ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

THÈSE PRÉSENTÉE À
L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

COMME EXIGENCE PARTIELLE
À L'OBTENTION DU
DOCTORAT EN GÉNIE

Ph.D.

PAR
Côn TRAN

ADAPTATION DE CAPACITÉ DANS LE RÉSEAU DÉDIÉ DE SERVICE POUR LA MAXIMISATION DU BÉNÉFICE

MONTRÉAL, LE 11 AOUT 2011

PRÉSENTATION DU JURY

CETTE THÈSE A ÉTÉ ÉVALUÉE

PAR UN JURY COMPOSÉ DE :

M. Zbigniew Dziong, directeur de thèse
Département de génie électrique à l'École de technologie supérieure

M. Pierre-Jean Lagacé, président du jury
Département de génie électrique à l'École de technologie supérieure

M. Michel Kadoch, membre du jury
Département de génie électrique à l'École de technologie supérieure

Dr. Jean Régnier, examinateur externe
Consultant sénior

ELLE A FAIT L'OBJET D'UNE SOUTENANCE DEVANT JURY ET PUBLIC

LE 20 JUILLET 2011

À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

REMERCIEMENTS

Je voudrais tout d'abord exprimer ici ma profonde reconnaissance envers mon directeur de thèse, le professeur Zbigniew Dziong, dont le support, les conseils et la patience m'ont permis de compléter ce travail. Son aide m'a grandement encouragé tout au long de cette entreprise.

Je voudrais ensuite remercier le directeur du laboratoire Lagrit de l'École de Technologie Supérieure, le professeur Michel Kadoch, qui m'a fourni les nombreuses ressources facilitant l'accomplissement de mon travail de recherche.

Je voudrais aussi remercier les membres de mon jury, les professeurs Pierre-Jean Lagacé et Michel Kadoch et le docteur Jean Régnier, qui voudront bien examiner et évaluer ce travail.

Aussi grande est mon affectueuse gratitude envers mon épouse Thi-Nga Vu et mes trois enfants chéris Luan, Lan et Quang pour leur patience, compassion et encouragements au cours de ces longues années où j'ai dû partager mon temps entre eux et mes études.

Je n'oublie pas mes confrères et consœurs chercheurs du laboratoire Lagrit. Notre esprit de camaraderie et d'entraide a rendu ce difficile travail bien plus plaisant.

ADAPTATION DE CAPACITÉ DANS LE RÉSEAU DÉDIÉ DE SERVICE POUR LA MAXIMISATION DU BÉNÉFICE

Côn TRAN

RÉSUMÉ

Les réseaux dédiés de service (*Service Overlay Network* – SON) sont formés en plaçant des nœuds dédiés dans l'Internet et en les reliant par des liens dédiés établis avec de la bande passante, incluant garantie de Qualité de Service (QoS), achetée d'une multitude de Systèmes Autonomes Internet (AS). La bande passante est obtenue par accords de niveau de service (*Service Level Agreement* – SLA) avec les fournisseurs de service Internet qui sont propriétaires des AS. Par sa couverture sur la multitude de AS, le SON peut alors fournir la QoS de bout en bout aux connexions de services en temps réel qu'il admet.

La quantité de largeur de bande achetée a un effet à la fois sur le trafic admis et sur le coût du réseau, affectant ainsi le bénéfice de l'opérateur du réseau. Cela donne à ce dernier la possibilité d'optimiser son bénéfice en adaptant ses ressources de réseau aux conditions changeantes de la demande de trafic et des coûts de SLA. Deux éléments sont requis pour réaliser l'optimisation : une méthode efficace d'estimation en temps réel de la demande de trafic, et une approche d'adaptation de capacité optimale qui sera alimentée par l'estimation de la demande. Dans cette thèse, nous proposons des nouvelles approches pour répondre à ces besoins identifiés.

Notre approche d'adaptation de capacité maximise le bénéfice du réseau en changeant lorsque requis les termes des SLA, afin d'adapter l'attribution de la largeur de bande des liens dédiés à de fréquentes estimations de demande de trafic. Pendant que s'effectue l'adaptation de capacité, le degré de service du réseau, spécifié par des contraintes de blocage de connexions, doit être maintenu. En s'appuyant sur un modèle économique, l'adaptation des ressources de lien est intégrée à la politique de contrôle d'admission de connexions (CAC) et de routage en vigueur dans le réseau. Dans notre approche, nous proposons d'appliquer une politique de CAC et routage, fondée sur la théorie de décision de Markov, qui maximise la récompense du réseau, bien que d'autres politiques de routage soient aussi applicables. L'intégration mène à un algorithme itératif et distribué d'adaptation de capacité de lien, où la sensibilité du bénéfice du réseau aux dimensions de lien est calculée à partir de la moyenne du *shadow price* de lien, qui lui-même constitue un paramètre du routage.

Des approches d'estimation en temps réel de tendance de demande de trafic, fondée sur des mesures et convenant bien à l'adaptation de capacité du SON, sont ensuite proposées. Dans la première approche, le paramètre du modèle de lissage exponentiel (Exponential Smoothing – ES) est adapté à la tendance de trafic. La tendance dans ce cas est estimée en utilisant les fonctions d'autocorrélation et de distribution cumulée de mesures de taux d'arrivées de connexion. La deuxième approche applique un filtre de Kalman dont le modèle est construit à partir de données historiques de trafic. Dans ce dernier cas, la disponibilité de la

distribution de l'erreur d'estimation dans l'algorithme du filtre permet d'élaborer une méthode qui améliore le contrôle du degré de service durant l'adaptation de capacité.

L'analyse ainsi que la simulation des modèles proposés ont été effectuées pour évaluer la performance des approches. La maximisation du bénéfice du réseau par l'approche d'adaptation de capacité est confirmée avec une étude analytique d'exemples de petits réseaux. Ensuite, la simulation d'exemples réalistes de réseau démontre les meilleurs bénéfices et/ou degré de service obtenus par notre approche d'adaptation de capacité, quand on la compare à l'attribution fixe de capacités de liens.

L'étude de l'estimation de demande de trafic montre que la performance de l'adaptation de capacité est améliorée par l'usage de nos méthodes proposées. Les méthodes adaptatives procurent des bénéfices plus élevés que celles à paramètre fixe. L'estimation par lissage exponentiel adaptée par autocorrélation donne la meilleure performance combinée réponse-stabilité, quand on la compare aux autres lissages exponentiels. Enfin, les approches fondées sur filtre de Kalman améliorent l'adaptation de capacité, démontrant des réductions significatives de l'augmentation du blocage du réseau quand la demande de trafic augmente.

Mots clés : réseau dédié de service, gestion de ressources, adaptation de capacité, estimation de trafic, processus de décision de Markov, filtre de Kalman.

SERVICE OVERLAY NETWORK CAPACITY ADAPTATION FOR PROFIT MAXIMIZATION

Côn TRAN

ABSTRACT

Service Overlay Networks (SON) are formed by placing overlay nodes over the Internet and interconnecting them with overlay links established by leasing bandwidth with Quality of Service (QoS) guarantees from a multitude of Internet Autonomous Systems (AS). Bandwidth is leased through service level agreements (SLA) with Internet Service Providers owning the AS. By covering a multitude of AS, the SON can provide end-to-end QoS to real time service connections serviced by its network.

The amount of leased bandwidth influences both the admitted traffic and network cost, affecting the network profit. This gives the network operator the opportunity to optimize the profit by adapting the network resources to changing traffic and SLA cost conditions. To realize this optimization, two elements are required: an efficient real time estimation of traffic demand, and an optimal capacity adaptation that will be fed by the demand estimation. In this thesis, we propose novel approaches that address the mentioned elements.

The approach for capacity adaptation maximizes network profit by modifying the SLA terms as needed to adapt overlay links bandwidth allocation to frequently updated traffic demand estimates. While performing capacity adaptation, the network Grade of Service, specified by connection blocking constraints, must also be maintained. Using an economic model, the link resources adaptation is integrated with the connection admission control (CAC) and routing policy in effect in the network. In our proposal, we apply a reward maximizing CAC and routing policy that is derived from the Markov Decision Process theory, although the approach can be applied to other routing policies. This integration leads to a distributed iterative algorithm for link bandwidth adaptation, where network profit sensitivity to link dimensions is calculated from the average link shadow price, itself being a routing parameter.

Approaches for measurement-based online traffic trend estimation that fit the SON capacity adaptation are next proposed. In the first approach, the smoothing parameter of the exponential smoothing (ES) model is adapted to the traffic trend. Here, the trend is estimated using measured connection arrival rate autocorrelation or cumulative distribution functions. The second approach applies a Kalman filter whose model is built from historical traffic data. In this case, availability of the estimation error distribution provided by the filter algorithm allows for better control of the network Grade of Service during capacity adaptation.

Analytical models as well as simulation of measurement based implementations of the proposed models are used to evaluate the performance of the proposed approach. Profit maximization by the capacity adaptation approach is confirmed by analysis of small network examples. Simulations on realistic network examples demonstrates higher network profit

and/or better Grade of Service obtained by our adaptation proposal, when compared to fixed link capacities.

Study of traffic demand estimation methods shows that capacity adaptation performance is further improved with the use of our proposed methods. Adaptive estimation, as compared to fixed parameter estimation, provided higher profits. The proposed autocorrelation based ES estimation gives the best combined response and stability performances when compared to known ES methods. The proposed Kalman filter based approach improves capacity adaptation performance by significantly limiting network blocking increase when traffic demand increases.

Keywords: Service overlay network, resource management, capacity adaptation, traffic estimation, Markov decision process, Kalman filter.

TABLE DES MATIÈRES

Page

LISTE DES TABLEAUX

Page

LISTE DES FIGURES

Page

LISTE DES ABRÉVIATIONS, SIGLES ET ACRONYMES

acf	autocorrelation function (fonction d'autocorrélation)
cdf	cumulative distribution function (fonction de distribution cumulée)
AEES	Adaptive Extended Exponential Smoothing
AS	Autonomous System (système autonome)
CAC	Contrôle d'Admission de Connexions
DES	Double Exponential Smoothing (lissage exponentiel double)
DiffServ	Differentiated Service
ES	Exponential Smoothing (lissage exponentiel)
GoS	Grade of Service (degré de service)
IntServ	Integrated Service
IPv6	Internet Protocol version 6
ISP	Internet Service Provider (fournisseur de service Internet)
MA	Moving Average (moyenne mobile)
MDP	Markov Decision Process
MDPD	Markov Decision Process Decomposed
MPLS	Multiprotocol Label Switching
OD	Origine-Destination
OL	Overlay Link (lien dédié)
ON	Overlay Node (nœud dédié)
P2P	Peer to peer (pair à pair)
QoS	Quality of Service (qualité de service)
SES	Simple Exponential Smoothing (lissage exponentiel simple)

SLA	Service Level Agreement (accord de niveau de service)
SON	Service Overlay Network (réseau dédié de service)
TE	Traffic Engineering
TOD	Time of Day
VN	Virtual Network (réseau virtuel)
VoIP	Voice over Internet Protocol (Voix sur IP)
VPN	Virtual Private Network (réseau privé virtuel)

INTRODUCTION

L'Internet est maintenant universellement utilisé comme réseau de communications, autant par les usagers individuels que par les institutions telles que les gouvernements et universités ainsi que le monde industriel. Au point de vue organisationnel, le réseau est formé par l'interconnexion d'une multitude de Systèmes Autonomes (AS) (Internet Assigned Number Authority., 2005). Les fournisseurs de service Internet (ISP), tels que les opérateurs Bell, MCI ou Sprint, gèrent indépendamment les uns des autres un ou plusieurs de ces AS. Fonctionnellement, l'Internet a été conçu comme un réseau *best effort*, n'offrant aucune garantie de performance ou de Qualité de Service (QoS), cette dernière étant définie en général par des limites aux pertes de données, délais de transmission et gigues (Oodan, Books24x7 Inc. et Institution of Electrical Engineers., 2003). Bien que certaines techniques aient été développées plus récemment pour supporter la QoS, l'absence de collaboration entre les ISP en limite la portée au domaine individuel de l'ISP.

Aux simples applications initiales de courriers électroniques, transferts de fichiers ou recherches d'information sur le *web* s'est maintenant ajoutée une variété toujours grandissante et plus sophistiquée de nouvelles applications. Une récente catégorie est constituée par les applications interactives comme la Voix sur IP (VoIP), le *Streaming* Multimédia et les jeux interactifs. Pour une utilisation plus satisfaisante, ces applications en temps réel ont besoin d'un niveau garanti de QoS de bout en bout de la connexion. Aussi, avec la popularité de l'Internet, les usagers exigent de plus en plus une performance améliorée des nouvelles applications. Or, comme mentionné, une garantie de performance ou de QoS de bout en bout n'est pas disponible dans le réseau Internet natif.

Les réseaux dédiés (*overlay networks*) ont été conçus pour fournir une solution pratique aux problèmes de performance et de QoS de bout en bout de l'Internet. Parmi ceux-ci, nous retrouvons les réseaux pair à pair (P2P) et les réseaux dédiés de service (SON). Pour établir la capacité de son réseau dédié de service, un opérateur peut acheter de la bande passante aux AS par l'entremise de contrats de service appelés *Service Level Agreement* (SLA). Les SLA

spécifient, entre autres, la quantité de bande passante, la QoS associée, le prix ainsi que les termes de changements de la quantité achetée (Evans et Filsfils, 2004).

0.1 Problématique de recherche

Pour tenter d'offrir sur l'Internet la QoS de bout en bout requise par les applications en temps réel, plusieurs propositions ont été avancées, comme DiffServ, IntServ, MPLS-TE et IPv6. Cependant, la réalisation de ces propositions a jusqu'à maintenant été limitée au territoire individuel de chaque ISP, à cause de l'absence d'un protocole inter-AS de communication de spécifications QoS ainsi que du peu d'intérêts de collaboration entre les ISP. Une approche plus pratique ne nécessitant aucun changement à la couche Transport de l'Internet est d'utiliser un réseau dédié de service réalisé par dessus ce service de transport. Dans un tel SON, les liens sont établis en achetant la bande passante requise, incluant la QoS, aux divers AS par le biais des SLA. Comme l'opérateur du SON contrôle tous les liens couvrant globalement son réseau, il peut en contrôler la performance et y offrir la QoS de bout en bout à ses connexions.

Dans cette thèse, nous étudions le sujet de la gestion (*network management*) en temps réel du SON, en particulier de ses ressources. Ci-après, nous évoquons les problématiques reliées à ce sujet.

- Un objectif primordial d'un opérateur est l'exploitation profitable de son réseau. La profitabilité peut être réalisée, entre autres, par une provision judicieuse de la capacité des liens du réseau couplée à un routage efficace des connexions. La capacité provisionnée permettra un service avec QoS à la demande de trafic tout en évitant un gaspillage de la bande passante. De son côté, un routage efficace permettra une utilisation optimale de cette capacité. Ainsi, la première problématique peut s'exprimer comme suit :

  Dans le cas d'un SON où la capacité des liens est achetée des AS pour servir sa demande de trafic, *comment gérer ces capacités pour maximiser le bénéfice de l'opérateur, tout en*

garantissant la QoS des connections dans le réseau ? Et *comment effectuer le routage des connexions pour obtenir ce bénéfice maximisé* ? Traditionnellement, les liens sont provisionnés lors de la planification du réseau, en se basant sur des prévisions soit de la demande moyenne ou de la pointe de demande du trafic. Comme le niveau de demande peut varier de façon significative durant la journée, une allocation fixe de capacité basée sur la demande moyenne a le désavantage de causer une irritante congestion des connexions durant les périodes de pointe. Dans l'autre cas d'une allocation basée sur la pointe de demande, la capacité est gaspillée durant les périodes creuses qui en général occupent la majorité du temps. Avec l'utilisation plus récente du routage dynamique, l'inefficacité de l'allocation fixe est davantage amplifiée à cause de l'absence d'adaptation de la capacité au routage.

- Afin de pouvoir garantir la QoS, une nouvelle demande de connexion sera refusée si le réseau ne dispose pas de capacité libre suffisante pour son service. Ceci mène à une perception par les usagers d'un niveau de service, connu comme *Grade of Service* (GoS) et quantifié par le taux de blocage (refus) des demandes de connexion. La deuxième problématique considérée sera alors : *vu l'objectif de la maximisation du bénéfice, comment garder le taux de blocage des connexions en dedans de contraintes définies, afin de maintenir la satisfaction des usagers ?*

- Une solution optimale aux problématiques citées ci-dessus est dépendante d'une évaluation ponctuelle et précise de la demande de trafic offerte au réseau. Ainsi, une autre problématique reliée aux précédentes sera : *quelles seront les méthodes d'évaluation de demande de trafic qui conviendront efficacement à la gestion de la capacité du SON ? Aussi, devrait-on et peut-on tenir compte des imprécisions de l'évaluation ?*

- Enfin, un objectif aussi important de toute méthode de gestion de réseau est sa capacité d'application à des réseaux de grandes tailles, caractéristique communément connu sous le terme « extensibilité » (*scalability*). Cette problématique est également considérée dans notre étude.

Une approche de gestion en temps réel des ressources du réseau devra apporter une solution intégrée qui répond de façon transparente simultanément à toutes les problématiques mentionnées ci-dessus.

0.2 Contributions de recherche

Comme la QoS sur Internet est une exigence dans notre étude et que le SON constitue un moyen pratique de la fournir, nous nous concentrons sur ce type de réseau. Le concept de la QoS sur Internet par le SON est appuyé par la présence de bon nombre de fournisseurs commerciaux de réseaux privés virtuels (VPN) dont Claranet (www.uk.clara.net), Internap (www.internap.com), Virtela (www.virtela.com), et Azzurri (www.azzurricommunications.com) qui l'utilisent. Certains de ces fournisseurs publicisent l'utilisation d'algorithmes propriétaires qui surveillent les dorsales de l'Internet pour trouver les chemins optimaux avec QoS pour servir leurs connexions. Pour pouvoir garantir la QoS, nous choisissons d'étudier le SON dont les liens sont établis par l'achat de bande passante avec QoS aux AS. Nous assumons que la bande passante est disponible dans les réseaux AS.

Ci-après, nous énumérons les contributions originales apportées par notre recherche dans le domaine de la gestion de ressources du SON.

0.2.1 Adaptation de capacité fondée sur le processus de décision de Markov

Pour maximiser le bénéfice du SON, nous proposons une approche d'adaptation en ligne de la capacité du réseau à la demande de trafic et aux prix de SLA. L'approche, fondée sur un modèle économique, maintient la maximisation du bénéfice en s'appuyant sur des mesures fréquentes de trafic et de performance. L'adaptation de capacité est basée sur le concept du *shadow price* moyen découlant de l'utilisation de la théorie du processus de décision de Markov. Les capacités optimales déterminées sont alors attribuées aux liens en effectuant des changements à la largeur de bande achetée par SLA et/ou par l'établissement de nouveaux SLA.

0.2.2 Intégration de la politique de routage et de l'adaptation de capacité

Par l'utilisation commune du *shadow price* comme paramètre, l'adaptation de capacité est étroitement intégrée avec la politique d'admission et de routage des connexions, fondée sur le processus de décision de Markov, en vigueur dans le réseau. Comparée à d'autres approches où l'adaptation doit être fondée sur une politique de routage approximée, l'intégration de l'actuelle politique de routage permet une meilleure efficacité de la maximisation de bénéfice.

0.2.3 Réalisation distribuée de l'algorithme d'adaptation

L'adaptation de capacité est réalisée par un algorithme itératif distribué aux liens du réseau, où la sensibilité du bénéfice de réseau par rapport à la dimension de chaque lien est calculée à partir du *shadow price* moyen du lien. Cette approche distribuée est importante car elle permet de réduire la complexité du traitement reliée à la cardinalité considérable des états dans les grands réseaux, et ainsi de réaliser une implantation pratique de la solution.

0.2.4 Maintien de GoS par adaptation de paramètre de récompense

Pour maintenir la GoS, nous intégrons aux adaptations de capacité et de routage une approche de satisfaction de contraintes de blocage de connexions fondée sur l'adaptation du paramètre de récompense du routage MDP.

0.2.5 Estimation de la tendance de demande de trafic

Une demande de trafic typique subit des changements significatifs dans une journée. Pour obtenir une *évaluation ponctuelle et précise de cette demande* et ainsi une optimisation valable du SON, nous proposons une méthode novatrice d'estimation de la tendance de la demande, fondée sur la fonction d'autocorrélation et la fonction de distribution cumulée du processus d'arrivée des connexions.

0.2.6 Estimation de demande de trafic par filtre de Kalman

Alternativement, nous proposons une méthode d'estimation de la demande fondée sur le filtre de Kalman qui procure, concurremment avec la valeur estimée, une évaluation de l'erreur d'estimation.

0.2.7 Amélioration du respect de la GoS fondée sur l'erreur d'estimation

Avec la proposition d'estimation de trafic par filtre de Kalman, nous incluons une approche d'assurance de GoS fondée sur la connaissance de la distribution de l'erreur d'estimation.

0.3 Liste des publications produites

La recherche a donné lieu à la publication des articles suivants.

0.3.1 Journaux

Tran, Con, et Zbigniew Dziong. 2011. « Traffic trend estimation for profit oriented capacity adaptation in service overlay networks ». accepté le 23 juin 2011 pour publication, *IEEE Transactions on Network and Service Management*.

Tran, Con, et Zbigniew Dziong. 2010b. « Service overlay network capacity adaptation for profit maximization ». *IEEE Transactions on Network and Service Management,* vol. 7, n^{o} 2, p. 72-82.

0.3.2 Conférences

Tran, Con, et Zbigniew Dziong. 2010c. « Traffic dynamics online estimation based on measured autocorrelation ». In *7th International ICST Conference on Broadband Communications, Networks, and Systems, October 25, 2010 - October 27, 2010*. Coll. « 7th International ICST Conference on Broadband Communications, Networks, and Systems - Proceedings ». Athens, Greece: Springer.

Tran, Con, et Zbigniew Dziong. 2010a. « Kalman filter based capacity adaptation for network profit maximization ». In *2010 14th International Telecommunications Network Strategy and Planning Symposium, Networks 2010, September 27, 2010 - September 30, 2010*. Coll. « Proceedings of 2010 14th International Telecommunications Network Strategy and Planning Symposium, Networks 2010 ». Warsaw, Poland: IEEE Computer Society. <http://dx.doi.org/10.1109/NETWKS.2010.5624942>.

Tran, Con, et Zbigniew Dziong. 2008. « Resource adaptation for continuous profit optimization in Overlay and Virtual Networks ». In *4th EURO-NGI Conference on Next Generation Internet Networks, April 28, 2008 - April 30, 2008*. p. 131-138. Coll. « 4th EURO-NGI Conference on Next Generation Internet Networks - Proceedings ». Krakow, Poland: Inst. of Elec. and Elec. Eng. Computer Society.

Tran, Con, Jahangir Sarker et Zbigniew Dziong. 2007. « Distributed resource adaptation for virtual network operators ». In *4th International Conference on Distributed Computing and Internet Technology, ICDCIT 2007, December 17, 2007 - December 20, 2007*. Vol. 4882 LNCS, p. 146-157. Coll. « Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) ». Bangalore, India: Springer Verlag.

Dziong, Zbigniew, Con Tran et Jahangir Sarker. 2007. « Integrated economic model for design and adaptation of overlay networks ». In *2007 1st International Symposium on Advanced Networks and Telecommunications Systems, ANTS, December 17, 2007 - December 18, 2007*. Coll. « 2007 1st International Symposium on Advanced Networks and Telecommunications Systems, ANTS ». Mumbai (Bombay), India: Inst. of Elec. and Elec. Eng. Computer Society.

Tran, Con, Zbigniew Dziong et Michal Pioro. 2007. « SLA adaptation for service overlay networks ». In *6th International IFIP-TC6 Networking Conference - NETWORKING 2007 Ad Hoc and Sensor Networks, Wireless Networks, Next Generation Internet, May 14, 2007 - May 18, 2007*. Vol. 4479 LNCS, p. 691-702. Coll. « Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) ». Atlanta, GA, United states: Springer Verlag.

Tran, Con, et Zbigniew Dziong. 2006. « Service overlay network resource adaptations based on an economic model ». In *2006 Canadian Conference on Electrical and Computer Engineering, CCECE'06, May 7, 2006 - May 10, 2006*. p. 1821-1824. Coll. « Canadian Conference on Electrical and Computer Engineering ». Ottawa, ON, Canada: Institute of Electrical and Electronics Engineers Inc.

0.4 Méthodologie de recherche

Notre proposition de gestion de ressources comprend deux aspects développés en deux étapes : adaptation de capacité du réseau au trafic et estimation de la demande de trafic comme paramètre de l'adaptation.

Pour l'adaptation de capacité, nous développons en premier lieu un modèle théorique fondé sur la théorie du processus de décision de Markov. Utilisant un cadre économique établi pour le SON, l'approche développée intègre une politique d'admission et de routage dépendante d'état[1] maximisant le revenu, tirée de (Dziong, 1997), avec notre proposition d'adaptation de capacité. Un algorithme itératif de minimisation du gradient du bénéfice de réseau est utilisé pour converger les capacités des liens vers leurs valeurs optimales. Le bénéfice est déterminé par l'application de la politique d'admission et de routage. Nous désignons ce modèle procurant une solution exacte par adaptation de capacité MDP (Markov Decision Process).

Pour pouvoir s'appliquer à des grands réseaux où la cardinalité des états est très élevée, la complexité de l'implantation de la solution exacte de réseau devra être réduite. Utilisant une hypothèse d'indépendance de liens, nous effectuons une décomposition du problème réseau en problèmes de liens indépendants qui résulte en une solution approximative. Dans cette approche décomposée et désignée par adaptation de capacité MDPD (Markov Decision Process Decomposed), l'intégration du routage et de l'adaptation de capacité est réalisée par le *shadow price* de lien qui contrôle à la fois ces deux processus. Des analyses numériques comparatives sont alors effectuées sur les approches MDP et MDPD dans des réseaux de tailles limitées pour valider la maximisation du bénéfice par l'adaptation MDPD.

Pour satisfaire aux contraintes de GoS du SON, nous proposons d'agir sur les paramètres de récompense de l'approche MDPD, qui peuvent contrôler le routage et l'adaptation de capacité. Cette adaptation de récompense est intégrée à celles du routage et de capacité pour former une solution globale transparente à la gestion des ressources du SON.

[1] dépendante d'état indique que l'action est dépendante de l'état d'une entité (e.g. du lien, du réseau, etc.)

La performance de l'approche intégrée MDPD est vérifiée à l'aide d'un simulateur à événements de réseau. Bien qu'en général notre approche peut s'appliquer à des réseaux SON multi-classes hétérogènes, pour cause de simplification, nous en limitons l'implantation à des réseaux homogènes. Dans cette première phase du travail, nous vérifions l'approche en utilisant des méthodes de base pour l'estimation de la demande de trafic.

Dans la deuxième étape, nous élaborons des méthodes d'estimation de la demande de trafic qui permettront une adaptation plus précise et une probabilité améliorée de satisfaction de GoS. Un lissage exponentiel (ES) à coefficient variable pour tenir compte de la tendance du trafic est développé. La tendance est évaluée en utilisant les fonctions d'autocorrélation (acf) et de distribution cumulée (cdf) du processus d'arrivée des connexions. L'adaptation du coefficient permettra une meilleure stabilité de l'estimation dans les périodes de trafic stationnaire, couplée à une réponse plus rapide à la tendance en périodes non stationnaires.

Pour tenir compte des erreurs reliées à l'estimation qui peuvent provoquer des dépassements des contraintes de GoS, nous développons aussi une méthode alternative d'estimation fondée sur le filtre de Kalman. Dans le cadre d'un système linéaire perturbé par du bruit gaussien blanc, le filtre est reconnu comme un estimateur optimal qui également procure une évaluation de l'erreur d'estimation. Un ré-ajustage de capacité basé sur cette erreur est proposé pour améliorer la probabilité de conformité aux contraintes de GoS.

Enfin, la performance de notre approche intégrant les éléments cités ci-dessus est évaluée par comparaison à celles de méthodes connues. L'évaluation est effectuée à l'aide du simulateur sur un réseau de taille réaliste et supportant un nombre significatif de couples origine-destination **(**OD) différents. Des conclusions seront tirées de l'analyse des résultats obtenus.

0.5 Plan de la thèse

La présente thèse comprend quatre chapitres. Le CHAPITRE 1 introduit le problème de la gestion des ressources dans les réseaux dédiés de service. Nous commençons par un survol

des différentes classifications de réseaux dédiés implantés sur Internet. Ensuite, nous discutons du sujet de la gestion du réseau dédié de service avec comme objectif la maximisation du bénéfice. Une revue de littérature des travaux connexes sur le sujet est enfin présentée dans le chapitre.

Au CHAPITRE 2, nous présentons notre proposition d'approche de maximisation du bénéfice par adaptation de capacité du réseau dédié. L'approche de contrôle d'admission et de routage de connexions qui maximise la récompense du réseau (Dziong, 1997), sur laquelle s'appuie la proposition est d'abord résumée. Un cadre économique est ensuite conçu pour englober la maximisation du bénéfice, intégrant les adaptations du routage et de la capacité du réseau. Le chapitre continue avec la description détaillée des modèles proposés d'adaptation qui réalisent la maximisation du bénéfice et le maintien du degré de service. Nous terminons le chapitre avec la présentation des résultats d'analyse qui confirme la validité de l'approche et donne une évaluation de la performance obtenue par l'approche.

Comme l'adaptation de capacité est dirigée par la demande de trafic au réseau, nous proposons au CHAPITRE 3 deux approches d'estimation de la demande, dont la pertinence des valeurs estimées contribuent efficacement à la performance de l'adaptation de capacité. Nous offrons d'abord une revue de méthodes existantes se rapportant à nos approches. Ensuite, nous présentons et discutons de nos méthodes proposées d'estimation de tendance de trafic, fondées sur le lissage exponentiel adaptatif et sur le filtre de Kalman. La performance de ces méthodes est analysée selon des critères définis de stabilité et de réponse au changement de tendance.

Le CHAPITRE 4 présente enfin une évaluation globale de notre proposition de gestion de ressources du réseau dédié de service, intégrant les fonctions d'estimation de trafic, de contrôle d'admission et de routage, et d'adaptation de capacité de liens. L'évaluation montre les résultats obtenus au point de vue du bénéfice de réseau ainsi que du degré de service offert aux connexions des usagers.

La dissertation conclut par un résumé des contributions de la thèse et des indications sur la direction des recherches subséquentes.

CHAPITRE 1

GESTION DES RÉSEAUX DÉDIÉS DE SERVICE

1.1 Introduction

L'Internet représente maintenant un réseau universel de communication de données. Aux applications initiales simples comme le courrier électronique et le transfert de fichiers se sont ajoutées des applications nouvelles et avancées telles que la transmission multipoint, la distribution de contenu, la VoIP, le *Streaming* Multimédia, le jeu interactif, etc. Pour ces dernières qui sont souvent du type temps réel, le caractère *best effort* de l'Internet ne convient plus. Elles demandent une performance améliorée et/ou une QoS de la transmission dans leurs connexions.

Le sujet de la QoS sur Internet a été abordé par plusieurs auteurs, par exemple (Wang, 2001) et (Ferguson et Huston, 1998). Les techniques les plus communément mentionnées pour supporter cette QoS sont IntServ (Braden, Clark et Shenker, 1994), DiffServ (Blake *et al.*, 1998), *Multiprotocol Label Switching* (MPLS) (Rosen, A. Viswanathan et Callon, 2001) et (Guichard, Le Faucheur et Vasseur, 2005), et *Traffic Engineering* (TE) (Awduche *et al.*, 1999) et (Awduche et Jabbari, 2002). Cependant, comme l'Internet est constitué d'une multitude de AS interconnectés mais formant des domaines distincts gérés par différents fournisseurs de service Internet (ISP) indépendamment les uns des autres (Figure 1.1), une garantie de QoS offerte est en général restreinte au domaine de chaque ISP. La limitation est due à la difficulté d'implantation d'un protocole inter-AS de transmission de spécifications QoS combinée au peu d'intérêt de collaboration entre les ISP pour transmettre ces spécifications. Par contre, les réseaux dédiés formés par-dessus l'infrastructure de transport de l'Internet peuvent couvrir de multiples AS et ainsi rendre possible la QoS de bout en bout de leurs connexions.

Dans la suite de ce chapitre, nous présentons une revue de littérature, organisée comme suit : la Section 1.2 aborde les types de réseaux dédiés, la Section 1.3 introduit la gestion des

ressources du réseau dédié de service pour la maximisation de son bénéfice et la Section 1.4 présente la littérature connexe à cette gestion des ressources.

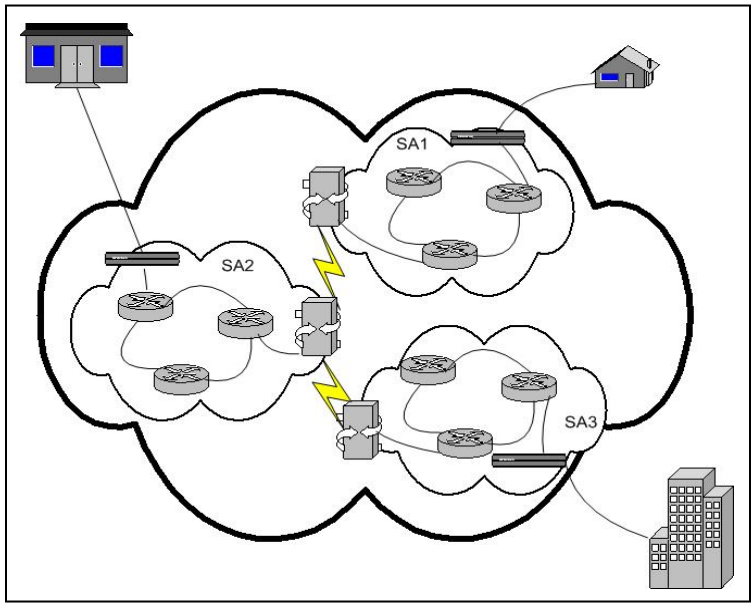

Figure 1.1 Schéma d'un Internet incluant les domaines SA1, SA2 et SA3.

1.2 Les réseaux dédiés

Le réseau dédié est formé par-dessus le réseau Internet, utilisant économiquement l'infrastructure de transport de ce dernier. Un nœud du réseau peut être réalisé par une application dans un nœud Internet existant (y compris un système d'utilisateur) ou par un nouveau nœud physique placé dans l'Internet. Un objectif majeur est l'amélioration de la performance du service sur Internet des applications du réseau dédié. Cela peut se réaliser par la recherche et l'utilisation judicieuse des ressources disponibles dans l'Internet. Couvrant de multiples AS, le réseau contourne aussi les problèmes reliés aux limitations inter-AS. Nous pouvons classifier ces réseaux dans deux grandes catégories fondées sur le genre de service fourni : les réseaux P2P et les réseaux d'opérateurs de service que nous désignerons par SON dans cette thèse. En général, les réseaux P2P sont réalisés pour améliorer l'efficacité d'applications d'usage individuel telles que le partage de contenu ou la transmission

multipoints. De leur côté, les SON servent surtout des entreprises, leur permettant d'obtenir la QoS de bout en bout de leurs connexions.

1.2.1 Réseau dédié P2P

Les premiers réseaux dédiés commerciaux sont apparus dans l'Internet comme réseaux P2P au début des années 2000, dont BitTorrent (BitTorrent Inc., 2001), Napster (Napster Web Team, 2003) et Kazaa (Kazaax) servant le partage et la distribution de contenu musical et multimédia, et Skype (Khamsi, 2004), (Evers, 2003) servant la téléphonie sur Internet. Dès ce moment, une définition du réseau P2P a été donnée dans (Schollmeier, 2002) distinguant son architecture de celle des réseaux clients/serveurs traditionnels. Un bon article de vue d'ensemble sur les réseaux P2P est (Eng Keong *et al.*, 2005), avec une mise à jour fournie par (Hyojin *et al.*, 2008). Ces réseaux servent des usagers individuels et les applications les plus communes incluent le partage de données, la distribution de contenu divers et la transmission multipoints. De nombreuses fonctions sont offertes, par exemple la recherche efficace de *peers* et de contenu, la duplication d'objets pour des transferts plus rapides, la multidiffusion (*multicasting*) efficace (Cui, Li et Nahrstedt, 2004), ainsi que la sécurité du réseau. Des réalisations de diverses qualités pour ces réseaux, comme la robustesse (Zhao *et al.*, 2004) et l'extensibilité (Gao et Steenkiste, 2004), ont été proposées.

Ces réseaux peuvent être classifiés selon leur architecture :

- centralisée, comme Napster. Dans cette architecture, la table d'index des objets du contenu est centralisée dans un serveur auquel sont dirigées toutes les requêtes. Ces réseaux sont en général limités par la capacité du serveur.
- distribuée non structurée, comme BitTorrent, Kazaa et Gnutella (Ripeanu, 2002). La table d'index est distribuée aux nœuds du réseau, ce qui facilite sa croissance. Le réseau ne présente pas de structure topologique particulière et les nœuds sont organisés dans un graphe aléatoire. La localisation d'un objet du contenu se fait bond par bond, avec un nombre de bonds indéterminé.

- distribuée structurée, comme Tapestry (Zhao *et al.*, 2004), Chord (Stoica *et al.*, 2003), CAN (Ratnasamy *et al.*, 2001) et Butterfly (Datar, 2002). Ici, la table d'index est distribuée, avec des clés assignant chaque objet à un nœud désigné. Le plus souvent, l'indexation des objets est réalisée utilisant une table de hachage distribuée (Distributed Haching Table). Cette technique permet de borner par la suite le nombre de bonds requis dans la localisation des objets.

Alternativement, les réseaux peuvent aussi se distinguer par leur appartenance à une génération donnée:

- Première génération : réseaux de base comme Napster et Gnutella. Ces réseaux présentent une performance et une capacité d'expansion réduites.
- Deuxième génération : un grand nombre de réseaux P2P appartiennent à cette génération. On en retrouve structurés en maille (Tapestry), en anneau (Chord et Viceroy (Malkhi, Naor et Ratajczak, 2002)), en Tore (CAN) ou en arbre K-aire (Fry et West, 2004). Des algorithmes plus efficaces de placement et localisation de contenu sont réalisés dans cette génération.
- Troisième génération : comme Butterfly et Low-Diameter (Pandurangan, Raghavan et Upfal, 2003). Cette génération est plus axée sur la robustesse du réseau, par exemple par la duplication d'objets et la recherche de chemins alternatifs de localisation.

De nombreux articles sur divers travaux concernant les réseaux P2P continuent d'apparaitre dans la littérature. La dernière vue d'ensemble peut être trouvée dans (Hyojin *et al.*, 2008).

1.2.2 Réseau SON

Contrastant avec les réseaux P2P qui servent chacun une application spécifique pour des usagers individuels, le réseau SON est un réseau dédié générique permettant le service concurrent de différentes applications. La propriété recherchée par le SON est la capacité d'offrir la QoS de bout en bout aux connexions de ses usagers. Ainsi, il peut bien supporter les services en temps réel tels que VoIP (Amir *et al.*, 2005), *streaming multimédia* ou jeux

interactifs. Aussi, il convient bien aux besoins en qualité de télécommunications d'entreprises. L'intérêt accru du milieu de la recherche pour le SON a suscité la publication d'un numéro spécial (Li *et al.*, 2004) de *IEEE Journal on Selected Areas in Communications* sur le sujet. Un SON est constitué en plaçant des nœuds *overlay* (ON) dans l'Internet et en les reliant avec des liens *overlay* (OL) établis utilisant la connectivité de l'Internet.

Diverses techniques ont été proposées pour supporter la QoS de bout en bout sur Internet. Nous distinguons pour le SON deux modèles d'utilisation du transport internet pour obtenir cette QoS :

- Modèle 1 : le SON achète la capacité d'accès à l'Internet pour les connexions de ses usagers. Le transport des données-usager est effectué par l'Internet *best effort*. Pour acheminer les données dans ce modèle, le SON doit trouver dans l'Internet le meilleur chemin pouvant supporter la QoS. Par exemple, le protocole de routage QRON utilisant le concept de l'*overlay broker* pour trouver des chemins supportant la QoS dans un réseau dédié a été proposé dans (Li et Mohapatra, 2004), et un algorithme de sélection de chemin considérant les délais sur les liens a été présenté dans (Fry et West, 2004). Un autre exemple d'amélioration de la QoS est donné par (Andersen, Snoeren et Balakrishnan, 2003), où les auteurs comparent le routage par sélection du meilleur chemin à celui comportant un dédoublement de transmission sur deux chemins différents. Comme il ne peut contrôler le trafic total dans le chemin utilisé, le SON doit continuellement en surveiller l'état et si une congestion survient, réadapter rapidement le chemin pour préserver la QoS. Le mécanisme de reroutage rapide présenté dans (Amir *et al.*, 2005) pourrait par exemple être utilisé à cette fin. Ce modèle est moins onéreux car il n'y a pas de coût pour le transport. Cependant, bien que la QoS puisse être améliorée de façon significative, la disponibilité d'un chemin avec QoS ne peut être garantie en tout temps.

- Modèle 2 : afin d'assurer la capacité de ses liens pour le trafic de ses connexions, le SON achète des ISP, par le biais des SLA, de la bande passante avec QoS dans les AS

concernés (Duan, Zhang et Hou, 2003). Ainsi, le SON peut fournir une couverture globale avec garantie de QoS en utilisant la bande passante achetée d'une multitude d'AS. Le réseau de Virtela a été construit de façon similaire à ce modèle (Allen, 2002). La quantité de bande passante achetée peut être statique comme dans un dimensionnement traditionnel. Cependant, avec la possibilité envisagée de changements aux termes des SLA, la quantité peut aussi être dynamique (Duan, Zhang et Hou, 2003) pour s'adapter à la demande variable de trafic. Ce modèle, couplé à un contrôle d'admission de connexion, permet une garantie de QoS en tout temps, cependant à un coût supérieur au Modèle 1. On pourrait aussi utiliser un modèle hybride plus économique, où le Modèle 1 serait utilisé quand la QoS est possible avec Internet *best effort,* mais qui serait remplacé par le Modèle 2 quand un chemin avec QoS n'est plus disponible. Dans cette thèse, nous nous concentrons sur ce modèle du SON établi par SLA avec QoS.

Un exemple de réseau SON établi sur l'Internet de la Figure 1.1, comportant quatre nœuds dédiés (ND1 à ND4) et quatre liens dédiés, est illustré à la Figure 1.2.

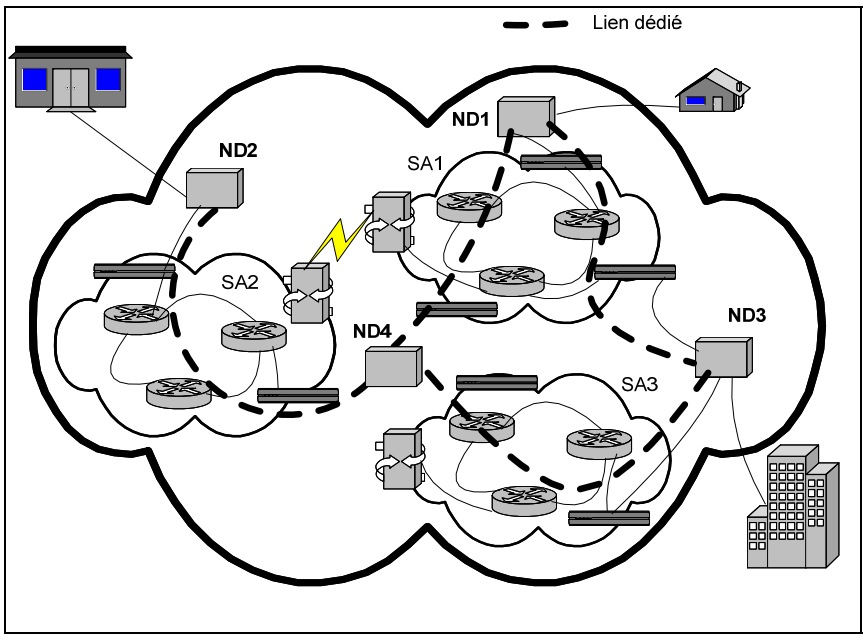

Figure 1.2 Réseau dédié de service établi sur Internet de la Figure 1.1.

1.3 Gestion du SON pour la maximisation du bénéfice d'exploitation

En plus de la QoS fournie aux connexions des usagers, l'autre objectif majeur de l'opérateur est la maximisation du bénéfice de son réseau. Ceci peut être obtenu par un dimensionnement optimal du réseau en fonction de la demande de trafic, couplé à un routage optimal des connexions. Pour garantir la QoS, la bande passante effective requise pour les connexions admises est réservée tout au long des chemisn des connexions. Dans le cas où la capacité résiduelle courante est insuffisante, la connexion est refusée, produisant un taux de blocage qui définit le GoS. Pour garder la satisfaction des clients, la maximisation de bénéfice doit s'effectuer tout en respectant les contraintes de GoS.

Les dimensionnements de réseaux sont fondés traditionnellement soit sur la demande maximale de trafic ou sur la demande moyenne. Comme la demande subit des variations journalières importantes reliées au schéma d'activités des usagers (Thompson, Miller et Wilder, 1997), (Roberts, 2004), un dimensionnement basé sur la demande maximale respectera la GoS, mais la capacité du réseau est gaspillée en périodes de faible demande. Un dimensionnement basé sur la demande moyenne produira une utilisation plus rationnelle de la capacité, mais le réseau subira en périodes de forte demande des congestions entrainant la détérioration de la GoS.

Les problèmes mentionnés peuvent être évités si les récents développements permettant les changements à court terme de capacité du réseau peuvent être appliqués. Cela nécessite la disponibilité de deux capabilités : une estimation efficace de la tendance de la demande de trafic et un algorithme d'adaptation optimale de capacité alimenté par la demande estimée. Nous introduisons dans les paragraphes suivants nos propositions, d'abord d'une approche de maximisation de bénéfice du réseau par adaptation de capacité appuyée par des mesures de trafic, ensuite d'une méthodologie d'estimation de la tendance de trafic qui s'intégrera à l'adaptation de capacité pour efficacement servir les objectifs de profitabilité et de GoS.

Dans le modèle de SON par SLA considéré et présumant que la bande passante des ISP est suffisamment disponible, l'opérateur du SON peut facilement adapter la capacité dédiée aux liens *overlay* en variant les SLA en vigueur ou en ajoutant des SLA couvrant de nouveaux AS. L'objectif de cette thèse est de rechercher un modèle, avec ses mécanismes associés, qui adaptera les SLA aux changements de trafic et de prix des SLA, dans le but de maximiser le bénéfice d'exploitation du SON tout en respectant le niveau requis de GoS. Le bénéfice ici est constitué par le revenu provenant du service des connexions usagers, moins les coûts d'exploitation du réseau. Nous proposons ainsi une approche fondée sur un modèle économique, qui maintient la maximisation du bénéfice en s'appuyant sur des mesures fréquentes de trafic et de performance du réseau. L'élément primordial dans le modèle économique proposé est l'intégration de l'adaptation de la bande passante des liens du SON avec une politique optimale de contrôle d'admission (CAC) et de routage des connexions du réseau fondée sur le processus de décision de Markov (MDP). Cette intégration est réalisée par un algorithme itératif et distribué d'adaptation de liens, où la sensibilité du bénéfice à la dimension d'un lien est obtenue à partir du *shadow price* moyen du lien, ce *shadow price* agissant aussi comme paramètre du routage. Les concepts de CAC et routage dérivés du processus de décision de Markov sont présentés dans le chapitre 5 de (Dziong, 1997). Notons que l'approche proposée ici peut aussi être appliquée à l'adaptation de ressources dans les réseaux virtuels (VN), constitués aussi par allocation logique de bande passante, en général cependant à l'intérieur du domaine d'un seul ISP.

Pour l'estimation de la demande de trafic, nous proposons deux alternatives d'approche fondées sur des concepts différents et procurant différents avantages pour la gestion de capacité. La première est réalisée dans le cadre du lissage exponentiel adapté (Gardner, 2006). Ici, nous adaptons le coefficient de lissage à la tendance estimée de la demande de trafic. Deux nouvelles méthodes sont proposées pour l'estimation de la tendance : l'une est fondée sur l'analyse de la fonction d'autocorrélation du processus d'arrivée du trafic, l'autre sur celle de la fonction de distribution cumulée du processus. L'objectif visé par cette approche est d'apporter une amélioration appréciable, par rapport aux méthodes ES connues, de la réponse de l'estimation aux changements de tendance, tout en préservant la stabilité de

l'estimation dans les périodes de trafic stationnaire. L'intégration de cette approche à l'adaptation de capacité peut procurer un meilleur bénéfice et un meilleur contrôle de GoS.

La motivation pour la seconde proposition d'estimation provient du fait que toute méthode d'estimation ponctuelle est sujette à une certaine erreur. Ainsi, une attribution optimale de capacité fondée uniquement sur une estimation ponctuelle de trafic peut entrainer une probabilité élevée de violation de contraintes GoS. Pour contenir ces violations à un niveau acceptable, nous proposons une estimation par intervalles qui utilise l'estimation de la répartition de l'erreur. Dans ce cas, nous appliquons un filtre de Kalman (Kalman, 1960) où le modèle de trafic est construit à partir de mesures passées. Ainsi avec cette approche, l'adaptation de capacité est effectuée en tenant compte de la variance de l'erreur d'estimation, ce qui procurera une conformité élevée au GoS.

1.4 Travaux connexes

Dans cette section, nous passons en revue la littérature sur les sujets connexes à notre travail. Les domaines de la maximisation de bénéfice du réseau, de l'adaptation de capacité supportée par mesures de trafic et de l'estimation de la demande de trafic ont été recherchés.

1.4.1 Maximisation de bénéfice du réseau

Une approche de CAC et routage avec pour objectif la maximisation de récompense dans le réseau a été présentée au chapitre 5 de (Dziong, 1997). La politique de CAC et routage, désignée par MDPD, est fondée sur la théorie de Décision de Markov qui, pour la simplification de l'approche, a été décomposée aux liens du réseau. Dans cette approche, un *shadow price* dépendant de l'état du lien et représentant le coût dynamique d'admission d'une connexion est calculé pour chaque lien du réseau. La politique utilise alors ces *shadow price* dans sa décision d'admission et de routage des connexions. Notre proposition de maximisation de bénéfice se sert de cette approche de CAC et routage MDPD comme fondation, sur laquelle nous ajoutons les considérations de coût réel des SLA et l'adaptation

de capacité. Nous donnerons un résumé plus détaillé de cette approche MDPD avec la présentation de notre proposition d'adaptation de capacité au CHAPITRE 2.

Plusieurs travaux (Girard, 1990), (Girard et Liau, 1993), (Pióro et Medhi, 2004) et (Lam, Dziong et Mason, 2007) ont traité du dimensionnement en différé de liens du réseau avec pour objectif la maximisation du bénéfice. Les algorithmes proposés peuvent potentiellement être utilisés dans une allocation adaptative de ressources, en les exécutant de façon périodique suite à des mesures de trafic. Cette approche présente cependant trois lacunes principales lorsque comparée à la notre proposée dans cette thèse. Primo, elle implique des méthodes centralisées qui nécessitent un mécanisme de messages de contrôle, en contraste avec notre méthode distribuée aux nœuds du réseau. Secundo, sa complexité est significative due au besoin d'un modèle de performance global. Enfin, la complexité du modèle oblige l'utilisation de mécanismes approximatifs de routage, tels que le *load sharing*, ce qui peut réduire la qualité des résultats.

D'autres approches d'optimisation de coûts dans les décisions de conception de topologie du réseau ont été proposées. (Vieira et Liebeherr, 2004) tient compte des coûts de liens d'accès et de transport du SON dans des conditions déterminées de trafic. Dans (Cohen et Kaempfer, 2000), des problèmes de théorie de graphe impliquant les coûts des liens et des routeurs sont définis, et des solutions heuristiques sont développées pour produire la topologie optimale du réseau. Aussi, quelques travaux sur la gestion et le contrôle des ressources, appliqués aux réseaux virtuels sont présentés dans (Jun et Leon-Garcia, 1998), (Kim *et al.*, 2005), (Liang *et al.*, 2002) et (Cheng *et al.*, 2005).

L'article se rapprochant le plus de notre approche est celui de (Duan, Zhang et Hou, 2003) portant sur le dimensionnement en ligne adaptatif de la bande passante du SON. Ce dimensionnement maximise le revenu net du SON tout en gardant l'utilisation de chaque lien en dedans d'un seuil défini pour assurer la QoS. Ceci entraine un certain sur-dimensionnement du lien. Dans ce modèle, à des intervalles réguliers de temps, la bande passante est répartie en se basant sur le trafic moyen mesuré au précédent intervalle. Le trafic

de chaque paire origine-destination est assigné à un chemin fixé d'avance. (Park et Choi, 2008) a ensuite enrichi ce travail en ajoutant formellement des processus stochastiques de demande de trafic à la formulation du problème. On trouve une solution approximative, fondée sur un algorithme du gradient appliqué à un problème distribué d'optimisation convexe. Notre approche diffère en plusieurs points de celles de Duan et Park, les principaux étant : a) aucun sur-dimensionnement, b) intégration des processus de routage et d'adaptation de capacité, leur procurant ainsi une précision accrue, c) utilisation de procédures adaptatives d'estimation du trafic améliorant l'efficacité de l'adaptation de capacité.

1.4.2 Adaptation de capacité supportée par mesures de trafic

Quelques approches d'adaptation de capacité supportées par mesures de trafic, pour différents types de réseaux, ont été proposées dans la littérature. Une proposée des réseaux Ad-Hoc ajuste la capacité du réseau à la charge de trafic (Seok *et al.*, 2003), où la charge sur chaque nœud est constituée par le total de la bande passante requise par les applications des sessions passant par ce nœud. Une autre pour des réseaux CDMA adapte dynamiquement la capacité de garde des cellules pour satisfaire la contrainte d'abandon de connexion durant le transfert intercellulaire (Wang, Ramjee et Viswanathan, 2005). Les taux d'abandon sont mesurés et leurs déviations du taux cible sont utilisées pour adapter la capacité de garde. Dans cette approche, la vitesse d'adaptation est contrôlée par un facteur appliqué à la déviation mentionnée. Dans (Recker, Ludiger et Geisselhardt, 2003), les variations de trafic dans les chemins et leurs effets sur le délai des paquets sont échantillonnés et lissés exponentiellement dans une fenêtre mobile de taille variable. Une technique de logique *fuzzy* est ensuite appliquée sur ces variations pour en déduire l'allocation minimale de ressources respectant les contraintes de QoS. La taille de la fenêtre est ajustée périodiquement, aussi par logique *fuzzy,* dans le but de capturer les variations importantes de trafic.

1.4.3 Estimation de demande de trafic

Plusieurs techniques d'estimation de demande de trafic ont été proposées dans la littérature. Une parmi les précurseurs a été proposée pour le réseau de téléphone public (Tu, 1994), où l'estimé des demandes de trafic point à point est obtenu par la résolution d'un ensemble d'équations linéaires de mesures de charges et de taux de blocage dans les liens du réseau. Avec la possibilité plus récente d'optimisation de réseau par allocation dynamique de capacité, plusieurs propositions d'estimation de trafic en temps réel ont été avancées. Une méthode simple de mesure et d'estimation supportant l'allocation dans le SON est présentée dans (Duan, Zhang et Hou, 2003). Une méthode de dimensionnement utilisant un filtre ARIMA est proposée dans (Krithikaivasan, Deka et Medhi, 2004), dans (Anjali *et al.*, 2004). Les auteurs présentent un algorithme approximatif de filtrage fondé sur le modèle du processus stochastique de naissance et de mort, et dans (Dasgupta, de Oliveira et Vasseur, 2008) une allocation en ligne de tunnels MPLS supportée par l'estimation par moyenne pondérée de la tendance de trafic est proposée.

Une proposition proche de notre approche d'estimation par lissage exponentiel adaptatif fondé sur la tendance du trafic est donnée dans (Ching, Scholtes et Zhang, 2004). Dans cette proposition, la moyenne de trafic est estimée par une procédure itérative impliquant un lissage exponentiel alimenté par un comptage régulier du trafic. La différence avec notre approche est que nous effectuons l'estimation par lissage adaptatif du trafic dans les liens, tandis que Ching utilise le lissage à paramètre fixe pour estimer le trafic des connexions. Dans les grands réseaux, l'estimation en ligne du trafic de connexions peut s'avérer accablante à cause de la complexité reliée à la grande cardinalité des chemins de connexions. L'estimation au niveau du lien est plus simple, permettant une opération en ligne plus rapide et assurant ainsi une meilleure extensibilité de l'approche. De son côté, le lissage adaptatif permettra, comme nous le démontrerons par la suite, une précision améliorée de l'estimation en ligne du trafic.

Quelques approches fondées sur le filtre de Kalman (Dziong, 1997), (Kolarov, Atai et Hui, 1994), (Anjali, Scoglio et Uhl, 2003), ont été proposées afin d'obtenir des estimations optimales dans divers problèmes de réseautique. Nous donnerons plus de détails sur ces approches avec la présentation de notre proposition d'estimation par filtre de Kalman au CHAPITRE 3. La nouveauté de notre approche réside dans l'intégration de l'estimation du trafic et de celle de la distribution de l'erreur d'estimation, obtenues par le filtre, dans l'adaptation de capacité des liens. Comme démontré par la suite, ceci permet la réalisation concurrente d'un bénéfice de réseau optimal et d'une conformité élevée aux contraintes de GoS.

1.5 Résumé

Dans ce chapitre, nous avons présenté un survol des réseaux dédiés et de la gestion de leurs ressources. Nous avons commencé par la motivation de fournir performance et QoS de bout en bout sur Internet par le réseau. Ces réseaux peuvent être classifiés en deux grandes catégories : les réseaux P2P et les réseaux SON procurant la QoS. Une revue de divers réseaux dédiés connus est donnée.

La présentation est ensuite concentrée sur le problème de gestion du SON pour maximiser son bénéfice. Un aperçu de notre proposition d'adaptation de capacité appuyée par l'estimation de la tendance de trafic est donné. La dernière section du chapitre présente une revue de littérature dans les domaines de la maximisation de bénéfice, de l'adaptation de capacité et de l'estimation de la demande de trafic. Au lieu de recourir à des modèles approximatifs de routage souvent proposés dans la littérature, notre adaptation de capacité est intégrée à l'actuel routage du réseau, ce qui assure plus d'exactitude à la maximisation du bénéfice. Aussi, la réalisation distribuée obtenue par la décomposition de l'adaptation aux liens du réseau réduira significativement la complexité du problème de réseau, permettant une opération en ligne. Enfin, nos approches d'estimation de trafic utilisant la tendance et l'erreur d'estimation permettront une meilleure réponse de l'estimation et une conformité améliorée du réseau aux contraintes de GoS.

CHAPITRE 2

MAXIMISATION DU BÉNÉFICE DU *SON* PAR ADAPTATION DE CAPACITÉ

2.1 Introduction

Dans ce chapitre, nous présentons notre proposition d'approche de gestion de ressources du SON. Le modèle du SON établi par ententes d'achat SLA de bande passante de fournisseurs de service Internet, propriétaires des systèmes autonomes, sera utilisé. La gestion de ressources vise les objectifs principaux suivants :

- Fournir la Qualité de Service de bout en bout aux connexions usagers acceptées dans le réseau, et
- Maximiser le bénéfice d'exploitation du réseau, tout en respectant les contraintes de blocage de connexions (GoS).

La QoS est obtenue en réservant assez de bande passante pour la connexion, tout au long des liens de son chemin. La largeur de bande requise peut être déterminée en se basant sur les besoins en bande effective de la connexion (Pechiar, Perera et Simon, 2002), (Kelly, 1996). Une demande de connexion sera rejetée si la largeur de bande résiduelle d'un lien du chemin est insuffisante. Pour la maximisation du bénéfice, notre proposition sera développée à partir d'une approche de maximisation de récompense de réseau présentée dans (Dziong, 1997), dont nous présentons un résumé à la Section 2.2 ci-après. Nous présentons ensuite notre proposition dans les sections suivantes. La Section 2.3 décrit le cadre économique dans lequel le problème de maximisation de bénéfice est posé et la solution par adaptations de routage, de capacité et de paramètres de récompense est établie. La Section 2.4 présente un modèle exact et un modèle approximatif d'adaptation de capacité, fondés respectivement sur l'approche MDP et MDP décomposé (MDPD), ainsi que le modèle d'adaptation de paramètres de récompense. La Section 2.5 démontre la validité du modèle d'adaptation de capacité MDPD par une analyse comparative avec le modèle exact MDP. La Section 2.6 présente la performance obtenue par le modèle d'adaptation MDPD et la Section 2.7 récapitule le chapitre.

2.2 Approche de CAC et routage pour la maximisation de récompense

Dans cette section, nous donnons un résumé de l'approche de CAC et routage ayant pour objectif la maximisation de récompense, exposée dans (Dziong, 1997). Cette approche est réalisée dans le cadre de la Théorie de Décision de Markov. La solution optimale fondée sur l'état du réseau est identifiée dans cette thèse comme *modèle de CAC et de routage MDP*. Une solution approximative à complexité réduite, obtenue par décomposition du problème aux liens, sera appelée *modèle de CAC et de routage MDPD*. Celles-ci seront transformées pour être intégrée à notre proposition d'adaptation de capacité pour maximisation de bénéfice du SON.

2.2.1 Solution optimale de réseau

Dans le réseau considéré, une connexion de classe j, j=1, 2, …, J, est définie par sa paire d'origine-destination, sa bande passante requise d_j, son taux d'arrivée λ_j, son temps moyen de service $1/\mu_j$ et son paramètre de récompense moyenne $r_j \geq 0$. L'ensemble des différents chemins que peut prendre la connexion est désigné par $\boldsymbol{W}_j$. On présume que chaque connexion admise dans le réseau y contribue une récompense au taux moyen de q_j pour la durée de service de la connexion :

$$q_j = r_j \mu_j . \tag{2.1}$$

L'objectif du CAC et routage MDP est de déterminer la politique optimale π^* qui maximise la récompense moyenne du réseau donnée par :

$$R(\pi) = \sum_{j=1}^{J} r_j \bar{\lambda}_j(\pi), \tag{2.2}$$

où $\bar{\lambda}_j$ représente le taux moyen d'admission des connexions de classe j.

L'état du réseau peut être représenté par une matrice $\mathbf{z} = [z_j^k]$ où z_j^k indique le nombre de connexions de classe *j* admises sur le chemin *k*, $k \in \mathbf{W}_j$. Le taux de récompense total du réseau dans un état **z** est donné par :

$$q(\mathbf{z}) = \sum_{j=1}^{J} \sum_{k \in \mathbf{W}_j} r_j z_j^k \mu_j \,, \tag{2.3}$$

L'état du réseau subit une transition quand la politique π admet une nouvelle connexion ou quand le service d'une connexion termine. Les taux de transition sont respectivement de λ_j et $z_j^k \mu_j$.

Par la théorie MDP, la politique optimale π^* est déterministe et peut être trouvée par un des algorithmes connus de programmation linéaire, d'itération de politique ou d'itération de valeur. Une solution appliquant l'algorithme d'itération de politique est donnée dans (Dziong, 1997).

2.2.2 Décomposition du problème aux liens du réseau

Pour la plupart des réseaux de télécommunications, le CAC et routage MDP fondée sur l'état exact du réseau est intraitable à cause de la très grande cardinalité de l'espace d'état. Pour réduire la complexité du traitement, une approche approximative est proposée, où les processus d'arrivée et de récompense des connexions sont décomposés respectivement en un ensemble de processus séparables d'arrivée et de récompense à chacun des liens du réseau. Cette approche représente le *modèle de CAC et de routage MDPD*.

Avec une hypothèse d'indépendance de liens communément utilisée dans les modèles de performance de réseaux, le processus d'arrivée du réseau peut être décomposé séparément aux liens *s*, *s*=1,2,…, *S*, du réseau. L'état de chaque lien est représenté par $\mathbf{x} = [x_j]$ où x_j indique le nombre de connexions de classe *j* admises dans le lien. Les taux d'arrivée et de départ des connexions au lien sont respectivement indiqués par $\lambda_j^s(\mathbf{x}, \pi)$ et $x_j \mu_j$.

L'évaluation de λ_j^s peut s'effectuer analytiquement par un modèle de performance dirigé par la politique π, ou par estimation à l'aide de mesures dans le réseau.

Pour la décomposition du processus de récompense, le paramètre de récompense de connexion r_j est distribué aux liens du chemin de la connexion, chaque lien étant assigné un paramètre de récompense $r_j^s \geq 0$, avec la contrainte :

$$r_j = \sum_{s \in S^k} r_j^s, \quad r_j^s \geq 0 \quad \forall j, s, \tag{2.4}$$

où S^k est l'ensemble des liens d'un chemin k.

Avec les décompositions, un gain net du lien, $g_j^s(\mathbf{x}, \pi)$, peut être défini comme étant l'espérance d'augmentation de la récompense du lien suite à l'admission d'une connexion de classe j. Avec le taux de récompense du lien dans un état **x** donné par :

$$q(\mathbf{x}) = \sum_{j \in J^s} r_j^s x_j \mu_j, \tag{2.5}$$

où J^s représente l'ensemble des classes de connexion admises au lien s, le gain net peut être obtenu par la résolution de l'ensemble d'équations :

$$R^s(\pi) = q(\mathbf{x}) + \sum_{j \in J^s} \lambda_j^s(\mathbf{x}, \pi) g_j^s(\mathbf{x}, \pi) - \sum_{j \in J^s} x_j \mu_j g_j^s(\mathbf{x} - \delta_j, \pi), \quad \mathbf{x} \in \mathbf{X}^s, \tag{2.6}$$

où R^s est la récompense moyenne du lien et δ_j est un vecteur de dimension J avec 1 dans l'élément j et 0 partout ailleurs. L'algorithme d'itération de valeur (Schweitzer et Federgruen, 1979), (Cavazos-Cadena, 2002), peut être utilisé pour la solution de (2.6). L'application de cet algorithme à temps discret nécessite une uniformisation du temps de transition par un temps moyen τ. La récurrence dans le modèle de lien pour trouver les fonctions de valeur $V_n^s(\mathbf{x}, \pi)$ est alors exprimée par :

$$\begin{aligned} V_n^s(\mathbf{x}, \pi) = {} & q(\mathbf{x})\tau + \sum_{j \in J^s} \lambda_j^s(\mathbf{x}, \pi)\tau \left[V_{n-1}^s(\mathbf{x} + \delta_j, \pi) - V_{n-1}^s(\mathbf{x}, \pi) \right] \\ & + \sum_{j \in J^s} x_j \mu_j \tau \left[V_{n-1}^s(\mathbf{x} - \delta_j, \pi) - V_{n-1}^s(\mathbf{x}, \pi) \right] + V_{n-1}^s(\mathbf{x}, \pi), \quad \mathbf{x} \in \mathbf{X}^s, \end{aligned} \tag{2.7}$$

où *n* est l'indice d'itération. Avec *n* est assez grand, on peut prouver que la différence $V_n^s(\mathbf{x},\pi) - V_{n-1}^s(\mathbf{x},\pi)$ sera aussi près que nécessaire de la récompense moyenne du lien $R^s\tau$. Ayant les fonctions de valeur, le gain net du lien peut être exprimé par :

$$g_j^s(\mathbf{x},\pi) = \lim_{n\to\infty}\left[V_n^s(\mathbf{x}+\delta_j,\pi) - V_n^s(\mathbf{x},\pi)\right]. \tag{2.8}$$

Dans cette approche, un *shadow price* dépendant d'état du lien, $p_j^s(\mathbf{x},\pi)$, est défini, en relation avec le gain net et la récompense de la connexion sur le lien, par :

$$p_j^s(\mathbf{x},\pi) = r_j^s - g_j^s(\mathbf{x},\pi). \tag{2.9}$$

On peut interpréter le *shadow price* comme étant l'espérance du prix de l'admission d'une connexion de classe *j* sur le lien *s* dans l'état ***x***. Ainsi, il ressort de (2.4), (2.8) et (2.9) que le gain net total procuré par l'admission de la connexion de classe *j* sur le chemin *k* est :

$$g_j^k(\mathbf{y},\pi) = r_j - \sum_{s\in S^k} p_j^s(\mathbf{x},\pi), \tag{2.10}$$

où $\mathbf{y} = \{\mathbf{x}^s\}$ indique l'état du réseau dans ce modèle décomposé. À chaque arrivée de connexion de classe *j*, la politique de routage choisit alors parmi tous les chemins possibles celui qui donne le gain net maximal :

$$g_{\max} = \max_{k\in W_j}\left[r_j - \sum_{s\in S^k} p_j^s(\mathbf{x},\pi)\right]. \tag{2.11}$$

Si aucun chemin ne donne un gain positif, la connexion est rejetée.

Une représentation de la décomposition du problème réseau en problèmes de lien est donnée à la Figure 2.1. Pour plus de détails sur cette approche de décomposition, veuillez vous référer à (Dziong, 1997).

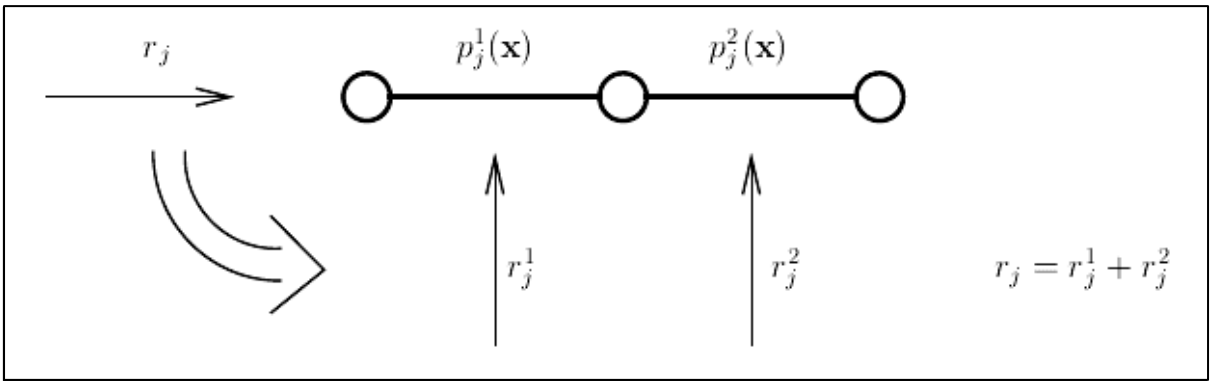

Figure 2.1 Décomposition du problème réseau.
Tirée de Dziong (1997, p. 107).

2.2.2.1 Simplifications du modèle

Quand l'espace d'état $\boldsymbol{X}^s$ est très grand, la résolution de (2.7) peut être affectée par des problèmes numériques ou de durée du traitement. Des simplifications, fondées respectivement sur une agrégation de classes de connexion, des *steady arrival rates* et une modification de structure du processus de lien, sont proposées dans (Dziong, 1997).

Dans la proposition d'adaptation de capacité de cette thèse, nous ferons usage de la première méthode mentionnée. Avec celle-ci, les classes de connexion ayant les mêmes exigences de largeur de bande et temps de service moyen sont agglomérées dans le lien en une classe commune *i*, avec le paramètre de récompense défini par :

$$r_i^s = \frac{\sum_{j \in J_i} r_j^s \overline{\lambda}_j^s}{\sum_{j \in J_i} \overline{\lambda}_j^s}, \tag{2.12}$$

où $\overline{\lambda}_j^s$ est le taux moyen d'admission des connexions de classe *j* au lien *s*. Dans ce cas, le processus aggloméré est statistiquement proche de l'original et on peut présumer que :

$$p_j^s(\mathbf{x}, \pi) \cong p_i^s(\mathbf{x}, \pi), \quad j \in J_i, \tag{2.13}$$

où J_i représente l'ensemble des classes agglomérées dans la classe *i*.

2.2.2.2 Shadow price moyen

Il a aussi été démontré (voir Section 5.2.4 de Dziong, 1997) qu'on peut exprimer la sensibilité de la récompense moyenne du réseau $\overline{R}(\pi)$ par rapport aux taux d'arrivée λ_j des classes de connexions comme une fonction du *shadow price* moyen $\overline{p}^s$. Dans le cas d'un réseau à taux unique ($d_j = 1$ et $\mu_j = 1$ pour tout *j*), où toute connexion est admise quand la capacité résiduelle est suffisante, on peut démontrer que les *shadow price* moyens constituent la solution unique des équations :

$$\overline{p}^s = \frac{\lambda^s}{\overline{\lambda}^s}\left[E(\lambda^s, N^s - 1) - E(\lambda^s, N^s)\right]\sum_{j \in J^s} \overline{\lambda}_j^k r_j^s , \tag{2.14}$$

où $\lambda^s, \overline{\lambda}^s$ sont les taux superposés, respectivement de toutes les connexions offertes et admises au lien *s*, N^s est la capacité du lien en nombre de connexions, $\overline{\lambda}_j^k$ est le taux des connexions classe *j* admises sur le chemin *k*, et :

$$E(\lambda, N) = \frac{\lambda^N / N}{\sum_{x=0}^{N} \left(\lambda^x / x\right)} \tag{2.15}$$

est la formule d'Erlang qui donne la probabilité de blocage lorsque λ Erlangs de trafic sont offerts à un lien de capacité *N* (en présumant un taux Poissonien d'arrivée de connexions).

L'expression de $\overline{p}^s$ (2.14) sera utilisée par la suite dans le développement de notre proposition d'adaptation de capacité pour la maximisation du bénéfice.

2.3 Cadre économique pour la maximisation de bénéfice du SON

Pour le transport dans le SON, nous présumons que les liens reliant les nœuds du réseau sont établis en achetant, par contrats SLA, de la bande passante dans les Systèmes Autonomes sous jacents des fournisseurs ISP. Les termes du SLA spécifient, pour chaque lien *s*, sa capacité N_s, ses paramètres QoS et son coût par unité temps C_s. Nous présumons qu'on peut modifier les SLA à court terme, permettant ainsi à l'opérateur du SON de changer au besoin l'attribution de la capacité de ses liens. Il peut même réduire le coût de la bande passante en

passant par un AS alternatif si celui-ci offre un SLA plus avantageux. Comme dans la Section 2.2.1, une connexion de classe *j* dans le SON est caractérisée par sa paire OD, sa bande passante requise d_j, son taux d'arrivée λ_j, son temps moyen de service $1/\mu_j$ et son paramètre de récompense moyenne r_j. r_j représente, dans ce cadre économique, le revenu moyen, au taux de $q_j = r_j \mu_j$, obtenu par le SON pour chaque connexion admise. Ceci mène à la définition de notre modèle économique, montré à la Figure 2.2, où les flèches en continu et en tirets représentent respectivement les prestations monétaires et de service.

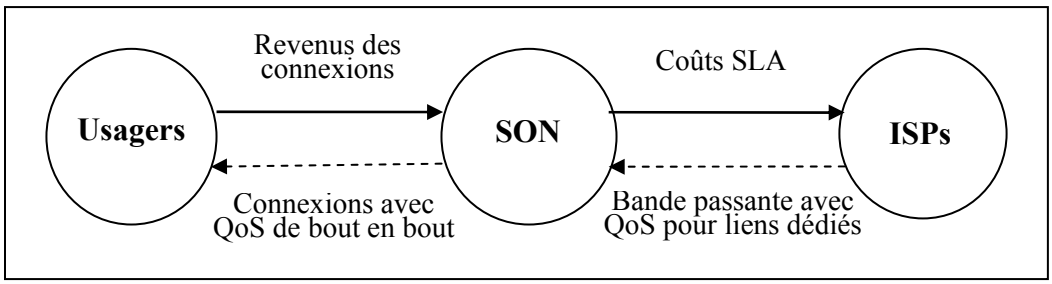

Figure 2.2 Modèle économique du SON.

Dans ce cadre, le taux du bénéfice d'exploitation du SON est exprimé par :

$$P = R(\mathbf{N}) - \sum_s C_s(N_s) = \sum_j \overline{\lambda}_j(\mathbf{N}) r_j - \sum_s C_s(N_s) \tag{2.16}$$

où *R* est le taux de revenu du réseau, $\mathbf{N}=(N_1, N_2, ..., N_S)$, et $\overline{\lambda}_j$ le taux moyen d'admission des connexions de classe *j*. Nous présumons que l'objectif primordial de l'opérateur du SON est la maximisation du bénéfice *P* respectivement à **N**, à la condition que les probabilités de blocage de connexions B_j respectent les contraintes correspondantes B_j^c. Le problème d'optimisation suivant peut alors être posé pour décrire la maximisation du bénéfice:

$$\max_{\mathbf{N}} P = \sum_j \overline{\lambda}_j(\mathbf{N}) r_j - \sum_s C_s(N_s) \tag{2.17}$$

$$\text{s. t.} \quad B_j(\mathbf{N}) = 1 - \frac{\overline{\lambda}_j(\mathbf{N})}{\lambda_j} \le B_j^c, \quad j = 1, 2, ..., J. \tag{2.18}$$

Bien que ce système d'équations représente une formulation classique d'un problème d'attribution de capacité, dans cette thèse nous nous concentrons sur le maintien dans le

temps du profit maximal. Pour cela, nous considérons l'adaptation de capacité des liens aux changements de trafic et de coûts des SLA.

Une possibilité serait d'effectuer périodiquement des redimensionnements du réseau en résolvant (2.17) et (2.18) par des méthodes telles que décrites dans (Girard et Liau, 1993) et (Lam, Dziong et Mason, 2007). Cette approche de redimensionnements comporte cependant plusieurs faiblesses. D'abord, du point de vue opérationnel, elle demande une mise en œuvre centralisée exigeant un mécanisme de collection de données du réseau, qui peut être susceptible aux pannes de réseau. En plus, la complexité des modèles impliqués est considérable due au besoin d'un modèle de performance global. Cette complexité impose le recours à des mécanismes approximatifs de routage, tel que par exemple le *load sharing*, qui peuvent réduire de façon significative l'exactitude de la solution.

Pour pallier aux faiblesses citées, nous proposons une approche distribuée où, de façon périodique, chaque nœud du réseau adapte localement la capacité de ses liens sur la base de mesures locales de trafic. Par cette approche, la complexité est réduite de façon significative. Aussi, en impliquant des mesures en temps réel du trafic au lien qui reflètent la véritable politique de routage, l'exactitude de la maximisation du bénéfice est améliorée. De plus, cette approche intègre la maximisation de bénéfice obtenue par adaptation de capacité avec celle obtenue par la politique de routage par leur usage commun du concept de *shadow price*, comme nous l'expliquerons ci-après.

2.3.1 Adaptation de la politique de CAC et de routage

Notre proposition pour adapter la politique de CAC et de routage au trafic est fondée sur le modèle MDPD réalisant la maximisation de récompense (Dziong, 1997), dont un résumé est donné à la Section 2.2. Pour s'appliquer à la maximisation de bénéfice du SON, le modèle est modifié par l'inclusion des coûts réels des SLA et une méthode de décomposition du paramètre de récompense sensible à ces coûts.

Avec l'inclusion des coûts des SLA dans le modèle économique (Figure 2.2), le taux de bénéfice dépendant d'état $\tilde{q}(\mathbf{z})$ du SON peut s'exprimer par rapport au taux de revenu q(**z**) (2.3) par :

$$\tilde{q}(\mathbf{z}) = q(\mathbf{z}) - \sum_{s=1}^{S} C_s = \sum_{j=1}^{J} \sum_{k \in \mathbf{W}_j} r_j z_j^k \mu_j - \sum_{s=1}^{S} C_s \,, \tag{2.19}$$

et de façon similaire, le taux de bénéfice décomposé aux liens sera pour chaque lien *s*:

$$\tilde{q}(\mathbf{x}) = q(\mathbf{x}) - C_s = \sum_{j \in J^s} r_j^s x_j \mu_j - C_s \,. \tag{2.20}$$

De manière équivalente à la maximisation de récompense, on peut alors calculer les *shadow price* $p_j^s(\mathbf{x}^s)$ (2.9) des liens en substituant $q(\mathbf{x})$ par $\tilde{q}(\mathbf{x})$ dans (2.7) et, à l'arrivée d'une connexion, choisir le chemin de routage procurant le gain maximal (2.11). Le *shadow price* représente le coût dynamique de l'admission de la connexion *j* dans le lien *s*. Il correspond à l'espérance de perte de revenu causée par la probable réjection de futures connexions, suite à l'admission de la connexion courante.

En plus des coûts dynamiques de lien, nous intégrons dans le modèle les coûts statiques d'achat de bande passante C_s, en effectuant une décomposition du paramètre de récompense r_j qui est proportionnelle à ces coûts :

$$r_j^s = r_j \frac{C_s / N_s}{\sum_{o \in S^k} C_o / N_o} \,. \tag{2.21}$$

Cette décomposition attribue une récompense plus élevée aux liens qui ont un coût de bande passante plus élevé. Une répartition du bénéfice aux liens selon leurs coûts est ainsi réalisée.

Les *shadow price* optimaux sont évalués de manière itérative pour aborder la dépendance fonctionnelle des *shadow price* au trafic offert au lien et vice versa. Cette méthode correspond à l'algorithme bien connu d'itération de politique, qui peut aussi être interprété comme une solution par substitutions répétitives sur un ensemble d'équations de point fixe décrivant les fonctions de *shadow price* et de charge de lien :

$$\mathbf{p}^s = f_p(\boldsymbol{\lambda}^s), \quad \boldsymbol{\lambda}^s = f_l(\boldsymbol{\Pi}, \mathbf{N}); \quad s = 1,2,...,S \,, \tag{2.22}$$

où $\mathbf{p}^s = (p_j^s, j = 1,2,...,J)$ est l'ensemble des *shadow price* du lien *s*, $\boldsymbol{\lambda}^s = (\lambda_j^s, j = 1,2,...,J)$ est l'ensemble des taux d'arrivée de connexions (charge) au lien et $\boldsymbol{\Pi} = (\mathbf{p}^s, s = 1,2,...,S)$ est l'ensemble de tous les *shadow price*. (Dziong, 1997) a montré qu'en choisissant bien la période de raffraichissement (mise à jour) de la politique, que nous appelerons t_p, la convergence de la politique est très rapide (deux itérations dans les examples cités). Puisque les taux d'arrivée $\boldsymbol{\lambda}^s$ peuvent être évalués à l'aide de mesures de trafic de liens, cette méthode n'est pas dépendante de l'utilisation d'un modèle de performance global de réseau. En plus, le caractère itératif de l'algorithme implique que les changements dans la tendance mesurée du trafic offert et dans les prix de SLA seront suivis.

2.3.2 Adaptation de capacité

Dans cette section, nous présentons notre proposition de maximisation de bénéfice du SON par adaptation de capacité. Dans ce cas, la condition nécessaire pour une solution optimale est d'avoir le gradient du bénéfice de réseau par rapport aux capacités de lien égal au vecteur zéro :

$$\nabla \mathbf{P}(\mathbf{N}) = \left(\frac{\partial P}{\partial N_1}, \frac{\partial P}{\partial N_2}, ..., \frac{\partial P}{\partial N_S} \right) = \mathbf{0}. \tag{2.23}$$

Le calcul exact d'un tel gradient est une opération complexe à cause de la cardinalité de l'état du réseau. Pour réduire la complexité, nous utilisons la décomposition du processus de récompense du réseau (Section 2.2.2) qui donne

$$R = \sum_s R_s \ , \tag{2.24}$$

où R_s est le taux de revenu du lien *s*. Utilisant (2.24) pour substituer *R* dans (2.16) et effectuant la différenciation de P donne :

$$\frac{\partial P}{\partial N_s} = \frac{\partial R}{\partial N_s} - \frac{\partial C_s}{\partial N_s} = \frac{\partial R_s}{\partial N_s} + \sum_{o=(1,2,...,S)\setminus s} \frac{\partial R_o}{\partial N_s} - \frac{\partial C_s}{\partial N_s}. \tag{2.25}$$

Pour des réseaux typiques, nous présumons que la sensibilité directe du revenu d'un lien à la capacité du lien, $\frac{\partial R_s}{\partial N_s}$, est beaucoup plus significative que sa sensibilité indirecte, $\frac{\partial R_o}{\partial N_s}$ où $o \neq s$. Alors, la condition (2.25) peut être approximée par :

$$\frac{\partial P}{\partial N_s} \cong \frac{\partial R_s}{\partial N_s} - \frac{\partial C_s}{\partial N_s} = \frac{\partial P_s}{\partial N_s} = 0, \quad s = 1,2,...,S\,, \tag{2.26}$$

où P_s est le bénéfice du lien *s*, exprimé par :

$$P_s = R_s(N_s) - C_s(N_s) = \sum\nolimits_j \overline{\lambda}_j^s(N_s) r_j^s - C_s(N_s)\,, \tag{2.27}$$

où $\overline{\lambda}_j^s$ est le taux des connexions de classe *j* admises au lien *s*. La condition (2.26) exprime que le bénéfice du réseau sera maximisé quand les bénéfices de tous les liens, chacun considéré séparément des autres, sont maximisés.

Dziong a aussi montré (Dziong, 1997, page 235) que le *shadow price* moyen correspond au différentiel de la récompense du lien, et ainsi il peut être utilisé pour approximer la dérivée de la récompense du lien par rapport à sa capacité :

$$\overline{p}^s = R_s(N_s) - R_s(N_s - 1) \cong \frac{\partial R_s}{\partial N_s}\,. \tag{2.28}$$

Mettant (2.28) dans (2.26) donne la condition suivante d'optimalité du bénéfice de lien qui dirigera notre adaptation de capacité :

$$\frac{\partial P_s}{\partial N_s} \cong \overline{p}^s - \frac{\partial C_s}{\partial N_s} = 0\,. \tag{2.29}$$

Les capacités optimales de lien N_s sont ainsi déterminées, sur la base de (2.29), en utilisant la moyenne des *shadow price* obtenus par la fonction de *shadow price* $f_p(\boldsymbol{\lambda}^s)$ (2.22). Nous désignons cette fonction par fonction de capacité de lien $N_s = f_c(f_p(\boldsymbol{\lambda}^s))$. Cette fonction et la fonction de charge de lien forment un ensemble d'équations similaire aux équations de point fixe :

$$N_s = f_c(f_p(\boldsymbol{\lambda}^s)), \quad \boldsymbol{\lambda}^s = f_l(\boldsymbol{\Pi}, \mathbf{N})\,; \quad s = 1,2,...,S\,. \tag{2.30}$$

Ces équations peuvent alors être résolues de façon itérative par substitutions répétitives pour aborder l'interdépendance des capacités de liens et des taux d'arrivée aux liens.

Notons que dans le cas d'un réseau à taux unique, le *shadow price* moyen utilisé dans la détermination de N_s peut aussi être obtenu par (2.14) qui représente également une fonction des taux d'arrivée $\boldsymbol{\lambda}^s$. Dans ce cas, nous obtenons également un ensemble d'équations formé d'une fonction de capacité et d'une fonction de charge interdépendantes qui peut être résolu de la même manière que pour (2.30).

La caractéristique importante de cette approche est que les deux moyens utilisés pour la maximisation de bénéfice, routage MDPD et adaptation de capacité, sont intégrés à travers le concept du *shadow price*. En effet, l'adaptation de capacité est fondée sur les valeurs de *shadow price* qui sont elles mêmes calculées, suite à des mesures, et utilisées dans le routage. Par conséquent, l'adaptation de capacité proposée peut être distribuée dans des problèmes de liens séparés et, avec les *shadow price* disponibles du routage, il n'est plus requis d'utiliser un modèle de performance de réseau. En plus, l'utilisation commune du concept de *shadow price* dans les deux algorithmes de routage et d'adaptation de capacité procure une consistance dans leur objectif commun. Dans cette approche, la période de rafraichissement de la capacité, t_c, doit évidemment être plus grande que celle du *shadow price*, t_p, puisque la fonction de capacité utilise le résultat de la fonction de *shadow price* qui doit avoir accompli sa convergence à ce moment.

2.3.3 Adaptation de paramètre de récompense

La maximisation du bénéfice présentée jusqu'à maintenant n'a pas tenu compte explicitement des contraintes de blocage (2.18). Bien qu'avec des prix raisonnables de service SON et de SLA, on peut présumer qu'en général cette solution pourra satisfaire aux contraintes, celles ci peuvent cependant être excédées dans certains scénarios. Comme solution à ce problème, nous proposons dans cette section une approche qui est aussi intégrée dans notre cadre économique.

On peut montrer facilement en effet qu'en augmentant le paramètre de récompense d'une classe de connexion, les mécanismes de routage et d'adaptation de capacité pour la maximisation de bénéfice donneront plus de ressources à cette classe, et que par conséquent sa probabilité de blocage sera réduite. Ceci indique que si la contrainte de blocage d'une classe *j* est excédée ($B_j > B_j^c$), le bénéfice de connexions de cette classe est sous-évalué et son paramètre de récompense r_j devrait être augmenté.

Ci-après, nous considérons donc une approche dans laquelle les paramètres de récompense peuvent être adaptés ($r_j \rightarrow \hat{r}_j$). Nous présumons toutefois, pour des raisons de marché, que l'augmentation de revenu offert total, $\Delta\hat{R}$, devrait être maintenu au minimum. Pour ce faire, nous envisageons aussi la diminution des paramètres de récompense des classes de connexion dont $B_j < B_j^c$. Ceci mène alors à la formulation suivante du problème :

$$\max_{\mathbf{N}} P = \sum_j \overline{\lambda}_j(\mathbf{N}, \hat{r}_j)\hat{r}_j - \sum_s C_s(N_s) - \min_{\hat{\mathbf{r}}} \sum_j \lambda_j(\hat{r}_j - r_j), \tag{2.31}$$

$$t.q. \quad B_j = 1 - \frac{\overline{\lambda}_j(\mathbf{N}, \hat{r}_j)}{\lambda_j} \le B_j^c, \quad j = 1, 2, ..., J, \tag{2.32}$$

$$\Delta\hat{R} = \sum_j \lambda_j(\hat{r}_j - r_j) \ge 0. \tag{2.33}$$

Pour résoudre ce problème, on peut formuler un ensemble d'équations implicites formé par une fonction de paramètre de récompense $\hat{r}_j = f_r(\mathbf{B}, \mathbf{B^c})$, où $\mathbf{B} = (B_j, j = 1, 2, ..., J)$ et $\mathbf{B}^C = (B_j^C, j = 1, 2, ..., J)$, et une fonction de blocage $B_j = f_b(\hat{\mathbf{r}})$. La première équation ajuste le paramètre de récompense en fonction de la différence entre les valeurs des blocages et de leurs contraintes, tout en tenant compte de la minimisation de (2.33). La deuxième équation calcule les valeurs des blocages en fonction des paramètres de récompenses ajustés. Puisque les paramètres de récompense influencent les capacités des liens ainsi que la politique de routage, la fonction de blocage peut aussi s'exprimer comme une fonction de l'ensemble des fonctions de capacité de lien $\mathbf{f_c}$ et de l'ensemble des fonctions de charge de lien $\mathbf{f_l}$. $\mathbf{f_c}$ dépend de l'ensemble des fonctions de *shadow price* $\mathbf{f_p}$ qui lui-même dépend des fonctions de charge de lien. Ceci mène finalement au système :

$$\hat{r}_j = f_r(\mathbf{B}, \mathbf{B}^{\mathbf{c}}), \tag{2.34}$$

$$B_j = f_b\left(\mathbf{f_c}(\mathbf{f_p}(\mathbf{f_l}(\mathbf{\Pi}, \mathbf{N}, \hat{\mathbf{r}}))), \mathbf{f_l}(\mathbf{\Pi}, \mathbf{N}, \hat{\mathbf{r}})\right). \tag{2.35}$$

Ce système d'équations peut être résolu de manière itérative par substitutions répétitives, avec une période de rafraichissement appelée t_r, jusqu'à la convergence des paramètres de récompense. L'équation (2.35) indique que la solution de B_j requiert les solutions convergées de $\mathbf{f_c}$ (2.30) et $\mathbf{f_p}$ (2.22). Ainsi, la relation entre les trois périodes de rafraichissement impliquées devra être $t_p < t_c < t_r$, ou même $t_p < t_c << t_r$ si l'évaluation de la fonction de blocage est effectuée par mesures, puisque les statistiques de blocage ont besoin d'un temps plus long. Bien qu'en général, si les B_j sont mesurés, la solution de ce système d'équations peut être distribuée dans les nœuds originaires, notons qu'un certain niveau d'échanges d'information entre les nœuds est requis pour pouvoir tenir compte de $\Delta\hat{R}$ (2.33). Nous donnerons plus de détails de cette procédure à la Section 2.4.

Cette adaptation de paramètre de récompense mène à une solution qui peut être utilisée de deux manières. La première établit que les nouvelles valeurs des paramètres de récompense soient utilisées pour ajuster les prix de service du SON afin de refléter les montants requis de ressources correspondants à chaque classe de connexion. La deuxième manière tient compte du fait que les prix de service doivent aussi dépendre de la compétition dans le marché, et qu'ainsi leurs ajustements peuvent être limités à cause de la possibilité de pertes de clients. Dans ce dernier cas, les nouveaux paramètres de récompense seront utilisés par l'opérateur du SON comme l'un des nombreux facteurs qui détermineront les prix de service.

Dans les approches présentées dans cette section et la Section 2.3.2, nous avions présumé l'existence d'une solution unique dans chaque cas ainsi que la convergence des algorithmes. En général, il est difficile d'étudier analytiquement ces aspects à cause de la complexité des modèles impliqués et de la multitude de possibilités de scénario de trafic. Bien que pour des problèmes similaires de dimensionnement de réseau, (Dziong, 1997) avait montré l'unicité de la solution, dans quelques scénarios de trafic spécifiques, en se basant sur des conditions Jacobiennes, ce résultat ne peut pas être élargi au cas général. Par conséquent, afin de valider

les modèles d'adaptation proposés, nous effectuons une analyse numérique extensive, utilisant également des plateformes analytique et de simulation. Cette analyse est présentée dans les Sections 2.5 et 2.6.

2.4 Modèles d'adaptation de capacité

Comme la solution obtenue avec notre modèle distribué d'adaptation de capacité, fondé sur les *shadow price* moyens, est approximative (Section 2.3.2), nous devons valider l'utilisation du modèle en comparant sa performance et sa convergence avec celles d'un modèle exact d'adaptation de capacité fondé sur le modèle exact de *CAC et routage MDP* (Section 2.2.1). Nous désignons respectivement ces modèles approximatif et exact d'adaptation par modèles *d'adaptation de capacité MDPD et MDP*. Comme dans cette section, nous nous concentrons seulement sur l'aspect de l'adaptation de capacité, les paramètres de trafic des liens, mesurés dans le modèle MDPD, ainsi que la performance dans les deux modèles sont évalués ici utilisant un modèle de performance analytique exact. Les modèles d'adaptation de capacité sont détaillés dans les Sections 2.4.1 et 2.4.2. Nous présentons aussi, à la Section 2.4.3, la réalisation du modèle d'adaptation de paramètre de récompense pour la satisfaction des contraintes de blocage. Les résultats numériques obtenus sont analysés à la Section 2.5.

Dans la suite, nous présumons que les taux d'arrivée des connexions λ_j et les paramètres de récompense r_j sont donnés. Bien que les concepts proposés ici soient applicables aux services multi-classes ayant des spécifications différentes de largeur de bande, nous nous limitons dans la suite de cette thèse, aux réseaux à connexions homogènes où toutes les classes ont la même largeur de bande requise et le même temps moyen de service. Nous présumons aussi un modèle de prix de SLA linéaire où le coût de SLA d'un lien est proportionnel à sa capacité, $C_s=c_sN_s$, où c_s représente le coût unitaire de largeur de bande du lien.

2.4.1 Modèle d'adaptation de capacité MDP

Dans le modèle d'adaptation de capacité MDP, la condition nécessaire d'optimalité du bénéfice de réseau (2.23) est accomplie en utilisant un algorithme itératif de minimisation du gradient comme celle décrite dans (Pióro et Medhi, 2004). L'algorithme effectue successivement des projections du gradient du bénéfice $\nabla \mathbf{P}(\mathbf{N})$ pour le converger au vecteur $\mathbf{0}$. À chaque itération, un facteur de pas (*step-size*) $\overline{\tau}$ redimensionne le changement des capacités projetées $\Delta \mathbf{N}=(\Delta N_1, \Delta N_2, \ldots, \Delta N_S)$ pour accélérer la convergence. Une méthode, telle que par exemple les approximations linéaires successives de Newton, peut être utilisée pour approximer à chaque itération la solution de $\frac{\partial P}{\partial N_s} = 0$ et trouver ΔN_s :

$$\Delta N_s = -\frac{\partial P / \partial N_s}{\partial^2 P / \partial N_s^2}. \tag{2.36}$$

Soient $\mathbf{N}^n$ et $P(\mathbf{N}^n)$, respectivement, l'ensemble des capacités des liens et le bénéfice de réseau à l'itération n, et soit $\boldsymbol{\delta}_s$ le vecteur de dimension S avec l'élément s égal à 1 et tous les autres éléments égaux à 0. Les première et seconde dérivées du bénéfice de réseau par rapport à la capacité du lien s, $\frac{\partial P}{\partial N_s}$ et $\frac{\partial^2 P}{\partial N_s^2}$, peuvent être approximée comme suit :

$$\frac{\partial P}{\partial N_s} \cong P(\mathbf{N}^n + \boldsymbol{\delta}_\mathbf{s}) - P(\mathbf{N}^n), \tag{2.37}$$

$$\begin{aligned} \frac{\partial^2 P}{\partial N_s^2} &\cong P(\mathbf{N}^n + \mathbf{2}\boldsymbol{\delta}_\mathbf{s}) - P(\mathbf{N}^n + \boldsymbol{\delta}_\mathbf{s}) - \left[P(\mathbf{N}^n + \boldsymbol{\delta}_\mathbf{s}) - P(\mathbf{N}^n)\right] \\ &= P(\mathbf{N}^n + \mathbf{2}\boldsymbol{\delta}_\mathbf{s}) - 2P(\mathbf{N}^n + \boldsymbol{\delta}_\mathbf{s}) + P(\mathbf{N}^n). \end{aligned} \tag{2.38}$$

Utilisant (2.37) et (2.38) dans (2.36) donne :

$$\Delta N_s = -\frac{P(\mathbf{N}^n + \boldsymbol{\delta}_\mathbf{s}) - P(\mathbf{N}^n)}{P(\mathbf{N}^n + 2\boldsymbol{\delta}_\mathbf{s}) - 2P(\mathbf{N}^n + \boldsymbol{\delta}_\mathbf{s}) + P(\mathbf{N}^n)}. \tag{2.39}$$

Dans ce modèle itératif, pour calculer le taux de revenu moyen du réseau, $R(\mathbf{N})$, dirigé par une politique de routage optimale (Section 2.2.1), nous utilisons un modèle exact de

performance fondé sur l'état complet du réseau. Ce modèle de performance est fondé sur l'algorithme d'itération de valeurs MDP (Schweitzer et Federgruen, 1979), (Cavazos-Cadena, 2002) qui, pour les conditions données de taux d'arrivée, de paramètres de récompense et de capacités de lien, détermine en même temps la politique de routage optimale et le taux de revenu $R(\mathbf{N})$ (Dziong, 1997). Mettant le résultat de $R(\mathbf{N})$ dans (2.16) produit le taux de bénéfice maximisé du réseau $P(\mathbf{N})$.

Appliquant le modèle ci-dessus, l'algorithme d'adaptation suivant est exécuté pour obtenir les capacités optimales de lien, à une précision relative ε spécifiée. Dans l'algorithme, tous les bénéfices $P(\mathbf{N})$ sont évalués utilisant le modèle exact de performance réseau et (2.16).

Algorithme d'adaptation de capacité MDP

0. Initialiser $n = 0$ et $\mathbf{N}^0$ à un vecteur initial de capacités de lien de valeurs arbitraires. Évaluer $P(\mathbf{N}^0)$.
1. Pour chaque lien s, évaluer $P(\mathbf{N}^n + \boldsymbol{\delta}_s)$ et $P(\mathbf{N}^n + 2\boldsymbol{\delta}_s)$, ensuite ΔN_s par (2.39). Ceci procure le vecteur $\Delta\mathbf{N}=(\Delta N_1, \Delta N_2, \ldots, \Delta N_S)$.
2. Effectuer une recherche unidimensionnelle pour trouver le pas scalaire $\bar{\tau}$ tel que $P(\mathbf{N}^n + \bar{\tau}\Delta\mathbf{N}) = \max_{\tau} P(\mathbf{N}^n + \tau\Delta\mathbf{N})$.
3. Assigner $\mathbf{N}^{n+1} = \mathbf{N}^n + \bar{\tau}\Delta\mathbf{N}$. Si $\left|P(\mathbf{N}^{n+1}) - P(\mathbf{N}^n)\right| \le \varepsilon\, P(\mathbf{N}^n)$, alors terminer; sinon assigner $n = n+1$ et retourner à l'étape 1.

2.4.2 Modèle d'adaptation de capacité MDPD

Nous débutons cette section par une description détaillée de la réalisation, fondée sur des mesures de trafic, du modèle d'adaptation de capacité MDPD introduit à la Section 2.3.2. Dans ce modèle, les capacités optimales des liens sont déterminées en résolvant de manière itérative les fonctions de capacité et de charge de lien (2.30). La fonction de charge de lien $f_l(\mathbf{\Pi}, \mathbf{N})$ évalue les taux d'arrivée aux liens, λ_j^s, et la fonction de capacité de lien

$f_c(f_p(\boldsymbol{\lambda}^s))$ trouve les capacités optimales des liens N_s en réalisant la condition d'optimalité (2.29) fondée sur le *shadow price* moyen.

La fonction de charge est réalisée en mesurant les taux de trafic admis au lien $\overline{\lambda}_j^s$ et en appliquant la formule d'Erlang B de probabilité de blocage :

$$\lambda_j^s = f_l(\boldsymbol{\Pi}, \mathbf{N}) = \frac{\overline{\lambda}_j^s}{1 - E_b(\sum_j \lambda_j^s, N_s)} . \tag{2.40}$$

Pour un trafic admis de lien $\overline{\lambda}^s = \sum_j \overline{\lambda}_j^s$ et une capacité N_s donnés, le trafic offert au lien $\lambda^s = \sum_j \lambda_j^s$ et la probabilité de blocage $E_b(\sum_j \lambda_j^s, N_s)$ peuvent être déterminés, par exemple par l'une des méthodes de recherche linéaire comme la recherche de bissection.

Les *shadow price* dépendant d'état, paramètres la fonction de capacité, sont obtenus par la fonction de *shadow price* de liens $f_p(\boldsymbol{\lambda}^s)$ (2.22). Pour simplifier la réalisation du modèle, nous utilisons l'agrégation de classes de connexion (Section 2.2.2.1) tirée de (Dziong, 1997), où les classes ayant les mêmes exigence de largeur de bande et temps de service moyen sont agglomérées en une classe avec un paramètre de récompense (2.12). Dans le cas considéré de connexions homogènes, toutes les classes sont agglomérées en une classe unique avec comme paramètre de récompense :

$$r^s = \sum_j \overline{\lambda}_j^s r_j^s \Big/ \overline{\lambda}^s . \tag{2.41}$$

On peut dans ce cas obtenir efficacement les *shadow price* des liens en appliquant une solution récursive présentée dans (Dziong, 1997) qui en même temps produit le taux maximisé de revenu de lien R_s. R_s est utilisé dans (2.27) pour le calcul du taux de bénéfice de lien P_s. Les *shadow price* moyens $\overline{p}^s$ peuvent alors être obtenus, en effectuant la moyenne dans le temps de $p^s(\mathbf{x}^s)$.

Alternativement dans ce cas de connexions homogènes, comme mentionné à la Section 2.3.2, les *shadow price* moyens peuvent aussi être déterminés par la formule (2.14) tirée de (Dziong, 1997). En intégrant la définition de r^s (2.41) dans (2.14), nous obtiendrons donc :

$$\overline{p}^{\,s} = (E_b(\lambda^s, N_s - 1) - E_b(\lambda^s, N_s))\lambda^s r^s . \tag{2.42}$$

Comme mentionné au début de cette Section 2.4, dans la réalisation analytique de la fonction de charge du modèle d'adaptation de capacité MDPD, utilisée pour l'analyse comparative avec le modèle MDP, nous substituons les mesures de taux admis de trafic $\overline{\lambda}_j^s$ par les valeurs obtenues à partir du modèle exact de performance réseau décrit à la Section 2.4.1. La politique de routage MDP est appliquée dans ce modèle de performance.

La réalisation analytique de la fonction de capacité du modèle MDPD demeure la même que celle fondée sur les mesures. En présumant un prix de SLA linéaire, la condition d'optimalité de bénéfice d'un lien *s* (2.26) devient :

$$\overline{p}^{\,s}(N_s) - c_s = 0 , \tag{2.43}$$

et substituant $\overline{p}^{\,s}$ dans (2.43) par (2.42) donne :

$$E_b(\lambda^s, N_s - 1) - E_b(\lambda^s, N_s) = c^s / \lambda^s r^s . \tag{2.44}$$

Une procédure itérative peut alors être utilisée pour converger N_s à la réalisation de (2.44). À chaque itération, la capacité optimale du lien est trouvée en effectuant une recherche linéaire d'un entier N_s qui réalise de plus près (2.44).

Soient $\mathbf{N}^{\mathrm{n}}$ et $\boldsymbol{\lambda}(\mathbf{N}^n) = (\lambda_j^s : j = 1,2,...,J; s = 1,2,...,S)$ respectivement les capacités et les taux d'arrivée évalués des liens à l'itération *n*. Alors l'algorithme suivant, appliquant le modèle décrit ci-dessus, est exécuté pour atteindre les capacités optimales de lien, à une précision relative ε spécifiée.

Algorithme d'adaptation de capacité MDPD

0. Initialiser n=0 et $\mathbf{N}^0$ à un vecteur initial de capacités de lien de valeurs arbitraires.
   Pour chaque lien s : évaluer $\boldsymbol{\lambda}^s(\mathbf{N}^0)$ (2.30), puis évaluer $P_s(N_s^0)$ (2.27).
1. Pour chaque lien, calculer sa nouvelle capacité N_s^{n+1} en se basant sur (2.44).
2. Pour chaque lien, évaluer $\boldsymbol{\lambda}^s(\mathbf{N}^{n+1})$ (2.30).
3. Pour chaque lien, évaluer $P_s(N_s^{n+1})$ (2.27).
4. Si $P_s(N_s^{n+1}) - P_s(N_s^n) \le \varepsilon P_s(N_s^n)$ pour tous les liens, alors terminer;
   sinon assigner $n = n+1$ et retourner à l'étape 1.

2.4.3 Modèle d'adaptation de paramètre de récompense

Comme mentionné à la Section 2.3.3, quand des contraintes de blocage de connexions ne sont pas respectées, les paramètres de récompense correspondants devraient être augmentés, tout en minimisant l'augmentation du revenu offert total. Pour réaliser cet objectif, nous proposons une approche approximative qui peut s'effectuer en deux phases.

Dans la première phase, nous considérons seulement la probabilité de blocage moyenne du réseau B_T et sa contrainte B_T^c, définies par :

$$B_T = \frac{\sum_j \lambda_j B_j}{\sum_j \lambda_j} \le B_T^c = \frac{\sum_j \lambda_j B_j^c}{\sum_j \lambda_j}. \qquad (2.45)$$

Si la condition (2.45) n'est pas respectée, nous augmentons les valeurs de tous les paramètres de récompense en appliquant un multiplicateur commun, $\gamma > 1$, jusqu'à l'égalité dans la condition ($B_T = B_T^c$). Avec l'égalité, nous présumons que la contrainte (2.33) du problème formulé (Section 2.3.3) est satisfaite et minimisée. Les nouveaux paramètres de récompense sont ainsi produits :

$$\hat{r}_j = \gamma r_j,\ j = 1,2,...,J\ . \qquad (2.46)$$

Dans la deuxième phase, nous vérifions le respect des contraintes de blocage de chaque classe de connexion. Alors, nous augmentons les paramètres de récompense des classes dont la contrainte n'est pas satisfaite, et en même temps nous réduisons ceux des classes dont la probabilité de blocage est inférieure à la contrainte correspondante, de manière à préserver le revenu offert total du réseau.

Notons que dans la première phase, l'adaptation de capacité est fortement affectée par l'adaptation de paramètres de récompense, à cause de l'augmentation du revenu offert total provenant de tous les liens. De son côté, la politique de routage n'est pas significativement affectée. On peut en effet montrer que, avec toutes autres conditions inchangées, le routage MDPD est sensible seulement aux changements relatifs, entre les classes, des paramètres de récompense. Dans la deuxième phase, la situation est inversée, puisque les changements relatifs impliqués des paramètres de récompense affecteront grandement la politique de routage, tandis que les capacités de lien ne seront pas beaucoup affectées car le revenu offert total reste en moyenne inchangé. Comme dans cette thèse nous nous concentrons sur l'adaptation de capacité, nous y limitons donc la description détaillée à la première phase de l'adaptation de paramètre de récompense.

Soit r_T le paramètre de récompense moyen du réseau, représentant le revenu moyen provenant de l'admission d'une connexion au réseau :

$$r_T = \sum_j \overline{\lambda}_j r_j \Big/ \sum_j \overline{\lambda}_j \ . \tag{2.47}$$

Pour trouver le multiplicateur γ accomplissant l'égalité $B_T = B_T^c$, nous appliquons les itérations de Newton. À chaque itération *n*, la nouvelle valeur de $\hat{r}_T$ sera :

$$\hat{r}_T^{n+1} = \hat{r}_T^n + \frac{B_T^c - B_T^n}{(\partial B_T^n / \partial \hat{r}_T^n)}. \tag{2.48}$$

Pour simplifier les calculs, nous utilisons le modèle de performance à un moment, fondé sur la présomption d'indépendance de liens, et la formule Erlang B, $E_b(\lambda^s, N_s)$. La probabilité de blocage d'une connexion de classe *j* est ainsi approximée par :

$$B_j = \Pi_{k \in W_j} B^k = \Pi_{k \in W_j} \left[1 - \Pi_{s \in S^k} (1 - B^s)\right], \tag{2.49}$$

où B^k est la probabilité de blocage sur le chemin k et B^s est celle du lien s. La probabilité de blocage moyenne du réseau en fonction des probabilités de blocage des liens s'obtient alors en utilisant (2.49) dans (2.45), ce qui permet de déterminer la dérivée $\partial B_T^n / \partial \hat{r}_T^n$ à partir des dérivées de lien $\partial B_s / \partial r_s$. Chacune de ces dernières est exprimée approximativement par :

$$\frac{\partial B^s}{\partial r^s} = \frac{\partial B^s}{\partial N_s}\frac{\partial N_s}{\partial r^s} \cong \left[E_b(\lambda^s, N_s) - E_b(\lambda^s, N_s - 1)\right]\frac{\partial N_s}{\partial r^s}. \tag{2.50}$$

De (2.44) qui réalise la fonction de capacité de lien dépendante du paramètre de récompense, r^s peut être déduit :

$$r^s = \frac{c_s}{\lambda^s\left[E_b(\lambda^s, N_s - 1) - E_b(\lambda^s, N_s)\right]}, \tag{2.51}$$

ce qui permet l'approximation de sa dérivée par rapport à la capacité du lien :

$$\begin{aligned}\frac{\partial r^s}{\partial N_s} \cong r^s(N_s) - r^s(N_s - 1) &= r^s - \frac{c_s}{\lambda^s\left[E_b(\lambda^s, N_s - 2) - E_b(\lambda^s, N_s - 1)\right]} \\ &= r^s\left[1 - \frac{E_b(\lambda^s, N_s - 1) - E_b(\lambda^s, N_s)}{E_b(\lambda^s, N_s - 2) - E_b(\lambda^s, N_s - 1)}\right].\end{aligned} \tag{2.52}$$

L'utilisation de (2.52) dans (2.50) mène alors à l'expression suivante de la dérivée de la probabilité de blocage du lien par rapport à son paramètre de récompense :

$$\begin{aligned}\frac{\partial B^s}{\partial r^s} \cong &\left[E_b(\lambda^s, N_s) - E_b(\lambda^s, N_s - 1)\right] * \\ &\frac{\left[E_b(\lambda^s, N_s - 2) - E_b(\lambda^s, N_s - 1)\right]}{r^s\left[E_b(\lambda^s, N_s - 2) - 2E_b(\lambda^s, N_s - 1) + E_b(\lambda^s, N_s)\right]}.\end{aligned} \tag{2.53}$$

Finalement, en utilisant (2.45), (2.49) et (2.53) dans (2.48), nous obtenons $\hat{r}_T^{n+1}$ et par conséquent le multiplicateur à l'itération n, donné par $\gamma = \hat{r}_T^{n+1} / \hat{r}_T^n$.

Comme mentionné à la fin de la Section 2.3.3, le processus itératif d'adaptation de paramètres de récompense décrit ici est intégré à ceux des adaptations de capacité (Section 2.4.2) et de routage (Section 2.3.1) en un processus global, où chaque algorithme d'adaptation est exécuté à son propre intervalle de temps choisi.

2.5 Analyse comparative des modèles d'adaptation

À cause de la complexité du modèle d'adaptation de capacité MDP, la validation par comparaison du modèle d'adaptation MDPD ne peut seulement s'effectuer que sur des petits réseaux. Dans cette section, nous présentons les résultats de cette validation impliquant deux exemples de réseaux simples, un réseau à trois liens et un autre à cinq liens, représentés à la Figure 2.3 et identifiés respectivement par *3L* et *5L*. Chaque réseau comporte trois classes de connexions dont les paires origine-destination correspondantes sont : pour *3L*, *ON1-ON2, ON2-ON3 et ON3-ON1*; et pour *5L, ON1-ON2, ON1-ON3 et ON1-ON4*. Toutes les classes sont définies ayant un paramètre de récompense unitaire, un besoin de largeur de bande unitaire et un temps de service exponentiel de moyenne unitaire. Dans les deux cas de réseaux, les demandes de connexions offertes ont des taux d'arrivée Poissoniens, respectifs aux origine-destination données ci-haut, de λ_1, λ_2 et λ_3. Les coûts unitaires de largeur de bande dans tous les liens sont fixés à 0,2. Dans les deux réseaux, les connexions peuvent être établies sur un chemin à un ou deux liens.

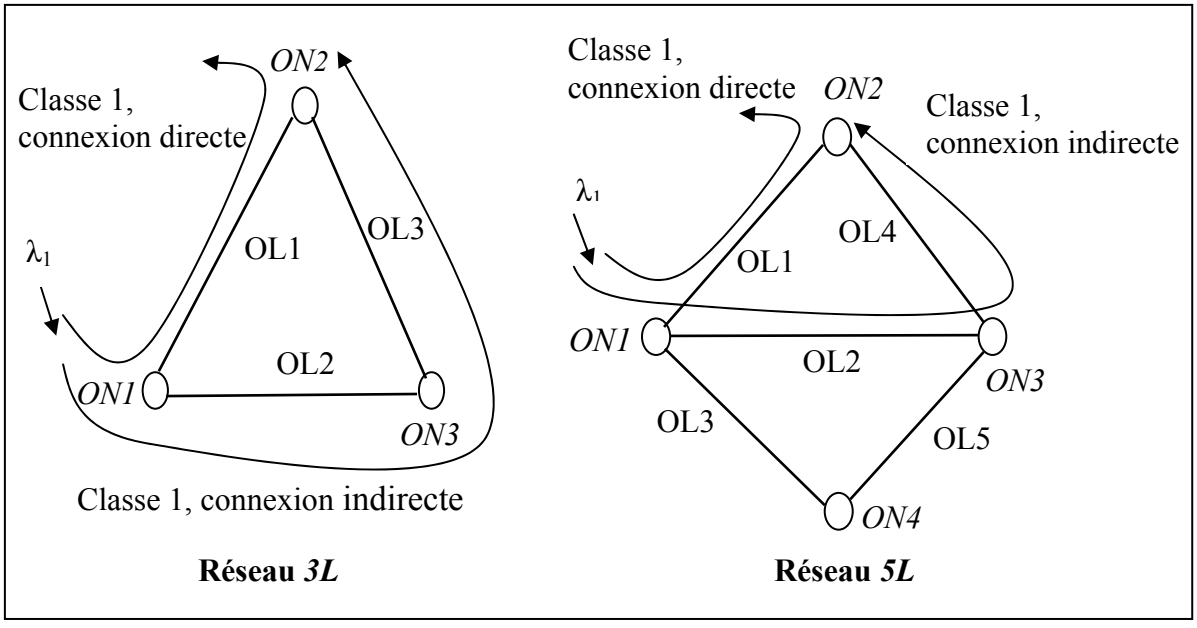

Figure 2.3 Réseaux exemples *3L* et *5L*.

2.5.1 Adaptation de capacités de lien

Le modèle d'adaptation de capacité MDPD est vérifié sur le réseau *3L* pour les cas de taux d'arrivée symétriques et asymétriques des classes de connexion, respectivement $\boldsymbol{\lambda} = (3, 3, 3)$ et $\boldsymbol{\lambda} = (4, 3, 2)$. Les capacités de liens après convergence de l'algorithme et les bénéfices de réseaux correspondants, obtenus par les deux modèles MDP et MDPD, sont données au Tableau 2.1. Dans le cas de trafic symétrique, la convergence est atteinte en une seule itération pour chacun des deux modèles. Dans le cas asymétrique, l'algorithme d'adaptation MDPD converge en deux itérations et celui d'adaptation MDP en trois itérations. Dans les deux scénarios de trafic, les deux modèles produisent des capacités optimales de liens identiques.

Tableau 2.1 Comparaison des adaptations de capacité MDP et MDPD, réseau *3L*

	$\boldsymbol{\lambda} = (3, 3, 3)$		$\boldsymbol{\lambda} = (4, 3, 2)$	
	N	$P(\mathbf{N})$	**N**	$P(\mathbf{N})$
Conditions initiales	(3, 3, 3)	4,106	(5, 5, 5)	5,164
Adaptation MDP	(5, 5, 5)	5,254	(7, 5, 4)	5,285
Adaptation MDPD	(5, 5, 5)	5,254	(7, 5, 4)	5,285

Dans le scenario pour le réseau *5L* un peu plus grand, dont les résultats d'adaptation sont donnés au Tableau 2.2, la convergence est atteinte en une itération dans le modèle MDP et en deux itérations dans le modèle MDPD. De nouveau, les capacités indiquées après convergence sont similaires dans les deux modèles. Elles sont identiques pour les liens directs (N_1, N_2 *et* N_3) et diffèrent seulement d'une unité de largeur de bande pour les liens servant uniquement aux chemins indirects. De plus, dans ce cas, le taux résultant de bénéfice de réseau du modèle MDPD ne diffère que de 0,4% du taux optimal obtenu par le modèle MDP. Durant cet exercice de validation de modèle, nous avons aussi vérifié par inspection que les solutions MDP obtenues dans les différents scénarios sont en effet toutes optimales.

Tableau 2.2 Comparaison des adaptations de capacité MDP et MDPD, réseau *5L*

	$\boldsymbol{\lambda} = (3, 4, 5)$	
	N	$P(\mathbf{N})$
Conditions initiales	(5, 5, 5, 1, 1)	4,106
Adaptation MDP	(5, 6, 8, 1, 1)	7,076
Adaptation MDPD	(5, 6, 8, 0, 0)	7,046

2.5.2 Adaptation de paramètres de récompense

Dans les exemples présentés ci-dessus, les adaptations de capacité entrainent les probabilités de blocage au réseau de 8,2%, 5,7% et 6%, pour les cas respectifs de trafic symétrique au réseau *3L*, trafic asymétrique au réseau *3L* et réseau *5L*. Pour ramener ces probabilités de blocage à leurs contraintes spécifiées, la procédure d'adaptation de paramètre de récompense décrite à la Section 2.4.3 est appliquée.

La Figure 2.4 illustre cette adaptation dans le cas de trafic asymétrique au réseau 3L, avec une contrainte de blocage au réseau B_T^c = 1%. Comme montré dans la partie gauche de la figure, la contrainte est satisfaite après quatre itérations. Notons que la probabilité de blocage est déjà devenue proche de la contrainte après la première itération, les itérations suivantes servant principalement à raffiner l'adaptation. Ce comportement est justifié par la granularité grossière des capacités des liens dans l'exemple considéré.

Les capacités des liens et les taux de bénéfice de réseau correspondant aux itérations de l'adaptation de paramètre de récompense sont montrés dans la partie droite de la figure. Les résultats confirment que l'augmentation des paramètres de récompense entraine l'augmentation de capacités de lien. Dans notre exemple où ce paramètre représente le revenu moyen de connexion, l'adaptation de capacité résultante entraine l'augmentation de bénéfice montrée.

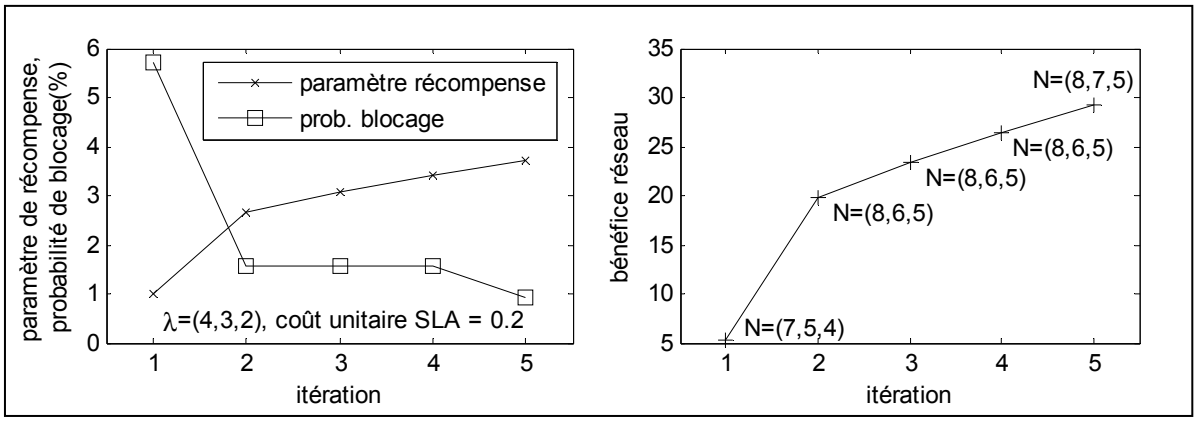

Figure 2.4 Adaptation de paramètres de récompense, réseau *3L*.

2.5.3 Conclusion de l'analyse

En résumé, la comparaison ci-dessus des résultats donnés par le modèle d'adaptation de capacité MDPD avec ceux du modèle exact d'adaptation MDP, bien qu'effectuée sur des exemples de petits réseaux, supporte bien la validité du modèle MDPD. En particulier, ces résultats démontrent que les solutions MDPD obtenues sont soit optimales ou soit très proches de l'optimal. De plus, ces solutions sont atteintes suite à un nombre très limité d'itérations, indiquant ainsi une très bonne caractéristique de convergence de l'algorithme proposé.

2.6 Analyse de performance du modèle d'adaptation MDPD

Dans cette section, nous analysons la performance des modèles d'adaptation de capacité MDPD et d'adaptation de paramètre de récompense présentés aux Sections 2.4.2 et 2.4.3. L'analyse est effectuée sur un exemple de réseau SON comportant un nombre réaliste de nœuds, liens et classes de connexion. Nous commençons par une description de l'estimation appuyée par des mesures des paramètres de trafic requis pour l'exécution des modèles d'adaptation. Ensuite nous proposons un ensemble de métriques qui permettent l'évaluation, aux points de vue de convergence et de stabilité, de l'adaptation de capacité. Enfin, à l'aide de ces métriques et de simulations à événements discrets, nous analysons la performance des adaptations dans des conditions de variations significatives de trafic et de prix SLA.

2.6.1 Estimation du trafic admis de lien

Comme indiqué dans les Sections 2.4.2 et 2.4.3, la réalisation des modèles d'adaptation fondés sur MDPD nécessite une estimation, appuyée par des mesures, des taux moyens d'admission de connexion aux liens $\overline{\lambda}_j^s$ et $\overline{\lambda}^s$, et des paramètres de récompense agglomérée de lien r^s. Pour réaliser cette estimation, nous estimons en premier les quantités moyennes de connexions admises aux liens $\overline{A}_j^s$ et $\overline{A}^s$, ainsi que les taux de service correspondants μ^s. Pour cela, nous mesurons les occupations moyennes et les temps de service moyens des liens, à intervalles réguliers de durée t_m. Ces mesures constituent des échantillons sur lesquels nous appliquons ensuite une méthode choisie d'estimation.

Pour l'analyse dans ce chapitre, nous considérons deux options simples de méthode d'estimation. Nous élaborerons plus tard sur des méthodes plus avancées dans le chapitre suivant traitant de ce sujet. La première méthode est fondée sur le filtre par lissage exponentiel (ES), réalisant une moyenne pondérée où un poids a_{es} est appliqué au plus récent échantillon. La deuxième est fondée sur une moyenne mobile (MA), obtenue en faisant la moyenne des échantillons dans une fenêtre de w_{ma} éléments. À chaque intervalle t_m, la fenêtre avance d'un échantillon.

Aussi, étant donné que les adaptations de capacité entrainent des changements correspondants dans l'admission de connexions aux liens, quand un changement de capacité de lien survient, les valeurs dernièrement estimées sont réajustées d'après ce qui suit :

$$\overline{A}_{adj} = A\left[1 - E_b(A, N + \Delta N)\right] = \frac{\overline{A}}{1 - E_b(A, N)}\left[1 - E_b(A, N + \Delta N)\right], \tag{2.54}$$

où ΔN est le changement de capacité du lien et $\overline{A}_{adj}$ est l'occupation réajustée du lien. Ainsi, nous obtenons enfin :

$$\overline{\lambda}_j^s = \overline{A}_j^s \mu^s, \tag{2.55}$$

$$\overline{\lambda}^s = \overline{A}^s \mu^s. \tag{2.56}$$

Pour avoir une bonne précision pour $\overline{\lambda}_j^s$ et $\overline{\lambda}^s$, l'estimation de μ^s est effectuée de façon consistante avec celles de $\overline{A}_j^s$ et $\overline{A}^s$ avec la même valeur du paramètre d'estimation (a_{es} ou w_{ma}). Introduisant les valeurs estimées de $\overline{\lambda}_j^s$ et $\overline{\lambda}^s$ dans (2.41), nous obtenons alors celle de r^s.

2.6.2 Métriques pour la performance de l'adaptation de capacité

Dans la présente analyse, la performance de l'adaptation de capacité, quand le trafic offert est périodiquement augmenté et diminué de façon significative, est évaluée. Pour cette évaluation, nous introduisons ici deux types de métriques. Les métriques de convergence mesurent la rapidité avec laquelle, suite à un changement de trafic, l'adaptation atteint les nouvelles capacités optimales de lien. Celles de stabilité mesurent la variation des capacités de lien dans des périodes de trafic stationnaire, après que la convergence a été atteinte.

2.6.2.1 Métriques de convergence

Les métriques de lien suivantes sont définies pour la convergence de l'adaptation de capacité :

- Durée de convergence d_C^s :

$$d_C^s = t_f^s - t_0^s, \tag{2.57}$$

  où t_0^s est le temps du changement de trafic et t_f^s est le temps où la capacité du lien *s* complète la convergence à sa valeur optimale $\overline{N}_s$.

- Écart de convergence σ_C^s :

$$\sigma_C^s = t_m \sum_{n=1..d_C} \left| N_s^n - \overline{N}_s \right| \Big/ \overline{N}_s \,, \tag{2.58}$$

  où N_s^n est la capacité du lien *s* à l'intervalle de mesure d'indice *n* dans la période de convergence. Cette métrique reflète l'écart du cas idéal où la convergence est accomplie instantanément.

Les métriques globales de réseau correspondantes, de durée et d'écart de convergence, sont définies par :

$$\overline{d}_C = \frac{1}{L}\sum_s d_C^s, \tag{2.59}$$

$$\overline{\sigma}_C = \frac{1}{\sum_s \overline{N}_s}\sum_s \sigma_C^s \overline{N}_s, \tag{2.60}$$

où L représente le nombre de liens du réseau.

Pour toutes ces métriques, une évaluation plus basse représente une meilleure convergence.

2.6.2.2 Métriques de stabilité

Dans une période de trafic stationnaire, la métrique de *gigue de stabilité* de lien α_S^s mesure le nombre de changements de capacité du lien s par unité de temps. Concernant les écarts d'amplitude, nous proposons la métrique d'*écart de stabilité* σ_S^s, définie par l'écart type relatif de la capacité de lien avec la capacité optimale $\overline{N}_s$:

$$\sigma_S^s = \sqrt{\frac{1}{l}\sum_{n=1..l}\left(N_s^n - \overline{N}_s\right)^2} \Big/ \overline{N}_s , \tag{2.61}$$

où l est le nombre d'intervalles de mesure dans la période considérée.

De manière équivalente aux métriques de convergence, les métriques de stabilité globales de réseau sont définies par :

$$\overline{\alpha}_S = \frac{1}{L}\sum_s \alpha_S^s, \tag{2.62}$$

$$\overline{\sigma}_S = \frac{1}{\sum_s \overline{N}_s}\sum_s \sigma_S^s \overline{N}_s. \tag{2.63}$$

De même pour ces métriques, une évaluation plus basse représente une meilleure stabilité.

2.6.3 Évaluation de performance

Nous présentons dans cette section l'évaluation de la performance de notre approche d'adaptation de ressources MDPD, effectuée à l'aide d'un simulateur à événements de réseau.

Durant la simulation, des demandes de connexion avec des périodes inter-arrivées et des temps de service distribués exponentiellement sont offertes au réseau SON simulé. Les paramètres requis de trafic de liens sont obtenus en utilisant le modèle d'estimation appuyé par des mesures décrit à la Section 2.6.1. Les connexions sont admises et acheminées selon la politique de CAC et de routage MDPD maximisant le bénéfice décrite à la Section 2.3.1. Quand des changements de trafic et/ou de conditions de prix SLA sont injectés dans la simulation, les capacités de lien et les paramètres de récompense sont réadaptés pour maintenir la maximisation du bénéfice, en suivant les modèles d'adaptation présentés aux Sections 2.4.2 et 2.4.3.

Les simulations sont exercées sur un exemple de réseau SON, appelé NEA_20L et montré à la Figure 2.5, qui relie des grandes villes du Nord Est de l'Amérique. Ce réseau inclut 10 nœuds et 20 liens, ce qui produit un total possible de 45 paires d'origine-destination, chacune représentant une classe de connexion. Le routage dans chaque classe peut sélectionner tout chemin possible, jusqu'à un maximum de cinq bonds. Dans les exemples considérés, les connexions sont homogènes, i. e. elles ont toutes un besoin de largeur de bande unitaire et leurs temps de service moyen t_s sont tous d'une unité de temps. Les durées assignées aux périodes de rafraichissement de la politique de routage et des capacités de lien, ainsi qu'à l'intervalle de mesure de trafic, sont $t_p = t_c = t_m = 2{,}5t_s$ (t_p peut être choisi égal à t_c car la convergence de la politique de routage est très rapide).

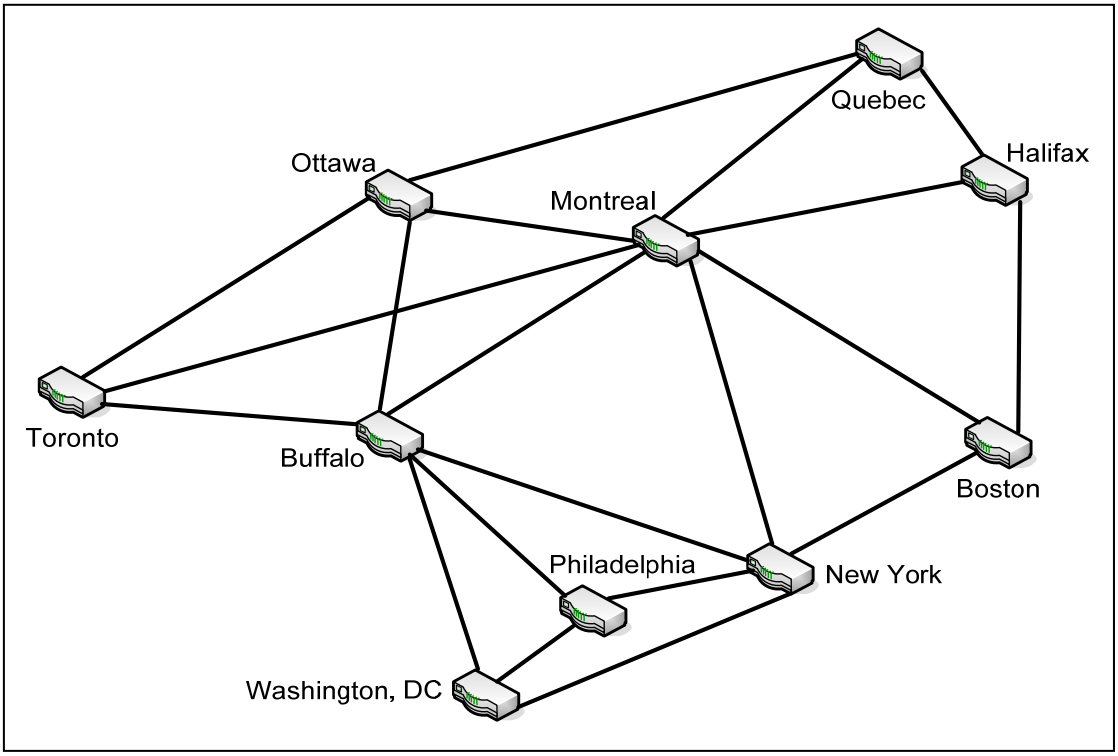

Figure 2.5 Exemple de réseau considéré NEA_20L.

2.6.3.1 Performance de l'adaptation de capacité

Dans le scénario considéré, le tiers des classes de connexion ont un taux d'arrivée de connexion initial de $20/t_s$ et le restant des classes un taux d'arrivée initial de $13/t_s$. Le coût unitaire de largeur de bande est fixé à 0,014 et les liens sont dimensionnés initialement pour réaliser le bénéfice optimal sous les conditions initiales de trafic, avec des capacités s'échelonnant de 25 à 62 unités de largeur de bande.

La simulation de réseau est d'abord exécutée avec les niveaux initiaux de trafic pendant une période de $500t_s$ pour atteindre la stabilité du système. À $500t_s$, tous les taux d'arrivée sont augmentés de 20% pour constituer des hauts niveaux de trafic, puis à $1250t_s$, ils sont tous ramenés aux niveaux initiaux de bas trafic pour une période supplémentaire de $750t_s$. Un ensemble de trois cycles de ces hauts et bas trafics est simulé pour obtenir des résultats statistiquement valides.

Nous évaluons d'abord la convergence et la stabilité de l'adaptation de capacité quand chacune des deux méthodes d'estimation de trafic, ES et MA (Section 2.6.1), est utilisée. Durant cette évaluation, les paramètres de récompense de toutes les classes sont fixés à 7 unités. Les comparaisons des deux méthodes d'estimation et l'analyse de compromis entre la convergence et la stabilité sont présentées avec le Tableau 2.3 et la Figure 2.6. Comme on peut le prévoir, quand on améliore la convergence, par une réduction de la fenêtre MA ou une augmentation du coefficient de lissage ES, la stabilité est détériorée de manière correspondante. Les valeurs de a_{es}=0,1 et w_{ma}=50 donnent des compromis raisonnables, respectivement dans les cas d'estimation ES et MA. La comparaison, dans la Figure 2.6, des deux méthodes d'estimation donne un léger avantage au lissage exponentiel. Les valeurs de bénéfice montrées au Tableau 2.3 représentent des taux moyens calculés pour la durée entière de la simulation. Les taux obtenus sont proches les uns des autres, toujours avec un léger avantage du lissage exponentiel sur la moyenne mobile.

Tableau 2.3 Convergence et stabilité des adaptations de capacité MDPD, Augmentation de traffic de 20%

Méthode d'estimation		Augmentation de traffic de 20%				Bénéfice réseau P
		$\overline{d}_c$	$\overline{\sigma}_C$	$\overline{\sigma}_S$	$\overline{\alpha}_S$	
MA, taille de fenêtre w_{ma} (en t_s)	*20*	43	16	0,0398	0,0905	7,82
	50	106	39	0,0236	0,0200	7,92
	100	209	71	0,0153	0,0047	7,93
ES, coefficient a_{es}	*0.2*	45	16	0.0308	0.0961	7.94
	0.1	116	32	0,0221	0,0295	7,97
	0.05	221	57	0,0170	0,0070	7,94

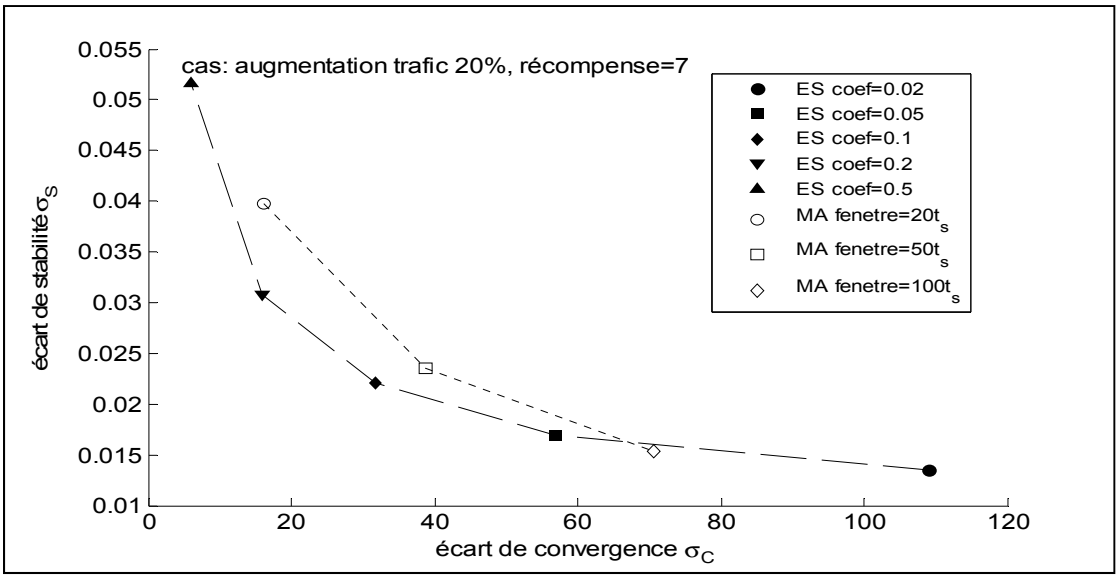

Figure 2.6 Convergence vs. Stabilité, comparaison des estimations ES et MA.

La comparaison des Figure 2.7 et Figure 2.8 confirme l'influence du paramètre d'estimation sur les caractéristiques de l'adaptation de capacité d'un lien à l'augmentation du trafic offert. Les exemples de moyenne mobile avec des tailles de fenêtres respectives de 20 t_s et 100 t_s sont montrés. Nous constatons bien qu'avec la plus petite fenêtre, la capacité converge beaucoup plus rapidement vers sa valeur adaptée, cependant cette capacité est bien plus instable que dans le cas de la plus grande fenêtre.

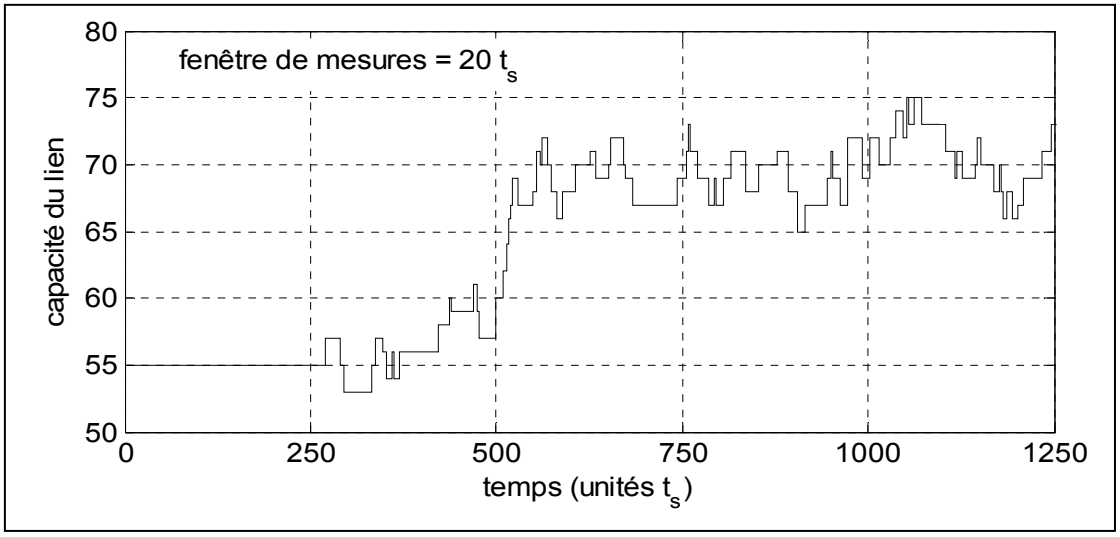

Figure 2.7 Adaptation de capacité du lien MTL-QUE (MA, fenêtre de 20 t_s).

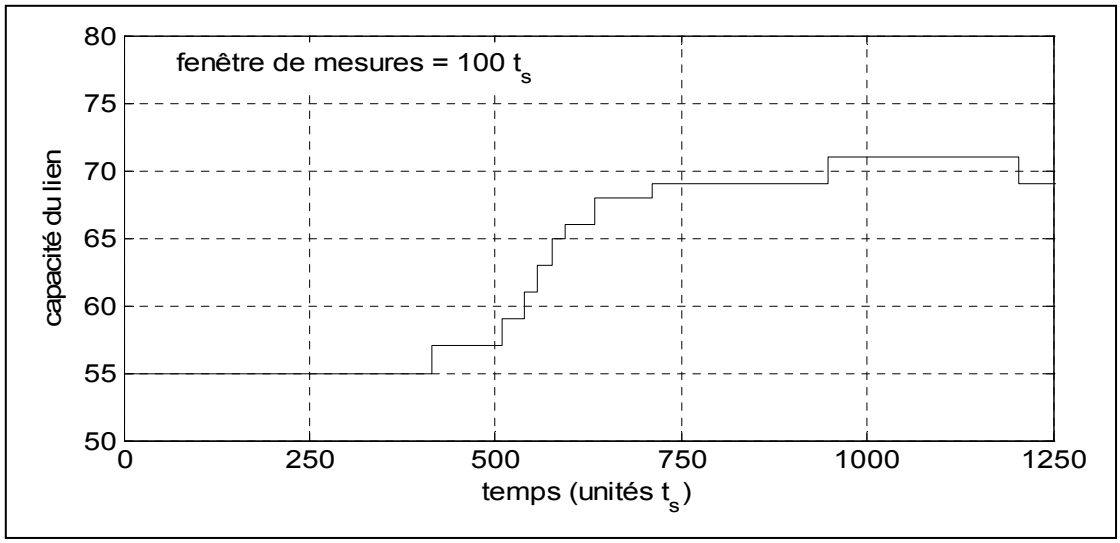

Figure 2.8 Adaptation de capacité du lien MTL-QUE (MA, fenêtre de 100 t_s).

Nous démontrons maintenant les gains en bénéfice du réseau ainsi que les probabilités de blocage du réseau correspondants, obtenus par l'adaptation de capacité des liens. Pour cela, les performances dans le cas d'un réseau à capacité fixe sont utilisées comme repère de comparaison. Dans ce réseau fixe, les liens sont dimensionnés avec les capacités optimales $\overline{N}_s$ trouvées pour le niveau (haut ou bas) de trafic choisi. Dans cet exercice, les paramètres de récompense de toutes les classes sont fixés à 10 afin de réaliser des probabilités de blocage du réseau raisonnables de ≈1.5%.

Comme montré au Tableau 2.4, quand les capacités du réseau fixe ont été optimisées pour le haut niveau de trafic, le taux de bénéfice du réseau fixe en période de bas trafic est de 23% inférieur au cas du réseau adaptatif. Dans l'autre cas où les capacités fixes sont optimisées pour le bas niveau de trafic, le bénéfice du réseau fixe est proche de celui du réseau adaptatif, cependant la probabilité de blocage du réseau fixe est augmentée à 11% en période de haut trafic, ce qui est inacceptable. Par contre avec notre approche proposée de capacités adaptatives, le bénéfice du réseau est maintenu optimisé aux deux niveaux de trafic, avec des probabilités de blocage toujours gardées au niveau désiré.

Tableau 2.4 Comparaison des bénéfices réseau et des probabilités de blocage

	Périodes de haut trafic		Périodes de bas trafic		Toutes périodes
	Bénéfice P	*Blocage* B_T	Bénéfice P	*Blocage* B_T	Bénéfice P
Capacités fixes, optimisées pour haut trafic	20,68	*1,12 %*	12,53	*0,01 %*	16,61
Capacités fixes, optimisées pour bas trafic	20,49	*11,27 %*	16,28	*1,43 %*	18,39
Capacités adaptatives	20,74	*1,48 %*	16,23	*1,17 %*	18,49

2.6.3.2 Performance de l'adaptation de paramètres de récompense

Les probabilités de blocage du SON pour le scénario d'adaptation de capacité considéré à la section précédente sont illustrées à la Figure 2.9. La contrainte de blocage du réseau de 1,5% est respectée durant les périodes d'avant et d'après la convergence de l'adaptation de capacité. Ceci indique qu'avec des paramètres de récompense de connexion bien conçus en fonction des coûts SLA, les probabilités de blocage demeurent à un niveau acceptable même quand les capacités de lien changent à cause des fluctuations de trafic.

La situation est cependant différente quand les prix SLA de la bande passante des liens augmentent. Dans ce cas, pour maintenir la maximisation du bénéfice face à l'augmentation des coûts, l'algorithme d'adaptation de capacité réduit la capacité des liens, ce qui entraine une augmentation des probabilités de blocage. Quand un problème de dépassement des contraintes de blocage survient, l'algorithme d'adaptation des paramètres de récompense (Section 2.4.3), intégré à celui d'adaptation de capacité, est utilisé pour le résoudre. Nous vérifions dans la suite de cette section la performance de cet algorithme.

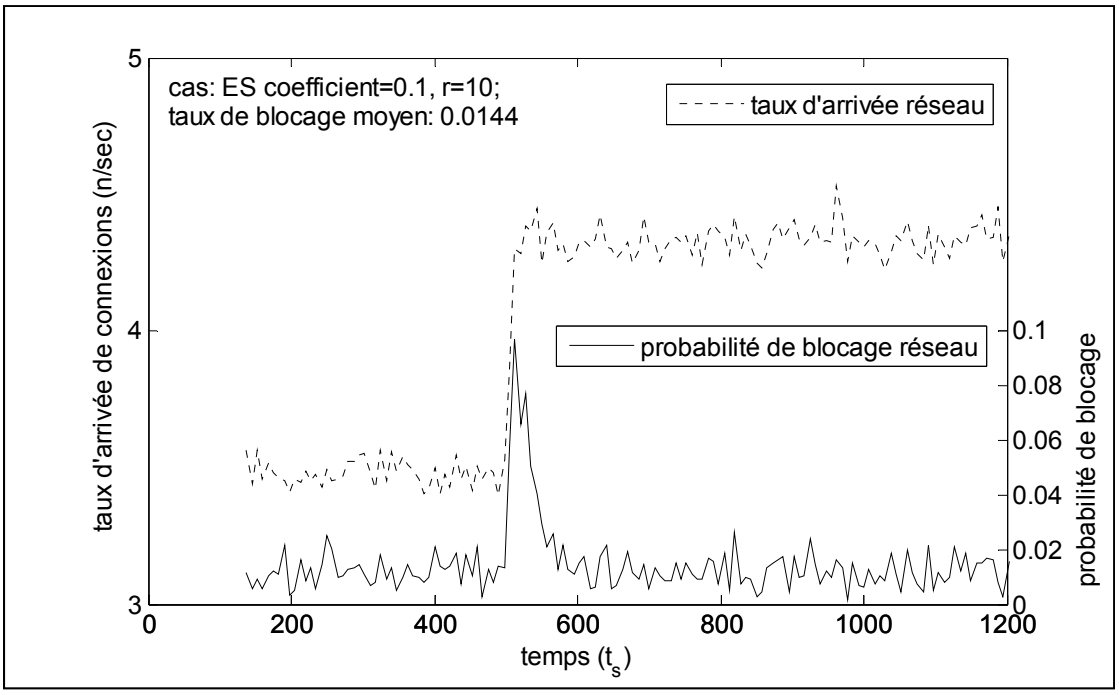

Figure 2.9 Trafic simulé et probabilités de blocage dans réseau NEA_20L.

Dans le scénario considéré, la contrainte de blocage du réseau, B_T^c, est fixée à 1% et, au temps 500 t_s, une augmentation de 30% du prix des SLA est appliquée à la bande passante de tous les liens. Nous assignons $t_r = 5t_m$ à la période de rafraichissement des paramètres de récompense. Comme indiqué à la Section 2.4.3, l'adaptation des paramètres de récompense vise l'égalité $B_T = B_T^c$, et nous permettons une tolérance de ±0.5% à cet objectif.

Les Figure 2.10 et Figure 2.11 montrent les résultats obtenus, respectivement pour les cas sans adaptation et avec adaptation de paramètres de récompense. La comparaison des graphes de blocage démontre clairement que l'algorithme d'adaptation de paramètres de récompense ramène rapidement les probabilités de blocage à des valeurs satisfaisant la contrainte, tandis que l'absence de l'algorithme entraine des dépassements significatifs de la contrainte.

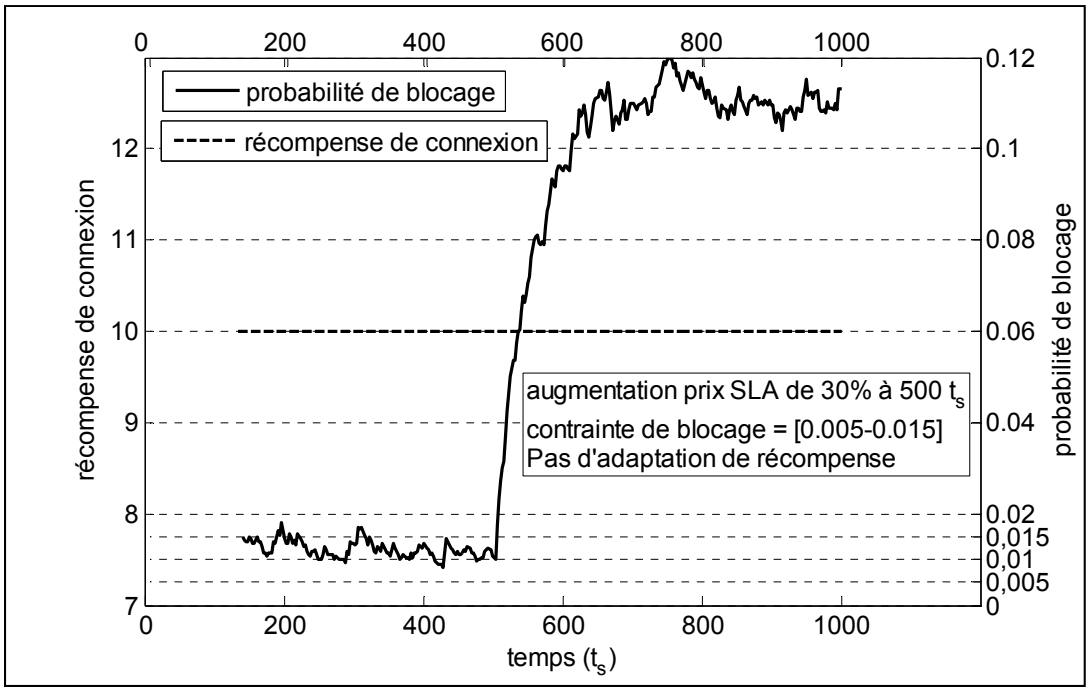

Figure 2.10 Récompense de connexion et probabilité de blocage
(augmentation du prix SLA de 30%, sans adaptation de paramètre de récompense).

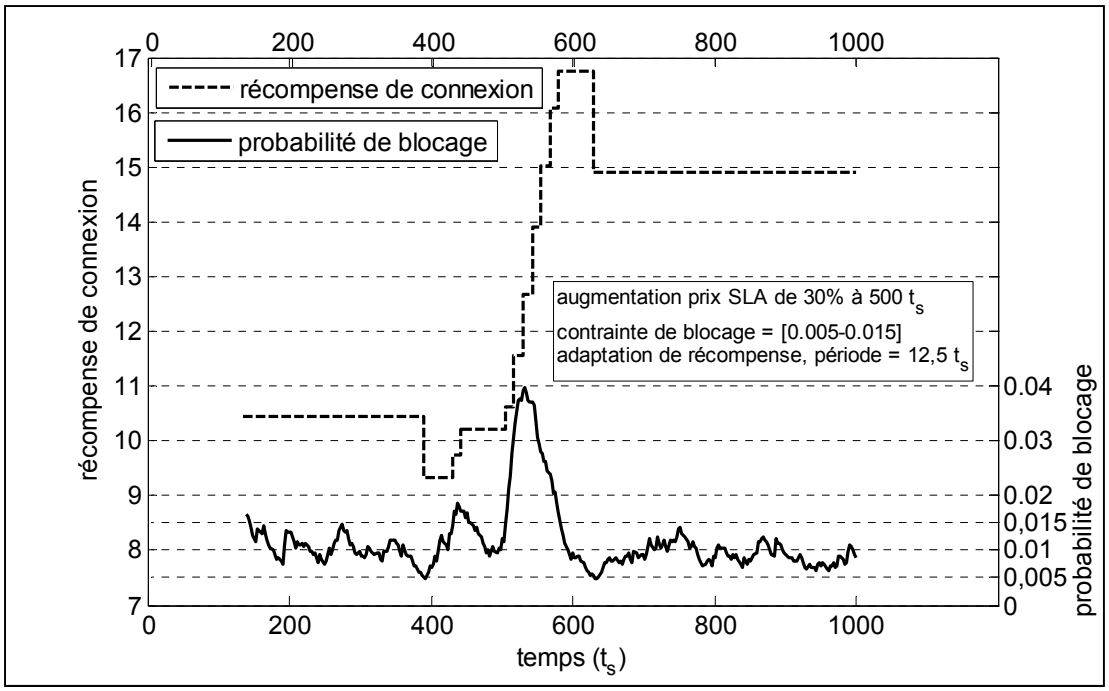

Figure 2.11 Adaptation de paramètre de récompense, $t_r = 5t_m$.

2.7 Résumé

Dans ce chapitre, nous avons proposé une approche d'adaptation de capacité s'appliquant aux réseaux dédiés de service. Les liens de ces réseaux sont établis en achetant de la bande passante de fournisseurs de service Internet, par le biais de SLA dont les termes, incluant la largeur de bande achetée, peuvent être modifiés à brève échéance. L'objectif de l'approche est de maintenir la maximisation du bénéfice du réseau tout en respectant les contraintes de probabilités de blocage, par des adaptations de largeur de bande et de récompenses de connexions.

La maximisation de bénéfice repose sur un modèle économique, à partir duquel nous établissons les conditions d'optimalité du bénéfice. Ce modèle économique, fondé sur la théorie de décision de Markov, fournit un cadre qui permet l'intégration de l'adaptation de capacité et de celle de récompenses de connexions avec le CAC et routage optimal du réseau. L'élément primordial dans cette intégration est la décomposition du problème réseau en un ensemble de problèmes séparés de liens, menant au concept de *shadow price* de lien qui fournit une base économique consistante à toutes les composantes de l'approche. Cette décomposition permet une solution distribuée, fondée sur la mesure et l'estimation de paramètres de lien.

La validité de l'approche a été confirmée à l'aide d'une comparaison analytique des résultats obtenus avec ceux d'une solution exacte, impliquant des exemples de petits réseaux. La comparaison démontre l'obtention de capacités optimales semblables dans les deux solutions, avec une différence des bénéfices maximisés de moins que 0,5%. Les simulations effectuées sur des exemples réalistes de réseaux ont démontré de bonnes performances de convergence de l'approche. Les essais ont montré que le compromis entre la convergence et la stabilité de l'adaptation de capacité peut être contrôlé par les paramètres d'estimation. Comparé au cas de capacités fixes provisionnées pour le trafic crête, le bénéfice obtenu avec les capacités adaptatives est supérieur d'environ 11%. Si les capacités fixes étaient provisionnées pour augmenter le bénéfice en considérant plutôt le bas trafic, le taux de blocage augmenterait à

11%, alors que l'approche d'adaptation de capacité maintient continuellement ce taux à moins que 1,5%.

Les résultats confirment aussi que les contraintes de blocage peuvent être satisfaites par l'adaptation de paramètres de récompense proposée. Bien que l'approche proposée ici ait été démontrée utilisant des connexions homogènes, sa couverture peut présumément aussi s'appliquer aux connexions *multi-rate*.

CHAPITRE 3

ESTIMATION DU TRAFIC EN TEMPS RÉEL POUR LA GESTION DES RESSOURCES

3.1 Introduction

Dans le chapitre précédent, nous avions présenté une proposition de gestion des ressources du SON avec comme objectif la maximisation du bénéfice du réseau, le tout dans le respect des contraintes de degré de service (GoS) associées au blocage de connexions. En particulier, le bénéfice est maintenu maximisé face aux variations journalières de trafic (Thompson, Miller et Wilder, 1997) et (Roberts, 2004), par une approche d'adaptation optimale de capacité des liens au trafic. L'algorithme d'adaptation MDPD, fondée sur la théorie de décision de Markov, obtient ses données de trafic par un processus d'estimation appuyé par des mesures fréquentes de trafic.

Dans ce précédent CHAPITRE 2, l'algorithme d'adaptation a été vérifié en utilisant des méthodes simples d'estimation (ES et MA) incorporant des paramètres d'estimation (respectivement coefficient de lissage et taille de fenêtre mobile) fixes. Avec ces méthodes simples, un compromis entre les qualités de convergence et de stabilité de l'adaptation était nécessaire (Section 2.6.3.1). Aussi, il est naturel que l'efficacité de l'approche d'adaptation, tant pour le maintien du bénéfice maximisé que pour celui de la GoS, dépend de la qualité de l'estimation de trafic.

L'objectif poursuivi dans ce présent chapitre est donc de rechercher une méthodologie d'estimation efficace de tendance de trafic qui, une fois intégrée à l'algorithme d'adaptation de capacité, procurera des gains additionnels de bénéfice et de GoS. En particulier, nous proposons deux options d'approche, fondées sur des concepts différents, qui procurent différentes capabilités à l'adaptation de capacité.

La première approche s'établit dans le cadre du lissage exponentiel adaptatif, où le coefficient de lissage est adapté à la tendance estimée de la demande de trafic. Pour l'estimation de la tendance, nous proposons deux nouvelles méthodes : la première est fondée sur l'analyse de la fonction d'autocorrélation du processus d'arrivée de trafic et la deuxième sur l'analyse de la fonction de distribution cumulée du même processus. Comme ce sera démontré plus tard dans ce chapitre, quand elle est comparée à d'autres méthodes de lissages exponentiels connues, cette approche améliore de façon significative la réponse aux changements de tendance du trafic, tout en préservant la stabilité de l'estimation en périodes de trafic stationnaire. Avec l'intégration de l'approche à l'adaptation de capacité, les caractéristiques citées se traduisent par de meilleurs bénéfices et un meilleur contrôle de la GoS.

La deuxième approche est motivée par le fait que toute méthode d'estimation ponctuelle est sujette à des erreurs d'estimation. Ainsi, une allocation optimale de capacité fondée seulement sur un estimé ponctuel peut entrainer des probabilités élevées de violations de contraintes de GoS. Pour restreindre ces violations à un niveau acceptable, nous proposons d'utiliser une méthode d'estimation par intervalles qui en même temps donne une distribution de l'erreur d'estimation. La méthode est réalisée par l'application du filtre de Kalman dans lequel le modèle de trafic est obtenu à partir de mesures historiques de trafic. Ainsi dans cette approche, l'adaptation de capacité est effectuée en tenant compte de la variance de l'erreur d'estimation. Les résultats numériques des essais démontreront que cette approche procure un haut degré de respect des objectifs de GoS couplé à des bénéfices de réseau élevés.

Dans la suite de ce chapitre, nous commençons par rapporter à la Section 3.2 des méthodes d'estimation connexes à nos approches et trouvées dans la littérature. Ensuite, notre approche d'estimation de tendance fondée sur le lissage exponentiel adaptatif est présentée à la Section 3.3. Celle fondée sur le filtre de Kalman est décrite à la Section 3.4. Enfin à la Section 3.5, la performance de nos approches proposées est analysée.

3.2 Méthodes connues d'estimation de la demande de trafic

Comme rapporté par (Thompson, Miller et Wilder, 1997) et (Roberts, 2004), l'intensité d'un trafic subit typiquement des variations journalières significatives qui correspondent au gabarit de l'activité des usagers du réseau. Notre modèle d'adaptation en ligne de capacité pour la maximisation de bénéfice (Section 2.3) nécessite alors une connaissance en temps réel des taux d'arrivée de connexions aux liens λ_j^s.

Les taux sont estimés à des instants discrets *k* séparés par des intervalles de temps réguliers ΔT. Comme dans notre approche la demande agglomérée de trafic au lien est estimée et que les processus d'arrivées de connexions aux liens sont présumés indépendants, nous omettons maintenant les indices *j* et *s* de λ_j^s pour simplifier la notation et désignons le taux d'arrivée aggloméré au lien, à chaque instant *k*, par λ_k. La demande de trafic peut alors être formulée par :

$$\lambda_k = \lambda_{k-1} + \Delta\lambda_k + e_k, \tag{3.1}$$

où $\Delta\lambda_k$ est l'espérance du taux de variation de λ_k dans l'intervalle (*k-1, k*), qui représente la tendance de trafic, et e_k est une déviation aléatoire de la valeur espérée de cette tendance. Nous présentons ci-après quelques approches d'estimation de trafic et de sa tendance trouvées dans la littérature.

3.2.1 Estimation par lissage exponentiel

Le lissage exponentiel (ES) représente une technique d'estimation bien connue et communément utilisée. Il a l'avantage d'être une méthode simple, utilisant une formule récursive qui ne nécessite aucune mémoire de mesures passées. Il peut aisément s'adapter avec le temps aux changements de la variable estimée. Par contre, les faiblesses de cette méthode proviennent de l'absence d'un modèle statistique formel et de la non-disponibilité d'une évaluation de l'erreur d'évaluation. Une vue d'ensemble de l'état de la recherche sur le

lissage exponentiel est donnée dans (Gardner, 2006) où différentes variantes de la méthode sont documentées.

Appliquée dans notre cas à l'estimation du taux d'arrivée des connexions, la forme du lissage exponentiel simple (SES) est définie par :

$$\hat{\lambda}_k = \alpha_k \tilde{\lambda}_k + (1-\alpha_k)\hat{\lambda}_{k-1}, \quad k = 1,2,... \tag{3.2}$$

où $\tilde{\lambda}_k$ et $\hat{\lambda}_k$ représentent respectivement les valeurs mesurée et estimée du taux d'arrivée au temps *k*, et α_k le paramètre de lissage, $0 < \alpha_k < 1$. La valeur estimée réalise une moyenne pondérée de toutes les mesures courante et passées. En plus de cette forme simple, il existe aussi une forme de lissage exponentiel double (DES) où la tendance estimée est explicitement évaluée. Un ensemble de deux équations exprime ce lissage double :

$$\hat{\lambda}_k = \alpha_k \tilde{\lambda}_k + (1-\alpha_k)(\hat{\lambda}_{k-1} + \hat{\Delta}\lambda_{k-1}), \quad k = 1,2,... \tag{3.3}$$

$$\hat{\Delta}\lambda_k = \gamma_k(\hat{\lambda}_k - \hat{\lambda}_{k-1}) + (1-\gamma_k)\hat{\Delta}\lambda_{k-1}, \quad k = 1,2,... \tag{3.4}$$

où $\hat{\Delta}\lambda_k$ représente la valeur estimée de la tendance et γ_k son paramètre de lissage, $0 < \gamma_k < 1$.

Les paramètres de lissage dans (3.2), (3.3) et (3.4) influencent les caractéristiques de l'estimation telles que sa réponse et sa stabilité. Ils peuvent être choisis comme étant fixes ou adaptatifs. Un total de 24 techniques ES ont été rapportées en 1982 dans (Makridakis *et al.*, 1982). Par la suite, *Adaptive Extended ES* (AEES (Mentzer, 1988) et AEES-C (Mentzer et Gomes, 1994)) ont été développés pour procurer une précision accrue de l'estimation. Ci-après, nous résumons une sélection de deux méthodes SES pour leur simplicité et de deux méthodes DES pour leur précision améliorée. Ces méthodes seront utilisées comme étalons de comparaison dans l'évaluation de la performance de nos méthodes proposées d'estimation fondée sur le lissage exponentiel, présentées plus tard à la Section 3.3.

3.2.1.1 SES-f : α_k fixe

Dans cette méthode de base, le paramètre α_k est fixé. Il peut être déterminé par ajustement de modèle (*model fitting*) sur la base des données historiques de la série temporelle. Une

valeur réduite de α_k amortira efficacement les variations aléatoires de $\tilde{\lambda}_k$ et procurera une meilleure stabilité à l'estimation. Par contre, une valeur plus élevée de α_k permettra au lissage exponentiel de mieux suivre les changements de tendance et ainsi de présenter une meilleure réponse de l'estimation (Section 2.6.3.1).

3.2.1.2 SES-a : α_k adaptatif selon AEES

Dans cette méthode, α_k est réajusté à chaque instant d'estimation. Nous choisissons d'utiliser la formule d'adaptation proposée pour AEES par (Mentzer, 1988) à cause de sa meilleure précision. L'adaptation est fondée sur la différence relative entre les valeurs estimée et mesurée :

$$\alpha_k = \left| (\hat{\lambda}_{k-1} - \tilde{\lambda}_k) / \tilde{\lambda}_k \right|. \tag{3.5}$$

3.2.1.3 DES-a1 : α_k adaptatif selon AEES, γ_k fixe

Ce lissage exponentiel double, (3.3) et (3.4), tient compte explicitement de la tendance de la variable estimée. La tendance est lissée exponentiellement avec un paramètre γ_k fixe déterminé par ajustement de modèle. Le lissage de la variable estimée est adaptatif avec α_k déterminé par (3.5).

3.2.1.4 SES-a2 : α_k , γ_k selon AEES-C

Dans cette variante, α_k est toujours adapté par (3.5), et en plus γ_k est aussi adapté. Cette dernière adaptation est fondée sur le changement relatif de la tendance estimée depuis la dernière estimation, comme proposée dans l'approche AEES-C :

$$\gamma_k = \left| (\hat{\Delta}\lambda_{k-1} - \hat{\Delta}\lambda_{k-2}) / \hat{\Delta}\lambda_{k-2} \right| \tag{3.6}$$

(Mentzer et Gomes, 1994) ont rapporté que, bien que les comparaisons de précision des résultats obtenus avec AEES et AEES-C n'étaient pas consistantes, les résultats de AEES-C ont été supérieurs dans 10 des 14 cas d'essais cités.

3.2.2 Utilisation du filtre de Kalman

Le filtre de Kalman a déjà été utilisé dans divers problèmes de réseaux pour obtenir des estimés optimaux. (Kolarov, Atai et Hui, 1994) l'ont appliqué aux réseaux ATM, pour minimiser la moyenne quadratique de l'erreur dans l'estimation du nombre de sources de trafic actives dans les chemins virtuels. Dans (Anjali, Scoglio et Uhl, 2003), le filtre est appliqué pour estimer l'utilisation d'un lien inter-domaine par le trafic. Dans ce cas, une série temporelle de mesures instantanées de charges de trafic est utilisée pour l'estimation. Dans (Dziong, 1997), une approche d'attribution de ressources au trafic aggloméré, fondée sur le filtre de Kalman, a été présentée. Pour garder les violations de QoS à un niveau acceptable, l'approche proposée ajoute une largeur de bande additionnelle, réservée à l'erreur d'estimation, à la bande attribuée au trafic estimé.

3.3 Estimation de tendance pour lissage exponentiel adaptatif

Dans cette section, nous proposons deux nouvelles méthodes de lissage exponentiel simple, où l'adaptation du paramètre de lissage est fondée sur l'estimation de la tendance de la variable estimée (dans notre cas, la demande de trafic). L'objectif de cette approche adaptative fondée sur la tendance est d'améliorer la précision de l'estimation du trafic dans un environnement où à la fois des périodes stationnaires et non-stationnaires de trafic coexistent. Le modèle d'estimation incluant les sous-modèles connectés de système de trafic, de mesures et de filtre SES est illustré à la Figure 3.1.

Le sous-modèle de système caractérise la demande de trafic aux liens du réseau. Des études de trafic Internet par (Thompson, Miller et Wilder, 1997) ont révélé que, dans la majorité des traces de trafic relevées, le volume des flots ainsi que des données sont non-stationnaires. Ces

volumes suivent des gabarits bien établis sur 24 heures, correspondant aux activités des utilisateurs, qui se répètent quotidiennement durant la semaine. Par conséquent, notre système de trafic sera modelé en cycles de trafic de 24 heures, chaque cycle étant formé d'instants discrets *k=1, 2, ..., K*, où *K=24heures*/ ΔT . À chaque instant *k*, la variation du taux d'arrivée réel λ_k , due à la dynamique de la demande par rapport à *k-1*, est représentée par la somme de sa tendance $\Delta\lambda_k$ et d'une déviation aléatoire désignée par e_k, comme formulée par (3.1). Nous présumons que la variable e_k présente une distribution Gaussienne de moyenne zéro et de variance $v_{e,k}$.

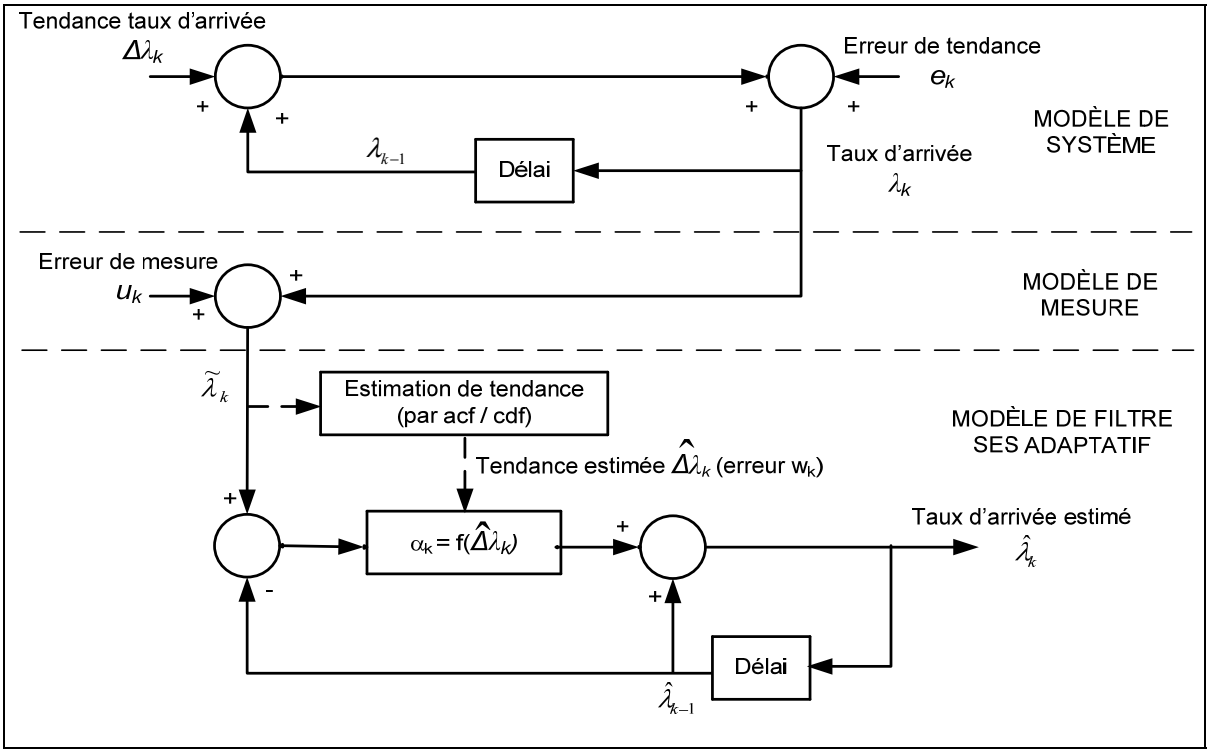

Figure 3.1 Modèle d'estimation SES fondée sur la tendance.

Dans le sous-modèle de mesure, le nombre d'arrivées $\tilde{x}_k$ dans chaque intervalle de temps régulier ΔT est échantillonné pour produire le taux d'arrivée mesuré $\tilde{\lambda}_k = \tilde{x}_k / \Delta T$. Ces taux sont affectés par les erreurs de mesure u_k qui sont présumées de distribution Gaussiennes de moyenne zéro et de variance $\tilde{v}_{e,k}$. Ainsi, la mesure du taux d'arrivée au temps *k* peut s'exprimer comme :

$$\tilde{\lambda}_k = \lambda_k + u_k . \tag{3.7}$$

Notons maintenant que l'expression du taux d'arrivée estimé par SES (3.2) peut aussi être reformulée comme suit :

$$\hat{\lambda}_k = \hat{\lambda}_{k-1} + \alpha_k \left(\tilde{\lambda}_k - \hat{\lambda}_{k-1} \right), \tag{3.8}$$

ce qui correspond à la dynamique représentée par les lignes dirigées en continu dans le sous-modèle du filtre SES de la Figure 3.1. Considérant (3.2) et (3.7) dans (3.8) et introduisant l'erreur d'estimation de la tendance w_k, ($\hat{\Delta}\lambda_k = \Delta\lambda_k + w_k$), nous obtenons :

$$\hat{\lambda}_k = \hat{\lambda}_{k-1} + \alpha_k (\Delta\lambda_k + e_k + u_k + w_k) . \tag{3.9}$$

Cette expression mène aux recommandations suivantes dans le choix des valeurs de α_k :

- Dans les périodes de trafic stationnaire, nous avons $\Delta\lambda_k \cong 0$ (tendance nulle). Nous pouvons alors assigner une valeur réduite à α_k dans ces périodes, afin d'atténuer les erreurs de système et de mesure et ainsi réaliser une bonne stabilité de l'estimation. Dans ce cas, la valeur réduite de α_k ne nuit pas à la sensibilité de l'estimation au changement du taux d'arrivée puisque la tendance est négligeable.
- Quand la tendance (positive ou négative) s'intensifie, $\Delta\lambda_k$ devient plus important pour l'estimation et la valeur de α_k doit augmenter de pair avec l'intensification de la tendance. Ceci procure une bonne sensibilité au changement dynamique du taux d'arrivée sans pour autant nuire à la stabilité, puisque la tendance devient l'élément dominant du facteur de α_k dans (3.9).

L'argumentation ci-dessus mène à la conclusion que α_k devrait être une fonction de la tendance, cette dernière pouvant être estimée en se basant sur des mesures du taux d'arrivée, comme illustré à la Figure 3.1. Pour réaliser cette approche, nous proposons ci-après deux méthodes d'estimation de tendance, fondées respectivement sur les fonctions d'autocorrélation (acf) et de distribution cumulée (cdf) des taux d'arrivées de connexions.

3.3.1 Estimation de tendance par fonction d'autocorrélation

La tendance dans une série temporelle, formée par *M* mesures de taux d'arrivée à des instants successifs aboutissant à l'instant *k* (instants *k-M+1* à *k*), peut être indiquée par le coefficient d'autocorrélation à décalage 1 de la série, $\rho_k(1)$. Comme seulement le décalage 1 est considéré dans cette méthode proposée, nous abandonnons le paramètre de décalage de la notation pour la simplifier. La moyenne $\breve{\lambda}_k$ des M mesures est donnée par :

$$\breve{\lambda}_k = \frac{1}{M}\sum_{m=k-M+1}^{k} \tilde{\lambda}_m \ . \tag{3.10}$$

et le coefficient d'autocorrélation est obtenu par :

$$\rho_k = \frac{\sum_{m=k-M+1}^{k-1}(\tilde{\lambda}_m - \breve{\lambda}_k)(\tilde{\lambda}_{m+1} - \breve{\lambda}_k)}{\sum_{m=k-M+1}^{k}(\tilde{\lambda}_m - \breve{\lambda}_k)^2}. \tag{3.11}$$

Avec *M* suffisant, le dénominateur de ρ_k est positif dû aux variations stochastiques de $\tilde{\lambda}_m$. Quand la série temporelle présente une tendance, on peut montrer que ρ_k est fonction croissante de l'amplitude de la tendance $|\Delta\lambda_k|$, comme illustré à la Figure 3.2.

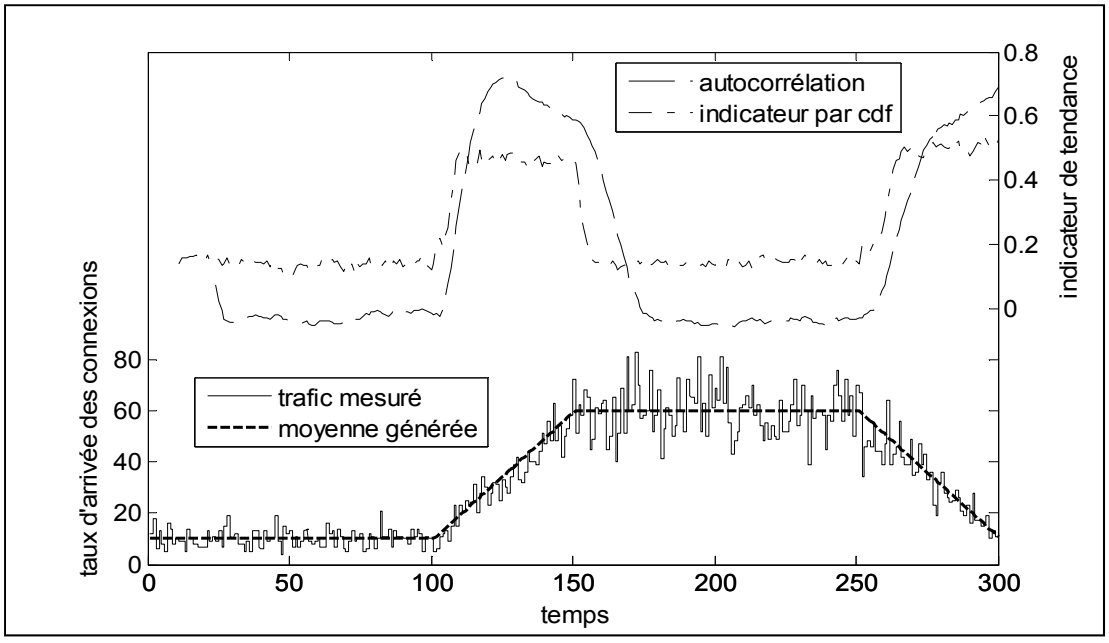

Figure 3.2 Estimation de tendance fondée sur acf et cdf.

3.3.2 Estimation de tendance par fonction de distribution cumulée

Pour réaliser cette méthode, nous définissons la probabilité suivante, fondée sur la fonction de distribution cumulée des taux d'arrivée de connexions :

$$p_k(\tilde{\lambda}_k) = \Pr\left(x > \tilde{\lambda}_k \mid \tilde{\lambda}_k > \hat{\lambda}_{k-1}\right) = \frac{1 - \Pr\left(x \le \tilde{\lambda}_k\right)}{1 - \Pr\left(x \le \hat{\lambda}_{k-1}\right)} = \frac{1 - cdf(\tilde{\lambda}_k)}{1 - cdf(\hat{\lambda}_{k-1})}, \tag{3.12}$$

dans le cas où $\tilde{\lambda}_k > \hat{\lambda}_{k-1}$, et

$$p_k(\tilde{\lambda}_k) = \Pr\left(x \le \tilde{\lambda}_k \mid \tilde{\lambda}_k \le \hat{\lambda}_{k-1}\right) = cdf(\tilde{\lambda}_k) / cdf(\hat{\lambda}_{k-1}), \tag{3.13}$$

dans le cas contraire où $\tilde{\lambda}_k \le \hat{\lambda}_{k-1}$.

En présumant que le processus d'arrivée des connexions est Poissonien, les taux d'arrivées mesurés dans un processus stationnaire tendent à se concentrer près de la moyenne, estimée ici par $\hat{\lambda}_{k-1}$. La probabilité $p_k(\tilde{\lambda}_k)$, définie ci-dessus, tend vers 1 quand $\tilde{\lambda}_k$ tend vers $\hat{\lambda}_{k-1}$ et elle diminue quand $\tilde{\lambda}_k$ s'éloigne de $\hat{\lambda}_{k-1}$. Une probabilité p_k plus élevée correspond alors à une plus grande probabilité que le processus d'arrivée est stationnaire, et une valeur plus élevée de son complément $1\text{-}p_k$ correspond à une plus grande probabilité d'une tendance non-nulle du taux d'arrivée. Nous pouvons donc utiliser $1\text{-}p_k$ comme un indicateur de tendance.

Le domaine des deux indicateurs de tendance fondés sur *acf* et *cdf*, respectivement ρ_k et $1\text{-}p_k$, est [0, 1]. Les mesures en ligne de ces indicateurs, prises dans un lien soumis à un profil de trafic incluant des périodes stationnaires et tendancielles, sont illustrées à la Figure 3.2. Nous pouvons bien y constater que les deux indicateurs se déplacent vers des niveaux élevés quand le trafic transitionne d'une période stationnaire à une période tendancielle, et vice versa.

3.3.3 Adaptation du coefficient de lissage fondée sur l'estimation de tendance

Disposant d'estimés de la tendance de trafic, nous devons maintenant choisir une fonction qui adaptera le coefficient de lissage α_k aux changements de l'indicateur de tendance, représenté par ρ_k (méthode appelée *SES-acf*) ou $1\text{-}p_k$ (méthode *SES-cdf*), et désigné ici de façon commune par θ_k. Dans le domaine des valeurs de l'indicateur, une fonction lentement croissante de θ_k, telle que $\alpha_k = \theta_k$ ou une fonction exponentielle, entraine des réponses lentes aux transitions entre les périodes stationnaire et tendancielle de trafic. Avec le dual objectif de stabilité et de rapidité de réaction selon le cas de trafic, un changement plus rapide de α_k est requis lors de ces transitions.

La fonction logistique peut bien répondre à l'exigence mentionnée car sa pente et son point d'inflexion peuvent être facilement contrôlés par les paramètres de la fonction. Suite à nos expériences, nous proposons d'utiliser la fonction logistique suivante :

$$\alpha_k = 0.05 + logistic(\theta_k) = 0.05 + \left[0.85 / \left(1 + l_a e^{-l_b \theta_k}\right)\right], \tag{3.14}$$

où l_a et l_b sont les paramètres de contrôle de la fonction logistique. Le coefficient de lissage réalisé par (3.14) présente des asymptotes à 0,05 et 0,90, et ses paramètres l_a et l_b peuvent être choisis pour obtenir son point d'inflexion désiré, $\left(\frac{\ln(l_a)}{l_b}, 0.05 + \frac{0.85}{2}\right)$, et par conséquent sa pente en se basant sur les données historiques de l'indicateur de tendance.

Par exemple, les fonctions logistiques utilisées avec les indicateurs de tendance montrés à la Figure 3.2 sont tracées à la Figure 3.3. Nous constatons bien que les fonctions présentent des niveaux distincts de α_k pour les valeurs basses et les valeurs élevées de l'indicateur de tendance, avec une transition rapide entre les niveaux. Aussi, les points de début d'augmentation de α_k sont différents pour les cas de l'indicateur *acf* et de l'indicateur *cdf*, pour ainsi correspondre aux caractéristiques de l'indicateur concerné.

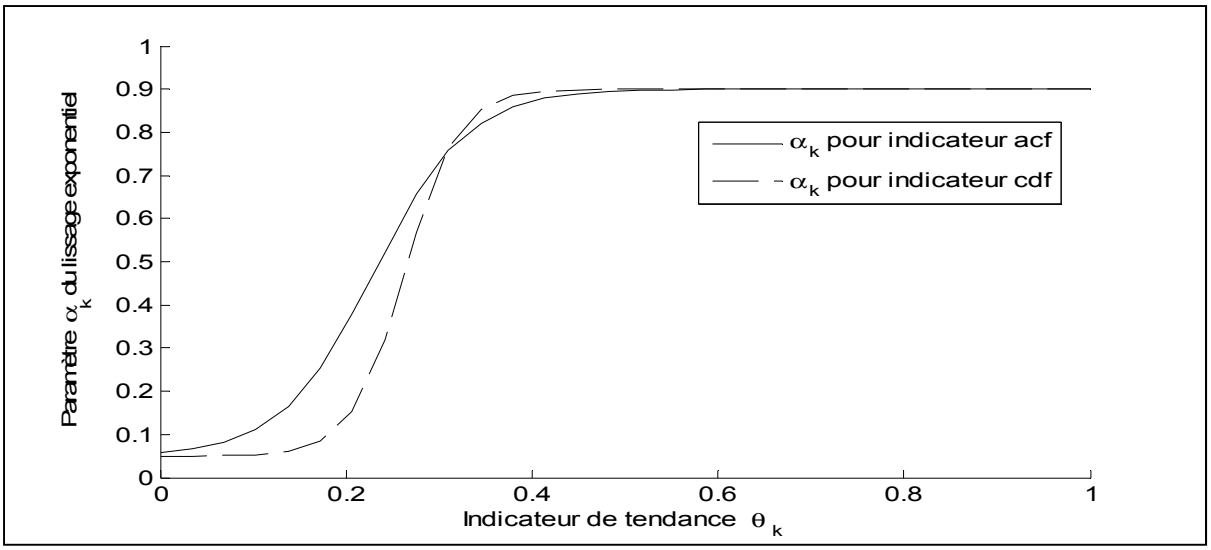

Figure 3.3 Paramètre α_k comme fonction logistique de la tendance.

3.4 Approche fondée sur le filtre de Kalman

Toute méthode d'estimation produit des valeurs estimées qui incluent des déviations statistiques des valeurs réelles. À cause de ces erreurs, les probabilités de violation des contraintes de GoS du réseau peuvent être élevées si les capacités des liens étaient attribuées uniquement sur la base des valeurs estimées. Pour atténuer ces probabilités de violations, nous proposons une application du filtre discret de Kalman qui procure, concurremment avec la valeur estimée, une évaluation de la distribution de l'erreur d'estimation. Dans cette application proposée du filtre, les paramètres du modèle de tendance de trafic sont établis à l'aide de mesures historiques de trafic. Dans la suite de cette section, nous présentons le modèle proposé du filtre de Kalman, avec son intégration au modèle d'adaptation de capacité (Section 2.4.2) dans le but de limiter les violations de GoS du réseau.

3.4.1 Modèle proposé de filtre de Kalman

Dans cette sous-section, nous donnons une description succincte de l'algorithme du filtre de Kalman, suivie de détails sur la détermination des paramètres requis pour la réalisation du modèle proposé.

3.4.1.1 Algorithme du filtre

La dynamique du modèle d'estimation de la demande de trafic fondé sur le filtre de Kalman est montrée à la Figure 3.4. Les valeurs mesurées de trafic $\tilde{\lambda}_k$ sont procurées au filtre à travers les mêmes sous-modèles de système (3.1) et de mesure (3.7) de trafic décrits auparavant pour le modèle d'estimation par lissage exponentiel (Section 3.3, Figure 3.1). À chaque instant *k*, le filtre produit un estimé $\hat{\lambda}_k$ du taux d'arrivée des connexions λ_k. La valeur estimée est donnée par :

$$\hat{\lambda}_k = \hat{\lambda}_k^e + K_k(\tilde{\lambda}_k - \hat{\lambda}_k^e), \tag{3.15}$$

où K_k est le gain du filtre et $\hat{\lambda}_k^e$ est la valeur extrapolée de l'estimée obtenue par :

$$\hat{\lambda}_k^e = Max\left(\hat{\lambda}_{k-1} + \hat{\Delta}\lambda_k, 0\right), \tag{3.16}$$

où $\hat{\Delta}\lambda_k$ représente l'estimé de la tendance qui doit être fourni au filtre. Le gain K_k est défini par :

$$K_k = P_k^e / (P_k^e + \tilde{v}_{e,k}), \tag{3.17}$$

où $\tilde{v}_{e,k}$ est la variance de l'erreur de mesure u_k, et P_k^e est la variance projetée de l'erreur d'estimation du taux d'arrivée donnée par :

$$P_k^e = P_{k-1} + \hat{v}_{e,k}, \tag{3.18}$$

où P_{k-1} est la variance de l'erreur d'estimation du taux d'arrivée à l'instant *k-1*, et $\hat{v}_{e,k}$ représente la variance de l'erreur d'estimation de la tendance $\hat{\Delta}\lambda_k$, qui doit aussi être fournie au filtre. Enfin, P_{k-1} est donné par :

$$P_{k-1} = (1 - K_{k-1})P_{k-1}^e. \tag{3.19}$$

Ainsi, pour obtenir les estimés du taux d'arrivée $\hat{\lambda}_k$ et de son erreur d'estimation P_k, la séquence (3.19), (3.18), (3.17), (3.16) est exécutée à chaque instant *k*. Comme indiqué ci-dessus, l'exécution de la séquence nécessite la disponibilité des estimés de $\hat{\Delta}\lambda_k$, $\hat{v}_{e,k}$ et $\tilde{v}_{e,k}$, en plus des mesures $\tilde{\lambda}_k$ définies auparavant.

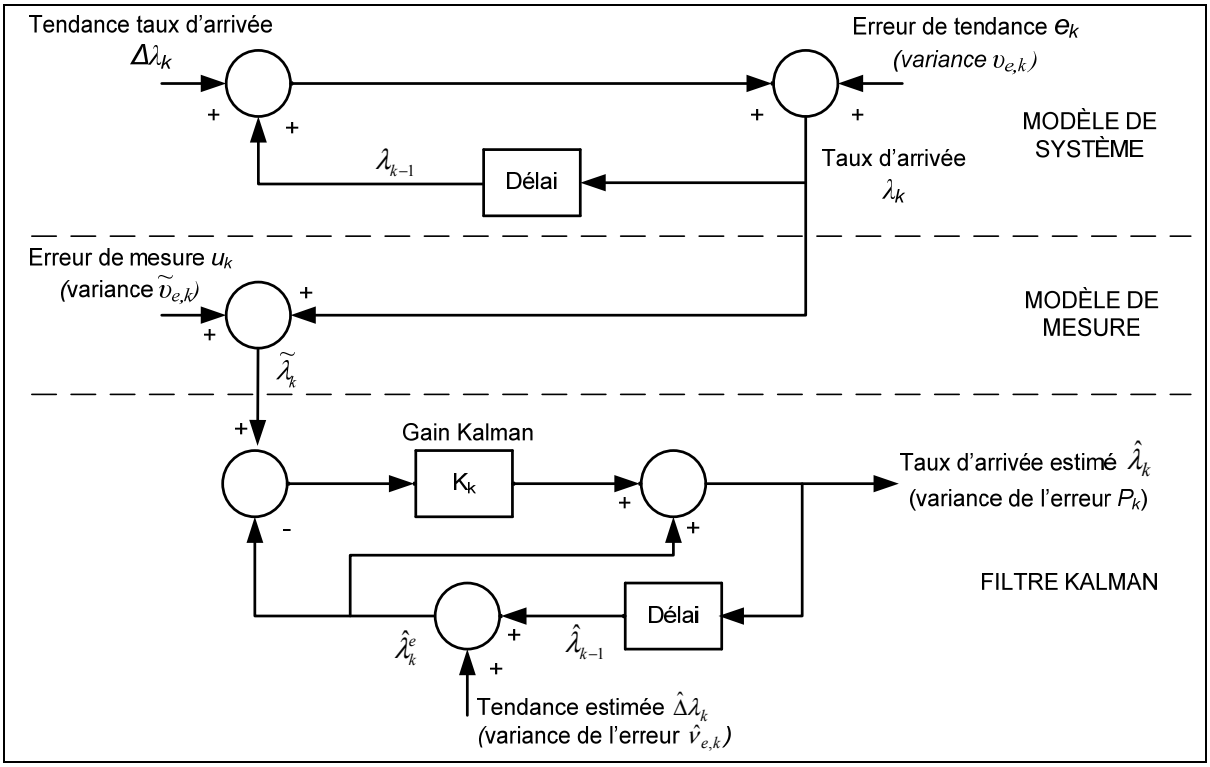

Figure 3.4 Modèle d'estimation fondée sur filtre de Kalman.

3.4.1.2 Détermination des paramètres d'entrée du filtre

3.4.1.2.1 Estimation de paramètres du sous-modèle de système

Les estimés de la tendance du taux d'arrivée $\hat{\Delta}\lambda_k$ et de la variance de son erreur $\hat{v}_{e,k}$, provenant du sous-modèle de système, sont requis pour l'exécution de l'algorithme du filtre. Ces paramètres peuvent être estimés à partir de statistiques historiques des taux d'arrivée de connexions aux liens, obtenues sur une période de *D* jours.

Soit $\hat{\lambda}_{k,d}$ l'estimé du taux d'arrivée recueilli au temps *k* du jour *d*, et soit $\hat{\delta}_{k,d}$ le changement correspondant du taux dans la période de *k-1* à *k* :

$$\hat{\delta}_{k,d} = \hat{\lambda}_{k,d} - \hat{\lambda}_{k-1,d} . \tag{3.20}$$

Afin d'obtenir une meilleure stabilité des valeurs estimées, nous effectuons un filtrage des fluctuations stochastiques de $\hat{\delta}_{k,d}$ en lui appliquant une moyenne mobile calculée sur des fenêtres de taille M_F. L'estimé de la tendance au temps k du jour d est ainsi produit :

$$\hat{\Delta}\lambda_{k,d} = \frac{1}{M_F} \sum_{m=k-M_F/2}^{k+M_F/2} \hat{\delta}_{m,d} \,. \tag{3.21}$$

L'estimé de la tendance au temps k, $\hat{\Delta}\lambda_k$, est alors obtenu pour le filtre de Kalman en effectuant la moyenne des $\hat{\Delta}\lambda_{k,d}$ recueillis sur D jours :

$$\hat{\Delta}\lambda_k = \frac{1}{D} \sum_{d=1}^{D} \hat{\Delta}\lambda_{k,d}, \quad k = 1,2,..,K. \tag{3.22}$$

Les erreurs d'estimation de la tendance au temps k, aux différents jours d, sont $\hat{\Delta}\lambda_{k,d} - \hat{\Delta}\lambda_k$. En présumant que la distribution des erreurs a une moyenne nulle, nous pouvons obtenir la variance de ces erreurs pour utilisation dans le filtre avec :

$$\hat{v}_{e,k} = \frac{1}{D} \sum_{d=1}^{D} \left(\hat{\Delta}\lambda_{k,d} - \hat{\Delta}\lambda_k \right)^2, \quad k = 1,2,...,K. \tag{3.23}$$

3.4.1.2.2 Estimation de paramètres du sous-modèle de mesure

Les paramètres requis par le sous-modèle de mesure sont les mesures du taux d'arrivée $\tilde{\lambda}_k$ et l'estimé $\tilde{v}_{e,k}$ de la variance de l'erreur de mesure.

Comme définie à la Section 3.3, les mesures du taux d'arrivée sont données par $\tilde{\lambda}_k = \tilde{x}_k / \Delta T$, où $\tilde{x}_k$ représente la mesure du nombre d'arrivées de connexions durant l'intervalle de mesure ΔT. $\tilde{x}_k$ est présumé de distribution Poissonnienne.

L'application de la propriété de la variance d'une distribution Poissonnienne donne :

$$Var(\tilde{x}_k) = \lambda_k \Delta T \,, \tag{3.24}$$

et par conséquent :

$$\tilde{v}_{e,k} = E[u_k - E(u_k)]^2 = Var(\tilde{\lambda}_k) = Var\left(\frac{\tilde{x}_k}{\Delta T}\right) = \frac{1}{(\Delta T)^2} Var(\tilde{x}_k) = \frac{\lambda_k}{\Delta T}, \qquad (3.25)$$

où $u_k = \tilde{\lambda}_k - \lambda_k$. Comme le taux réel λ_k est inconnu, nous l'approximons par la valeur extrapolée du taux estimé $\hat{\lambda}_k^e$, ce qui mène à l'expression de la variance de l'erreur de mesure :

$$\tilde{v}_{e,k} = \frac{\hat{\lambda}_k^e}{\Delta T} = \frac{\hat{\lambda}_{k-1} + \hat{\Delta}\lambda_k}{\Delta T}. \qquad (3.26)$$

3.4.2 Application au degré de service du réseau

Afin de garantir la Qualité de Service, les nouvelles connexions ne sont admises que lorsqu'assez de largeur de bande résiduelle est disponible, sinon elles sont bloquées. Pour présenter un degré de service acceptable aux usagers, le taux de blocage des connexions doit être limité à des contraintes choisies. Dans notre approche d'adaptation de capacité et de routage pour la maximisation du bénéfice présentée au CHAPITRE 2, la probabilité de blocage du réseau B_T est gardée inférieure à sa contrainte B_T^C par l'adaptation des paramètres de récompense de connexion r_j (Section 2.4.3).

Soit $B_{k,T}$ une variable représentant le taux de blocage du réseau dans l'intervalle de temps *k-1* à *k*. Comme mentionné, l'estimation de la demande de trafic à chaque instant *k*, servant à l'adaptation de capacité, est affectée par une erreur e_k qui entraine une déviation de la demande réelle. Avec cette déviation, la probabilité que $B_{k,T}$ viole sa contrainte de GoS, B_T^C, peut être élevée. Pour améliorer le service, une contrainte à la probabilité de violation de GoS d'intervalle, ε, peut être spécifiée :

$$\Pr\{B_{k,T} > B_T^C\} \le \varepsilon, \quad k = 1,2,.... \qquad (3.27)$$

Comme l'estimation de trafic est effectuée à chaque lien, décomposons maintenant la contrainte (3.27) en contraintes de lien. Présumant un trafic non nul, le taux de blocage du réseau $B_{k,T}$ est déterminé par les taux de blocage $B_{k,j}$ des différentes classes de connexions *j* :

$$B_{k,T} = \frac{\sum_j \lambda_{k,j} B_{k,j}}{\sum_j \lambda_{k,j}}. \tag{3.28}$$

Comme dans la Section 2.4.3, nous utilisons le modèle de performance à un moment fondé sur l'hypothèse d'indépendance de liens pour simplifier les calculs et ainsi arriver à l'approximation de $B_{k,j}$ similaire à (2.49) :

$$B_{k,j} \cong \prod_{i \in \mathbf{W}_j} B_{k,i} = \prod_{i \in \mathbf{W}_j} [1 - \prod_{s \in \mathbf{S}^i} (1 - B_{k,s})], \tag{3.29}$$

où $\boldsymbol{W}_j$ est l'ensemble des chemins de la classe j, $\boldsymbol{S}^k$ est l'ensemble des liens du chemin i et $B_{k,i}$ et $B_{k,s}$ sont respectivement les taux de blocage du chemin i et du lien s. En combinant (3.28) et (3.29), la relation entre $B_{k,T}$ et $B_{k,s}$, $s=1,2,...,S$, peut être approximativement déterminée. Alors (3.27) peut être distribuée de manière heuristique dans un ensemble de spécifications indépendantes de violations de GoS aux liens :

$$\Pr\{B_{k,s} > B_s^C\} \le \varepsilon, \quad k = 1,2,...; s = 1,2,..., S. \tag{3.30}$$

Dans la suite de cette section, nous abandonnons l'indice s de la notation pour la simplifier. D'après le modèle du filtre de Kalman, le taux d'arrivée λ_k peut être exprimé en termes de son estimé $\hat{\lambda}_k$ et de l'erreur d'estimation $\hat{e}_k$ par $\lambda_k = \hat{\lambda}_k + \hat{e}_k$, où $\hat{e}_k$ est de distribution Gaussienne de moyenne zéro et de variance P_k (3.19). En se basant sur P_k, nous pouvons alors facilement déterminer, à chaque instant k, la valeur L_k qui satisfait à la condition suivante :

$$\Pr\{\lambda_k > \hat{\lambda}_k + L_k\} = \varepsilon. \tag{3.31}$$

Le paramètre L_k est maintenant intégré comme suit dans la procédure d'adaptation de capacité pour assurer un niveau élevé de conformité aux contraintes de GoS. En premier lieu, la contrainte de blocage $B_k \le B^C$ est vérifiée en utilisant B_k donné par l'expression :

$$B_k = E_b(\frac{\hat{\lambda}_k + L_k}{\mu_k}, N_k) \tag{3.32}$$

où E_b indique la formule d'Erlang B de probabilité de blocage, μ_k est le taux de service moyen des connexions et N_k est la capacité de lien optimale. N_k a été déterminée par l'adaptation de capacité sur la base de la demande de trafic estimée $\hat{\lambda}_k$. Notons que l'évaluation du taux de blocage (3.32) a pris en considération la possibilité d'un excédant de trafic L_k sur la demande de trafic estimée. Si la contrainte de blocage est satisfaite, la capacité N_k réalise à la fois un bénéfice maximisé et la conformité à la GoS.

Dans le cas de violation de la contrainte, la capacité du lien est augmentée à N_k^{GoS} calculée avec :

$$E_b(\frac{\hat{\lambda}_k + L_k}{\mu_k}, N_k^{GoS}) = B^C, \tag{3.33}$$

qui permettra de se conformer à la spécification de GoS (3.30).

3.5 Analyse de performance

Dans cette section, la performance de nos méthodes d'estimation de demande de trafic, fondées sur le lissage exponentiel adapté à la tendance (Section 3.3) et sur le filtre de Kalman (Section 3.4), est évaluée dans des conditions diverses de niveaux et de tendances de trafic. Les critères d'évaluation considérés ici sont la stabilité de l'estimation en périodes de trafic stationnaire, et la réponse de l'estimation en périodes de tendance de trafic. La formulation des métriques pour ces critères est donnée dans des sous-sections ci-après. L'analyse est effectuée en appliquant les estimations sur un lien unique soumis à des scénarios de trafic incluant des périodes stationnaires et tendancielles. Dans chaque scénario, nos méthodes proposées sont comparées à des méthodes connues de lissage exponentiel sélectionnées à la Section 3.2.1.

L'analyse de la performance GoS obtenue avec le filtre de Kalman sera présentée au chapitre suivant, dans le cadre de l'évaluation finale du réseau où sont intégrées l'estimation de trafic et l'adaptation de capacité.

3.5.1 Métriques de performance

3.5.1.1 Stabilité d'estimation en trafic stationnaire

Dans les périodes de trafic stationnaire, la stabilité de l'estimation des taux d'arrivées est avantageuse car elle permet d'éviter des modifications inutiles au routage et à l'attribution de capacité des liens, qui pourraient nuire à la stabilité au réseau. La mesure de cette stabilité, désignée par σ_S, est définie par l'écart type du taux d'arrivée estimé $\hat{\lambda}_k$ par rapport au taux réel (généré par l'application d'essai) λ_k, normalisé par l'écart type du taux d'arrivée mesuré $\tilde{\lambda}_k$:

$$\sigma_S = \frac{\sqrt{\frac{1}{n_s}\sum_{k=1}^{n_s}(\hat{\lambda}_k - \lambda_k)^2}}{\sqrt{\frac{1}{n_s}\sum_{k=1}^{n_s}(\tilde{\lambda}_k - \lambda_k)^2}} = \sqrt{\frac{\sum_{k=1}^{n_s}(\hat{\lambda}_k - \lambda_k)^2}{\sum_{k=1}^{n_s}(\tilde{\lambda}_k - \lambda_k)^2}}, \tag{3.34}$$

où n_s représente le nombre d'échantillons prélevés durant la période de trafic stationnaire considérée. Avec n_s suffisant, le dénominateur de σ_S est non nul. Une valeur plus réduite de σ_S indique une déviation moindre, donc une meilleure stabilité.

3.5.1.2 Réponse de l'estimation en trafic tendanciel

Dans les périodes de trafic tendanciel, les valeurs estimées des taux d'arrivées devraient converger rapidement vers les valeurs réelles. Cette qualité de l'estimation est particulièrement importante dans les périodes de croissance de demande de trafic, car une réponse lente de l'estimation entraine un délai de réaction dans la gestion de réseau, causant des pertes de connexions et de données d'usagers. La mesure de la réponse de l'estimation, désignée par σ_C, est définie par le délai moyen des taux estimés $\hat{\lambda}_k$ par rapport aux taux réels λ_k (un délai positif indiquant que $\lambda_k > \hat{\lambda}_k$), normalisé par λ_k :

$$\sigma_C = \frac{1}{n_c}\sum_{k=1}^{n_c}(\lambda_k - \hat{\lambda}_k)/\lambda_k , \tag{3.35}$$

où n_c représente le nombre d'échantillons prélevés durant la période de trafic tendanciel considérée. Avec cette métrique, plus l'amplitude du délai σ_C est réduite, meilleure est la réponse de l'estimation.

3.5.2 Résultats de l'analyse

Les algorithmes d'estimation de trafic sélectionnés sont analysés sur un lien servant un trafic avec une trajectoire qui reflète des comportements réels rapportés dans la littérature. En général, les traces de trafic réel relevées montrent une répétition de gabarits journaliers qui incluent deux niveaux de trafic, élevé durant le jour et bas durant la nuit, avec une différence pouvant atteindre jusqu'à 300% - 500%. Les périodes de transition entre les niveaux peuvent s'étendre sur quelques heures (Thompson, Miller et Wilder, 1997). La Figure 3.5 montre des traces d'une journée de trafic de divers types, enregistrées à divers emplacements des États-Unis. Ces traces ainsi que d'autres trouvées sur le site Web WAND WITS (WAND Network Research Group, 2010, 27 juillet) confirment le gabarit journalier.

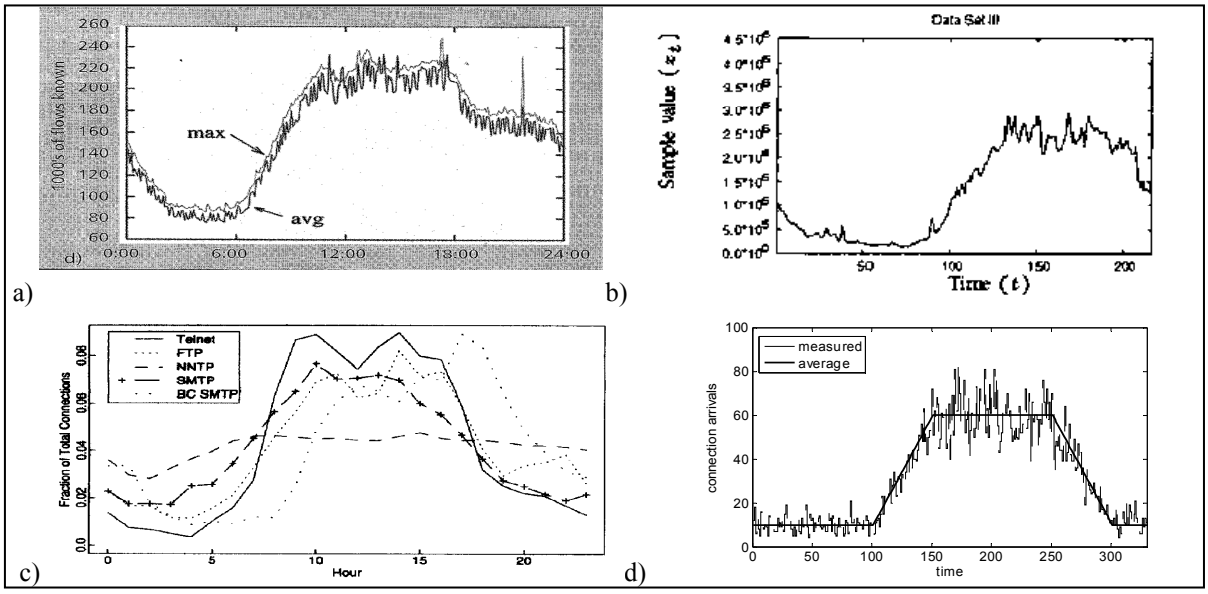

Figure 3.5 Traces d'une journée de trafic Internet :
a) volume de flots d'un lien, cote Est E.U., tiré de (Thompson, Miller et Wilder, 1997),
b) trafic de lien, Univ. Missouri-KansasCity, de (Krithikaivasan, Deka et Medhi, 2004),
c) taux d'arrivées de connexion au Lawrence Berkeley Lab, de (Paxson et Floyd, 1995),
d) modèle de simulation du gabarit des arrivées de connexion.

En nous inspirant de ces traces, nous construisons le scénario de trafic illustré à la partie d) de la Figure 3.5, qui inclut deux périodes stationnaires, une de bas et une de haut niveau de trafic, séparées par deux périodes tendancielles de transition. Dans la simulation, les taux d'arrivée de connexions λ_k suivent la trajectoire de ce scénario, et les échantillons de mesure $\tilde{\lambda}_k$ sont générés avec une distribution de Poisson définie par le taux d'arrivée λ_k.

Dans nos essais, chacune des périodes stationnaires et tendancielles comporte respectivement 100 et 50 échantillons (Figure 3.5, d). Le taux d'arrivée durant la période stationnaire initiale de bas trafic est fixé à 10, et trois cas de tendances sont examinés, $\Delta\lambda_k$ = 0,5, 1,0 et 1,5. Ces tendances mènent respectivement aux trois cas de taux de période stationnaire de haut trafic de 35, 60 et 85.

Pour notre méthode proposée *SES-acf*, le coefficient d'autocorrélation ρ_k (3.11) est calculé en utilisant une fenêtre de M=30 échantillons pour obtenir un résultat statistiquement valide. Les paramètres (l_a, l_b) des fonctions logistiques d'adaptation du coefficient de lissage (3.14) sont fixés à (100, 20) pour *SES-acf* et à (100000, 15) pour *SES-cdf*, afin de réaliser les courbes montrées à la Figure 3.3. Dans le cas de notre méthode fondée sur le filtre de Kalman, les paramètres du sous-modèle de système sont estimés sur la base de 30 cycles (jours) de données historiques de trafic.

Dans l'évaluation de la réponse de l'estimation, nous nous concentrons sur les tendances croissantes de trafic puisque la réponse à cette catégorie de tendances est critique dans la limitation des pertes de connexion. Les résultats de stabilité et de réponse d'estimation sont respectivement donnés aux Tableau 3.1 et Tableau 3.2, et la Figure 3.6 montre une représentation graphique de la performance combinée stabilité-réponse. Les cas des valeurs ($\Delta\lambda_k$; λ) de (0,5; 35), (1,0; 60) et (1,5; 85) sont montrés dans la figure. Pour la plupart des méthodes, le compromis de performance apparait clairement, comme une meilleure stabilité est compensée par une réponse plus lente et vice-versa. *SES-cdf* procure une bonne réponse mais sa stabilité demeure seulement moyenne. Les meilleures performances combinées sont

obtenues par la méthode proposée fondée sur le *filtre de Kalman*, suivie de *SES-acf*. Avec les répétitions journalières du gabarit de trafic, la meilleure performance du filtre peut s'expliquer par son exploitation de données historiques de plusieurs journées de trafic.

Tableau 3.1 Mesures de stabilité d'estimation de demande de trafic

	Stabilité d'estimation de trafic σ_S (confidence : 90%)			
Méthode	*λ = 10*	*λ = 35*	*λ = 60*	*λ = 85*
SES-f, α=0.1	0,225 ± 0,009	0,218 ± 0,004	0,218 ± 0,006	0,217 ± 0,006
SES-f, α=0.5	0,576 ± 0,004	0,570 ± 0,004	0,574 ± 0,006	0,572 ± 0,005
SES-a	0,713 ± 0,008	0,487 ± 0,014	0,399 ± 0,015	0,347 ± 0,013
SES-acf	**0,272 ± 0,018**	**0,261 ± 0,016**	**0,274 ± 0,013**	**0,269 ± 0,021**
SES-cdf	**0,495 ± 0,011**	**0,489 ± 0,012**	**0,494 ± 0,020**	**0,489 ± 0,009**
DES-a1	0,753 ± 0,008	0,549 ± 0,015	0,470 ± 0,019	0,421 ± 0,017
DES-a2	1,026 ± 0,014	0,885 ± 0,024	0,832 ± 0,024	0,819 ± 0,027
Kalman	**0,121 ± 0,008**	**0,109 ± 0,003**	**0,113 ± 0,005**	**0,113 ± 0,005**

Tableau 3.2 Mesures de réponse d'estimation de demande de trafic

	Réponse d'estimation de traffic σ_C (confidence : 90%)		
Méthode	*$\Delta\lambda_k$ = 0,5*	*$\Delta\lambda_k$ = 1,0*	*$\Delta\lambda_k$ = 1,5*
SES-f, α=0.1	0,161 ± 0,005	0,215 ± 0,004	0,247 ± 0,003
SES-f, α=0.5	0,022 ± 0,006	0,030 ± 0,005	0,037 ± 0,004
SES-a	0,069 ± 0,007	0,083 ± 0,005	0,097 ± 0,004
SES-acf	**0,048 ± 0,007**	**0,031 ± 0,005**	**0,026 ± 0,005**
SES-cdf	**0,031 ± 0,006**	**0,041 ± 0,005**	**0,051 ± 0,006**
DES-a1	0,037 ± 0,008	0,030 ± 0,007	0,029 ± 0,005
DES-a2	-0,034 ± 0,007	-0,034 ± 0,007	-0,025 ± 0,008
Kalman	**0,027 ± 0,004**	**0,015 ± 0,003**	**0,015 ± 0,003**

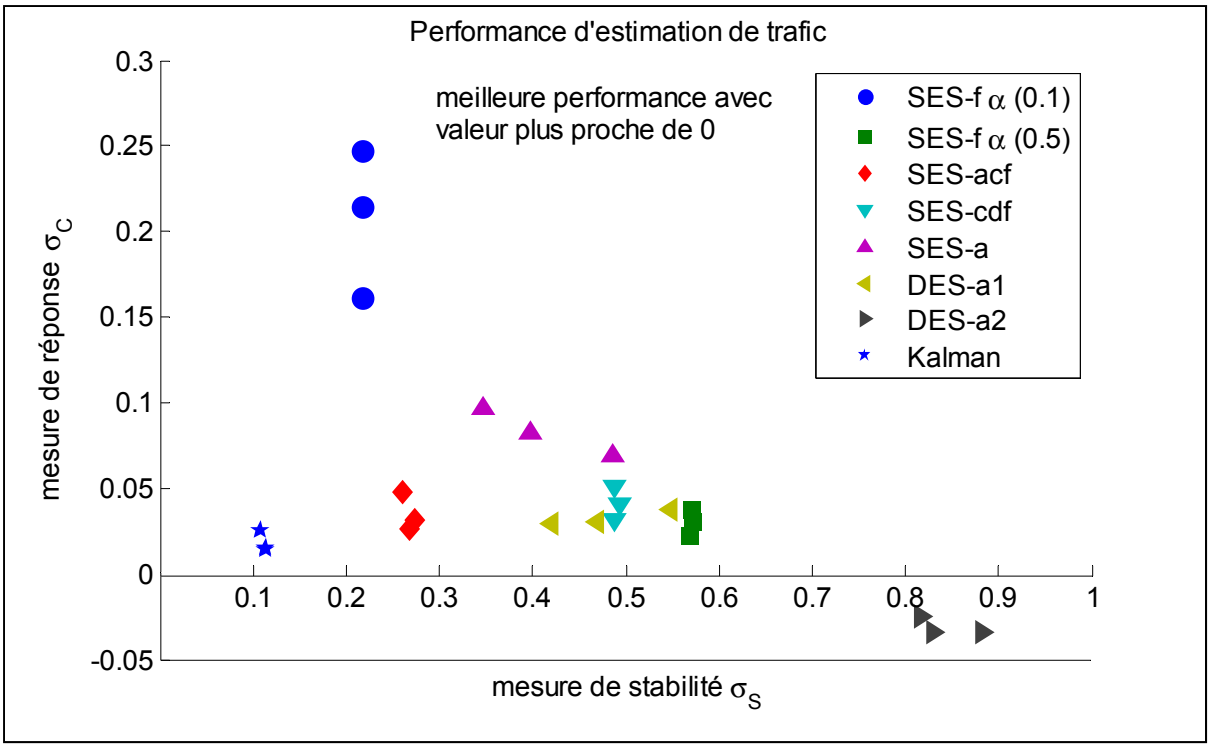

Figure 3.6 Performance combinée stabilité-réponse d'estimation de demande de trafic.

Pour illustrer la performance des estimations, les Figure 3.7 à Figure 3.10 montrent les traces des taux d'arrivées de connexions estimés et mesurés pour différentes méthodes d'estimation. *SES-f* avec $\alpha=0.1$ (Figure 3.7) démontre une très bonne stabilité d'estimation, cependant les délais durant les périodes de tendance croissante comme décroissante de trafic sont clairement visibles. La réponse de l'estimation par *DES-a2* (lissage adaptatif, Figure 3.8) aux tendances de trafic est très rapide, par contre sa stabilité en périodes stationnaires est médiocre. Les exemples de ces deux méthodes démontrent bien le nécessaire compromis stabilité-réponse des lissages exponentiels traditionnels. La méthode proposée *SES-acf* (Figure 3.9) offre un bon compromis, où une bonne réponse aux tendances est démontrée en même temps qu'une stabilité raisonnable en périodes de trafic stationnaire. Enfin, c'est notre approche par *filtre de Kalman* (Figure 3.10) qui procure la meilleure combinaison des performances démontrant à la fois les meilleures stabilité et réponse.

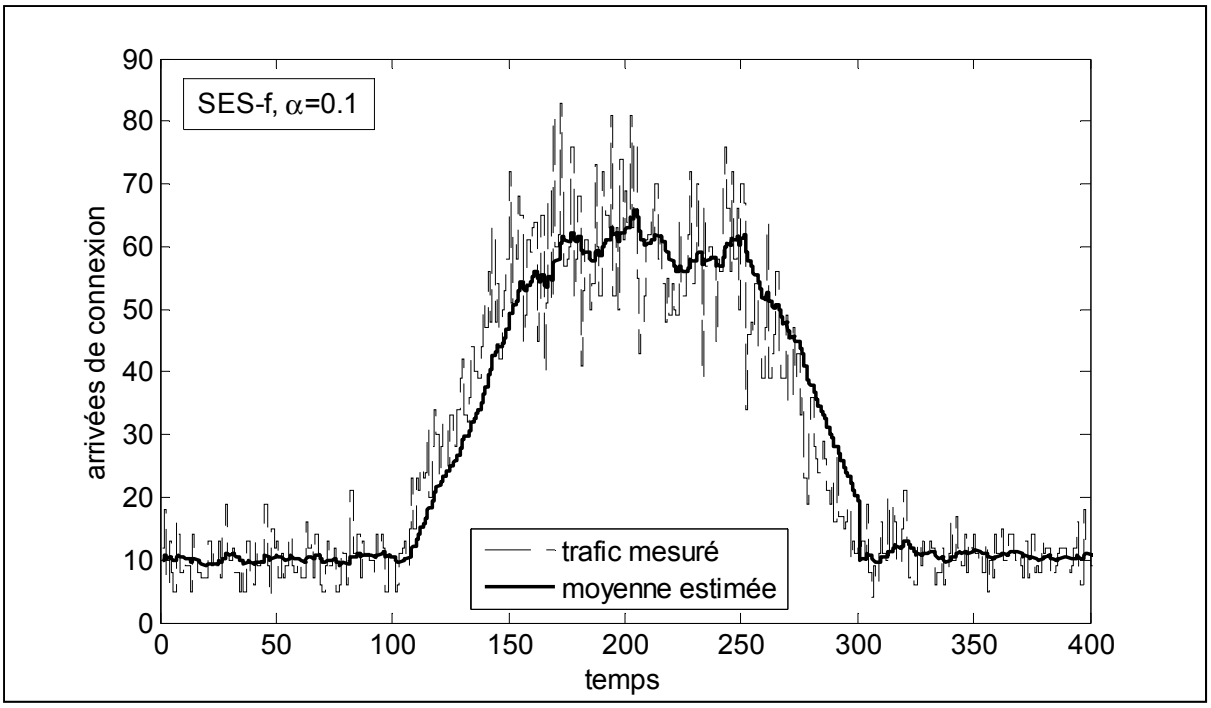

Figure 3.7 Estimation de trafic par méthode *SES-f*.

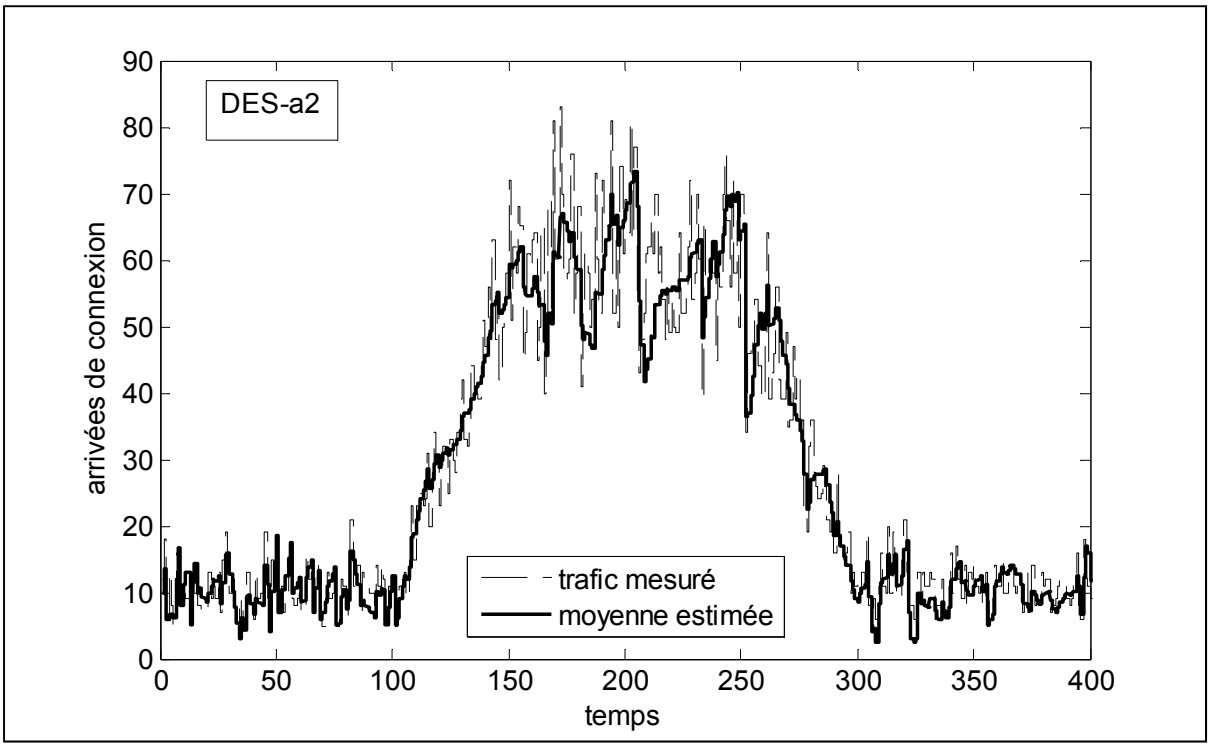

Figure 3.8 Estimation de trafic par méthode *DES-a2*.

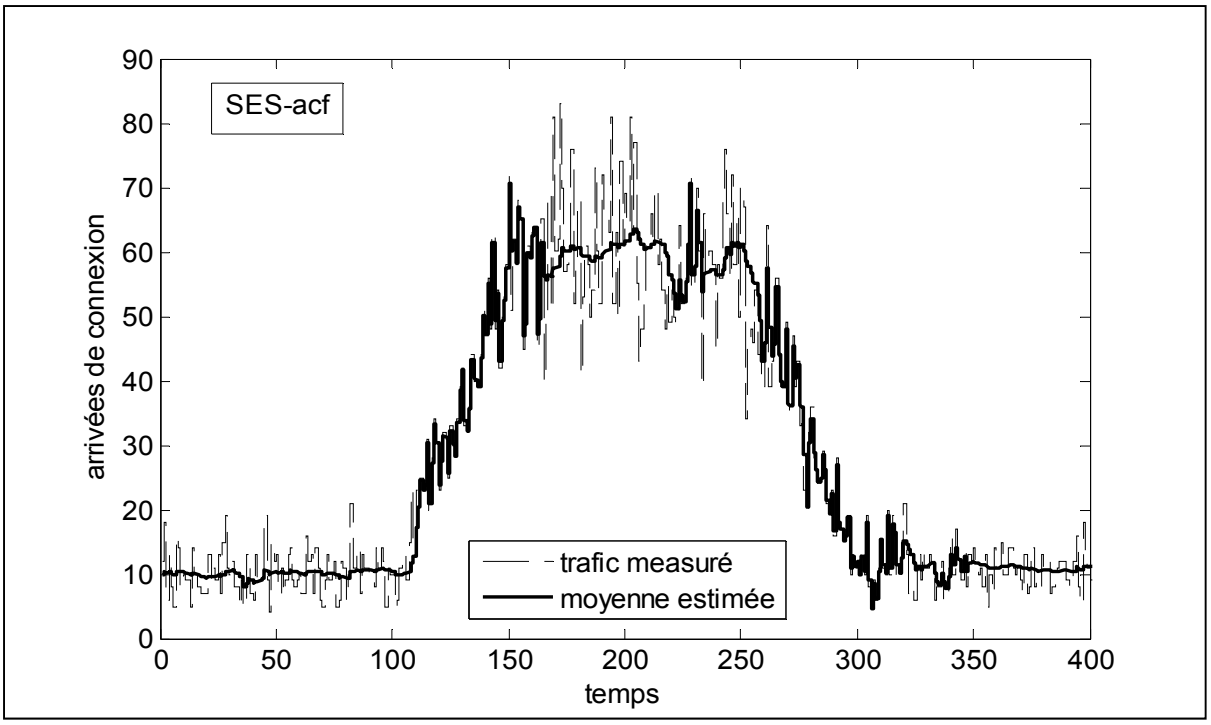

Figure 3.9 Estimation de trafic par méthode *SES-acf*.

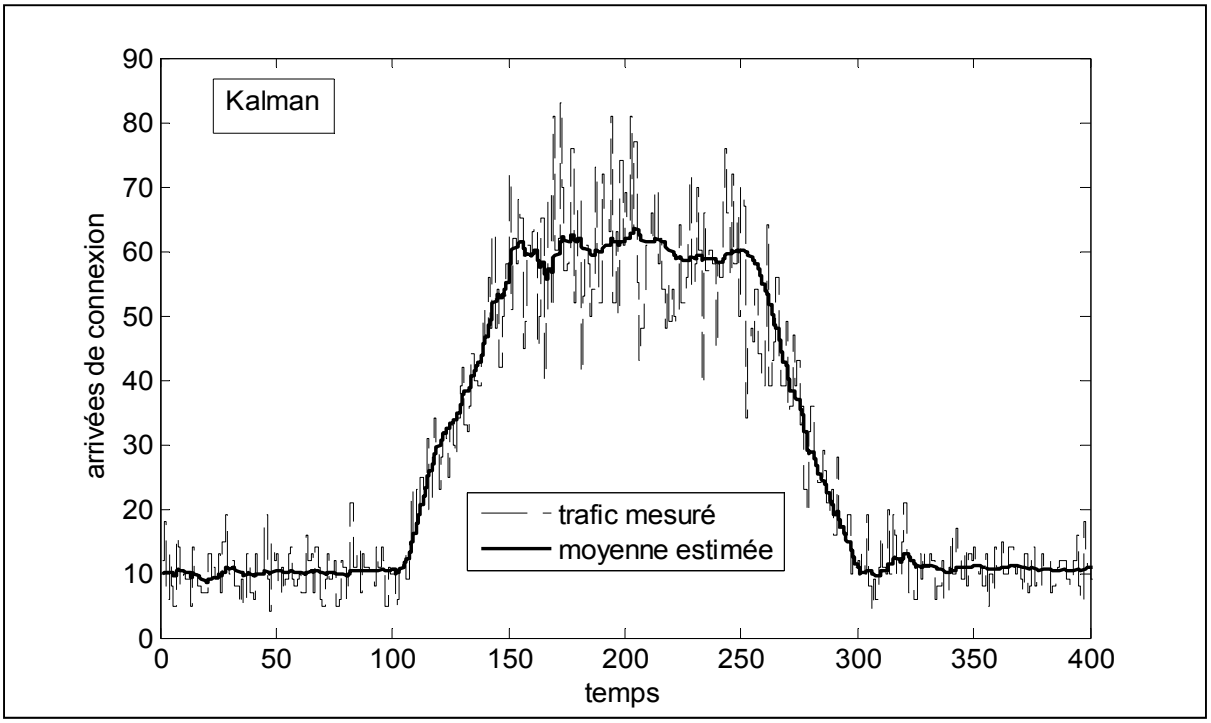

Figure 3.10 Estimation de trafic par *filtre de Kalman*.

3.6 Résumé

Dans ce chapitre, nous avons proposé deux options d'approche pour l'estimation en ligne de la tendance de trafic, utilisée pour supporter l'adaptation de capacité du réseau dédié de service.

La plus simple approche est fondée sur le lissage exponentiel (SES) adaptatif où le coefficient de lissage est adapté à la tendance courante de demande de trafic. La tendance est évaluée en utilisant soit la fonction d'autocorrélation (*acf*) ou soit la fonction de distribution cumulée (*cdf*) du processus d'arrivées de connexion. Cette approche peut facilement être réalisée pour de grands réseaux, sans la nécessité d'établir un modèle de demande de trafic. La deuxième approche utilise un filtre de Kalman où le modèle du système est obtenu à partir de mesures passées de trafic.

La performance des méthodes d'estimation est évaluée à l'aide de métriques définies de stabilité et de réponse. Les méthodes connues de lissage exponentiel présentent un choix de compromis entre les performances de stabilité et de réponse, le meilleur étant réalisé par *SES-a* (AEES), (Mentzer, 1988). Comparés à ceux de *SES-a*, des résultats de stabilité environ 1,5 fois meilleurs et de réponse environ 2 fois meilleurs sont réalisés par l'approche *SES-acf*. La meilleure performance a été obtenue par notre méthode fondée sur le *filtre de Kalman* où la stabilité est environ 4 fois meilleure que *SES-a* et la réponse 3 à 5 fois meilleure.

Nous avons aussi présenté une approche de ré-ajustage de capacité, qui profite de la disponibilité de la variance de l'erreur d'estimation fournie par le filtre de Kalman pour améliorer la GoS du réseau. Dans le prochain chapitre, nous présenterons l'évaluation de la performance du réseau SON obtenue avec l'intégration des approches proposées d'estimation de trafic et d'adaptation de capacité.

CHAPITRE 4

ÉVALUATION DE PERFORMANCE

4.1 Introduction

Dans ce chapitre, nous présentons l'évaluation de la performance de notre approche de gestion de ressources du SON. Les objectifs de l'approche sont la garantie de QoS de bout en bout des connexions et la maximisation du bénéfice du SON, sous la contrainte du maintien d'un degré de service (GoS) acceptable. L'approche est réalisée par l'intégration des éléments suivants :

- Une estimation en temps réel de la demande de trafic à acheminer dans le réseau. Les propositions sur cet élément ont été présentées au CHAPITRE 3.
- Une adaptation de capacité des liens du réseau à la demande de trafic estimée ci-dessus qui maximise le bénéfice. Pour cet élément, le modèle d'adaptation de capacité MDPD a été présenté aux Sections 2.3.2 et 2.4.2.
- Une adaptation du contrôle d'admission et du routage des connexions à la demande de trafic estimée et à la capacité des liens, qui aussi maximise le bénéfice et en même temps assure la QoS. Cet élément est réalisé par le modèle de CAC et routage MDPD (Dziong, 1997) adapté pour le SON à la Section 2.3.1.

La performance est évaluée pour les cas des méthodes proposées d'estimation de demande de trafic qui ont été jugées les meilleures, soient *SES-acf* et estimation fondée sur *filtre de Kalman*. Les résultats obtenus de GoS et de bénéfice du réseau sont comparés à ceux obtenus quand les méthodes d'estimation connues *SES-f* et *SES-a* sont utilisées. Comme la maximisation du bénéfice pour une méthode d'estimation donnée a déjà été vérifiée aux Sections 2.5.1 et 2.6.3, nous nous concentrons ici sur les améliorations additionnelles apportées par une estimation plus précise du trafic. Nous démontrerons aussi l'amélioration de la GoS résultant de l'application de notre approche de ré-ajustage de capacité fondée sur l'erreur d'estimation fournie par le filtre de Kalman (Section 3.4.2).

Les essais de cette évaluation ont tous été effectués à l'aide d'un simulateur à événements discrets développé dans le cadre de ce projet. Le réseau simulé est celui décrit à la Section 2.6.3 et illustré à la Figure 2.5.

4.2 Demande de trafic

La demande de trafic sur le réseau est simulée en suivant le gabarit quotidien des taux d'arrivées de connexions qui inclut des périodes de bas et de haut trafic, tel qu'illustré à la partie d) de la Figure 3.5. Le gabarit est défini par les paramètres suivants :

- Λ_l : taux d'arrivée moyen dans la période stationnaire de bas trafic,
- Λ_h : taux d'arrivée moyen dans la période stationnaire de haut trafic,
- T_0 : heure moyenne du début de la période de bas trafic,
- T_1 : heure moyenne de la fin de la période de bas trafic,
- T_2 : heure moyenne du début de la période de haut trafic,
- T_3 : heure moyenne de la fin de la période de haut trafic.

Les niveaux de trafic durant les jours de fin de semaine sont en général moindres que ceux des autres jours de la semaine. Pour accélérer l'exercice, nous excluons les fins de semaine de notre évaluation de performance. Si requis, il sera techniquement facile d'étendre l'étude pour couvrir cette exclusion. Étant donné que la population d'usagers du réseau ne change pas, nous présumons que Λ_l et Λ_h ne changent pas d'un jour à l'autre. Cependant les heures de débuts et de fins des périodes de bas et de haut trafic, t_0, t_1, t_2, t_3, montreront des déviations d'un jour à l'autre, par rapport à leurs moyennes respectives T_0 à T_3, représentées par :

$$t_i = T_i + \tau_i, i = 0..3, \tag{4.1}$$

où τ_i est la déviation que nous présumons être de distribution Gaussienne de moyenne zéro et de variance v_τ.

Ainsi, le taux d'arrivées de connexion au temps *k*, λ_k, peut être spécifié comme suit pour les différentes périodes dans un cycle de trafic journalier :

$$k \in [t_0, t_1[: \quad \lambda_k = \Lambda_l + e_k, \tag{4.2}$$

$$k \in [t_1, t_2[: \quad \lambda_k = \Lambda_l + (\Lambda_h - \Lambda_l)\frac{k - t_1}{t_2 - t_1} + e_k, \tag{4.3}$$

$$k \in [t_2, t_3[: \quad \lambda_k = \Lambda_h + e_k, \tag{4.4}$$

$$k \in [t_3, t_0 + 24h[: \quad \lambda_k = \Lambda_h + (\Lambda_l - \Lambda_h)\frac{k - t_3}{t_0 + 24h - t_3} + e_k, \tag{4.5}$$

où e_k est une déviation, présumée de distribution Gaussienne de moyenne zéro et de variance $v_{e,k}$, provenant de la nature aléatoire de λ_k.

Des arrivées de connexion Poissoniennes sont alors générées, suivant les taux λ_k définis ci-dessus, pour simuler la demande de trafic sur le réseau. Ce trafic servira à l'évaluation de performance ainsi qu'à la collection de statistiques de trafic de liens pour établir le modèle du filtre de Kalman (Section 3.4.1.2.1).

4.3 Résultats de l'évaluation

Dans les scénarios considérés, les paramètres du gabarit de la demande journalière de trafic sont assignés comme suit :

- Les taux d'arrivées moyens Λ_l sont de $4/t_s$ pour le tiers des classes de connexions et de $2{,}6/t_s$ pour le restant des classes;
- Taux d'arrivées moyens $\Lambda_h = 5\,\Lambda_l$;
- Heures des changements de trafic : T_0=0, T_1=11, T_2=14, T_3=21 (la dernière période dure trois heures);
- Écart type des temps T_i, $i=0..3$: $\sqrt{v_\tau}$ = 1000 secondes;
- Écart type des taux d'arrivée λ_k : $\sqrt{v_{e,k}}$ = 10% de λ_k .

La contrainte du taux de blocage moyen du réseau et l'objectif de la GoS sont fixés respectivement à:

$$B_T \le B_T^C = 1\%, \tag{4.6}$$

$$\Pr\{B_{k,T} > B_T^C\} \le 10\%. \tag{4.7}$$

Dans la suite, nous présenterons premièrement une analyse des résultats de GoS obtenus quand les différentes méthodes d'estimation sont utilisées, et nous montrerons ensuite les résultats de bénéfices du réseau correspondants.

4.3.1 Analyse de degré de service

Dans cette analyse, nous comparons la GoS fournie par le réseau dans les cas où l'estimation de la demande de trafic utilise chacune des méthodes SES-acf, filtre de Kalman et filtre de Kalman avec approche d'amélioration de GoS (Section 3.4.2). Nous identifions cette dernière méthode par $Kalman_{GoS}$. Le taux de blocage de réseau à chaque instant k d'une journée, $B_{k,T}$, que nous appellerons taux de blocage Time of Day (TOD), est calculé en intervalles de cinq minutes. Les valeurs moyennes $\overline{B}_{k,T}$ de ces taux sont obtenues en effectuant une moyenne sur 20 jours.

La Figure 4.1 montre les résultats d'une journée de simulation obtenus dans le cas de l'utilisation de l'estimation *SES-acf*. $\overline{B}_{k,T}$ se maintient à moins de 2% pour la plupart des temps k, excepté pour la période initiale de tendance positive des taux d'arrivées où il atteint temporairement un sommet à 15%. Nous constatons aussi que les violations TOD de l'objectif de GoS, montrées avec le graphe de $\Pr\{ B_{k,T} > 0{,}01 \}$, apparaissent dans la majorité des temps.

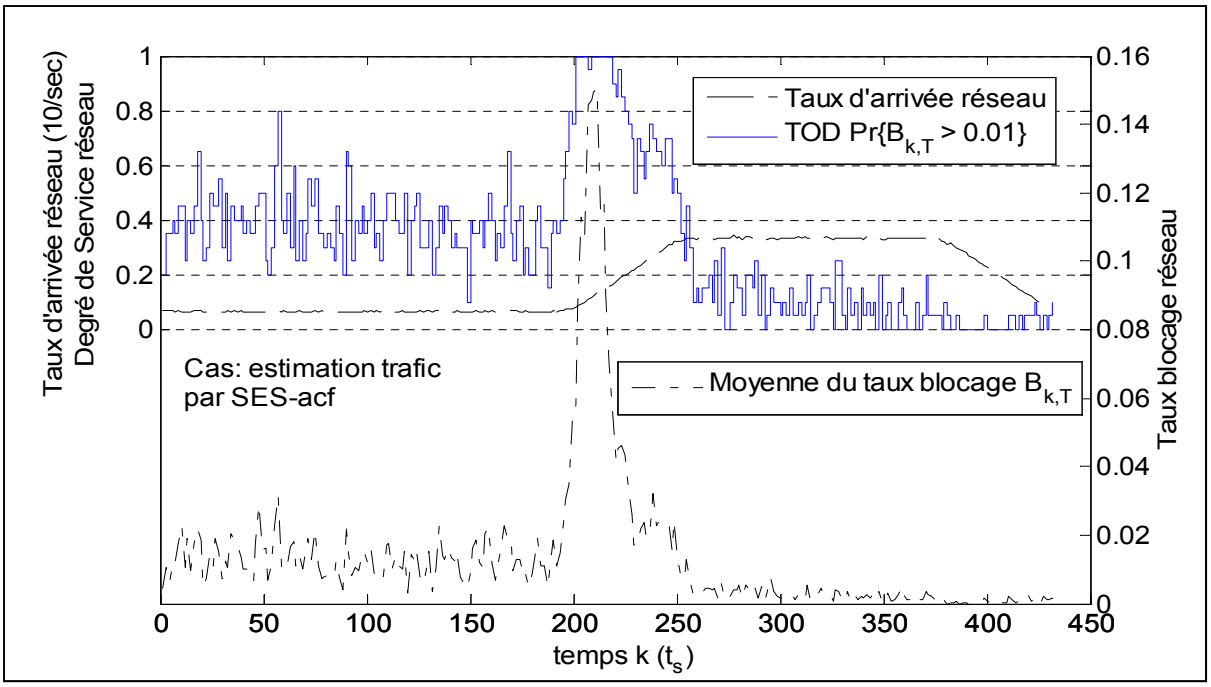

Figure 4.1 Trafic offert et taux de blocage TOD du réseau,
adaptation de capacité utilisant estimation SES-acf.

La Figure 4.2 montre les résultats de la journée de simulation avec la méthode proposée d'estimation par filtre de Kalman. Dans ce cas ci, on ne tient pas compte de la variance de l'erreur d'estimation pour le ré-ajustage des capacités. Maintenant, $\overline{B}_{k,T}$ se maintient la plupart du temps en bas de 1% et son sommet temporaire au début de la tendance ascendante de trafic est limité à 2%. Cependant, les violations de GoS demeurent fréquentes.

Pour contrôler les violations de GoS, nous appliquons maintenant l'approche de ré-ajustage de capacité proposée à la Section 3.4.2 imposant la contrainte de probabilité de violation de GoS fixée par (4.7). Pour le scénario du réseau considéré, nous avons déterminé de façon heuristique que la contrainte de blocage de réseau (4.6) est respectée quand le taux de blocage de chacun des liens obéit à la contrainte $B_s^C = 0.1$ (3.30). Exécutant le ré-ajustage de capacité par (3.31) à (3.33) sur tous les liens, les taux de blocage du réseau et les probabilités de violation de GoS sont réduits, satisfaisant ainsi l'objectif de GoS du réseau comme montrés à la Figure 4.3 et au Tableau 4.1.

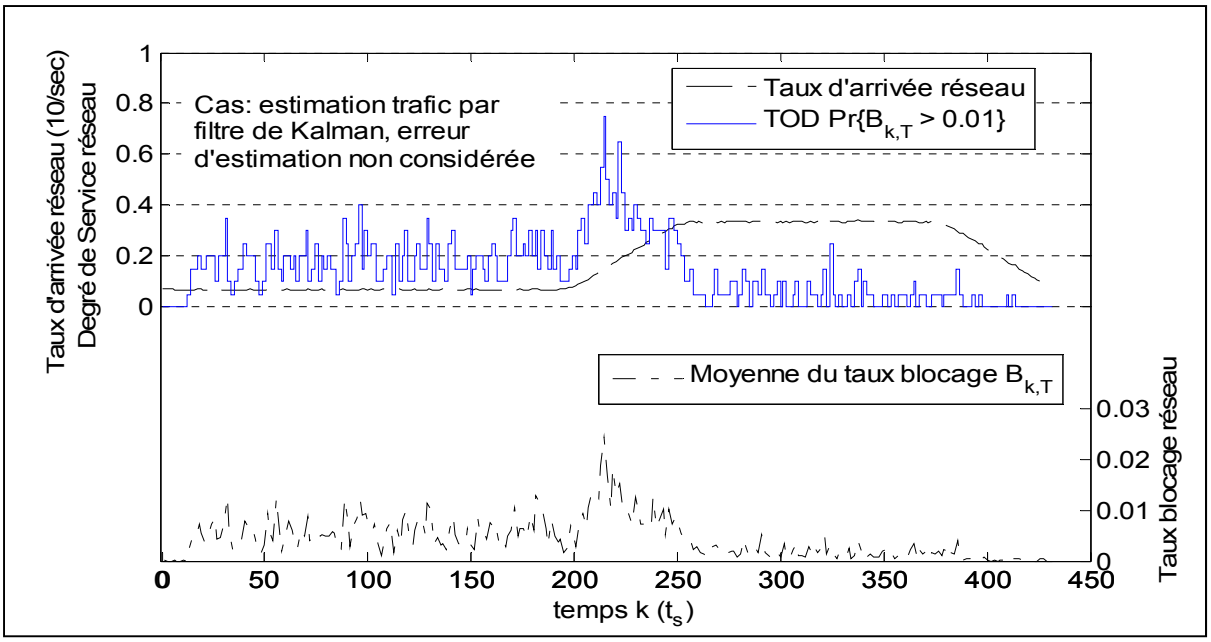

Figure 4.2 Trafic offert et taux de blocage TOD du réseau,
adaptation de capacité utilisant estimation par filtre de Kalman,
GoS obtenu sans considérer les erreurs d'estimation.

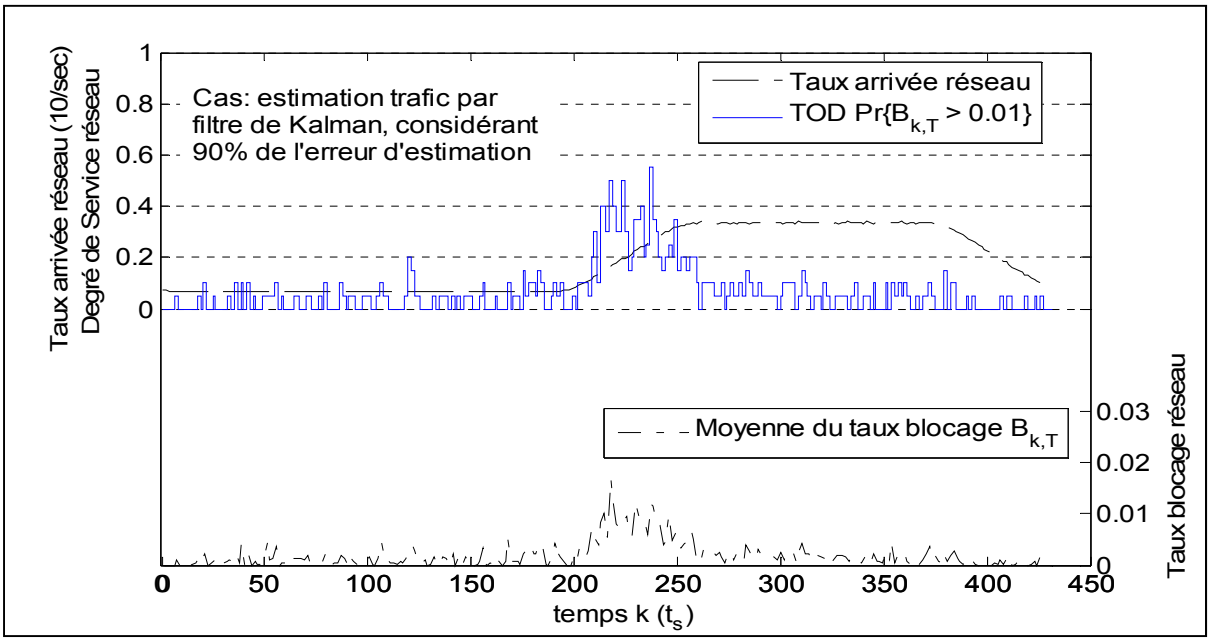

Figure 4.3 Trafic offert et taux de blocage TOD du réseau,
adaptation de capacité utilisant estimation par filtre de Kalman,
GoS obtenu en tenant compte de 90% de l'erreur d'estimation.

Un aperçu des performances sur le blocage de réseau, résultant des différentes méthodes d'estimation, est illustré avec les graphes de fonction de distribution cumulée (cdf) de $B_{k,T}$ à la Figure 4.4. Pour chaque méthode, la courbe correspondante indique les proportions des taux de blocage qui sont inférieurs aux niveaux donnés. La figure confirme clairement l'avantage procuré par l'estimation par filtre de Kalman, en particulier quand l'approche de ré-ajustage de capacité $Kalman_{GoS}$ est appliquée.

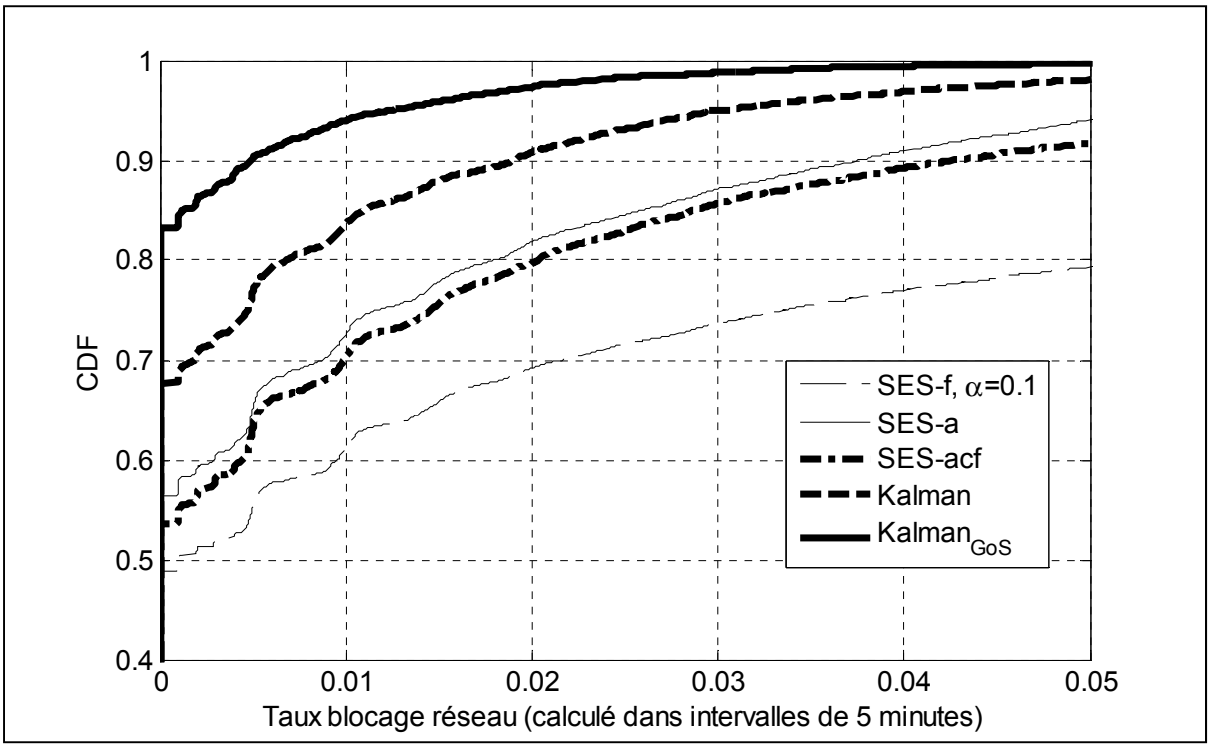

Figure 4.4 Fonction de distribution cumulée du taux de blocage de réseau.

Nous présentons une vue d'ensemble des taux de blocage correspondant aux méthodes d'estimation au Tableau 4.1. B_T mesure la moyenne générale du taux de blocage, calculée sur toute la durée de 20 jours de la simulation. Les probabilités de violation de GoS, $\Pr\{ B_{k,T} > 1\% \}$, sont aussi calculées sur la durée totale de simulation.

Les meilleures moyennes de blocage sont obtenues avec l'usage de l'estimation par filtre de Kalman. En effet, celles-ci sont d'au moins de moitié inférieures à celles obtenues par les méthodes *SES* adaptatifs et de dix fois inférieures à celle par *SES* fixe.

4.3.2 Bénéfice du réseau

Pour présenter une image complète des résultats, le Tableau 4.1 inclut aussi une vue des taux de bénéfice de réseau réalisés dans les différents cas de méthodes d'estimation. Les bénéfices plus élevés produits dans les cas d'estimation par filtre de Kalman et par *SES* adaptatifs démontrent l'avantage relié à l'usage de méthodes d'estimation adaptative. Dans ces cas, les taux de bénéfice entre eux sont comparables, se retrouvant en dedans de leurs intervalles de confiance respectifs.

Les résultats confirment également que l'application de l'approche $Kalman_{GoS}$ réalise l'objectif GoS (4.7) en réduisant la probabilité de violations de moitié par rapport à celle du filtre de Kalman seul. Il est intéressant de noter que cet objectif est réalisé avec une dépense de bénéfice de moins de 1%.

Tableau 4.1 Résultats de performance, à 90% d'intervalle de confiance

Méthode	*Moyenne du taux de blocage* B_T *(%)*	$\Pr\{B_{k,T} > 1\%\}$ *(%)*	*Taux de bénéfice*
SES-f, α=0.1	4,085 ± 0,053	39,6 ± 1,9	15,733 ± 0,027
SES-a	1,062 ± 0,021	28,5 ± 1,9	16,325 ± 0,026
SES-acf	1,392 ± 0,055	30,8 ± 2,1	16,279 ± 0,028
Kalman	0,440 ± 0,026	13,9 ± 0,8	16,332 ± 0,031
$Kalman_{GoS}$	0,197 ± 0,012	6,7 ± 0,5	16,238 ± 0,028

4.4 Résumé

L'évaluation de notre proposition de gestion de ressources du SON a été présentée dans ce chapitre. La solution intégrant nos approches proposées d'estimation de demande de trafic et d'adaptation de capacités de liens à cette demande a été évaluée par simulation dans un réseau de taille réaliste. La demande de trafic simulée suit un gabarit de trafic avec des variations quotidiennes similaires à celles observées dans des traces de trafic réel publiées dans la littérature.

Les résultats montrent que les bénéfices du réseau sont les plus élevés dans les cas d'adaptation de capacité où la demande de trafic est estimée soit par le filtre de Kalman ou soit par lissage exponentiel adapté à la tendance. En plus du gain de 11% réalisé avec l'adaptation de capacité par rapport la capacité fixe (Section 2.7), l'utilisation d'une estimation adaptative proposée procure une amélioration additionnelle du bénéfice d'environ 4% sur celle d'une estimation par lissage exponentiel à paramètre fixe.

Au sujet du respect des contraintes de GoS, les meilleurs résultats sont obtenus quand notre méthode par filtre de Kalman est adoptée pour l'estimation de la demande de trafic. L'utilisation de base du filtre (sans ré-ajustage supplémentaire de capacité) donne un taux de blocage moyen de 2 à 9 fois inférieur à ceux donnés avec lissage exponentiel, et une probabilité de violations temporelles de GoS de 2 à 3 fois moindre. Avec l'intégration de notre approche de ré-ajustage de capacité fondé sur l'erreur d'estimation, une réduction supplémentaire de 50% est obtenue, tant dans le taux de blocage que dans les violations de GoS. Cette réduction est réalisée avec une dépense de moins de 1% en bénéfice.

CONCLUSION

A. Sommaire et avantages de notre approche de gestion de ressources

En même temps que la qualité de service fournie aux clients, notre approche de gestion de ressources du réseau dédié de service présentée dans cette thèse procure à l'opérateur un bénéfice d'exploitation maximisé, ainsi que le maintien d'un degré de service spécifié qui assure la satisfaction des clients du réseau. Cette approche est fondée sur les deux éléments suivants : premièrement, un modèle d'adaptation de capacité des liens du réseau à la demande courante de trafic, qui maximise le bénéfice en tenant compte des revenus de connexions et des coûts du réseau; ensuite, une méthode efficace d'estimation en temps réel de la demande de trafic qui alimente l'adaptation de capacité.

La conception du modèle d'adaptation de capacité a été effectuée à partir d'un encadrement économique où le bénéfice du réseau provient de la différence du revenu des connexions admises et des coûts du réseau impliquant principalement les coûts SLA de la bande passante des liens dédiés. Pour pouvoir se réaliser dans des réseaux de tailles réalistes, le modèle d'adaptation est distribué aux liens du réseau, produisant ainsi des conditions d'optimalité du bénéfice de lien indépendantes. Chaque condition est fondée sur la moyenne du *shadow price* du lien, ce *shadow price* provenant de la politique de contrôle d'admission et de routage de connexion MDPD (Dziong, 1997) appliquée dans le réseau. La politique de routage MDPD, qui est fondée sur la théorie du processus de décision de Markov, procure la maximisation du bénéfice dans un réseau où la capacité a été déterminée. À travers ce *shadow price* qui assigne un coût dynamique d'utilisation de bande passante du lien par la connexion, l'adaptation de capacité est étroitement intégrée au contrôle d'admission et routage de connexion, ce qui assure l'efficacité de la maximisation du bénéfice. La validité du modèle d'adaptation de capacité pour la maximisation du bénéfice a été vérifiée sur des réseaux simples (voir CHAPITRE 2).

Dans le modèle d'adaptation de capacité, le degré de service désiré du réseau est réalisé par le choix des valeurs données aux paramètres de récompense des connexions (CHAPITRE 2). Si un fournisseur Internet change momentanément le coût SLA de sa bande passante, l'algorithme d'adaptation peut changer la capacité de lien pour compenser le nouveau coût et ainsi maintenir la maximisation du bénéfice. Ceci peut provoquer une détérioration du degré de service du réseau. Pour maintenir la contrainte de degré de service, nous avons aussi proposé une approche d'adaptation de paramètre de récompense. Cette dernière adaptation est également intégrée à l'adaptation de capacité et au routage cités précédemment pour constituer une approche unifiée de gestion de ressources.

Dans la partie suivante de la thèse, nous avons proposé deux méthodes alternatives d'estimation en temps réel de demande de trafic, fondées sur l'estimation de sa tendance, pour servir les algorithmes d'adaptation de capacité et de routage. Comme la demande de trafic traverse typiquement des périodes stationnaires et tendancielles, l'estimation de trafic fondée sur sa tendance permet une meilleure stabilité des valeurs estimées en période stationnaire et une réponse plus rapide en période de tendance.

La première méthode d'estimation proposée est réalisée par lissage exponentiel adaptatif où le coefficient de lissage est adapté à la tendance de trafic. Cette tendance de trafic est estimée par l'analyse soit de la fonction d'autocorrélation ou soit de la fonction de distribution cumulée du processus d'arrivées de connexion.

La deuxième méthode utilise un filtre de Kalman où les paramètres du filtre sont déterminés à partir de données statistiques de la demande de trafic du réseau. Cette méthode présente l'avantage de pleinement tenir compte de l'historique de la demande de trafic pour en permettre des estimations plus précises. L'avantage majeur de la méthode est cependant la disponibilité de la distribution de l'erreur d'estimation procurée par l'algorithme. Nous avons ainsi pu concevoir un mécanisme de ré-ajustage de capacité fondée sur l'erreur d'estimation, pour améliorer la probabilité de satisfaction de contrainte de degré de service. Quand l'estimation par filtre de Kalman est utilisée, ce mécanisme de ré-ajustage de capacité peut

aussi être utilisé, en alternative à l'adaptation de paramètre de récompense mentionnée plus haut, comme approche de réalisation du degré de service du réseau.

L'évaluation de notre approche intégrée de gestion des ressources du réseau dédié de service a été effectuée par simulation d'un réseau homogène de taille réaliste. Les résultats de nos exemples démontrent que l'adaptation de capacité a procuré des bénéfices de réseau de 11% à 15% supérieurs à ceux du même réseau avec des capacités de liens fixes. L'évaluation des méthodes d'estimation de trafic a été effectuée utilisant des métriques définies de performances en stabilité et en réponse. Les résultats de nos méthodes par lissage exponentiel adaptatif, fondé sur la fonction d'autocorrélation (*SES-acf*), et par filtre de Kalman ont démontré des améliorations de ces performances, respectivement d'environ 2% et 4% par rapport à celles obtenues par lissage exponentiel AEES (Mentzer, 1988) connu. Aussi, en intégrant dans l'adaptation de capacité le mécanisme proposé de ré-ajustage fondé sur l'erreur d'estimation du filtre de Kalman, le taux de violations de GoS dans le réseau a été réduit à 4 fois moins que celui montré par AEES.

B. Travaux subséquents et direction future de recherche

La philosophie d'adaptation de capacité pour la maximisation du bénéfice développée pour le SON dans cette thèse peut être adaptée pour s'appliquer à d'autres types de réseaux dont l'attribution de capacité peut être variable, comme par exemple les réseaux virtuels. Une première adaptation de l'approche s'appliquant à des réseaux maillés sans fil cognitifs a été proposée dans (Amini et Dziong, 2010). La maximisation du bénéfice dans ce cas est obtenue par une approche d'attribution des canaux entre les nœuds qui utilise les valeurs moyennes de *shadow price* des nœuds.

Une direction de recherche future serait d'élargir l'étude d'adaptation de capacité au-delà des limites assumées dans cette thèse. Une première limitation concerne l'application de notre approche à des réseaux homogènes. Comme les réseaux de service devraient en général pouvoir supporter des connexions avec différents besoins en largeur de bande et temps de

service moyen, il est intéressant d'étendre l'étude à des réseaux *multi-rate*. Une deuxième simplification est la présomption d'une structure de prix de SLA linéaire, avec une résolution accommodant l'acquisition de bande passante par unité de connexion. Il serait aussi intéressant de comprendre les changements nécessaires à notre approche dans le cas d'une structure plus générale, avec un prix non linéaire et une acquisition par blocs de multiples connexions.

LISTE DE RÉFÉRENCES BIBLIOGRAPHIQUES

Allen, Doug 2002. « Virtela's IP VPN Overlay Networks. ». En ligne. <http://business.highbeam.com/787/article-1G1-81216931/virtela-ip-vpn-overlay-networks>. Consulté le 28 juillet 2010.

Amini, Reza Mossanen, et Zbigniew Dziong. 2010. « A framework for routing and channel allocation in cognitive wireless mesh networks ». In *2010 7th International Symposium on Wireless Communication Systems, ISWCS'10, September 19, 2010 - September 22, 2010*. p. 1017-1021. Coll. « Proceedings of the 2010 7th International Symposium on Wireless Communication Systems, ISWCS'10 ». York, United kingdom: IEEE Computer Society.

Amir, Yair, Claudiu Danilov, Stuart Goose, David Hedqvist et Andreas Terzis. 2005. « 1-800-Overlays: Using overlay networks to improve VoIP quality ». In *15th International Workshop on Network and Operating Systems Support for Digital Audio and Video, NOSSDAV 2005, June 13, 2005 - June 14, 2005*. p. 51-56. Coll. « Proceedings of the International Workshop on Network and Operating System Support for Digital Audio and Video ». Stevenson, WA, United states: Association for Computing Machinery.

Andersen, David G., Alex C. Snoeren et Hari Balakrishnan. 2003. « Best-path vs. multi-path overlay routing ». In *Proceedings of the 2003 ACM SIGCOMM Internet Measurement Conference, IMC 2003, October 27, 2003 - October 29, 2003*. p. 91-100. Coll. « Proceedings of the ACM SIGCOMM Internet Measurement Conference, IMC ». Miami Beach, FL, United states: Association for Computing Machinery.

Anjali, T., C. Bruni, D. Iacoviello, G. Koch et C. Scoglio. 2004. « Filtering and forecasting problems for aggregate traffic in Internet links ». *Performance Evaluation,* vol. 58, n° 1, p. 25-42.

Anjali, Tricha, Caterina Scoglio et George Uhl. 2003. « A new scheme for traffic estimation and resource allocation for bandwidth brokers ». *Computer Networks,* vol. 41, n° 6, p. 761-777.

Awduche, D. , J. Malcolm, J. Agogbua, M. O'Dell et J. McManus. 1999. *Requirements for Traffic Engineering Over MPLS.* RFC 2702. Internet Engineering Task Force (IETF). <http://www.apps.ietf.org/rfc/rfc2702.html>. Consulté le 15 juillet 2010.

Awduche, Daniel O., et Bijan Jabbari. 2002. « Internet traffic engineering using multi-protocol label switching (MPLS) ». *Computer Networks,* vol. 40, n° 1, p. 111-129.

BitTorrent Inc. 2001. « BitTorrent ». En ligne. <http://www.bittorrent.com/>. Consulté le 26 juillet 2010.

Blake, S. , D. Black, M. Carlson, E. Davies, Z. Wang et W. Weiss. 1998. *An Architecture for Differentiated Services*. RFC 2475. Internet Engineering Task Force (IETF). <http://www.apps.ietf.org/rfc/rfc2475.html>. Consulté le 15 juillet 2010.

Braden, R., D. Clark et S. Shenker. 1994. *Integrated Services in the Internet Architecture: an Overview*. RFC 1633. Internet Engineering Task Force (IETF). <http://www.apps.ietf.org/rfc/rfc1633.html>. Consulté le 15 juillet 2010.

Cavazos-Cadena, Rolando. 2002. « Value iteration and approximately optimal stationary policies in finite-state average Markov decision chains ». *Mathematical Methods of Operations Research,* vol. 56, n° 2, p. 181-196.

Cheng, Yu, Ramy Farha, Ali Tizghadam, Myung Sup Kim, Massoud Hashemi, Alberto Leon-Garcia et James Won-Ki Hong. 2005. « Virtual network approach to scalable IP service deployment and efficient resource management ». *IEEE Communications Magazine,* vol. 43, n° 10, p. 76-84.

Ching, Wai-Ki, Stefan Scholtes et Shu-Qin Zhang. 2004. « Numerical algorithms for dynamic traffic demand estimation between zones in a network ». *Engineering Optimization,* vol. 36, n° 3, p. 379-400.

Cohen, R., et G. Kaempfer. 2000. « On the cost of virtual private networks ». *IEEE/ACM Transactions on Networking,* vol. 8, n° 6, p. 775-84.

Cui, Yi, Baochun Li et Klara Nahrstedt. 2004. « oStream: Asynchronous streaming multicast in application-layer overlay networks ». *IEEE Journal on Selected Areas in Communications,* vol. 22, n° 1, p. 91-106.

Dasgupta, S., J. C. de Oliveira et J. P. Vasseur. 2008. « Trend based bandwidth provisioning an online approach for traffic engineered tunnels ». In *2008 Next Generation Internet Networks (NGI '08), 28-30 April 2008*. p. 53-60. Coll. « 2008 Next Generation Internet Networks (NGI '08) ». Piscataway, NJ, USA: IEEE.

Datar, M. 2002. « Butterflies and peer-to-peer networks ». In *Algorithms - ESA 2002. 10th Annual European Symposium. Proceedings, 17-21 Sept. 2002*. p. 310-22. Coll. « Algorithms - ESA 2002. 10th Annual European Symposium. Proceedings (Lecture Notes in Computer Science Vol.2461) ». Berlin, Germany: Springer-Verlag.

Duan, Zhenhai, Zhi-Li Zhang et Yiwei Thomas Hou. 2003. « Service Overlay Networks: SLAs, QoS, and Bandwidth Provisioning ». *IEEE/ACM Transactions on Networking,* vol. 11, n° 6, p. 870-883.

Dziong, Zbigniew. 1997. *ATM network resource management*. New York: McGraw-Hill, 315 p.

Eng Keong, Lua, J. Crowcroft, M. Pias, R. Sharma et S. Lim. 2005. « A survey and comparison of peer-to-peer overlay network schemes ». *Communications Surveys & Tutorials, IEEE,* vol. 7, n° 2, p. 72-93.

Evans, J., et C. Filsfils. 2004. « Deploying Diffserv at the network edge for tight SLAs, part I ». *IEEE Internet Computing,* vol. 8, n° 1, p. 61-5.

Evers, Joris. 2003. « Kazaa makers move into IM, Net telephony ». En ligne. <http://www.infoworld.com/t/networking/kazaa-makers-move-im-net-telephony-044?r=802>. Consulté le 23 juillet 2010.

Ferguson, Paul, et Geoff Huston. 1998. *Quality of service : delivering QoS on the Internet and in corporate networks*. New York: Wiley. <http://www.books24x7.com/marc.asp?bookid=505>.

Fry, G., et R. West. 2004. « Adaptive routing of QoS-constrained media streams over scalable overlay topologies ». In *Proceedings. RTAS 2004. 10th IEEE Real-Time and Embedded Technology and Applications Symposium, 25-28 May 2004*. p. 518-25. Coll. « Proceedings. RTAS 2004. 10th IEEE Real-Time and Embedded Technology and Applications Symposium ». Los Alamitos, CA, USA: IEEE Comput. Soc.

Gao, Jun, et Peter Steenkiste. 2004. « Design and evaluation of a distributed scalable content discovery system ». *IEEE Journal on Selected Areas in Communications,* vol. 22, n° 1, p. 54-66.

Gardner, E. S., Jr. 2006. « Exponential smoothing: the state of the art - Part II ». *International Journal of Forecasting,* vol. 22, n° 4, p. 637-66.

Girard, A., et B. Liau. 1993. « Dimensioning of adaptively routed networks ». *IEEE/ACM Transactions on Networking,* vol. 1, n° 4, p. 460-8.

Girard, André. 1990. *Routing and dimensioning in circuit-switched networks*. Coll. « Addison-Wesley series in electrical and computer engineering. Telecommunications. ». Reading, Mass.: Addison-Wesley, xiv, 556 p. p.

Guichard, Jim, François Le Faucheur et Jean-Philippe Vasseur. 2005. *Definitive MPLS network designs*. Indianapolis, Ind.: Cisco ;, xxv, 516 p. p.

Hyojin, Park, Yang Jinhong, Park Juyoung, Gak Kang Shin et Kyun Choi Jun. 2008. « A survey on peer-to-peer overlay network schemes ». In *2008 10th International Conference on Advanced Communication Technology, Febrary 17, 2008 - Febrary 20, 2008*. Vol. 2, p. 986-988. Coll. « International Conference on Advanced

Communication Technology, ICACT ». Phoenix Park, Korea, Republic of: Institute of Electrical and Electronics Engineers Inc.

Internet Assigned Number Authority. 2005. « Number Resources ». En ligne. <http://www.iana.org/numbers/>. Consulté le 01 mars 2011.

Jun, A. D. S., et A. Leon-Garcia. 1998. « Virtual network resources management: a divide-and-conquer approach for the control of future networks ». In *IEEE GLOBECOM 1998, 8-12 Nov. 1998*. Vol. vol.2, p. 1065-70. Coll. « IEEE GLOBECOM 1998 (Cat. NO. 98CH36250) ». Piscataway, NJ, USA: IEEE. <http://dx.doi.org/10.1109/GLOCOM.1998.776890>.

Kalman, R. E. 1960. « New approach to linear filtering and prediction problems ». *American Society of Mechanical Engineers -- Transactions -- Journal of Basic Engineering Series D,* vol. 82, n° 1, p. 35-45.

Kazaax. « Peer-To-Peer (P2P) and How Kazaa Works ». En ligne. <http://www.kazaax.com/>. Consulté le 23 juillet 2010.

Kelly, F.P. 1996. « Notes on effective bandwidth ». In *Kelly, F.P., Zachary, S.,Ziedins, I.B. (Eds.), Stochastic Networks: Theory and Applications*. p. 141-168. Oxford: Oxford University Press.

Khamsi, Roxanne. 2004. « Skype beyond the hype ». *Technology Review,* vol. 107, n° 5, p. 44-47.

Kim, Myung Sup, Ali Tizghadam, Alberto Leon-Garcia et James Won-Ki Hong. 2005. « Virtual network based autonomic network resource control and management system ». In *GLOBECOM'05: IEEE Global Telecommunications Conference, 2005, November 28, 2005 - December 2, 2005*. Vol. 2, p. 1075-1079. Coll. « GLOBECOM - IEEE Global Telecommunications Conference ». St. Louis. MO, United states: Institute of Electrical and Electronics Engineers Inc. <http://dx.doi.org/10.1109/GLOCOM.2005.1577802>.

Kolarov, A., A. Atai et J. Hui. 1994. « Application of Kalman filter in high-speed networks ». In *1994 IEEE GLOBECOM. Communications: The Global Bridge, 28 Nov.-2 Dec. 1994*. Vol. 1, p. 624-628. Coll. « 1994 IEEE GLOBECOM. Communications: The Global Bridge. Conference Record (Cat. No.94CH34025) ». New York, NY, USA: IEEE. <http://dx.doi.org/10.1109/GLOCOM.1994.513593>.

Krithikaivasan, Balaji, Kaushik Deka et Deep Medhi. 2004. « Adaptive bandwidth provisioning envelope based on discrete temporal network measurements ». In *IEEE INFOCOM 2004 - Conference on Computer Communications - Twenty-Third Annual Joint Conference of the IEEE Computer and Communications Societies, March 7,*

2004 - March 11, 2004. Vol. 3, p. 1786-1796. Coll. « Proceedings - IEEE INFOCOM ». Hongkong, China: Institute of Electrical and Electronics Engineers Inc.

Lam, Ngok, Zbigniew Dziong et Lorne G. Mason. 2007. « Network capacity allocation in service overlay networks ». In *20th International Teletraffic Congress, ITC20 2007, June 17, 2007 - June 21, 2007*. Vol. 4516 LNCS, p. 224-235. Coll. « Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) ». Ottawa, ON, Canada: Springer Verlag.

Li, B., Y. T. Hou, K. Sohraby, M. Ulema et Z. Zhang. 2004. « Recent Advances in Service Overlay Networks [Numéro spécial] ». *IEEE Journal on Selected Areas in Communications,* vol. 22, n° 1.

Li, Zhi, et Prasant Mohapatra. 2004. « QRON: QoS-aware routing in overlay networks ». *IEEE Journal on Selected Areas in Communications,* vol. 22, n° 1, p. 29-40.

Liang, Huan, O. Kabranov, D. Makrakis et L. Orozco-Barbosa. 2002. « Minimal cost design of virtual private networks ». In *IEEE CCECE2002. Canadian Conference on Electrical and Computer Engineering. Conference Proceedings, 12-15 May 2002*. Vol. vol.3, p. 1610-15. Coll. « IEEE CCECE2002. Canadian Conference on Electrical and Computer Engineering. Conference Proceedings (Cat. No.02CH37373) ». Piscataway, NJ, USA: IEEE. <http://dx.doi.org/10.1109/CCECE.2002.1012997>.

Makridakis, S., A. Andersen, R. Carbone, R. Fildes, M. Hibon, R. Lewandowski, J. Newton, E. Parzen et R. Winkler. 1982. « The accuracy of extrapolation (time series) methods: Results of a forecasting competition ». *Journal of Forecasting,* vol. 1, n° 2, p. 111-153.

Malkhi, Dahlia, Moni Naor et David Ratajczak. 2002. « Viceroy: A scalable and dynamic emulation of the butterfly ». In *Proceedings of the Twenty - First Annual ACM Symposium on Principles of Distributed Computing PODC 2002, July 21, 2002 - July 24, 2002*. p. 183-192. Coll. « Proceedings of the Annual ACM Symposium on Principles of Distributed Computing ». Monterey, CA, United states: Association for Computing Machinery.

Mentzer, John. 1988. « Forecasting with adaptive extended exponential smoothing ». *Journal of the Academy of Marketing Science,* vol. 16, n° 3, p. 62-70.

Mentzer, John, et Roger Gomes. 1994. « Further extensions of adaptive extended exponential smoothing and comparison with the M-Competition ». *Journal of the Academy of Marketing Science,* vol. 22, n° 4, p. 372-382.

Napster Web Team. 2003. « Napster.ca - All the music you want. Any way you want it. ». En ligne. <http://www.napster.ca/>. Consulté le 22 juillet 2010.

Oodan, A. P., Books24x7 Inc. et Institution of Electrical Engineers. 2003. *Telecommunications quality of service management [ressource électronique] : from legacy to emerging services*. Coll. « IEE telecommunications series », 48. London: Institution of Electrical Engineers. <http://www.books24x7.com/marc.asp?isbn=0852964242>.

Pandurangan, G., P. Raghavan et E. Upfal. 2003. « Building low-diameter peer-to-peer networks ». *IEEE Journal on Selected Areas in Communications,* vol. 21, n° 6, p. 995-1002.

Park, Kyung-Joon, et Chong-Ho Choi. 2008. « Optimization driven bandwidth provisioning in service overlay networks ». *Computer Communications,* vol. 31, n° 14, p. 3169-3177.

Paxson, Vern, et Sally Floyd. 1995. « Wide area traffic: The failure of Poisson modeling ». *IEEE/ACM Transactions on Networking,* vol. 3, n° 3, p. 226-244.

Pechiar, Juan, Gonzalo Perera et Maria Simon. 2002. « Effective bandwidth estimation and testing for Markov sources ». *Performance Evaluation,* vol. 48, n° 1-4, p. 157-175.

Pióro, Michal, et Deepankar Medhi. 2004. *Routing, flow, and capacity design in communication and computer networks*. Coll. « The Morgan Kaufmann series in networking ». Amsterdam ; Boston: Elsevier/Morgan Kaufmann, xxviii, 765 p. p.

Ratnasamy, Sylvia, Paul Francis, Mark Handley, Richard Karp et Scott Schenker. 2001. « A scalable content-addressable network ». In *ACM SIGCOMM 2001- Applications, Technologies, Architectures, and Protocols for Computers Communications-, August 27, 2001 - August 31, 2001*, 4. Vol. 31, p. 161-172. Coll. « Computer Communication Review ». San Diego, CA, United states: Association for Computing Machinery.

Recker, S., H. Ludiger et W. Geisselhardt. 2003. « Dynamic adaptation of virtual network capacity for deterministic service guarantees ». In *Quality of Service in Multiservice IP Networks. Second International Workshop, QoS-IP 2003. Proceedings, 24-26 Feb. 2003*. p. 689-703. Coll. « Quality of Service in Multiservice IP Networks. Second International Workshop, QoS-IP 2003. Proceedings (Lecture Notes in Computer Science Vol.2601) ». Berlin, Germany: Springer-Verlag.

Ripeanu, M. 2002. « Peer-to-peer architecture case study: Gnutella network ». In *Proceedings First International Conference on Peer-to-Peer Computing, 27-29 Aug. 2001*. p. 99-100. Coll. « Proceedings First International Conference on Peer-to-Peer Computing ». Los Alamitos, CA, USA: IEEE Comput. Soc.

Roberts, J. W. 2004. « Internet traffic, QoS, and pricing ». *Proceedings of the IEEE,* vol. 92, n° 9, p. 1389-99.

Rosen, E., A. Viswanathan et R. Callon. 2001. *Multiprotocol Label Switching Architecture*. RFC 3031. Internet Engineering Task Force (IETF). <http://www.apps.ietf.org/rfc/rfc3031.html>. Consulté le 15 juillet 2010.

Schollmeier, R. 2002. « A definition of peer-to-peer networking for the classification of peer-to-peer architectures and applications ». In *Proceedings First International Conference on Peer-to-Peer Computing, 27-29 Aug. 2001*. p. 101-2. Coll. « Proceedings First International Conference on Peer-to-Peer Computing ». Los Alamitos, CA, USA: IEEE Comput. Soc.

Schweitzer, P. J., et A. Federgruen. 1979. « Geometric convergence of value-iteration in multichain Markov decision problems ». *Advances in Applied Probability,* vol. 11, n° 1, p. 188-217.

Seok, Yongho, Nakjung Choi, Yanghee Choi et Jihong Kim. 2003. « Application-driven network capacity adaptation for energy efficient ad-hoc networks ». In *14th IEEE 2003 International Symposium on Personal, Indoor and Mobile Radio Communications. Proceedings, 7-10 Sept. 2003*. Vol. 1, p. 211-15. Coll. « 14th IEEE 2003 International Symposium on Personal, Indoor and Mobile Radio Communications. Proceedings (IEEE Cat. No.03TH8677) ». Piscataway, NJ, USA: IEEE.

Stoica, Ion, Robert Morris, David Liben-Nowell, David R. Karger, M. Frans Kaashoek, Frank Dabek et Hari Balakrishnan. 2003. « Chord: A scalable peer-to-peer lookup protocol for Internet applications ». *IEEE/ACM Transactions on Networking,* vol. 11, n° 1, p. 17-32.

Thompson, K., G. J. Miller et R. Wilder. 1997. « Wide-area Internet traffic patterns and characteristics ». *IEEE Network,* vol. 11, n° 6, p. 10-23.

Tu, M. 1994. « Estimation of point-to-point traffic demand in the public switched telephone network ». *IEEE Transactions on Communications,* vol. 42, n° 2-4, p. 840-5.

Vieira, S. L., et J. Liebeherr. 2004. « Topology design for service overlay networks with bandwidth guarantees ». In *2004 Twelfth IEEE International Workshop on Quality of Service, 7-9 June 2004*. p. 211-20. Coll. « 2004 Twelfth IEEE International Workshop on Quality of Service (IEEE Cat. No.04EX790) ». Piscataway, NJ, USA: IEEE.

WAND Network Research Group. 2010, 27 juillet. « WITS: Waikato Internet traffic storage ». En ligne. <http://www.wand.net.nz/wits/>. Consulté le 21 janvier 2011.

Wang, Xin, R. Ramjee et H. Viswanathan. 2005. « Adaptive and predictive downlink resource management in next-generation CDMA networks ». *IEEE Journal on Selected Areas in Communications,* vol. 23, n° 6, p. 1219-32.

Wang, Zheng. 2001. *Internet QoS : architectures and mechanisms for quality of service*. Coll. « Morgan Kaufmann series in networking ». San Francisco: Morgan Kaufmann, 239 p.

Zhao, Ben Y., Ling Huang, Jeremy Stribling, Sean C. Rhea, Anthony D. Joseph et John D. Kubiatowicz. 2004. « Tapestry: A resilient global-scale overlay for service deployment ». *IEEE Journal on Selected Areas in Communications,* vol. 22, n^{o} 1, p. 41-53.