Titre: Title:	Amélioration coût-performance et modèles de tarification dans les réseaux intégrés de troisième génération
Auteur: Author:	Fabien Houeto
Date:	2003
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Houeto, F. (2003). Amélioration coût-performance et modèles de tarification dans les réseaux intégrés de troisième génération [Ph.D. thesis, École Polytechnique de Montréal]. PolyPublie. https://publications.polymtl.ca/7171/

URL de PolyPublie: PolyPublie URL:	https://publications.polymtl.ca/7171/
Directeurs de recherche: Advisors:	Samuel Pierre
Programme: Program:	Génie informatique

In compliance with the
Canadian Privacy Legislation
some supporting forms
may have been removed from
this dissertation.

While these forms may be included
in the document page count,
their removal does not represent
any loss of content from the dissertation.

UNIVERSITÉ DE MONTRÉAL

AMÉLIORATION COÛT-PERFORMANCE ET MODÈLES DE TARIFICATION DANS LES RÉSEAUX INTÉGRÉS DE TROISIÈME GÉNÉRATION

OKOUN'OLA ERNST FABIEN HOUETO
DÉPARTEMENT DE GÉNIE INFORMATIQUE
ÉCOLE POLYTECHNIQUE DE MONTRÉAL

THÈSE PRÉSENTÉE EN VUE DE L'OBTENTION
DU DIPLÔME DE PHILOSOPHIAE DOCTOR
(GÉNIE INFORMATIQUE)
AVRIL 2003

Your file Votre référence
ISBN: 0-612-86446-4
Our file Notre référence
ISBN: 0-612-86446-4

UNIVERSITÉ DE MONTRÉAL
ÉCOLE POLYTECHNIQUE DE MONTRÉAL

Cette thèse intitulée :

AMÉLIORATION COÛT-PERFORMANCE ET MODÈLES DE TARIFICATION DANS LES RÉSEAUX INTÉGRÉS DE TROISIÈME GÉNÉRATION

présentée par : HOUETO Okoun'ola Ernst Fabien
en vue de l'obtention du diplôme de : PHILOSOPHIAE DOCTOR
a été dûment acceptée par le jury d'examen composé de :

M. PESANT Gilles, Ph.D., président
M. PIERRE Samuel, Ph.D., directeur de recherche et membre
M. CONAN Jean, Ph.D., membre
M. OROZCO-BARBOSA Luis, Ph.D., membre externe

À ma famille

REMERCIEMENTS

Je tiens en premier lieu à remercier Dr. Samuel Pierre dont le soutien, les conseils et l'exemple m'ont guidé tout au long de ma recherche.

Je tiens également à remercier le Conseil de Recherche en Sciences Naturelles et en Génie (CRSNG) pour la bourse qu'il m'a octroyée afin de me permettre de me consacrer à mes études.

Je remercie aussi mes parents, frère et sœurs pour leur soutien constant et indéfectible.

Je remercie tous mes collègues et le personnel du Laboratoire de Recherche en Réseautique et Informatique Mobile (LARIM) pour leur collaboration et surtout pour l'ambiance de travail chaleureuse.

Enfin, je remercie tous ceux que je n'ai pas eu la possibilité de nommer et qui m'ont soutenu, aidé ou encouragé d'une manière ou d'une autre durant ce long processus.

RÉSUMÉ

La tendance dans les réseaux de 3ème génération est à l'intégration des services. Ainsi, des normes et protocoles tels que l'UMTS et le Wireless Access Protocol (WAP) devraient permettre à tout usager mobile, en plus de la téléphonie, d'accéder à Internet et de faire de la vidéoconférence. L'intégration de services soulève cependant un certain nombre de problèmes. En effet, si dans les télécommunications traditionnelles, les modèles de trafic utilisés ont pu être testés et éprouvés pendant de longues années, l'avènement de nouveaux services (tel Internet) les a remis en question. De nouveaux modèles doivent être élaborés pour prendre en compte des types de trafic plus complexe. Pour les réseaux de troisième génération, plusieurs études se sont penchées sur les performances de leurs technologies radio, mais aucune étude ne s'est vraiment intéressée aux performances de ces réseaux d'un point de vue « applicatif », lorsque différents types de trafic sont multiplexés. Un autre problème soulevé par la combinaison de différents types de trafic est celui de la tarification. Avec des trafics ayant des exigences différentes, les opérateurs se heurtent au problème du choix de type et de la sorte de tarification à adopter. De manière générale, le rôle de la tarification n'est pas seulement de générer des revenus, elle introduit aussi une rétroaction qui permet d'exercer un certain contrôle sur les usagers en les incitant à réagir aux tarifs par une minimisation de leurs coûts.

Cette thèse vise donc, dans un premier temps, à étudier les performances de la partie accès (ou boucle locale) des réseaux cellulaires multiservices dans un contexte où sont multiplexés plusieurs types de service ou de trafic présentant souvent une dépendance (ou corrélation) à long terme et un caractère fractal et en rafales. Dans un deuxième temps, on cherche également à proposer un modèle de tarification amélioré pour permettre une prise en compte efficace des contraintes des usagers et des caractéristiques du trafic ainsi qu'une gestion judicieuse des ressources du réseau.

Pour y parvenir, nous avons d'abord recensé les différentes applications à modéliser. Ensuite, nous avons adapté les modèles existants ou en avons proposé de

nouveaux qui sont basés sur les spécifications de UMTS et parfois aussi sur des observations réelles. Les modèles obtenus ont été calibrés et validés par comparaison avec des traces réelles, puis utilisés pour simuler un réseau UMTS avec l'agrégation de différents types de trafic. Ceci définit un modèle de référence dont les performances ont été évaluées en fonction de certaines métriques de qualité de service. Des politiques de gestion et de contrôle pour assurer une meilleure qualité de service ont été suggérées à partir des observations faites et des résultats obtenus. Un modèle séparé a été développé pour l'évaluation de l'impact de diverses topologies sur la fiabilité d'une implémentation d'un réseau de troisième génération. Finalement, après un survol des différentes méthodes de tarification proposées dans la littérature, nous avons défini un certain nombre de critères requis pour notre modèle de tarification. À partir de ces critères, nous avons analysé les modèles existants et proposé une amélioration qui satisfait nos spécifications et qui s'adapte au cadre des réseaux de troisième génération qui intègrent plusieurs types de trafic. Plus précisément, nous avons introduit une extension de la tarification basée sur le débit effectif qui regroupe d'une part la proposition d'un critère pour prendre en compte la contrainte sur la variation de délai lors du contrôle d'admission et, d'autre part, l'utilisation d'une nouvelle famille de fonctions de tarification. Ces fonctions tiennent compte des incertitudes des usagers, de la sporadicité du trafic et offrent aussi une plus grande précision tout en gardant les bonnes propriétés sur les incitatifs.

Les principales contributions de cette thèse se situent au niveau des modèles de trafic, de réseau et de coût développés et des évaluations de performance réalisées. De manière générale, les résultats obtenus montrent qu'avec une bonne conception, les performances des réseaux de troisième génération, pour le trafic envisagé, sont acceptables. Il faut cependant définir avec soin les profils de QoS associés à chaque classe ou application en utilisant judicieusement la sur-souscription. Pour la classe « en flux », il faudrait mettre en place des tampons adéquats pour assurer la qualité de service notamment par rapport aux délais. Aussi, la prise en compte supplémentaire de la gestion de la QoS par un mécanisme d'ordonnancement au niveau des couches

supérieures permet une amélioration des performances. Bien que les résultats dépendent d'un grand nombre de paramètres, nous avons cependant pu établir l'impact d'un certain nombre de politiques comme la mise en place de la priorité au niveau IP et la sur-souscription des liaisons.

Au niveau de la fiabilité, différentes topologies ont été étudiées. Avec la robustesse des équipements disponibles, la redondance nécessaire pour assurer une grande disponibilité est minimale et peut être atteinte avec des topologies très simples. En effet, une topologie d'interconnexion des RNC en arbre avec quelques liens redondants fournit la disponibilité souhaitée avec des coûts relativement moindres.

Le critère de prise en compte de la gigue au niveau du contrôle d'admission a été validé par des simulations. Il permet de réaliser une amélioration pouvant aller jusqu'à 10% sur les bornes et par conséquent sur les admissions. La nouvelle famille de fonctions de tarification permet aussi d'obtenir des tarifs moins élevés sans augmenter le contrôle exercé par le réseau. Pour toutes ces améliorations, le système de facturation ne contrôle, tout comme dans le schéma classique, que le débit moyen de chaque connexion. Pour l'usager, les économies reliées à la nouvelle fonction de tarification croissent avec la taille des tampons disponibles pour le trafic au niveau des nœuds du réseau.

ABSTRACT

Third generation networks tend to integrate more and more services. Thus, standards and protocols such as the UMTS and Wireless Access Protocol (WAP) should allow any mobile user, in addition to telephony, to navigate the Web or have a videoconference call. However, the integration of services raises a certain number of problems. For example, the advent of new services (such as the Internet) has questioned the traditional models of traffic used in the telecommunication world. New models have then been proposed to take into account more complex traffic. But in third generation networks, even if several works have been dedicated to the study of the radio technologies performance, few have been done on the performance of these networks from an application perspective, when various types of traffic are multiplexed. Another problem raised by the combination of various traffic types is that of pricing. With traffic having different requirements, the operators must choose a type of pricing, which generates income, introduces a feedback and allows a certain control on the users by inciting them to minimize their costs.

This thesis has, as initial objective, the performance evaluation of the access part (or local loop) of cellular multiservice networks in a context where several types of service or traffic exhibiting self-similarity or long-range dependence are multiplexed. The second objective is to propose an improved model of pricing to efficiently take into account the users' constraints, the characteristics of the traffic and a judicious management of the network resources.

For that purpose, we first listed the various applications to be modeled. Then, we adapted the existing models or proposed new ones, which are based on the specifications of UMTS and sometimes, also on real observations. The models obtained were calibrated and validated by comparison with real traces and used to simulate a UMTS network aggregating various types of traffic. This defines a reference model whose performance was evaluated according to certain metrics of quality of service (QoS). Based on the results obtained, management and control policies were suggested to

ensure a better quality of service. A separate model was developed for assessing the impact of various topologies on the reliability of a third generation network implementation. Finally, after a review of the various methods of pricing suggested in the literature, we defined a certain number of criteria necessary in the framework of third generation networks. Based on these criteria, we analyzed the existing models and proposed an improved one which satisfies our specifications. More precisely, we proposed an extension of the effective bandwidth pricing scheme which encompasses the proposal of a criterion to take into account a jitter constraint in the admission control process and the use of a new family of pricing functions. These functions consider the uncertainties of the users, the burstiness of the traffic and also offer a higher degree of accuracy while keeping good incentive properties.

The main contributions of this thesis include the different models of traffic, network and pricing, as well as the evaluation of their performance. In a general way, with a good design, the performance of the network with the traffic considered is acceptable. However, it is necessary to carefully define the QoS profiles associated with each class or application and judiciously oversubscribe the radio link. For the streaming class, the traffic should be buffered to ensure an adequate level of QoS, particularly in terms of delay. The implementation of additional QoS management mechanisms at the higher levels also allows an improvement of the performance. Although the results obtained depend on a great number of parameters, we have assessed the impact of a certain number of policies like the installation of priority management at the IP level and the oversubscription of the connections.

Various network topologies were studied to estimate reliability. With the highly reliable equipment available, the redundancy necessary to ensure a great availability is minimal and can be reached with very simple topologies. A tree topology with some redundant links indeed suffices to interconnect the RNC and provides the desired availability with relatively low costs.

Our criterion to take into account the jitter constraint in the admission control process was validated by simulations. It allows improvements up to 10% on the bounds

and consequently on the number of connections admitted. The new family of pricing functions also yields lower costs for the user with no additional monitoring from the network. For the user, the savings due to the new pricing function increase with the size of the buffers used in the network nodes.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES SIGLES ET ABRÉVIATIONS

2G :	Seconde génération
3G :	Troisième génération
AICH :	Acquisition Indicator Channel
ATM :	Asynchronous Transfer Mode
CBR :	Constant Bit Rate
cdf :	cumulative distribution function
CDMA :	Code Division Multiple Access
CPCH :	Common Packet Channel
CS :	Circuit Switched
DCH :	Dedicated Channel
DCT :	Direct Cosine Transform
DSCH :	Downlink Shared Channel
DSLAM :	Digital Subscriber Line Access Multiplexer
DTC :	Délai de Transfert de la Cellule
FACH	Forward Access Channel
FDD :	Frequency Division Duplex
fdp :	fonction de densité de probabilité
fdr :	fonction de densité de répartition
FTP :	File Transfer Protocol
GGSN :	Gateway GPRS Serving/Support Node
GPRS :	General Packet Radio Service
GSM :	Global System for Mobile Communications
HIMM	High Interactive Multimedia
HLR :	Home Location Register
HMM :	High Multimedia
HTTP :	Hypertext Transfer Protocol
IP :	Internet Protocol

LMDS :	Local Multipoint Distribution Service
MAC :	Medium Access Control
MMDS :	Multichannel Multipoint Distribution Service
MMM :	Medium Multimedia
MPEG :	Moving Picture Experts Group
MSC :	Mobile Switching Center
MTBF :	Mean Time Between Failures
MTTF :	Mean Time To Failure
MTTR :	Mean Time To Repair
NNTP :	Network News Transport Protocol
nrt-VBR :	non real time Variable Bit Rate
NSS :	Network Switching Subsystem
PDP :	Packet Data Protocol
PDU :	Protocol Data Unit
PLMN :	Public Land Mobile Network
PS :	Packet Switched
QoS :	Quality of Service
RAB :	Radio Access Bearer
RACH :	Random Access Channel
RLC :	Radio Link Control
RNC :	Radio Network Controller
RNS :	Radio Network Subsystem
RRC :	Radio Resource Control
SD	Switched Data
SGSN :	Serving GPRS Support Node
SM :	Simple Messaging
SMS :	Single Messaging System
SMTP :	Simple Mail Transfer Protocol
TCP :	Transmission Control Protocol

TDD :	Time Division Duplex
TR :	Terminaison de Réseau
TTI :	Transmission Time Interval
UBR :	Unspecified Bit Rate
UDD :	Unconstrained Delay Data
UE :	User Equipment
UIT :	Union Internationale des Télécommunications
UM :	Unité Mobile
UMTS :	Universal Mobile Telecommunications System
UTRA :	Universal Terrestrial Radio Access
UTRAN :	Universal Terrestrial Radio Access Network
VBR :	Variable Bit Rate
VLR :	Visitor Location Register
VoD :	Video on Demand
WCDMA :	Wideband Code Division Multiple Access

CHAPITRE 1

INTRODUCTION

Un réseau de télécommunications est un ensemble d'équipements de communications reliés entre eux par des liaisons. Ces équipements et liaisons permettent l'échange de messages par des utilisateurs géographiquement dispersés. Pour que le réseau soit fonctionnel, son architecture doit être adaptée au service fourni. Malheureusement, ce n'est pas toujours le cas, parce que les composants de réseau ne sont pas toujours fiables [4,62,63]ou ont été conçus en ne tenant pas compte des caractéristiques particulières des nouveaux types de trafic. En effet, ces dernières années, une variété d'ordinateurs mobiles équipés de dispositifs de transmission sans fil sont devenus populaires et la révolution dans les réseaux, qui a commencé avec les technologies à large bande, a rendu possible de nombreux services et applications, tels que le transfert de fichier, la vidéoconférence, le courrier électronique (courriel), les interfaces graphiques d'utilisateur, les systèmes de fichiers à distance, etc. Ces développements ont permis d'envisager une évolution des systèmes cellulaires mobiles de deuxième génération vers une troisième génération plus performante en termes de capacité, de couverture et de qualité de service. Il en résulte un éventail de nouveaux services et de nouvelles applications, ce qui implique de nouveaux problèmes et défis liés à la qualité de service, à la performance et à la facturation et qui font l'objet de cette thèse. Dans ce chapitre d'introduction, après avoir défini les concepts de base, nous préciserons les éléments de la problématique ainsi que les objectifs de recherche. Par la suite, nous présenterons une esquisse de la méthodologie que nous appliquerons pour atteindre ces objectifs, suivie d'une synthèse des principales contributions escomptées et des grandes lignes du mémoire.

1.1 Définitions et concepts de base

Les réseaux cellulaires multiservices désignent des réseaux transportant à la fois de la voix, des données, du trafic multimédia et vidéo [34,39]. Leur architecture intègre des réseaux d'accès sans fil à large bande qui sont plus adaptés au trafic multimédia et aux applications qui génèrent un trafic en rafales [4,19]. Le concept de réseau cellulaire multiservice est alors employé pour dénoter l'infrastructure intégrée destinée à supporter ces applications et services dans un environnement de mobilité. Un exemple de réseau cellulaire multiservice est le système UMTS (Universal Mobile Telecommunications System) qui est un système de communications mobiles de troisième génération introduit par l'Union Internationale des Télécommunications (UIT) [124].

En fait, dans le cadre des systèmes de troisième génération, l'UIT a introduit six technologies d'accès radio dont l'UMTS qui, à son tour, regroupe les technologies *UTRA/FDD* (*Universal Terrestrial Radio Access/Frequency Division Duplex*) et *UTRA/TDD* (*Universal Terrestrial Radio Access/Time Division Duplex*). De plus, la majorité de ces techniques reposent sur une méthode d'accès de type *CDMA* (*Code Division Multiplex Access*) [118]. Le CDMA est une technique de multiplexage basée sur le principe de l'étalement de spectre. L'étalement de spectre est une méthode de transmission dans laquelle le signal occupe une largeur de bande plusieurs fois supérieure à la largeur de bande minimale nécessaire pour envoyer l'information. L'étalement est réalisé à l'aide d'un code qui est indépendant des données, et un récepteur synchronisé avec l'émetteur est utilisé pour récupérer le signal originel [128]. Pour plus d'informations sur le système CDMA, le lecteur pourra se référer au chapitre 2 et à [118,128].

Dans cette thèse, nous nous intéresserons plus spécifiquement à la technologie UMTS. Selon le forum UMTS, cette technologie a été conçue pour être un système global, avec des composants terrestres et satellites. La couverture est organisée de manière hiérarchique, avec des *picocellules* à l'intérieur des immeubles, des *microcellules* en environnement urbain dense, des *macrocellules* en environnement suburbain ou rural, et même des *cellules globales* desservies par des satellites. L'UMTS

jouera un rôle principal dans le futur marché grand public pour les transmissions sans fil de haute qualité. Ce marché est estimé à environ 2 milliards d'utilisateurs dans le monde entier pour l'année 2010. Les débits offerts par le système peuvent aller jusqu' à 2 Mbps [124]. L'accès radio dans l'UMTS fait appel à la technologie UTRA qui opère dans une largeur de bande de 5 MHz (en incluant les bandes de garde).

Comme l'illustre la Figure 1.1, un réseau UMTS comporte trois domaines : le *domaine de l'équipement usager*, le *domaine du réseau d'accès (UTRAN)* et le *domaine du réseau cœur*. Ces différents domaines sont détaillés au chapitre 2.

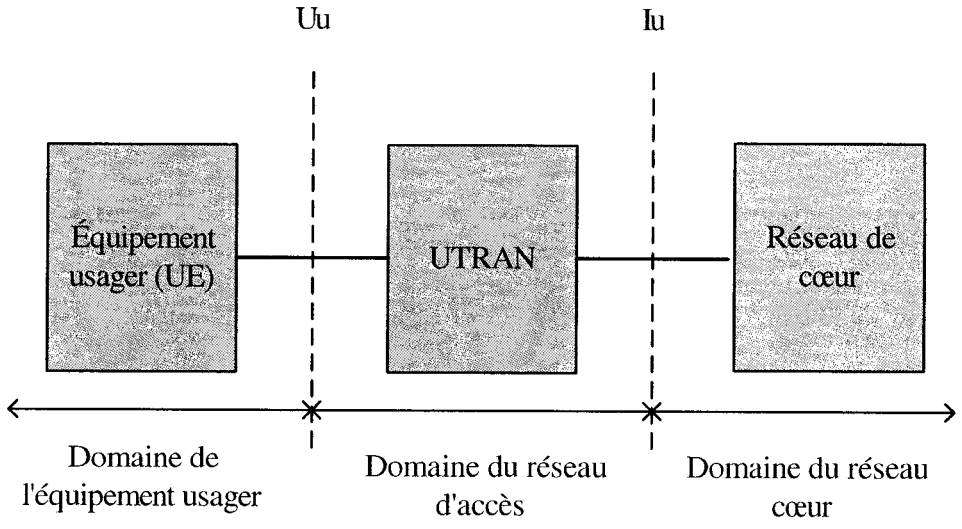


Figure 1.1 Les domaines d'un réseau UMTS

1.2 Éléments de la problématique

Dans les réseaux cellulaires multiservices, les caractéristiques particulières des nouveaux types de trafic posent de nouveaux défis [104]. En effet, la croissance exponentielle du trafic en rafales change la dynamique des réseaux et nécessite, lors de la conception des boucles d'accès, une bonne évaluation de différentes classes de service. De plus, la combinaison de la voix, des données, de la vidéo et du multimédia dans des réseaux sans fil à haut débit exige de nouvelles solutions aux problèmes de

facturation, de synchronisation dans les systèmes multimédia, de qualité de service, de technologies d'accès, de performance des services et des applications [34,77].

Pendant des décennies, le trafic des réseaux (détenus essentiellement par des opérateurs téléphoniques) se modélisait très bien par un ou plusieurs processus de Poisson. Ainsi, Habib et *al.* [50] proposaient en 1992 un modèle de source de voix qui était essentiellement un modèle ON-OFF avec des inter-arrivées distribuées de manière exponentielle. Même si ce modèle reste valable pour la voix sur les réseaux à commutation de paquets [7], l'avènement de nouvelles applications, l'intégration des services, les travaux de Leland et *al.* [74] sur le trafic dans les réseaux locaux et ceux de Paxson [107] sur le trafic dans les réseaux distribués ont ébranlé le modèle communément admis des processus de Poisson.

En effet, ce modèle implique qu'une agrégation du trafic sur un intervalle de temps suffisamment long donnerait un trafic plus régulier. Mais des mesures réelles [74,107] ont montré que le trafic présentait une variation significative sur plusieurs échelles de temps. Un tel trafic est qualifié d'*auto-similaire* ou *fractal* et présente des rafales (périodes prolongées d'activité bien au-dessus de la moyenne) sur plusieurs échelles de temps. De plus, un processus fractal peut aussi présenter une dépendance à long terme, i.e. caractérisée par le fait que les valeurs prises à un instant quelconque sont corrélées positivement et de façon significative avec les valeurs du processus à tout instant futur. Ceci implique que, contrairement au cas de données non corrélées, la variation de l'estimation de la moyenne d'une population de n échantillons ne décroît pas proportionnellement à $1/n$ mais plutôt proportionnellement à $n^{-\beta}$ $(0<\beta<1)$. Ainsi, la détermination de la moyenne de tels échantillons peut s'avérer délicate.

Soulignons qu'un trafic qui présente une dépendance à long terme est forcément auto-similaire, alors que l'inverse n'est pas toujours vrai. Il est montré par exemple dans [16,74,107] que le trafic vidéo ou le trafic dans un réseau local ou étendu présente une dépendance à long terme. Sur cette base, des études telles [47,101,102] montrent l'importance plus ou moins grande de la prise en compte des nouvelles caractéristiques du trafic dans les modèles de sources de trafic. De manière générale, les modèles ne

tenant pas compte de la dépendance à long terme génèrent des simulations qui sous-estiment de manière significative les mesures de performance telles le délai moyen des paquets ou la taille maximale du tampon [46,107,117,123]. De ce fait, la conception de réseaux intégrant du trafic fractal doit être soigneusement menée.

Avec le développement des réseaux de troisième génération, plusieurs études [20] se sont penchées sur les performances de leurs technologies radio. Cependant, aucune étude ne s'est vraiment intéressée aux performances de ces réseaux d'un point de vue « applicatif », lorsque différents types de trafic sont multiplexés. Il s'agit donc, dans un premier temps, d'étudier les performances de la partie accès (ou boucle locale) des réseaux cellulaires multiservices dans un contexte où sont multiplexés plusieurs types de service ou de trafic présentant souvent une dépendance (ou corrélation) à long terme et un caractère fractal et en rafales. De cette étude, seront déduits des indices de qualité de service permettant de déterminer la viabilité de solutions (comme l'UMTS) actuellement proposées pour les réseaux mobiles de troisième génération.

Un autre problème soulevé par la combinaison de différents types de trafic est celui de la tarification. En effet, avec des trafics ayant des exigences différentes, les opérateurs se heurtent au problème du choix de type et de la sorte de tarification à adopter. De manière générale, le rôle de la tarification n'est pas seulement de générer des revenus. Elle introduit aussi une rétroaction qui permet d'exercer un certain contrôle sur les usagers en les incitant à réagir aux tarifs par une minimisation de leurs coûts. Par exemple, une politique de tarification appropriée devrait jouer - même partiellement - le rôle d'un mécanisme de contrôle de congestion en encourageant une utilisation efficace des ressources du réseau, et aussi inciter les utilisateurs du réseau à regrouper leurs besoins dans des classes de service appropriées. La mise en place d'une politique de tarification qui transmet les bons incitatifs aux usagers et qui satisfait un certain nombre de propriétés constitue donc une préoccupation pertinente et un défi de recherche.

1.3 Objectifs de recherche

L'objectif principal de cette thèse est d'étudier les performances de la partie accès (ou boucle locale) des réseaux cellulaires multiservices dans un contexte où sont multiplexés plusieurs types de service ou de trafic présentant souvent une dépendance (ou corrélation) à long terme et un caractère fractal et en rafales. Il en résultera des indices de qualité de service permettant de déterminer la viabilité de solutions (comme l'UMTS) actuellement proposées pour les réseaux mobiles de troisième génération. Cette thèse cherche également à proposer un modèle de tarification amélioré pour permettre une prise en compte efficace des contraintes des usagers et une gestion judicieuse des ressources du réseau.

Plus spécifiquement, nous visons les objectifs suivants :

1. élaborer, par une caractérisation théorique du trafic, un modèle de simulation pour représenter les divers types de trafic en jeu ;
2. utiliser les modèles de trafic pour simuler le cas des réseaux UMTS avec l'agrégation de différents types de trafic ;
3. proposer des métriques de qualité de service et les utiliser pour évaluer la viabilité de la solution UMTS ;
4. suggérer des politiques de gestion et de contrôle pour assurer une meilleure qualité de service ;
5. proposer un modèle de tarification amélioré pour permettre une prise en compte efficace des contraintes des usagers et une gestion judicieuse des ressources du réseau.

1.4 Esquisse méthodologique

Pour atteindre l'objectif spécifique 1, qui consiste à élaborer des modèles pour représenter les divers types de trafic en jeu, nous commencerons d'abord par recenser les différentes applications à modéliser. Cette énumération se fera en se fondant sur différentes spécifications de UMTS et une revue sélective de littérature des modèles

existants. Ensuite, nous adapterons les modèles existants ou en proposerons de nouveaux en nous basant encore sur les spécifications de UMTS et parfois aussi sur des observations réelles. Les modèles obtenus seront calibrés et validés par comparaison avec des traces réelles et seront utilisés dans la réalisation de l'objectif spécifique 2. Il s'agira de simuler un réseau UMTS avec l'agrégation de différents types de trafic. Nous recenserons donc un certain nombre de paramètres d'installation en analysant l'architecture UMTS (types de trafic, proportions, type de clientèle visé, caractéristiques techniques, etc.) pour en proposer ou adapter un modèle. Quelques éléments d'intérêt dans ce cas seront les modèles de propagation, de transmission, de gestion de priorité qui sont détaillés à l'Annexe I. Ensuite, nous établirons un modèle de référence. Ce modèle de référence sera une implémentation aussi fidèle que possible des spécifications de l'UMTS [124].

Les performances du modèle de référence seront estimées en fonction de certaines métriques de qualité de service pour évaluer la viabilité de la solution UMTS. Des politiques de gestion et de contrôle pour assurer une meilleure qualité de service seront suggérées à partir des observations faites et des résultats obtenus. Nous étudierons, entre autres, l'impact de divers niveaux et politiques de QoS et l'évolution des performances en fonction de la charge. Au niveau des performances, nous ferons une distinction entre la fiabilité et les autres paramètres de performance. Nous développerons donc un modèle séparé pour une étude de la fiabilité d'une implémentation d'un réseau de troisième génération. L'accent sera mis sur l'impact de diverses topologies sur les performances de fiabilité et les résultats seront obtenus par simulation.

Finalement, après un survol des différentes méthodes de tarification proposées dans la littérature, nous définirons un certain nombre de critères requis pour notre modèle de tarification. À partir de ces critères, nous analyserons les modèles existants et essayerons de proposer un modèle amélioré qui satisfait nos spécifications et qui s'adapte au cadre des réseaux de troisième génération qui intègrent plusieurs types de trafic. La méthodologie proposée sera mise en œuvre et les résultats validés. Nous la

comparerons aussi à d'autres approches résultant de travaux antérieurs dans la littérature.

1.5 Principales contributions

Les principales contributions de cette thèse s'articulent autour de deux grands axes : la modélisation de performance à des fins d'évaluation d'un réseau mobile multiservice de troisième génération, la proposition d'un modèle de tarification adapté aux réseaux 3G. De manière plus spécifique, nous avons adopté une approche qui consiste à étudier les performances d'un point de vue « applicatif » d'un réseau 3G qui intègre différents types de trafic et de services. Ceci constitue une nouvelle approche puisque les travaux antérieurs s'intéressaient plus aux performances de la couche radio des différentes technologies d'accès 3G. L'objectif visé est d'avoir une meilleure vision des capacités des réseaux de troisième génération du point de vue de l'utilisateur et de proposer des politiques de gestion et de contrôle pour assurer une meilleure qualité de service. Pour cela, un certain nombre d'apports secondaires ont été faits. En effet, les modèles de trafic ont été adaptés au cadre des réseaux de troisième génération par un processus de calibrage et de validation. Pour l'application Fax, aucun modèle n'existait dans la littérature. Un modèle simple certes, mais nouveau, a donc été proposé. Finalement, sur le plan de la fiabilité, diverses topologies de réseaux de troisième génération ont été étudiées.

En ce qui a trait au modèle de tarification, après avoir analysé les meilleurs modèles présents dans la littérature, nous avons présenté une nouvelle extension du schéma de tarification basée sur le débit effectif. Cette extension consiste en une famille de fonctions de tarification qui améliorent globalement le schéma en tenant compte des incertitudes des usagers et de la sporadicité du trafic. Ces fonctions offrent aussi une plus grande précision tout en gardant les bonnes propriétés sur les incitatifs. En outre, pour toutes ces améliorations, le système de facturation ne contrôle, tout comme dans le schéma classique, que le débit moyen de chaque connexion. Ensuite, un critère pour

intégrer dans le schéma de tarification des contraintes sur la variation de délai a été introduit. En effet, dans le schéma classique, la gigue est au mieux traitée comme un délai, ce qui sous-estime la zone d'admission. Les extensions et le critère proposés ont été validés à travers des simulations et les résultats obtenus comparés avec ceux du schéma classique. On obtient des améliorations de la zone d'admission pouvant aller jusqu'à 10 %.

1.6 Plan de la thèse

Cette thèse comprend six chapitres. Suivant ce premier chapitre d'introduction, le chapitre 2 présente une revue de littérature qui couvre les domaines de la modélisation d'applications et de la tarification. Le chapitre 3 décrit les différents modèles d'applications et de trafic, ainsi que l'adaptation du modèle UMTS. Il présente aussi le plan d'expérience pour l'évaluation de performance et analyse les résultats obtenus. Dans le chapitre 4, nous exposons et analysons la fiabilité d'une implémentation d'un réseau de troisième génération en fonction des diverses topologies adoptées. Le chapitre 5 introduit les critères d'évaluation des modèles de tarification, le modèle de tarification proposé et son extension ainsi qu'une analyse de ses performances. Enfin, le chapitre 6 résume les principaux résultats obtenus, les limitations de nos travaux ainsi que les extensions possibles.

CHAPITRE 2

TRAFIC ET TARIFICATION DANS LES RÉSEAUX

Traditionnellement, l'évaluation de performance est considérée comme une étape de la conception des réseaux et est en général traitée séparément des problèmes de tarification qui sont plutôt de nature économique. De ce fait, ce chapitre présentera d'abord quelques définitions et concepts de base qui seront utiles pour le reste de cette thèse. Ensuite, nous ferons une revue des problèmes de modélisation de trafic, puis ceux de tarification. Cette revue nous permettra d'introduire les problèmes que nous aborderons ainsi que les apports escomptés.

2.1 Définitions et concepts de base

Ces dernières années, la révolution dans les réseaux qui a commencé avec les technologies à large bande a rendu possible une grande variété de services et d'applications, tels que le transfert de fichier, la vidéoconférence, le courrier électronique, les interfaces graphiques d'utilisateur, les systèmes de fichiers à distance, etc. Ces développements ont permis d'envisager une évolution des systèmes cellulaires de deuxième génération vers une troisième génération plus performante tant en termes de capacité et de couverture que de variété et de qualité de service.

Les technologies introduites par l'IUT dans le cadre des systèmes de troisième génération (3G) sont assez variées. Cependant, elles satisfont les caractéristiques techniques suivantes [24,118] :

- un seul système intégré donnant un accès facile et uniforme aux différents services ;
- une différentiation entre les services offerts par différents réseaux dans différents environnements ;

- des services à haut débit avec un minimum obligatoire de 144 kbps pour tous les types d'environnement. Dans des environnements intérieurs avec une mobilité restreinte, le débit peut atteindre 2 Mbps ;
- une qualité de parole comparable à celle des réseaux câblés ;
- une transmission aussi bien symétrique qu'asymétrique des données ;
- des services de commutation de circuits et de commutation de paquets ;
- une capacité et une efficacité spectrale supérieures à celles des systèmes cellulaires de seconde génération ;
- la possibilité d'offrir des services multimédias avec des qualités de service différentes ;
- une compatibilité avec les réseaux d'accès de deuxième génération (2G) ;
- la capacité d'itinérance (roaming) entre les différents systèmes de troisième génération ;
- une couverture universelle associant des satellites aux réseaux terrestres.

La majorité de ces techniques reposent sur une méthode d'accès de type *WCDMA* (*Wideband Code Division Multiple Access*) [118]. Le CDMA est une technique de multiplexage basée sur le principe de l'étalement de spectre. L'étalement de spectre est une méthode de transmission dans laquelle le signal occupe une largeur de bande plusieurs fois supérieure à la largeur de bande minimale nécessaire pour envoyer l'information. Selon [118], les principales qualités de l'étalement de spectre sont :

- la robustesse vis-à-vis des interférences intentionnées et non intentionnées ;
- la sécurisation des communications et faible probabilité de détection ;
- la robustesse vis-à-vis des dégradateurs engendrés par les trajets multiples grâce au gain de traitement.

L'étalement est réalisé à l'aide d'un code qui est indépendant des données, et un récepteur synchronisé avec l'émetteur est utilisé pour récupérer le signal originel [128]. Dans un système à étalement de spectre, on accroît la largeur spectrale du signal afin de réduire le rapport signal à bruit au minimum et obtenir de bonnes performances. Le rapport entre la largeur de bande occupée par le signal après et avant étalement est

appelé *gain de traitement*. L'étalement se fait par une réduction de la densité spectrale du signal. Cette réduction correspond à un facteur égal au gain de traitement. Dans un système CDMA à séquences directes tel que l'UTRA, le signal d'information est directement modulé par un code qui possède certaines propriétés statistiques. Un générateur de codes produit un code d'étalement dont les composantes ont une durée T_c et une amplitude de +1 ou de –1. Chaque élément du code est appelé *chip*. Le message à transmettre est constitué de symboles d'information de durée T_s. Le débit chip est souvent fixe et égal à 3.84 Mcps dans la technologie UTRA. Une fois le code généré, l'étalement consiste à multiplier les symboles d'informations par les chips du code. On passe donc d'un signal bande étroite à un signal large bande, étant donné que $T_s > T_c$. Finalement, le signal étalé est modulé en fréquence radio et mis sur une porteuse de fréquence f_c. Le *facteur d'étalement* (SF : Spreading Factor) est égal au rapport entre le débit chip et le débit symbole et est égal au gain de traitement dans le cas où le débit binaire utile du signal d'information est égal au débit symbole.

Dans un système CDMA, on distingue en fonction de la largeur de bande occupée par le signal étalé, les systèmes CDMA à large bande (*WCDMA*) et ceux à bande étroite. Les avantages des systèmes WCDMA tels l'UTRA sont :

- un gain de traitement plus élevé ;
- la possibilité d'offrir des services à haut débit ;
- de meilleures performances pour détecter des trajets multiples ;
- la possibilité de déploiement dans un spectre de fréquence déjà utilisé.

Cependant, ces avantages ne vont pas sans un certain nombre d'inconvénients :

- une complexité du support matériel et logiciel étant donné que le débit chip est plus élevé ;
- une interférence mutuelle entre les utilisateurs d'une même cellule ;
- la nécessité d'une synchronisation temporelle plus précise ;
- la nécessité d'un contrôle de puissance rapide.

Le contrôle de puissance est réalisé au niveau physique en général. La Figure 2.1 présente l'architecture en couche des protocoles sur l'interface radio.

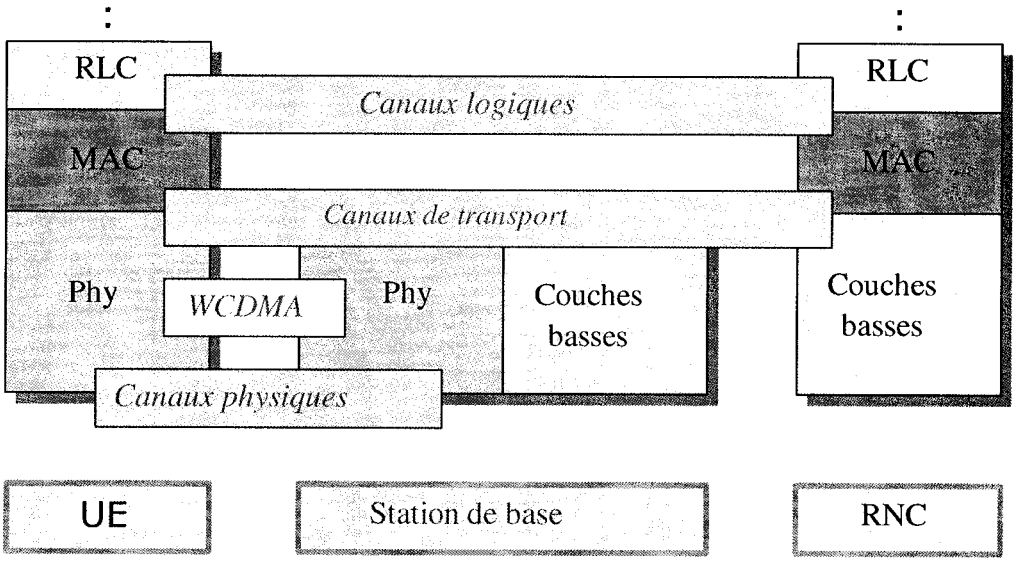


Figure 2.1 Architecture en couche des protocoles sur l'interface radio

Les couches *RLC* (*Radio Link Control*) et *MAC* (*Medium Access Control)* constituent la couche liaison du protocole d'accès de l'UTRA. La couche RLC utilise les services de la couche MAC par l'intermédiaire de *canaux logiques* qui sont définis par le type d'information (information de contrôle ou données de l'usager) transportée. La couche RLC fournit trois types de services :

- Un service transparent où les unités de données (PDU pour Protocol Data Unit) des couches supérieures sont segmentées et transmises de manière transparente aux couches inférieures et vice versa pour le réassemblage des PDU des couches inférieures. Dans ce mode, la couche RLC n'ajoute pas d'entête aux segments générés.
- Un service non acquitté où les PDU sont segmentés (respectivement réassemblés), une entête RLC est ajoutée (respectivement enlevée) à chaque segment. Chaque segment possède un numéro de séquence et les segments manquants ne sont pas retransmis.
- Un service acquitté similaire au service non acquitté, sauf que les segments manquants sont retransmis.

La couche MAC met en correspondance les canaux logiques et les canaux de transport qui sont des services offerts par la couche physique à la couche MAC. Un canal de transport est unidirectionnel et est caractérisé par la façon dont les données sont transmises au niveau de l'interface radio, alors qu'un canal physique est défini par une fréquence, un facteur d'étalement, un code d'embrouillage. Le flot d'information est protégé au niveau physique par des codes de convolution et ensuite découpé en un ensemble de blocs de transport qui seront transférés sur le canal de transport à un intervalle régulier dénommé TTI (Transmission Time Interval). Les codes de convolution introduisent tout simplement une redondance au niveau des données transmises.

L'implémentation de la qualité de service se fait par l'assignation de niveaux de priorité aux différentes classes ou services. Ces priorités sont gérées au niveau de la couche MAC. Les couches situées au-dessus la couche RLC sont en général des couches correspondant au niveau 3 (ou plus) du modèle OSI. Ces couches peuvent aussi mettre en place des mécanismes de différenciation ou de réservation. Par exemple, dans un réseau de troisième génération basée sur la commutation de paquets, on utilise le protocole TCP/IP au dessus de la couche RLC.

Comme illustré à la Figure 2.2, un réseau UMTS comporte trois domaines : le *domaine de l'équipement usager*, le *domaine du réseau d'accès (UTRAN)* et le *domaine du réseau cœur*. L'interface *Uu* relie le domaine de l'équipement usager au domaine du réseau d'accès, tandis que l'interface *Iu* connecte le réseau d'accès au réseau cœur.

L'équipement usager (*UE*) regroupe les différents types de terminaux que l'usager peut utiliser (PDAs, ordinateurs, téléphones, etc.) pour accéder à l'infrastructure et aux services par le biais de l'interface radio Uu.

Le réseau d'accès fournit à l'UE les ressources nécessaires pour accéder au réseau cœur. Plus précisément, il se charge de transférer les données générées par l'UE vers le réseau cœur, de la gestion de l'admission au réseau et de la mobilité, du contrôle de congestion et de la diffusion des informations systèmes. L'UTRAN s'occupe aussi du chiffrement et du déchiffrement des données, des fonctions liées à la gestion de la

ressource radio, ainsi que de la synchronisation. Dans l'UTRAN, on retrouve des équipements tels que le *nœud B* et le *contrôleur du réseau radio* (*RNC*).

Le nœud B assure la transmission et la réception radio entre l'UTRAN et un ou plusieurs équipements usagers. Il peut desservir plus d'une cellule si on utilise des antennes sectorielles. Dans ce cas, le nœud B gère la *relève simple* (*soft-handover*) au niveau des UE qui traversent ces cellules. Il s'occupe aussi du contrôle de puissance de l'UE et de l'adaptation du trafic de telle sorte qu'il puisse être acheminé par l'interface Uu. Il est relié par une interface *Iub* au RNC.

Le RNC équivaut au *contrôleur de stations de base* des réseaux de deuxième génération. Chaque nœud B est associé à un RNC qui gère l'ensemble des ressources radio des nœuds B sous son contrôle. Le RNC se charge de la gestion de l'admission au réseau et des liens radio. Il intervient aussi dans la gestion de la relève, dans le mécanisme de contrôle de puissance dit « hors boucle » et dans les mécanismes de relocalisation. Il peut être relié à un autre RNC par une interface *Iur*. Un RNC avec l'ensemble des nœuds B qui lui sont associés constituent un *sous-système du réseau radio* (*RNS*).

Le réseau cœur est composé de l'ensemble des équipements qui assurent des fonctions telles que le contrôle des appels, le contrôle de la sécurité, la gestion de l'interface avec les réseaux externes. Il permet aux usagers de communiquer à l'intérieur d'un même réseau de téléphonie mobile, assure l'interconnexion avec les réseaux externes et permet de maintenir des communications sécuritaires même lorsque l'usager est en déplacement. Le réseau cœur est subdivisé en domaines à *commutation de cir*cuits (*CS*) et à *commutation de paquets* (*PS*). Chacun de ces domaines est relié à l'UTRAN à travers une interface *Iu-CS* pour le domaine à commutation de circuits et *Iu-PS* pour le domaine à commutation de paquets. Pour des raisons de compatibilité, le domaine à commutation de circuits ressemble beaucoup au *sous-système de commutation et d'acheminement* (*NSS*) des systèmes de deuxième génération. Quant au domaine PS, ces principaux composantes sont : le *nœud de service GPRS* (*SGSN*), le *nœud passerelle du GPRS* (*GGSN*) ainsi que le *répertoire de réseau*, appelé *HLR* (Home Location Register).

Le nœud de service GPRS peut être comparé, du point de vue fonctionnel, à l'ensemble MSC/VLR des systèmes de deuxième génération. Il est en charge de l'acheminement des paquets de données et s'occupe du routage, de la gestion de la mobilité (procédure d'attachement et de détachement) et des procédures d'authentification.

Le nœud passerelle du GPRS assure l'interconnexion entre le réseau cœur et les réseaux à commutation de paquets externes. Le GGSN et le SGSN sont interconnectés par un réseau fédérateur (*Intra-PLMN*) qui repose sur le protocole IP.

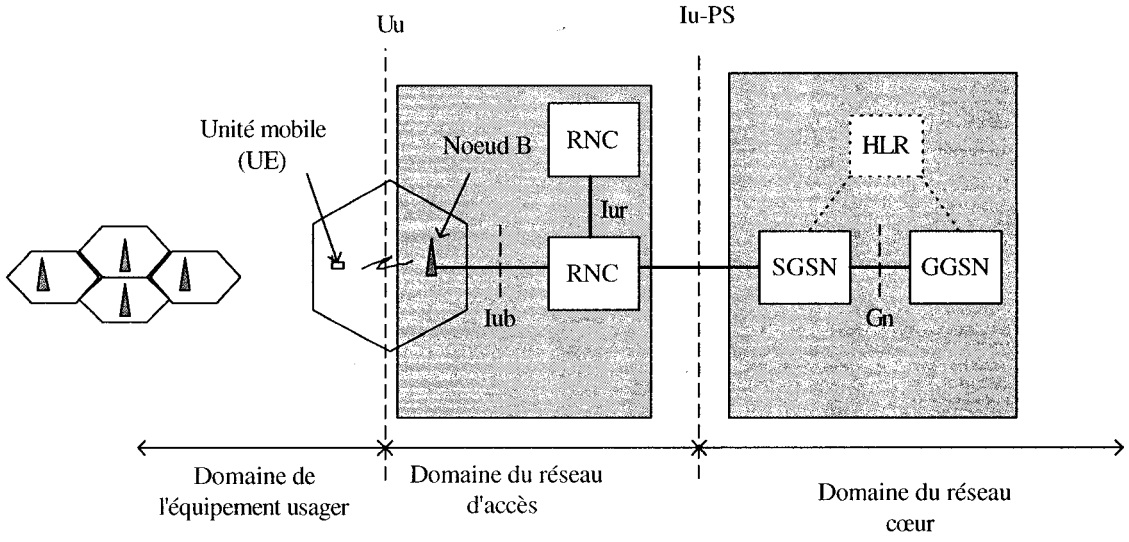


Figure 2.2 Architecture détaillée d'un réseau UMTS

Il est à noter que les liens câblés du réseau d'accès utilisent la technologie ATM pour assurer le transport à l'intérieur du réseau de flots d'information dont les débits et exigences de *qualité de service* (*QoS*) varient. En effet, dans les réseaux intégrés tels UMTS [124] ou ATM, l'utilisateur peut établir avec le réseau un contrat de service qui définit les niveaux de service convenus. Ce contrat de service comprend les éléments suivants :

- Un ensemble de descripteurs de la connexion : l'ensemble de descripteurs permet de situer le trafic de la connexion dans des classes de trafic ou de

service prédéfinies. Ces classes de trafic fournissent des qualités de service différentes au niveau du débit, des garanties de délai, etc. L'idée derrière ces catégories de service est d'obtenir un meilleur mécanisme pour contrôler les ressources limitées du réseau, telles que la capacité de la liaison ou la taille des tampons. On suppose que l'abonné ou le fournisseur de service peut associer le trafic qu'il génère à une des catégories prédéfinies. Sachant la catégorie de service, le réseau a plus d'informations sur le profil du trafic et peut ainsi, lors de la demande de connexion, prendre une meilleure décision quant à la disponibilité de ressources pour fournir le service demandé.

- Un ensemble de paramètres de qualité de service qui donne à l'utilisateur une idée du niveau de service qu'il peut s'attendre à recevoir du réseau.
- Un ensemble de règles de vérification de conformité qui servent à assurer le respect du contrat de connexion. Les utilisateurs qui, d'après les règles de conformité établies, ne respectent pas le contrat initial sont susceptibles d'être pénalisés.

L'UMTS prévoit quatre classes de service en fonction de la QoS. Par ordre décroissant de priorité, on retrouve : la classe des services conversationnels, celle des services « en flux » (streaming), celle des services interactifs et finalement les services « en arrière-plan » (background). La différence essentielle entre ces différentes classes repose sur leur tolérance au délai et à la variation de délai.

2.1.1 Services conversationnels

Cette classe de trafic est destinée à des applications interactives en temps réel qui nécessitent des contraintes strictes de délai et de variation de délai dictées par la perception humaine. Les contraintes sont très strictes car leur violation entraîne une dégradation inacceptable de la qualité. Des exemples d'applications sont la téléphonie, la voix sur IP, la vidéoconférence.

2.1.2 Services « en flux »

L'une des caractéristiques essentielles de cette classe de service est qu'elle correspond à des applications qui envoient en temps réel des données principalement dans un seul sens. Le niveau d'interactivité requis est donc largement inférieur au cas précédent. Les données transférées sont souvent destinées à un être humain, et de ce fait, les contraintes sur la variation de délai sont très strictes même si celles sur le délai ne le sont pas. En général, le délai est limité par les performances de l'équipement et de l'application plutôt que par les limites de la perception humaine. L'exemple typique d'application est la vidéo sur demande (VoD).

La vidéo sur demande est un terme regroupant un large ensemble de technologies permettant à des personnes de choisir des vidéos à partir d'un serveur central [34]. La vidéo sur demande peut être utilisée dans un but récréatif, pour l'éducation, ou pour enrichir des présentations avec des clips vidéos. Puisque notre cible principale est le marché grand public, nous considérerons que la VoD est principalement utilisée dans un but récréatif. Les applications suggérées sont des clips vidéo sur demande, des clips sonores sur demande, de courtes séquences vidéos pour mettre en valeur une présentation [124]. Un système de vidéoconférence peut être employé pour tenir une conférence par l'intermédiaire d'un téléphone ou d'une connexion réseau. Une fois qu'une vidéoconférence est établie, un groupe peut partager des applications et marquer des informations sur un tableau commun. La différence fondamentale entre VoD et vidéoconférence se situe dans le fait que la VoD est typiquement un trafic asymétrique alors que la vidéoconférence est symétrique. Cependant, au niveau du lien descendant (*DL*), les caractéristiques des deux types de trafic sont assez similaires.

2.1.3 Services interactifs

Contrairement aux services conversationnels, les services interactifs fonctionnent en mode requête/réponse. L'usager peut être une machine ou un humain qui espère recevoir la réponse à sa requête dans un délai donné. Grâce aux exigences plus souples de délai, le taux d'erreur est en général très bas. Parmi les applications de cette classe,

nous pouvons citer la navigation sur Internet, l'accès à des bases de données et à des serveurs.

2.1.4 Services « en arrière-plan »

L'usager est typiquement une machine qui effectue un certain nombre de tâches en arrière-plan. La contrainte de délai est donc quasi-inexistante. Toutefois, comme dans le cas des services interactifs, le taux d'erreur est en général très bas. Le courriel, les SMS (Short Messaging Service) sont quelques exemples d'applications en arrière-plan.

Les spécifications du forum UMTS [124] retiennent en général 6 types de services représentatifs des quatre classes de service. Le Tableau 2.1 présente les divers types de service, des exemples d'application dans chacun des types de service et une correspondance possible entre les applications et les classes de service.

Tableau 2.1 Types de service

Type	Definition	Symétrique/ asymétrique	Exemples	Classe de service
HIMM	High Interactive Multimedia	symétrique	vidéoconférence, video -téléphonie, téléprésence	service conversationnel ou en flux
HMM	High Multimedia	asymétrique	video-clips sur demande, audio-clips sur demande, achats en ligne	service en flux
MMM	Medium Multimedia	asymétrique	Accès Internet, jeux interactifs	service interactif
SD	Switched Data	symétrique	Fax, Accès Internet	service interactif
SM	Simple messaging	symétrique	Courriel	service en arrière-plan
S	Speech	symétrique	téléconférence, voix	service conversationnel

2.2 Évaluation de performance et modélisation de réseaux

Avant l'avènement des réseaux à commutation de paquets tels le réseau IP, le trafic des réseaux constitué essentiellement de trafic de voix se modélisait très bien par un ou plusieurs processus de Poisson. Dans les années 90, les travaux de Leland et *al.* [74] sur le trafic dans les réseaux locaux et ceux de Paxson [107] sur le trafic dans les réseaux distribués ont ébranlé le modèle communément admis des processus de Poisson. De nouvelles caractéristiques comme l'auto-similarité et la dépendance à long terme sont découvertes pour le trafic de paquets. Des travaux tels ceux de [47] se sont ensuite

penchés sur l'impact de la prise en compte de ces nouvelles caractéristiques dans les modèles de sources de trafic. De manière générale, les modèles ne tenant pas compte de la dépendance à long terme génèrent des simulations qui sous-estiment de manière significative les mesures de performance telles le délai moyen des paquets ou la taille maximale du tampon [107]. Plus précisément :

- L'agrégation de plusieurs trafics auto-similaires présentera aussi un caractère auto-similaire. Le trafic agrégé présentera des variations importantes sur plusieurs échelles de temps (même si ces variations sont plus ou moins atténuées par rapport à celle du trafic initial [122]) et des corrélations positives, ce qui affectera la qualité de service aux usagers.
- Le taux de perte des paquets peut être de plusieurs ordres de grandeur supérieur aux prévisions obtenues avec les modèles markoviens.
- L'augmentation de la taille des tampons n'a pas un effet significatif sur la réduction du taux de perte des paquets. En général, les trafics fractals exigent des tampons très grands qui affectent le délai des cellules. Cela suppose, au niveau de la qualité de service, un compromis entre délai et taux de perte des paquets.
- La qualité de service se dégrade très rapidement avec l'augmentation de l'utilisation du réseau.

De ce fait, plusieurs études se sont penchées sur les performances de réseaux multiplexant des trafics fractals ou présentant une dépendance à long terme. Dans cette revue, nous nous intéresserons à des trafics fractals qui pourraient se retrouver sur des réseaux de troisième génération. À titre d'illustration, citons le trafic vidéo (Vidéo sur Demande ou vidéoconférence) et le trafic Internet (Navigation sur le web).

Dans [69], les caractéristiques et le multiplexage de plusieurs sources vidéo MPEG-1 ont été étudiés. Le modèle proposé est fondé sur un extrait du film « Le Magicien d'Oz ». Les résultats obtenus valident le modèle pour un multiplexage allant jusqu'à 10 sources vidéo en ce qui a trait à la taille des tampons et au taux de perte des paquets. Dans [71], Krunz et Tripathi proposent un modèle auto-régressif d'ordre 2 pour

capturer les variations de débit à long terme. Giordano et *al.* [46], quant à eux, proposent un nouveau modèle d'admission de connexions tenant compte du caractère fractal du trafic VBR vidéo dans les réseaux ISDN à large bande. L'admission d'une connexion se fait sur la base des caractéristiques statistiques du trafic véhiculé par celle-ci. Ces caractéristiques, comme par exemple le paramètre de Hurst, la moyenne et la variance du trafic, sont obtenues par des estimateurs en temps réel ou par une analyse différée. Les auteurs évaluent le gain de multiplexage obtenu en agrégeant N sources homogènes de trafic fractal obtenues à partir d'un mouvement brownien fractionnaire.

Heyman et Lakshman [54] proposent une méthodologie pour déduire des modèles pour des séquences vidéo avec des changements fréquents de scène. Les modèles sont en général basés sur les chaînes de Markov et les processus discrets auto-régressifs (DAR). La pertinence des modèles est évaluée par rapport à leur capacité à simuler fidèlement les pertes de cellules dans les réseaux ATM. Les mêmes auteurs [73] utilisent de longues séquences de vidéoconférence pour étudier les propriétés statistiques et proposer des modèles basés sur des processus autorégressifs d'ordre 1 et 2 et sur des chaînes de Markov. Dans [52], le modèle est généralisé pour le cas de l'utilisation de différents codecs comme H.261. Les différents types de vidéo sont étudiés et classifiés dans [23]. Les auteurs notent que les différents types de vidéos (sport, film, vidéoconférence) peuvent être regroupés en catégories selon le rapport écart-type/moyenne des différents types de trames MPEG.

Une autre contribution importante dans le domaine de la caractérisation et de la modélisation du trafic vidéo vient de Rose [115], qui propose deux modèles analytiques de différentes complexités pour simuler le trafic MPEG-1 vidéo à débit variable. Cependant, la principale contribution est la mise en ligne par l'auteur d'une banque de traces de séquence vidéo qui constitue aujourd'hui une référence dans le domaine. Dans la même lancée, Fitzek et Reisslein [44] proposent une archive de traces de vidéos MPEG-4 et H.263. Ils analysent aussi les propriétés statistiques desdites traces. Pour une revue assez complète des différents modèles de source vidéo à débit variable, le lecteur pourra se référer à [61].

L'autre type de trafic qui exhibe un caractère fractal est le trafic Internet. En effet, avec la croissance de l'Internet, le trafic sur les réseaux distribués a augmenté de façon considérable et constituera la majeure partie de la charge sur les réseaux futurs, y compris les réseaux mobiles. Paxson [106-108] a été l'un des premiers à étudier le trafic dans des réseaux distribués et à montrer que celui-ci présente des caractéristiques d'auto-similarité. Les trafics étudiés par Paxson variaient de sessions TELNET et FTP à NNTP et SMTP. Les traces récoltées par Paxson sont disponibles sur le site de « Internet Traffic Archive » [60]. Barford et Crovella [10,11] ont, quant à eux, étudié spécifiquement le modèle HTTP. Ils proposent un outil appelé SURGE pour modéliser des charges Web représentatives. Liu et *al.* [81] étudient aussi de leur côté des traces de trafic HTTP et arrivent à un modèle similaire à celui implémenté dans SURGE. Dans [10], les auteurs étudient l'évolution des caractéristiques du trafic HTTP sur des ensembles de données de 1995 et 1998. De toutes ces études, il ressort que le type de distribution qui modélise la taille des documents (ou fichiers) est relativement invariant malgré l'évolution de l'Internet. Finalement, les études de traces HTTP les plus récentes ont été menées par Arlitt et Jin [8] qui ont analysé les traces des serveurs de la coupe du Monde 98 en France. Là aussi le modèle de distribution pour la taille des documents (ou fichiers) reste toujours la même.

Le modèle généralement admis pour la voix dans les réseaux à commutation de paquets est un modèle ON-OFF où les périodes ON et OFF ont des durées suivant une distribution exponentielle et où la période ON correspond aux impulsions vocales. Ce modèle a été proposé entre autres par [21] et utilisé dans plusieurs articles dont [28,30,31].

Les réseaux de troisième génération, malgré leur avènement relativement récent, ont fait l'objet de nombreuses spécifications [1,124] et de nombreux articles tant au niveau de la technologie CDMA qui sous-tend la couche radio de UMTS [3,5,20,86,90] que de l'architecture et du déploiement des réseaux UMTS [72,75,91,113]. Cependant, de manière générale, la communauté scientifique s'est intéressée aux performances radio

et non aux performances du système vu d'un niveau applicatif. Par conséquent, les modèles de trafic utilisés sont souvent simplifiés [3,20,80].

À l'opposé, il existe toute une série d'articles qui étudient et analysent les performances de réseaux fixes (de type ATM ou IP) d'un point de vue applicatif lorsque plusieurs types de trafic (dont principalement des trafics auto-similaires) sont multiplexés. En général, ces études servent à déduire un mécanisme de contrôle des admissions de connexions. Norros [98,99] étudie de manière théorique les performances (utilisation, débit, taille des tampons) d'un système recevant un agrégat de trafic auto-similaire. Son modèle de trafic est fondé aussi sur un mouvement brownien fractionnaire. De plus, il suppose un trafic entrant suffisamment gaussien (combinaison d'un assez grand nombre de sources auto-similaires) et utilise une distribution de Weibull pour estimer la queue (quantile supérieur) de la distribution de la taille du tampon. Il en déduit la largeur de bande effective dans le cas d'un trafic fractal.

Dans [12,126], les auteurs étudient le multiplexage statistique de sources vidéo auto-similaires et homogènes. Ils proposent un algorithme d'admission de connexion et utilisent les résultats théoriques obtenus dans [98,99] avec la notion de largeur de bande effective ainsi qu'un modèle de trafic vidéo pour évaluer les performances de leur algorithme. Elwalid et Mitra [38] s'intéressent aussi au multiplexage et à l'algorithme d'admission de connexions dans le cas de sources de trafic vidéo. Pour eux, le trafic vidéo peut être classé en trafic pouvant être statistiquement multiplexé et en trafic ne pouvant l'être. Quand ces deux types de trafic vidéo sont agrégés, des interactions non linéaires dégradent les performances.

Stathis et Maglaris [122] utilisent les résultats de [98,99] pour présenter une méthode de dimensionnement de réseau et d'admission de connexion dans le cas de trafic fractal dans un réseau distribué. Ils proposent, entre autres, un modèle de trafic fondé sur des traces réelles, le comparent aux données réelles et présentent une étude théorique du gain de multiplexage obtenu avec plusieurs sources conçues selon leur modèle. Lindberger propose dans [79] le multiplexage de deux types de trafic : un trafic

à débit constant et un trafic à débit variable, et estime les gains résultant de ce multiplexage.

Finalement, dans [37], les auteurs comparent plusieurs stratégies d'agrégation dans les réseaux d'entreprises. Ils estiment les gains en ressources dus à l'agrégation des trafics et étudient le délai de bout-en-bout et le taux de perte des paquets pour évaluer la qualité de service fournie aux applications temps réel. De manière générale, ils recommandent l'agrégation des trafics en deux classes : le trafic en temps réel et le trafic non temps réel.

En ce qui concerne les réseaux mobiles, Aguado et *al.* [3] ont étudié l'impact de diverses agrégations de trafic de voix et paquets UDD 144 sur le débit total et le délai dans l'UTRA. Selon les proportions de trafic dans l'agrégation, on peut assister à une dégradation des performances du système. La mise en place d'une stratégie de réservation permet d'améliorer l'utilisation totale du système au prix d'une augmentation du délai. Menth [91], quant à lui, s'est intéressé aux performances d'un réseau UMTS sur IP (partie câblée) qui multiplexe de la voix et des données en commutation de circuits. Il propose des modèles pour les sources de voix et de données en commutation de circuits à 64 Kbps. Différents schémas de multiplexage et de tunnellisation (tunneling) des données temps réel dans un paquet IP ainsi que différents types de contrats sont étudiés. Les indices de performances sont le délai et le taux de perte des paquets. En général, le multiplexage est plus efficace que la tunnellisation.

Dans [5], les auteurs soulignent le fait que beaucoup de travaux se sont intéressés au contrôle d'admission et à l'allocation des ressources dans un réseau intégrant plusieurs types de trafic avec des exigences variées de QoS, alors que très peu d'articles se sont penchés sur l'intégration de plusieurs types de services dans un réseau WCDMA. Ils étudient donc dans leur article l'effet de l'application d'un ordre de priorité sur les différents services dans un réseau WCDMA UMTS. Des services CBR et interactifs NRT (non temps réel) sont considérés. Les auteurs proposent un mécanisme d'admission basé sur la priorité et les résultats montrent que les performances globales du système sont meilleures quand le trafic temps réel à faible débit a une priorité plus élevée.

Dans les différents modèles introduits jusqu'ici, on utilise souvent des modèles stochastiques pour représenter les paramètres de la charge de travail et leurs fluctuations. Cette approche permet de réduire de manière considérable la quantité d'information manipulée. On obtient donc un simulateur par distribution. Une *distribution* se définit comme la répartition sous forme de table numérique de la fréquence d'un évènement en fonction d'un caractère. Elle est associée à une *variable aléatoire* qui est en fait une fonction de l'espace échantillonnal dans l'ensemble des réels. Plus précisément, la distribution d'une variable aléatoire est définie par la donnée de la *fonction de densité de probabilité (fdp)* ou la *fonction cumulative de distribution (cdf) ou fonction de densité de répartition (fdr).* Des exemples de distribution sont la distribution lognormale, la distribution de Weibull, la distribution de Pareto, etc. Les caractéristiques de ces différentes distributions sont données dans les tableaux 2.2 à 2.4 où Φ est la fonction de densité d'une loi normale standardisée et G la fonction gamma.

Tableau 2.2 Loi de Pareto

Paramètres	β = paramètre de forme (shape), $\beta > 0$; a = paramètre d'échelle (scale), $a > 0$
Domaine	$a \leq x \leq \infty$
Densité	$f(x) = \beta a^{\beta} x^{-(\beta+1)}$
fdr	$F(x) = 1 - (a/x)^{\beta}$
Moyenne	$a\beta/(\beta-1)$, $\beta > 1$ (∞ sinon)
Variance	$a^2\beta/(\beta-1)^2(\beta-2)$, $\beta > 2$ (∞ sinon)

La distribution d'une variable aléatoire X est *sous-exponentielle* si elle vérifie la propriété suivante :

$$\forall \varepsilon > 0, \lim_{x \to \infty} e^{\varepsilon x} P[X > x] = \infty \qquad (2.1)$$

La distribution de X est dite à queue *lourde* (heavy-tailed) si elle vérifie la formule suivante :

$$P[X > x] \underset{x \to \infty}{\sim} L(x) x^{-\beta} \qquad (2.2)$$

où L est une fonction variant lentement à l'infini (i.e. $\lim_{t\to\infty} L(xt)/L(t) = 1$), $f(x) \sim g(x)$ signifie que $\lim_{x\to\infty} f(x)/g(x) = 1$ et $\beta \in]1,2]$.

Tableau 2.3 Loi lognormale

Paramètres	m = moyenne du log, m >0 ; σ = écart-type du log, σ>0
Domaine	$0 \le x \le \infty$
Densité	$f(x) = \sqrt{\frac{1}{2\pi\sigma^2 x^2}} \exp\{-((\ln(x) - m)^2 / 2\sigma^2)\}$
Moyenne	$\exp\left[m + \frac{\sigma^2}{2}\right]$
Variance	$\exp[2m + \sigma^2] * (\exp[\sigma^2] - 1)$

Tableau 2.4 Loi de Weibull

Paramètres	b = paramètre de forme (shape), b >0 ; a = paramètre d'échelle (scale), a>0
Domaine	$0 \le x \le \infty$
Densité	$f(x) = (bx^{b-1} / a^b) e^{-(x/a)^b}$
fdr	$F(x) = 1 - e^{-(x/a)^b}$
Moyenne	$(a/b) G(\frac{1}{b})$
Variance	$(a/b)^2 \left(2bG(\frac{2}{b}) - \left[G(\frac{1}{b}) \right]^2 \right)$

Cette dernière définition implique une espérance finie, mais une variance infinie. Un exemple de distribution à queue lourde est la loi de Pareto lorsque $\beta > 2$. Notons finalement qu'une distribution à queue lourde est sous-exponentielle, mais la réciproque est fausse.

Les distributions à queue lourde sont utilisées surtout pour modéliser certains types de fichiers précis ou les tailles des fichiers transférés sur un réseau local ou étendu,

comme c'est le cas par exemple sur l'Internet. Dans le cas de la navigation Internet, parmi les autres facteurs pouvant influencer les performances, on peut citer les paramètres du protocole TCP (slow-start ou démarrage à froid, les algorithmes de contrôle de flux utilisés, les autres algorithmes comme celui de Nagle) et la version de HTTP utilisée. En effet, la version 1.1 de HTTP introduit la notion de connexion persistante qui permet d'utiliser la même connexion pour envoyer plusieurs requêtes, alors que dans la version 1.0, une connexion est établie juste pour le transfert de chaque objet incorporé.

De manière plus spécifique, l'accès des terminaux mobiles au réseau Internet et les travaux sur le WAP (W*ireless Application Protocol*) ont permis de quantifier et de proposer des configurations de TCP et HTTP adaptées à l'environnement mobile. Ainsi, le forum WAP [127] a développé des spécifications pour les protocoles HTTP et TCP dans un environnement sans fil (voir la suite de spécifications WAP-225 et WAP 229). Nous nous inspirerons fortement de ces spécifications pour nos configurations du protocole TCP.

De manière générale, les études précédentes restent assez restreintes soit au niveau des type d'applications ou de trafics considérés [3,20,80], soit au niveau des performances recueillies [3]. En effet, les modèles de trafic sont soit simplifiés, soit ne couvrent pas une gamme représentative des applications qui pourraient être exécutées dans un réseau de troisième génération. De plus, les performances ne prennent souvent pas en compte les usagers (distribution, habitudes d'utilisation des services) et se limitent dans la plupart des cas à une analyse des performances du canal radio. La qualité de service, telle que perçue par les usagers, est rarement abordée. Souvent aussi, le modèle UMTS est simplifié ou limité à une partie de l'architecture (réseau d'accès) sans tenir compte du fait que le réseau UMTS combine à la fois une partie sans fil et une partie câblée. Enfin, l'impact de l'environnement de déploiement du réseau est aussi souvent négligé.

L'un des objectifs de cette thèse est d'étudier de manière plus globale et du point de vue des applications ou de l'utilisateur, les performances d'un réseau UMTS dans un

contexte où sont multiplexés un échantillon des services ou des trafics représentatifs, présentant souvent une dépendance à long terme et un caractère fractal et en rafales. Plus précisément, nous nous intéresserons aux réseaux UMTS basés sur des réseaux à commutation de paquets et implémentant un mécanisme de qualité de service et ce, dans divers environnements (intérieur, piétonnier, véhiculaire).

2.3 Tarification

Une fois l'étude des performances réalisée et le réseau implanté, l'opérateur doit choisir des tarifs pour son ou ses services afin de couvrir ses coûts. Dans le domaine des réseaux intégrés tels ATM ou dans le cas de l'Internet qui évolue vers un réseau intégré avec garantie de Qualité de Service, plusieurs approches sont proposées pour la tarification. Pour se rendre compte de la diversité des conceptions, un rapide survol des opérateurs montréalais de réseaux mobiles offrant, en plus du service traditionnel de voix, un service de navigation sur Internet, montre une variété de tarifications : tarification à la durée, au volume et à la transaction. De plus, cette diversité est exacerbée par le fait que les réseaux intégrés mettent en contact le monde de la commutation de circuits habitué à une tarification à l'usage et le monde de la commutation de paquets plus familier avec la tarification au taux forfaitaire (le « flat-rate » d'Internet).

De manière générale, les usagers peuvent être facturés selon plusieurs facteurs allant du type de service à l'utilisation, en passant par les ressources allouées (ou une mesure telle que le débit effectif), la durée de l'appel, le débit, le début de l'appel, la distance, le nombre d'appels, etc. Souvent, le prix est déterminé en fonction d'une combinaison de plusieurs de ces facteurs.

Les approches tarifaires peuvent être regroupées en deux catégories :

- Tarification statique : ce genre de tarification impose un prix forfaitaire indépendant de l'usage ou un prix fixe par unité de ressource. Dans les deux

cas, le prix ne dépend pas de l'état du réseau et ne varie pas dans le temps. Le prix fixe peut aussi être une moyenne résultant d'une tarification dynamique.

- Tarification dynamique : le prix est en général par unité de ressource et varie avec l'état du réseau.

Le problème de la tarification est devenu crucial avec la fin du financement public d'Internet. En effet, la transition vers un Internet commercial a fait ressortir le besoin de recouvrir les coûts des infrastructures à travers un coût d'accès et possiblement un coût basé sur l'usage. Une politique de tarification appropriée devrait jouer partiellement le rôle d'un mécanisme de contrôle de congestion en encourageant une utilisation efficace de la capacité et aussi inciter les utilisateurs du réseau à classer leurs besoins dans les classes de service appropriées. Dans les nouveaux réseaux tels les réseaux de troisième génération, une politique de tarification est aussi un outil de marketing et de concurrence. Dans un environnement de libéralisation économique, les prix ne doivent pas être trop élevés pour ne pas perdre des clients, ni trop bas pour assurer la rentabilité et le développement du réseau.

2.3.1 Tarification statique

Tarification forfaitaire

Actuellement utilisée surtout pour Internet [85,89], elle consiste en un prix forfaitaire pour se connecter au réseau. L'usager n'est facturé ni pour le nombre de bits transmis ni pour le temps d'utilisation. Cette tarification a l'avantage d'être très simple et facile à implémenter. Elle n'introduit pas dans le réseau une surcharge liée à la gestion de la tarification. En effet, l'usager paie un prix forfaitaire pour la connexion et l'utilise comme il veut. Le gestionnaire de réseau n'est pas obligé de maintenir un système de tarification qui génère un trafic supplémentaire et consomme une partie des ressources. Pour le système téléphonique par exemple, la surcharge liée à la gestion de la tarification est évaluée à plus de 50% de la facture de téléphone. Ce genre de tarification est aussi utilisé pour les appels locaux en Amérique du Nord.

Une variante de la tarification forfaitaire est la tarification au débit garanti où l'usager paie un tarif dont une partie dépend du débit de la connexion qui est le débit maximum faisable et l'autre partie du débit minimal garanti à l'usager. La tarification forfaitaire est possible quand le coût marginal d'envoi ou de réception d'un bit est quasi-nul [6]. Dans ce dernier article, Anania et *al.* ont étudié l'évolution de la tarification dans les réseaux pour recommander une tarification forfaitaire pour le réseau RNIS. La tarification à l'usage, au volume ou par transaction présente selon eux une incitation à la fraude ou à l'arbitrage. Quant à la tarification basée sur les coûts, elle ne garantit pas l'accès à un nombre suffisant d'usagers et ne génère pas un capital d'investissement suffisant.

Cependant, la tarification forfaitaire ne fournit pas un mécanisme économique de contrôle de congestion pour l'allocation des ressources. En effet, dans un réseau intégré qui fournit des niveaux de qualité de service différents aux paquets, si les tarifs ne sont pas fonction des ressources mobilisées par l'usager, tout le monde choisirait le meilleur service, ce qui entraînerait un gaspillage des ressources. De plus, en général une tarification forfaitaire génère moins de surplus social tant du point de vue des utilisateurs que de celui des opérateurs [49,87]. Finalement, les usagers ne générant pas beaucoup de trafic sont pénalisés.

Tarification au taux fixe

Dans ce cas, l'usager paie un taux fixe par unité de volume de données envoyées [27,114]. La facture totale est proportionnelle au nombre de bits envoyés. Les différentes classes de service ne sont pas prises en compte.

Tarification prioritaire

Cocchi et *al.* [27] proposent une tarification qui tient compte de la priorité et des classes de services. Dans ce schéma, les usagers paient un tarif plus élevé par octet pour des données requérant une plus grande priorité. Pour avoir une meilleure qualité de service, l'usager utilise une priorité plus élevée pour ses données, mais en retour paie pour la contrainte imposée au réseau et aux autres usagers. Les auteurs arrivent à la

conclusion qu'un mécanisme de tarification est nécessaire seulement quand le réseau est sous haute charge. Cependant, leurs hypothèses assument que le trafic généré par les usagers ne change pas et que seuls les niveaux de service requis le font.

Un modèle similaire est proposé dans [49]. Les prix pour les différentes classes sont dérivés d'un modèle Arrow-Debreu et sont attribués en fonction de la priorité, de la taille de la transaction, du taux d'arrivée pour cette classe de service et du coût dû au délai sur le service.

Tarification selon le débit convenu

Dans ce modèle, Clark [26] propose un compromis entre les services garanti et non garanti. Dans un contrat initial, l'usager convient d'un débit avec l'opérateur. Un contrôle de trafic (inspiré des mécanismes de contrôle de trafic de ATM) basé sur le débit convenu est mis en place et marque les paquets qui ne respectent pas le contrat initial. Ce modèle ressemble beaucoup au modèle de gestion de trafic sur ATM. Cependant, ici, le contrat initial qui fixe le débit convenu est un contrat à long terme. L'usager paie donc un taux forfaitaire pour un certain débit, dépendant des paramètres (taille du seau et taux de fuite du seau dans le mécanisme de « leaky bucket » par exemple) du contrat initial. Les valeurs du contrat ne sont pas des bornes supérieures et ne sont utilisées en fait que dans le cas de congestion. De plus, le modèle ne fournit aucune garantie.

Miah et Cuthbert [92] proposent une approche similaire où le coût d'une connexion est donné par :

$$C = \alpha D.t \tag{2.3}$$

où α est un facteur multiplicatif constant, t est la durée de la connexion et D correspond au débit du régulateur de trafic dédié à la connexion. En fait, l'usager demande un certain débit pour son régulateur de trafic. Le régulateur de trafic est implémenté de telle sorte que le flot à sa sortie soit borné par une enveloppe de Poisson (qui possède les propriétés classiques d'additivité). L'utilisateur est supposé adapter sa demande de débit à son type de trafic. Miah et Cuthbert [92] généralisent cette approche au cas de différentes catégories de trafic avec différents niveaux de QoS. Pour le trafic CBR/VBR,

le coût est similaire à la relation (2.1), alors que pour le trafic UBR, le coût est juste proportionnel au volume de trafic.

Tarification du métro de Paris (ou tarification basée sur la réservation)

Cette méthode est inspirée de l'ancien modèle de tarification dans le métro de Paris où un nombre identique de places est offert en première et deuxième classe mais avec des prix différents. La première classe se retrouve donc moins chargée. Dans le cas d'un réseau, on suppose que cela engendrerait une meilleure qualité de service pour le trafic de première classe. L'intérêt de ce modèle [100] réside dans sa simplicité d'implémentation surtout dans Internet.

Un mécanisme similaire est proposé par Parris et *al.* [103]. Les usagers choisissent une classe de service parmi deux possibilités et s'engagent pour une certaine durée d'utilisation. Un mécanisme d'acceptation de connexion simplifié est utilisé pour accepter les connexions qui doivent défrayer des frais de connexion. L'usager paye ensuite un taux fixe (plus élevé pour la classe de plus forte priorité) par unité de volume et par classe de service. Cependant, ce modèle pénalise les utilisateurs plus pauvres ou qui ont des conversations plus courtes à cause des frais de connexion.

Tarification selon les moments de la journée

Souvent implémentée par les compagnies de téléphone, elle consiste à offrir des réductions durant les périodes creuses et à augmenter le tarif pendant les périodes de pointe. Parris et *al.* [103] ont proposé une extension de leur modèle précédent pour inclure des tarifs selon les moments de la journée. En général, ces tarifs permettent de mieux étaler les demandes des utilisateurs et résultent donc en une utilisation plus uniforme des ressources.

Autres types de tarification

Da Silva [36] présente un modèle de tarification statique où il y a des frais de connexion pour chaque appel. Les connexions à débit constant sont facturées au débit alloué, celles à débit variable ou disponible au débit alloué et/ou utilisé, et celles à débit non spécifié ne sont pas facturées à l'usage.

2.3.2 Tarification dynamique

Tarification sensible

Proposé par MacKie-Mason et *al.* [83,84] dans le cadre de la tarification d'Internet, ce mécanisme introduit une rétroaction à laquelle les usagers réagissent, d'où son nom de « responsive pricing ». Les auteurs introduisent le principe de « marché intelligent » où, avant toute transmission, les usagers informent le réseau du prix qu'ils sont prêts à payer pour la transmission d'un paquet. Les paquets sont ensuite admis si les prix soumis dépassent un certain seuil déterminé par le coût marginal de congestion relié à un paquet additionnel. Les usagers ne paient pas le prix soumis, mais plutôt le prix seuil correspondant au paquet le moins cher admis. Ce schéma, même s'il n'est pas simple à implémenter, résout élégamment le problème des externalités de congestion dans un réseau à une seule classe de service.

Tarification selon le débit effectif

Basée sur la notion de capacité effective, cette approche a été introduite par plusieurs auteurs dont Kelly, Courcoubetis, Siris et Lindberger [28,29,67,78,121]. L'idée de base est d'appliquer un tarif comprenant un prix par unité de temps, un prix par unité de volume et des frais de connexion. L'usager déclare à la connexion un débit moyen qui détermine, selon le principe de la capacité effective, les différents tarifs qui lui seront appliqués. Si l'usager envoie plus ou moins que le débit déclaré, il paie un prix plus élevé. Même si dans [121] les auteurs montrent la pertinence de ce modèle du point de vue concurrentiel, on note quelques désavantages : l'usager doit déclarer son débit moyen, l'algorithme d'admission de connexion utilise les paramètres déclarés par l'usager et ce dernier se retrouve pénalisé à la fois par la perte de ses paquets et par un tarif plus élevé, s'il ne respecte pas le contrat.

Une variante est proposée par Lindberger [78], où chaque connexion encourt un tarif C défini par :

$$C=K_{L,T}.\ d.\ t$$

où $K_{L,T}$ est un facteur dépendant du moment de la journée et de la distance, d est une estimation a priori du débit effectif de la connexion et t est la durée de la connexion. La limitation principale ici est la détermination du débit effectif.

Vente aux enchères de la capacité

Introduit par Hayer [51], ce modèle (Transport auction) ressemble au modèle de tarification sensible de MacKie-Mason et *al.* [83,84]. Chaque usager a un agent qui accepte un certain trafic et le prix maximum que l'usager est prêt à payer. Le réseau fixe, en fonction de sa charge et de sa capacité, un prix seuil et accepte le trafic des agents offrant un prix supérieur à son prix-seuil. L'usager peut ajouter un délai limite au bout duquel le trafic est offert au réseau, quel que soit le coût. Ce mécanisme, contrairement à celui du « marché intelligent », n'amène pas les usagers à révéler la véritable valeur qu'ils attachent au trafic. De plus, le mécanisme est assez complexe et les prix doivent être mis à jour assez régulièrement pour justifier l'utilisation de l'agent.

Allocation dynamique du débit

Dans ce schéma [95], les usagers possèdent une fonction de demande de la capacité qui est utilisée pour prédire la demande de capacité pour un intervalle de temps donné. Les usagers ne tiennent compte que de leurs propres intérêts. Le réseau définit une fonction de coût dépendant de la capacité et du débit total utilisé. Le prix pour le prochain intervalle de temps est calculé en fonction du coût marginal pour envoyer le trafic prédit et du coût courant. L'un des éventuels problèmes avec ce modèle est l'existence d'un équilibre instable.

Autres types de tarification

Beaucoup d'autres types de tarification dynamiques ont été proposés. Dans [18,35], les auteurs présentent des revues des principales méthodes de tarification proposées dans la littérature.

2.3.3 Intégration des tarifications

Une fois la structure de la tarification définie, l'opérateur doit établir ses prix au niveau du réseau. Pour ce faire, plusieurs approches sont possibles. L'opérateur peut essayer de maximiser ses revenus [13] pour le réseau ou d'amortir les coûts d'infrastructure et d'opération du réseau sur une période donnée [42,45,58,93]. Dans tous les cas, l'opérateur doit déterminer ses revenus en se basant sur la structure de tarification définie, la fonction de demande des usagers qui reflète le taux d'arrivée des appels en fonction du tarif et la probabilité de blocage. Les coûts du réseau peuvent être plus ou moins complexes à déterminer. Par exemple dans [13], les auteurs décomposent le coût du service Internet pour différents types d'usager en cinq catégories :

- les coûts de l'équipement qui regroupent les coûts liés au matériel et au logiciel nécessaires pour le fonctionnement du réseau ;
- les coûts de transport qui comprennent les coûts des interconnexions et des lignes louées ;
- les coûts pour le service à la clientèle qui correspondent aux coûts pour les locaux et le personnel qui fournira le service à la clientèle ;
- les coûts d'opération dans lesquels on regroupe les coûts de facturation, d'entretien des équipements et des locaux et les coûts pour le personnel d'opération ;
- les autres coûts comme par exemple les coûts administratifs, les coûts généraux, les coûts de vente et de marketing ;

Ces catégories s'appliquent aussi aux réseaux mobiles. Par exemple, Gavish et Sridhar [45] prennent en compte les coûts de mise en place, de maintenance et d'opération des tours (stations de base) et le coût associé à chaque canal disponible dans la cellule. On peut aussi prendre en compte divers facteurs d'économie comme une économie d'échelle sur les canaux ou sur le volume de trafic (lignes louées) [42].

De manière générale, les différents travaux présentés n'adoptent pas une approche de tarification intégrée du niveau de l'usager au niveau du réseau. L'autre objectif de cette thèse est de proposer un modèle de tarification amélioré permettant de

déterminer les coûts liés aux différents niveaux de service et d'assurer une meilleure intégration au réseau par la prise en compte de contraintes supplémentaires telles une borne sur la gigue.

2.4 Conclusion

Pour atteindre nos deux objectifs (évaluation des performances et tarification), nous aborderons dans les prochains chapitres les performances des réseaux de type UMTS en nous concentrant sur des performances observables sur une petite échelle de temps comme le débit, le délai, le temps de réponse, etc. Ensuite, nous étudierons les problèmes de fiabilité qui eux en général impliquent une échelle de temps plus grande. Finalement, les résultats ainsi obtenus seront utilisés pour essayer de mettre en place un système de tarification efficace et équitable.

CHAPITRE 3

ASPECTS DE MODÉLISATION D'UN RÉSEAU UMTS ET DE SA CHARGE POUR L'ÉVALUATION DE LA QUALITÉ DE SERVICE

La méthodologie d'une évaluation de performance comprend en général six étapes [41] : l'identification des problèmes, la formulation des objectifs, la modélisation du système et de sa charge, l'élaboration du plan d'expérience, l'implémentation du plan et la collecte des résultats, l'interprétation des résultats. La formulation des objectifs nécessite la définition et la sélection des indices de performance d'intérêt. Une fois les indices sélectionnés, il faut définir un modèle du système et de sa charge, l'implémenter et recueillir les résultats par l'application d'un plan d'expérience. Dans ce chapitre, nous présenterons d'abord les indices de performance découlant de la formulation de nos objectifs. Ensuite, nous nous intéresserons à la modélisation du système et de sa charge qui constitue l'une des phases les plus importantes d'une évaluation de performance.

3.1 Indices de performance

Au sens le plus simple, la qualité de service (QoS) est l'indicateur ou l'ensemble des indicateurs permettant de fournir de façon constante un service de transmission de données conforme aux exigences de l'utilisateur du réseau [25]. En d'autres termes, la qualité de service est la capacité d'un élément de réseau (par exemple, une application, un routeur) d'avoir une certaine assurance que ses exigences de trafic et de service peuvent être satisfaites. Ceci exige la coopération de toutes les couches du réseau, de la plus basse à la plus haute, et de tous les éléments du réseau, d'une extrémité à une autre. La garantie globale de QoS est déterminée par le maillon le plus faible de la « chaîne » entre l'émetteur et le récepteur. Pour évaluer la QoS, nous retenons, par souci de

simplicité, cinq paramètres qui revêtent une importance plus ou moins grande selon les contextes: *le délai, la variation de délai, la capacité ou le débit, le taux de perte de paquets et la disponibilité.*

Le *délai* peut être défini comme le temps pris par un message pour aller d'un nœud à un autre. Dans les réseaux de commutation de paquets, la *variation de délai* est une variation du temps d'arrivée entre deux paquets consécutifs comparé au temps entre ces deux mêmes paquets lors de la transmission initiale. Au chapitre 5, nous aborderons plus en détail la définition de la variation de délai. Une telle variation est particulièrement préjudiciable pour le trafic multimédia [116]. La *capacité* ou le débit mesure le taux d'information qu'une liaison peut véhiculer. Le *taux de perte de paquets* est un rapport du nombre de paquets perdus pendant une transmission au nombre de paquets transmis. La *disponibilité* est une mesure du « temps » pendant lequel le réseau ou la connexion peut fonctionner de manière continue sans aucune panne. Elle peut être exprimée par le rapport MTBF/(MTBF + MTTR), où MTBF désigne la durée moyenne entre deux pannes et MTTR le temps moyen de réparation. Le taux de perte de paquets et la disponibilité permettent de caractériser la fiabilité d'une connexion.

3.2 Modélisation du système et de sa charge

La charge du système revient essentiellement au trafic généré par les utilisateurs. Le trafic de contrôle sera considéré comme une surcharge sur celui généré par les utilisateurs. Pour nos prévisions de trafic, nous adoptons les classes du forum UMTS. Le Tableau 3.1 présente une estimation du trafic offert par usager pour les liaisons montante (*UL*) et descendante (*DL*).

Tableau 3.1 Trafic offert par usager effectif par heure de pointe (kbits/h/user)

Services	Liaison descendante	Liaison montante
HIMM (High Interactive Multimedia)	36864	36864
HMM (High Multimedia)	213200	1066
MMM (Medium Multimedia)	10675.2	277.55
SD (Switched Data)	6552	6552
SM (Simple Messaging)	840	840
S (Speech)	1680	1680

Le Tableau 3.2 fournit une vue d'ensemble des niveaux de QoS définis pour les différentes classes de trafic [124].

Tableau 3.2 Qualité de Service

Type ou classe de trafic	Service	Délai	Variation de délai	Fiabilité
Interactif en temps réel	Conversation téléphonique	< 150 ms	< 1 ms	< 3% TET[1]
	Vidéophone	< 150 ms		< 1% TET
	Télémétrie (contrôle)	< 250 ms		≈ 0% TET
	Jeux	< 250 ms		< 3 % TET
Interactif	Messagerie vocale	< 1 sec	< 1 ms	< 3% TET
	Navigation Web	4 sec / page		
	commerce électronique	4 sec		
Flot continu	Flot audio continu	< 10 sec	< 1 ms	< 1% TET
	Vidéo	< 10 sec		< 1% TET
	Télémétrie (surveillance)	< 10 sec		≈ 0% TET

[1] Taux d'erreur par trame

Le système UMTS est globalement asymétrique. Le Tableau 3.1 illustre cet aspect. En effet, on note que les trafics offerts sur les liaisons descendante et montante ne sont pas toujours égaux.

3.2.1 Modèle général du système

Notre modélisation se fera au niveau session ou application avec 6 types de trafic principaux: la voix, le courrier électronique (ou courriel) et la radiomessagerie, le fax et l'accès commuté, Internet (en général), la vidéo sur demande, la téléconférence. Comme nous nous intéressons seulement au réseau d'accès, nous considérons, à toutes fins utiles, le sous-réseau de transport comme un ensemble d'applications (ou serveurs) génératrices de trafic. La Figure 3.1 présente le modèle général du système.

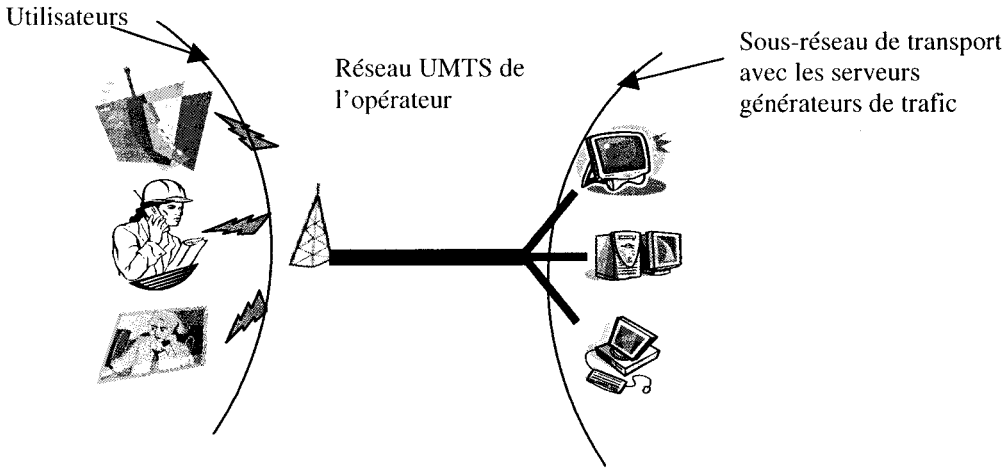


Figure 3.1 Modèle général du système

3.2.2 Modèle UMTS

Au niveau UMTS, l'architecture de qualité de service telle que proposée par le Forum 3GPP [2] inclut un ensemble de couches qui se relaient pour assurer un service de bout en bout. En effet, pour accéder au service, un terminal mobile doit tout d'abord s'attacher au réseau par une procédure appelée « *GPRS Attach* ». Cette procédure identifie le mobile et établit un lien logique le reliant au noeud de service SGSN. La procédure inverse est la procédure « *GPRS Detach* ». Pour ensuite envoyer et recevoir des données, le mobile doit activer le *contexte PDP* (Packet Data Protocol) qu'il veut utiliser lors de la procédure « *PDP Context Activation* ».

Un *contexte PDP* est un ensemble d'informations qui caractérise un service de transmission de base. Il regroupe des paramètres qui permettent à un abonné de communiquer avec une adresse PDP définie, selon un protocole spécifique (IP ou X.25) et suivant un profil de Qualité de service déterminé (débit, délai, priorité...). La procédure "PDP Context Activation", déclenchée à l'initiative de l'abonné mobile, permet au terminal d'être connu de la passerelle GGSN qui réalise l'interconnexion avec le réseau PDP externe demandé par l'abonné. La transmission de données entre le réseau UMTS et le réseau PDP externe peut alors débuter. La procédure inverse de « *PDP Context Activation* » est la procédure « *PDP Context Deactivation* ».

La qualité de service de bout en bout utilise un service support local équipement terminal (ou terminal mobile) et un service support UMTS. Les paramètres de QoS du service support UMTS ou *profil de QoS* sont les suivants :

- la classe de trafic : le genre d'application pour lequel le service support UMTS est optimisé ;
- le débit maximal ;
- le débit garanti : les indices de performance sont garantis pour un trafic de débit moins élevé que le débit garanti ;
- la livraison en séquence des paquets ;
- la taille maximale des unités de données de service (SDU : Service Data Unit) ;
- l'information sur le format des SDU : la liste des tailles exactes possibles des SDU ;
- le taux de SDU erronés ;
- le taux d'erreur binaire résiduel ;
- la livraison des SDU erronés ;
- le délai de transfert : le délai maximum pour le $95^{ème}$ centile de la distribution du délai de tous les SDU transmis pendant la durée du service support ;
- la priorité du trafic : l'importance relative des SDU de ce service support comparativement aux SDU d'autres services support ;
- la priorité d'allocation/rétention : l'importance relative comparativement aux autres services support UMTS de l'allocation/rétention du service support UMTS considéré.

La configuration générale de ces différents paramètres pour les différentes classes de service est spécifiée dans [2]. Ces différents paramètres sont présents dans le logiciel *Opnet®* pour permettre une configuration adéquate de la qualité de service. Cependant, il faut noter d'une part que le logiciel n'offre la possibilité que d'un seul profil par classe de trafic. Le modèle UMTS dans *Opnet* implémente aussi toutes les composantes d'une architecture de réseau UMTS typique telle que décrite au chapitre 2.

3.2.3 Modèle de trafic

Les modèles génériques pour les 6 types d'applications retenues sont les suivants.

Voix (S)

Le modèle de trafic de voix peut être décrit par un processus stochastique, avec des temps d'arrivée correspondant aux débuts de sessions. Chaque session décrit un appel téléphonique complet et contient des périodes ON-OFF. Les périodes ON correspondent aux impulsions vocales. Cette alternance entre périodes ON et OFF est obtenue par :

- la détection du bruit de fond au niveau de l'émetteur afin de transmettre seulement les paramètres importants au récepteur ;
- la détection du silence au niveau de l'émetteur ;
- une trame de silence (ou d'inactivité) qui est générée localement au niveau du récepteur en l'absence de trames normales de voix.

D'après les spécifications de UMTS [124], le codec de la voix est un codec GSM d'une taille de 20 ms à un débit de 12.2 kbps. La détection de silence est implémentée par la possibilité que le codec, par une signalisation adéquate, change son débit à chaque trame. Ceci permet d'avoir un facteur d'activité de l'ordre de 50%, i.e. la période d'activité couvre à peu près la moitié de la durée de la session. Le taux d'erreur par trame est de l'ordre de 3% soit un taux d'erreur par bit de l'ordre de 10^{-4}. L'excès de trafic dû au codage est de l'ordre de 75%, i.e. l'ensemble des mesures (redondance, checksum, entêtes, etc.) nécessaires pour garantir une certaine qualité de service génère un volume supplémentaire de données égal aux trois-quarts du volume initial. De plus, on suppose que le facteur d'efficacité dû à la retransmission des paquets erronés est de 0.75 (i.e. les paquets sont retransmis en moyenne une fois sur quatre). Le trafic de voix est supposé symétrique, i.e. la quantité d'information sur la liaison descendante est du même ordre que celle sur la liaison montante. La Figure 3.2 illustre des sessions d'appel.

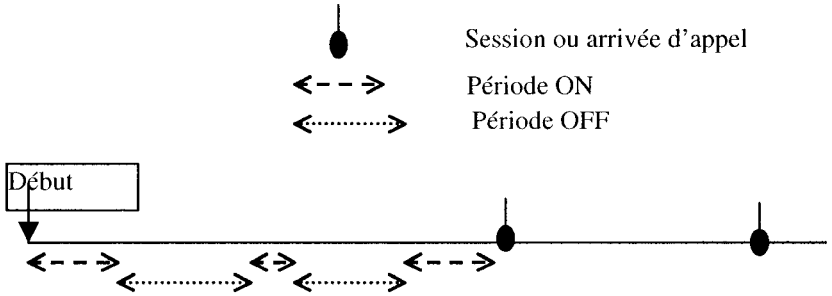


Figure 3.2 Modèle d'appel téléphonique

Dans [50], le modèle de voix proposé est une source ON-OFF où la période ON suit une distribution exponentielle de moyenne 352 ms. La période OFF suit aussi une loi exponentielle de moyenne 650 ms. Avec un facteur d'efficacité des paquets de 0.75, le débit réel de la voix est de 16 kbps (12.2*(1+(1-0.75))) au lieu des 12.2 kbps traditionnellement associés au GSM. De ce fait, la durée des périodes ON et OFF sera proportionnellement ajustée pour générer le volume de trafic spécifié dans le Tableau 3.1.

Courriel (SM)

Le modèle proposé est inspiré de [22,105]. Contrairement au modèle de trafic de voix, le modèle de courriel est décrit directement par l'arrivée de messages électroniques. La Figure 3.3 en est une illustration.

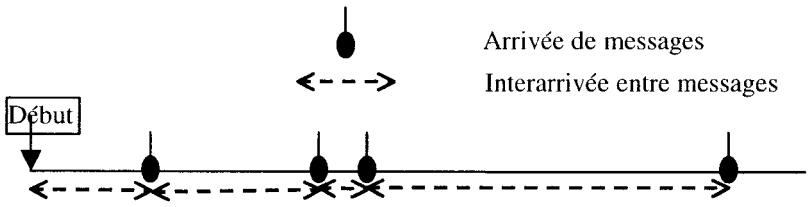


Figure 3.3 Modèle de courriel

D'après les spécifications de UMTS [124], une session de messagerie transfère à peu près 40 Ko sur une période de 30 sec. Le facteur d'efficacité des paquets est aussi de 0.75. L'excès de trafic dû au codage est de l'ordre de 2. On suppose aussi un trafic symétrique (sur une longue période d'observation).

Dans [107], les auteurs montrent que le trafic de courrier électronique a deux pointes, une le matin et l'autre l'après-midi. De plus, les interarrivées peuvent être considérées comme exponentielles sur une courte période.

La distribution de taille de courrier est (comme on aurait pu deviner) bimodale. Chaque mode correspond aux messages sans et avec pièce jointe. Paxson [105] propose deux lois lognormales. La première a une moyenne de 1324 octets et un écart-type de 2.75 octets. Cette distribution donne la taille des fichiers appartenant au premier mode, i.e. les fichiers sans pièce jointe. Ces fichiers ont une taille inférieure à 2048 octets et représentent les fichiers du quantile (inférieur) d'ordre 80% des données transmises. Le deuxième mode est décrit par une autre distribution lognormale de moyenne de 662 octets et un écart-type de 3 octets. Cette distribution correspond aux fichiers avec pièce jointe dont la taille est supérieure à 2048 octets (ce qui équivaut au quantile supérieur d'ordre 80% des données transmises). La moyenne donnée inclut une entête obligatoire de 300 octets, ce qui revient essentiellement à imposer une taille minimale aux messages.

Le modèle suggéré dans [105] a été proposé en 1994 et ne correspond plus très bien à la réalité actuelle ou future (par exemple, les spécifications UMTS proposent des fichiers de 40 Ko en moyenne [124], alors que la moyenne du modèle de [105] est d'environ 1.7 Ko). Cependant, Paxson [105] a constaté que le modèle lognormal pour la source de trafic de courriel est évolutif. Cela signifie que la distribution lognormale est une distribution robuste pour modéliser le trafic de courriel et fonctionnera même si les paramètres employés ne sont pas ceux présentés dans [105]. Nous gardons donc le même modèle que dans [105], mais rajustons les paramètres des distributions de chaque mode.

Pour cela, nous avons analysé, avec les outils statistiques de *Matlab*, 600 courriels reçus par 3 personnes sur des périodes d'un mois pour chaque personne. Comme suggéré dans [105], le premier mode est décrit par une loi lognormale de moyenne 3923 octets et d'écart-type 3087 octets, tandis que le deuxième mode a une moyenne de 152373 octets et un écart-type de 711224 octets. La limite entre les deux

modes est choisie de telle sorte que le nombre de courriels sans pièces jointes soit 3 fois plus important que celui de courriels avec pièce jointe pour avoir une moyenne globale de 40 Ko (ce qui correspond à peu près à la taille moyenne spécifiée par le groupe UMTS). Soulignons cependant que, contrairement au modèle précédent, nous n'imposons pas de taille minimale au courrier électronique, ni ne rajoutons pas d'entête.

Internet (MMM)

Le modèle d'Internet est inspiré de [8,10,11,81]. Comme le trafic de voix, il peut être décrit par des sessions contenant des périodes ON-OFF. Chaque session décrit une série de clics de l'utilisateur sur le même serveur par exemple. Chaque clic produira un certain nombre de demande d'objets incorporés. Les demandes d'objets incorporés sont séparées par des périodes OFF actives. La période inactive entre la fin des demandes d'objets incorporés et le clic suivant est une période OFF inactive [11,81]. La Figure 3.4 illustre une session Internet.

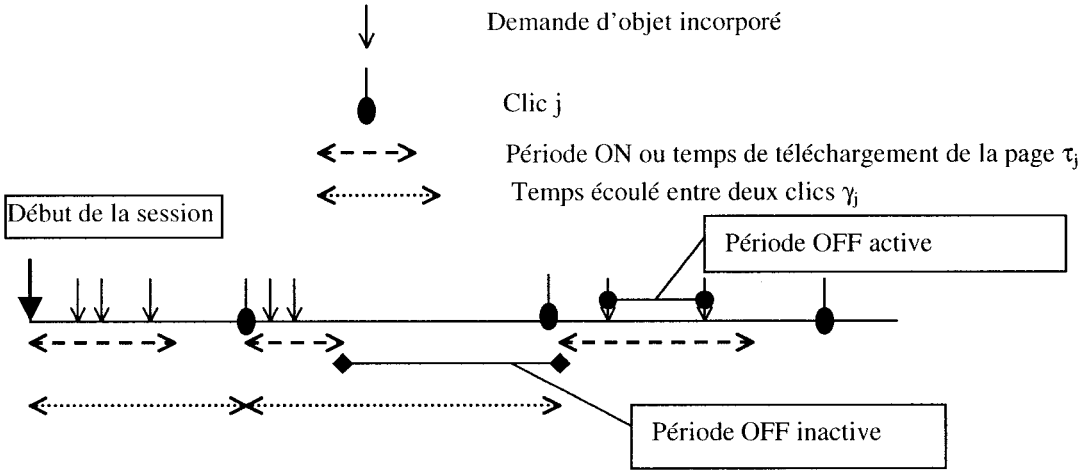


Figure 3.4 Session Internet

D'après les spécifications de UMTS [124], une session Internet transfère à peu près 0.5 Mo sur une période de 14 sec. Cette période correspond à la vraie période active. Ainsi, la durée de la session sera supérieure à 14 sec car elle inclut des périodes d'inactivité. En fait, la durée de la session sera ajustée pour obtenir le trafic spécifié par

[124]. Le facteur d'efficacité des paquets est aussi de 0.75. L'excès de trafic dû au codage est de l'ordre de 2. On suppose que le trafic Internet est asymétrique avec un facteur d'asymétrie de l'ordre de 0.026.

D'après [11,116], le trafic Internet présente un caractère fractal (ou autosimilaire). Dans [17,33], les causes de cette autosimilarité ont été analysées et reliées essentiellement à la nature des distributions (à queue lourde) des fichiers web et des temps d'inactivité. De plus, dans [17], il est proposé une étude systématique des différents modèles ON-OFF qui présentent un caractère autosimilaire. Finalement, dans [8,10,11,81], différentes traces ont été étudiées et montrent que le caractère autosimilaire ne dépend ni du protocole ni de l'évolution du trafic dans le temps et constitue de ce fait, une caractéristique fondamentale du trafic Internet.

Les divers paramètres du modèle d'Internet se présentent comme suit [10,11,81]:

- La taille des fichiers peut être modélisée avec deux distributions : une distribution lognormale dont la normale associée a une moyenne de 7.64 et un écart-type de 1.705 pour la partie principale de la distribution des tailles de fichiers, et une distribution de Pareto avec comme paramètre de position (location) 3328 octets et paramètre de forme (shape) 1.383 pour la queue de la distribution (88 % des données sont assignées à la distribution lognormale, soit les fichiers de taille inférieure à 15.4 Ko). Le fait que la queue de la distribution de la taille des fichiers suive une loi de Pareto explique en grande partie le caractère autosimilaire du trafic Internet.
- Les périodes OFF inactives suivent une distribution de Pareto avec comme paramètre de position 1 et comme paramètre de forme 1.5. Cependant, Niclausse [81] a constaté qu'une loi lognormale ou une loi inverse gaussienne pouvaient aussi être appropriées.
- Les périodes OFF actives d'inter-arrivées suivent une distribution de Weibull avec comme paramètre de position 1.46 et paramètre d'échelle (ou de dispersion) 0.382.

- Le nombre de références incorporées suit une loi de Pareto avec comme paramètre de position 1 et paramètre de forme 2.43.
- Le nombre de clics par session peut être modélisé par une distribution de Pareto avec un paramètre de position d'environ 0.8 et un paramètre de forme d'environ 1.16 [81]. Il a été aussi constaté qu'une loi inverse gaussienne peut parfois être employée. Cependant, nous avons préféré la loi de Pareto.

Vidéo sur demande (VoD) (HMM)

Le modèle de la vidéo sur demande (VoD) est plus complexe. D'après les spécifications de UMTS [124], une session VoD génère à peu près 10 Mo sur une période de 53 sec. Cette période correspond à la vraie période active. Le facteur d'efficacité des paquets est aussi de 0.75. Le taux d'erreur par trame est de l'ordre de 1%. L'excès de trafic dû au codage est de 2. On suppose que le trafic est asymétrique avec un facteur d'asymétrie de l'ordre de 0.005.

Une revue générale des modèles de vidéos à débit variable est présentée dans [61]. En général, les données vidéo sont fortement corrélées. Les corrélations découlent des ressemblances entre les images (corrélation inter-trame) ou entre des parties d'images (corrélation intra-trame). En raison des débits élevés exigés par les flots vidéo non comprimés, des algorithmes de compression sont employés. Le plus utilisé est l'algorithme MPEG. Dans cet algorithme, une séquence vidéo consiste en une série de trames, chacune contenant un tableau bidimensionnel de pixels. Pour chaque pixel, la luminance et la chrominance sont stockées. L'algorithme de compression sert à réduire le débit de données avant la transmission de la vidéo. Dans une séquence vidéo MPEG-1, 2 ou 4 [88], on distingue trois types de trames, chacune utilisant un modèle de codage légèrement différent:

- les trames I emploient seulement le codage intra-trame, basé sur la transformation discrète de cosinus (DCT) et sur le codage d'entropie ;
- les trames P emploient un algorithme de codage semblable aux trames I, mais avec une compensation de mouvement par rapport à la trame I ou P précédente ;

- les trames B sont semblables aux trames P, sauf que la compensation de mouvement peut être effectuée par rapport à la trame I ou P ou une interpolation des deux.

Typiquement, les trames I ont une plus grande taille que les trames P. Les trames B exigent en général le débit le plus faible. Les propriétés statistiques varient en général selon le type de trame puisque les approches de codage y sont différentes. Après le codage, les trames sont arrangées dans un ordre périodique déterminé, par exemple "IBBPBB" ou "IBBPBBPBBPBB", qui est appelé *groupe d'images* (GOP). Les modèles de vidéo proposés essayent généralement de reproduire les caractéristiques vidéo au niveau des scènes ou des trames.

La norme MPEG-4 conçue pour le web et la mobilité permet d'obtenir des séquences de qualité supérieure à des débits équivalents à ceux de MPEG-1 ou 2 ou bien de compresser les séquences à qualité inchangée avec des débits beaucoup plus faibles tout en apportant une grande interactivité. Des mécanismes de correction et de protection d'erreurs sont aussi intégrés pour permettre une adaptation au manque de fiabilité des environnements mobiles. La norme MPEG-4 peut être vue comme une généralisation des normes MPEG-1 et 2. En effet, dans ces deux dernières normes, les séquences sont décomposées hiérarchiquement en groupes d'images, images, tranches, macro-blocs et blocs, où chaque bloc est un rectangle qui représente l'objet de base. La norme MPEG-4 généralise la notion d'objet à la vidéo et à l'audio en définissant des « objets média » qui peuvent être d'origine naturelle ou synthétique. Ces objets peuvent être ensuite combinés pour former une scène ou une trame. Par rapport aux standards MPEG-1 et 2, qui sont basés sur les trames, MPEG-4 spécifie des outils qui permettent d'encoder les objets, de les combiner et de les sauvegarder. La reconstruction de la scène dans MPEG-4 se fait au niveau du terminal usager, ce qui permet d'obtenir des débits assez faibles. Une autre raison de l'efficacité de MPEG-4 est l'utilisation des objets évolutifs, qui sont encodés une seule fois mais dont le décodage s'adapte au débit du réseau. En outre, la norme fournit un ensemble d'outils pour encoder efficacement des grilles 2 ou 3-D, animer des personnages et des visages, ce qui fournit une compression très élevée.

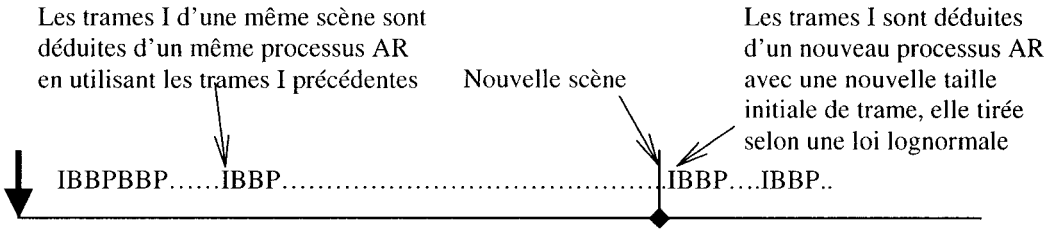


Figure 3.5 Session Video

Notre modèle s'inspire de [69,70]. C'est une modélisation au niveau trame. Une distribution lognormale (dont les paramètres sont déterminés avec des estimateurs de vraisemblance maximale) est utilisée pour caractériser les tailles de chaque type de trame. Les tailles des trames P et B sont supposées indépendantes et identiquement distribuées (i.i.d.) et suivant une loi lognormale. Quant aux trames I, leur taille moyenne d'une scène à une autre suit aussi une loi lognormale, mais à l'intérieur d'une scène donnée, les trames I sont modélisées à l'aide d'un processus autorégressif d'ordre 2 (AR2). Par conséquent, à l'intérieur d'une scène donnée, la taille des trames I fluctue autour de la taille moyenne déterminée pour cette scène à partir de la distribution lognormale. Le choix des types de trame est fait selon le groupe d'images choisi.

Dans [71,116], la taille des trames P et B (observées pour un certain nombre de films) peut être modélisée par des distributions lognormales dont les distributions normales associées ont respectivement des moyennes de 4 et 3, et des écarts-types de 0.6 et 0.47 approximativement. Quant aux moyennes des trames I, la distribution normale associée à leur distribution lognormale a une moyenne de 5.19 et un écart-type de 0.4. Les unités des valeurs spécifiées ci-dessus sont des cellules. Un facteur d'unité et un facteur d'échelle seront par la suite appliqués aux données pour les rendre conformes aux spécifications de [124]. Sur la liaison montante, les requêtes sont tout simplement modélisées par des trames de taille fixe proportionnelle à la moyenne des trames sur la liaison descendante.

L'inter-arrivée des trames est constante et égale à 41.7 ms (24 trames par seconde). Le nombre de trames par session est proportionnel à la longueur de session.

Une session VoD peut être une session de film ou une session de clip vidéo. Le groupe d'images utilisé sera "IBBPBBPBBPBB".

Les modèles utilisés dans cette thèse ont été dérivés de séquences vidéo MPEG-1. Cependant, nous estimons que le modèle sera aussi valable pour la norme MPEG-4 avec un ajustement d'échelle si nécessaire (pour s'adapter au volume de trafic spécifié), car il tient bien compte des propriétés d'auto-similarité du trafic vidéo et la structure des GOP ne change pas entre les normes MPEG-1, 2 et 4. L'idée est que la corrélation inter-trame dépend principalement de la séquence du groupe d'images. Par contre, la corrélation intra-trame n'est pas forcément bien modélisée par cette adaptation d'échelle, mais heureusement le niveau de détail de notre modèle s'arrête aux trames.

Téléconférence (HIMM)

D'après les spécifications de UMTS, une session de visioconférence dure en moyenne 180 sec avec un débit de 128 kbps et un facteur d'occupation de 0.8. Le taux d'erreur par trame est de l'ordre de 1%. L'excès de trafic dû au codage est de 2. On suppose que le trafic est symétrique.

Pour la liaison descendante, le modèle de visioconférence est directement déduit du modèle de la vidéo sur demande avec quelques petits ajustements. Canonico [23] a étudié des sources vidéo multiples et les a regroupées en trois classes, selon le rapport (σ/μ) des trames I et B-P, où σ est l'écart-type de la distribution de la taille des trames et μ la moyenne. Pour des films et vidéo-clips, nous avons $(\sigma/\mu)_I \in]0.2, 0.4]$ et $(\sigma/\mu)_{PB} \in]0, 1]$, tandis que pour les visioconférences $(\sigma/\mu)_I \in]0, 0.1]$ et $(\sigma/\mu)_{PB} \in]0, 0.6]$.

Le modèle VoD présenté dans la section précédente est caractérisé par des rapports $(\sigma/\mu)_P=0.66$, $(\sigma/\mu)_I=0.42$ et $(\sigma/\mu)_B=0.48$. Le modèle de visioconférence est donc déduit en ajustant d'abord la taille moyenne des différents types de trames conformément au rapport de proportion entre les débits moyens de la VoD et de la visioconférence. Ensuite, en utilisant les rapports (σ/μ) typiques aux différentes trames de visioconférence, on calcule l'écart-type des trames. La séquence du GOP sera "IBBPBBPBBPBB". L'inter-arrivée des trames est constante et égale à 41.7 ms (24

trames par seconde) et on suppose que, dans un modèle de téléconférence, il n'y a pas de changement de scène [23,53].

La durée de la session de vidéoconférence sera aussi choisie conformément à [124].

Fax (SD)

Le modèle proposé pour le fax est simple. La durée moyenne de la session est 156 s à un débit de 14.4 Kbps [124]. Cela donne une moyenne de 281000 octets transmis. À chaque session de fax, un message dont la taille sera tirée d'une distribution exponentielle de moyenne 281000 octets est transmis. L'excès de trafic dû au codage est de 3. On suppose que le trafic est symétrique. Le processus d'inter-arrivée des sessions de fax sera modélisé selon une loi exponentielle.

3.3 Implémentation du modèle

L'implémentation du modèle a été réalisée avec *Opnet* ®. C'est un environnement de développement de technologie de réseau utilisé dans l'industrie et qui permet de concevoir et d'étudier les réseaux, les dispositifs, les protocoles et les applications de transmission. Le logiciel utilise une approche orientée objet pour refléter la structure des réseaux réels ainsi que leurs composants. Le noyau de Opnet est basé sur des machines à états finis (MEF) utilisées pour créer des processus, qui à leur tour, peuvent être combinés pour générer des nœuds. La simulation dans Opnet se fait par événements, ce qui garantit une grande fiabilité, mais au prix d'un effort de calcul (et par conséquent d'un temps de simulation) important. De plus, Opnet fournit le code source, ce qui permet à l'utilisateur d'implémenter ses propres extensions. La version utilisée est 8.0 C. Les aspects de Opnet qui nous intéressent plus spécifiquement sont la configuration des applications et profils dans le module général et le module UMTS. Les autres aspects, notamment la modélisation de liens radio sont abordés dans l'Annexe I.

Le module UMTS implémente les différents paramètres de QoS définis précédemment ainsi que les composantes d'un réseau UMTS typique, tels que décrits au

chapitre 2. En particulier, le multiplexage et la gestion des priorités des différentes classes de trafic se font au niveau MAC.

Les applications dans Opnet sont organisées de manière hiérarchique. En effet, chaque serveur ou station de travail (ou tout autre équivalent) peut supporter un certain nombre de profils. Les profils correspondent à un ensemble particulier d'applications modélisant les habitudes d'un groupe d'usagers. On peut ainsi avoir un profil pour le département de gestion et un autre pour le département d'ingénierie. Les profils sont munis d'un motif de répétition indiquant le nombre de répétitions ainsi que le temps entre deux répétitions. À l'intérieur de chaque profil, on indique l'ensemble d'applications, le schéma de répétition ainsi que le mode d'exécution (sériel ou parallèle). Ainsi, une exécution du profil génère a priori plusieurs exécutions des applications, selon le schéma de répétition spécifié au niveau des applications, et une simulation génère a priori plusieurs exécutions du profil selon le schéma de répétition propre au profil. En outre, chaque application peut être divisée en un certain nombre de tâches s'exécutant séquentiellement ou en parallèle. Finalement, les tâches correspondent à un certain nombre de phases de transfert ou de traitement de données. Ces phases peuvent aussi s'exécuter séquentiellement ou concurremment.

Le logiciel fournit un certain nombre de « méta-applications standards » telles FTP, courriel, vidéoconférence, HTTP, voix, que l'utilisateur peut adapter à sa convenance. Si les applications fournies ne lui conviennent pas, l'utilisateur dispose d'une application sur mesure (custom application) qu'il peut configurer selon la hiérarchie phases/tâches/application/profil. L'utilisateur peut aussi implémenter ses propres modèles à partir d'un code en *C*.

Pour pouvoir simuler le modèle, certaines hypothèses générales ont été faites :

1. Nous ne considérons pas les problèmes de routage, ni ne prenons en compte la mobilité des usagers. Ceci a comme conséquence de réduire le trafic de signalisation sans affecter autrement le trafic de données. Nous nous limitons à l'étude d'une cellule car nous estimons que le goulot d'étranglement sera la liaison radio.

2. Les environnements d'intérêt sont les environnements intérieur, piétonnier et véhiculaire. Le type de cellule, le nombre de secteurs et la superficie des cellules dans ces différents environnements sont donnés au Tableau 3.3 pour l'année 2005.

Tableau 3.3 Caractéristiques des cellules pour l'an 2005

	secteurs	type de cellule	rayon R (km)	Aire des cellules sectorisées (km^2)
Intérieur	3	micro/**pico**	0.075	0.005
Urbain-piétonnier	3	macro/**micro**	0.7	0.424
Urbain-véhiculaire	3	macro/micro	0.7	0.424

3. La pénétration des divers types de trafic varie selon l'année envisagée et selon le type de service [124]. De ce fait, le modèle doit tenir compte des proportions d'usagers telles qu'indiquées au Tableau 3.4, ce qui est implémenté au niveau du modèle par une configuration adéquate des profils ou du nombre d'usagers par réseau.

Tableau 3.4 Nombre d'usagers par environnement par km^2 et par service pour l'an 2005

Services	Intérieur	Urbain piétionnier	Urbain véhiculaire
HIMM	216	32.4	0.0556
HMM	1080	304.56	0.52264
MMM	1728	498.96	0.85624
SD	1080	324	0.556
SM	2700	810	1.39
S	108000	38880	1000.8

4. Nous considérons que les surcharges de signalisation et les messages de l'interface radio (trafic de contrôle) pour l'ensemble des types de trafic sont de 20 % [124].
5. La voix, la vidéo, la téléconférence et le fax utilisent le protocole de transport UDP; le trafic HTTP, le courriel utiliseront le protocole TCP [24].
6. Nous utiliserons des canaux dédiés (DCH) sur les liaisons montantes et descendantes.

7. Dans le modèle UMTS du simulateur, on a un seul profil de QoS par classe d'application. De ce fait, pour une classe qui intègre des applications de types différents, nous choisirons le plus grand débit parmi les applications intégrées pour les liaisons montantes et descendantes. Dans le cas de la classe interactive par exemple, on regroupe des applications comme le fax (trafic symétrique de 14 Kbps) et la navigation sur Internet (trafic asymétrique de 10 Kbps sur la liaison montante et 384 kbps sur la liaison descendante). Le profil associé à la classe interactive devrait donc avoir 14 Kbps sur la liaison montante et 384 Kbps sur la liaison descendante. Les profils de QoS par classe ainsi que les services RLC associés à chaque classe sont résumés au Tableau 3.5.

Tableau 3.5 Profil de QoS et modes RLC

Classe	Débit liaison montante/ service RLC	Débit liaison descendante/ service RLC
Conversationnelle	128 Kbps/ Non acquitté	128 Kbps/ Non acquitté
En flux	10 Kbps/ Non acquitté	2000 Kbps/ Transparent
Interactive	14 Kbps/ Acquitté	384 Kbps/ Acquitté
En arrière plan	14 Kbps/ Acquitté	14 Kbps/ Acquitté

8. Nous ne modélisons pas spécifiquement le facteur d'efficacité des paquets, mais le considérons déjà inclus au volume de données. Par contre, le facteur de codage sera explicitement modélisé.
9. Compte tenu de la limitation du simulateur par rapport au nombre d'unités mobiles qui peuvent être modélisés par Nœud B (Node B) et compte tenu du nombre élevé d'usagers potentiels du service de voix, nous avons regroupé une vingtaine d'applications potentielles de voix par unité mobile.
10. Au niveau du sous-réseau ATM, la voix et la téléconférence auront un service CBR; la vidéo sur demande un service de nrt-VBR; l'Internet et le fax ainsi que le courriel, un service UBR [4,124].

Le modèle, tel qu'il se présente dans *Opnet,* est reproduit aux figures 3.6 et 3.7.

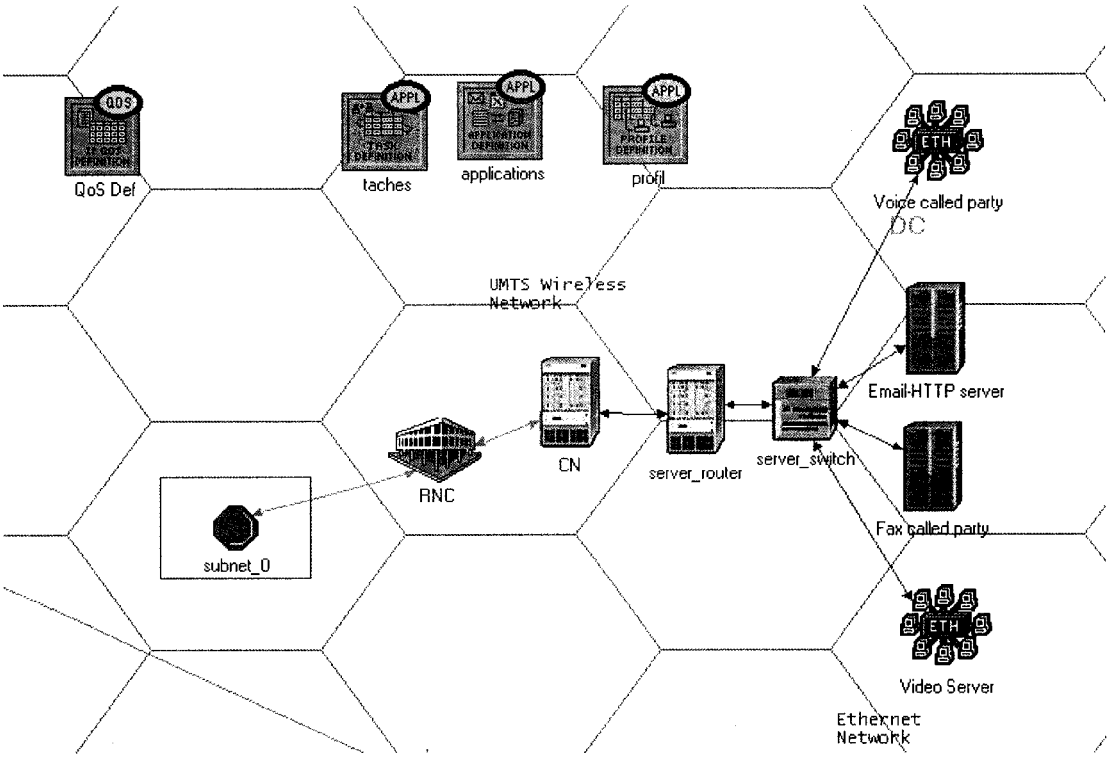


Figure 3.6 Topologie du réseau UMTS

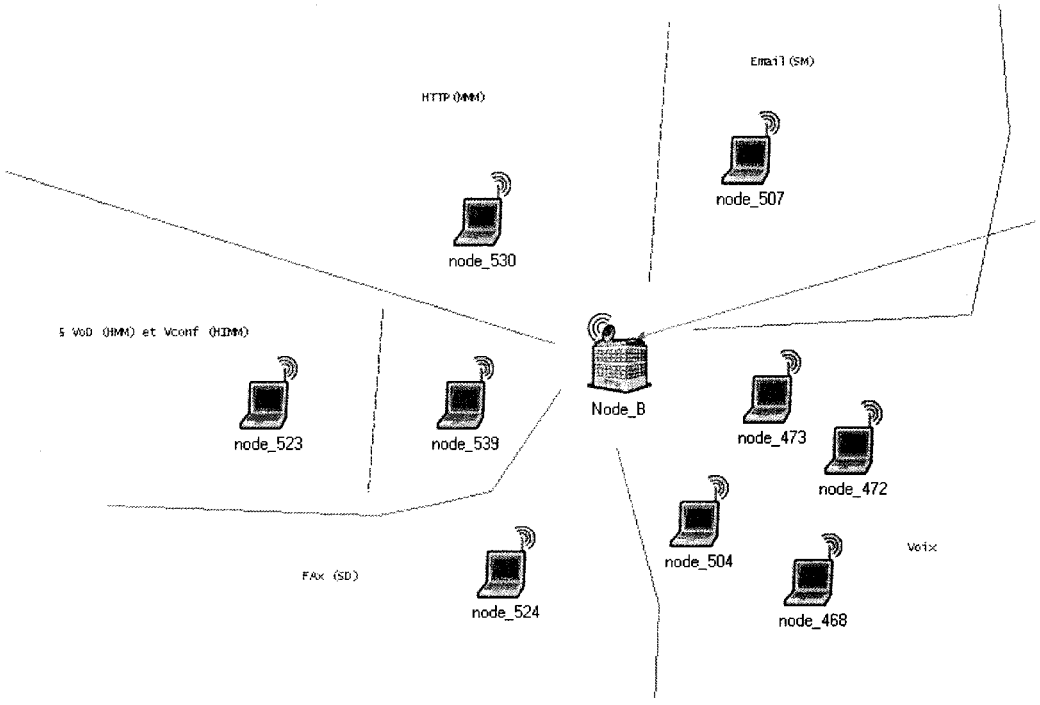


Figure 3.7 Détails d'une cellule

Finalement les paramètres de configuration complémentaires sont présentés au Tableau 3.6.

Tableau 3.6 Paramètres de configuration complémentaires

Modèle de propagation: 3GPP TR 25.942 inspiré du modèle COST 231
Antennes: omnidirectionnelles
Contrôle de puissance: Boucle externe +1dB par paquet erroné; –0.01dB par paquet correct
Transmission: Basée sur les TTI (*Transmission Time Interval)* avec la pile de protocole UMTS
Puissance de l'UE: 0.2W
Gestion de la priorité au niveau MAC et basée sur les classes de trafic
Facteur d'étalement : de 4 à 512
TTI de 625 µs
Mode FDD
Pas de gestion de QoS au niveau IP dans le modèle de base
Réseau ATM surdimensionné
Utilisation des canaux de transport dédiés (DCH)

3.4 Résultats et analyse

Dans cette partie, nous présentons les résultats de l'évaluation de performance dont la méthodologie a été décrite précédemment. L'approche adoptée consiste dans un premier temps à établir un modèle de référence qui sera calibré et validé d'abord par une évaluation de son bien-fondé, ensuite en remplaçant un certain nombre de trafics simulés par des traces réelles. L'évaluation du bien-fondé du modèle de référence permettra d'établir de manière plus ou moins absolue ses performances (par exemple, par rapport à un niveau de service donné ou aux requis moyens des utilisateurs). Dans un second temps, nous étudierons l'impact de divers paramètres au niveau du réseau et des usagers sur l'amélioration relative des performances par rapport au modèle de référence.

Les résultats sont présentés pour l'environnement intérieur. Dans cet environnement, les unités mobiles sont dans un rayon de 75 m autour du Nœud B qui a une puissance de 5W. Pour une telle taille de cellule, la densité des usagers fournies dans le Tableau 3.4 prévoit 1 usager potentiel pour la vidéoconférence (HIMM), 5 pour la VoD (HMM), 5 pour le fax (SD), 13 pour le courriel (SM), 8 pour HTTP (MMM) et 500 pour la voix (S).

3.4.1 Calibrage des applications

Pour adapter les modèles génériques définis à la section 3.2.3 aux spécifications du Tableau 3.1, nous avons effectué des simulations sur des durées de 1 à 3 jours, ce qui correspond à peu près à 24 à 72 répétitions de l'heure de pointe. Les paramètres de calibrage des applications sont les suivants.

Voix

L'application qui génère la voix a été configurée avec les paramètres du Tableau 3.7.

Tableau 3.7 Paramètres de configuration de l'application de voix

Application utilisée	Voix standard
Durée des périodes ON	Exponentiel (moyenne=0.4215 sec)
Durée des périodes OFF	Exponentiel (moyenne=0.5785 sec)
Type de codec	GSM avec détection de silence
Nombre de trames de voix par paquet	1

Le profil utilise le schéma de répétition suivant spécifié dans le Tableau 3.8 :

Tableau 3.8 Paramètres de configuration du profil de voix

Application	Voix
Durée de l'application	Exponentiel (moyenne=120 sec)
Nombre de répétitions de l'application	Illimité
Temps inter-application	Exponentiel (moyenne=3600 sec)
Durée du profil	Toute la simulation
Nombre de répétitions du profil	Illimité
Temps inter-profil	0

L'application est symétrique et associée à une classe conversationnelle.

Courriel

L'application qui génère le courriel a été configurée avec les paramètres suivants du Tableau 3.9. La distribution spéciale « Email-dist » est une distribution que nous avons ajoutée au logiciel. Elle consiste à choisir la taille des messages à partir d'une distribution selon une distribution lognormale de moyenne 3923 et d'écart-type 3087 dans 75% des cas et selon une autre lognormale de moyenne 152373 et d'écart-type

711224 dans le reste des cas, et ceci conformément aux spécifications introduites à la section 3.2.3.

Tableau 3.9 Paramètres de configuration de l'application courriel

Application utilisée	Courriel standard
Temps interarrivée à l'émission	Exponentiel (moyenne=300 sec)
Nombre de messages groupés à l'émission	1
Temps interarrivée à la réception	Exponentiel (moyenne=300 sec)
Nombre de messages groupés à la réception	1
Taille des messages	Email-dist

Le profil utilise le schéma de répétition du Tableau 3.10 :

Tableau 3.10 Paramètres de configuration du profil de courriel

Application	Email
Durée de l'application	Exponentiel (moyenne=42 sec)
Nombre de répétitions de l'application	Une fois
Temps inter-application	N/A
Durée du profil	Exponentiel (moyenne=3600 sec)
Nombre de répétitions du profil	Illimité
Temps inter-profil	0

L'application est symétrique et associée à une classe en arrière plan.

Internet

Pour le trafic HTTP, les paramètres sont spécifiés dans le Tableau 3.11.

Tableau 3.11 Paramètres de configuration de l'application HTTP

Application utilisée	HTTP modifiée
Specification HTTP	HTTP 1.1
Temps inactif OFF	Pareto (position=1; forme=1.5)

Dans le modèle d'application HTTP standard proposée par OPNET, on doit spécifier le temps d'interarrivée des pages (γ_j) sur la Figure 3.4. Pour être conforme au modèle adopté, nous avons modifié l'application HTTP standard de OPNET pour pouvoir plutôt spécifier le temps inactif OFF. Par rapport au modèle introduit au point 3.2.3, les temps actifs OFF n'ont pas été modélisés. Toutes les requêtes incorporées d'une page sont envoyées en parallèle au client sur la même connexion. Chaque page est composée d'un certain nombre d'objets spécifiés par le Tableau 3.12.

Tableau 3.12 Paramètres de configuration des composantes d'une page HTTP

Taille de l'objet	Nombre par page
500 octets	1
Internet file size	Pareto (position=1; forme=2.43)

La distribution spéciale « Internet file size » est une distribution que nous avons ajoutée au logiciel. Elle consiste à choisir la taille des messages selon une distribution lognormale de moyenne 8897 et d'écart-type 37009 pour les fichiers dont la taille est inférieure à 15.4 Ko et selon une distribution de Pareto avec comme paramètre de position (location) 3328 octets et paramètre de forme (shape) 1.383 pour les autres fichiers, et ceci conformément aux spécifications introduites à la section 3.2.3. En fait, les spécifications de [10] ont été légèrement modifiées pour normaliser la combinaison des deux distributions. Le profil utilise le schéma de répétition du Tableau 3.13 :

Tableau 3.13 Paramètres de configuration du profil HTTP

Application	HTTP
Durée de l'application	Exponentiel (moyenne=360 sec)
Nombre de répétitions de l'application	Une fois
Temps inter-application	N/A
Durée du profil	3600 sec
Nombre de répétitions du profil	Illimité
Temps inter-profil	0

L'application est asymétrique et associée à une classe interactive.

VoD

Le trafic VoD est modélisé en modifiant la méta-application « videoconferencing » pour permettre l'envoi de trames selon un GoP. Les propriétés des différentes trames sont données dans le Tableau 3.14 :

Tableau 3.14 Paramètres de configuration de l'application VoD

Type de trame	Taille
Trame I	Lognormale (moyenne=42075, écart-type=17755)
Trame P	Lognormale (moyenne=14130, écart-type=10350)
Trame B	Lognormale (moyenne=4770, écart-type=2362.5)
a1 [71]	0.4
a2	0.11
Résidu des trames I [71]	Normale (moyenne=0, écart-type=521.1)
Longueur des scènes [71]	Géométrique (moyenne=10)

Comme indiqué à la section 3.2.3, un facteur d'unité et un facteur d'échelle ont été appliqués aux données initiales afin de les rendre conformes aux spécifications de [124], i.e. obtenir le trafic stipulé par les spécifications de UMTS. Le profil utilise le schéma de répétition spécifié dans le Tableau 3.15 :

Tableau 3.15 Paramètres de configuration du profil VoD

Application	Videoconférence modifiée
Durée de l'application	Triangulaire (min=27 sec; max=79.8 sec)
Nombre de répétitions de l'application	Une fois
Temps inter-application	N/A
Durée du profil	3600 sec
Nombre de répétitions du profil	Illimité
Temps inter-profil	0

L'application est asymétrique et associée à une classe « en flux ». Plus précisément, la liaison montante utilise les tailles suivantes 211 octets, 71 octets et 24 octets respectivement pour les trames de taille fixe I, P et B.

Téléconférence

Comme pour le trafic VoD, on utilise une version modifiée de la méta-application « videoconferencing » pour permettre l'envoi de trames selon un GoP. Les propriétés des différentes trames sont celles citées dans le Tableau 3.16 :

Tableau 3.16 Paramètres de configuration de l'application Téléconférence

Type de trame	Taille
Trame I	Lognormale (moyenne=2730, écart-type=273)
Trame P	Lognormale (moyenne=917, écart-type=102)
Trame B	Lognormale (moyenne=308, écart-type=92.4)
a1 [53]	1.2
a2	-0.22
Résidu des trames I	Normale (moyenne=0, écart-type=43.4)
Longueur des scènes [23,53]	Infinie

Les valeurs dans le Tableau 3.16 sont obtenues comme indiqué à la section 3.2.3 à partir du modèle de VoD. Nous avons pris un rapport de 12 entre les tailles moyennes des images pour la téléconférence et la VoD; ce rapport reflète la proportionnalité qui existe entre les débits de VoD et de téléconférence tels que spécifiés dans [124]. Ensuite, pour déterminer les écarts-types, nous avons considéré des rapports $(\sigma/\mu)_P=(\sigma/\mu)_B=0.3$ et

$(\sigma/\mu)_I$=0.1. L'écart-type des fluctuations de la taille des trames I est aussi réduit de 12. Le profil utilise le schéma de répétition du Tableau 3.17 :

Tableau 3.17 Paramètres de configuration du profil Téléconférence

Application	Vidéoconférence modifiée
Durée de l'application	exponentiel (moyenne=144 sec)
Nombre de répétitions de l'application	Une fois
Temps inter-application	N/A
Durée du profil	3600 sec
Nombre de répétitions du profil	Illimité
Temps inter-profil	0

Pour la durée, nous avons choisi une distribution exponentielle car la visioconférence ou téléconférence se rapproche dans son utilisation plus d'une session de voix. L'application est symétrique et associée à une classe conversationnelle.

Fax

Le service de fax est modélisé par une adaptation de l'application sur mesure (custom application). Une tâche consiste à envoyer en un paquet, les pages que le fax avait préalablement scannées en mémoire. Nous n'avons pas spécifiquement modélisé les paquets de contrôle (établissement de connexion, négociation des paramètres, etc.) étant donné que leur volume peut être considéré comme négligeable par rapport aux données. Les spécifications du trafic entre la source et la destination sont données dans le Tableau 3.18 :

Tableau 3.18 Paramètres de configuration de l'application Fax

Application utilisée	Application sur mesure (custom)
Nombre de requêtes	1
Temps entre les requêtes	N/A
Taille des requêtes	Exponentiel (moyenne = 281000)

Le profil utilise le schéma de répétition décrit dans le Tableau 3.19 :

Tableau 3.19 Paramètres de configuration du profil Fax

Application	Fax
Durée de l'application	Exponentiel (moyenne=156 sec)
Nombre de répétitions de l'application	Une fois
Temps inter-application	N/A
Durée du profil	3600 sec
Nombre de répétitions du profil	Illimité
Temps inter-profil	0

L'application est asymétrique et associée à une classe interactive.

3.4.2 Validation

Elle se fait à deux niveaux. D'abord au niveau du réseau UMTS par inspection, par comparaison avec d'autres modèles UMTS et par vérification du fonctionnement et de la pertinence des résultats obtenus. Ainsi, une analyse des résultats obtenus nous a permis de détecter une mauvaise configuration du sous-réseau ATM, de la puissance du nœud B et du nombre d'utilisateurs potentiels et de la représentation des usagers de voix. Ensuite, le bien-fondé des niveaux de services atteints par la simulation est analysé et la validation est complétée en remplaçant un certain nombre de types de trafic générés stochastiquement par des traces de trafic réel de même type. Les résultats obtenus dans les deux cas sont finalement comparés. Finalement, ajoutons que le logiciel Opnet étant un logiciel très utilisé en recherche et en industrie, il existe de fait, par la pratique, une certaine validation des composantes de base du logiciel. Nous ne reprendrons donc pas une pareille tâche. Une validation préliminaire des modèles a été aussi faite avec le logiciel Comnet [56].

Modèle de référence

Pour l'établissement du modèle de référence, il fallait concilier des objectifs contradictoires. Le modèle devait être à la fois précis et fiable, tout en étant rapide et léger. Il faut donc trouver un compromis entre d'une part, la précision et la fiabilité qui sont directement reliées au niveau de détail et d'autre part, la rapidité d'exécution qui tend à décroître avec un modèle plus détaillé. D'un autre côté, le temps d'exécution du modèle croît fortement avec le nombre d'usagers mobiles modélisés. Comme les usagers

qui génèrent du trafic de voix sont majoritaires, nous les avons regroupé en lots de 20 par unité mobile ou équipement d'usager mobile (UE). Ceci allège le modèle et le rend plus rapide, mais a sûrement un impact sur les performances de ce type d'application.

Pour évaluer cet impact, nous avons simulé un réseau donné avec une charge donnée de trafic pour différentes tailles de regroupement des UE. Les résultats obtenus pour le délai moyen sur la liaison montante et la probabilité que le délai moyen bout-à-bout sur la liaison descendante soit inférieur à 100 ms sont illustrés aux figures 3.8 et 3.9.

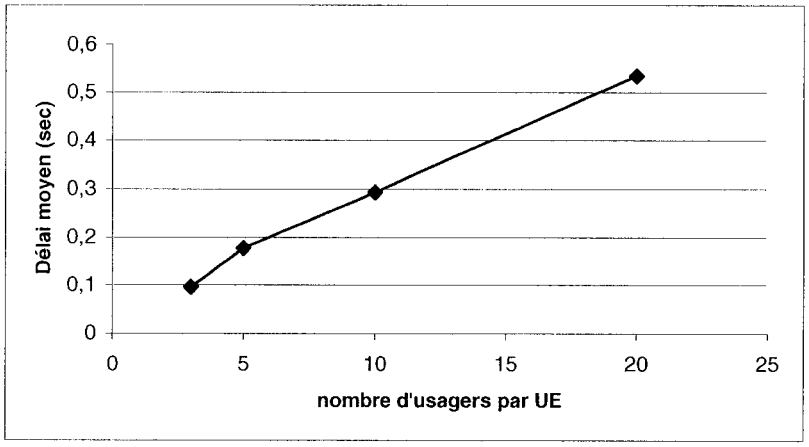


Figure 3.8 Délai moyen sur la liaison montante en fonction de la taille du regroupement d'usagers de voix par UE

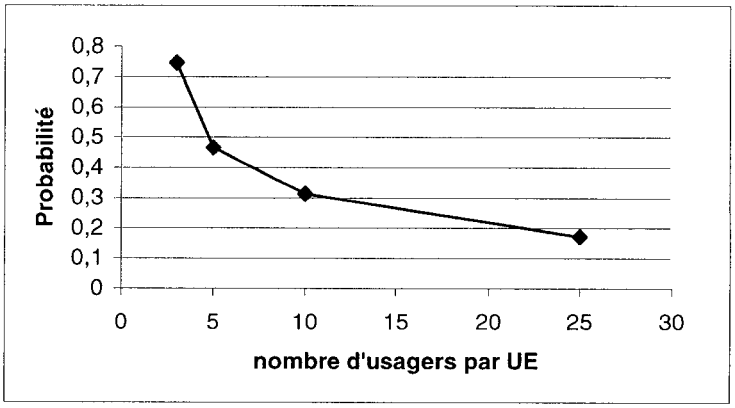


Figure 3.9 Probabilité que le délai moyen bout-à-bout sur la liaison descendante soit inférieur à 100 ms en fonction de la taille du regroupement d'usagers de voix par UE

On note une dégradation des performances du réseau quand la taille du regroupement d'usagers de voix par UE croît. Il faut donc évaluer cette erreur systématique et la corriger pour avoir des résultats fiables et significatifs. Pour cela, nous avons comparé les résultats obtenus pour des regroupements respectifs de 20, 10, 5, 3 et 1 usager de voix par UE sur un réseau qui regroupe toutes les applications. Le Tableau 3.20 montre les valeurs obtenues pour la gigue et les délais moyens sur les liaisons montante et descendante.

Tableau 3.20 Variation de la gigue et des délais moyens pour la voix selon la taille du regroupement

Liaison montante						
délai moyen pour la voix (sec)						**Correction**
profils	*20*	*10*	*5*	*3*	*1*	
délai	0.49101827	0.2696661	0.1530194	0.07404298	0.05442119	**0.43659708**
Liaison descendante						
délai moyen pour la voix (sec)						
profils	*20*	*10*	*5*	*3*	*1*	
délai	0.53430916	0.29352654	0.17615518	0.09601353	0.07569	**0.45861916**
gigue moyenne pour la voix						
profils	*20*	*10*	*5*	*3*	*1*	
gigue	0.19782395	0.09848714	0.04880137	0.01551477	3.43E-05	**1.97790E-01**

Nous déduisons donc une correction à apporter à notre modèle de référence qui regroupe 20 usagers potentiels de voix par UE. Pour la voix, sur les liaisons montante et descendante, cette correction est de l'ordre de 0.44 sec.

De manière générale, on peut noter au niveau des performances des applications de classe conversationnelle (essentiellement composé de la voix) que le délai moyen corrigé dans le modèle de référence est de l'ordre de 50 ms sur la liaison montante et 75 ms bout à bout pour la liaison descendante. En sachant que le délai moyen généralement admis pour la voix est situé entre 150 et 200 ms (voir Tableau 3.2), les performances pour le délai sont largement acceptables. Pour la gigue, on note aussi une bonne performance du modèle de référence (de l'ordre de quelques dizaines de microsecondes) par rapport aux spécifications du Tableau 3.2.

Il faut aussi pour la validation s'assurer que la simplification introduite pour la voix n'a pas un impact trop grand sur les autres classes d'application. Les figures 3.10 à 3.13 montrent les délais moyens pour les classes interactive et en arrière-plan en fonction de la taille des regroupements sur les liaisons montantes et descendantes. On note qu'il n'y a pas de corrélation nette et directe (comme c'est le cas pour la classe conversationnelle) entre la taille des regroupements et les délais obtenus pour les autres classes d'application.

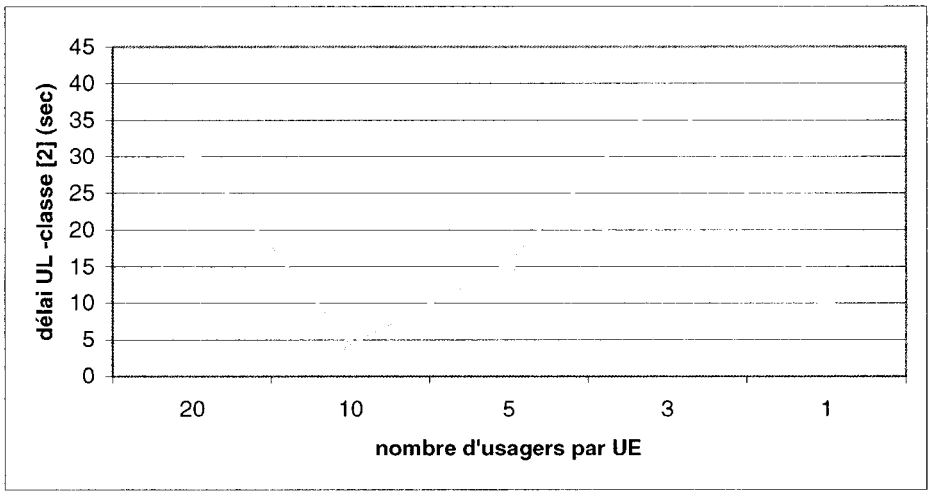


Figure 3.10 Délai moyen sur la liaison montante pour les applications de classe « interactive » en fonction de la taille du regroupement d'usagers de voix par UE

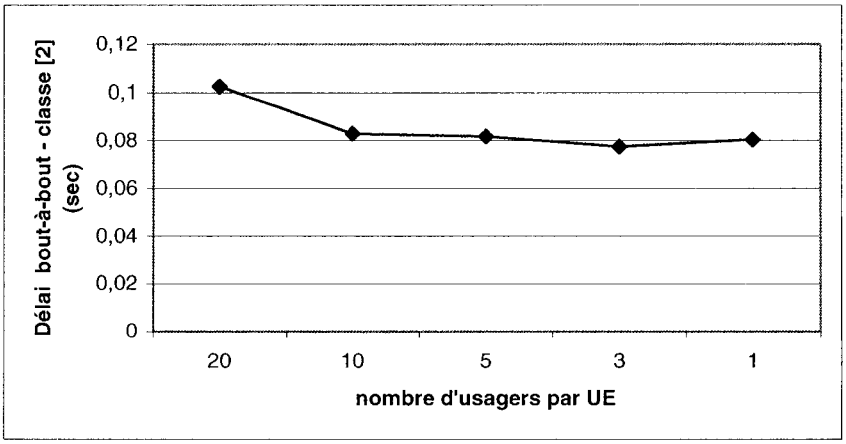


Figure 3.11 Délai moyen bout à bout sur la liaison descendante pour les applications de classe « interactive » en fonction de la taille du regroupement d'usagers de voix par UE

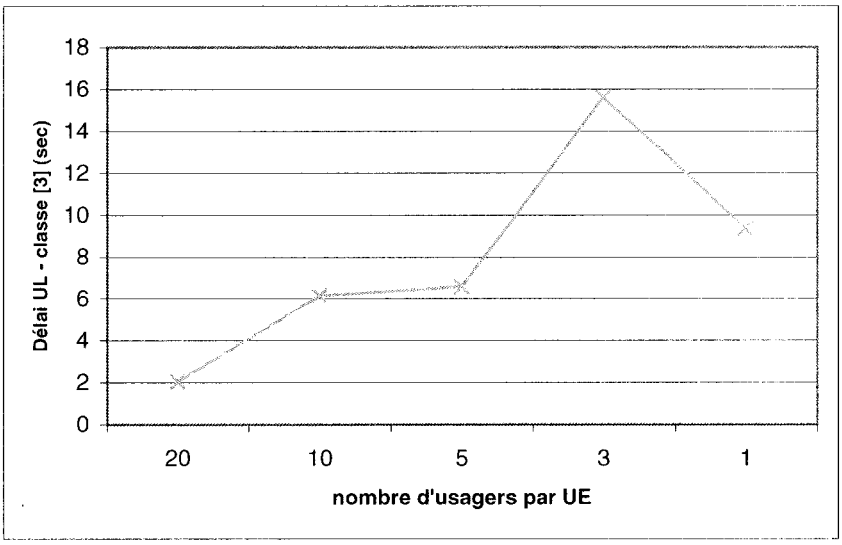


Figure 3.12 Délai moyen sur la liaison montante pour les applications de classe « en arrière-plan » en fonction de la taille du regroupement d'usagers de voix par UE

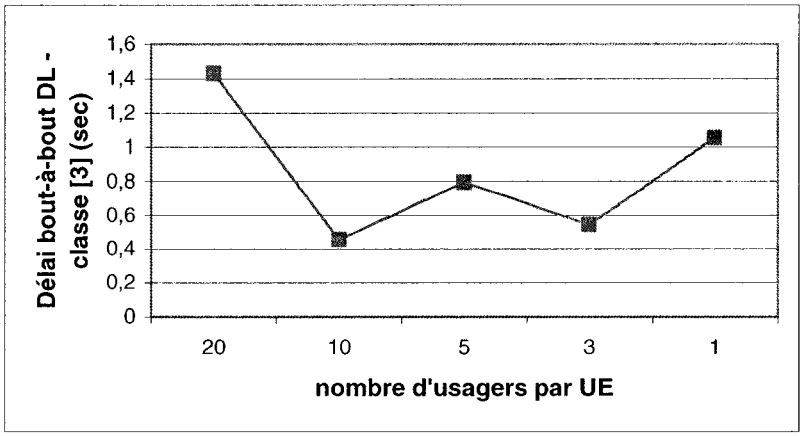


Figure 3.13 Délai moyen bout à bout sur la liaison descendante pour les applications de classe « en arrière-plan » en fonction de la taille du regroupement d'usagers de voix par UE

La variation observée est plus liée à la variation de la séquence de nombres aléatoires, qu'à l'impact de la taille des regroupements.

Impact de la séquence aléatoire

Comme nous l'avions déjà remarqué dans la section précédente, la séquence aléatoire utilisée pour générer les trafics simulés peut avoir un impact non négligeable

sur les valeurs de performance obtenues pour le réseau. Dans cette section, nous essayons d'évaluer cet impact dans le cas de notre modèle de référence.

Pour cela, nous faisons varier la séquence de nombres aléatoires en changeant la semence initiale des générateurs. Les figures 3.14 et 3.15 montrent pour un réseau donné les délais moyens obtenus pour différentes valeurs de la semence initiale.

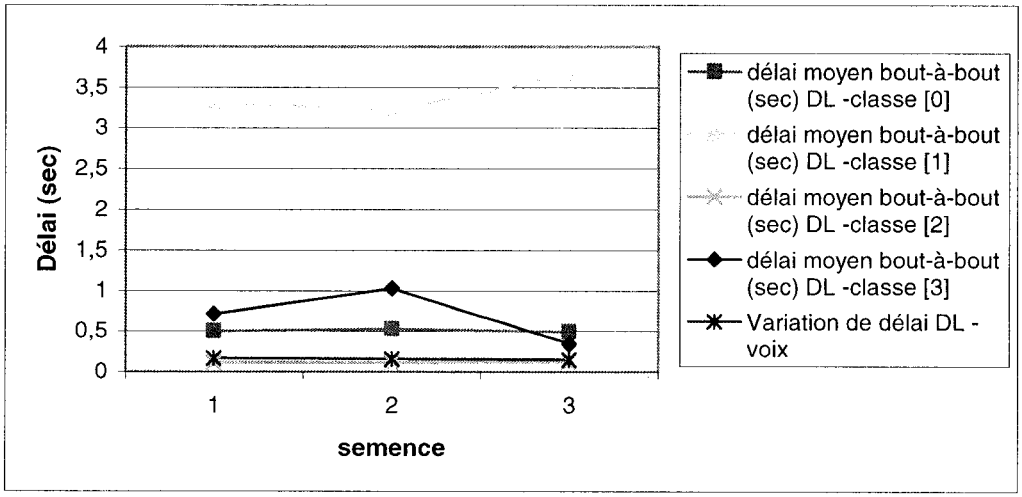


Figure 3.14 Influence de la semence sur les valeurs de performance pour la liaison descendante

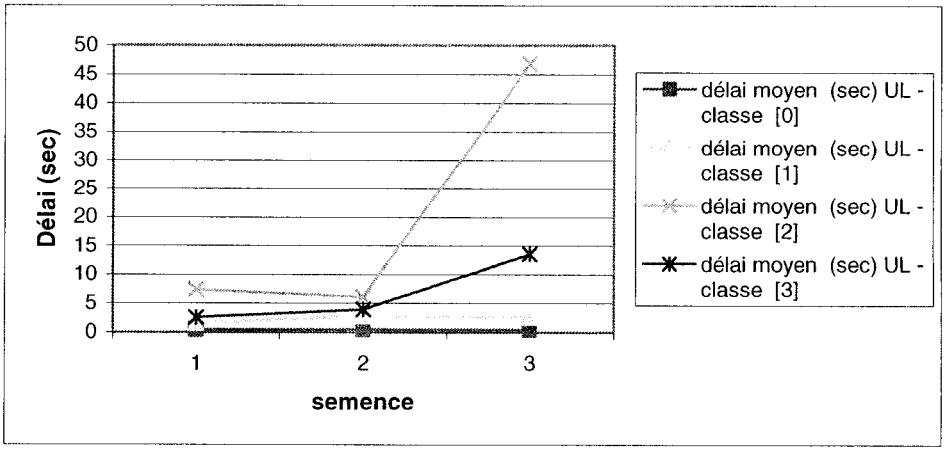


Figure 3.15 Influence de la semence sur les valeurs de performance pour la liaison descendante

On note une variabilité assez grande des résultats obtenus pour les classes de moindre priorité (interactive et arrière-plan) en fonction de la semence, alors que pour les classes de priorité plus élevée, on a un comportement plutôt stable. Nous considérerons donc, surtout pour les classes de moindre priorité, une moyenne des paramètres de performance prise sur plusieurs exécutions avec des semences différentes.

Pertinence des résultats du modèle de référence

Comme nous l'avions déjà vu dans les sections précédentes, les performances du réseau pour les applications de la classe conversationnelle sont largement acceptables, une fois que le biais systématique introduit par la simplification de représentation a été corrigé. Pour les applications des autres classes, le Tableau 3.21 illustre les délais moyens obtenus pour le modèle de référence sur un ensemble de trois semences différentes.

Tableau 3.21 Délai moyen pour le modèle de référence

	Modèle de base
délai moyen bout-à-bout (sec) DL -classe [0]	0.51
délai moyen bout-à-bout (sec) DL -classe [1]	3.40
délai moyen bout-à-bout (sec) DL -classe [2]	0.12
délai moyen bout-à-bout (sec) DL -classe [3]	0.70
Variation de délai DL - voix	0.16
délai moyen (sec) UL -classe [0]	0.46
délai moyen (sec) UL -classe [1]	2.30
délai moyen (sec) UL -classe [2]	20.10
délai moyen (sec) UL -classe [3]	6.67
Délai moyen vidéoconférence	1.63
Délai moyen pour une page HTTP	14.13

Pour la navigation Internet, on a une moyenne de 14 sec par page. Ce qui est assez proche des 10 sec du Tableau 3.2. De manière générale, les délais pour les trames de classe 2 sont de l'ordre de 20 sec sur la liaison montante et 100 ms sur la liaison descendante. De même pour la vidéoconférence, on a des délais de l'ordre de 1.5 sec, ce qui est relativement plus grand que les 150 ms visées.

Pour pouvoir analyser les résultats, nous avons aussi mesuré les délais de référence obtenus en simulant chaque type d'application toute seule (par exemple 20 usagers de voix seuls, une application fax seule, etc.) pour une durée constante.

Tableau 3.22 Délai indicatif pour des applications seules

Voix seule	
délai moyen bout-à-bout (sec) DL -classe [0]	0.41
délai moyen (sec) UL -classe [0]	0.41
Variation de délai DL - voix	0.16
Vidéoconférence seule	
délai moyen bout-à-bout (sec) DL -classe [0]	1.7
délai moyen (sec) UL -classe [0]	0.0026
VoD seule	
délai moyen bout-à-bout (sec) DL -classe [1]	3.4
délai moyen (sec) UL -classe [1]	1.76
HTTP seul	
délai moyen bout-à-bout (sec) DL -classe [2]	0.082
délai moyen (sec) UL -classe [2]	0.37
Délai moyen pour une page HTTP	13.07
Fax seul	
délai moyen bout-à-bout (sec) DL -classe [2]	0.156
délai moyen (sec) UL -classe [2]	52.8

En comparant les tableaux 3.21 et 3.22, on note que l'agrégation de trafic n'a pas un impact significatif sur le délai. Les différences entre les délais obtenus dans le contexte agrégé et le contexte où les applications sont seules ne sont pas très significatives. Ceci pourrait s'expliquer par le fait qu'avec des canaux dédiés si la liaison n'est pas *sur-souscrite*, chaque application a une garantie de bande passante, ce qui l'isole de l'influence des autres applications. En effet, dans le réseau UMTS, les demandes de connexion sont accordées par un algorithme d'acceptation de connexion [55]. Cet algorithme estime le facteur de charge sur la liaison radio et accepte ou refuse la connexion. On dit que la liaison est *sur-souscrite* si le total des demandes de réservation de canaux accordées est supérieur à la capacité de la liaison.

Les mauvaises performances obtenues dans le cas du réseau simulé ici sont donc reliées plus au contexte PDP utilisé. En effet, pour nos contextes PDP, nous avons utilisé les recommandations en terme de bande passante du forum UMTS pour réserver les

canaux sur les liaisons montantes et descendantes. De manière générale, les débits réservés pour les canaux sont du même ordre que le débit moyen des applications. Par exemple, une application de vidéoconférence, on a un débit de 128 kbps avec un facteur d'occupation de 0.8, ce qui donne un débit réel moyen de 102.4 kbps que l'on cherche à véhiculer sur un canal de 128 kbps. De ce fait, les délais obtenus sont énormes. Le Tableau 3.23 montre, pour le cas de la vidéoconférence, les délais moyens obtenus pour différentes tailles de canaux.

Tableau 3.23 Impact de la taille du canal dédié sur les performances

Taille du canal	128	256
Délai moyen (en sec)	1.70	0.1

Il apparaît donc que le profil de QoS associé à chaque application est très important. Ceci explique par exemple le fait que pour la classe 2, les délais soient aussi différents sur les liaisons montantes et descendante. En effet, à cause de l'agrégation de deux types différents de trafic (voir Tableau 3.5) le canal descendant est beaucoup plus large que le canal montant et fournit donc des délais largement inférieurs.

Dans le Tableau 3.24, on montre les résultats pour une simulation du modèle de référence avec, pour la classe interactive, un profil de QoS où les liaisons descendante et montante ont la même capacité de 384 kbps. Dans la suite, nous désignerons ce modèle par *modèle de référence modifié.*

Tableau 3.24 Délais moyens (en sec) pour le modèle de référence modifié

délai moyen Email	125.926439
délai moyen HTTP	12.4458829
délai moyen VoD	3.38709209
délai moyen Vidéoconférence	1.69792746
délai moyen Voix	0.51079719
Gigue (sec) - DL- voix	0.20802118
délai moyen (sec) -UL- classe [0]	0.48208875
délai moyen (sec) -UL- classe [1]	2.52268513
délai moyen (sec) -UL- classe [2]	0.43520919
délai moyen (sec) -UL- classe [3]	2.83262011
délai moyen bout-à-bout (sec) - DL - classe [2]	0.10131472
délai moyen bout-à-bout (sec) - DL - classe [3]	0.68507507

On note tout de suite une différence au niveau des délais de la classe interactive. Le concepteur d'un réseau UMTS devrait donc bien caractériser les profils de QoS (débit, type de trafic) de chacune de ses applications pour assurer une bonne performance de son réseau. En règle générale, un canal dont le débit est du même ordre que le trafic moyen écoulé génère des délais inacceptables. Ceci est encore plus important dans un cadre où, comme dans notre simulation, on associe un contexte à chaque classe de trafic et non à chaque application. Aussi dans le cas de l'utilisation de canaux dédiés, la facturation de l'application dépendra directement du profil de QoS (ou contexte) qui lui est associé. Le chapitre 5 traite plus détail des problèmes liés à la facturation.

Importation de trafic

Dans la dernière étape de la validation, nous avons substitué à nos modèles de trafic des traces réelles pour les applications de voix et vidéo. Les traces utilisées ont été obtenues à partir de [43,60,125] et ensuite mises à l'échelle pour générer une charge globale équivalente à celle des modèles. Les résultats obtenus en remplaçant les applications de vidéoconférence, vidéo sur demande et de navigation Internet par les traces sont présentés à la Figure 3.16.

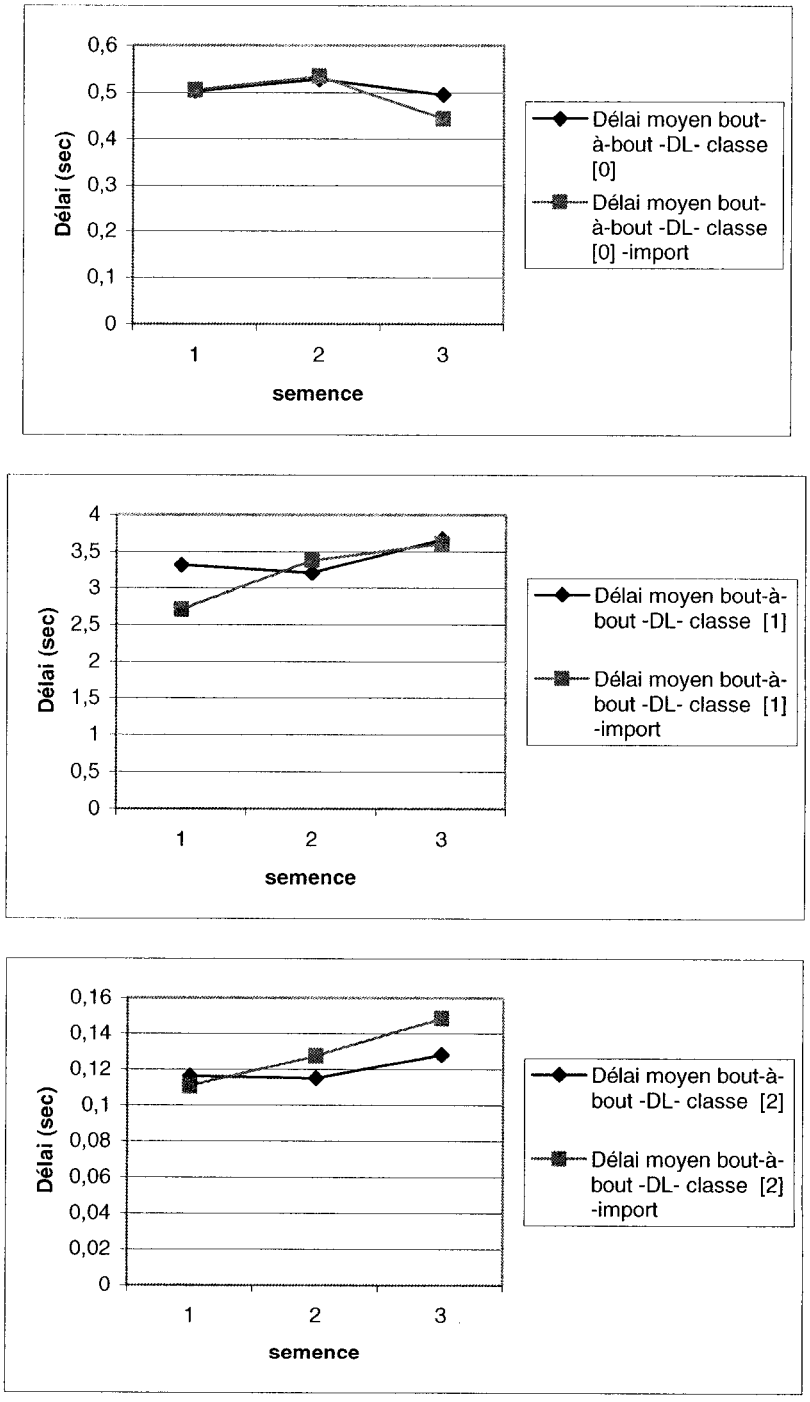


Figure 3.16 Comparaison des performances pour le trafic importé et le trafic simulé

Le Tableau 3.25 résume aussi les différences entre les cas de trafic simulé et importé dans le cas du modèle de référence modifié.

Tableau 3.25 Comparaison des performances pour le trafic importé et le trafic simulé

	Délai bout à bout DL- classe [0]	Délai bout à bout DL- classe [1]	Délai bout à bout DL- classe [2]	Délai bout à bout DL- classe [3]
importé	0.50761747	3.29519031	0.09586509	1.08033016
simulé	0.48305723	3.51382758	0.10380193	0.80513087
Écart relatif (%)	8.86693746	7.17658102	1.5362808	19.1752766

Les différences en valeurs relatives entre le cas où le trafic est importé de traces réelles et celui où il est simulé sont inférieures à 10 % en moyenne sauf pour la classe 3 (en arrière-plan). Cependant, pour cette classe on avait déjà noté une grande variabilité des résultats par rapport à la semence, ce qui ne nous fournissait pas une grande précision dès le départ.

3.4.3 Expérimentations

Impact du facteur de charge

Comme nous l'avons souligné, quand on utilise des canaux dédiés, si la liaison n'est pas sur-souscrite, chaque application a une garantie de bande passante, ce qui l'isole de l'influence des autres applications. Ici, nous étudions l'impact de la sur-souscription du lien sur les performances obtenues. Les figures 3.17 et 3.18 montrent l'évolution des délais moyens pour différentes classes et applications sur les liaisons montante et descendante pour le modèle de référence modifié en pour des facteurs de charge de plus en plus importants. Nous n'avons pas représenté la classe 3 à cause de sa trop grande variabilité.

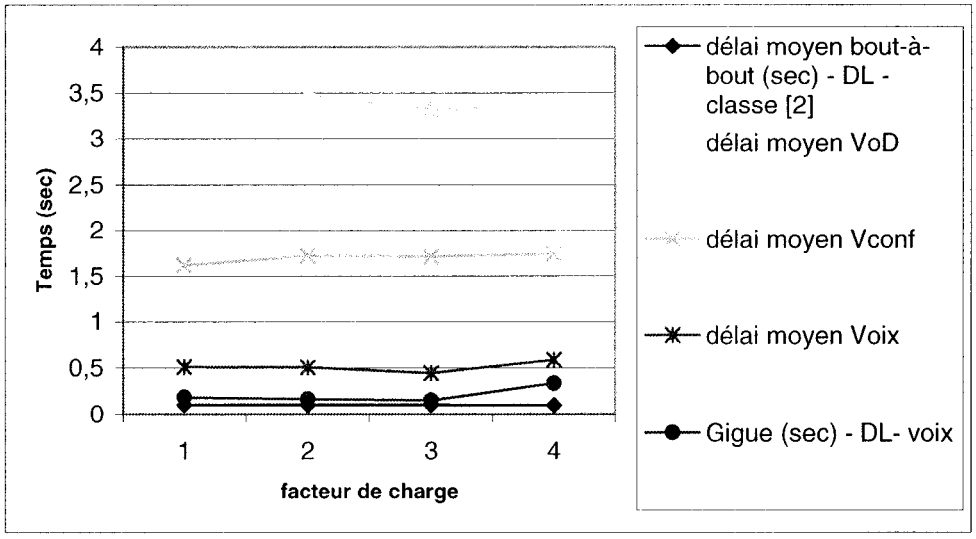


Figure 3.17 Délais moyens (sens descendant) en fonction du facteur de charge

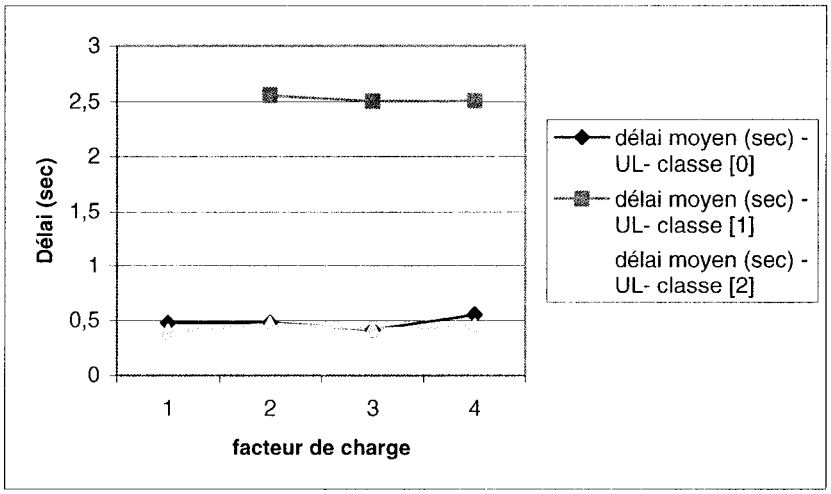


Figure 3.18 Délais moyens sur la liaison montante en fonction du facteur de charge

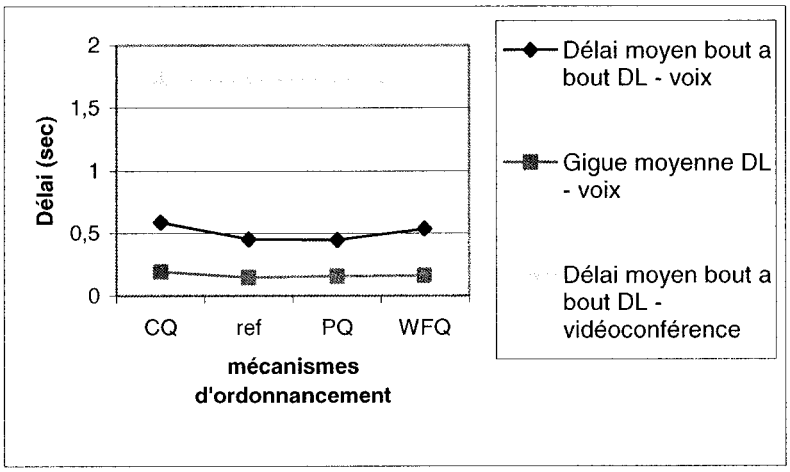
On note en général après une légère amélioration, que les performances se dégradent avec l'augmentation du facteur de charge surtout sur la liaison descendante. L'amélioration initiale pourrait s'expliquer par le fait que dans un premier temps, on admet plus de connexions dont les délais demeurent raisonnables (moins que les maximums), ce qui diminue le délai total moyen. Mais après un certain seuil de sur-

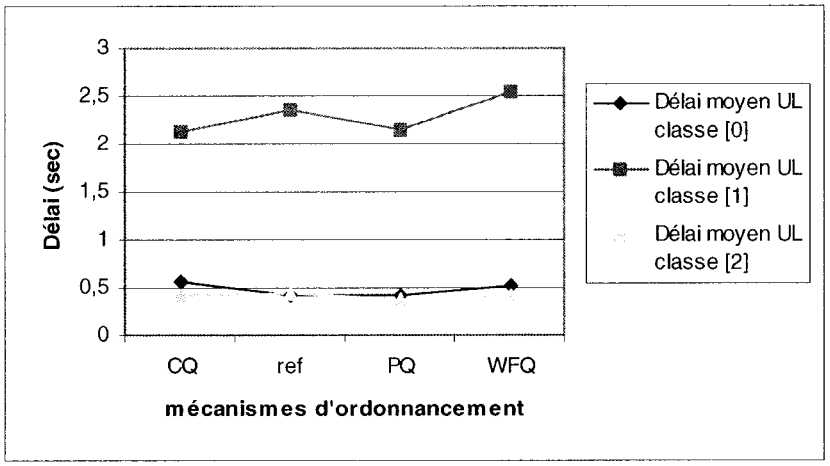
souscription, les performances se dégradent, car pour faire la sur-souscription, on ne respecte plus les canaux dédiés à chaque application. Le phénomène est plus notable sur la liaison descendante qui supporte une plus grande charge.

Impact de la qualité de service au niveau IP

Le modèle UMTS implémente la gestion de la qualité de service au niveau de la couche MAC. Cette gestion de QoS se réalise par la gestion des priorités des classes. Dans le modèle de référence developpé, la gestion de QoS repose donc essentiellement sur cette prioritisation des classes au niveau de la couche MAC et sur les profils de QoS. A priori, l'ajout d'un niveau de gestion de QoS supplémentaire au niveau de la couche IP devrait améliorer les performances du réseau. La Figure 3.19 compare les délais moyens pour les différentes applications dans le cas du modèle de référence (ref) et pour divers mécanismes d'ordonnancement de file d'attente, dont le partage équitable pondéré (WFQ), l'ordonnancement par priorité (PQ) et par classe (CQ). Ces mécanismes d'ordonnancement s'ajoutent bien entendu à l'ordonnancement initial établi par la priorisation au niveau MAC.

On note que le mécanisme d'ordonnancement par priorité permet une meilleure performance pour quasiment toutes les applications. En effet, ce mécanisme prolonge tout simplement la priorisation déjà implantée au niveau MAC de l'UE d'une part vers les couches supérieures et d'autre part dans le réseau cœur.





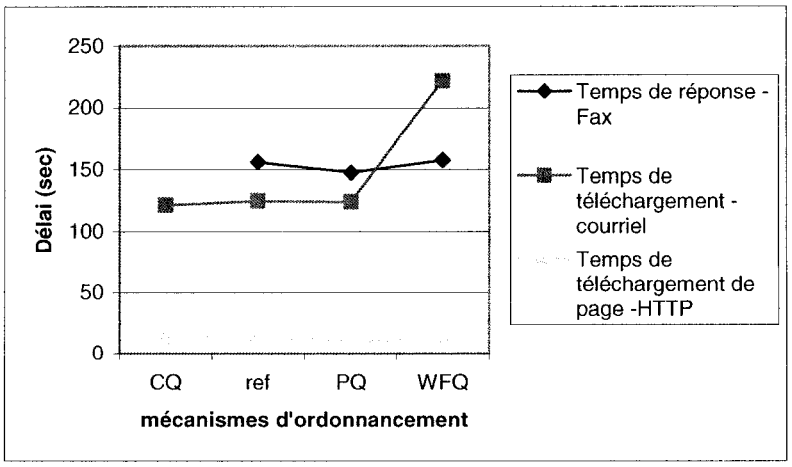


Figure 3.19 Délais moyens pour différentes applications et différents schémas d'ordonnancement

3.5 Conclusion

Dans ce chapitre, nous avons présenté notre modèle de référence pour l'étude des performances des réseaux UMTS. Le modèle résulte d'une combinaison de modèles de trafic et d'un modèle de réseau, tous les deux conformes aux spécifications du forum UMTS [124]. Dans l'implémentation, les performances sont moyennes car les débits suggérés par le forum UMTS sont un peu restrictifs. Il faut donc définir avec soin les profils de QoS associés à chaque classe ou application.

L'approche du forum 3GPP [1] qui propose, pour chaque classe, un ensemble de débits possibles est plus souple et efficace et permet une sur-souscription du lien pour avoir plus d'efficacité. Cependant, il faut déterminer le seuil à partir duquel la sur-souscription dégrade les performances. Pour la classe « en flux », il faudrait mettre en place des tampons adéquats pour assurer la qualité de service notamment par rapport aux délais. Finalement, la prise en compte supplémentaire de la gestion de la QoS par un mécanisme d'ordonnancement au niveau des couches supérieures permet une amélioration des performances.

CHAPITRE 4

MODÈLE DE FIABILITÉ

Dans le chapitre précédent, nous nous sommes intéressés à l'évolution de divers critères de performance en fonction des paramètres du réseau. À cause de l'échelle de temps des expériences précédentes, les indices (critères) externes étudiés sont des indices tels le débit, le délai, le temps de réponse, etc. Pour étudier des indices comme la fiabilité, il faut une autre échelle de temps à cause de la rareté des événements impliqués. Dans ce chapitre, nous aborderons le problème de la fiabilité dans une implémentation possible d'un réseau de troisième génération. Pour cela, nous commencerons par quelques concepts de base, ensuite nous présenterons l'organisation topologique globale d'un réseau de troisième génération. Par la suite, nous introduirons quelques considérations de fiabilité. Finalement, nous présenterons une approche analytique, les modèles de simulation et les résultats.

4.1 Définitions et concepts de base

Un graphe G est défini par un couple (N,A) où N désigne l'ensemble des *nœuds* et $A \subseteq N*N$ est l'ensemble des *liens* (ou *liaisons*) du graphe. On définit donc une liaison comme un couple (ou une paire) (m,n) où m et n sont les extrémités de la liaison. À chaque liaison, on associe un certain nombre de caractéristiques comme la *capacité*, qui est la quantité maximale d'information que peut véhiculer le lien par unité de temps, le *flot*, qui est la quantité réelle d'information véhiculée par le lien par unité de temps. À chaque couple de nœuds (m,n), on peut associer un *trafic* qui ne dépend pas de l'existence ou non d'une liaison entre le couple de nœuds et qui est défini comme le nombre de paquets échangés par unité de temps entre les deux nœuds.

Un graphe est *K-connexe au sens des liaisons* si et seulement si chaque paire de nœuds est reliée par au moins K *chemins disjoints de liaisons*. De façon similaire, on

définit un graphe *K-connexe au sens des nœuds* comme un graphe où chaque paire de nœuds est reliée par au moins *K chemins disjoints de nœuds.*

De manière générale, un réseau peut avoir une topologie en étoile, en arbre, en anneau ou une topologie maillée [110]. Dans une topologie en étoile, tous les nœuds sont reliés à un nœud central et toutes les communications sur un tel réseau transitent par le nœud central. Dans une topologie en arbre, le réseau a une structure hiérarchique. Le nœud principal de la topologie est le nœud racine. En général, les liens dans une telle topologie sont des câbles (avec plusieurs branches) auxquels on relie une ou plusieurs stations. Dans une topologie en anneau, tous les nœuds sont interconnectés de manière à former un anneau. Finalement, dans une topologie maillée, chaque paire de nœuds est reliée par plus d'un chemin. Si chaque paire possible de nœuds est reliée par une liaison directe, on parle de topologie totalement maillée ou topologie complète.

4.2 Topologie globale d'un réseau de troisième génération

Généralement, les réseaux de télécommunications sont classés en trois catégories selon la grandeur de l'étendue géographique desservie : les réseaux locaux (*LAN* ou *Local Area Network*), les réseaux métropolitains (*MAN* ou *Metropolitan Area Network*) et les réseaux étendus (*WAN* ou *Wide Area Network*). La Figure 4.1 illustre l'organisation topologique globale d'un réseau de troisième génération (3G). Dans un tel réseau, le réseau cœur est un réseau métropolitain ou un réseau étendu qui relie les *GGSN* ou *SGSN* de différents opérateurs, alors que le réseau d'accès est souvent un réseau local qui est constitué de concentrateurs d'accès équivalents en fait aux RNC (*Radio Network Controller*). Les RNC d'un même opérateur sont connectés entre eux par un réseau métropolitain souvent basé sur la technologie ATM. Ces RNC sont aussi reliés à un GGSN qui sert de passerelle entre le réseau de l'opérateur et d'autres réseaux externes à commutation de paquets. On parlera de point de présence (*PoP ou Point of Presence*) pour désigner cette interconnexion du réseau de l'opérateur au sous-réseau de transport. À l'intérieur d'une cellule multiservice, on retrouve les usagers mobiles et des

nœuds B. Selon le type de cellules, la structure du nœud B peut être plus ou moins élaborée. En général, un site de nœud B comportera une antenne et un module qui intègre le matériel radio constitué d'un émetteur-récepteur, d'un multiplexeur, d'un concentrateur et d'interfaces pour les liens ATM ou liens radio. Ces interfaces permettent aux nœuds B de communiquer aussi bien avec les usagers qu'avec les RNC. Les RNC peuvent aussi être reliés entre eux.

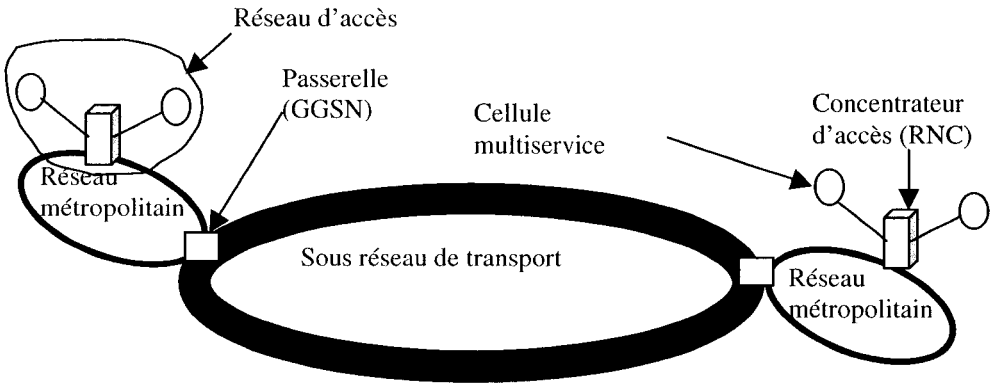


Figure 4.1 Organisation topologique globale d'un réseau cellulaire de 3ème génération

Dans la migration des réseaux 2G vers 3G, une solution d'architecture possible consiste à utiliser dans le réseau d'accès la technologie EDGE. Dans ce cas, on peut utiliser les technologies Wireless-DSLAM et LMDS ou MMDS au niveau du réseau métropolitain comme illustré à la Figure 4.2.

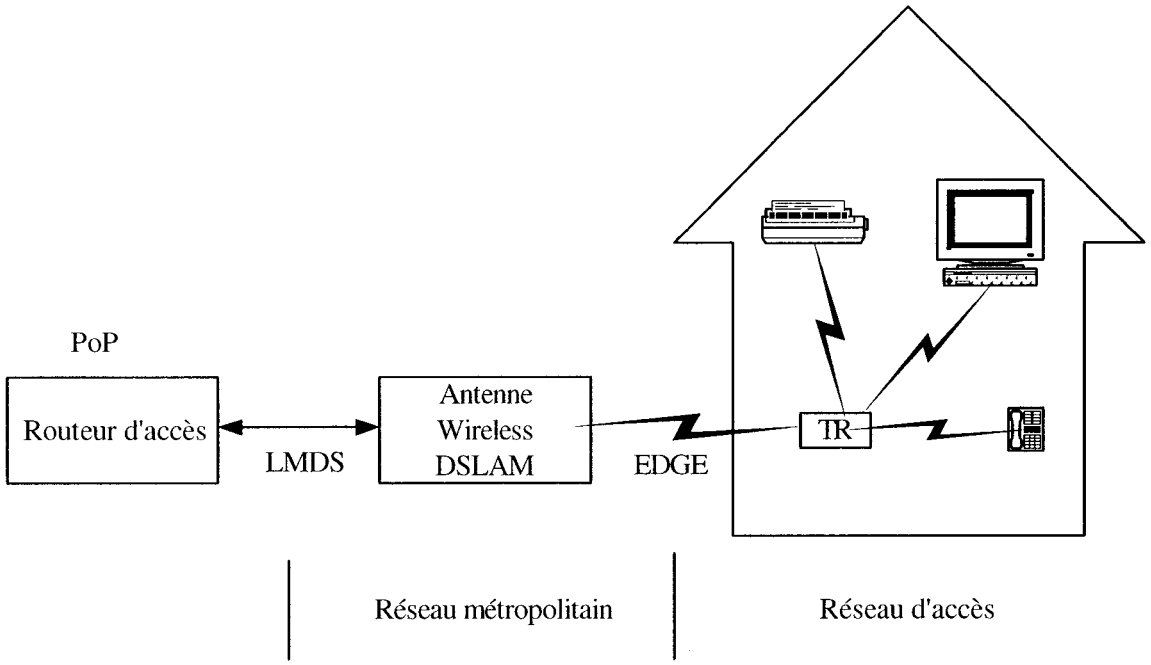


Figure 4.2 Architecture d'un système W-DSLAM

Le module DSLAM joue ici un rôle équivalent à celui du nœud B. Il met en relation l'interface radio de distribution et l'interface radio agrégée. Du côté agrégé, on peut atteindre des débits de l'ordre de 40 Mbps, tandis que l'interface de distribution qui relie le module DSLAM aux usagers peut, dans certains cas, supporter jusqu'à 300 usagers avec un débit de crête de 2 Mbps. Le routeur d'accès correspond au RNC et est relié au GGSN qui sera considéré comme le point de présence. Le TR (terminaison de réseau) permet de créer chez l'usager un réseau local avec une technologie du genre Bluetooth par exemple [51].

4.3 Considérations de fiabilité

Les principaux composants d'un réseau sont les nœuds et les liens. Les nœuds peuvent être des terminaux, des serveurs, des ordinateurs, des multiplexeurs, des concentrateurs, des routeurs, des équipements mobiles, etc. En général, dans un réseau, on s'intéresse à la capacité, au coût, à la disponibilité et à la fiabilité.

Nous définissons la *disponibilité* comme la probabilité que le réseau soit utilisable dans une période donnée en tenant compte des mécanismes de redondance, de détection de défaillance, de reconfiguration et des procédures de réparation [109,110]. La disponibilité D peut s'exprimer par la formule suivante :

$$D = \frac{MTBF}{MTBF + MTTR}$$

où $MTBF$ désigne le temps moyen entre les pannes (Mean Time Between Failures) et $MTTR$ le temps moyen qui s'écoule avant que la défaillance ne soit réparée (Mean Time To Repair). La *fiabilité* peut être définie comme la probabilité qu'au moins un chemin existe entre chaque paire de nœuds dans le réseau [65]. Elle dépend de la disponibilité et de la fiabilité des composantes du réseau. Il est donc nécessaire d'évaluer la fiabilité totale du réseau en prenant en compte, dès la phase de conception topologique, la possibilité de défaillance des nœuds ou liaisons [14,62,65,97,109,112]. Pour cela, plusieurs mesures de fiabilité ont été proposées [62,65,109,111]. La plus populaire parmi ces mesures est la *k-connexité* qui intègre à la fois les notions de connexité au sens des nœuds et au sens des arcs (ou liaisons). Cependant, la connexité au sens des nœuds est une contrainte plus forte que celle au sens des arcs. De ce fait, elle est plus appropriée pour mesurer la fiabilité du réseau et pour garantir un niveau de tolérance aux défaillances. Toutefois, garantir une forte fiabilité suppose l'utilisation de plus d'équipements redondants et l'augmentation du coût d'implantation du réseau.

Il s'agit donc de déterminer un compromis entre le coût du réseau et son niveau de fiabilité. Plus précisément, on suppose que les types de nœuds et leurs positions sont connus, on cherche alors la topologie qui minimise le coût total du réseau pour une capacité des nœuds et une structure de coût données, sous une contrainte de fiabilité qui peut être par exemple une probabilité de défaillance de 0.001% pour le trafic de voix.

On redéfinit ici un *point de présence* comme la localisation d'un point d'accès à l'Internet. Dans la Figure 4.1, le point de présence correspond à la localisation de la connexion du GGSN au réseau externe. Le réseau UMTS considéré est constitué de

« routeurs » d'accès AXI 540 qui correspondent aux RNC, de concentrateurs et de nœuds B interconnectés comme à la Figure 4.3.

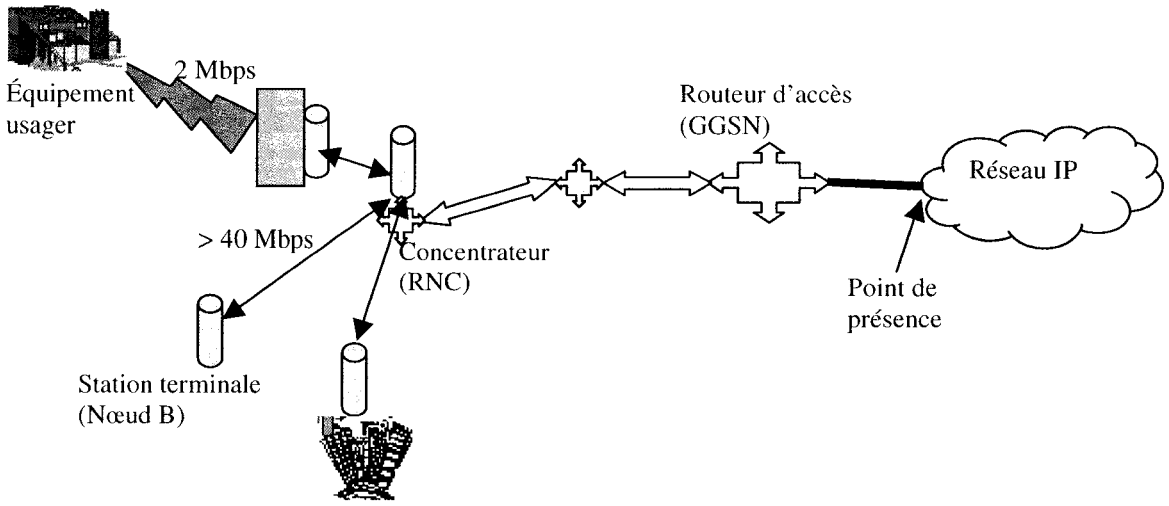


Figure 4.3 Réseau UMTS considéré

Les routeurs AXI sont modulaires et peuvent, avec l'ajout de quelques modules supplémentaires, jouer le rôle de RNC ou de nœud B-RNC intégré. Les statistiques de MTBF du routeur ont été calculées selon les standards de *Bellcore*™ [15] et le Tableau 4.1 présente les MTTF pour les différents composants (MTTF=MTBF si le temps de réparation est inférieure au MTTF).

Tableau 4.1 MTTF pour les composants de base du routeur AXI 540

Composant	MTTF
Matrice de commutation	30.8 ans= 269808 heures
Panneau arrière	113.4 ans= 993384 heures
Tableau de diodes	695.8 ans= 6095208 heures
Processeur de routage	389.48 ans= 3411144 heures

Les différents modules du routeur sont tous remplaçables « à chaud », minimisant ainsi le MTTR. De plus, tous les modules ont une redondance 1:1 (pour chaque composant actif, il y a un composant de réserve) et le remplacement « à chaud » des composants peut se faire sans arrêter le système. Un technicien avec un minimum de formation peut remplacer n'importe quel module du AXI 540 en l'espace de 10 minutes.

Cependant, nous supposons dans notre cas que le MTTR est de 2 heures, car cette valeur est celle qui est souvent considérée en pratique puisqu'elle tient compte par exemple du temps de déplacement du technicien.

Du point de vue logiciel, on peut avoir deux types de défaillance :

- une défaillance mineure qui entraîne un redémarrage du processus ;
- une défaillance majeure qui nécessite le redémarrage ou le remplacement du processeur.

Pour une défaillance mineure, la résolution est quasi-instantanée avec un redémarrage du processus dans un délai de 1 à 2 secondes. Les conséquences des défaillances logicielles majeures sont minimisées par la possibilité de remplacer les composants « à chaud » et la redondance matérielle 1:1 qui permet à un autre processeur de prendre le relais.

Si les distributions des défaillances sont indépendantes et distribuées exponentiellement avec un temps moyen d'activité égal au MTTF, la disponibilité de l'ensemble du routeur peut être approximée par les valeurs du Tableau 4.2.

Tableau 4.2 Disponibilité pour le routeur AXI 540

	MTTF (heures)	Non disponibilité (MTTR/MTTF)	Disponibilité
Matrice de commutation	269808	$7.41268*10^{-6}$	0.999992587
Panneau arrière	993384	$2.01332*10^{-6}$	0.999997987
Tableau de diodes	6095208	$3.28127*10^{-7}$	0.999999672
Processeur de routage	3411144	$5.86314*10^{-7}$	0.999999414
Total sans redondance		$1.03404*10^{-6}$	0.99998966
Total avec redondance 1:1		$1.0692387*10^{-10}$	0.999999999

On note que, sans la redondance, le routeur ne satisfait pas la disponibilité cible de 0.001 % de probabilité de défaillance pour le trafic de voix. De plus, en tenant compte de la redondance, la probabilité de défaillance matérielle du routeur est très faible et largement supérieure aux exigences de disponibilité.

De manière similaire, si on considère les équipements UMTS de divers fournisseurs [82], ces derniers sont conçus pour une fiabilité très grande. Par exemple, le nœud B proposé par la compagnie *Lucent*™, le *Flexent OneBTS Node B*™ utilise une

redondance de type *N+1*, i.e. pour *N* modules actifs, on a un module redondant. Ainsi, la défaillance d'un seul module n'entraîne pas de discontinuité dans le fonctionnement de l'équipement. Du point de vue logiciel aussi, les conséquences d'une défaillance logicielle sont minimisées par la possibilité de charger les logiciels en arrière-plan. De ce fait, nous supposerons que la fiabilité des nœuds est suffisante, même lorsqu'on considère un réseau constitué d'un ensemble de nœuds qui interagissent.

À cause de l'organisation hiérarchique du réseau UMTS, les équipements en amont tels le SGSN et le GGSN supportent une grande partie du réseau (des usagers du réseau) et, de ce fait, sont cruciaux pour le bon fonctionnement du réseau. Les équipements situés en aval ont des contraintes de fiabilité moindre car ils supportent moins d'usagers, tandis que ceux en amont tel le *Flexent GGSN* ™ de *Lucent*™ ont en général une fiabilité supérieure au 99.999 % requis pour la voix [82].

Nous supposerons donc que la fiabilité du chemin (nœuds et liens) qui va des usagers au RNC est suffisante et nous considérons seulement la fiabilité des liaisons au niveau du réseau qui relie les RNC entre eux et aux GGSN. Les topologies considérées varient de la topologie en arbre à la topologie partiellement maillée, comme illustré aux figures 4.4 à 4.6.

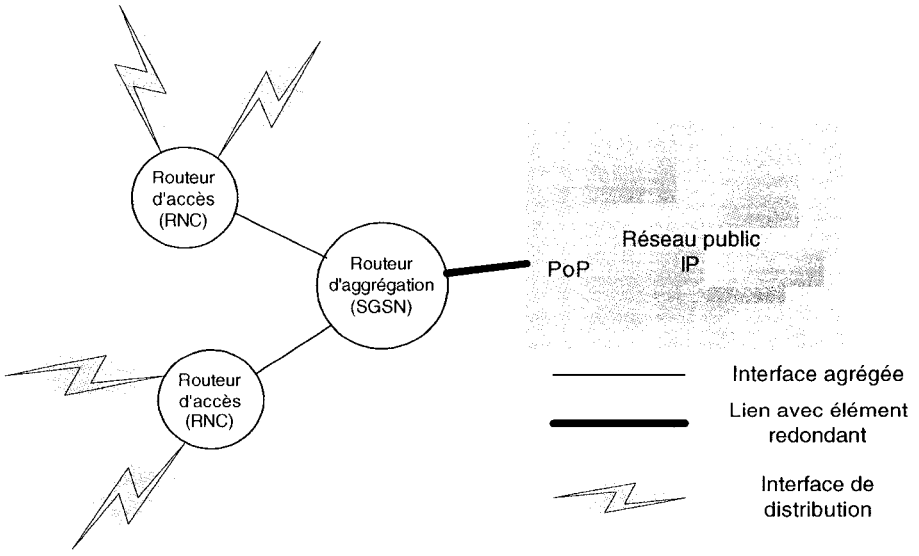


Figure 4.4 Topologie en arbre avec liens redondants parallèles

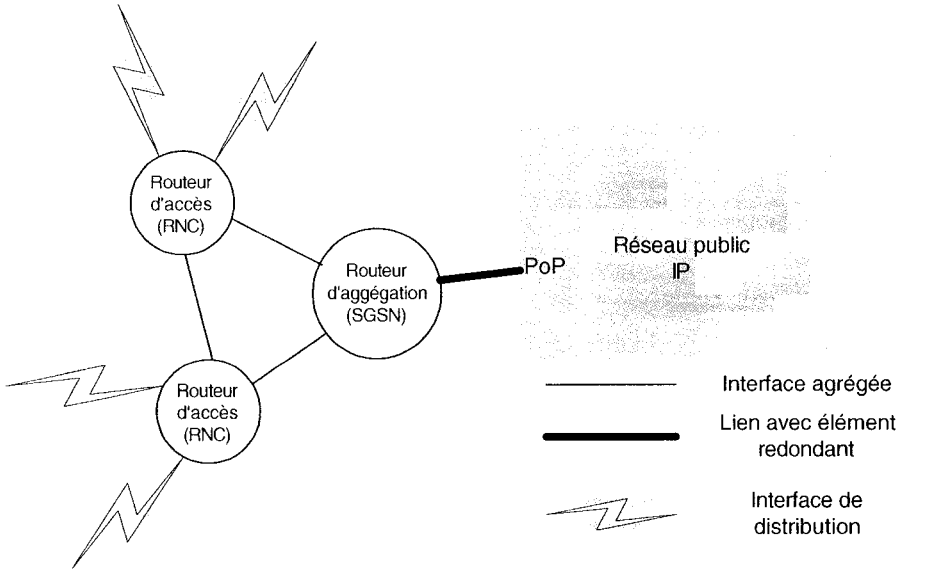


Figure 4.5 Topologie en anneau avec un accès au point de présence

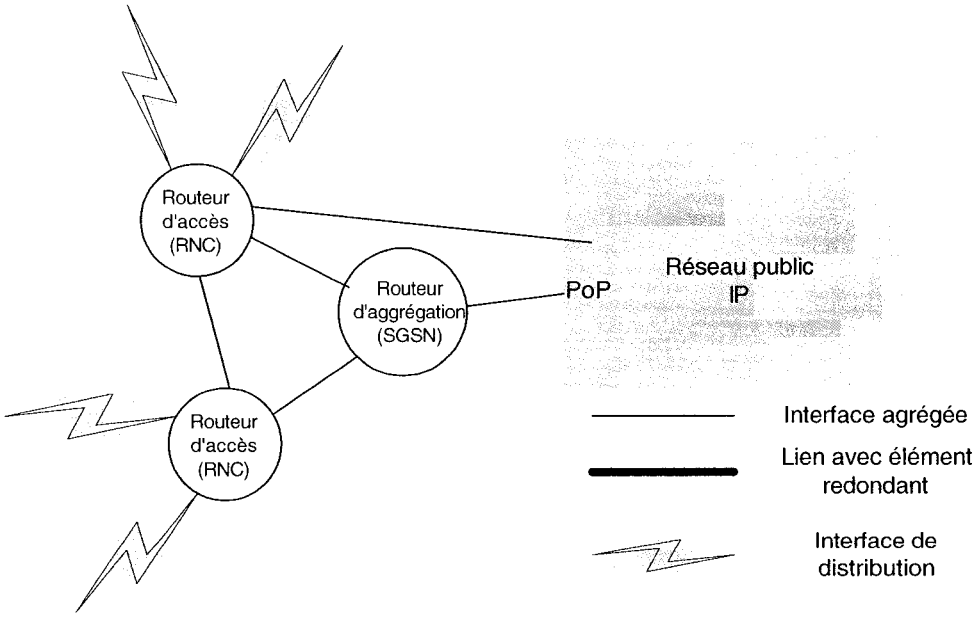


Figure 4.6 Topologie partiellement maillée

4.4 Approches analytiques

Considérons les liens du routeur d'accès au point de présence ou entre routeurs d'accès. Ces liens sont souvent des liens optiques à l'échelle métropolitaine. Si on suppose des liaisons moyennes de 120 km de longueur et si on considère les données de [96], on détermine une probabilité de défaillance de la liaison de $1.2*10^{-5}$. Cette même probabilité peut être retrouvée en supposant des liaisons plus courtes de 20 km de longueur en moyenne et en réduisant le MTTR de façon proportionnelle.

La probabilité de défaillance pour une liaison est donc supérieure à la probabilité cible qui est de 10^{-5}. Dans une topologie où les liaisons du routeur d'accès au point de présence ou les liaisons entre routeurs ne sont pas protégées, la fiabilité n'est donc pas assurée, d'autant plus que le nombre d'usagers dont la connexion passe par ce lien est en général très grand (les routeurs d'accès sont de grands points de concentration). Par exemple, le *Flexent RNC* de *Lucent* peut, avec un seul module, supporter 180 nœuds B, 310 Mbps de débit brut et 4000 Erlangs de trafic de voix [82]. Nous essayerons donc d'améliorer la fiabilité au niveau liaison en augmentant la connexité tout en essayant de maintenir un coût de revient raisonnable pour le réseau. Nous évaluerons donc différentes topologies dont la topologie en arbre avec liens redondants parallèles. Pour cette topologie, deux cas seront considérés :

- le réseau ne supporte aucune surcharge : dans le cas d'une défaillance, le réseau continuera son fonctionnement normal seulement s'il existe une liaison auparavant inutilisée pour remplacer la liaison défaillante ;
- le réseau supporte une redistribution de la charge : dans le cas d'une défaillance, le réseau continuera son fonctionnement normal s'il existe une liaison auparavant inutilisée pour remplacer la liaison défaillante ou si la charge jusque là véhiculée par la liaison défaillante peut être répartie sur d'autres liaisons du réseau sans dépasser un certain seuil.

Cas 1 : Le réseau ne supporte aucune surcharge

D'un routeur d'accès à un point de présence, supposons qu'on ait au moins 2 routes disjointes de liaison : une route primaire et une redondante. Notons *k :p* le degré de protection (*k* liens de redondance inactifs pour *p* liaisons actives). Si on note *f* la probabilité de défaillance d'une liaison, dans ce cas-ci, le système a un fonctionnement anormal si au moins *k+1* liaisons sont défaillantes (on considère des liaisons directes parallèles). La probabilité associée est donc :

$$V = \sum_{i=k+1}^{p+k} C^i_{p+k} f^i (1-f)^{p+k-i} = 1 - \sum_{i=0}^{k} C^i_{p+k} f^i (1-f)^{p+k-i} \text{ où } C^i_n = \frac{n!}{i!(n-i)!}$$

Le Tableau 4.3 montre quelques valeurs de *V* en fonction de *k* et *p*. À titre de comparaison, le taux de défaillance pour la voix est de 10^{-5}. L'examen du tableau montre que, pour une liaison directe du routeur d'accès au point de présence, une liaison redondante suffit largement.

Tableau 4.3 Probabilité de défaillance

k	p	V
1	3	8.66×10^{-10}
1	8	5.2×10^{-9}
1	16	1.97×10^{-8}
1	30	6.71×10^{-8}

Cas 2 : Le réseau supporte une redistribution de la charge

Dans ce cas, la défaillance de *k*+1 liaisons n'est pas considérée comme une panne du réseau mais entraîne une surcharge d'un facteur de *1/p-1* sur les liaisons restantes, si on suppose que la surcharge est uniformément redistribuée. Si on tolère par exemple un seuil de surcharge de *2/p-2*, le réseau tolérera la défaillance de *k+2* liaisons. Dans le cas où la surcharge tolérée par liaison est de *1/p-1*, le fonctionnement anormal du réseau correspondrait à la défaillance de *k+2* liaisons. On a donc une probabilité de défaillance :

$$V' = \sum_{i=k+2}^{p+k} C^{\,i}{}_{p+k} f^{\,i} (1-f)^{p+k-i} = 1 - \sum_{i=0}^{k+1} C^{\,i}{}_{p+k} f^{\,i} (1-f)^{p+k-i}$$

Et il est évident que $V' < V$.

4.5 Simulation

Une analyse plus générale de la topologie en arbre ou des autres topologies considérées devient très rapidement complexe. Par conséquent, nous ferons plutôt une étude comparative des performances de fiabilité des différentes topologies afin de déterminer une « meilleure » topologie.

Le logiciel de simulation utilisé est *Comnet III* [22]. Trois topologies ont été étudiées :

- l'arbre avec liens redondants parallèles ;
- l'anneau avec un accès au point de présence ;
- la topologie maillée.

Dans ce contexte, un trafic distribué selon une loi exponentielle est utilisé pour générer une utilisation des canaux de l'ordre de 15%. Le trafic exponentiel a été choisi par souci de simplicité et aussi en se basant sur l'hypothèse que la fiabilité d'une topologie donnée est indépendante de la nature du trafic. Le réseau métropolitain (voir Figure 4.1) est constitué de 3 routeurs d'accès. Le trafic total dans le réseau reste le même mais, dépendamment de la topologie, le trafic à travers chaque routeur variera. Les figures 4.7 à 4.9 présentent les modèles utilisés pour la simulation avec Comnet, tandis que la Figure 4.10 présente la structure interne du sous-réseau utilisé dans les figures 4.8 et 4.9.

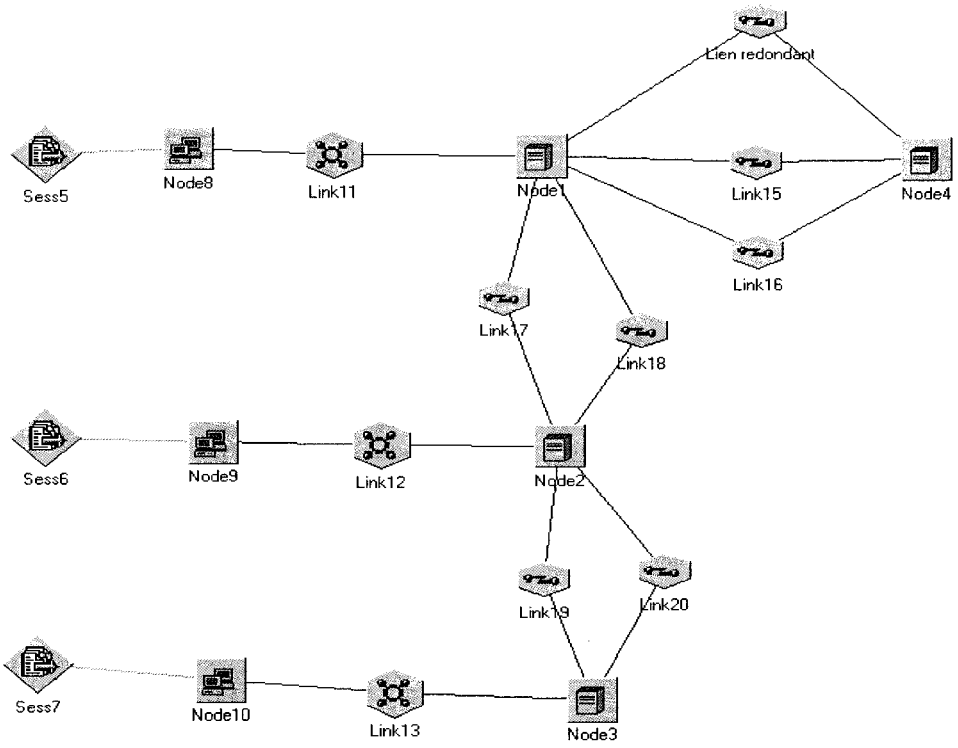


Figure 4.7 Modèle de topologie en arbre avec liens redondants parallèles

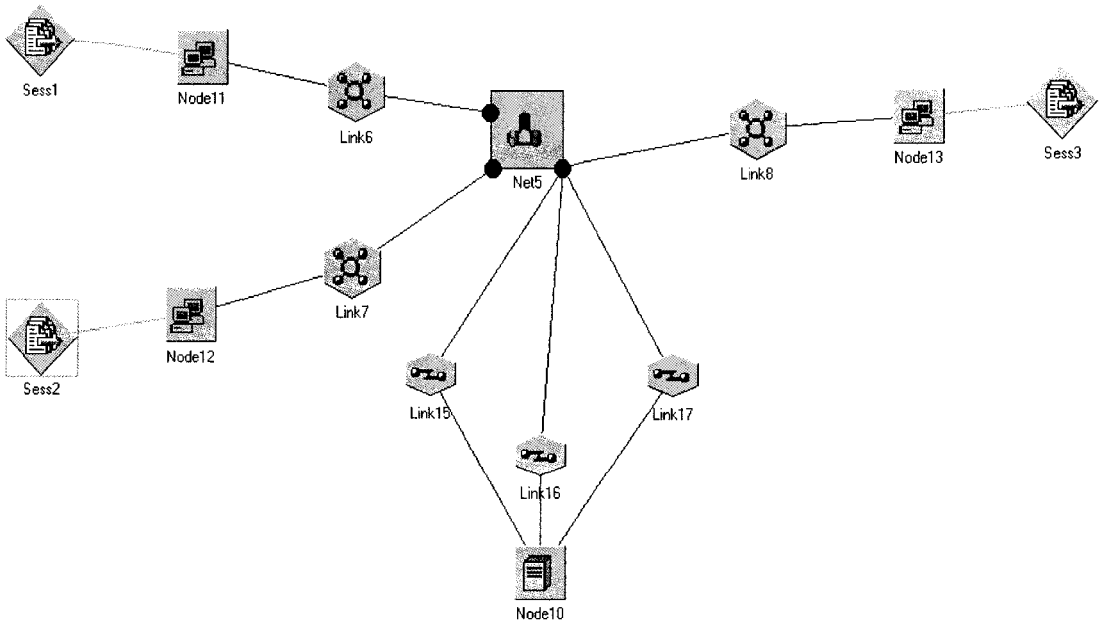


Figure 4.8 Modèle de topologie en anneau avec un accès au point de présence

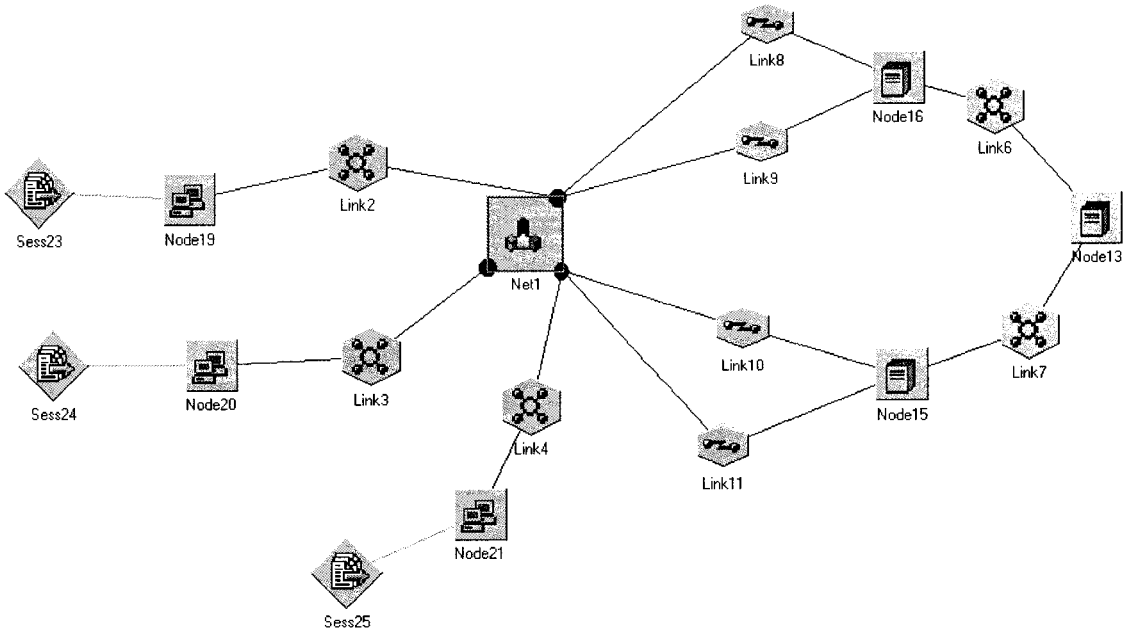


Figure 4.9 Modèle de topologie partiellement maillée

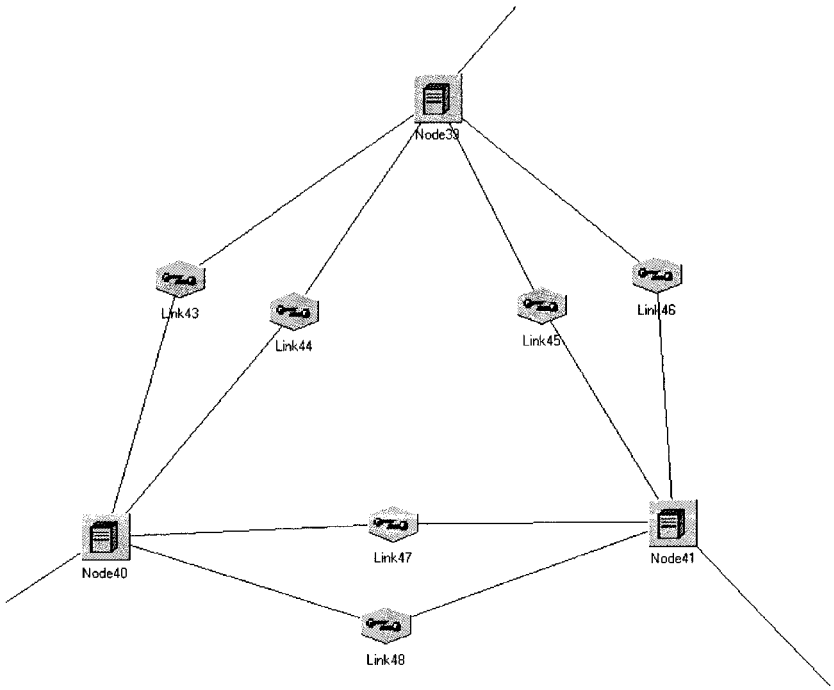


Figure 4.10 Détails du sous-réseau des figures 4.8 et 4.9

L'une des difficultés de la simulation est que les événements d'intérêt ont une très faible probabilité d'occurence. Il faudrait soit exécuter de très longues simulations ou augmenter la probabilité d'occurrence des événements. Comme nous faisons une étude comparative et ne voulons pas établir des valeurs absolues pour les performances de fiabilité des différentes topologies, nous avons choisi d'augmenter la probabilité d'occurrence des événements, l'idée étant que les performances relatives des diverses topologies conserveront alors les mêmes relations d'ordre. Le temps entre deux défaillances suit une loi exponentielle d'une moyenne de 2 minutes et le temps de réparation est aussi exponentiellement distribué avec une moyenne de 0.01 minute. Chaque lien entre routeurs d'accès ou entre un routeur d'accès et un point de présence est représenté par deux canaux indépendants pour plus de généralité. La liaison redondante, quant à elle, est représentée par un seul canal dans les topologies en anneau et en arbre. Le Tableau 4.4 résume le nombre total de liaisons par topologie. n désigne le nombre de routeurs d'accès, p le nombre total de canaux par liaison et k_T le nombre total de canaux redondants, l'idée étant que, pour passer d'une topologie à une autre, on ajoute soit un lien primaire, soit une liaison redondante supplémentaire. Notons que la topologie en arbre a une connexité de 2 au sens des arcs, mais une connexité de 1 au sens des nœuds. La topologie en anneau et la topologie maillée ont une connexité de 2 au sens des arcs et une connexité de 1 au sens des nœuds dans le sous-réseau reliant les RNC. Cependant, si on considère le réseau en anneau dans son ensemble, sa connexité au sens des nœuds tombe à 1.

Tableau 4.4 Nombre de canaux par topologie

Topologie	Arbre	Anneau	Partiellement maillée
Nombre de canaux	7	9	10
Généralisation	$np+k_T$	$(n+1)p+k_T$	$(n+1)p+k_T+1$

À cause de la faiblesse d'occurrence des événements considérés, nous avons simulé le modèle pour des durées croissantes de 6000, 10000, 18000 et 25000 secondes. Les résultats sont obtenus en faisant une moyenne des quatre expérimentations, ce qui

garantit une plus grande confiance dans les résultats obtenus. Les résultats sont présentés au Tableau 4.5

Tableau 4.5 Probabilités de blocage

Durée (en secondes)	Arbre	Anneau	Partiellement maillée
6000	$2.37 * 10^{-4}$	$8.52 * 10^{-4}$	$4.31 * 10^{-5}$
10000	$2.78 * 10^{-4}$	$5.69 * 10^{-4}$	$4.53 * 10^{-5}$
18000	$2.52 * 10^{-4}$	$3.74 * 10^{-4}$	$7.55 * 10^{-5}$
25000	$2.74 * 10^{-4}$	$3.15 * 10^{-4}$	$1.65 * 10^{-4}$
Moyenne	$2.60 * 10^{-4}$	$5.28 * 10^{-4}$	$8.23 * 10^{-5}$

Généralement, la probabilité de blocage pour la topologie en anneau est plus élevée que pour celle en arbre. La probabilité de blocage correspond dans ce cas au nombre de tentatives d'établissement de connexion qui ont échoué à cause de la défaillance d'une liaison. Dans la topologie en anneau, la conclusion est que la connexité plus élevée dans le réseau métropolitain (réseau inter-RNC) n'est pas très utile si la fiabilité des liens reliant les RNC (routeurs d'accès) au GGSN (point de présence) est faible. La topologie partiellement maillée présente une fiabilité meilleure à celle de l'arbre. Cependant, elle utilise plus de liaisons, ce qui a un impact sur le coût du réseau.

Le Tableau 4.6 donne les probabilités de blocage théoriques de la section 4.4 recalculées en fonction de la probabilité de défaillance de liaison réajustée pour la simulation.

Tableau 4.6 Probabilités de défaillance théoriques réajustées

Nombre total de liaisons	Nombre de liaisons redondantes	Topologie	Probabilité de défaillance
7	1	arbre	$5.1632 * 10^{-4}$
9	1	anneau	$8.7923 * 10^{-4}$
10	2	maillée	$5.9103 * 10^{-4}$

On note que, comme dans la simulation, la topologie en anneau a une probabilité plus élevée de défaillance. Par contre, les probabilités par simulation sont plus faibles car, contrairement à l'hypothèse de l'approche analytique, la défaillance de k liaisons n'entraîne pas forcément un blocage (toutes les liaisons ne sont pas forcément occupées au moment de la défaillance).

Le Tableau 4.7 compare les probabilités de déconnexion obtenues par simulation pour les trois topologies. La topologie en anneau est plus fiable que la topologie en arbre et ceci, en termes de sessions déconnectées à la suite d'une défaillance.

Tableau 4.7 Probabilités de déconnexion

Durée (en secondes)	Arbre	Anneau	Partiellement maillée
6000	$1.13 * 10^{-3}$	$9.925 * 10^{-4}$	$5.72 * 10^{-4}$
10000	$1.26 * 10^{-3}$	$9.83 * 10^{-4}$	$5.56 * 10^{-4}$
18000	$1.22 * 10^{-3}$	$9.24 * 10^{-4}$	$5.82 * 10^{-4}$
25000	$1.25 * 10^{-3}$	$9.18 * 10^{-4}$	$6.15 * 10^{-4}$
Moyenne	$1.21 * 10^{-3}$	$9.42 * 10^{-4}$	$5.81 * 10^{-4}$

Si nous reprenons la simulation précédente en essayant de re-router une seule fois les sessions bloquées, les probabilités de blocage avec possibilité de reroutage présentées au Tableau 4.8 sont approximativement égales aux probabilités de blocage. En effet, le temps de défaillance d'une liaison est d'un ordre de grandeur supérieur au temps nécessaire pour re-router une session. L'option de pouvoir re-router une session bloquée n'a donc pas un impact significatif sur les probabilités de blocage.

Tableau 4.8 Probabilités de blocage avec possibilité de reroutage

Arbre	Anneau	Partiellement maillée
$2.61 * 10^{-4}$	$5.28 * 10^{-4}$	$8.24 * 10^{-5}$

Le Tableau 4.9 présente les probabilités globales obtenues en additionnant les probabilités de déconnexion et de blocage.

Tableau 4.9 Probabilités globales de blocage

Durée (en secondes)	Arbre	Anneau	Partiellement maillée
6000	$1.367 * 10^{-3}$	$1.8445 * 10^{-3}$	$6.151 * 10^{-4}$
10000	$1.538 * 10^{-3}$	$1.5520 * 10^{-3}$	$6.0130 * 10^{-4}$
18000	$1.472 * 10^{-3}$	$1.2980 * 10^{-3}$	$6.5750 * 10^{-4}$
25000	$1.524 * 10^{-3}$	$1.2330 * 10^{-3}$	$7.8000 * 10^{-4}$
Moyenne	$1.4753 * 10^{-3}$	$1.4819 * 10^{-3}$	$6.6348 * 10^{-4}$

La topologie en arbre présente des résultats similaires à celle de la topologie en anneau en terme de fiabilité, mais est plus économique en terme du nombre de liaisons

utilisées. La fiabilité de la topologie partiellement maillée peut être améliorée au prix de l'ajout de liaisons supplémentaires.

Comme dans le cas de la fiabilité, le degré de performance d'un réseau a une influence très forte sur les coûts. L'opérateur cherche donc à trouver un compromis entre le coût du réseau et les performances de ce dernier. Ce compromis se fonde sur le comportement des usagers qui donne une bonne idée des rapports prix/performance acceptables. Le chapitre suivant se penchera donc sur les problèmes de coût et de tarification dans les réseaux de troisième génération.

CHAPITRE 5

CARACTÉRISATION DE LA GIGUE ET TARIFICATION

Au chapitre 3, l'étude de performance a prouvé que la liaison radio est souvent le goulot d'étranglement pour le réseau UMTS. Dans ce cas, un usager qui veut se connecter, envoie une requête de connexion qui est transférée au RNC. Le RNC (en collaboration avec le nœud B, si nécessaire) applique un algorithme d'admission de connexion pour vérifier la disponibilité des ressources. Quand les ressources sont disponibles, la connexion est établie, sinon elle est refusée. Dans le cas où des canaux de transport dédiés (DCH) sont utilisés, la requête de l'usager correspond à une demande de réservation de bande passante. Dans [55] par exemple, l'augmentation du bruit due à un usager supplémentaire sur la liaison montante est $(1+i)L_j$, où L_j est le rapport du signal reçu de l'usager j au signal total reçu sur la bande de fréquence, incluant le bruit thermique au niveau de la station de base et i est le rapport des interférences entre les usagers d'une même cellule aux interférences provenant des autres cellules. Dans ce cas, on a un contrôle d'admission qui repose sur la gestion des ressources de puissance. Dans le cas de connexions à débit variable, un tel schéma réserve de la puissance pour transmettre au débit maximal, ce qui engendre un gaspillage des ressources. Une autre alternative est de calculer un débit effectif, de réserver la puissance en se basant sur le débit effectif et de fournir ainsi à l'usager des garanties probabilistes. De manière similaire, dans le cas où on utilise des canaux de transport partagés (DSCH) ou bien une liaison sur-souscrite, il faudrait établir une tarification équitable qui reflète l'utilisation que chaque usager fait du canal. Cependant, le contrôle d'admission ou la tarification basée sur le débit effectif tels que présentés dans la littérature [29] présentent un certain nombre de lacunes. Dans ce chapitre, nous proposons un schéma amélioré qui pourrait s'adapter aux réseaux de troisième génération [57].

5.1 Définitions et concepts de base

Avec le développement de l'ATM, un grand nombre de travaux se sont consacrés à l'étude de la probabilité asymptotique de débordement du tampon qui constitue la base de la théorie du débit effectif. Dans la suite, nous résumons les résultats obtenus par Courcoubetis *et al.* [31], Courcoubetis et Weber [32] et Likhanov et Mazumdar [76].

5.1.1 Taux asymptotique de perte de cellules

Cette section traite plus spécifiquement des résultats obtenus par Likhanov et Mazumdar [76] et complétés par Courcoubetis et Weber [32]. Les détails qui y sont omis sont regroupés à l'Annexe II. Considérons N sources indépendantes, identiques, stationnaires et ergodiques, chacune générant des paquets au taux $\lambda_{n,j}$ où n se réfère au temps et j à la source. On assume que le temps est discrétisé et que les taux d'arrivées des paquets possèdent un certain nombre de propriétés de régularité [76]. Le nombre total de paquets générés par la source j durant l'intervalle de temps *[0,t)* est $X_{t,j} = \sum_{n=0}^{t-1} \lambda_{n,j}$ et le nombre total de paquets générés par toutes les sources est $X_t^{(N)} = \sum_{k=1}^{N} X_{t,k}$.

Soit $M_{t,1}(s)$ ou tout simplement $M_t(s)$ la fonction génératrice de moment de $X_{t,1}$, i.e. $M_t(s) = M_{t,1}(s) = E[e^{sX_{t,1}}]$. Supposons que les sources accèdent à un tampon de taille NB avec un débit de sortie de NC. La charge stationnaire de travail $W^{(N)}(C)$ est donnée par : $W^{(N)}(C) = \sup_{t\in\{1,\ldots\}} \left(X_{-t}^{(N)} - NCt\right)$ où $X_{-t}^{(N)}$ désigne le nombre total de cellules qui arrivent dans l'intervalle *(-t,0]*.

Alors, si $N \to \infty$,

$$P\left\{W^{(N)}(C) > NB\right\} = \frac{1}{\sqrt{2\pi N\sigma^2 s_0^2}} e^{-NI_{t_0,s_0}(C,B)}\left(1 + O\left(\frac{1}{N}\right)\right) \tag{5.1}$$

où s_0, t_0 sont les points critiques (points qui maximisent ou minimisent) de :

$$-I_{t_0,s_0}(C,B) = \ln(M_{t_0}(s_0)) - (Ct_0 + B)s_0 = \sup_t \inf_s [\ln(M_t(s)) - (Ct + B)s] \quad (5.2)$$

et

$$\sigma^2 = \frac{\partial^2}{\partial s^2} \ln\left(E[e^{sX_{t_0,1}}]\right) = \frac{M''_{t_0}(s_0)}{M(s_0)} - (Ct_0 + B)^2 \quad (5.3)$$

Dans la définition précédente, nous supposons par abus de notation que $0 \leq s_0, t_0 \leq +\infty$, i.e. que s_0 et t_0 ne sont pas forcément des valeurs finies.

Dans [94], le terme $\sigma^2 s_0^2$ est approximé par $2I_{t_0,s_0}(C,B)$. Ainsi, l'équation (5.1) devient :

$$P\{W^{(N)}(C) > NB\} = \frac{1}{\sqrt{4\pi N I_{t_0,s_0}(C,B)}} e^{-NI_{t_0,s_0}(C,B)} \left(1 + O\left(\frac{1}{N}\right)\right) \quad (5.4)$$

5.1.2 Contrôle d'admission et tarification

Le concept de débit effectif a été introduit par Hui [59] et Guerin *et al.* [48]. Combiné avec les travaux de [32,76,94], il a été utilisé par Kelly, Courcoubetis, Siris dans leur recherche d'une région linéaire d'acceptation pour certaines ressources [28,29,31,67] qui utilisent des tampons et dans la conception de schémas de tarification. Nous rappelons les résultats obtenus par Courcoubetis *et al.* [31].

Plusieurs types de sources sans priorité

Supposons que le processus d'arrivée à une liaison haut débit soit la superposition de J types de sources indépendantes identiquement distribuées. Soit $N_i = Nn_i$, $i=1,\ldots, J$ le nombre de sources de type i, et soit $n=(n_1, \ldots, n_i, \ldots, n_J)$ (les n_i ne sont pas nécessairement entiers). La liaison est desservie par un tampon NB au taux NC. Le paramètre N est un facteur d'échelle qui reflète la taille du système. Il est typiquement égal au plus petit nombre de sources de même type multiplexés, i.e. $N=min(N_i)$.

En utilisant les notations de la section précédente, nous définissons $X_{t,j}$ comme le trafic total généré par une source de type j durant l'intervalle $[0,t)$ et $M_{t,j}(s)$ comme la

fonction génératrice des moments de $X_{t,j}$, i.e. $M_{t,j}(s) = E[e^{sX_{t,j}}]$, où E est l'espérance. On rappelle que $X_{t,j}$ a des accroissements stationnaires. Alors, le *débit effectif* d'une source de type j est défini comme suit :

$$\alpha_j(s,t) = \frac{1}{st} \ln E\left[e^{sX_{t,j}}\right] = \frac{1}{st} \ln(M_{t,j}(s))$$

où s, t sont des paramètres du système définis par le contexte de la source. Une interprétation plus détaillée des paramètres s et t est fournie à l'Annexe II.

Plus précisément, s et t sont définis par l'équation (5.2), qui peut se réécrire :

$$-I_{t_0,s_0}(C,B) = s_0 t_0 \sum_{j=1}^{J} n_j \alpha_j(s_0,t_0) - (Ct_0 + B)s_0 = \sup_t \inf_s \left[st \sum_{j=1}^{J} n_j \alpha_j(s,t) - (Ct+B)s \right] \quad (5.5)$$

Considérons que la contrainte de qualité de service sur la probabilité de débordement du tampon soit $P(débordement) \leq e^{-\gamma}$ et posons $\gamma = N\gamma_0$. Si un point $(N_1, \ldots, N_J) = (Nn_1, \ldots, Nn_J)$ satisfait :

$$\sum_{j=1}^{J} n_j \alpha_j(s_0,t_0) \leq C + \frac{1}{t_0}\left(B - \frac{\gamma_0}{s_0}\right) = C^* \quad (5.6)$$

où s_0 et t_0 sont définis par l'équation (5.5), alors la contrainte de qualité de service sur la probabilité de débordement du tampon soit $P(débordement) \leq e^{-\gamma}$ est satisfaite. Ainsi, l'équation (5.6) définit une région d'acceptation.

En prenant en compte les résultats de la section précédente (équation (5.4)), la région d'acceptation peut être améliorée comme suit :

$$\sum_{j=1}^{J} n_j \alpha_j(s_0,t_0) \leq C + \frac{1}{t_0}\left(B - \frac{\gamma_0'}{s_0}\right) = C_{B-R}^* \quad (5.7)$$

où $\gamma_0' = \gamma_0 - \dfrac{\frac{1}{2}\log(4\pi N\gamma_0)}{N + \frac{1}{2\gamma_0}} \approx I_{t_0,s_0}(C,B)$

Supposons que nous ayons un type j_1 de trafic qui est facturé au taux f_1. Si l'opérateur possède n_1 connexions j_1, alors, pour satisfaire la contrainte $P(débordement) \leq e^{-\gamma}$, il choisit un prix f_1 qui vérifie :

(O1): $\max f_1 n_1$ sous la contrainte $n_1\alpha_1(s_1,t_1) \le C + \frac{1}{t_1}(B - \frac{\gamma'}{s_1}) = K_1$ avec

$$\gamma' = \gamma - \frac{\frac{1}{2}\log(4\pi\gamma)}{1+\frac{1}{2\gamma}}$$

La solution est $f_1 = \lambda_1\alpha_1$ où λ_1 est le coefficient de Lagrange associé à la contrainte.

Plus généralement, supposons que nous ayons deux types j_1 et j_2 de trafic. La contrainte d'admission de connexion peut être exprimée au moins localement comme une contrainte linéaire $n_1\alpha_1 + n_2\alpha_2 \le C$, où n_1 et n_2 sont les nombres de connexions de type j_1 et j_2, tandis que α_1, α_2 et C dépendent des ressources réseau et des caractéristiques statistiques des connexions de type j_1 et j_2. Les coefficients α_1, α_2 constituent les débits effectifs. Dans un équilibre compétitif, le bien-être social $u(n_1,n_2)$, i.e. la somme de tous les profits des différents acteurs, est maximisé sujet à la contrainte linéaire exprimée précédemment. En formulant le lagrangien du problème d'optimisation, i.e., maximiser $u(n_1,n_2) - \lambda\alpha_1 n_1 - \lambda\alpha_2 n_2$ par rapport à n_1 et n_2, on peut constater que l'optimum est atteint quand l'opérateur annonce les prix $\lambda\alpha_1$ et $\lambda\alpha_2$ et n_1 et n_2 sont choisis de manière décentralisée. Par conséquent, le tarif proposé devrait être proportionnel au débit effectif et de ce fait, aux ressources consommées.

Plusieurs types de sources avec priorité

Souvent, dans les réseaux à large bande, différents niveaux de qualité de service sont appliqués aux différentes classes de trafic. Supposons par exemple, que les classes de trafic soient partitionnées en deux sous-ensembles, J_1 et J_2. Le service est PAPS (Premier Arrivé, Premier Servi), excepté que les sources de trafic de type J_1 sont prioritaires par rapport aux sources de trafic de type J_2. La Figure 5.1 en est une illustration.

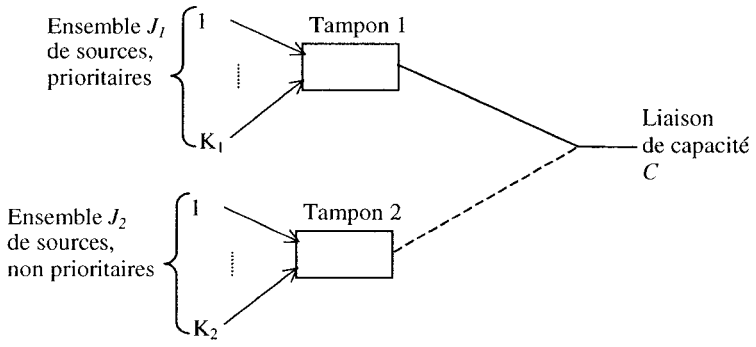


Figure 5.1 Système de files avec priorité

Dans la Figure 5.1, le tampon 1 est desservi au taux C, tant qu'il n'est pas vide ou tant qu'une source de type J_1 génère des paquets. Seule la bande passante résiduelle (qui varie dans le temps) est mise à la disposition du tampon 2. Dans [38], on montre que le système de la Figure 5.1 est équivalent à celui de la Figure 5.2.

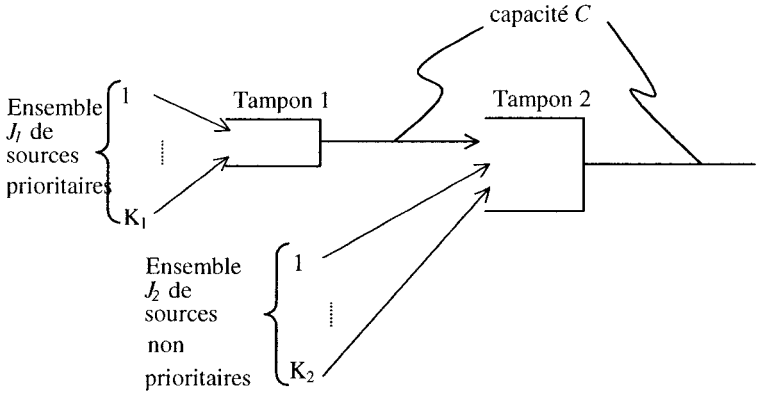


Figure 5.2 Système équivalent de files avec priorité

Si nous supposons que pour $i \in J_1$, il y a une garantie de qualité de service de la forme :

$$P(\textit{délai} > B_1 / C) \leq e^{-\gamma_1},$$

sur le délai et pour toutes les sources, une garantie sur le taux de perte de cellules :

$$P(\textit{débordement du tampon}) \leq e^{-\gamma_2},$$

et en remarquant que $P(délai > B_1 / C) = P(tampon\ B_1\ déborde)$, nous pouvons appliquer plusieurs fois les équations (5.6) ou (5.7) et obtenir deux contraintes de la forme:

$$\sum_{j \in J_1} n_j \alpha_j(s_1, t_1) \leq K_1 \tag{5.8}$$

$$\sum_{j \in J_1 \cup J_2} n_j \alpha_j(s_2, t_2) \leq K_2 \tag{5.9}$$

où K_1 et K_2 sont définis par l'équation (5.6) ou (5.7) et s_i, t_i sont déterminés à partir de l'équation (5.5).

Implémentation de la tarification selon le débit effectif

La tarification selon le débit effectif, bien que simple et efficace, n'est pas facile à implémenter d'une manière décentralisée. Dans [28,29], les auteurs proposent une implémentation simple dans laquelle le prix par unité de temps est une fonction linéaire de la forme :

$$f(X) = a_0 + a_1 g_1(X) + ... + a_L g_L(X) = a_0 + a^{\perp} g(X)$$

où $g_1(X), \ldots, g_L(X)$ sont des mesures déduites à partir de l'observation de $X = (X_1, ..., X_T)$ ou de fonctions de ces mesures, l'indice de X se référant au temps discrétisé *1,...,T*, et $^{\perp}$ est le symbole de la transposée. Dans ce cas, X et $a_0, \ldots, a_L$ dépendent de *j* et de ce fait, des paramètres de contrôle (policing parameters) tels le débit moyen des sources de type *j*. Cependant pour alléger la notation, nous n'écrirons pas explicitement cette dépendance.

Supposons que X possède des accroissements stationnaires, que $X \in \chi(h)$, pour un ensemble $\chi(h)$ paramétré par un vecteur h et que les mesures vérifient $E[g(X)] = m$ pour un tuple m donné. Soit $\overline{Q}_{g(X)}(m,h)$ la borne supérieure suivante :

$$\overline{Q}_{g(X)}(m,h) = \sup_{X: E(g(X))=m; X \in \chi(h)} \left\{ E[e^{sX[0,t]}] \right\} \tag{5.10}$$

où $g(X) = \frac{1}{T} \sum_{i=1}^{T} X_i$ et *X[0,t]* est la charge totale générée durant l'intervalle *[0,t]*. Soit

$$\overline{\alpha}(m,h) = \frac{1}{st} \ln \overline{Q}_{g(X)}(m,h) \qquad (5.11).$$

$\overline{\alpha}(m,h)$ est concave en m.

La Figure 5.3 montre comment le débit effectif peut être borné par une fonction linéaire du paramètre mesuré, $E(g(X))$. m et h peuvent être interprétés respectivement comme le débit moyen et le débit de crête de source.

Dans la suite du chapitre, nous nous limitons au cas où $L=1$ et $g_1(X) = \frac{1}{T}\sum_{i=1}^{T} X_i$ [28,29]. Dans ce cas, la charge totale est juste une fonction du nombre total de paquets véhiculés et aussi à travers a_0, de la durée de la connexion. Ainsi, l'opérateur de réseau publie un ensemble de tarifs possibles ($a_0(m_r)$, $a_1(m_r)$) qui définissent les tangentes à la borne du débit effectif $\overline{\alpha}(m_r,h)$ où m_r est le débit réel de l'usager tel que mesuré par le réseau. L'usager choisit, pour minimiser ses coûts, le tarif ($a_0(m)$, $a_1(m)$) correspondant au débit moyen estimé m de sa connexion. Cette paire ($a_0(m)$, $a_1(m)$) correspond aux paramètres de la tangente de $\overline{\alpha}$ au point $m = \frac{X_t}{t}$. Le taux horaire de l'usager est alors $\overline{\alpha}$.

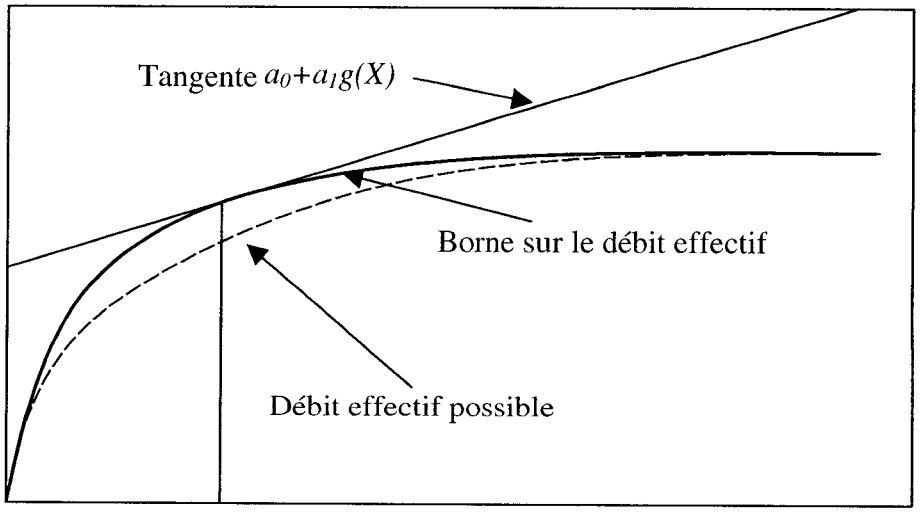


Figure 5.3 Débit effectif et ses bornes en fonction de *m*

Quelques propriétés des schémas de tarification

Le rôle de la tarification consiste non seulement à générer des revenus, mais également à introduire une rétroaction et un contrôle entre l'opérateur réseau et les usagers en incitant ces derniers à réagir aux prix et à essayer de minimiser leurs coûts. Par exemple, une politique appropriée de tarification devrait d'une part, jouer le rôle d'un mécanisme de contrôle de congestion en promouvant une utilisation efficace des ressources et d'autre part, encourager les usagers à établir une correspondance convenable entre leurs besoins et les classes de service disponibles. Plus généralement, un schéma de tarification devrait :

- (P1) *Être simple* : la charge due à l'implémentation de la tarification et les mesures nécessaires à son application ne devraient pas surcharger le réseau. Les mesures doivent aussi être assez simples pour que les usagers puissent comprendre le mécanisme de tarification.
- (P2) *Être équitable* : le prix payé par chaque client devrait être le reflet de son utilisation relative des ressources du réseau.
- (P3) *Fournir les bons incitatifs :* une fois les tarifs fixés, chaque usager qui essaye de minimiser ses coûts, devrait choisir un contrat et utiliser les ressources du réseau d'une manière qui maximise la performance globale du réseau.
- (P4) *Permettre de recouvrer les coûts et développer le réseau :* cet aspect dépend plus des décisions de gestion. Cependant, il peut être pris en compte en ajoutant des contraintes supplémentaires au problème.

Par exemple, une tarification forfaitaire fournit de mauvais incitatifs aux usagers et les amène à abuser des ressources du réseau. Dans la section suivante, nous analysons la tarification basée sur le débit effectif par rapport aux propriétés énumérées.

Analyse de la tarification basée sur le débit effectif

La tarification selon le débit effectif a été développée pour mesurer l'utilisation des ressources du réseau par une connexion au niveau d'un commutateur. Cependant, elle peut être aussi utilisée pour une route à travers le réseau parce que, souvent dans le réseau, il existe un point d'étranglement, qui génère la contrainte active. Dans le cas d'un réseau UMTS, ce point est souvent la liaison radio. En outre, le débit effectif de la liaison limitante n'est souvent pas affecté par le multiplexage qui a pu avoir lieu dans les tampons en amont. Finalement, les paramètres s et t peuvent être utilisés pour ajuster les mesures de débits effectifs et prendre en compte la performance globale du réseau [29].

La tarification basée sur le débit effectif a un certain nombre de propriétés intrinsèques d'équité et fournit les incitatifs adéquats (propriétés P2 et P3) aux usagers. Les propriétés d'équité découlent du fait que la tarification pour une connexion donnée est directement proportionnelle à la borne $\overline{\alpha}$ sur le débit effectif. Cependant, le calcul de $\overline{\alpha}$ est généralement très difficile. Par conséquent, en pratique, on utilise une approximation $\overline{\alpha}'$ de $\overline{\alpha}$. Dans [119], Siris présente trois approximations possibles de $\overline{\alpha}$: une borne on-off, qui dépend seulement du débit de crête h et du débit moyen m, une borne simple mais plus stricte dérivée de l'algorithme du seau percé et une approximation basée sur des fonctions en T inversé. Ces bornes seront présentées plus en détail à la section 5.2. Heureusement en pratique, l'utilisation de $\overline{\alpha}'$ à la place de $\overline{\alpha}$ préserve les propriétés d'équité.

Le schéma de tarification fournit les incitatifs adéquats aux usagers à travers un cycle. L'opérateur réseau publie des tarifs qui ont été calculés à partir du point d'opération actuel du réseau sur la base des paramètres s et t. Ces tarifs incitent les usagers à changer leurs contrats dans le but de minimiser les coûts anticipés. Avec les nouveaux contrats, le point d'opération du réseau change, puisque l'opérateur doit garantir les performances figurant dans les nouveaux contrats. Il recalcule donc de nouveaux tarifs pour le nouveau point d'opération et cette interaction continue jusqu'à l'équilibre. Dans [29], on montre que cet équilibre existe même lorsque $\overline{\alpha}$ est remplacé par $\overline{\alpha}'$ et qu'il constitue un optimum pour le bien-être social.

Finalement, la tarification basée sur le débit effectif est simple (propriété P1) car elle requiert seulement la surveillance du débit moyen de la connexion.

Néanmoins, dans le cas d'une connexion dont le débit moyen mesuré ne correspond pas au débit déclaré dans le contrat, celle-ci est doublement pénalisée : elle est facturée à un tarif plus élevé que le tarif fixé dans le contrat ($E(f_{m,h}(X)) \geq \overline{\alpha}(E(g(X)),h)$) et en plus, l'algorithme de conformité rejette un certain nombre de paquets de la connexion (violation de la propriété P2). Rappelons que le tarif plus élevé imposé à la connexion est, en outre, calculé sur la base d'une borne supérieure $\overline{\alpha}(E(g(X)),h)$ du débit effectif.

De plus, pour des connexions avec des contraintes sur la gigue, le contrôle d'admission qui sous-tend la tarification basée sur le débit effectif oblige à traiter la gigue comme un simple délai. En d'autres termes, il (le contrôle d'admission) ne fournit pas une manière claire et précise de prendre en compte des garanties sur la variation de délai.

Par rapport à toutes ces pénalités imposées à l'usager qui ne respecte pas rigoureusement son contrat, nous proposons une fonction de tarification qui pénalise moins les usagers qui demeurent dans un intervalle d'incertitude raisonnable autour du débit moyen déclaré dans leur contrat. Entre autres, nous essayerons de réduire le surcoût relié au non-respect du débit déclaré, puisque l'usager est déjà sanctionné par la perte éventuelle de ses paquets, tout en gardant les bonnes propriétés du schéma de tarification initial.

Ensuite, nous introduirons une contrainte supplémentaire pour prendre en compte des garanties sur la variation de délai, ce qui permettra d'étendre le schéma de tarification introduit à la section 5.1.2 et qui tenait compte seulement des garanties sur le délai et la perte des paquets (équations (5.8) et (5.9)).

5.2 Fonction de tarification proposée

Supposons que X ait des accroissements stationnaires, que $X \in \chi(h)$, pour un ensemble donné $\chi(h)$ paramétré par un certain vecteur h et que les mesures vérifient $E[g(X)] = m$. On rappelle que $\overline{Q}_{g(X)}(m,h)$ est la borne supérieure suivante :

$$\overline{Q}_{g(X)}(m,h) = \sup_{X:E(g(X))=m;X\in\chi(h)} \left\{E[e^{sX[0,t]}]\right\} \text{ avec } g(X) = \frac{1}{T}\sum_{i=1}^{T} X_i$$

$$\text{et } \overline{\alpha}(m,h) = \frac{1}{st}\ln \overline{Q}_{g(X)}(m,h) \text{ (s, t sont supposés fixés).}$$

On rappelle aussi que nous considérons d'abord une fonction de tarification de la forme:

$$f(X) = a_0 + a_1 g(X) \tag{5.12}$$

avec $a_0 = \overline{\alpha}(m,h) - m\frac{\partial}{\partial m}\overline{\alpha}(m,h)$, $a_1 = \frac{\partial}{\partial m}\overline{\alpha}(m,h)$. m et h peuvent respectivement être interprétés comme le débit moyen et le débit de crête de la source. On suppose que le débit de crête h est fixé par le type de la connexion comme spécifié dans les spécifications du groupe 3GPP [2] et l'usager choisit son débit moyen m. Par souci de simplicité, on note $\overline{Q}_{g(X)}(m) = \overline{Q}_{g(X)}(m,h)$ et $\overline{\alpha}(\overline{Q}_{g(X)}(m)) = \overline{\alpha}(m,h) = \frac{1}{st}\ln \overline{Q}_{g(X)}(m,h)$.

Le but est alors de proposer une fonction de tarification qui pénalise moins les usagers qui demeurent dans un intervalle d'incertitude raisonnable autour du débit moyen déclaré dans leur contrat tout en gardant les bonnes propriétés du schéma de tarification initial. Par conséquent, nous essayerons d'utiliser la concavité de la fonction logarithme pour développer une meilleure fonction de tarification qui sera moins sensible aux imprécisions raisonnables des usagers sans nécessiter des mesures supplémentaires sur l'état de la connexion.

Considérons

$$h_{\overline{Q},m}(m_r) = \overline{\alpha}(\overline{Q}_{g(X)}(m)) + \lambda_{1,\overline{Q}}(\overline{Q}_{g(X)}(m_r) - \overline{Q}_{g(X)}(m)) + \lambda_{2,\overline{Q}}(\overline{Q}_{g(X)}(m_r) - \overline{Q}_{g(X)}(m))^2 + \lambda_{3,\overline{Q}}(\overline{Q}_{g(X)}(m_r) - \overline{Q}_{g(X)}(m))^3 \quad (5.13)$$

avec

$$\lambda_{1,\overline{Q}} = \left.\frac{\partial\overline{\alpha}}{\partial\overline{Q}}\right|_{\overline{Q}_{g(X)}(m)} = \frac{1}{st\overline{Q}_{g(X)}(m)},$$

$$\lambda_{2,\overline{Q}} = \frac{1}{2!}\left.\frac{\partial^2\overline{\alpha}}{\partial\overline{Q}^2}\right|_{\overline{Q}_{g(X)}(m)} = -\frac{1}{2st(\overline{Q}_{g(X)}(m))^2},$$

$$\lambda_{3,\overline{Q}} = \frac{1}{3!}\left.\frac{\partial^3\overline{\alpha}}{\partial\overline{Q}^3}\right|_{\overline{Q}_{g(X)}(m)} = \frac{1}{3st(\overline{Q}_{g(X)}(m))^3}$$

où m est le débit moyen déclaré par l'usager au moment de l'établissement de la connexion, tandis que m_r est le débit moyen réel mesuré par le réseau. La Figure 5.4 montre comment notre fonction de tarification borne le débit effectif et la fonction linéaire de tarification du schéma initial introduit à la section 5.1.2 à l'intérieur d'un voisinage donné.

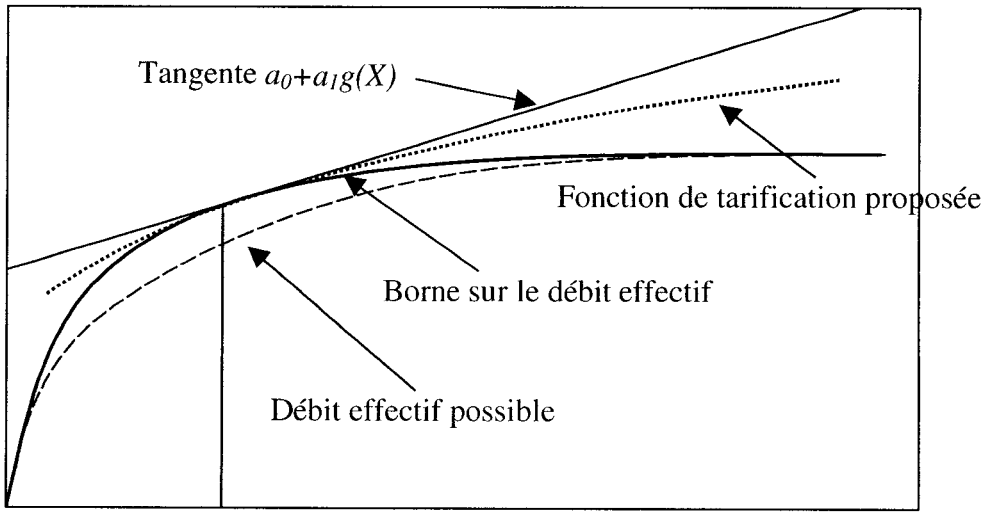


Figure 5.4 Fonction de tarification proposée

$h_{\overline{Q},m}$ est le développement de Taylor de $\overline{\alpha}$ au voisinage de $\overline{Q}_{g(X)}(m)$.

Notons que la fonction proposée est aussi « simple » en terme de mesures que la fonction initiale car elle ne dépend que de la mesure du débit moyen réel m_r de la connexion. En effet, si nous avons une expression pour $\overline{Q}_{g(X)}(m)$, nous pouvons calculer $h_{\overline{Q},m}$ directement à partir du débit moyen mesuré *g(X)* en utilisant une fonction polynomiale. La fonction polynomiale utilisée ici est certes de degré supérieur à celle de l'équation (5.12), mais reste une fonction seulement du débit *g(X).* Ainsi les coefficients de la fonction seront calculés au début de chaque cycle d'interaction opérateur - usager. En fait, on utilise un développement en série qui est une fonction de la borne sur le débit effectif avec le supremum $\overline{Q}_{g(X)}(m)$ de la fonction génératrice des moments comme paramètre libre au lieu du débit moyen *m*, parce que $\overline{Q}_{g(X)}(m)$ ne possède pas de forme fermée. Il est donc difficile d'en faire un développement en série. En pratique, plusieurs approximations de $\overline{Q}_{g(X)}(m)$ sont utilisées. Le fait que notre tarification dépend de $\overline{Q}_{g(X)}(m)$ permet de l'adapter aux différentes approximations. De plus, soulignons que l'équation (5.13) propose une fonction d'ordre 3. Cependant, cette fonction de tarification peut être étendue très facilement à des degrés plus élevés (impairs) tout en gardant les mêmes propriétés. Les autres propriétés de la fonction proposée sont :

Lemme 1 $h_{\overline{Q},m}(m_r)$ est supérieur ou égal à $\overline{\alpha}(\overline{Q}_{g(X)}(m_r))$, l'égalité étant atteinte lorsque $m_r = m$.

Preuve

Supposons que les conditions nécessaires de régularité sont réunies. Le développement en série de $\overline{\alpha}(\overline{Q}_{g(X)}(m_r)) = \overline{\alpha}(Y_r)$ au voisinage de $Y = \overline{Q}_{g(X)}(m)$ est

$$\overline{\alpha}(Y_r) = \overline{\alpha}(\overline{Q}_{g(X)}(m)) + \frac{1}{st\overline{Q}_{g(X)}(m)}(\overline{Q}_{g(X)}(m_r) - \overline{Q}_{g(X)}(m)) - \frac{1}{2st(\overline{Q}_{g(X)}(m))^2}(\overline{Q}_{g(X)}(m_r) - \overline{Q}_{g(X)}(m))^2$$

$$+ \frac{1}{3st(\overline{Q}_{g(X)}(m))^3}(\overline{Q}_{g(X)}(m_r) - \overline{Q}_{g(X)}(m))^3 - \frac{1}{4st(\overline{Q}_{g(X)}(m))^4}(\overline{Q}_{g(X)}(m^*) - \overline{Q}_{g(X)}(m))^4$$

où $\overline{Q}_{g(X)}(m) \le \overline{Q}_{g(X)}(m^*) \le \overline{Q}_{g(X)}(m_r)$

$$s>0,\ t>0 \Rightarrow \overline{\alpha}(Y_r) - h_{\overline{Q},m}(m_r) = -\frac{1}{4st(\overline{Q}_{g(X)}(m))^4}(\overline{Q}_{g(X)}(m^*) - \overline{Q}_{g(X)}(m))^4 \le 0, \forall m, \forall m_r$$

Ainsi, $h_{\overline{Q},m}(m_r) \ge \overline{\alpha}(\overline{Q}_{g(X)}(m_r))$, et l'égalité est atteinte si $m_r = m$. De ce fait, l'usager a un fort incitatif à déclarer son débit moyen réel dans le contrat.

Lemme 2 $h_{\overline{Q},m}(m_r)$ est croissant en m_r.

Preuve

Définissons $Y_r = \overline{Q}_{g(X)}(m_r)$ et $Y = \overline{Q}_{g(X)}(m)$.

$$\frac{\partial h_{\overline{Q},m}}{\partial Y_r} = \frac{1}{stY} - \frac{1}{stY^2}(Y_r - Y) + \frac{1}{stY^3}(Y_r - Y)^2 = \frac{1}{st}\left(\frac{Y_r^2 - 3YY_r + 3Y^2}{Y^3}\right) = \frac{1}{st}\left(\frac{(Y_r - \frac{3}{2}Y)^2 + \frac{3}{4}Y^2}{Y^3}\right)$$

Comme *s*, *t* et *Y* sont positifs, alors $\frac{\partial h_{\overline{Q},m}}{\partial Y_r} \ge 0$ et $h_{\overline{Q},m}(m_r)$ est croissant en $\overline{Q}_{g(X)}(m_r)$.

Pour montrer que $\overline{Q}_{g(X)}(m_r)$ est croissant en m_r, considérons deux réels m_1 et m_2 tels que $m_1 \le m_2$. Considérons aussi $X^S = \sup_{X \in \chi(h)} X$. Nous pouvons supposer que X^S existe (i.e. est fini), sinon le contrôle imposé par la contrainte $X \in \chi(h)$ est inefficace, mais la démonstration demeure similaire même si X^S est infini. Supposons que l'extrémum X^S est atteint (sinon, on peut toujours trouver une fonction X^a arbitrairement proche de X^S qui jouera un rôle similaire à celui de X^S dans cette preuve). Par définition de X^S et *g(X)*, nous avons $m_S = E(g(X)) = g(X^S) = \sup_{X \in \chi(h)} g(X)$. Par conséquent, pour $m_1, m_2 \in [0, m_S]$, nous avons $\overline{Q}_{g(X)}(m_1) = \sup_{X:E(g(X))=m_1; X \in \chi(h)} \left\{E[e^{sX[0,t]}]\right\}$ et $\overline{Q}_{g(X)}(m_2) = \sup_{X:E(g(X))=m_2; X \in \chi(h)} \left\{E[e^{sX[0,t]}]\right\}$. Pour toute fonction X_1 telle que $E(g(X_1)) = m_1; X_1 \in \chi(h)$, nous pouvons trouver une fonction X_2 telle que

$E(g(X_2)) = m_2; X_2 \in \chi(h)$. Cette fonction X_2 est obtenue en augmentant la fonction X_1 partout où $X_1<X^S$ tout en respectant la contrainte $X_2<X^S$ jusqu'à ce que $E(g(X_2))=m_2$. La fonction résultante est forcément un élément de $\chi(h)$ puisqu'elle correspond à une charge inférieure au maximum permis X^S. En outre, nous avons par construction :

$X_2 \geq X_1 \Rightarrow E[e^{sX_2[0,t]}] \geq E[e^{sX_1[0,t]}]$ (puisque l'espérance est prise dans le temps).

Par conséquent, en prenant le supremum des deux membres de l'inégalité, nous trouvons que $\overline{Q}_{g(X)}(m_1) \geq \overline{Q}_{g(X)}(m_2)$ et alors que $h_{\overline{Q},m}(m_r)$ est croissant en m_r.

Ce lemme signifie que $h_{\overline{Q},m}$ préserve une certaine équité puisque l'usager qui utilise plus de ressources du réseau paiera plus. La quantification de cette équité pourra faire l'objet de recherches futures. Cependant, comme nous l'avons déjà dit, en pratique, une approximation $\overline{\alpha}'$ de $\overline{\alpha}$ (et par conséquent une approximation $\overline{Q}'_{g(X)}(m)$ de $\overline{Q}_{g(X)}(m)$) est utilisée. Les approximations $\overline{Q}'_{g(X)}(m)$ de $\overline{Q}_{g(X)}(m)$ doivent donc être croissantes pour que $h_{\overline{Q}',m}(m_r)$ puisse toujours être garantie comme fonction croissante. Dans [119], Siris présente trois approximations simples $\overline{\alpha}'$ de $\overline{\alpha}$:

Borne on-off

La borne on-off [66] dépend seulement du débit de crête h de la connexion qu'on suppose contrôlée et du débit moyen m. Cette borne est donnée par :

$$\overline{\alpha}'_{on-off}(m,h) = \frac{1}{st}\ln\left[1+\frac{m}{h}(e^{sht}-1)\right] \qquad (5.14).$$

Par conséquent, on peut définir :

$$\overline{Q}'_{g(X)-OO}(m) = 1+\frac{m}{h}(e^{sht}-1) \qquad (5.15).$$

Cette borne est atteinte lorsque le flot de trafic prend deux valeurs extrêmes, un minimum de zéro (phase « off ») et un maximum de h (phase « on ») et que la durée de la phase « on » est relativement grande par rapport au temps probable d'occupation du tampon avant un débordement.

Borne simple

Quand une connexion , en plus d'avoir son debit de crête contrôlé, est soumis à un algorithme de conformité de type « seau percé » avec comme paramètres (β, ρ), une meilleure borne [28] sur le débit effectif est :

$$\overline{\alpha}'_{sb}(m,\beta,\rho,h) = \frac{1}{st}\ln\left[1+\frac{tm}{\overline{X}[0,t]}(e^{s\overline{X}[0,t]}-1)\right] \qquad (5.16),$$

où $\overline{X}[0,t] = \min\{ht, \beta+\rho t\}$, i.e., $\overline{X}[0,t]$ est le nombre maximal de paquets que la source peut générer dans une fenêtre de temps t.

Par conséquent, on peut définir

$$\overline{Q}'_{g(X)-sb}(m) = 1+\frac{tm}{\overline{X}[0,t]}(e^{s\overline{X}[0,t]}-1) \qquad (5.17).$$

Approximation en T inversé

L'approximation en T inversé [28,29] est calculée analytiquement en utilisant un motif en forme de T inversé.

On remarque bien que les différentes approximations sont en général croissantes en fonction de m.

Lemma 3 $h_{\overline{Q},m}(m_r) \le f_{m,h}(X)\big|_{m_r}$ pour m_r dans un voisinage donné de m.

Preuve

Pour cette preuve, nous commencerons par montrer que $\overline{Q}_{g(X)}(m)$ est concave en m.

Supposons que $X, Y \in \chi(h)$ et $E(g(X))=m_1$, $E(g(Y))=m_2$. Soit Z une variable égale à X avec une probabilité θ ou Y avec une probabilité $1-\theta$, où $0<\theta<1$. Ceci correspond au cas pratique où l'on n'est pas sûr du type de la connexion. Alors,

$$E(g(Z)) = \theta E(g(X)) + (1-\theta)E(g(Y)) = \theta m_1 + (1-\theta)m_2$$

Et par suite,

$\overline{Q}_{g(X)}(\theta m_1 + (1-\theta)m_2) \ge E[e^{sZ[0,t]}]$ par définition de $\overline{Q}_{g(X)}(m)$

$$= \theta E[e^{sX[0,t]}] + (1-\theta)E[e^{sY[0,t]}]$$

Puisque ceci est valide pour tout *X[0,t]* et *Y[0,t]* vérifiant les contraintes, nous avons, en maximisant le membre droit de l'inégalité :

$$\overline{Q}_{g(X)}(\theta m_1 + (1-\theta)m_2) \geq \theta \overline{Q}_{g(X)}(m_1) + (1-\theta)\overline{Q}_{g(X)}(m_2)$$

Ainsi, $\overline{Q}_{g(X)}(m)$ est concave en *m*.

Considérons maintenant la différence:

$$h_{\overline{Q},m}(m_r) - f_{m,h}(X)\Big|_{m_r} = \frac{1}{st\overline{Q}_{g(X)}(m)}(\overline{Q}_{g(X)}(m_r) - \overline{Q}_{g(X)}(m) - (m_r - m)\frac{\partial \overline{Q}}{\partial m}\Big|_m)$$

$$- \frac{1}{2st(\overline{Q}_{g(X)}(m))^2}(\overline{Q}_{g(X)}(m_r) - \overline{Q}_{g(X)}(m))^2$$

$$+ \frac{1}{3st(\overline{Q}_{g(X)}(m))^3}(\overline{Q}_{g(X)}(m_r) - \overline{Q}_{g(X)}(m))^3$$

Comme $\overline{Q}_{g(X)}(m)$ est concave en *m*, alors $\overline{Q}_{g(X)}(m_r) - \overline{Q}_{g(X)}(m) - (m_r - m)\frac{\partial \overline{Q}}{\partial m}\Big|_m \leq 0$.

Si on note $Y_r = \overline{Q}_{g(X)}(m_r)$ et $Y = \overline{Q}_{g(X)}(m)$, alors la somme des deux derniers termes de la différence a le même signe que $R = \frac{2Y_r - 5Y}{6stY}$ et est négative dans le voisinage $[0, \frac{5}{2}Y] = [0, \frac{5}{2}\overline{Q}_{g(X)}(m)]$ de $\overline{Q}_{g(X)}(m)$. Et comme $\overline{Q}_{g(X)}(m)$ est croissant en *m,* ceci correspond à un intervalle [0, *m'*] tel que $m \in [0, m']$. Ainsi, dans ce voisinage de *m*, $h_{\overline{Q},m}(m_r) \leq f_{m,h}(X)\Big|_{m_r}$ i.e. l'usager a une fonction de tarification meilleure que $f_{m,h}(X)$ sans coût additionnel sur les mesures prises par le réseau tout en permettant au réseau de fournir les incitatifs adéquats par rapport à l'utilisation des ressources. Cependant, dans le cas où le débit mesuré ne se situe pas dans l'intervalle [0, *m'*], alors l'usager paie un tarif plus élevé que $f_{m,h}(X)$. Par conséquent, même si l'usager n'est pas obligé de respecter rigoureusement le débit moyen *m* spécifié dans son contrat, il a un fort incitatif à demeurer dans un intervalle $[0, m']$ autour de *m*. En considérant les approximations $\overline{Q}'_{g(X)}(m)$ de $\overline{Q}_{g(X)}(m)$ présentées précédemment, on voit que pour les approximations

qui sont linéaires (et donc concaves), l'intervalle de tolérance est $[0,\frac{5}{2}m]$. On constate donc que même avec ces approximations simples (borne « on-off » et borne simple), la fonction de tarification proposée $h_{\overline{Q}',m}(m_r)$ vérifie les 3 lemmes précédents et par conséquent est meilleure que la fonction de tarification $f_{m,h}(X)$. Ainsi, même si au point m, la fonction proposée $h_{\overline{Q}',m}(m_r)$ et $f_{m,h}(X)$ sont égales, l'amélioration introduite par $h_{\overline{Q}',m}(m_r)$ se fait sans ajouter à la complexité de mesure, i.e. le réseau a seulement besoin de mesurer le débit moyen m_r de la connexion. L'effort de calcul supplémentaire est négligeable, à moins que les ressources de calcul du gestionnaire des comptes soient très limitées. En outre, cette amélioration facilite la tâche de l'usager tout en gardant l'incitatif à respecter le débit moyen déclaré.

5.3 Garantie sur la gigue et extension de la tarification basée sur le débit effectif

Dans la section 5.1.2, nous avons rappelé comment les contraintes peuvent être généralisées au cas du délai et du taux de perte garantis (équations (5.8) et (5.9)). Cependant, comme nous l'avions déjà souligné, il n'y a pas de contraintes pour prendre en compte une éventuelle garantie sur la gigue ou variation de délai. Dans cette section, nous proposons une contrainte pour le cas où on voudrait une garantie sur la variation de délai et ensuite nous élargissons la tarification basée sur le débit effectif au cas où délai, taux de perte et gigue sont garantis.

5.3.1 Borne sur la variation de délai à l'intérieur d'un commutateur

Pour les applications en temps réel, la variation de l'interarrivée entre paquets (encore appelée gigue) est un paramètre important et doit être un paramètre garanti par la qualité de service. De ce fait, il est important d'avoir une approximation assez précise de la variation de délai du flot d'information. Dans la littérature, la gigue a été définie de plusieurs manières. Dans [64], on l'utilise pour exprimer la sporadicité du trafic et elle

est définie comme le nombre maximum de paquets dans un intervalle de temps moyen. Dans [40], le terme est utilisé pour traduire l'intensité de la distorsion du motif de trafic causée par le réseau et est défini comme la différence maximale entre les délais subis par deux paquets quelconques de la même connexion. Dans [68], la gigue est définie comme la variation de délai crête-à-crête telle que spécifiée par le Forum ATM, i.e. le (1-α) quantile du délai de transfert de la cellule (DTC) moins le délai de transfert fixe (ou minimum) encouru par toute cellule de la connexion. Le terme crête-à-crête se réfère à la différence entre les meilleur et pire cas de DTC, où le meilleur cas est égal au délai de transfert fixe dû aux délais de propagation sur les liens, aux délais de commutation et aux délais de transmission, tandis que le pire cas est une valeur de délai que l'on dépasse avec une probabilité inférieure ou égale à α.

Dans cette thèse, nous définissons la variation de délai ou gigue comme le (1-α) quantile de la différence entre les délais subis par deux paquets consécutifs de la même connexion. La gigue locale est la variation de délai au niveau d'un commutateur particulier, alors que la variation de délai de bout-en-bout est la gigue subie par les cellules le long d'un chemin de connexion allant de la source à la destination. Des méthodes ont été développées dans la littérature pour relier gigue de bout-en-bout et gigue locale [68].

Soient B_1 et B_2 la taille des tampons des commutateurs respectivement à l'arrivée de deux paquets successifs *P1* et *P2* de la même connexion *j* dans un commutateur qui multiplexe *N+1* connexions. On suppose qu'on a un seul type de connexion. Si la taille des paquets est petite comparée à la taille des tampons, la variation de délai au niveau du commutateur pour les deux paquets peut être définie comme la valeur absolue $G = \left| \frac{B_2 - B_1}{(N+1)C} \right|$. En utilisant les notations de la section 5.1.1, nous avons :

$$|B_2 - B_1| \le \sup(0, \lambda_1^N - (n+1)C\Delta t) \le \sup_t (X_{-1}'^{(N)} - (N+1)Ct) = \sup_t (X_{-1}'^{(N)} - (N)(\frac{N+1}{N}C)t) = W^{(N)}(\frac{N+1}{N}C),$$

où $X_t'^{(N)} = \sum_{k=1}^{N} X_{t,k}$ est l'agrégat (charge totale) des N autres connexions différentes de la connexion j et Δt est le temps entre deux arrivées consécutives sur la même connexion. Ainsi en utilisant l'équation 5.1, la probabilité de « débordement » de la gigue peut être définie par :

$$P(G>\frac{NB}{(N+1)C}) = P(|B_2 - B_1| > NB) \le P(W^{(N)}(\frac{N+1}{N}C) > NB) = \frac{1}{\sqrt{2\pi N\sigma^2 s_0^2}} e^{-NI_{t_0,s_0}(\frac{N+1}{N}C,B)}\left(1+O\left(\frac{1}{N}\right)\right)$$

où:

$$-I_{t_0,s_0}(\frac{N+1}{N}C,B) = \ln(M_{t_0}(s_0)) - (\frac{N+1}{N}Ct_0 + B)s_0 = \sup_t \inf_s \left[\ln(M_t(s)) - (\frac{N+1}{N}Ct + B)s\right]$$

et

$$\sigma^2 = \frac{\partial^2}{\partial s^2}\ln\left(E[e^{sX_{t,1}}]\right) = \frac{M_t''(s_0)}{M(s_0)} - (\frac{N+1}{N}Ct + B)^2 .$$

En remplaçant $I_{t_0,s_0}(\frac{N+1}{N}C,B)$, nous pouvons simplement écrire:

$$P(G > \frac{NB}{C}) = P(|B_2 - B_1| > NB) \le P(W^{(N)}(\frac{N+1}{N}C) > NB) \approx e^{-NI_{t_0,s_0}(\frac{N+1}{N}C,B) - \frac{1}{2}\ln(2\pi N\sigma^2 s_0^2)}$$

$$\le e^{N\ln(M_{t_0}(s_0)) - (N+1)Cs_0t_0 - NBs_0 - \frac{1}{2}\ln(2\pi N\sigma^2 s_0^2)} = Le^{-s_0 NB} \qquad (5.18)$$

où $L = e^{N\ln(M_{t_0}(s_0)) - (N+1)Cs_0t_0 - \frac{1}{2}\ln(2\pi N\sigma^2 s_0^2)}$ est une constante qui dépend de NB. Si nous utilisons l'équation 5.4 à la place de l'équation 5.1, nous aurons $L = e^{N\ln(M_{t_0}(s_0)) - (N+1)Cs_0t_0 - \frac{1}{2}\ln(4\pi NI_{t_0,s_0}(\frac{N+1}{N}C,B))}$. Cette équation peut être généralisée au cas de plusieurs types de connexions d'une manière similaire à celle utilisée à la section 5.1.2.

5.3.2 Contrôle d'admission de connexion

Considérons la contrainte suivante de qualité de service sur la gigue :

$$P(G > \frac{NB}{C}) \le e^{-N\alpha_0} = e^{-N\alpha} .$$

Supposons que les sources accèdent à un tampon de taille *NB* avec un débit de sortie *NC*. On suppose aussi que nous avons un seul type de sources et que celles-ci sont indépendantes et identiquement distribuées. La section précédente a montré qu'une condition suffisante pour le respect de la contrainte de qualité de service sur la gigue est $P(W^{(N)}(\frac{N+1}{N}C) > NB) \le e^{-N\alpha_0}$. Par conséquent, les équations (5.6) et (5.7) donnent respectivement les contraintes suivantes pour la région d'admission :

$$\alpha(s_0,t_0) \le \frac{N+1}{N}C + \frac{1}{t}(B - \frac{\alpha_0}{s}) = C' \qquad (5.19)$$

$$\alpha(s_0,t_0) \le \frac{N+1}{N}C + \frac{1}{t}(B - \frac{\alpha_0'}{s}) = C'_{B-R} \qquad (5.20)$$

où $\alpha(s_0,t_0)$ est le débit effectif de l'une des *N+1* connexions, $I_{t_0,s_0}(\frac{N+1}{N}C,B) \approx \alpha_0 - \frac{\frac{1}{2}\ln(4\pi N\alpha_0)}{N+\frac{1}{2\alpha_0}} = \alpha_0'$ et s_0,t_0 sont les solutions de l'équation d'optimisation :

$$-I_{t_0,s_0}(\frac{N+1}{N}C,B) = \ln(M_{t_0}(s_0)) - (\frac{N+1}{N}Ct_0 + B)s_0 = \sup_t \inf_s \left[\ln(M_t(s)) - (\frac{N+1}{N}Ct + B)s\right]$$

Ce processus peut aussi être généralisé au cas où on a plusieurs types de connexions d'une manière similaire à celle de la section 5.1.2.

Ces algorithmes de contrôle d'admission peuvent être appliqués à la liaison radio qui constitue le goulot d'étranglement. Ils permettent dans le cas de canaux dédiés ou partagés de réserver (ou d'assurer) à l'usager des ressources proportionnelles à son débit effectif tout en lui donnant des garanties. On n'accepte l'usager que si les contraintes (5.19) ou (5.20), (5.8) et (5.9) sont respectées.

5.3.3 Tarification avec délai, gigue et taux de perte garantis

Supposons par exemple que les classes de trafic sont partitionnées en deux sous-ensembles J_1 et J_2. Le service est PAPS (Premier Arrivé, Premier Servi), excepté que les

sources de trafic de type J_1 sont prioritaires par rapport aux sources de trafic de type J_2. Supposons aussi que pour $i \in J_1$, il y a une garantie de qualité de service de la forme:

$$P(délai > B_1 / C) \leq e^{-\gamma_1},$$

sur le délai et une garantie de qualité de service sur la gigue de la forme :

$$P(gigue > B_1 / C) \leq e^{-\gamma_2}$$

De plus, pour toutes les sources on a une garantie sur le taux de perte de cellules :

$$P(débordement\ du\ tampon) \leq e^{-\gamma_2},$$

Comme dans la section 5.1.2, nous pouvons appliquer itérativement les équations (5.6) ou (5.7) et obtenir les contraintes suivantes :

$$\sum_{j \in J_1} \alpha_j(s_{1'}, t_{1'}) \leq K_{1'} \tag{5.21}$$

$$\sum_{j \in J_1} \alpha_j(s_{1''}, t_{1''}) \leq K_{1''} \tag{5.22}$$

$$\sum_{j \in J_1 \cup J_2} \alpha_j(s_2, t_2) \leq K_2 \tag{5.23}$$

Dépendamment des valeurs de $s_{1'}, t_{1'}, s_{1''}, t_{1''}, K_{1'}, K_{1''}$, la contrainte (5.21) ou (5.22) sera la contrainte active et déterminera ainsi la région d'admission à l'intérieur de laquelle le fournisseur réseau peut opérer [28].

Supposons qu'un opérateur réseau facture f_i par unité de temps pour une connexion de type i, $i=1,2$. Le revenu $n_1 f_1 + n_2 f_2$ est maximisé en opérant, si possible, en un point de la frontière de la zone d'admission. Si l'équation (5.23) est active, alors il serait approprié de facturer les connexions de type 1 et 2 à des taux respectivement proportionnels à $\alpha_1(s_2, t_2)$ et $\alpha_2(s_2, t_2)$. Si l'une des contraintes (5.21) ou (5.22) est active, alors on facture les connexions de type 1 à un taux proportionnel à $\alpha_1(s_1, t_1)$. Si l'une des contraintes (5.21) ou (5.22) est active en même temps que la contrainte (5.23), alors les connexions de type 1 seront facturées au taux $\lambda_1 \alpha_1(s_1, t_1) + \lambda_2 \alpha_1(s_2, t_2)$, où λ_1, λ_2 sont les prix virtuels associés aux contraintes (5.21) ou (5.22) d'une part et (5.23)

d'autre part, tandis que les connexions de type 2 sont toujours facturées au taux $\lambda_2\alpha_2(s_2,t_2)$.

Ce schéma de tarification peut être implémenté en utilisant la fonction de tarification proposée à la section 5.2. Si nous supposons par exemple que l'opérateur utilise une des approximations linéaires (5.14 ou 5.16), alors pour les connexions de type 2, l'opérateur propose un n-tuple *(a, b, c, d)* et l'usager paie pour la durée totale T de sa connexion :

$$h(m_r) = aT + bV + cVm_r + dVm_r^2,$$

où a, b, c, d dépendent de $m_1, h_1, m_2, h_2, s_2, t_2$ .Cette formule a été obtenue en multipliant la fonction de tarification par unité de temps $h_{\overline{Q}',m}(m_r)$ par la durée T de la connexion et ensuite en remplaçant la borne supérieure $\overline{Q}_{g(X)}(m)$ par son approximation $\overline{Q}'_{g(X)}(m)$ qui est une fonction linéaire de m. Par conséquent, nous avons une tarification qui est proportionnelle à la fois à la durée T et au volume V (les deux premiers termes), mais qui introduit aussi des corrections (les termes restants) qui dépendent du volume et du débit moyen mesuré pour tenir compte de l'incertitude de l'usager et de la sporadicité du trafic. Ces corrections donnent plus de flexibilité à l'usager comparativement au schéma classique de tarification basée sur le débit effectif.

Pour les connexions de type 1 qui sont facturées au taux $\lambda_1\alpha_1(s_1,t_1) + \lambda_2\alpha_1(s_2,t_2)$, la fonction de tarification pour une connexion qui dure T unités de temps et transfère un volume V de données est :

$$\begin{aligned} h(m_r) &= \lambda_1(a_1T + b_1V + c_1Vm_r + d_1Vm_r^2) + \lambda_2(a_2T + b_2V + c_2Vm_r + d_2Vm_r^2) \\ &= aT + bV + cVm_r + dVm_r^2 \end{aligned},$$

où a_1, b_1, c_1, d_1 dépendent de $m_1, h_1, s_{1'}, t_{1'}, s_{1''}, t_{1''}$ et a_2, b_2, c_2, d_2 dépendent de $m_1, h_1, m_2, h_2, s_2, t_2$. La forme de la fonction de tarification demeure donc la même.

5.4 Résultats

Dans un premier temps, nous testons les résultats de la section précédente relatifs à la contrainte de garantie sur la gigue (équation 5.18). Pour cela, nous utilisons le simulateur Comnet ®. Le modèle tel qu'il apparaît dans Comnet est montré à la Figure 5.5.

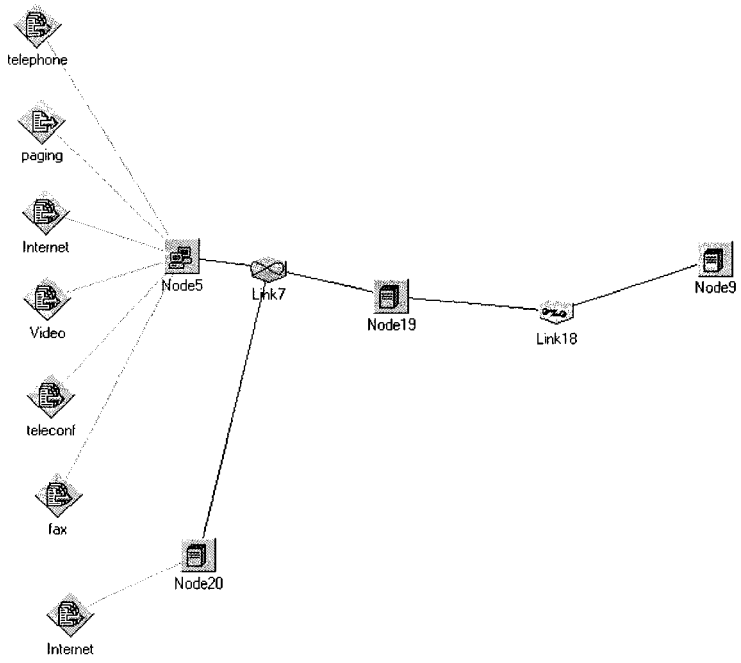


Figure 5.5 Modèle de test dans Comnet

Dans ce modèle, nous utilisons un groupe d'ordinateurs (*Node5*) pour modéliser plusieurs sources semblables. Un nœud de traitement (*Node20*) est utilisé pour modéliser la source dont on veut mesurer la gigue. Le trafic généré par toutes les sources est multiplexé au noeud *Node19* et ensuite envoyé au nœud de destination (*Node9*) à travers un lien (*Link18*), qui ici symbolise le lien radio. La capacité de ce lien est donc de 2 Mbps [124]. Le réseau supporte des paquets de 1518 octets maximum. Nous simulons une large gamme de modèles de trafic avec des distributions de Pareto, de Weibull, des distributions en rafales, etc. Le Tableau 5.1 résume les différentes distributions utilisées dans les modèles de session.

Tableau 5.1 Caractérisation des modèles de source

	Nombre de messages par session	Interarrivée des messages (sec)	Taille des messages (octets)
Application 1 (Internet)	Pareto (échelle=33.0, forme=1.16)	Sporadique: • Interarrivée des rafales Pareto (échelle=1.0, shape=1.5) • Interarrivée à l'intérieur des rafales: Weibull (forme=0.382, échelle=1.46) • Nombre de messages à l'intérieur des rafales: Pareto (position=1.0, forme=2.43)	Mixte : • 93% des messages: Lognormale (moyenne=8897.0, écart-type=37009.0) • Reste des messages Pareto (échelle=3328.0, forme=1.383)
Application 2 (Voix)	Géometrique (min=0.0, moyenne=3000.0)	Sporadique: • Interarrivée des rafales Exponentielle (moyenne=1.002) • Interarrivée à l'intérieur des rafales: 0.02 • Nombre de messages à l'intérieur des rafales: Géometrique (min=0.0, moyenne=17.0)	70
Application 3 (Fax)	Exponentielle (moyenne=1000.0)	Exponentielle (moyenne=0.156)	Exponentielle (moyenne=281.0)
Application 4 (Vidéo)	Triangulaire (min=648.0, mode=1282.0, max=1916.0)	0.04166666	Sporadique: • Premier message: Lognormale (moyenne=1897, écart-type=800) • Onze prochains messages: Lognormale (moyenne=637, écart-type=467)

La variation de délai est calculée au noeud *Node20* parce que ce noeud supporte une seule instance de l'application. Nous étudions à la fois le cas où la variation de délai est définie comme la variation entre deux paquets consécutifs quelconques de la même connexion et le cas où la gigue est définie comme variation relative par rapport au délai minimum des paquets de la connexion. Ces données sont utilisées pour déduire la probabilité $P(gigue > x)$ (qui est le complémentaire de la fonction cumulative de distribution (cdf)).

Pour obtenir les bornes théoriques (équation (5.18)), nous recueillons pour chaque type de trafic, une trace générée par une seule source et calculons les points d'extremum s_0, t_0, en utilisant le logiciel *msa* développé par Courcoubetis *et al.* [31], et disponible à http://www.ics.forth.gr/netgroup/msa/, avec les valeurs appropriées de capacité et de taille de tampon (*(N+1)C et (N+1)B pour N sources*). Nous supposons aussi que le débit moyen total des N sources est inférieur à la capacité de la liaison pour des raisons évidentes de stabilité et de stationnarité.

Les figures 5.6-5.10 montrent les différentes bornes ainsi que la probabilité $P(gigue > x)$ obtenue par la simulation (et notée *Psimul* dans le cas où la gigue est définie comme la variation entre deux paquets consécutifs quelconques et *Psimul_abs* dans le cas où les variations sont prises par rapport au délai minimum des paquets de la connexion) comme une fonction de x pour différents types de trafics avec différents nombres de connexions. Ces figures montrent aussi les trois approximations de la borne $Le^{-s_0 NB}$ (équation (5.18)). La première approximation, *Psimple*, est obtenue avec $L = e^{N\ln(M_{t_0}(s_0))-(N+1)Cs_0t_0}$, i.e., *Psimple* est égal à $e^{-NI_{t_0,s_0}(\frac{N+1}{N}C,B)}$; la deuxième approximation *B-R-approx* est calculée avec $L = e^{N\ln(M_{t_0}(s_0))-(N+1)Cs_0t_0-\frac{1}{2}\log(4\pi NI_{t_0,s_0}(\frac{N+1}{N}C,B))}$, i.e., en utilisant l'amélioration de Bahadur-Rao et l'approximation de [94]. Finalement, la dernière approximation, *B-R*, est obtenue avec $L = e^{N\ln(M_{t_0}(s_0))-(N+1)Cs_0t_0-\frac{1}{2}\log(2\pi N\sigma^2 s_0^2)}$.

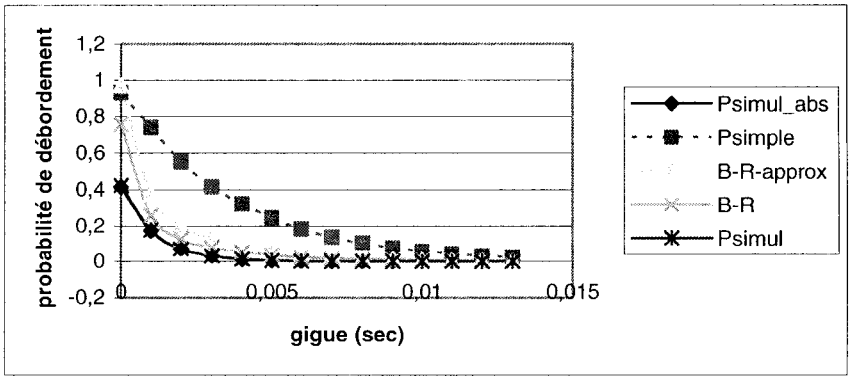


Figure 5.6 Probabilité de « débordement » de la gigue pour « l'application 3 », 25 connexions et un lien de 2 Mbps

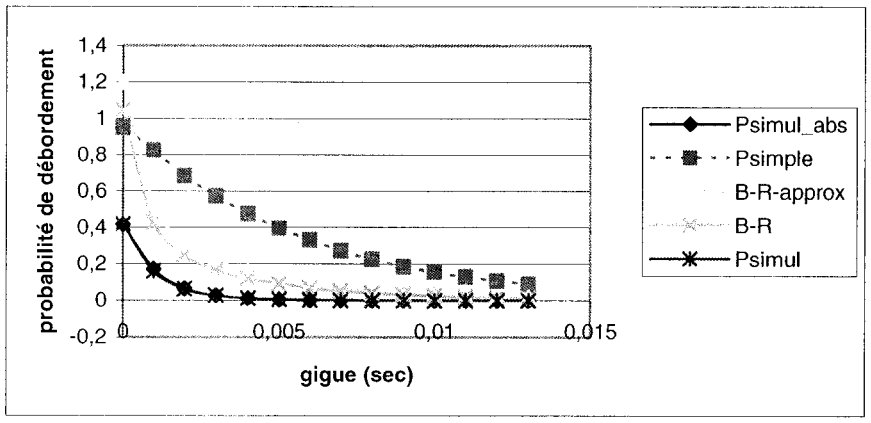


Figure 5.7 Probabilité de « débordement » de la gigue pour « l'application 3 », 40 connexions et un lien de 2 Mbps

Dans un premier temps, on note que les deux définitions de la variation de délai donnent approximativement les mêmes résultats en terme de probabilité de « débordement ». Pour le type de trafic utilisé pour les deux figures précédentes, les bornes sont plus strictes pour de plus petites valeurs de *N*. Notons aussi que de manière absolue, les bornes sont plus précises pour de grandes valeurs de gigues. Finalement, pour les petites valeurs de gigue, les bornes *B-R* and *B-R-approx* ne sont pas très utiles car elles sont trop grandes. Dans ce cas, la borne simple reste encore la meilleure.

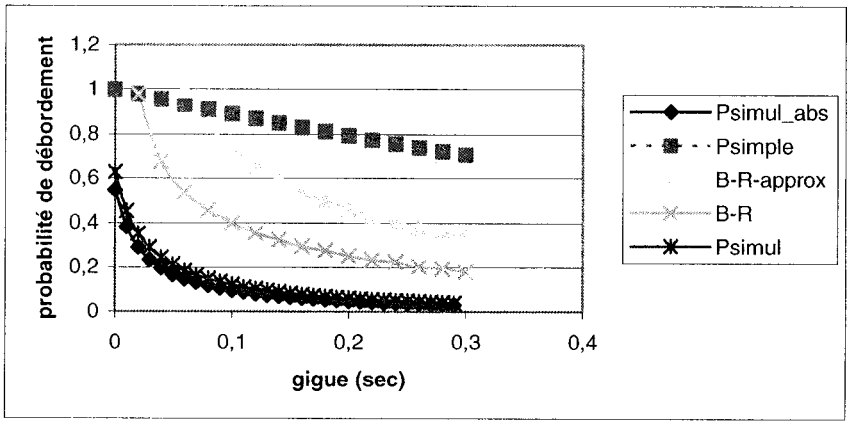


Figure 5.8 Probabilité de « débordement » de la gigue pour « l'application 1 », 8 connexions et un lien de 2 Mbps

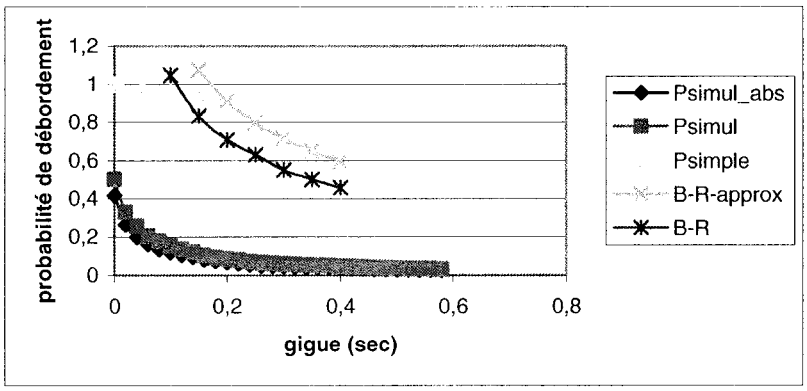


Figure 5.9 Probabilité de « débordement » de la gigue pour « l'application 1 », 50 connexions et un lien de 2 Mbps

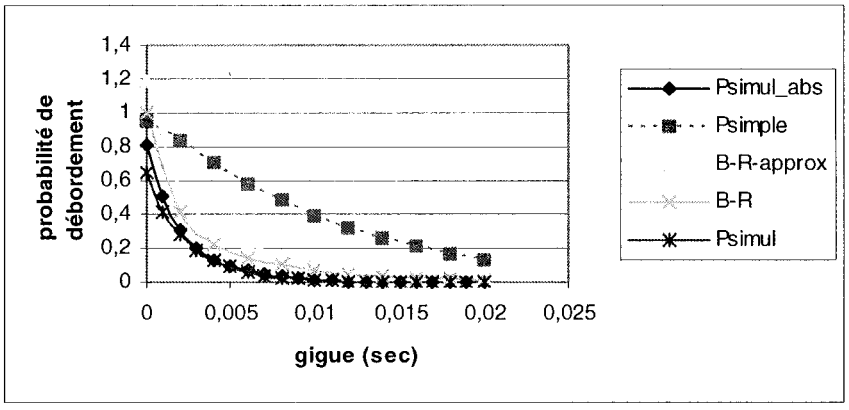


Figure 5.10 Probabilité de « débordement » de la gigue pour « l'application 4 », 5 connexions et un lien de 2 Mbps

Ces dernières figures confirment le fait que nos bornes sont meilleures pour de petites valeurs de *N*.

Une autre comparaison, digne d'intérêt, est d'évaluer l'amélioration de nos bornes par rapport à la borne *borne-délai* obtenue en traitant la gigue comme un simple délai (*NC et NB pour N sources*). Les figures 5.11-5.13 montrent l'amélioration absolue obtenue en utilisant nos bornes pour les applications 1 et 3.

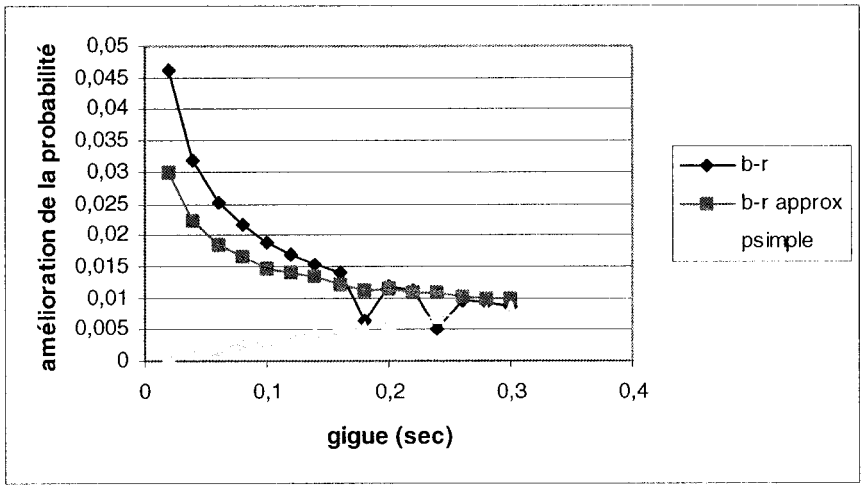


Figure 5.11 Amélioration absolue sur la « borne-délai » pour « l'application 1 » avec 8 connexions et un lien de 2 Mbps

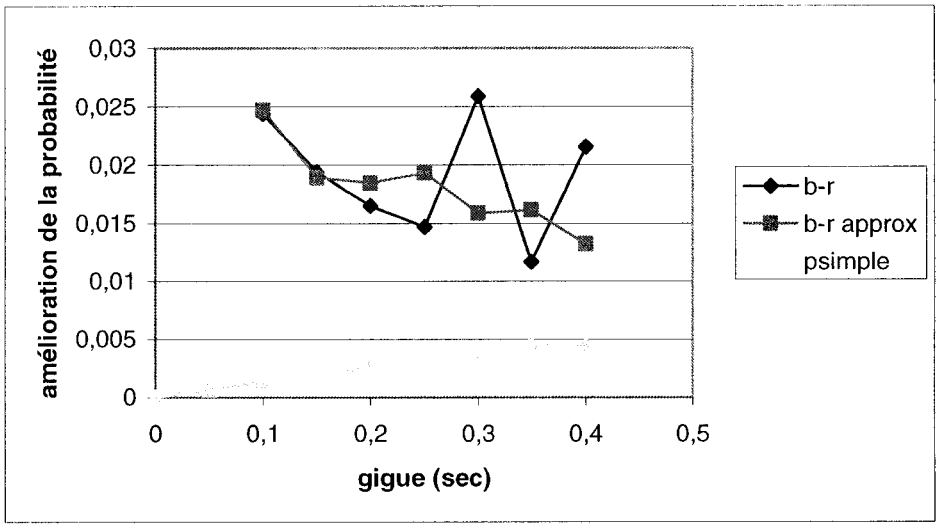


Figure 5.12 Amélioration absolue sur la « borne-délai » pour « l'application 1 » avec 50 connexions et un lien de 2 Mbps

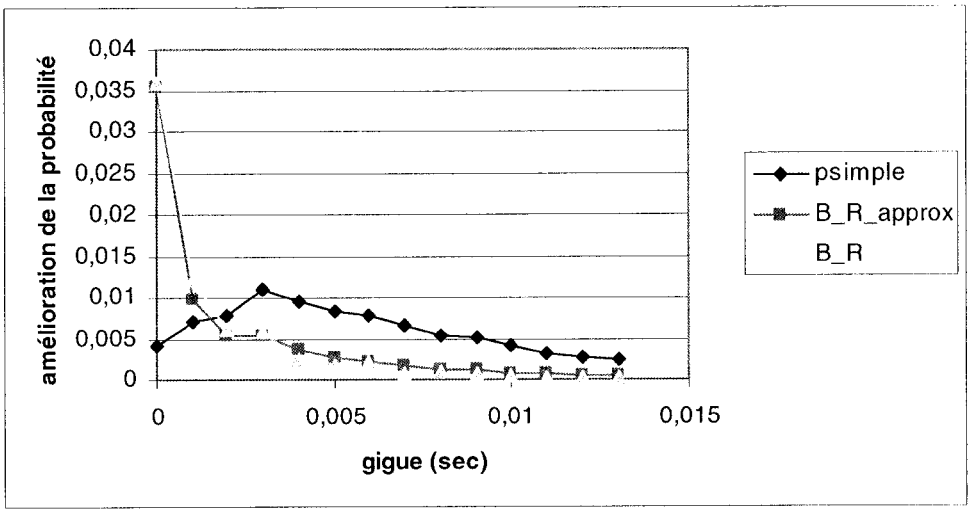


Figure 5.13 Amélioration absolue sur la « borne-délai » pour « l'application 3 » avec 25 connexions et un lien de 2 Mbps

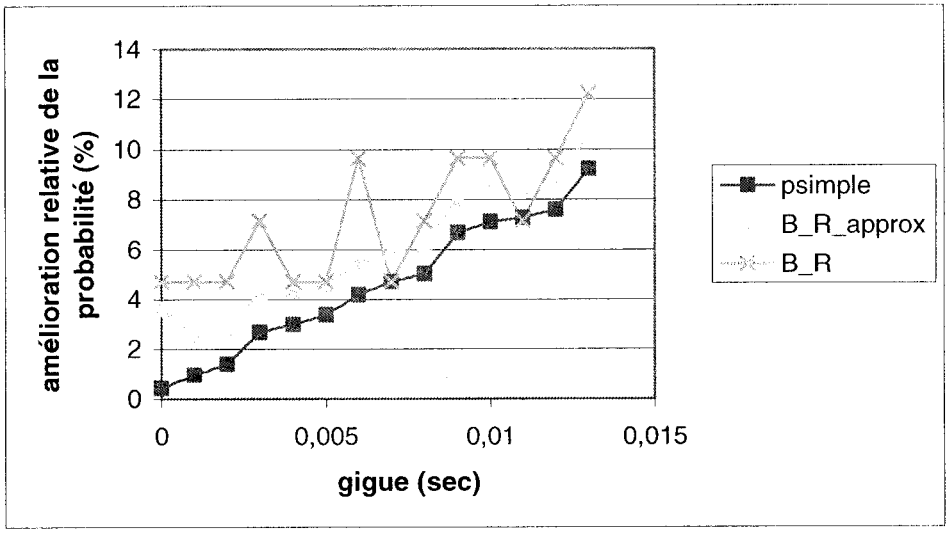


Figure 5.14 Amélioration relative sur la « borne-délai » pour « l'application 3 » avec 25 connexions et un lien de 2 Mbps

En termes d'amélioration absolue, les bornes proposées ne semblent pas être assez efficaces. Cependant, quand on les ramène à des valeurs relatives, on note que l'amélioration apportée peut être conséquente (jusqu'à 10%) comme le montre la Figure 5.14. La meilleure amélioration relative est presque toujours obtenue avec la borne de Bahadur-Rao et est meilleure quand N croît.

Dans le processus d'évaluation des bornes théoriques, nous devons déterminer la granularité des époques utilisées pour effectuer la capture de la trace. Ceci a une influence directe sur la détermination du débit effectif par le logiciel *msa* et sur le calcul du point d'extremum de t [31]. Un compromis doit être fait entre le niveau de détail du fichier de trace (qui ne doit pas être trop grand) et la précision des résultats obtenus (qui dépend directement du niveau de détail). Quand la valeur de t est petite, les détails du trafic sont bien pris en compte et l'algorithme de résolution de l'équation (5.5) couvre un large domaine de t. Les résultats obtenus sont donc plus précis. Aussi, a priori, les variations rapides ont une influence importante sur la gigue. De ce fait, une bonne valeur empirique pour le niveau de détail de la trace est de 1 ou 2 ms. La Figure 5.15 montre les bornes pour l'application de voix avec deux granularités temporelles différentes.

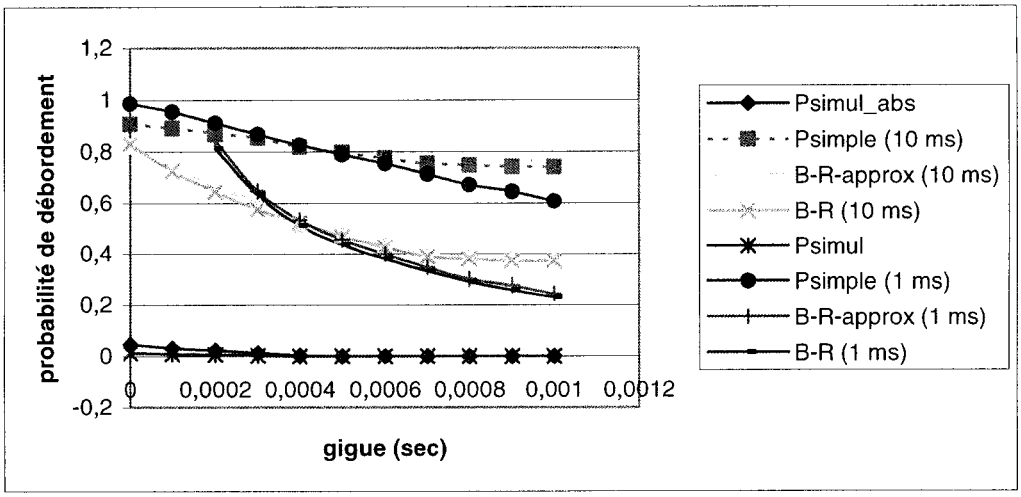


Figure 5.15 Bornes pour « l'application 2 » avec 120 connexions, un lien de 2 Mbps et deux granularités temporelles

Les bornes obtenues avec des granularités temporelles plus fines sont meilleures que celles obtenues avec des granularités plus grossières pour les grandes valeurs de gigue. Il faudrait aussi noter que les bornes obtenues sont assez relâchées. Ceci est dû au fait que le nombre de connexions impliquées est assez grand. En effet, nous avions déjà mentionné que les bornes étaient plus strictes pour de plus petites valeurs de N. Ce comportement peut s'expliquer par le fait que les bornes théoriques considèrent le pire cas où toutes les N sources injectent des données dans le tampon. Alors qu'en pratique

quand N croît, seule une petite proportion des sources sont actives à la fois, ce qui donne des variations moins importantes sur le délai.

Une autre condition pour obtenir des bornes acceptables est que le fichier de trace utilisé par le logiciel *msa* doit couvrir une assez grande période.

Finalement, nous avons testé la fonction de tarification proposée. À l'aide du logiciel Matlab ®, nous avons tracé sur le même graphe la borne supérieure du débit effectif noté *alpha*, la fonction linéaire de tarification et la fonction de tarification proposée. La borne supérieure utilisée est la borne ON-OFF introduite dans [119] (équations 5.14 et 5.15). Comme déjà mentionné, le calcul de la fonction de tarification dépend seulement du débit moyen mesuré. Comme la fonction de tarification proposée peut être facilement générée, nous avons généré à la fois des fonctions de troisième et cinquième degré. Les figures 5.16-5.18 montrent les résultats pour les applications 2 et 4 avec différentes tailles de tampon.

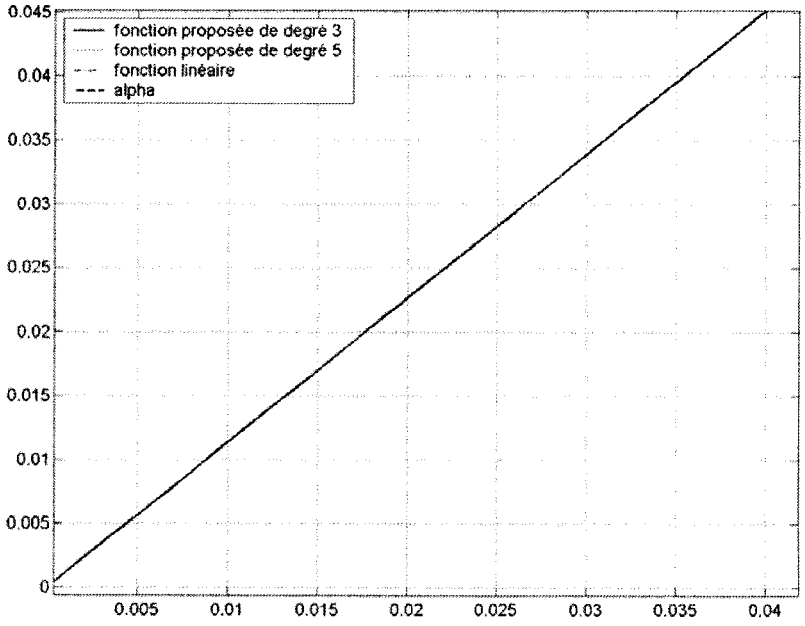


Figure 5.16 Fonction de tarification pour « l'application 2 » sans tampon

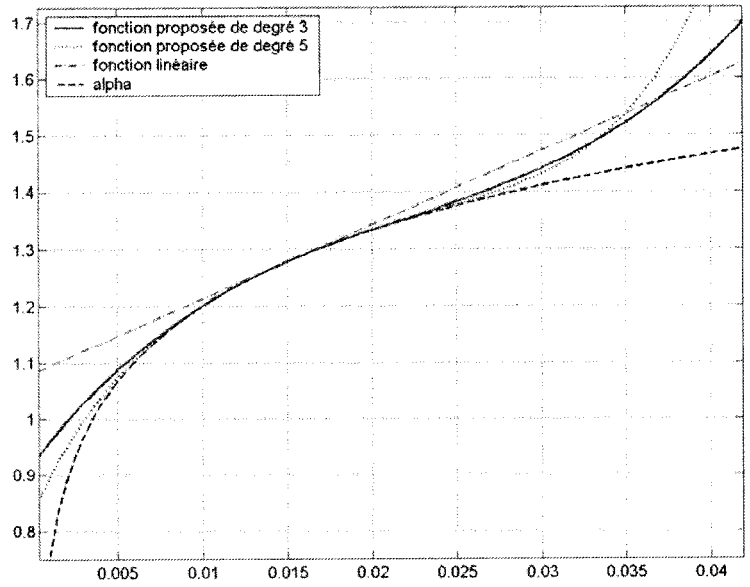


Figure 5.17 Fonction de tarification pour « l'application 2 » et un tampon de 2000 octets

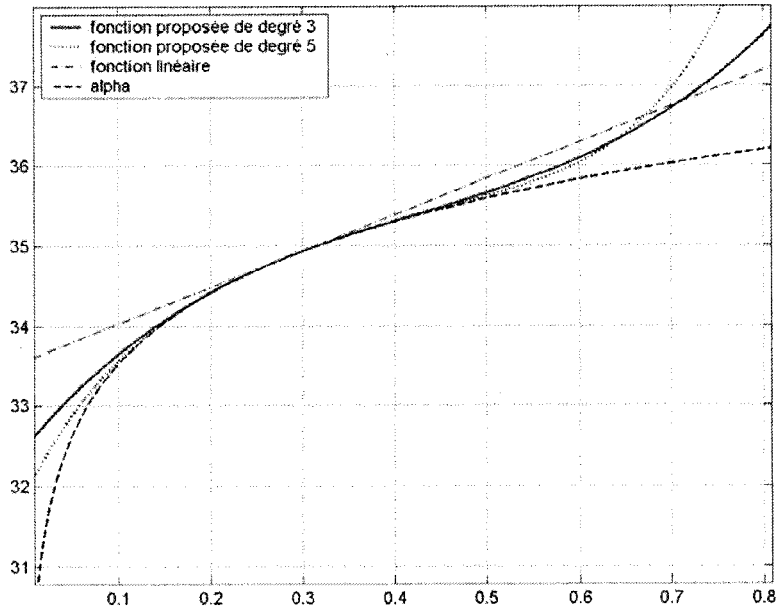


Figure 5.18 Fonction de tarification pour « l'application 4 » et un tampon de 5000 octets

Pour de petites tailles de tampon, il n'y a pas de différence entre les fonctions de tarification. Par contre, quand la taille du tampon croît, notre fonction de tarification reflète plus précisément, à l'intérieur du voisinage défini dans le lemme 3, la borne sur le débit effectif. Finalement, l'utilisation d'une fonction de tarification de degré plus grand n'engendre pas une amélioration notable. Cependant, comme l'accroissement de la complexité due à une fonction de tarification de degré plus grand est minimal, l'opérateur pourra choisir un compromis adéquat.

Dans ce chapitre, nous avons analysé le cadre de la tarification dans les réseaux multiservice et avons proposé un schéma amélioré de tarification basé sur le concept de débit effectif et qui prend en compte des paramètres supplémentaires de qualité de service comme par exemple la variation de délai. Ce schéma donne aussi plus de flexibilité à l'usager.

CHAPITRE 6

CONCLUSION

Les récents développements technologiques ont permis une évolution des systèmes cellulaires mobiles de deuxième génération vers une troisième génération plus performante en termes de capacité, de couverture et de qualité de service. Il en résulte un éventail de nouveaux services et de nouvelles applications, ce qui implique de nouveaux problèmes et défis liés à la qualité de service, à la performance et à la facturation. Dans cette thèse, nous avons essayé d'aborder ces différents problèmes en les définissant, en les caractérisant et finalement en proposant des approches de solutions. Dans ce chapitre final, nous présentons une synthèse générale des travaux avant d'aborder les limitations de notre recherche ainsi que les indications de recherche future.

6.1 Synthèse des travaux

Les réseaux cellulaires multiservices désignent des réseaux transportant à la fois de la voix, des données, du trafic multimédia et vidéo [34,39]. Leur architecture intègre des réseaux d'accès sans fil à large bande qui sont plus adaptés au trafic multimédia et aux applications qui génèrent un trafic en rafales [4,19]. Un exemple de réseau cellulaire multiservice est le système UMTS (Universal Mobile Telecommunications System) qui est un système de communications mobiles de troisième génération introduit par l'Union Internationale des Télécommunications (UIT) [124].

Dans les réseaux cellulaires multiservices, les caractéristiques particulières des nouveaux types de trafic posent de nouveaux défis [104]. En effet, la croissance exponentielle du trafic en rafales change la dynamique des réseaux et nécessite, lors de la conception des boucles d'accès, une bonne évaluation de différentes classes de service. De plus, la combinaison de la voix, des données, de la vidéo et du multimédia dans des réseaux sans fil à haut débit exige de nouvelles solutions aux problèmes de

facturation, de synchronisation dans les systèmes multimédia, de qualité de service, de technologies d'accès, de performance des services et des applications envisagées, et d'affectation dynamique des canaux [34,77].

Avec le développement des réseaux de troisième génération, plusieurs études [20] se sont penchées sur les performances de leurs technologies radio. Cependant, aucune étude ne s'est vraiment intéressée aux performances de ces réseaux d'un point de vue « applicatif », lorsque différents types de trafic sont multiplexés. Nous avons donc, dans un premier temps, étudié les performances de la partie accès (ou boucle locale) des réseaux cellulaires multiservices dans un contexte où sont multiplexés plusieurs types de service ou de trafic présentant souvent une dépendance (ou corrélation) à long terme et un caractère fractal et en rafales. Pour cela, nous avons d'abord retenu un certain nombre de paramètres de qualité de service ainsi que d'applications qui présentaient un intérêt particulier pour notre étude. Ensuite, nous avons élaboré des modèles de trafic, présenté le modèle UMTS ainsi que l'outil de simulation utilisé et finalement adapté les modèles de trafic à l'outil de simulation. Les résultats obtenus ont été analysés et l'impact de divers paramètres et politiques a été étudié.

De manière générale, avec une bonne conception, les performances des réseaux de troisième génération avec le trafic envisagé sont acceptables. Il faut cependant définir avec soin les profils de QoS associés à chaque classe ou application en utilisant judicieusement la sur-souscription. Pour la classe « en flux », il faudrait mettre en place des tampons adéquats pour assurer la qualité de service notamment par rapport aux délais. Finalement, la prise en compte supplémentaire de la gestion de la QoS par un mécanisme d'ordonnancement au niveau des couches supérieures permet une amélioration des performances. Bien que les résultats dépendent d'un grand nombre de paramètres, nous avons cependant pu établir l'impact d'un certain nombre de politiques comme la mise en place de la priorité au niveau IP, la sur-souscription des liaisons.

Au niveau de la fiabilité, différentes topologies ont été étudiées. Avec la robustesse des équipements disponibles, la redondance nécessaire pour assurer une grande disponibilité est minimale et peut être atteinte avec des topologies très simples.

Une topologie d'interconnexion des RNC en arbre avec quelques liens redondants fournit la disponibilité souhaitée avec des coûts relativement moindres.

Un autre problème soulevé par la combinaison de différents types de trafic est celui de la tarification. Avec des trafics ayant des exigences différentes, les opérateurs se heurtent au problème du choix de type et de la sorte de tarification à adopter. En effet, de manière générale, le rôle de la tarification n'est pas seulement de générer des revenus. Elle introduit aussi une rétroaction qui permet d'exercer un certain contrôle sur les usagers en les incitant à réagir aux tarifs par une minimisation de leurs coûts. Par exemple, une politique de tarification appropriée devrait jouer -même partiellement- le rôle d'un mécanisme de contrôle de congestion en encourageant une utilisation efficace des ressources du réseau, et aussi inciter les utilisateurs du réseau à regrouper leurs besoins dans des classes de service appropriées. Nous avons aussi étudié les différents types de tarification disponibles dans les réseaux à commutation et proposé une extension de la tarification basée sur le débit effectif. Cette extension regroupe d'une part la proposition d'un critère pour prendre en compte la contrainte sur la variation de délai. En effet, le modèle original ne proposait pas la prise en compte de la gigue ou tout au plus la considérait comme un délai. Notre critère a été validé par des simulations. Il permet de réaliser une amélioration pouvant aller jusqu'à 10% sur les bornes et par conséquent sur le nombre d'admission. D'autre part, nous avons introduit une famille de fonctions de tarification qui tiennent compte des incertitudes des usagers, de la sporadicité du trafic et qui offrent aussi une plus grande précision tout en gardant les bonnes propriétés sur les incitatifs. En outre, pour toutes ces améliorations, le système de facturation ne contrôle, tout comme dans le schéma classique, que le débit moyen de chaque connexion. Pour l'usager, les économies reliées à la nouvelle fonction de tarification croissent avec la taille des tampons disponibles pour le trafic au niveau des nœuds du réseau. Par exemple, pour un trafic sporadique qui oblige l'utilisation de grands tampons, la différence sera plus notable.

6.2 Limitations des travaux

Dans l'analyse de performance, malgré les résultats satisfaisants obtenus, le logiciel de simulation utilisé nous a amené à faire quelques restrictions. Nous avons dû simplifier l'implémentation du modèle de voix au niveau du nombre d'usagers représentés, nous avons aussi un seul profil de QoS par classe de trafic, ce qui nous offre un contrôle moins précis et nous a amené à modifier les profils associés à chaque type de trafic pour les transformer en profils liés aux classes. Pour les usagers, nous avons aussi simulé une seule application par unité mobile (sauf dans le cas des applications de voix) alors qu'en réalité, il est fort probable qu'un même usager exécute plusieurs applications (vidéoconférence et navigation sur Internet par exemple) simultanément. Dans ce cas, on ne peut prédire quel serait le délai encouru par chaque application. Dans la mise en oeuvre de notre étude de performance, nous nous sommes aussi finalement restreint à un sous-ensemble de paramètres de qualité de service. La complexité du modèle nous a aussi amené à ne pas considérer ni les relèves, ni le routage.

Dans l'analyse de fiabilité, les résultats établis sont relatifs et non absolus. En effet, la rareté des événements simulés nous a amené à modifier les données réelles pour avoir en un temps raisonnable une simulation significative. De plus, seules trois topologies différentes ont été étudiées.

Dans le processus de caractérisation de la gigue, et de manière générale pour la caractérisation par le débit effectif, on suppose l'existence et la connaissance du profil du trafic. La variation du profil demanderait de recalculer les points de fonctionnement du réseau. De plus, le critère sur la variation de délai est valide seulement localement (au point où se situe le goulot d'étranglement) et non de bout en bout. Quant à la fonction de tarification, elle n'apporte des améliorations substantielles que dans un voisinage retreint de la moyenne déclarée par l'usager. Ceci force certes d'une part l'usager à respecter plus ou moins le profil déclaré, mais oblige l'opérateur à réaliser une conception très soigneuse de sa grille de tarifs sinon il risque de perdre les avantages liés à la nouvelle fonction en sortant par exemple du voisinage privilégié. Comme nous l'avions aussi noté, les bénéfices liés à l'utilisation d'une fonction de plus haut degré diminuent avec

l'augmentation du degré. L'utilisation de fonctions de degré de plus en plus élevés n'est donc pas forcément rentable. Finalement, le schéma de tarification proposé permet seulement de fixer des prix relatifs entre les différentes classes de trafic ou les différentes applications. Pour aboutir à des prix absolus, il faudrait prendre en compte des données supplémentaires comme le bien-être social, le coût du réseau, la période d'amortissement, le contrôle d'admission, les stratégies de gestion [45].

6.3 Indications de recherche future

Les réseaux de troisième génération sont des réseaux relativement nouveaux et encore très peu implantés. De ce fait, les avenues de recherche sont encore nombreuses. Dans le domaine de l'analyse de performance, avec l'apparition de nouvelles applications et l'évolution des habitudes des usagers, la caractérisation du trafic demeure un domaine en perpétuel renouvellement. Il faut en permanence affiner les modèles existants ou en développer de nouveaux. Il en est de même pour le modèle UMTS. L'implémentation actuelle du modèle implique un certain nombre de restrictions sur le nombre d'usagers qu'on pourrait raisonnablement modéliser, la granularité des profils de QoS, les canaux utilisés, etc. On pourrait donc se pencher sur des modèles qui utilisent une simulation (ou un simulateur) plus efficace (par exemple, les nouvelles versions de Opnet intègrent des mécanismes pour rendre plus efficace la simulation des liens radio) pour avoir un modèle sans simplification sur le nombre d'usagers. On pourrait aussi revoir les performances dans le cas où chaque application implémente son profil de QoS propre au lieu de l'hériter directement de la classe de trafic. L'impact de nombreux autres facteurs comme la classe de trafic affectée à chaque application, les caractéristiques de l'interface radio, l'environnement considéré pourrait aussi être considéré. À partir de la contrainte locale sur la gigue, on pourrait développer une extension prenant en compte la variation de délai de bout en bout. Toutes les contraintes pourraient ensuite être intégrées avec les considérations de gestion, de bien-être social,

de coût du réseau, d'amortissement, de contrôle d'admission pour obtenir une grille complète de tarification.

RÉFÉRENCES

[1] 3GPP Forum, "3GPP home page", *http://www.3gpp.org*, 2002.

[2] 3GPP Technical Specification Group Services and System Aspects, *QoS Concept and Architecture*, 3GPP TS 23.107 v5.3.0 (2002-01), 2002.

[3] Aguado, L. E., O'Farrell, T. et Harris, J. W., "Evaluation of impact of mixed traffic on UTRA performance", *IEE Electronics Letters*, vol. 36, no. 22, pp. 1876-1877, 2000.

[4] Ahn, H. et Kim, J. K., "Shifting-Level Process as a Lrd Video Traffic Model and Related Queuing Results", *Computer Communications*, vol. 23, pp. 371-378, Feb. 2000.

[5] Akhtar, S., Malik, S. A. et Zeghlache, D., "Prioritized admission control for mixed services in UMTS WCDMA networks", *12th International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC 2001)*, pp. B-133-B-137, vol. 1, 2001.

[6] Anania, L. et Solomon, R. J., "Flat: the minimalist price", In: *Internet Economics*, McKnight, L. W. et Bailey, J. P. (eds.), Cambridge, Mass.: MIT Press, 1997.

[7] Aras, C. M., Kurose, J. F., Reeves, D. S. et Schulzrinne, H., "Real-time communication in packet-switched networks", *Proceedings of the IEEE*, vol. 82, no. 1, pp. 122-139, Jan. 1994.

[8] Arlitt, M. et Jin, T., "A workload characterization study of the 1998 World Cup Web site", *IEEE Network*, vol. 14, no. 3, pp. 30-37, 2000.

[9] Bahadur, R. R. et Ranga Rao R., "On Deviations on the Sample Mean", *Annals of Mathematical Statistics*, vol. 31, pp. 1015-1027, 1960.

[10] Barford, P. , Bestavros, A., Bradley, A. et Crovella, M., "Changes in Web client access patterns: characteristics and caching implications", *World Wide Web* , vol. 2, no. 1-2, pp. 15-28, 1999.

[11] Barford, P. et Crovella, M., "Generating representative Web workloads for

network and server performance evaluation", *SIGMETRICS '98/PERFORMANCE'98, Joint International Conference on Measurement and Modeling of Computer Systems,* Madison, WI, USA, pp. 151-160, vol. 26, June 1998.

[12] Bashforth, B. N. et Williamson, C. L., "Statistical multiplexing of self-similar video streams: simulation study and performance results", *Proceedings of MASCOTS '98: 6th International Symposium on Modeling, Analysis and Simulation of Computer and Telecom Systems,* Montreal, Que., Canada, pp. 119-126, 1998.

[13] Bean, N. G., Brown, D. R. et Taylor, P. G., "Maximal profit dimensioning and tariffing of loss networks with cross-connects", *Mathematical and Computer Modelling: Workshop on Stochastic Models in Engineering, Technology, and Management,* pp. 21-30, 2000.

[14] Beaubrun, R. et Pierre, S., "Routing and delay analysis models for high-speed networks", *Computers and Electrical Engineering*, vol. 27, no. 1, pp. 37-53, 2001.

[15] Bellcore , Reliability Prediction Procedure for Electronic Equipment, Bellcore, Technical report TR-332.

[16] Beran, J., Sherman, R., Taqqu, M. S. et Willinger, W., "Long-range dependence in variable-bit-rate video traffic", *IEEE Transactions on Communications*, vol. 43, no. 2-4, pp. 1566-1579, 1995.

[17] Bo, R. et Lowen, S., "Fractal traffic models for Internet simulation", *Proceedings of 5th IEEE Symposium on Computer and Communications (ISCC 2000),* Antibes-Juan les Pins, France, pp. 200-206, 2000.

[18] Bodamer, S. , *Charging in Multi-Service Networks*, Institute of Communication Networks and Computer Engineering, University of Stuttgart, Stuttgart, Germany, Internal Report n° 29, 1998.

[19] Borella, M. S., "Source Models of Network Game Traffic", *Computer Communications*, vol. 23, pp. 403-410, Feb. 2000.

[20] Borgonovo, F., Capone, A., Cesana, M. et Fratta, L., "Packet service in UMTS: delay-throughput performance of the downlink shared channel", *Computer Networks*, vol. 38, no. 1, pp. 43-59, Jan. 2002.

[21] Brady, P. T., "A model for generating on-off speech patterns in two-way conversation", *Bell System Technical Journal*, vol. 48, no. 7, pp. 2445-2472, 1969.

[22] CACI Products Company, *COMNET III: Reference Guide*, La Jolla CA, USA, 1998.

[23] Canonico, R., D'ardia, L. et Ventre, G., "A Simulation Based Approach for Network Resources Dimensioning in Video Delivery Systems", *Telecommunication Systems*, vol. 11, pp. 187-199, 1999.

[24] Castro , J. P., The UMTS network and radio access technology air interface techniques for future mobile systems, Chichester, England, Toronto: Wiley, 2001.

[25] Chang, Y.-C., Tse, D. et Messerschmitt, D. G., "Multimedia CDMA wireless network design: the link layer perspective", *1999 IEEE International Conference on Communications,* pp. 1421-1425, vol. 3, 1999.

[26] Clark, D. D., "A Model for Cost Allocation and Pricing in the Internet", In: *Internet Economics*, McKnight, L. W. et Bailey, J. P. (eds.), Cambridge, Mass.: MIT Press, 1997.

[27] Cocchi, R., Shenker, S., Estrin, D. et Lixia Zhang, "Pricing in computer networks: motivation, formulation, and example", *IEEE/ACM Transactions on Networking*, vol. 1, no. 6, pp. 614-627, 1993.

[28] Courcoubetis, C., Kelly, F. et Weber, R., "Measurement-based usage charges in communications networks", *Operations Research*, vol. 48, no. 4, pp. 535-548, July 2000-Aug. 2000.

[29] Courcoubetis, C., Kelly, F. P., Siris, V. A. et Weber, R., "A Study of Simple Usage-Based Charging Schemes for Broadband Networks", *Telecommunication Systems*, vol. 15, pp. 323-343, 2000.

[30] Courcoubetis, C., Siris, V. A. et Stamoulis, G. D., "Application and evaluation of large deviation techniques for traffic engineering in broadband networks", *SIGMETRICS '98/PERFORMANCE'98, Joint International Conference on Measurement and Modeling of Computer Systems,* pp. 212-221, 1998.

[31] Courcoubetis, C., Siris, V. A. et Stamoulis, G. D., "Application of the many sources asymptotic and effective bandwidths to traffic engineering", *Telecommunication Systems - Modeling, Analysis, Design and Management*, vol. 12, no. 2-3, pp. 167-191, 1999.

[32] Courcoubetis, C. et Weber, R., "Buffer overflow asymptotics for a buffer handling many traffic sources", *Journal of Applied Probability*, vol. 33, no. 3, pp. 886-903, Sept. 1996.

[33] Crovella, M. E. et Bestavros, A., "Self-Similarity in World Wide Web Traffic: Evidence and Possible Causes", *IEEE/ACM Transactions on Networking*, vol. 5, pp. 835-846, Dec. 1997.

[34] Cuomo, F. et Listanti, M., "Performance Analysis of a Prototypal Multimedia Service in an Intelligent Broadband Network", *Computer Communications*, vol. 23, pp. 341-361, Feb. 2000.

[35] DaSilva, L. A., "Pricing for QoS-enabled Networks: a Survey", *IEEE Communication Surveys and Tutorials*, vol. 3, pp. 14-20, 2000.

[36] DaSilva, L. A., Petr, D. W. et Akar, N., "Static Pricing and Quality of Service in Multiple Service Networks", *Proceedings of the 5th Joint Conf. on Information Sciences (JCIS'00),* Atlantic City (NJ), pp. 355-358, vol. 1, 2000.

[37] Dolzer, K., Payer, W. et Eberspacher, M., "A simulation study on traffic aggregation in multi-service networks", *ATM 2000, Proceedings of the IEEE Conference on High Performance Switching and Routing,* pp. 157-165, 2000.

[38] Elwalid, A. I. et Mitra, D., "Effective bandwidth of general Markovian traffic sources and admission control of high speed networks", *IEEE/ACM Transactions on Networking*, vol. 1, no. 3, pp. 329-343, 1993.

[39] Feng, G. et Yum, T. S. P., "Bifurcated-M routing for multi-point

videoconferencing", *Computer Communications*, vol. 23, pp. 362-370, 2000.

[40] Ferrari, D. , "Client requirements for real-time communication services", *IEEE Communications Magazine*, vol. 28, no. 11, pp. 65-72, 1990.

[41] Ferrari , D., *Computer systems performance evaluation*, Englewood Cliffs, N.J. : Prentice-Hall, 1978.

[42] Fishburn, P. C. et Odlyzko, A. M., "Dynamic behavior of differential pricing and quality of service options for the Internet", *Decision Support Systems, First International Conference on Information and Computation Economies (ICE '98),* pp. 123-136, 2000.

[43] Fitzek, F. et Reisslein, M., "MPEG-4 and H.263 Video Traces for Network Performance Evaluation", http://www-tkn.ee.tu-berlin.de/research/trace/trace.html, 2003.

[44] Fitzek, F. H. P. et Reisslein, M., "MPEG-4 and H.263 video traces for network performance evaluation", *IEEE Network*, vol. 15, no. 6, pp. 40-54, 2001.

[45] Gavish, B. et Sridhar, S., "Economic aspects of configuring cellular networks", *Wireless Networks*, vol. 1, no. 1, pp. 115-128, 1995.

[46] Giordano, S., Pagano, M., Pannocchia, R. et Russo, F. , "A New Call Admission Control Scheme Based On The Self Similar Nature of Multimedia Traffic", *Proc. IEEE Intl. Conf. Communications (ICC),* pp. 1612-1618, 1996.

[47] Grossglauser, M. et Bolot, J.-C., "On the relevance of long-range dependence in network traffic", *Proc. ACM SIGCOMM'96,* Stanford, CA, USA, pp. 15-24, 1996.

[48] Guerin, R., Ahmadi, H. et Naghshineh, M., "Equivalent capacity and its application to bandwidth allocation in high-speed networks", *IEEE Journal on Selected Areas in Communications*, vol. 9, no. 7, pp. 968-981, 1991.

[49] Gupta, A., Stahl, D. O. et Whinston, A. B., "Priority pricing of integrated services networks", In: *Internet Economics*, McKnight, L. W. et Bailey, J. P. (eds.), Cambridge, Mass.: MIT Press, 1997.

[50] Habib, I. et Saadawi, T., "Multimedia Traffic Characteristics in Broadband

Networks", *IEEE Communications Magazine*, vol. 30, pp. 48-54, July 1992.

[51] Hayer, J., *Transport Auction: A New Service Concept*, TRLabs Edmonton, Technical report TR-93-05, 1993.

[52] Heyman, D., Lakshman, T. V., Tabatabai, A. et Heeke, H., "Modeling teleconference traffic from VBR video coders", Conference Record - International Conference on Communications: Proceedings of the 1994 IEEE International Conference on Communications, pp. 1744-1748, vol. 3, 1994.

[53] Heyman, D. P., Tabatabai, A. et Lakshman, T. V., "Statistical analysis and simulation study of video teleconference traffic in ATM networks", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 2, no. 1, pp. 49-59, Mar. 1992.

[54] Heyman, D. P. et Lakshman, T. V., "Source models for VBR broadcast-video traffic", *IEEE/ACM Transactions on Networking*, vol. 4, no. 1, pp. 40-48, 1996.

[55] Holma, H. et Toskala, A., *WCDMA for UMTS : Radio Access for Third Generation Mobile Communications*, Chichester, Eng., New York : Wiley, 2000.

[56] Houeto, F. et Pierre, S., "Some Performance and QoS issues in Third-generation Wireless Networks", *Proceedings of the 21st Biennial Symposium on Communications,* Kingston, Canada, pp. 36-40, 2002.

[57] Houeto, F. and Pierre, S., "Jitter Characterization in Admission Control and Pricing Issues in Integrated Multiservice Networks", submitted to *IEEE/ACM Transactions on Networking*, 2003.

[58] Houeto, F., Pierre, S., Beaubrun, R. et Lemieux, Y., "Reliability and cost evaluation of third-generation wireless access network topologies: a case study", *IEEE Transactions on Reliability*, vol. 51, no. 2, pp. 229-239, 2002.

[59] Hui, J. Y., "Resource allocation for broadband networks", *IEEE Journal on Selected Areas in Communications*, vol. 6, no. 9, pp. 1598-1608, 1988.

[60] Internet Traffic Archive, "Traces In The Internet Traffic Archive", *http://ita.ee.lbl.gov/html/traces.html*, 2003.

[61] Izquierdo, M. R. et Reeves, D. S., "A survey of statistical source models for

variable-bit-rate compressed video", *Multimedia Systems*, vol. 7, no. 3, pp. 199-213, 1999.

[62] Jan, R. H., Hwang, F. J. et Cheng, S. T., "Topological Optimization of a Communication Network Subject to Reliability Constraints", *IEEE Trans. on Reliability*, vol. 42, pp. 63-70, 1993.

[63] Jelenkovic, R. P., Lazar, A. A. et Semret, N., "Multiple Time Scales and Subexponentiality in MPEG Video Streams", *Proc. of the Intl. IFIP-IEEE Conf. on Broadband Comm.*, pp. 64-75, 1996.

[64] Kalmanek, C. R., Kanakia, H. et Keshav, S., "Rate controlled servers for very high-speed networks", *Proceedings of GLOBECOM '90: IEEE Global Telecommunications Conference and Exhibition*, pp. 12-20, vol. 1, 1990.

[65] Ke, W.-J. et Wang, S.-D., "Reliability Evaluation for Distributed Computing Networks with Imperfect Nodes", *IEEE Trans. on Reliability*, vol. 46, pp. 342-349, 1997.

[66] Kelly, F. P., "On tariffs, policing and admission control for multiservice networks", *Operations Research Letters*, vol. 15, no. 1, pp. 1-9, 1994.

[67] Kelly, F. P., "Charging and Accounting for Bursty Connections", In: *Internet Economics*, McKnight, L. W. et Bailey, J. P. (eds.), Cambridge, Mass.: MIT Press, 1997.

[68] Korpeoglu, I., Tripathi, S. K. et Xiaoqiang Chen, "Estimating end-to-end cell delay variation in ATM networks", *ICCT'98: 1998 Int'l Conference on Communication Technology Proceedings*, pp. 472-483, vol. 1, 1998.

[69] Krunz, M., Sass, R. et Hughes, H., "Statistical characteristics and multiplexing of MPEG streams", *Proceedings of IEEE INFOCOM '95*, pp. 455-462, vol. 2, 1995.

[70] Krunz, M., Sass, R. et Hughes, H., "Study of VBR MPEG-coded video traffic and associated multiplexing performance", *Computer Systems Science and Engineering*, vol. 11, no. 3, pp. 135-143, 1996.

[71] Krunz, M. et Tripathi, S. K., "On the characterization of VBR MPEG streams", Performance Evaluation Review: 1997 ACM International Conference on

Measurement and Modeling of Computer Systems (SIGMETRICS 97), pp. 192-202, 1997.

[72] Kwon, T., Kapoor, R., Lee, Y. et Gerla, M., "A hybrid architecture of UMTS and Bluetooth for indoor wireless/mobile communications", *Proceedings of the International Conference on Wireless LANS and Home Networks,* pp. 273-282, 2001.

[73] Lee, D. C., Park, S. J. et Song, J. S., "Performance analysis of queueing strategies for multiple priority calls in multiservice personal communications services", *Computer Communications*, vol. 23, no. 11, pp. 1069-1083, 2000.

[74] Leland, W. E., Taqqu, M. S., Willinger, W. et Wilson, D. V., "On the self-similar nature of Ethernet traffic (extended version)", *IEEE/ACM Transactions on Networking*, vol. 2, no. 1, pp. 1-15, Feb. 1994.

[75] Lennox, J., Murakami, K., Karaul, M. et La Porta, T. F., "Interworking Internet telephony and wireless telecommunications networks", *Computer Communication Review*, vol. 31, no. 5, pp. 25-36, 2001.

[76] Likhanov, N. et Mazumdar, R. R., "Cell loss asymptotics in buffers fed with a large number of independent stationary sources", *Proceedings of IEEE INFOCOM '98,* pp. 339-346, vol. 1, 1998.

[77] Lin, Y. B., Mohan, S. et Noeopel, A., "Queueing priority channel assignment strategies for PCS hand-off and initial access", *IEEE Trans. on Vehicular Technology*, vol. 43, pp. 704-712, 1994.

[78] Lindberger, K., "Cost-based charging principles in ATM networks", . *Proceedings of the 15th International Teletraffic Congress - ITC 15,* pp. 771-780, vol. 2, 1997.

[79] Lindberger, K., "Balancing quality of service, pricing and utilisation in multiservice networks with stream and elastic traffic", *Proceedings of the International Teletraffic Congress - ITC-16,* pp. 1127-1136, vol. 2, 1999.

[80] Litjens, R. et Boucherie, R. J., "Performance analysis of fair channel sharing policies in an integrated cellular voice/data network", *Telecommunication*

Systems - Modeling, Analysis, Design and Management, vol. 19, no. 2, pp. 147-186, 2002.

[81] Liu, Z., Niclausse, N. et Jalpa-Villanueva, C., "Traffic model and performance evaluation of Web servers", *Performance Evaluation*, vol. 46, no. 2-3, pp. 77-100, 2001.

[82] Lucent Technologies, "Lucent - Home page", *http://www.lucent.com*, 2002.

[83] MacKie-Mason, J. K., Murphy, L. et Murphy, J., "The role of responsive pricing in the Internet", In: *Internet Economics*, McKnight, L. W. et Bailey, J. P. (eds.), Cambridge, Mass.: MIT Press, 1997.

[84] MacKie-Mason, J. K. et Varian, H. R., "Pricing congestible network resources", *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 7, pp. 1141-1149, 1995.

[85] MacKie-Mason, J. K. et Varian, H. R., "Economic FAQs About the Internet", In: *Internet Economics*, McKnight, L. W. et Bailey, J. P. (eds.), Cambridge, Mass.: MIT Press, 1997.

[86] Malik, S. A., Akhtar, S. et Zeghlache, D., "Performance of prioritized resource control for mixed services in UMTS WCDMA networks", *IEEE Vehicular Technology Conference: IEEE 54th Vehicular Technology Conference (VTC FALL 2001),* pp. 1000-1004, vol. 2, 2001.

[87] Mandjes, M. et Van Foreest, N., "Aspects of pricing in an integrated services network", *Proceedings of the International Teletraffic Congress - ITC-16,* pp. 1331-1340, vol. 2, 1999.

[88] Matrawy, A. , Lambadaris, L. et Changcheng Huang, "MPEG4 traffic modeling using the transform expand sample methodology", *Proceedings 2002 IEEE 4th International Workshop on Networked Appliances,* pp. 249-256, 2002.

[89] McKnight, L. W. et Bailey, J. P., "Introduction to Internet Economics", In: *Internet Economics*, McKnight, L. W. et Bailey, J. P. (eds.), Cambridge, Mass.: MIT Press, 1997.

[90] Melis, B. et Romano, G., "UMTS W-CDMA: evaluation of radio performance by

means of link level simulations", *IEEE Personal Communications*, vol. 7, no. 3, pp. 42-49, 2000.

[91] Menth, M., "Analytical performance evaluation of low-bitrate real-time traffic multiplexing in UMTS over IP networks", *Journal of Interconnection Networks*, vol. 2, no. 1, pp. 147-174, 2001.

[92] Miah, B. et Cuthbert, L. G., "Charging and Billing in ATM Networks", *Proceedings of the 6th IFIP Workshop on Performance Modeling and Evaluation of ATM Networks,* Ilkley, UK, pp. 92\1-92\9, 1998.

[93] Mmaduka Ufongene, C., "A model for the cost analysis of wireless access architectures", *Bell Labs Technical Journal*, vol. 4, no. 3, pp. 134-154, 1999.

[94] Montgomery, M. et De Veciana, G., "On the relevance of time scales in performance oriented traffic characterizations", *Proceedings IEEE INFOCOM '96,* pp. 513-520, vol. 2, 1996.

[95] Murphy, J., Murphy, L. et Posner, E. C., "Distributed pricing for embedded ATM networks", *Proceeding of the 14th International Teletraffic Congress - ITC 14,* pp. 1053-1063, vol. 2, 1994.

[96] Network Reliability and Interoperability Council, "New Wireline Access Technologies Subteam Final Report ", http://www.nric.org/pubs/nric2/fg3/index.html, 2003.

[97] Newport, K. T. et Varshney, P. K., "Design of survivable communications networks under performance constraints", *IEEE Transactions on Reliability*, vol. 40, no. 4, pp. 433-440, 1991.

[98] Norros, I., "A storage model with self-similar input", *Queueing Systems Theory and Applications*, vol. 16, no. 3-4, pp. 387-396, 1994.

[99] Norros, I., "On the use of fractional Brownian motion in the theory of connectionless networks", *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 6, pp. 953-962, Aug. 1995.

[100] Odlyzko, A., "Paris Metro pricing: the minimalist differentiated services solution", *1999 Seventh International Workshop on Quality of Service,*

IWQoS'99, pp. 159-161, 1999.

[101] Park, K., Kim, G. et Crovella, M., "On the Relation between File Sizes, Transport Protocols, and Self-Similar Network Traffic", *Proc. IEEE Int'l Conf. Network Protocols,* pp. 171-180, 1996.

[102] Park, K., Kim, G. et Crovella, M., "On the Effect of Traffic Self-Similarity on Network Performance", *Proc. SPIE Int'l Conf. Perf. And Control of Network Sys,* pp. 296-310, 1997.

[103] Parris, C. , Keshav, S. et Ferrari, D., *A framework for the study of pricing in integrated networks*, International Computer Science Institute, Berkeley, CA, Technical Report TR-92-016, Mar. 1992.

[104] Pavlzdou, F. N., "Mixed media cellular systems", *IEEE Trans. on Communications*, vol. 42, pp. 1505-1511, 1994.

[105] Paxson, V. , "Empirically derived analytic models of wide-area TCP connections", *IEEE/ACM Transactions on Networking*, vol. 2, no. 4, pp. 316-336, 1994.

[106] Paxson, V. , "Growth trends in wide-area TCP connections", *IEEE Network*, vol. 8, no. 4, pp. 8-17, July 1994-Aug. 1994.

[107] Paxson, V. et Floyd, S., "Wide area traffic: the failure of Poisson modeling", *IEEE/ACM Transactions on Networking*, vol. 3, no. 3, pp. 226-244, June 1995.

[108] Paxson, V. et Floyd, S., "Why we don't know how how simulate the Internet", *Proceedings of 1997 Winter Simulation Conference,* Atlanta, GA, USA, pp. 1037-1044, 1997.

[109] Pierre, S. et Beaubrun, R., "Integrating Routing and Survivability in Fault-Tolerant Computer Network Design", *Computer Communications*, vol. 23, pp. 317-327, Feb. 2000.

[110] Pierre, S. et Gharbi, I., "A Generic Object-Oriented Model for Representing Computer Network Topologies", *International Journal of Advances in Engineering Software*, vol. 32, pp. 95-110, 2001.

[111] Pierre, S. et Legault, G., "A genetic algorithm for designing distributed computer

network topologies", *IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics)*, vol. 28, no. 2, pp. 249-258, 1998.

[112] Pierre, S. et Elgibaoui, A., "Tabu-search approach for designing computer-network topologies with unreliable components", *IEEE Transactions on Reliability*, vol. 46, no. 3, pp. 350-359, 1997.

[113] Quessette, F., Troubnikoff, A. et Valois, F., "Modeling and analysis of UMTS hierarchical networks", *Proceedings 8th International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems,* pp. 432-437, 2000.

[114] Roberts, J. W., "Quality-of-service guarantees and charging in multiservice networks", *IEICE Transactions on Communications*, vol. E81-B, no. 5, pp. 824-831, 1998.

[115] Rose, O., "Discrete-time analysis of a finite buffer with VBR MPEG video traffic input", *Teletraffic Contributions for the Information Age. Proceedings of the 15th International Teletraffic Congress - ITC 15,* pp. 413-422, vol. 1, 1997.

[116] Sahinoglu, Z. et Tekinay, S., "On multimedia networks: self-similar traffic and network performance", *IEEE Communications Magazine*, vol. 37, no. 1, pp. 48-52, 1999.

[117] Saito, H. et Tsuchiya, T., "Upper bound of loss probability for self-similar traffic", *Proceedings of ICC/SUPERCOMM '96 - International Conference on Communications,* Dallas, TX, USA, pp. 1624-1629, vol. 3, 1996.

[118] Sanchez, J. et Thioune, M., *UMTS services, architecture et WCDMA*, Paris: Hermes Science, 2001.

[119] Siris, V. A., *Performance Analysis and Pricing in Broadband Networks*, These de doctorat, 1997, University of Crete, Heraklion, Crete.

[120] Siris, V. A., "Some FAQs for the *s*, *t* parameters", http://www.ics.forth.gr/netgroup/msa/faq_s_t.html, 2003.

[121] Siris, V. A., Songhurst, D. J., Stamoulis, G. D. et Stoer, M., "Usage-based charging using effective bandwidths: studies and reality", *Proceedings of 16th*

International Teletraffic Congress (ITC-16), Edinburgh, UK, pp. 929-940, vol. 2, 1999.

[122] Stathis, C. et Maglaris, B., "Modelling and call admission control for self-similar WAN traffic", *Proceedings of the Applied Telecommunications Symposium (ATS'99),* pp. 60-65, 1999.

[123] Tsybakov, B. et Georganas, N. D., "Overflow Probability in an ATM Queue with Self-Similar Traffic", *Proceedings of IEEE International Conference on Communications (ICC'97),* Montreal, (Qc), Canada, pp. 822-826, vol. 2, 1997.

[124] UMTS Forum, "UMTS Forum Web Page", *http://www.umts-forum.org*, 2002.

[125] Visual Communications Lab, "Video traces", *http://viscom.kaist.ac.kr/video-trace.html*, 2003.

[126] Wang, Y., Williamson, C. et Doerksen, J., "CAC performance with self-similar traffic: simulation study and performance results", *Proceedings of the Seventh Int'l Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems - MASCOTS '99,* pp. 102-111, 1999.

[127] WAP Forum, "WAP Forum Specifications", http://www.wapforum.org/what/technical.htm, 2002.

[128] Yang, S. C., *CDMA RF system engineering*, Boston : Artech House, 1998.

ANNEXE I

Opnet : aperçu général

Opnet est un simulateur par événements. Le logiciel est organisé en trois niveaux : le niveau du réseau, le niveau du nœud et le niveau des processus. Un objet au niveau réseau est une instanciation d'un modèle de nœud tandis qu'un nœud est lui aussi une instanciation d'un modèle de processus. À l'intérieur d'un nœud, les processus peuvent communiquer par des canaux de paquets et de statistiques et sont organisés selon les différentes couches du modèle OSI.

L'usager spécifie le comportement au niveau processus par une machine à états finis et des blocs de code associés à chaque état. Ces blocs définissent le comportement du processus lorsqu'il se trouve dans un état donné. Le noyau du simulateur peut envoyer à chaque instance de processus des interruptions qui correspondent aux événements. Ces événements sont gardés dans une liste et classés par ordre chronologique. À chaque événement est donc associé un temps qui permet de faire avancer l'horloge de la simulation.

Module UMTS

Le modèle UMTS de Opnet est basé sur la version de 1999 des standards du groupe 3GPP.

Quand un usager allume son UE, le modèle assume que la synchronisation et une connexion de signalisation à commutation de paquets est établie. Cette connexion de signalisation est gardée durant toute la durée de la connexion. De ce fait, l'usager peut immédiatement accomplir une procédure de « GPRS attach » avec le SGSN et ainsi accéder aux services GPRS.

Les paquets reçus des couches supérieures sont mis dans une file d'attente. Comme chaque usager supporte quatre profils de QoS, le trafic est donc dirigé vers l'une des quatre files de QoS. Si aucun contexte PDP n'avait été activé, une demande d'activation de contexte PDP est envoyée au SGSN. Ce message d'activation contient le

profil de QoS demandé par l'usager. Le modèle assume que le SGSN, après avoir consulté le RNC, accorde la totalité de la QoS demandée ou rejette la demande. Ainsi, le profil de QoS n'est pas négocié dans le modèle.

À la réception de la demande d'activation de contexte PDP, le SGSN envoie une demande d'assignation de RAB au RNC avec le profil de QoS demandé. L'UTRAN détermine alors, à l'aide de l'algorithme d'admission de connexion si la requête sera acceptée. Si les liaisons montante et descendante ont suffisamment de ressources, la demande est acceptée et dans ce cas, le RNC envoie une demande d'établissement de support radio à l'UE. À la réception de la requête de demande d'établissement de support radio (RAB), l'UE établit le canal tel que spécifié dans la requête et envoie une confirmation de support radio (Radio Bearer Complete) au RNC. Quand il reçoit la confirmation, le RNC envoie une réponse d'assignation de RAB incluant le profil QoS accordé au SGSN. Le SGSN envoie alors à l'UE le message d'acceptation de l'activation de contexte PDP, qui inclut aussi le niveau de QoS accordé.

L'UE peut dès lors envoyer des paquets à la destination. Avant d'atteindre leur destination les paquets sont d'abord tunellisés à travers le RNC et le SGSN/GGSN et ensuite routés à travers le réseau IP. Si le réseau de destination est aussi un réseau UMTS, alors ils sont mis dans une file au niveau du nœud SGSN/GGSN de destination et sont transférés à la destination dès qu'un canal est établi vers ce dernier.

La couche GMM contient des fonctions des couches GMM, GSM et RRC. Il contient des fonctions de gestion de mobilité (comme le « GPRS attach »), de gestion de session (l'activation de contexte PDP) et de contrôle des ressources radio (l'établissement et la libération des supports radio). La couche RLC/MAC inclut la gestion de la priorité des flux de données, les trois modes RLC ainsi que la segmentation/réassemblage des paquets des couches supérieures.

Les liens entre l'émetteur/receveur radio et la couche RLC/MAC constituent les canaux de transport. Sur la liaison montante, on a un canal d'accès aléatoire (RACH), un canal commun de paquet (CPCH), un canal dédié (DCH) qui véhicule les données et la signalisation. Chaque canal dédié de transport a un code d'étalement unique qui le

distingue des autres canaux de transport. Sur la liaison descendante, il y a un canal d'accès direct (FACH), un canal d'accès partagé (DSCH), un canal indicateur d'acquisition (AICH), un canal dédié de signalisation par usager et jusqu'à quatre canaux de données. Un code d'étalement différent est assigné à chaque canal et le trafic peut être envoyé simultanément sur tous les canaux.

La couche GMM contient 4 files d'attente, une pour chaque classe de QoS. Quand un paquet de la couche application arrive au niveau de la couche GMM, il est transféré à la couche RLC/MAC si un canal avait déjà reçu un message d'établissement de RAB pour le support radio de la classe de QoS du paquet. Sinon, le paquet est mis, au niveau de la couche GMM, dans la file correspondant à son profil de QoS. La couche RLC/MAC utilise les files pour transmettre les paquets provenant des couches supérieures, retransmettre les paquets dans le mode RLC acquitté, recevoir les paquets provenant des couches inférieures et les rassembler pour construire les PDU. Pour chacune de ces fonctions, on a une file pour le trafic de signalisation et quatre pour chaque classe de QoS du trafic de données.

Modélisation radio

Dans Opnet, le lien radio n'est pas une entité représentée physiquement. Il naît de la mise en présence de deux entités dotées de possibilités de communication radio. Les caractéristiques du lien radio peuvent varier dans le temps en fonction de facteurs tels la mobilité, la modification des attributs du transmetteur ou du receveur et les interférences provenant des autres transmissions simultanées.

Le logiciel utilise un modèle en 13 étapes pour modéliser le lien radio. Comme ce dernier est essentiellement un médium de diffusion, chaque transmission peut potentiellement affecter plusieurs receveurs. En outre, pour une transmission donnée, le lien vers chaque receveur peut avoir des caractéristiques différentes. De ce fait, la série des 13 étapes est exécutée pour chaque récepteur éligible.

Dans un premier temps, le délai de transmission est calculé à partir du débit du canal et de la longueur du paquet. Ensuite, on évalue la connexité entre toutes les paires

d'émetteur/récepteur radio en présence. Cette évaluation se fait sur la base du modèle de ligne de vue (line of sight). Si le segment reliant l'émetteur au récepteur coupe la surface terrestre, il n'y a pas de connexité. À la troisième étape, on détermine la compatibilité entre l'émetteur et le récepteur en considérant les attributs de chaque extrémité de la liaison radio que sont la fréquence, la largeur de bande, le débit, le code d'étalement et la modulation. Si les bandes de transmission des deux extrémités ne se recoupent pas, les paquets sont tout simplement ignorés. Par contre, s'il y a recoupement des bandes sans qu'il y ait une correspondance totale des autres attributs des extrémités de la liaison, les paquets sont considérés comme de l'interférence. À ce titre, ils sont pris en compte par la modélisation du lien puisqu'ils affectent les paquets valides. Le gain de transmission est ensuite calculé. Il est nul pour les antennes omnidirectionnelles.

L'étape suivante calcule le délai de propagation du paquet selon la distance séparant l'émetteur du récepteur et selon la vitesse de propagation du signal qui est considérée comme étant la vitesse de la lumière dans le vide. À l'étape 6, on calcule le gain de réception. Le niveau moyen de chaque signal pertinent (y compris les signaux d'interférence) reçu par le canal de réception radio est ensuite calculé et servira plus tard à évaluer le rapport signal/bruit et à déduire le taux d'erreur par bloc. Ce calcul inclut des modèles d'affaiblissement de propagation et d'évanouissement qui dépendent du type d'environnement. Le modèle (d'affaiblissement) de propagation est basé sur les spécifications 3GPP TR 25.942 inspirées du modèle COST 231. Ce modèle tient compte du type d'environnement. Par exemple, dans un environnement urbain intérieur, l'atténuation maximale est donnée par la formule :

$$L_p = 30\log_{10}(1000R) = 18.3n^{\left(\frac{n+2}{n+1}-0.46\right)} + 37$$

où R est la distance en kilomètres entre l'usager et la station de base, n est le nombre d'étages traversés par le signal.

Au niveau suivant, on calcule le bruit de fond qui prend en compte les sources de bruit qui ne sont pas explicitement modélisées tels le bruit thermal, le bruit urbain et le bruit provenant de l'environnement du receveur radio. On passe ensuite au calcul du

niveau d'interférence créée par des transmissions concurrentes. Dans le modèle actuel, Opnet ne calcule pas les interférences « intra et extra-cellules ». Le rapport signal/bruit est calculé comme le rapport de la puissance moyenne du signal à la puissance moyenne cumulée de toutes les sources d'interférence et de bruit de fond. Ce rapport est déterminé pour chaque segment de paquet envoyé pendant une période où le nombre de sources d'interférence demeure constant. À la fin de la réception de chaque segment, on déduit le taux d'erreur par bloc en fonction du schéma et du taux de codage et du TTI de chaque canal de transport. Les taux d'erreur des différents segments sont ensuite utilisés pour calculer le nombre total de bits erronés par paquet. Les paquets qui ont pu être complètement transmis sont acceptés si le taux d'erreur du paquet est inférieur au seuil de correction du receveur. Dans ce cas, ils sont transmis à la couche supérieure pour la suite du traitement.

ANNEXE II

Borne de Chernoff

Soit $S = \sum_{k=1}^{K} X_k$, où X_1, ..., X_K sont des variables aléatoires indépendantes et non négatives. Alors la borne de Chernoff pour S est $P\{S \geq C\} \leq E[e^{s(S-C)}]$, où $s \geq 0$.

Une application directe de la borne de Chernoff est : $\log P[S \geq C] \leq s(\sum_k \alpha_k(s) - C)$ où $\alpha_k(s) = s^{-1} \log E[e^{sX_k}]$.

Conditions de régularité du taux d'arrivée des paquets

Pour les résultats de la section 5.1, les conditions de régularité sont :

- le temps est discrétisé et segmenté;
- à chaque instant discrétisé n, un nombre fini $\lambda_{n,j}$ de cellules est émis par la source j;
- Le tampon de taille NB est fini et se vide au taux constant NC;
- Si le tampon est plein, alors les cellules supplémentaires sont perdues;
- $\{\lambda_{n,j}\}$ est un processus stationnaire avec $E[\lambda_{n,j}] < C$.

Théorème de Bahadur-Rao

On considère N sources indépendantes, identiques, stationnaires et ergodiques, chacune générant des paquets au taux $\lambda_{n,j}$ où n se réfère au temps et j à la source respectant les propriétés citées précédemment.

Le nombre total de paquets générés par la source j durant l'intervalle de temps $[0,t)$ est $X_{t,j} = \sum_{n=0}^{t-1} \lambda_{n,j}$ et le nombre total de paquets générés par toutes les sources est $X_t^{(N)} = \sum_{k=1}^{N} X_{t,k}$.

Soit $M_{t,1}(s)$ ou tout simplement $M_t(s)$ la fonction génératrice de moment de $X_{t,1}$, i.e. $M_t(s) = M_{t,1}(s) = E[e^{sX_{t,1}}]$. Supposons que les sources accèdent à un tampon de taille NB avec un débit de sortie de NC. Bahadur et Rao [9] ont montré le théorème suivant :

Théorème

Pour $N \to \infty$, pour tout $B>0$

$$P\left\{X_t^{(N)}(C) > N(Ct+B)\right\} = \frac{1}{\sqrt{2\pi N\sigma^2 s_0^2}} e^{-NI_{t,s}(C,B)}\left(1+O\left(\frac{1}{N}\right)\right)$$

où $-I_{t,s_0}(C,B) = \ln(M_t(s_0)) - (Ct+B)s_0$, s_0 est l'unique solution de $\frac{M_t'(s_0)}{M(s_0)} = (Ct+B)$ et $\sigma^2 = \frac{\partial^2}{\partial s^2}\ln\left(E[e^{sX_{t,1}}]\right) = \frac{M_t''(s_0)}{M(s_0)} - (Ct+B)^2$.

Formule de perte asymptotique à la limite de Courcoubetis et Weber

La charge stationnaire de travail $W^{(N)}(C)$ est donnée par : $W^{(N)}(C) = \sup_{t\in\{1,\ldots\}}\left(X_{-t}^{(N)} - NCt\right)$ où $X_{-t}^{(N)}$ désigne le nombre total de cellules qui arrivent dans l'intervalle *(-t,0]*. Ici, on utilise les propriétés de stationnarité pour remplacer l'intervalle *(0,t]* par l'intervalle *(-t,0]*. En effet alors, $X_{-t}^{(N)}$ a la même distribution que $X_t^{(N)}$.

La formule de Courcoubetis et Weber donne le taux de perte asymptotique à l'infini et s'écrit :

$$\lim_{N\to\infty}\frac{1}{N}\ln(P\{W^{(N)}(C) > NB\}) = -\inf_t \sup_s [s(B+Ct) - t\rho_t(s)] = -I_{t_o,s_o}(C,B)$$

Preuve

Puisque $X_{-t}^{(N)}$ a la même distribution que $X_t^{(N)}$, alors :

$P\left\{X_{t_0}^{(N)} > (Ct_0+B)N\right\} \le P\left\{W^{(N)}(C) > NB\right\}$ par définition de $W^{(N)}$

$$\le \sum_{t=1}^{\infty} P\left\{X_t^{(N)} > (Ct+B)N\right\} \text{ par sommation des probabilités} \qquad \text{(A.1).}$$

Définissons $\rho_t(s)=\frac{1}{t}\ln(M_t(s))$ et $\rho(s)=\lim_{t\to\infty}\rho_t(s)$. Choisissons θ_1 tel que $C\theta_1-\rho_t(\theta_1)>2\varepsilon$. Puisque $\rho(s_1)=\lim_{t\to\infty}\rho_t(s_1)$, alors il existe un t_1 tel que $\forall t>t_1$, on ait $Cs_1-\rho_t(s_1)>\varepsilon$ et $s_1B+t\varepsilon>\sup_s[sC-\rho_1(s)]$.

Une borne de Chernoff pour tout $s>0$ est :

$$P\left\{X_t^{(N)}>(Ct+B)N\right\}\le e^{-NI_t(C,B)}=e^{-N((Ct+B)s-\ln(M_t(s)))}$$

$$=e^{-N((Ct+B)s-t\rho_t(s))}$$

$$P\left\{W^{(N)}(C)>NB\right\}\le\sum_{t=1}^{\infty}P\left\{X_t^{(N)}>(Ct+B)N\right\}$$

$$=\sum_{t=1}^{t_1-1}P\left\{X_t^{(N)}>(Ct+B)N\right\}+\sum_{t=t_1}^{\infty}P\left\{X_t^{(N)}>(Ct+B)N\right\}$$

$$\le\sum_{t=1}^{t_1-1}e^{-N\sup_s[s(B+Ct)-t\rho_t(s)]}+\sum_{t=t_1}^{\infty}e^{-N(Bs_1+t(Cs_1-\rho_t(s_1)))}$$

$$\le\sum_{t=1}^{t_1-1}e^{-N\sup_s[s(B+Ct)-t\rho_t(s)]}+\sum_{t=t_1}^{\infty}e^{-N(Bs_1+t\varepsilon)}$$

$$\le(t_1-1)e^{-N\min_{t<t_1}\sup_s[s(B+Ct)-t\rho_t(s)]}+\frac{e^{-N(Bs_1+t_1\varepsilon)}}{1-e^{-N\varepsilon}}$$

Par conséquent, $\limsup_{N\to\infty}\frac{1}{N}\ln(P\left\{W^{(N)}(C)>NB\right\})\le-\min_{t<t_1}\sup_s[s(B+Ct)-t\rho_t(s)]$

$$\le-\inf_t\sup_s[s(B+Ct)-t\rho_t(s)].$$

Pour la borne inférieure, on a :

$P\left\{W^{(N)}(C)>NB\right\}\ge P\left\{X_t^{(N)}>(Ct+B)N\right\}$ par définition de $W^{(N)}$ et alors

$\forall t>1,\ P\left\{W^{(N)}(C)>NB\right\}\ge P\left\{X_t^{(N)}>(Ct+B)N\right\}/t$.

En utilisant la borne de Cramer pour la somme de N variables i.i.d. $\liminf_{N\to\infty} \frac{1}{N}\ln(P\{W^{(N)}(C) > NB\}) \geq -\sup_s[s(B+Ct) - t\rho_t(s)] \quad \forall m$. D'où

$$\lim_{N\to\infty} \frac{1}{N}\ln(P\{W^{(N)}(C) > NB\}) = -\inf_t \sup_s[s(B+Ct) - t\rho_t(s)] = -I_{t_0,s_0}(C,B)$$

Formule de perte asymptotique de cellules de Likhanov et Mazumdar

La formule de Likhanov et Mazumdar donne une expression asymptotique exacte de la probabilité de débordement d'un tampon multiplexant un grand nombre de sources indépendantes et stationnaires et constitue une généralisation de la formule précédente proposée par Courcoubetis et Weber.

Perte asymptotique

Si on garde les mêmes notations que précédemment et si on suppose en plus des conditions de régularité que :

a) il existe un unique $t_0 < \infty$ tel que

$$-I_{t_0,s_0}(C,B) = \ln(M_{t_0}(s_0)) - (Ct_0 + B)s_0 = \sup_t \inf_s [\ln(M_t(s)) - (Ct+B)s]$$

b) $$\liminf_{t\to\infty} \frac{I_{t,s}(C,B)}{\log(t)} > 0,$$

alors, si $N \to \infty$,

$$P\{W^{(N)}(C) > NB\} = \frac{1}{\sqrt{2\pi N\sigma^2 s_0^2}} e^{-NI_{t_0,s_0}(C,B)}\left(1 + O\left(\frac{1}{N}\right)\right) \tag{5.1}$$

où s_0, t_0 sont les points critiques (points qui maximisent ou minimisent) de :

$$-I_{t_0,s_0}(C,B) = \ln(M_{t_0}(s_0)) - (Ct_0 + B)s_0 = \sup_t \inf_s [\ln(M_t(s)) - (Ct+B)s] \tag{5.2}$$

et

$$\sigma^2 = \frac{\partial^2}{\partial s^2}\ln\left(E[e^{sX_{t,1}}]\right) = \frac{M''_{t_0}(s_0)}{M(s_0)} - (Ct_0 + B)^2 \tag{5.3}$$

Preuve :

La minimisation par rapport à s est déjà faite et on connaît s_o. Puisque les notations et les conditions sont les mêmes, on récrit l'équation (A.1)

$$P\left\{X_{t_0}^{(N)} > (Ct_0 + B)N\right\} \leq P\left\{W^{(N)}(C) > NB\right\} \leq \sum_{t=1}^{\infty} P\left\{X_t^{(N)} > (Ct + B)N\right\}.$$

En appliquant la borne de Chernoff, aux termes de la sommation, on obtient pour chaque terme, $P\left\{X_t^{(N)} > (Ct + B)N\right\} \leq e^{-NI_t(C,B)}$.

De plus, par unicité de t_o, il existe $\varepsilon > 0$ tel que $I_{t_o,s_o}(C,B) + \varepsilon \leq I_{t,s_0}(C,B)$ pour tout $t \neq t_0$. De ce fait, dans [76], on montre que par application de l'amélioration de Bahadur-Rao pour de grands N que

$$\frac{P\left\{X_t^{(N)} > (Ct + B)N\right\}}{P\left\{X_{t_0}^{(N)} > (Ct_0 + B)N\right\}} \sim O(e^{-N\varepsilon}) \qquad \text{(A.2)}.$$

Maintenant, l'hypothèse $\liminf_{t\to\infty} \frac{I_{t,s_0}(C,B)}{\log(t)} > 0$ implique qu'il existe $t_1 > t_0$ et $\alpha > 0$ tel que, pour tout $t \geq t_1$, on a :

$$I_{t,s_0}(C,B) > \alpha \log t > I_{t_0,s_0}(C,B) \text{ par définition de } t_0 \qquad \text{(A.3)}.$$

Par conséquent pour $N > 1/\alpha$, on a :

$$\sum_{t=1}^{\infty} P\left\{X_t^{(N)} > (Ct + B)N\right\} = \sum_{t=1}^{t_1} P\left\{X_t^{(N)} > (Ct + B)N\right\} + \sum_{t=t_1+1}^{\infty} P\left\{X_t^{(N)} > (Ct + B)N\right\}$$

$$\leq P\left\{X_{t_0}^{(N)} > (Ct_0 + B)N\right\}\left(1 + t_1 e^{-N\varepsilon} + \frac{e^{(-N\alpha+1)\log(1+t_1)}}{N\alpha - 1}\right)$$

En utilisant encore l'amélioration de Bahadur-Rao, pour $P\left\{X_{t_0}^{(N)} > (Ct_0 + B)N\right\}$, on a :

$$P\left\{W^{(N)}(C) > NB\right\} \leq \frac{1}{\sqrt{2\pi N\sigma^2 s_0^2}} e^{-NI_{t_o,s_o}(C,B)}\left(1 + O\left(\frac{1}{N}\right)\right)\left(1 + t_1 e^{-N\varepsilon} + \frac{e^{(-N\alpha+1)\log(1+t_1)}}{N\alpha - 1}\right).$$

Par conséquent pour de grands N, le second terme entre parenthèses à droite est $1 + O(e^{-N})$ et ainsi le terme $O(1/N)$ domine et on a :

$$P\{W^{(N)}(C) > NB\} \le \frac{1}{\sqrt{2\pi N\sigma^2 s_0^2}} e^{-NI_{t_0,s_0}(C,B)}\left(1+O\left(\frac{1}{N}\right)\right).$$

Donc $P\{X_{t_0}^{(N)} > (Ct_0 + B)N\} \le P\{W^{(N)}(C) > NB\} \le \frac{1}{\sqrt{2\pi N\sigma^2 s_0^2}} e^{-NI_{t_0,s_0}(C,B)}\left(1+O\left(\frac{1}{N}\right)\right)$

et par conséquent $P\{W^{(N)}(C) > NB\} = \frac{1}{\sqrt{2\pi N\sigma^2 s_0^2}} e^{-NI_{t_0,s_0}(C,B)}\left(1+O\left(\frac{1}{N}\right)\right)$

Lien avec la borne de Chernoff et les conditions d'admission

Les deux preuves précédentes fournissent le lien entre la borne de Chernoff et les formules de perte asymptotique. Pour aboutir à la condition d'admission (5.6), il suffit d'utiliser la formule de perte de Courcoubetis et Weber. Comme

$$\lim_{N\to\infty} \frac{1}{N} \ln(P\{W^{(N)}(C) > NB\}) = -\inf_t \sup_s [s(B+Ct) - t\rho_t(s)] = -I_{t_0,s_0}(C,B)$$

si on veut que $P\{W^{(N)}(C) > NB\} \le e^{-N\gamma_0}$, alors on a comme condition d'admission :

$$e^{-NI_{t_0,s_0}} \le e^{-N\gamma_0} \Rightarrow \ln(M_{t_0}(s_0)) - (Ct_0 + B)s_0 \le -\gamma_0$$

Nous rappelons que le débit effectif est : $\alpha(t,s) = \frac{1}{s}\rho_t(s) = \frac{1}{st}\ln(M_t(s))$. De ce fait, la condition d'admission s'écrit :

$$\alpha(s_0,t_0) \le C + \frac{1}{t_0}(B - \frac{\gamma_0}{s_0}) \qquad \text{(A.4).}$$

Pour aboutir à l'inéquation (5.6), il suffit de considérer le cas où on multiplexe plusieurs types de sources. Supposons donc que le processus d'arrivée à une liaison haut débit soit la superposition de J types de sources indépendantes identiquement distribuées. Soit $N_i=Nn_i$, $i=1,\ldots, J$ le nombre de sources de type i, et soit $n=(n_1, \ldots, n_i, \ldots, n_J)$ (les n_i ne sont pas nécessairement entiers). La liaison est desservie par un tampon NB au taux NC. Le paramètre N est un facteur d'échelle qui reflète la taille du système. Alors si on « définit » une *source de base* comme la superposition de n_i sources de type i, $i=1,\ldots, J$, le système revient donc à une superposition de N sources de base. Une

source de base est caractérisée par le débit effectif $\sum_{i=1}^{J}\alpha_i(s,t)$, où $\alpha_i(s,t)$ est le débit effectif d'une source de type *i*. On retrouve donc en récrivant (A.4)

$$\sum_{j=1}^{J} n_j \alpha_j(s_0,t_0) \le C + \frac{1}{t_0}(B - \frac{\gamma_0}{s_0})$$

Pour obtenir l'inéquation (5.7), la démarche est exactement la même, à la seule différence qu'on utilise la formule asymptotique de Likhanov et Mazumdar au lieu de celle de Courcoubetis et Weber.

Interprétation des paramètres *s* et *t*

Les interprétations fournies sont extraites de [31,120]. Nous rappelons que si nous définissons $X_{t,j}$ comme le trafic total généré par une source de type *j* durant l'intervalle *[0,t)* et $M_{t,j}(s)$ comme la fonction génératrice des moments de $X_{t,j}$, i.e. $M_{t,j}(s) = E[e^{sX_{t,j}}]$, où *E* est l'espérance, alors, le *débit effectif* d'une source de type *j* est défini comme suit :

$$\alpha_j(s,t) = \frac{1}{st}\ln E\left[e^{sX_{t,j}}\right] = \frac{1}{st}\ln(M_{t,j}(s))$$

Les paramètres *s* et *t* sont définis par le contexte de la source, i.e. les caractéristiques du trafic multiplexé, ses exigences de QoS ainsi que les ressources (capacité et tampon) disponibles. À partir de l'équation (5.2), on déduit que le paramètre *t* (ou plus précisément t_0), mesuré par exemple en millisecondes, correspond à la longueur la plus probable de la période d'occupation du tampon avant un débordement. Par conséquent, *t* indique les échelles de temps qui influencent le débordement du tampon. En effet, $\alpha_j(s,t)$ dépend de la charge produite par la source durant la période *t*. Ainsi, une petite valeur de *t* indique que les variations rapides du trafic sont responsables du débordement du tampon. En outre, le paramètre *t* montre la granularité minimale à considérer pour les traces afin d'être en mesure de saisir les propriétés statistiques qui affectent le débordement du tampon. Le paramètre *s* est un paramètre *spatial* (mesuré par exemple en kb^{-1}), qui donne une indication du degré de multiplexage et dépend entre

autres du rapport des débits de crête des différentes sources multiplexées au débit de la liaison. En particulier pour une liaison de capacité beaucoup plus grande que les débits de crête des sources multiplexées, s tend vers zéro et $\alpha_j(s,t)$ tend vers le débit moyen de la source. Par contre, lorsque le lien n'a pas une très grande capacité par rapport aux débits de crête des sources, $\alpha_j(s,t)$ se rapproche de la valeur maximale de la variable aléatoire $X_j[0,t]/t$.

Pour un flux de trafic donné, le débit effectif $\alpha_j(s,t)$ est une fonction où l'on doit choisir les paramètres s et t de manière à fournir une mesure correcte de l'utilisation effective due à la source pour le point d'opération choisi. Bien que la valeur du point d'opération dépende de la source individuelle, on peut en pratique l'ignorer pour les systèmes qui multiplexent beaucoup de sources.

Mathématiquement, posons $P\{W^{(N)}(C) > NB\} = e^{-N\gamma_0}$. En utilisant la formule de perte de Courcoubetis et Weber, on a :

$$e^{-NI_{t_0,s_0}} = e^{-N\gamma_0} \Rightarrow \ln(M_{t_0}(s_0)) - (Ct_0 + B)s_0 = -\gamma_0$$

En dérivant l'équation précédente, on a :

$$s_0 = \frac{\partial \gamma_0}{\partial B} \text{ et } s_0 t_0 = \frac{\partial \gamma_0}{\partial C}$$

Ainsi, le paramètre s_0 est égal au taux auquel le logarithme de la probabilité de débordement décroît en fonction de la taille du tampon pour une capacité donnée. D'autre part, le produit s_0t_0 est égal au taux auquel le logarithme de la probabilité de débordement décroît en fonction de la capacité de la liaison pour un tampon de taille donnée.