaite* [180] où, par simulation inverse dans le temps, il est possible de simuler la loi stationnaire *exacte* d'un processus de Markov. Cette technique, pour être efficace en pratique, nécessite des propriétés telle que la monotonie [180, 239] ou encore des majorations/minorations [114] de manière à ne pas devoir traiter tout l'espace d'états ; ces propriétés demandent à être affaiblies. De plus, l'utilisation de suites à discrépance faible au lieu de nombres pseudo aléatoires devrait permettre une meilleure répartition de l'échantillon sélectionné. Enfin, l'application de cette méthode à des problèmes d'estimation d'événements rares reste un problème ouvert, car il semble difficile de la combiner avec de l'échantillonnage préférentiel par exemple.

Les méthodes de quasi-Monte Carlo randomisées ont connu depuis une dizaine d'année un développement important en raison de leur efficacité et de leur grande simplicité d'utilisation. Nous avons déjà appliqué ces méthodes à la simulation des réseaux de files d'attente, mais souhaitons développer et démocratiser leur application dans le domaine des télécommunications (par le calcul d'espérances mathématiques). Dans le cas des modèles statiques, l'utilisation des méthodes RQMC et/ou mixte si la dimension mathématique du problème est grande, devrait s'avérer efficace. Dans le cas dynamique, on s'intéressera plus particulièrement à la méthode quasi-Monte Carlo et la version randomisée associée que nous avons développé pour les chaînes de Markov. Elle s'avère très efficace si les états peuvent être totalement ordonnés et sont représentés de manière unidimensionnelle. Il existe cependant des applications pour lesquelles ces hypothèses ne sont pas adaptées (comme dans le cas de la simulation de nos systèmes hautement fiables par exemple). Il s'agira donc d'étendre la méthode à ces applications.

Nos travaux autour de la tarification ont été initiés en 2000, et n'ont connu un réel développement qu'à partir de 2002. De nombreux travaux restent encore à faire. Par exemple, la détermination de nouveaux modèles est toujours d'actualité, surtout dans le cas de ré-

seaux de topologie générale, avec des fournisseurs de service en compétition. Nous commençons également à nous intéresser aux problèmes d'implémentation pratique (protocoles...). De même, nous cherchons à développer des méthodes spécifiques pour les technologies sans fils pour lesquels la ressource est et restera probablement limitée. Plus particulièrement, nous souhaitons nous intéresser aux réseaux basés sur la technologie CDMA (voir [4, 133, 149, 194, 202] pour quelques travaux de la littérature sur le sujet) que l'on retrouvera dans les réseaux UMTS, où l'externalité est le niveau d'interférence induite par une communication, les réseaux WiFi (voir [74, 161]) ou les réseaux ad hoc (voir [192, 58, 108, 182]), où l'objectif est d'inciter les utilisateurs à collaborer pour transmettre les paquets des autres utilisateurs, avec les questions de routage associés. La tarification inter-domaines, afin que chaque fournisseur puisse acheminer son trafic jusqu'à destination, est aussi une thématique ayant pris de l'importance ces dernières années, et est le sujet du projet AUCTIONS du réseau d'excellence européen EuroNGI dans lequel nous sommes impliqués.

Tous les modèles de tarification de la littérature utilisent des fonctions d'utilité pour représenter la réaction des utilisateurs face aux prix. Les fonctions utilisées l'ont été pour leurs propriétés mathématiques intéressantes et/ou leur validité a priori pour certaines applications. Cependant, en comparaison avec la littérature sur les méthodes de tarification, peu d'effort ont porté sur la détermination ou la validation de ces fonctions d'utilité. Quelques projets ont cependant été développés à travers le monde afin de mieux comprendre le comportement des utilisateurs face à la tarification : on peut citer le projet INDEX à Berkeley [23, 22, 66, 234], le projet M3I [1]... L'un de nos objectifs est, par le biais de tests et selon les applications, de déterminer la loi de la demande pour des paramètres de qualité de service donnée et un prix donné. Nous pourrons pour cela utiliser les compétences du projet ARMOR sur l'évaluation quantitative de la qualité [158]. Notons que, de manière anecdotique, les méthodes de simulation quasi-Monte Carlo peuvent être utiles aux modèles de tarification des réseaux qui utilisent des modèles de Logit ou Probit (c'est à dire des modèles de choix discrets) [27] qui nécessitent le calcul de coefficients représentés par des intégrales. Ce problème a commencé à être étudié dans [26], mais pas pour notre domaine d'application.

Le cas de la tarification pour les réseaux pair à pair est aussi un thème qui se développe actuellement. Ce sujet, en fait très proche de notre problématique, mérite de l'attention.

La problématique de la tarification ne peut être dissociée de deux autres thèmes auxquels nous nous sommes encore peu intéressés : la métrologie et la facturation. La métrologie doit être utilisée pour comptabiliser (probablement de manière statistique) le trafic envoyé par chaque usager, mais aussi pour mesurer la qualité de service fournie dans le cas d'une tarification dynamique dépendant du délai ou du taux de perte par exemple. La métrologie, utilisant des outils mathématiques puissants, a connu un essor très important ces dernières années et a donné lieu à des implémentations disponibles sur le commerce [34, 64, 109, 204]. De la même manière, établir les prix et mesurer sont inutiles s'il n'y a pas de mode de facturation. La facturation des réseaux pose de nombreux problèmes ouverts auxquels il faut répondre. Ainsi on peut envisager qu'un usager changera de domaine régulièrement (fixe, UMTS...) de sorte que la facturation globale devra être vue comme une agrégation de factures. Il faudra donc chercher à définir une architecture de facturation, distribuée, utilisant éventuellement des micro paiements. Pour quelques références

sur la problématique de facturation, voir [17, 20, 37, 49].

Un dernier problème pour un opérateur, très lié à la tarification, est celui de la planification des capacités (dans un contexte concurrentiel). Ainsi un opérateur a t'il intérêt à surdimensionner son réseau (de sorte que les clients ne paieraient plus d'externalité) ? De même, le coût d'une installation diminuant avec le temps, à quel moment faut-il la changer ? Ces questions commencent à être abordées dans les télécommunications dans un cadre général [62, 156], mais doivent être abordées plus précisément dans le cadre des modes de tarification que nous avons étudiés.

Enfin, un de nos objectifs est de rendre disponibles sur le web les logiciels et solutions développés ou en cours de développement, et d'améliorer leur convivialité.

Bibliographie

[1] M3I : Market Managed Multiservice Internet. `http://www.m3i.org/`.

[2] O. Ait-Hellal, E. Altman, D. Elouadghiri, M. Erramdani, and N. Mikou. Performance of tcp/ip : the case of two controlled sources. In *Proceedins of the International Conference for Computer Communications - ICCC'97*, pages 469–477, Cannes, France, November, 19-21 1997.

[3] M. Ajmone Marsan, G. Balbo, G. Conte, S. Donatelli, and G. Franceschinis. *Modelling with Generalized Stochastic Petri Nets*. John Wiley & Sons, 1995.

[4] T. Alpcan, T. Başar, R. Srikant, and E. Altman. CDMA uplink power control as a noncooperative game. *Wireless Networks*, 2002.

[5] E. Altman, C. Barakat, E. Laborde, P. Brown, and D. Collange. Fairness analysis of TCP/IP. In *Proceedings of IEEE Conference on Decision and Control (CDC'00)*, Sydney, Australia, December 2000.

[6] E. Altman, D. Barman, R. El Azouzi, D. Ros, and B. Tuffin. Pricing Differentiated Services : A Game-Theoretic Approach. In Nikolas Mitrou, Kimon P. Kontovasilis, George N. Rouskas, Ilias Iliadis, and Lazaros F. Merakos, editors, *Networking 2004*, volume 3042 of *Lecture Notes in Computer Science*, Athens, Greece, may 2004. Springer.

[7] E. Altman, D. Barman, R. El Azouzi, D. Ros, and B. Tuffin. Pricing Differentiated Services : A Game-Theoretic Approach. *Computer Networks*, 5 :982–1002, 2006.

[8] E. Altman, D. Barman, B. Tuffin, and M. Vojnović. Parallel TCP Sockets : Simple Model, Throughput and Validation. In *IEEE INFOCOM 2006*, Barcelona, Spain, April 2006.

[9] E. Altman, R. El Azouzi, D. Ros, and B. Tuffin. Loss Strategies for Competing TCP/IP Connections. In Nikolas Mitrou, Kimon P. Kontovasilis, George N. Rouskas, Ilias Iliadis, and Lazaros F. Merakos, editors, *Networking 2004*, volume 3042 of *Lecture Notes in Computer Science*, Athens, Greece, may 2004. Springer.

[10] E. Altman, R. El Azouzi, D. Ros, and B. Tuffin. Loss strategies for competing TCP/IP connections. *Computer Networks*, 2005 (to appear).

[11] E. Altman, T. Jiménez, and D. Kofman. DPS Queues with Stationary Ergodic Service Times and the Performance of TCP in Overload. In *proceedings of IEEE INFOCOM*, 2004.

[12] R. Alur, C. Courcoubetis, T.A. Henzinger, and P.H. Ho. Hybrid Automata : An Algorithmic Approach to the Specification and Verification of Hybrid Systems. In Anders P. Ravn Robert L. Grossman, Anil Nerode and Hans Rischel, editors, *Hybrid Systems*, volume 736 of *Lecture Notes in Computer Science*, pages 209–229. Springer-Verlag, 1991.

[13] R. Alur and D.L. Dill. A theory of timed automata. *Theoretical Computer Science*, 126 :183–235, 1994.

[14] L. Anania and R.J. Solomon. Flat- The Minimalist Price. In Lee W. McKnight and Joseph P. Bailey, editors, *Internet Economics*, pages 91–118. MIT Press, 1997.

[15] K.D. Ansell, P.S. Glazebrook and I. Mitrani. Threshold policies for a single-server queuing network. *Probability in the Engineering and Informational Sciences*, 15 :15–33, 2001.

[16] I.A.. Antonov and Saleev V.M. An economic method of computing lp_τ-sequences. *USSR Computational Math. and Math. Phys.*, 19(1) :252–256, 1979.

[17] N. Asokan, P. Janson, M. Steiner, and M. Waidner. The state of the art in electronic payment systems. *IEEE Computer*, 30 :28–35, 1997.

[18] S. Athuraliya and S.H. Low. Optimization Flow Control, II : Implementation. Technical report, 2000.

[19] S. Athuraliya and S.H. Low. Optimization Flow Control with Newton-Like Algorithm. *Telecommunication Systems*, 13, 2000.

[20] Y.A. Au and R.J. Kauffman. Should we wait ? network externalities, compatibility, and electronic billing adoption. *BIJournal of Management Information Systems*, 18(2) :47–64, 2001.

[21] F. Baccelli and D. Hong. AIMD, Fairness and Fractal Scaling of TCP Traffic. In *Proceedings of IEEE INFOCOM 02*, April 2002.

[22] W. Beckert. Estimation of stochastic preferences : An empirical analysis of demand for internet services. Technical report, Department of Economics, University of Berkeley, 2000.

[23] W. Beckert. *Stochastic Demand Analysis*. PhD thesis, UC Berkeley, 2000.

[24] V. Bentkus and F. Götze. The Berry-Esseen bound for Student's statistic. *The Annals of Probability*, 24(1) :491–503, 1996.

[25] L. Bernstein. Managing the last mile. *IEEE Communications Magazine*, 35(10) :72–76, Oct 1997.

[26] C.R. Bhat. Simulation estimation of mixed discrete choice models using randomized and scrambled halton sequences. *Transportation Research Part B*, 37 :837–855, 2003.

[27] M. Bierlaire. Discrete choice models. In K.Tanczos M. Labb, G. Laporte and Ph. Toint, editors, *Operations Research in Traffic and Transportation Management, Vol. 166 of NATO ASI Series, Series F : Computer and Systems Sciences*, pages 203–227. Springer Verlag, 1998.

[28] C. Bisdikian, K. Maruyama, D. I. Seidman, and D. N. Serpanos. Cable access beyond the hype : On residential broadband data services over HFC networks. *IEEE Communications Magazine*, 34(11) :128–135, Nov 1996.

[29] R. Bohn, H.W. Braun, K.C. Claffy, and S. Wolff. Mitigating the Coming Internet Crunch : Multiple service levels via Precedence. Technical report, University of California - San Diego, 1993.

[30] T. Bonald and L. Massoulié. Impact of Fairness on Internet Performance. In *Proceedings of ACM Sigmetrics 2001*, 2001.

[31] J.P. Borel, G. Pagès, and Y. Xiao. *Probabilités numériques*, chapter suites à discrépance faible et intégration numérique. INRIA, 1991. collection didactique.

[32] N. Bouleau and D. Lépingle. *Numerical methods for stochastic processes*. John Wiley & Sons, 1994.

[33] E. Braaten and G. Weller. An Improved Low-Discrepancy Sequence for Multidimensional Quasi-Monte Carlo Integration. *J. Comput. Phys.*, 33 :249–258, 1979.

[34] ENST Bretagne. *Page Internet du projet SATURNE*. http ://saturne.ipv6.rennes.enst-bretagne.fr/.

[35] P. Brown. Resource sharing of TCP connections with different round trip times. In *IEEE Infocom*, Mar 2000.

[36] T. Bu and D. Towsley. Fixed point approximations for tcp behavior in an aqm network. In *Proceedings of the 2001 ACM SIGMETRICS international conference on Measurement and modeling of computer systems*, pages 216–225. ACM Press, 2001.

[37] L. Buttyan and N. Ben Salem. A payment scheme for broadcast multimedia streams. In *Proceedings of the Sixth IEEE Symposium on Computers and Communications*, 2001.

[38] H. Cancela. *Évaluation de la sûreté de fonctionnement : Modèles combinatoires et Markoviens*. PhD thesis, Université de Rennes 1, December 1996.

[39] H. Cancela, G. Rubino, and B. Tuffin. Fast Monte Carlo Methods for evaluating Highly Dependable Markovian Systems. In *Second Internatinal Conference on Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing*, Salzburg, July 1996.

[40] H. Cancela, G. Rubino, and B. Tuffin. MTTF estimation using importance sampling on Markov models. In *Proceedings of ICIL2001*, 2001.

[41] H. Cancela, G. Rubino, and B. Tuffin. MTTF estimation by Monte Carlo methods using Markov models. *Monte Carlo Methods and Applications*, 8(4) :312–341, 2002.

[42] H. Cancela, G. Rubino, and B. Tuffin. Bounded relative efficiency in rare event simulations. In *Proceedings of SAINT'05 workshops*, Trento, Italy, January 2005.

[43] H. Cancela, G. Rubino, and B. Tuffin. New measures of robustness in rare event simulation. In F.B. Armstrong M.E. Kuhl, N.M. Steiger and J.A. Joines, editors, *Proceedings of the 2005 Winter Simulation Conference*, pages 519–527, 2005.

[44] J. A. Carrasco. Failure distance based on simulation of repairable fault tolerant systems. *Proceedings of the 5th International Conference on Modeling Techniques and Tools for Computer Performance Evaluation*, pages 351–365, 1992.

[45] C.G. Cassandras. *Discrete Event Systems : Modeling and Performance Analysis*. Aksen Associates Incorporated Publishers and Irwin, 1993.

[46] B.D. Choi, S.H. Choi, B. Kim, and D.K. Sung. Analysis of priority queueing system based on thresholds and its application to signaling system no. 7 with congestion control. *Computer Networks*, 32 :149–170, 2000.

[47] H. Choi, V.G. Kulkarni, and K.S. Trivedi. Markov Regenerative Stochastic Petri Nets. *Performance Evaluation*, 20(1-3) :337–357, 1993.

[48] H. Choi, V.G. Kulkarni, and K.S. Trivedi. Transient analysis of deterministic and stochastic petri nets. In M. Ajmone Marsan, editor, *Application and Theory of Petri Nets 1993*, volume 691 of *Lecture Notes in Computer Science*, pages 166–185. Springer Verlag, 1993.

[49] J. Chomicki, S. Naqvi, and M.F. Pucci. Decentralized micropayment consolidation. In *Proceedings of the 18th International Conference on Distributed Computing Systems*, 1998.

[50] G. Ciardo, D.M. Nicol, and K.S. Trivedi. Discrete-Event Simulation of Fluid Stochastic Petri-Nets. *IEEE Transactions on Software Engineering*, 25(2) :207–217, 1999.

[51] D. Clark and W. Fang. Explicit Allocation of Best-Effort Packet Delivery Service. *IEEE/ACM Transactions on Networking*, 6 :362–373, 1998.

[52] R. Cocchi, D. Estrin, S. Shenker, and L. Zhang. A Study of Priority Pricing in Multiple Service Class Networks. In *Proceedings of SIGCOMM'91*, pages 123–130, 1991.

[53] R. Cocchi, D. Estrin, S. Shenker, and L. Zhang. Pricing in Computer Networks : Motivation, Formulation and Example. *IEEE/ACM Transactions on Networking*, 1(6) :614–627, 1993.

[54] S. Collas and B. Tuffin. Creation of a dynamic model language and application of Monte Carlo methods for the reliability analysis of industrial complex systems. In *Proceedings of Lambda-Mu 13*, Lyon, March 2002.

[55] C. Courcoubetis, M.P. Dramitinos, and G.D. Stamoulis. An auction mechanism for bandwidth allocation over paths. Technical report, Athens University of Economics and Business, 2001.

[56] C. Courcoubetis and R. Weber. *Pricing Communication Networks—Economics, Technology and Modelling*. Wiley, 2003.

[57] R. Cranley and T.N.L. Patterson. Randomization of number theoretic methods for multiple integration. *SIAM J. Numer. Anal.*, 13(6) :904–914, December 1976.

[58] J. Crowcroft, R. Gibbens, F. Kelly, and S. Östring. Modelling incentives for collaboration in mobile ad hoc networks. In *Proceedings of WoOpt'03, Sophia Antipolis, France*, 2003.

[59] L.A. DaSilva. Pricing of QoS-Enabled Networks : A Survey. *IEEE Communications Surveys & Tutorials*, 3(2), 2000.

[60] A. Delenda, P. Maillé, and B. Tuffin. Reserve price in progressive second price auctions. In *Proceedings of the 9th IEEE Symposium on Computers and Communications*, June 2004.

[61] V. Demers, P. L'Ecuyer, and B. Tuffin. A combination of randomized quasi-monte carlo with splitting for rare-event simulation. In *Proceedings of the 2005 European Simulation and Modelling Conference (ESM'2005)*, Porto, Portugal, October 2005.

[62] Y. d'Halluin, Peter A. Forsyth, and Kenneth R. Vetzal. Managing capacity for telecommunications networks under uncertainty. *IEEE/ACM Transactions on Networking (TON)*, 10(4) :579–587, 2002.

[63] P. Dolan. Internet Pricing. is the end of the World Wide Wait in view ? *Communications & Strategies*, 37 :15–46, 2000.

[64] A.B. Downey. Using pathchchar to estimate internet link characteristics. *ACM SIGCOMM'99*, 2000.

[65] M. Drmota and R.F. Tichy. *Sequences, Discrepancies and Applications*, volume 1651 of *Lecture Notes in Mathematics*. Springer Verlag, Heidelberg, 1997.

[66] R. Edell and P. Varaiya. Providing internet access : What we learn from index. *IEEE Network*, 13(5), 1999.

[67] M. Falkner, M. Devetsikiotis, and I. Lambadaris. An Overview of Pricing Concepts for Broadband IP Networks. *IEEE Communications Surveys & Tutorials*, 3(2), 2000.

[68] Z. Fan. Pricing and provisioning for guaranteed internet services. In P. Lorenz, editor, *ICN 2001*, volume 2093 of *Lecture Notes in Computer Science*, pages 55–64. Springer-Verlag, 2001.

[69] H. Faure. Discrépances de suites associées à un système de numération (en dimension s). *Acta Arith.*, 41 :337–351, 1982.

[70] H. Faure. Good Permutations for Extreme Discrepancy. *Journal of Number Theory*, 42 :47–56, 1992.

[71] G.S. Fishman. *Monte Carlo : Concepts, Algorithms and Applications*. Springer-Verlag, 1996.

[72] S. Floyd and V. Jacobson. Random Early Dectection Gateways for Congestion Avoidance. *IEEE/ACM Transactions on Networking*, 1(4) :397–413, 1993.

[73] B. L. Fox, P. Bratley, and H. Niederreiter. Implementation and tests of low discrepancy sequences. *ACM Trans. Model. Comput. Simul.*, 2(3) :195–213, July 1992.

[74] E. Friedman and D. Parkes. Pricing WiFi at Starbucks – Issues in online mechanism design. In *Proc. of 4th ACM Conf. on Electronic Commerce (EC'03)*, 2003. Extended version at http ://www.eecs.harvard.edu/econs/pubs/online.pdf.

[75] M. J. J. Garvels. *The Splitting Method in Rare Event Simulation*. PhD thesis, Faculty of mathematical Science, University of Twente, The Netherlands, 2000.

[76] R. Gibbens, R. Mason, and R. Steinberg. Internet service classes under competition. *IEEE Journal on Selected Areas in Communications*, 18(12) :2490–2498, 2000.

[77] R.J. Gibbens and F.P. Kelly. Measurement-based connection admission control. In *Proceedings of the 15th International Teletraffic Congress*, 1997.

[78] R.J. Gibbens and F.P. Kelly. Distributed connection acceptance control for a connectionless network. In *Proceedings of the 16th International Teletraffic Congress*, 1999.

[79] R.J. Gibbens, S.K. Sargood, F.P. Kelly, H. Azmoodeh, R. Macfadyen, and N. Macfadyen. An Approach to Service Level Agreements for IP networks with Differential Services. Technical report, Statistical laboratory, University of Cambridge, January 2000.

[80] P. Glasserman, P. Heidelberger, P. Shahabuddin, and T. Zajic. Splitting for rare event simulation : analysis of simple cases. In D.T. Brunner J.M. Charnes, D.J. Morrice and J.J. Swain, editors, *Proceedings of the 1996 Winter Simulation Conference*, pages 302–308, 1996.

[81] P. Glasserman, P. Heidelberger, P. Shahabuddin, and T. Zajic. A look at multilevel splitting. In G. Larcher H. Niederreiter, P. Hellekalek and P. Zinterhof, editors, *Second Internatinal Conference on Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing*, volume 127 of *Lecture Notes in Statistics*, pages 98–108. Springer Verlag, 1997.

[82] P. Glasserman, P. Heidelberger, P. Shahabuddin, and T. Zajic. A large deviations perspective on the efficiency of multilevel splitting. *IEEE Transactions on Automatic Control*, 43(12) :1666–1679, 1998.

[83] P. Glasserman, P. Heidelberger, P. Shahabuddin, and T. Zajic. Multilevel splitting for estimating rare event probabilities. *Operations Research*, 1999. To appear.

[84] A. Goyal, L. Lavenberg, and K. Trivedi. Probabilistic Modeling of Computer System Availability. *Annals of Operations Research*, 8 :285–306, 1987.

[85] A. Goyal, P. Shahabuddin, P. Heidelberger, V.F. Nicola, and P.W. Glynn. A Unified Framework for Simulating Markovian Models of Highly Dependable Systems. *IEEE Transactions on Computers*, 41(1) :36–51, January 1992.

[86] A. Gupta, D.O. Stahl, and A.B. Whinston. Priority Pricing of Integrated Services Networks. In Lee W. McKnight and Joseph P. Bailey, editors, *Internet Economics*, pages 323–352. MIT Press, 1997.

[87] A.Y. Ha. Optimal pricing that coordinates queues with customer-chosen service requirements. *Management Science*, 47(7) :915–930, 2001.

[88] J. M. Hammersley and D. C. Handscomb. *Monte Carlo Methods*. Methuen, London, 1964.

[89] Z. Haraszti and J.K. Townsend. The theory of direct probability redistribution and its application to rare event simulation. *ACM Transactions on Modeling and Computer Simulation*, 9(2) :105–140, 2000.

[90] R. Hassin and M. Haviv. *To Queue or Not To Queue*. Kluwer's INTERNATIONAL SERIES, 2003.

[91] Y. Hayel, P. Maillé, and B. Tuffin. Modelling and analysis of Internet pricing : introduction and challenges. In *Proceedings of the International Symposium on Applied Stochastic Models and Data Analysis (ASMDA 2005)*, May 2005.

[92] Y. Hayel, M. Ouarraou, and B. Tuffin. Optimal Measurement-Based Pricing for an M/M/1 Queue. *Networks and Spatial Economics*, 2006. To appear.

[93] Y. Hayel, V. Ramos, and B. Tuffin. Optimal Static Pricing of Reverse-link DS-CDMA Multiclass Traffic. Technical Report 1731, IRISA, July 2005.

[94] Y. Hayel, D. Ros, and B. Tuffin. Less-than-Best-Effort Services : Pricing and Scheduling. In *IEEE INFOCOM 2004*, Hong-Kong, China, March 2004.

[95] Y. Hayel and B. Tuffin. Optimisation d'un modèle de tarification de l'internet. In *5ème congrès de la Société Française de Recherche Opérationnelle et d'Aide à la Décision (ROADEF'2003)*, Avignon, France, Février 2003.

[96] Y. Hayel and B. Tuffin. A Mathematical Analysis of the Cumulus Pricing Scheme. *Computer Networks*, 47 :907–921, 2005.

[97] Y. Hayel and B. Tuffin. Pricing for heterogeneous services at a discriminatory processor sharing queue. In *4th IFIP-TC6 Networking Conference*, Waterloo, Canada, June 2005.

[98] Y. Hayel and B. Tuffin. An Optimal Congestion and Cost-Sharing Pricing Scheme for Multiclass Services. *Mathematical Methods of Operations Research (To appear)*, 2006.

[99] P.E. Heegaard. *Efficient simulation of network performance by Importance Sampling*. PhD thesis, Norwegian University of Science and Technology, May 1998.

[100] P. Heidelberger. Fast Simulation of Rare Events in Queueing and Reliability Models. *ACM Transactions on Modeling and Computer Simulation*, 5(1) :43–85, January 1995.

[101] T. Henderson, J. Crowcroft, and S. Bhatti. Congestion Pricing. Paying Your Way in Communication Networks. *IEEE Internet Computing*, September/October :85–89, 2001.

[102] T. Henzinger, P. Kopke, A. Puri, and P. Varaiya. What's decidable about hybrid automata ? In *Proc. 27th Symposium on the Theory of Computing*, pages 373–382, 1995.

[103] F. J. Hickernell. Obtaining $o(n^{-2+\epsilon})$ convergence for lattice quadrature rules. In K.-T. Fang, F. J. Hickernell, and H. Niederreiter, editors, *Monte Carlo and Quasi-Monte Carlo Methods 2000*, pages 274–289, Berlin, 2002. Springer-Verlag.

[104] C. Hirel, B. Tuffin, and K.S. Trivedi. SPNP Version 6.0. In B.R. Haverkort, H.C. Bohnenkamp, and C.U. Smith, editors, *Computer performance evaluation : Modelling tools and techniques ; 11th International Conference ; TOOLS 2000, Schaumburg, Il., USA*, volume 1786 of *Lecture Notes in Computer Science*, pages 354–357. Springer Verlag, 2000.

[105] M.L. Honig and K. Steiglitz. Usage-based pricing of packet data generated by a heterogeneous user population. In *Proceedings of IEEE INFOCOM 95*, pages 867–874, 1995.

[106] O.C. Ibe, H. Choi, and K.S. Trivedi. Performance Evaluation of Client-Server Systems. *IEEE Transactions on Parallel and Distributed Systems*, 4(11) :1217–1229, 1993.

[107] O.C. Ibe and J. Keilson. Multi-server threshold queues with hysteresis. *Performance Evaluation*, 21 :185–213, 1995.

[108] O. Ileri, S.C. Mau, and N.B. Mandayam. Pricing for enabling forwarding in self-configuring ad hoc networks. *Submitted to IEEE Journal on Selected Areas in Communications (JSAC)*, 2004.

[109] M. Jain and C. Dovrolis. Pathload : A measurement tool for ent-to-end available bandwidth. In *PAM'02*, pages 14–25, Fort Collins, CO, March 2002.

[110] C. Kelling. A Framework for Rare Event Simulation of Stochastic Petri Nets Using "RESTART". In D.T. Brunner J.M. Charnes, D.J. Morrice and J.J. Swain, editors, *Proceedings of the 1996 Winter Simulation Conference*, pages 317–324, 1996.

[111] F.P. Kelly. Mathematical modelling of the Internet. In *Proceedings of the Fourth International Congress on Industrial and Applied Mathematics*, 2000.

[112] F.P. Kelly, P.B. Key, and S. Zachary. Distributed Acceptance Control. *IEEE Journal on Selected Areas in Communications*, 18, 2000.

[113] F.P. Kelly, A.K. Mauloo, and D.K.H. Tan. Rate control in communication networks : shadow prices, proportional fairness and stability. *Journal of the Operational Research Society*, 49 :237–252, 1998.

[114] J. Kendall, W.S. an Moller. Perfect simulation using dominating processes on ordered state spaces, with application to locally stable point processes. *Advances in Applied Probability*, 32(3) :844–865, 2000.

[115] T.V. Lakshman and U. Madhow. he performance of TCP/IP for networks with high bandwidth-delay products and random loss. *IEEE/ACM Transactions on Networking*, Jun 1007.

[116] J.P. Lambert. Quasi-Monte Carlo, low discrepancy sequences, and ergodic transformations. *Journal of Computational and Applied Mathemathics*, 12&13 :419–423, 1985.

[117] B. Lapeyre and G. Pagès. Familles de suites à discrépance faible obtenues par itération de transformations de [0,1]. *C. R. Acad. Sci. Paris*, 308(I) :507–509, 1989.

[118] A.A. Lazar and N. Semret. Design and Analysis of the Progressive Second Price Auction for Network Bandwidth Sharing. *Telecommunication Systems - Special Issue on Network Economics*, 13, 1999.

[119] L.M. Le Ny and B. Tuffin. Modeling and analysis of multi-class threshold-based queues with hysteresis using Stochastic Petri Nets. In *Petri Nets 2002*, Lecture Notes in Computer Science. Springer Verlag, 2002.

[120] L.M. Le Ny and B. Tuffin. A simple analysis of heterogeneous multi-server threshold queues with hysteresis. In *Proceedings of Applied Telecommunication Symposium (ATS)*, San Diego, USA, April 2002.

[121] C. Lécot and B. Tuffin. Simulating markov chains with quasi-monte carlo methods. In *Vth IMACS Seminar on Monte Carlo Methods MCM-2003*, Berlin, September 2003.

[122] C. Lécot and B. Tuffin. Quasi-Monte Carlo Methods for Estimating Transient Measures of Discrete Time Markov Chains. In H. Niederreiter, editor, *Monte Carlo and Quasi-Monte Carlo Methods 2002*, pages 329–344. Springer, Berlin, 2004.

[123] P. L'Ecuyer. Random number generation. In J.E. Gentle, Haerdle W., and Y. Mori, editors, *Handbook of Computational Statistics*. Springer-Verlag, 2004.

[124] P. L'Ecuyer, P. Lécot, and B. Tuffin. Quasi-monte carlo simulation of markov chains with randomized copies of a two-dimensional highly-uniform point set. In *Fifth International Conference on Monte Carlo and quasi-Monte Carlo Methods*, June 2004.

[125] P. L'Ecuyer, P. Lécot, and B. Tuffin. A new randomized quasi-monte carlo approach for markov chains. In *INFORMS Applied Probability Conference*, July 2005.

[126] P. L'Ecuyer, P. Lécot, and B. Tuffin. A Randomized Quasi-Monte Carlo Simulation Method for Markov Chains. Technical Report 1709, IRISA, 2005.

[127] P. L'Ecuyer and C. Lemieux. Variance reduction via lattice rules. *Management Science*, 49(6) :1214–1235, 1998.

[128] J.Y. Lee and Y.H. Kim. Performance analysis of a hybrid priority control scheme for input and output queueing ATM switches. In *Proceedings of IEEE INFOCOM 98*, pages 1470–1477, March 1998.

[129] C. Lemieux and P. L'Ecuyer. Efficiency improvement by lattice rules for pricing asian options. In *Proceedings of the 1998 Winter Simulation Conference*, pages 579–585, 1998.

[130] C. Lemieux and P. L'Ecuyer. A comparison of Monte Carlo, lattice rules and other low-discrepancy point sets. In H. Niederreiter and J. Spanier, editors, *Monte Carlo and Quasi-Monte Carlo Methods 1998*, pages 326–340, Berlin, 2000. Springer-Verlag.

[131] E. E. Lewis and F. Böhm. Monte Carlo Simulation of Markov Unreliability Models. *Nuclear Engineering and Design*, 77 :49–62, 1984.

[132] X. Lin and N.B. Shroff. Pricing-based control of large networks. In S. Palazzo, editor, *IWDC 2001*, volume 2170 of *Lecture Notes in Computer Science*, pages 212–231. Springer-Verlag, 2001.

[133] P. Liu, M.L. Honig, and S. Jordan. Forward-link CDMA resource allocation based on pricing. In *Proceedings of the 2000 IEEE Wireless Communications and Networking Conference*, pages 1410–1414, 2000.

[134] S.H. Low. Optimization Flow Control with On-line Measurement or Multiple Paths. In *Proceedings of the 16th International Teletraffic Congress*, 1999.

[135] S.H. Low and D.E. Lapsley. Optimization Flow Control, I : Basic Algorithm and Convergence. *IEEE/ACM Transactions on Networking*, 7(6), 1999.

[136] S.H. Low, F. Paganini, and J.C. Doyle. Internet Congestion Control. *IEEE Control Systems Magazine*, 2002.

[137] P. Maillé and B. Tuffin. Mécanisme d'enchères au second prix pour l'allocation de bande passante dans les réseaux. In *5ème congrès de la Société Française de Recherche Opérationnelle et d'Aide à la Décision (ROADEF'2003)*, Avignon, France, Février 2003.

[138] P. Maillé and B. Tuffin. The progressive second price mechanism in a stochastic environment. *Netnomics*, 5(2) :119–147, 2003.

[139] P. Maillé and B. Tuffin. A progressive second price mechanism with a sanction bid from excluded players. In *Third International Workshop on Internet Charging and QoS Technology, ICQT'03*, Lecture Notes in Computer Science. Springer Verlag, 2003.

[140] P. Maillé and B. Tuffin. Multi-bid auctions for bandwidth allocation in communication networks. In *IEEE INFOCOM 2004*, Hong-Kong, China, March 2004.

[141] P. Maillé and B. Tuffin. Multi-bid versus Progressive Second Price Auctions in a Stochastic Environment. In *Fourth International Workshop on Internet Charging and QoS Technology, ICQT'04*, Lecture Notes in Computer Science. Springer Verlag, 2004.

[142] P. Maillé and B. Tuffin. An auction-based pricing scheme for bandwidth sharing with history-dependent utility functions. In *Proceedings of the First International Workshop on Incentive Based Computing (IBC'05)*. IEEE CS Press, september 2005.

[143] P. Maillé and B. Tuffin. Pricing the Internet with multi-bid auctions. *IEEE/ACM Transactions on Networking*, 2006. (To appear).

[144] M. Malhotra, J.K. Muppala, and K.S. Trivedi. Stiffness-tolerant methods for transient analysis of stiff Markov chains. *Microelectronics Reliability*, 34(11) :1825–1841, 1994.

[145] M. Mandjes. Pricing Strategies under Heterogeneous Service Requirements. In *IEEE INFOCOM*, 2003.

[146] P. Marbach. Differentiated Services Networks : Pricing and Software Agents. Technical Report CSRG-422, Department of Computer Science, University of Toronto, 2001.

[147] P. Marbach. A price-based resource allocation mechanism for priority services. Technical Report CSRG-421, Department of Computer Science, University of Toronto, 2001.

[148] P. Marbach. Pricing Differentiated Services Networks : Bursty Traffic. In *Proceedings of IEEE INFOCOM 2001*, 2001.

[149] P. Marbach and R. Berry. Downlink resource allocation and pricing for wireless networks. In *Proc. IEEE INFOCOM*, 2002.

[150] L. Massoulié and J. Roberts. Arguments in favour of admission control for TCP flows. In *Proceedings of the 16th International Teletraffic Congress*, 1999.

[151] L. Massoulié and J. Roberts. Bandwidth sharing : objectives and algorithms. In *Proceedings of IEEE INFOCOM 99*, 1999.

[152] J. Matoušek. On the L_2-discrepancy for anchored boxes. *Journal of Complexity*, 14 :527–556, 1998.

[153] J.K. McKie-Mason and H.R. Varian. Some Economics of the Internet. Technical report, University of Michigan, November 1993. `http://wueconb.wustl.edu:8089/eps/comp/papers:9401/9401001.pdf`.

[154] J.K. McKie-Mason and H.R. Varian. Pricing Congestible Network Resources. *IEEE Journal on Selected Areas in Communications*, 13 :1141–1149, 1995.

[155] H. Mendelson and S. Whang. Optimal incentive-compatible priority pricing for the M/M/1 queue. *Operations Research*, 38(5) :870–883, 1990.

[156] D. Mitra and Q. Wang. Stochastic engineering with applications to network revenue management. In *Proceedings of IEEE INFOCOM 03*, April 2003.

[157] J. Mo and J. Walrand. Fair end-to-end window-based congestion control. In *Proceedings of SPIE'98*, October 1998.

[158] S. Mohamed, G. Rubino, and M. Varela. Performance Evaluation of Real-time Speech Through a Packet Network : a Random Neural Networks Based Approach. *Performance Evaluation (to appear)*, 2003.

[159] W. J. Morokoff and R. E. Caflisch. Quasi-Random Sequences and their Discrepancies. *SIAM Journal on Scientific Computing*, pages 1571–1599, December 1994.

[160] R.R. Muntz, E. de Souza e Silva, and A. Goyal. Bounding Availability of Repairable Computer Systems. *IEEE Transactions on Computers*, 38(12) :1714–1723, 1989.

[161] J. Musacchio and J. Walrand. Game theoretic modeling of WiFi pricing. In *Allerton 2003*, Oct 2003.

[162] M. K. Nakayama. General Conditions for Bounded Relative Error in Simulations of Highly Reliable Markovian Systems. *Advances in Applied Probability*, 28 :687–727, 1996.

[163] H. Niederreiter. *Random Number Generation and Quasi-Monte Carlo Methods*. CBMS-NSF, SIAM, Philadelphia, 1992.

[164] A. Odlyzko. Paris Metro Pricing for the Internet. In *ACM Conference on Electronic Commerce (EC'99)*, pages 140–147, 1999.

[165] A.M. Odlyzko. The current state and likely evolution of the Internet. In *Proceedings of Globecom'99*, pages 1869–1875, 1999.

[166] G. Ökten. A probabilistic result on the discrepancy of a hybrid-Monte Carlo sequence and applications. *Monte Carlo Methods and Applications*, 2(4) :255–270, 1996.

[167] G. Ökten and B. Tuffin. A central limit theorem for a hybrid-Monte Carlo method. In *Fifth International Conference on Monte Carlo and quasi-Monte Carlo Methods*, June 2004.

[168] G. Ökten, B. Tuffin, and V. Burago. A central limit theorem and improved error bounds for a hybrid-Monte Carlo sequence with applications in computational finance, 2006. To appear.

[169] A. B. Owen. Randomly permuted (t, m, s)-nets and (t, s)-sequences. In Harald Niederreiter and Peter Jau-Shyong Shiue, editors, *Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing*, volume 106 of *Lecture Notes in Statistics*, pages 299–317. Springer, 1995.

[170] A. B. Owen. Monte Carlo Variance of Scrambled Net Quadrature. *SIAM Journal of Numerical Analysis*, 34 :1884–1910, 1997.

[171] A. B. Owen. Scrambled Net Variance for Integrals of Smooth Functions. *Annals of Statistic*, 25 :1541–1562, 1997.

[172] A. B. Owen and D. Tavella. Scrambled nets for value at risk calculations. In S. (Ed.) Grayling, editor, *VaR : Understanding and Applying Value-at-Risk*. RISK, London, 1997.

[173] A.B. Owen. Monte Carlo, quasi-Monte Carlo, and randomized quasi-Monte Carlo. In H. Niederreiter and J. Spanier, editors, *Monte Carlo and Quasi-Monte Carlo Methods 1998*, pages 86–97. Springer, 2000.

[174] J. Padhye, V. Firoiu, J. Kurose, and D. Towsley. Modeling TCP Throughput : A Simple Model and its Empirical Validation. In *Proceedings of ACM SIGCOMM 98*, pages 303–314, 1998.

[175] G. Pagès and Y.J. Xiao. Sequences with low discrepancy and pseudo-random numbers : theoretical results and numerical tests. *J. Statist. Comput. Simul.*, 56 :163–188, 1997.

[176] C. Papadopoulos. A New Technique for MTTF Estimation in Highly Reliable Markovian Systems. *Monte Carlo Methods and Applications*, 4(2) :95–112, 1998.

[177] I.Ch. Paschalidis and Y. Liu. Pricing in Multiservices Loss Networks : Static Pricing, Asymptotic Optimality, and Demand Substitution Effects. *IEEE/ACM Transactions on Networking*, 10(3), 2002.

[178] I.Ch. Paschalidis and J.N. Tsitsiklis. Congestion-Dependent Pricing of Network Services. *IEEE/ACM Transactions on Networking*, 8(2) :171–184, 2000.

[179] J.L. Peterson. *Petri nets and the Modeling of Systems*. Prentice-Hall, Englewood-Cliffs, NJ, 1981.

[180] J. Propp and D. Wilson. Exact sampling with coupled Markov chaines and applications to statistical mechanisms. *Random Structures and Algorithms*, 9 :223–252, 1996.

[181] A. Puri and P. Varaiya. Decidability of hybrid systems with rectangular inclusions. In D. Dill, editor, *Computer Aided Verification, CAV'94*, volume 818 of *Lecture Notes in Computer Science*, pages 95–104. Springer-Verlag, 1994.

[182] Y. Qiu and P. Marbach. Bandwidth allocation in ad hoc networks : A price-based approach. In *Proceedings of IEEE INFOCOM*, march 2003.

[183] P. Reichl, P. Flury, J. Gerke, and B. Stiller. How to overcome the feasibility problem for tariffing internet services : the cumulus pricing scheme. In *Proceedings ofIEEE ICC 2001, vol. 7*, pages 2079–2083, 2001.

[184] P. Reichl and B. Stiller. Edge pricing in space and time : Theoretical and practical aspects of the cumulus pricing scheme. In *Proceedings of the 17th International Teletraffic Congress*, 2001.

[185] P. Reichl, B. Stiller, and S. Leinen. Auction Models for Multiprovider Internet Connections. In *Proc. Messung, Modellierung und Bewertung MMB'99. Trier (Germany)*, 1999.

[186] J.W. Roberts. Quality of Service Guarantees and Charging in Multiservice Networks. *IEICE Trans. Commun.*, E81(5) :824–831, 1998.

[187] D. Ros. *Étude de réseaux haut débit via la simulation de modèles fluides*. PhD thesis, Institut National des Sciences Appliquées (INSA) de Rennes, January 2000.

[188] D. Ros and B. Tuffin. Charging in packet networks using the paris metro pricing scheme. In *Optimization Days 03*, Montréal, May 2003.

[189] D. Ros and B. Tuffin. A mathematical model of the Paris Metro Pricing scheme for charging packet networks. *Computer Networks*, 46(1) :73–85, September 2004.

[190] K.W. Ross and J. Wang. Implementation of Monte Carlo Integration for the Analysis of product-form queueing networks. *Performance Evaluation*, 29(4) :273–292, 1997.

[191] D. Sahu, S. Towsley and J. Kurose. A Quantitative Study of Differentiated Services for the Internet. *Journal of Communications and Networks*, 2 :127–137, 2000.

[192] Naouel Ben Salem, Levente Buttyan, Jean-Pierre Hubaux, and Markus Jakobsson. A charging and rewarding scheme for packet forwarding in multi-hop cellular networks. In *Proceedings of the 4th ACM international symposium on Mobile ad hoc networking & computing*, pages 13–24. ACM Press, 2003.

[193] W. Sandmann. Relative error and asymptotic optimality in estimating rare event probabilities by importance sampling. In *Proceedings of the OR Society Simulation Workshop (SW04) held in cooperation with the ACM SIGSIM, Birmingham, UK, March 23–24, 2004*, pages 49–57. The Operational Research Society, 2004.

[194] C.U. Saraydar, N.B. Mandayam, and D.J. Goodman. Efficient power control via pricing in wireless data networks. *IEEE Transactions on Communications*, 50(2) :291–303, 2002.

[195] F. Schreiber and C. Görg. Rare Event Simulation : a modified RESTART Method using LRE-Algorithm. In D.T. Brunner J.M. Charnes, D.J. Morrice and J.J. Swain, editors, *Proceedings of the 1996 Winter Simulation Conference*, pages 390–397, 1996.

[196] L.W. Schruben. A coverage function for interval estimators of simulation response. *Management Science*, 26(1) :18–27, 1980.

[197] N. Semret. *Market Mechanisms for Network Resource Sharing*. PhD thesis, Columbia University, 1999.

[198] N. Semret, R.R.-F. Liao, A.T. Campbell, and A.A. Lazar. Market Pricing of Differentiated Internet Services. In *Proceedings of the 7th International Workshop on Quality of Service*, 1999.

[199] B. Sericola and B. Tuffin. A Fluid Queue Driven by a Markovian Queue. *Queuing Systems : Theory and Applications*, pages 253–264, 1999.

[200] P. Shahabuddin. Importance Sampling for the Simulation of Highly Reliable Markovian Systems. *Management Science*, 40(3) :333–352, March 1994.

[201] J. Shu and P. Varaiya. Pricing network services. In *Proceedings of IEEE INFOCOM*, march 2003.

[202] V.A. Siris. Resource control for elastic traffic in CDMA networks. In *8th international conference on Mobile computing and networking*, pages 193–204, Atlanta, USA, 2002. ACM Press.

[203] I. H. Sloan and P. J. Kachoyan. Lattice Methods for Multiple Integration : Theory, Error Analysis and Examples. *SIAM Journal of Numerical Analysis*, 24(1) :116–128, February 1987.

[204] R. Sommer and A. Feldmann. Netflow : Information loss or win. In *Proc. of ACM SIGCOMM'02 Internet Measurement Workshop*, Pittsburg, PA, August, 19-23 2002.

[205] D. Songhurst (ed.). *Charging Communication Networks : from Theory to practice*. Elsevier, Amsterdam, 1999.

[206] J. Spanier. Quasi-Monte Carlo Methods for Particle Transport Problems. In Harald Niederreiter and Peter Jau-Shyong Shiue, editors, *Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing*, volume 106 of *Lecture Notes in Statistics*, pages 121–148. Springer, 1995.

[207] J. Spanier and L. Li. Quasi-Monte Carlo methods for integral equations. In H. Niederreiter, P. Hellekalek, G. Larcher, and P. Zinterhof, editors, *Monte Carlo and Quasi-Monte Carlo Methods 1996*, Lecture Notes in Statistics, pages 398–414. Springer, 1998.

[208] B. Stiller, P. Reichl, and S. Leinen. Pricing and Cost Recovery for Internet Services : Practical Review, Classification, and Application of Relevant Models. *Netnomics*, 2(1), 2000.

[209] J. Struckmeier. Fast generation of low-discrepancy sequences. *Journal of COmputational and Applied Mathematics*, 61 :29–41, 1995.

[210] B. Tuffin. Improvement of Halton sequences distribution. Technical Report 998, IRISA, March 1996.

[211] B. Tuffin. On the Use of Low Discrepancy Sequences in Monte Carlo Methods. *Monte Carlo Methods and Applications*, 2(4) :295–320, 1996.

[212] B. Tuffin. Variance reduction technique for a cellular system with dynamic resource sharing. In *Proceedings of the 10th European Simulation Multiconference*, pages 467–471, Budapest, June 1996.

[213] B. Tuffin. A new Permutation Choice in Halton Sequences. In G. Larcher H. Niederreiter, P. Hellekalek and P. Zinterhof, editors, *Second Internatinal Conference on Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing*, volume 127 of *Lecture Notes in Statistics*. Springer Verlag, 1997.

[214] B. Tuffin. Réduction des bornes de l'erreur d'approximation des méthodes de quasi-Monte Carlo. Technical Report 1133, IRISA, October 1997.

[215] B. Tuffin. *Simulation accélérée par les méthodes de Monte Carlo et quasi-Monte Carlo : théorie et applications*. PhD thesis, Université de Rennes 1, Octobre 1997.

[216] B. Tuffin. Variance Reductions applied to Product-Form Multi-Class Queuing Network. *ACM Transactions on Modeling and Computer Simulation*, 7(4) :478–500, 1997.

[217] B. Tuffin. Variance reduction order using good lattice points in Monte Carlo methods. *Computing*, 61(4) :371–378, 1998.

[218] B. Tuffin. Bounded Normal Approximation in Simulations of Highly Reliable Markovian Systems. *Journal of Applied Probability*, 36(4) :974–986, 1999.

[219] B. Tuffin. A randomized Quasi-Monte Carlo method for the simulation of product-form loss networks. In *Proceedings of the 9th International Conference on Telecommunication Systems*, pages 468–478, Dallas, USA, March 2001.

[220] B. Tuffin. Pricing schemes in telecommunication networks. In *4th Latino-american days on Economic Theory*, Mexico City, October 2002.

[221] B. Tuffin. Revisited Progressive Second Price Auctions for Charging Telecommunication Networks. *Telecommunication Systems*, 20(3) :255–263, 2002.

[222] B. Tuffin. Charging the Internet without bandwidth reservation : an overview and bibliography of mathematical approaches. *Journal of Information Science and Engineering*, 19(5) :765–786, 2003.

[223] B. Tuffin. Some new research directions about hybrid monte carlo methods. In *Vth IMACS Seminar on Monte Carlo Methods MCM-2003*, Berlin, September 2003.

[224] B. Tuffin. On numerical problems in simulations of Highly Reliable Markovian Systems. In *Proceedings of the 1st International Conference on Quantitative Evaluation of SysTems (QEST)*, University of Twente, Enschede, the Netherlands, September 2004. IEEE CS Press.

[225] B. Tuffin. Coverage function of randomized quasi-monte carlo methods. In *INFORMS Applied Probability Conference*, July 2005.

[226] B. Tuffin, D.S. Chen, and K.S. Trivedi. Comparison of Hybrid Systems and Fluid Stochastic Petri Nets. *Discrete Event Dynamic Systems*, 11(1& 2) :77–96, 2001.

[227] B. Tuffin, C. Hirel, and K.S. Trivedi. Simulation versus analytic-numeric methods : a petri net example. En soumission.

[228] B. Tuffin and L-M. Le Ny. Parallélisation d'une combinaison des méthodes de monte carlo et quasi-monte carlo et application aux réseaux de files d'attente. *RAIRO : Recherche Opérationnelle*, 34(1) :85–98, 2000.

[229] B. Tuffin and L-M. Le Ny. Modeling and analysis of threshold queues with hysteresis using stochastic Petri nets : the monoclass case. In *Proceedings of Petri Nets and Performance Models*, pages 175–184. IEEE CS Press, 2001.

[230] B. Tuffin, P. Rodriguez, and H. Cancela. End-to-end reliability-dependent pricing of network services. In *Proceedings of the 12th Latin-Ibero-American Conference on Operations Research (CLAIO'04)*, 2004.

[231] B. Tuffin and K.S. Trivedi. Implementation of importance splitting techniques in stochastic Petri net package. In B.R. Haverkort, H.C. Bohnenkamp, and C.U. Smith, editors, *Computer performance evaluation : Modelling tools and techniques ; 11th International Conference ; TOOLS 2000, Schaumburg, Il., USA*, volume 1786 of *Lecture Notes in Computer Science*, pages 216–229. Springer Verlag, 2000.

[232] B. Tuffin and K.S. Trivedi. Importance sampling for the simulation of stochastic Petri nets and fluid stochastic petri nets. In *Proceedings of High Performance Computing (HPC)*, pages 228–235, Seattle, USA, April 2001.

[233] P. Varaiya. Design, simulation and implementation of hybrid systems. In Susanna Donatelli and Jetty Kleijn, editors, *Application and Theory of Petri Nets 1999*, volume 1639 of *Lecture Notes in Computer Science*, pages 1–5. Springer-Verlag, 1999.

[234] H.R. Varian. The Demand for Bandwidth : Evidence from the INDEX Experiment. Technical report, UC Berkeley, School of Information Management & Systems, 2001.

[235] M. Villen-Altamirano, A. Martinez-Marron, J. Gamo, and M. Fernandez-Cuesta. Enhancement of the accelerated simulation method restart by considering multiple thresholds. In Elsevier J. Labetoulle, J.W. Roberts, editor, *Proceedings of the 14th International Teletraffic Congress, the Fundamental Role of Teletraffic in the Evolution of Telecommunications Networks*, pages 787–810, 1994.

[236] M. Villen-Altamirano and J. Villen-Altamirano. RESTART : A Method for Accelerating Rare Event Simulations. In J.W. Cohen and C.D. Pack, editors, *Proceedings of the 13th International Teletraffic Congress, Queueing Performance and Control in ATM*, pages 71–76, 1991.

[237] M. Villen-Altamirano and J. Villen-Altamirano. RESTART : A Straightforward Method for Fast Simulation of Rare Event. In D.A. Sadowski J.D. Tew, S. Manivannan and A.F. Seila, editors, *Proceedings of the 1994 Winter Simulation Conference*, pages 282–289, 1994.

[238] M. Villen-Altamirano and J. Villen-Altamirano. RESTART : An Efficient and General Method for Fast Simulation of Rare Event. Technical Report 7, Departamento de Maetmatica Aplicada, E.U. Informática, Universidad Politécnica de Madrid, 1997.

[239] J.M. Vincent and C. Marchand. On the exact simulation of functionals of stationary markov chains. In *Fourth International Conference on the Numerical Solution of Markov Chains (NSMC'03)*, pages 77–97, 2003.

[240] Q. Wang, J.M. Peha, and M.A. Sirbu. Optimal Pricing for Integrated Services Networks. In Lee W. McKnight and Joseph P. Bailey, editors, *Internet Economics*, pages 353–376. MIT Press, 1997.

[241] X. Wang and F.J. Hickernell. Randomized Halton sequences. *Mathematical and Computer Modelling*, 32 :887–899, 2000.

[242] L.S. Zhang, D. Deering, S. Shenker, and D. Zappala. RSVP : A resource ReSerVation Protocol. *IEEE Network Magazine*, 1993.