POLYPUBLIE

Gestion de la mobilité et ingénierie de trafic en conception des réseaux mobiles de troisième génération

Ronald Beaubrun

Document en libre accès dans PolyPublie
Open Access document in PolyPublie

Date	2002
Type	Mémoire ou thèse / Dissertation or Thesis Thèse de doctorat/Ph.D. thesis
URL de PolyPublie PolyPublie URL	https://publications.polymtl.ca/7094/
Conditions d'utilisation Terms of use	© Tous droits réservés / All rights reserved
Référence Citation	Beaubrun, R. (2002). Gestion de la mobilité et ingénierie de trafic en conception des réseaux mobiles de troisième génération [Thèse de doctorat, École Polytechnique de Montréal]. PolyPublie. https://publications.polymtl.ca/7094/

Détails de la thèse
Thesis details

Directeurs de recherche Advisors	Samuel Pierre, & Jean Conan
Programme Program	Génie électrique

Ce fichier a été téléchargé à partir de PolyPublie, le dépôt institutionnel de Polytechnique Montréal
This file has been downloaded from PolyPublie, the institutional repository of Polytechnique Montréal
https://publications.polymtl.ca

UNIVERSITÉ DE MONTRÉAL

GESTION DE LA MOBILITÉ ET INGÉNIERIE DE TRAFIC EN CONCEPTION DES RÉSEAUX MOBILES DE TROISIÈME GÉNÉRATION

RONALD BEAUBRUN
DÉPARTEMENT DE GÉNIE ÉLECTRIQUE
ÉCOLE POLYTECHNIQUE DE MONTRÉAL

THÈSE PRÉSENTÉE EN VUE DE L'OBTENTION
DU DIPLÔME DE PHILOSOPHIAE DOCTOR (Ph.D)
(GÉNIE ÉLECTRIQUE)
NOVEMBRE 2002

© Ronald Beaubrun, 2002.

National Library
of Canada

Acquisitions and
Bibliographic Services

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale
du Canada

Acquisisitons et
services bibliographiques

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence
ISBN: 0-612-81713-X
Our file Notre référence
ISBN: 0-612-81713-X

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou aturement reproduits sans son autorisation.

Canada

UNIVERSITÉ DE MONTRÉAL

ÉCOLE POLYTECHNIQUE DE MONTRÉAL

Cette thèse intitulée :

GESTION DE LA MOBILITÉ ET INGÉNIERIE DE TRAFIC EN CONCEPTION DES RÉSEAUX MOBILES DE TROISIÈME GÉNÉRATION

présentée par : BEAUBRUN, Ronald

en vue de l'obtention du diplôme de : Philosophiae Doctor

a été dûment acceptée par le jury d'examen constitué de:

M. PESANT, Gilles, Ph.D., président

M. PIERRE, Samuel, Ph.D., directeur de recherche

M. CONAN, Jean, Ph.D., codirecteur de recherche

M. CHAMBERLAND, Steven, Ph.D., membre

M. BOUTABA, Raouf, Ph.D., membre externe

REMERCIEMENTS

Je tiens à remercier tous ceux qui m'ont supporté dans la réalisation de cette thèse. Plus particulièrement, j'exprime ma gratitude à mon directeur de recherche, le professeur Samuel Pierre, directeur du laboratoire de réseautique et d'informatique mobile (LARIM) de l'École Polytechnique de Montréal. Ce dernier s'est particulièrement dévoué pour me procurer une aide financière adéquate et un encadrement de qualité sans lequel cette thèse n'aurait pas vu le jour, ce qui fut pour moi une grande source de motivation. Les fructueuses discussions que j'ai régulièrement eues avec lui en rapport avec cette recherche m'ont beaucoup éclairé.

Mes remerciements s'adressent également à mon codirecteur de recherche Jean Conan, professeur titulaire au département de génie électrique de l'École Polytechnique de Montréal, qui a su mettre à ma disposition les ressources matérielles nécessaires à l'accomplissement de mes travaux de recherche. Je dois également remercier la professeure Paola Flocchini pour l'aide financière qu'elle m'a accordée, ainsi que les professeurs Gilles Pesant et Benjamin Smith, tous deux membres du jury de mon examen de synthèse, pour leurs commentaires et suggestions.

Je m'en voudrais de ne pas remercier les membres du Centre de recherche LICEF, ainsi que ceux du LARIM, pour leurs encouragements qui se sont surtout manifestés pendant les durs moments du travail. Je remercie également Madeleine Proulx et Marie-Hélène Dupuis, toutes deux de la bibliothèque de l'École Polytechnique de Montréal, de m'avoir si bien supporté dans la recherche documentaire. Enfin, je remercie tous ceux de ma famille qui n'ont jamais arrêté de m'encourager et de me supporter tant financièrement que moralement, me permettant ainsi d'atteindre ce but ultime.

RÉSUMÉ

Dans cette thèse, nous apportons trois principales contributions. Tout d'abord, nous avons conçu et présenté une approche modulaire de planification qui essaie d'optimiser conjointement les coûts de planification et la capacité des systèmes mobiles de la prochaine génération, en se basant sur l'identification de cinq facteurs fondamentaux : l'ingénierie du trafic, l'architecture du réseau, l'allocation des ressources, la gestion de l'itinérance globale et la couverture radio. Chaque facteur est pris en compte par un module qui traite d'un aspect particulier du problème de planification. Ces modules interviennent de façon parallèle dans le processus de planification, et la prise en compte des interactions entre eux permet une gestion simultanée de plusieurs sous-problèmes ayant des objectifs éventuellement conflictuels les uns par rapport aux autres. Ainsi, la méthodologie préconisée contribue à concevoir un système qui supporte le plus d'usagers possible et qui répond à leurs exigences, en termes de disponibilité du réseau, de la qualité de la communication et de délai de connexion.

Les résultats portant sur le compromis existant entre la capacité du réseau et la puissance minimale requise par utilisateur pour maintenir une communication de qualité ont montré que, pour une valeur donnée du rapport signal à bruit, la puissance minimale requise par usager décroît lorsque la capacité de la cellule augmente. En outre, l'analyse du comportement du rapport signal à bruit minimal reçu par utilisateur en fonction du niveau d'interférence a montré que plus le trafic dans le système est intense, plus le système doit fournir de la puissance pour maintenir une communication de qualité. La comparaison du rapport signal à bruit minimal par usager pour les macro-cellules, micro-cellules et pico-cellules a également révélé qu'un tel rapport est plus élevé pour les macro-cellules que pour les autres types de cellules, et que l'utilisation de micro-cellules ou de pico-cellules aide à augmenter la capacité du système.

Ensuite, nous avons conçu et mis en œuvre une approche de gestion de mobilité globale basée sur l'utilisation d'un équipement spécial d'interconnexion appelé WING

(*Wireless INterworking Gateway*) qui facilite l'interopérabilité des composantes hétérogènes du réseau. Cette base de données enregistre le profil, ainsi que les informations relatives aux sessions des usagers qui traversent les frontières des systèmes adjacents pour accélérer la télé-recherche et l'acheminement des communications. De plus, pour évaluer la quantité de trafic de signalisation généré, nous avons défini les séquences d'opérations mises en œuvre au niveau du WING lors des changements de sous-systèmes, et modélisé chaque WING par une file d'attente de type M/G/1. Les performances de l'approche proposée ont été évaluées pour deux modèles de mobilité : le modèle fluide et le modèle de gravité.

L'évaluation de performance nous a permis de tirer plusieurs conclusions. D'abord, pour un taux fixé d'appels émis ou reçus, le trafic de signalisation généré au niveau des bases de données tend à augmenter lorsque le degré de mobilité globale des abonnés augmente, mais ne dépend pas du modèle de mobilité utilisé. Les résultats ont aussi révélé que le temps de réponse aux requêtes décroît linéairement en fonction d'un paramètre p qui caractérise la stratégie de stockage d'informations et qui indique la probabilité qu'une information requise lors d'une requête soit accessible au niveau d'un VLR (*Visitor Location Register*). Dans le même ordre d'idées, nous avons considéré deux sous-systèmes i et j, et analysé, pour chaque modèle de mobilité, l'influence de la répartition des usagers sur le trafic de signalisation, ainsi que sur le temps de réponse du réseau. Les résultats obtenus de cette analyse ont montré que l'influence de la répartition des usagers sur la qualité de service dépend du modèle de mobilité considéré. Dans tous les cas, nous avons trouvé que, peu importe le comportement ou la répartition des abonnés, la stratégie de stockage ou le modèle de mobilité considéré, l'approche préconisée dans cette thèse contribue à améliorer significativement les performances du réseau, en termes de trafic de signalisation et de délai associé à la télé-recherche et à la localisation des utilisateurs.

En outre, nous avons proposé deux modèles de files d'attente pour caractériser le trafic dans chaque cellule des systèmes mobiles de la prochaine génération : le M/G/∞ (ou M/M/∞ sous certaines conditions) et le G/G/c/c. Le premier modèle se base sur la

capacité élastique de la technologie CDMA pour modéliser chaque cellule par un système n'ayant aucune limite théorique sur le nombre de communications à gérer. Ce modèle nous a permis d'analyser, entre autres, l'effet de la taille des cellules sur l'intensité de trafic. Les résultats de cette analyse ont montré que, pour les macro-cellules, l'intensité du trafic est plus faible que pour les micro-cellules ou les pico-cellules qui, quant à elles, sont généralement considérées comme des zones plus denses. Il en résulte que le risque d'interférence est plus élevé dans les pico-cellules que dans les micro-cellules ou les macro-cellules.

Quant au modèle G/G/c/c, son analyse étant complexe, nous avons appliqué les principes d'entropie maximale pour le résoudre. Cela nous a alors permis d'évaluer de manière réaliste à la fois la distribution du trafic et la probabilité de blocage d'appels dans chaque cellule, en plus de permettre l'analyse du comportement de tout système qui peut être modélisé par une file d'attente de type G/G/c/c. En termes de résultats, le modèle G/G/c/c nous a permis d'analyser, d'une part, l'effet des coefficients de variation des distributions d'arrivée et de service C_a et C_s sur la distribution du trafic $p(n)$ et, d'autre part, celui des coefficients de variation sur la probabilité de blocage p_B. Pour l'analyse de l'effet de C_a et de C_s sur $p(n)$, nous avons trouvé que le coefficient de variation de la distribution du trafic d'arrivée a plus d'influence sur la distribution du nombre de terminaux actifs que celui de la distribution du temps d'occupation des canaux. Les résultats ont aussi permis de valider que la distribution du trafic est insensible au temps de service pour tout modèle de file d'attente de type M/G/c/c.

Nous avons enfin entrepris d'analyser l'effet de C_a et de C_s sur p_B de manière à déterminer, en fonction de C_a et de C_s, les conditions selon lesquelles le réseau offre de faibles probabilités de blocage, c'est-à-dire une meilleure qualité de service. Nous avons trouvé que, pour des valeurs fixées de C_a, une meilleure qualité de service peut être offerte pour des valeurs élevées de C_s, alors qu'une meilleure qualité de service est offerte dans chaque cellule pour des valeurs faibles de C_a lorsque C_s est fixe. Nous avons toutefois constaté que p_B est plus sensible aux variations de C_a qu'à celles de C_s. De plus, nous avons vérifié que, pour un trafic d'arrivée poissonien, p_B reste invariante pour tout

$C_s \geq 1$, ce qui valide que la probabilité de blocage est insensible à la distribution du temps de service dans un modèle de type M/G/c/c.

ABSTRACT

The Next-Generation (NG) wireless systems are intended to unify many current systems into a seamless infrastructure, capable of offering a wide range of services to both mobile and fixed users. Planning such systems remains a challenging process which requires coping simultaneously with a large number of different aspects, such as the system performance, the network capacity, the radio coverage analysis, the signaling traffic, as well as the mobility management and the access technology. In this thesis, we bring several contributions related to planning the NG wireless systems. First, we propose a modular approach which identifies the key planning factors and their interactions in order to jointly optimize the infrastructure costs and the system capacity. Results concerning the system capacity have shown that an increase in the traffic intensity requires an increase in the transmitted power in order to maintain the quality of service. Also, it has been shown that the required minimum signal to interference ratio per user is higher for macrocells than for microcells and picocells. As a result, picocells require better coverage than microcells or macrocells, while using picocells or microcells enables to increase the system capacity.

Secondly, in the NG wireless systems, mobile users will be able to move across various heterogeneous networks while using their mobile terminals to communicate, which significantly increases the network-signaling traffic. In this context, we propose an efficient approach that uses a special gateway called *Wireless INterworking Gateway* (WING) to facilitate interoperability between heterogeneous subsystems of the NG wireless systems. Results reveal that such an approach significantly improves the results obtained from existing methods, in terms of generated signaling traffic and response time during the global roaming process.

Furthermore, the NG wireless systems raise challenging problems in teletraffic modeling, due to the random aspect of generating and terminating calls, as well as the user mobility. In this context, we propose to model each cell within the service area as an

M/G/∞ or a G/G/c/c queueing system. The first model has permitted to compare the traffic distribution within picocells, microcells and macrocells, and to confirm that for macrocells, the expected traffic intensity is lower than for microcells or picocells. As for the G/G/c/c model, we have considered, for analysis purposes, that arrivals from new and handoff calls are characterized by a mean rate of λ and a squared coefficient of variation of C_a, whereas service distribution has a mean rate of μ and a squared coefficient of variation of C_s. As this model has not been explicitly addressed in the literature, we have applied maximum entropy principles for the evaluation of both traffic distribution and blocking probability within the service area. The general solution has permitted to not only derive results previously obtained from M/M/c/c models, but also analyze the behavior of any system which may be modeled as a G/G/c/c queue. The G/G/c/c model has also permitted to validate that traffic distribution is insensitive to the service distribution when call arrivals follow a Poisson process. Result analysis has shown that variation of C_a has more impact over both traffic distribution and blocking probability than variation of C_s. Furthermore, this analysis has revealed that the system offers better quality of service (*i.e.*, lower blocking probability) for high values of C_s when C_a is constant, or for small values of C_a when C_s remains constant.

TABLE DES MATIÈRES

Page

REMERCIEMENTS.....iv

RÉSUMÉ.....v

ABSTRACT.....ix

TABLE DES MATIÈRES.....xi

LISTE DES TABLEAUX.....xv

LISTE DES FIGURES.....xvi

SIGLES ET ABRÉVIATIONS.....xviii

CHAPITRE 1 : INTRODUCTION.....1

1.1 Définitions et concepts de base.....2

1.2 Éléments de problématique.....5

1.3 Objectifs de recherche et résultats anticipés.....7

1.4 Principales contributions.....9

1.5 Plan de la thèse.....10

CHAPITRE 2 : ÉLÉMENTS DE PLANIFICATION DES RÉSEAUX MOBILES.....12

2.1 Architecture de base.....12

2.1.1 Sous-systèmes radio.....13

2.1.2 Sous-système réseau.....15

2.2 Systèmes mobiles de troisième génération.....16

2.2.1 Architecture de l'UMTS.....16

2.2.2 L'architecture CDMA2000.....18

2.3 Paramètres de planification.....19

2.3.1 Charge du système.....20

2.3.2 Paramètres de qualité de service.....23
2.3.3 Contraintes de qualité de service.....25
2.3.4 Mobilité des usagers.....25
2.4 Formulation du problème de planification.....28
2.5 Étapes de planification.....31
2.5.1 Planification de la partie radio.....32
2.5.2 Planification de la partie réseau.....33
2.6 Synthèse des approches traditionnelles de planification.....34

CHAPITRE 3 : MÉTHODOLOGIE DE PLANIFICATION PROPOSÉE.....37
3.1 Caractérisation de la méthodologie.....37
3.2 Description des modules.....39
3.2.1 Couveture radio.....39
3.2.2 Architecture.....42
3.2.3 Allocation des ressources.....50
3.2.4 Itinérance globale.....54
3.2.5 Ingénierie du trafic.....56
3.3 Résultats obtenus.....56

CHAPITRE 4 : APPROCHE PROPOSÉE POUR LA GESTION DE LA MOBILITÉ GLOBALE.....61
4.1 Gestion de la mobilité globale.....61
4.2 Approche proposée.....66
4.2.1 Principes généraux.....66
4.2.2 Séquence des opérations.....69
4.3 Évaluation de performance.....72
4.3.1 Définition des paramètres.....73
4.3.2 Modèles de mobilité.....74
4.3.3 Taux de requêtes et de mises à jour.....75

4.3.4 Temps de réponse du réseau.....78
4.4 Résultats obtenus et analyse.....80
4.4.1 Modèle de mobilité fluide.....80
4.4.2 Modèle de gravité.....84

CHAPITRE 5 : INGÉNIERIE DU TRAFIC.....93
5.1 Paramètres de modélisation.....93
5.1.1 Principes de modélisation des cellules.....94
5.1.2 Caractérisation du trafic d'arrivée.....95
5.1.3 Durée d'occupation des canaux.....96
5.1.3.1 Évaluation.....97
5.1.3.2 Durée d'occupation exponentielle.....98
5.1.3.3 Durée d'occupation de loi générale.....99
5.2 Modèle théorique d'évaluation du trafic.....107
5.3 Proposition d'un modèle général.....110
5.3.1 Caractérisation du modèle.....110
5.3.2 Principes de résolution du modèle.....111
5.3.3 Évaluation de la distribution du trafic et de la probabilité de blocage...115
5.3.4 Résolution du modèle M/M/c/c à partir du modèle présenté.....118
5.4 Résultats obtenus et analyse.....121
5.4.1 Influence de la taille des cellules sur l'intensité de trafic.....122
5.4.2 Influence des coefficients de variation sur la distribution du trafic.....123
5.4.3 Influence de C_a et de C_s sur la probabilité de blocage.....129
5.5 Adaptabilité des modèles proposés au trafic multimédia.....133

CHAPTER 6 : CONCLUSION.....136
6.1 Synthèse des résultats.....136
6.2 Limitations des travaux et orientations de recherches.....140

BIBLIOGRAPHIE.....142

LISTE DES TABLEAUX

Page

2.1 Contraintes de délais en fonction du type de service..... 26
3.1 Rapport signal à bruit acceptable en fonction des interférences.....59
3.2 Paramètres d'évaluation du trafic.....60
4.1 Paramètres d'analyse pour le modèle fluide..... 82
4.2 Temps moyen de traitement au niveau des bases de données.....84
4.3 Paramètres d'analyse pour le modèle de gravité.....86
5.1 Paramètres d'analyse de trafic.....123
5.2 Valeur maximale de $p(n)$ et de p_B en fonction de C_s.....126
5.3 Valeur maximale de $p(n)$ et de p_B en fonction de C_a.....128
5.4 Valeur maximale de $p(n)$ et de p_B en fonction de $C_a = C_s$.....129

LISTE DES FIGURES

Page

1.1 Vision globale des réseaux mobiles de la prochaine génération 4
1.2 Formes de mobilité dans différents réseaux 6
2.1 Architecture de base des réseaux mobiles.....13
2.2 Architecture de l'UMTS.....17
2.3 Architecture CDMA2000.....20
2.4 File d'attente modélisant une cellule.....22
2.5 Approche traditionnelle de planification.....36
3.1 Méthodologie de planification proposée.....38
3.2 Opérations du module *Couverture radio*.....42
3.3 Séquence des opérations de la méthode d'affectation proposée.....48
3.4 Opérations du module *Architecture*.....50
3.5 Prédiction de déplacement pour l'allocation des ressources.....51
3.6 Opérations du module *Allocation des ressources*.....54
3.7 Opérations du module *Itinérance globale*.....55
3.8 Opérations du module *Ingénierie du trafic*.....57
3.9 Rapport signal à bruit minimal par usager.....58
3.10 Rapport signal à bruit acceptable par usager selon le type de cellules.....60
4.1 Réseau global constitué de sous-systèmes hétérogènes.....62
4.2 Subdivision de la zone de service de chaque sous-système en LAs.....63
4.3 Architecture centralisée pour la gestion de la mobilité globale.....64
4.4 Schéma d'interconnexion des HLRs selon l'architecture proposée.....67
4.5 Processus d'enregistrement dans un nouveau sous-système.....70
4.6 Processus de mise à jour de localisation.....71
4.7 Processus d'acheminement d'une communication.....73
4.8 Influence du comportement des usagers sur le taux de requêtes.....83
4.9 Influence du comportement des usagers sur le taux de mises à jour.....84

4.10 Influence de la distribution des usagers sur le taux de requêtes.....85
4.11 Influence de la distribution des usagers sur le taux de mises à jour.....86
4.12 Influence de la distribution des usagers sur le temps de réponse.....87
4.13 Influence du paramètre p sur le temps de réponse.....88
4.14 Influence du GCMR sur le taux de requêtes (modèle de gravité).....89
4.15 Influence du GCMR sur le taux de mises à jour (modèle de gravité).....90
4.16 Influence de la distribution des abonnés sur le taux de requêtes (modèle de..... gravité).....91
4.17 Influence de la distribution des usagers sur le temps de réponse (modèle de..... gravité).....92
5.1 Détermination de la durée d'occupation d'un canal.....98
5.2 Distribution exponentielle de T_c.....100
5.3 Distribution hyperexponentielle de T_c avec $k = 2$, $\mu_1 = 1$, $\mu_2 = 4$.....101
5.4 Distribution normale de T_c avec $m = 5$.....103
5.5 Distribution lognormale de T avec $m = 0$.....104
5.6 Durée d'occupation de loi Gamma avec $\alpha = 1$.....106
5.7 Modélisation d'une cellule par une file d'attente à capacité infinie.....109
5.8 Distribution des terminaux actifs pour plusieurs types de cellules.....124
5.9 Distribution des terminaux actifs avec $C_a = 1.5$.....125
5.10 Distribution des terminaux actifs pour $C_s = 1.5$.....127
5.11 Distribution des terminaux actifs avec C_a et C_s variables.....128
5.12 Évolution de la probabilité de blocage avec $C_a = 1.5$.....130
5.13 Évolution de la probabilité de blocage avec $C_s = 1.5$.....131
5.14 Évolution de la probabilité de blocage avec C_a et C_s variables.....132

SIGLES ET ABRÉVIATIONS

3GPP	: Third-Generation Partnership Project
AAA	: Authentication, Authorization, Accounting
ANSI	: American National Standards Institute
AMPS	: Advanced Mobile Phone System
ATM	: Asynchronous Transfer Mode
BER	: Bit Error Rate
BHCA	: Busy Hour Call Attempt
BIU	: Boundary Interworking Unit
BLA	: Boundary Location Area
BLR	: Boundary Location Register
BS	: Base Station
BSC	: Base Station Controller
BSS	: Base Station Sub-System
BTS	: Base Transceiver Station
CDMA	: Code Division Multiple Access
EIA	: Electronic Industry Associations
ETSI	: European Telecommunications Standards Institute
FDMA	: Frequency Division Multiple Access
FIFO	: First In First Out
GCMR	: Global Call-to-Mobility Ratio
GE	: General Exponential
GGSN	: Gateway GPRS Support Node
GMSC	: Gateway MSC
GPRS	: General Packet Radio Systems
GSM	: Global System for Mobile communications
GSM MAP	: GSM Mobile Application Part
HLR	: Home Location Register

HSS	: Home Subscriber Server
IS-41	: Interim Standard 41
IS-95	: Interim Standard 95
LA	: Location Area
LCMR	: Local Call-to-Mobility Ratio
ME	: Maximum Entropy
MGW	: Media Gateway
MHLR	: Multi-tier Home Location Register
MR	: Multiple Registration
MSC	: Mobile Switching Center
MT	: Mobile Terminal
MU	: Mobile User
NG	: Next-Generation
NM	: New Methodology
NSS	: Network Sub-System
OMC-N	: Operating and Maintenance Center-Network
OMC-R	: Operating and Maintenance Center-Radio
PDSN	: Packet Data Serving Node
PHS	: Personal Handyphone System
PN	: Personal Number
QoS	: Quality of Service
RA	: Registration Area
RNC	: Radio Network Controller
RNS	: Radio Network Subsystem
RTCP	: Réseau Téléphonique Commuté Public
SGSN	: Serving GPRS Support Node
SOHYP	: Sum of Hyperexponentials
SR	: Single Registration
SS7	: Signaling System 7

TDMA	: Time Division Multiple Access
TIA	: Telephone Industry Associations
UE	: User Equipment
UIT	: Union Internationale des Télécommunications
UMTS	: Universal Mobile Telecommunications System
UPT	: Universal Personal Telecommunications
UTRAN	: Universal Terrestrial Radio Access Network
VRL	: Visitor Location Register
WATM	: Wireless ATM
WCDMA	: Wideband CDMA
WING	: Wireless INterworking Gateway

CHAPITRE 1

INTRODUCTION

Depuis le début des années 90, nous assistons à un véritable engouement pour le développement des réseaux mobiles. En effet, le nombre d'abonnés de ces systèmes est passé de 11 millions en 1990 à plus de 500 millions à la fin de l'an 2000 (Spilling *et al.*, 2000). En 1998 seulement, on a dénombré 100 millions de nouveaux abonnés dans le monde (Varshney *et al.*, 1999). En outre, on s'attend à avoir un total de 830 millions d'abonnés dans le monde d'ici la fin de 2003 et 1,8 milliards d'ici l'an 2010 (Prasad, 1999; Spilling *et al.*, 2000). Dans certains pays, le nombre d'abonnés de réseaux mobiles dépasse déjà celui des réseaux fixes (Pandya, 1999). Pourtant, malgré cette popularité, les réseaux mobiles actuels, appelés *systèmes de la première* ou de la *deuxième génération* et initialement conçus pour la transmission de la voix, fournissent une couverture limitée à une région géographique donnée. Aussi sont-ils incapables de répondre aux exigences actuelles du trafic multimédia. Il s'avère alors important de mettre au point de nouveaux systèmes, que nous appelons la prochaine génération de réseaux mobiles, capables de satisfaire aux nouveaux besoins des utilisateurs. De tels systèmes, également connus sous le vocable de *systèmes de troisième génération*, visent à offrir non seulement des services mobiles multimédia de qualité, en mettant à contribution aussi bien des composantes fixes et mobiles que des satellites, mais aussi une couverture globale, en unifiant l'ensemble des systèmes actuels en une infrastructure unique, capable d'interopérer avec divers environnements radio. Cette infrastructure, de par sa complexité, doit faire l'objet d'une bonne planification, de manière à optimiser les coûts d'investissement, tout en respectant les multiples contraintes liées à la qualité de service à offrir. Cette thèse vise justement à apporter une série de contributions relatives à deux aspects importants de la planification des systèmes mobiles de la prochaine (ou troisième) génération : la mobilité globale des usagers et le trafic multimédia.

1.1 Définitions et concepts de base

Dans un réseau mobile, la zone de service est constituée de stations de base (BS : Base Station) qui fournissent les liens radio aux usagers mobiles pour la communication. Chaque BS dessert une zone géographique appelée *cellule*. En général, on distingue trois catégories de cellules: les macro-cellules, les micro-cellules et les pico-cellules (Coombs et Steel, 1999). Les macro-cellules ont un rayon allant de 1 à 30 km et desservent principalement les autoroutes où la vitesse des usagers est relativement élevée (Tabbane, 2000). Pour leur part, les micro-cellules utilisent des BSs de faible puissance et qui sont généralement installées dans les zones à forte concentration d'abonnés (tels que le centre-ville, les aéroports et les centres d'achat) pour desservir surtout les abonnés à faible vitesse (comme les piétons) dans un rayon allant de 100 à 300 mètres. Quant aux pico-cellules, elles sont principalement utilisées pour desservir les édifices à bureaux. Elles sont en mesure d'offrir des vitesses de transmission pouvant atteindre les 2 Mbps et une couverture radio qui s'étend sur plusieurs étages, dans un rayon de 10 à 100 mètres (Castro, 2001).

Par ailleurs, les réseaux mobiles actuels peuvent être divisés en systèmes de la première génération et ceux de la deuxième génération (Pandya, 1999). La première génération a été lancée au début des années 80, utilisant le mode de transmission analogique et la technologie FDMA (Frequency Division Multiple Access), dans la bande de fréquences 800-900 MHz pour la transmission de la voix (Lin et Chlamtac, 1996). Actuellement, le système AMPS (Advanced Mobile Phone System) demeure parmi les systèmes les plus utilisés de la première génération, avec plus de 50 millions d'abonnés dans le monde, principalement aux Etats-Unis (Agrawal et Famolari, 1999). Au début des années 90, la deuxième génération a été commercialement lancée, utilisant la transmission numérique et le mode d'accès TDMA (Time Division Multiple Access) pour la transmission de la voix et des données (Sollenberger *et al.*, 1999). Actuellement, le GSM (Global System for Mobile Communications), le CDMA IS-95 (Code Division Multiple Access), le TDMA IS-136 et le PHS (Personal Handyphone System) constituent

des exemples de systèmes numériques de la deuxième génération, parmi les plus utilisés dans le monde (Ojanperä et Prasad, 1998).

Toutefois, nous nous rendons compte que les réseaux de la deuxième génération sont limités par les exigences du trafic actuel. En effet, leur interface radio est principalement optimisée pour le transport de la voix et offre des services précaires de transfert de données, où la taille des messages est limitée à 160 caractères et le débit offert est inférieur à 9.6 kbps (Prasad, 1999). En outre, l'itinérance globale, permettant à un usager mobile de passer d'un réseau à un autre de manière transparente, n'y est généralement pas supportée. C'est dans ce contexte qu'au début des années 90, l'UIT (*Union Internationale des Télécommunications*) a entrepris de participer activement au développement de standards qui visent la mise en place des réseaux mobiles de la prochaine génération (Schwartz, 1995; Buchanan *et al.*, 1997; Aghvami et Jafarian, 2000).

Cette génération se propose de regrouper les divers environnements mobiles et incompatibles en une infrastructure capable d'offrir, avec une bonne qualité de service, toute une gamme de services de télécommunications à grande échelle. Ses principales caractéristiques sont les suivantes (Buchanan *et al.*, 1997; Pandya, 1999; Aghvami et Jafarian, 2000) :

- spectre radio commun partout dans le monde (bande 1.8-2.2 GHz);
- support de la recherche globale d'usagers dans le réseau (itinérance ou mobilité globale);
- intégration et compatibilité des services des différents réseaux fixes et mobiles;
- débit élevé, pouvant atteindre les 10 Mbps et supportant des applications multimédia (accès rapide à l'Internet, traitement d'images, vidéoconférence);
- sécurité accrue.

Il en résultera une amélioration significative par rapport aux systèmes existants, en termes de mobilité globale des usagers et des services offerts.

La vision globale des réseaux mobiles de la prochaine génération est illustrée à la Figure 1.1. Trois idées clés y sont présentées : l'intégration des réseaux fixes et satellitaires aux réseaux mobiles, la flexibilité et l'adaptativité des terminaux à plusieurs types de systèmes, la présence d'intelligence aux différents nœuds du réseau, ce qui permet d'offrir aux usagers la mobilité personnelle, ainsi que la portabilité des services.

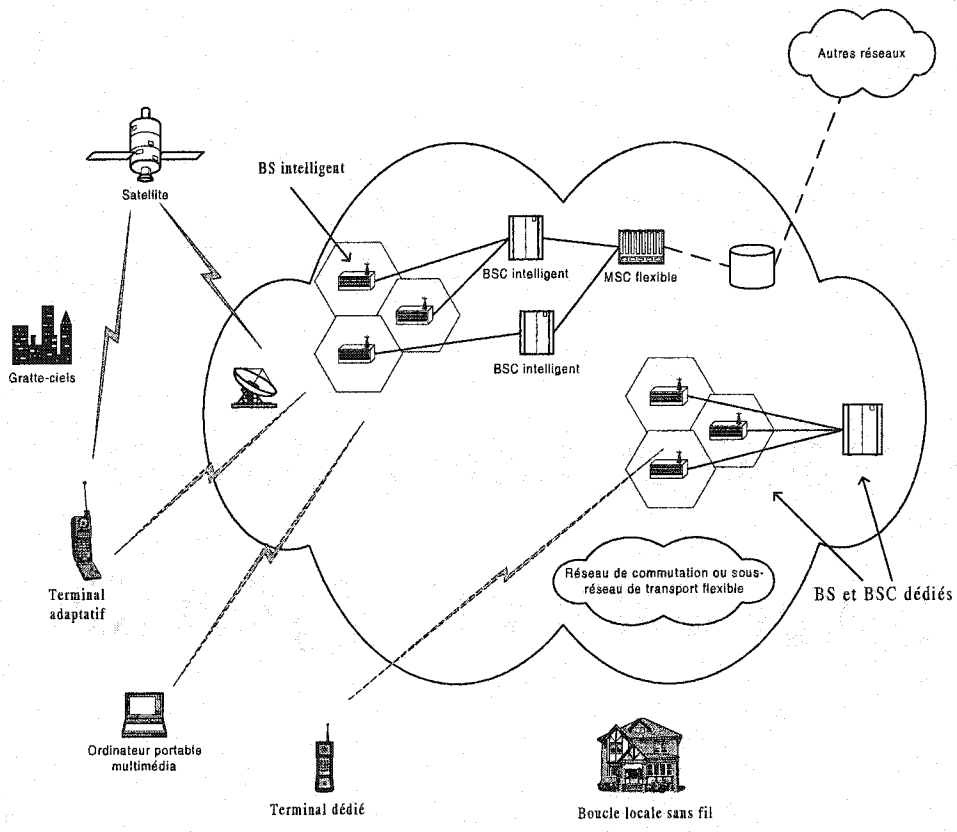


Figure 1.1 Vision globale des réseaux mobiles de la prochaine génération

Par ailleurs, le concept d'un appareil de communications à la fois petit, léger et flexible joue un rôle fondamental dans le développement des systèmes de la prochaine génération. Cet appareil offre à la fois la mobilité du terminal, complémentaire à la mobilité personnelle, et le service de gestion du profil fourni par l'UPT (*Universal Personal Telecommunications*). Le concept d'UPT permet aux usagers de bénéficier

d'un ensemble de services personnalisés, d'émettre et de recevoir des appels sur la base d'un numéro personnel unique (PN : *Personal Number*), appelé « numéro UPT », transparent au réseau (Pandya, 1999). Grâce aux services offerts par l'UPT, chaque abonné est décrit au sein du système par un profil de services. Ce profil contient des informations statiques (numéro UPT, options de services, configuration du terminal, ...) et des informations dynamiques (localisation de l'usager, durée des communications, ...).

La mobilité du terminal est caractérisée par la possibilité pour un abonné d'être localisé et identifié pendant le déplacement, et d'accéder aux services à partir d'un point quelconque du réseau. À ce niveau, même si la relation entre le réseau et le terminal est dynamique, la relation entre un usager et son terminal demeure statique et la gestion des communications se base sur l'identification du terminal. Quant à la mobilité personnelle, elle est caractérisée par la possibilité d'identifier les utilisateurs lors de leur déplacement, de leur permettre de gérer leurs communications et d'accéder aux services auxquels ils sont abonnés, à partir de n'importe quel terminal (fixe ou mobile). De ce fait, la mobilité personnelle se réfère à la portabilité des services et repose sur l'association dynamique entre l'usager et son terminal, de sorte que la gestion des communications se base sur le numéro personnel. La gestion de la mobilité (ou de l'itinérance) consiste à déterminer à tout moment la position de chaque abonné dans le système, de manière à pouvoir le rejoindre lors d'un appel. La Figure 1.2 illustre les formes de mobilité offertes par l'UPT (Tabbane, 1997).

1.2 Éléments de la problématique

De manière générale, la planification d'un système consiste à se baser sur une série de données pour faire les meilleurs choix technologiques possibles de manière à optimiser le coût de l'infrastructure ou l'utilisation des ressources du système, en s'assurant qu'une série de contraintes liées à la qualité de service seront respectées. Dans le contexte des réseaux mobiles de la prochaine génération, la planification constitue une tâche critique et un processus délicat qui doit tenir compte, entre autres, du niveau de

trafic et du profil des usagers pour optimiser à la fois le coût d'exploitation et la capacité du système. À ce niveau, plusieurs méthodologies et outils ont été développés pour faciliter la mise en place, réduire le temps et le coût de planification des systèmes mobiles actuels (Tutschku et Tran-Gia, 1998; Bahai et Aghvami, 2000; Spilling *et al.*, 2000). Mais, de tels modèles font abstraction de la mobilité personnelle et de la distribution des usagers mobiles, en plus de faire abstraction des interactions entre les différents facteurs du processus de planification. Par conséquent, ils peuvent dégénérer en un processus long et coûteux qui ne garantit pas pour autant le respect des contraintes relatives au trafic multimédia.

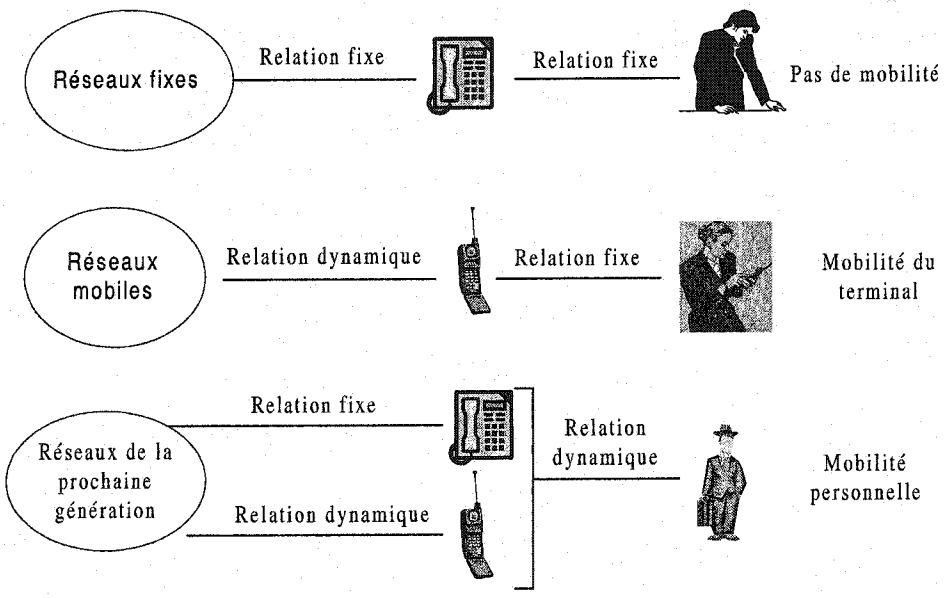


Figure 1.2 Formes de mobilité dans différents réseaux

Par ailleurs, les modèles actuels de conception ne sont pas évolutifs, en ce sens qu'à chaque période de 10 - 15 ans, le processus de planification et de mise en place d'une nouvelle génération de systèmes se met en branle. C'est d'ailleurs pourquoi, entre autres, on parle de première, deuxième, troisième et de quatrième génération. À ce niveau, il importe de se demander s'il est possible de mettre en place une méthodologie

qui permettrait de concevoir une génération de réseaux mobiles en nécessitant de simples réajustements de paramètres plutôt qu'une reprise complète du processus de conception, lors de changements majeurs d'ordre technologique ou enregistrés dans le comportement des usagers. Cela permettrait alors de parler simplement de prochaine génération.

Toutefois, la planification d'un tel système, pour être efficace, nécessite la modélisation adéquate du trafic, en tenant compte de ses diverses composantes multimédia, de la mobilité globale (ou personnelle), ainsi que de l'interopérabilité des sous-réseaux faisant partie du système global. Comme il n'existe pas vraiment de modèle général de trafic qui tient compte de toutes ces contraintes, certains chercheurs ont essayé d'adapter les modèles des réseaux fixes à ceux des réseaux mobiles (Jabbari, 1996; Spilling *et al.*, 2000), alors que d'autres n'ont fait que décrire la distribution temporelle du trafic en un point donné, en supposant une répartition uniforme des usagers mobiles dans le système (Tutschku et Tran-Gia, 1998). Malheureusement, ces modèles sont incapables de prédire le comportement des usagers dans une zone de service donnée, en plus de faire abstraction de leur distribution dans le système. Par conséquent, ils doivent être améliorés pour être applicables aux systèmes de la prochaine génération.

D'autre part, la gestion de la mobilité implique que chaque terminal reste en état de veille, de façon à signaler ses mouvements au système et ce, même en l'absence de communication. Il en résulte un trafic de signalisation important qui, même s'il utilise les ressources du réseau, ne peut être facturé aux abonnés. Ainsi, l'une des préoccupations des concepteurs de systèmes mobiles demeure la minimisation du coût de gestion de la mobilité (ou de l'itinérance). À ce propos, une série de recherches ont été entreprises, mais les méthodes préconisées ne sont valides que pour les systèmes de la première ou de la deuxième génération (Park et Lin, 1997; Akyildiz *et al.*, 1998; Safa, 2000).

1.3 Objectifs de recherche et résultats anticipés

L'objectif principal de cette thèse est de concevoir une méthodologie qui essaie d'optimiser conjointement les coûts de planification et la capacité du système, en se

basant sur cinq facteurs fondamentaux d'ingénierie de systèmes : l'ingénierie du trafic, l'architecture du réseau, l'allocation des ressources, la gestion de l'itinérance globale et la couverture radio. Ces facteurs sont appelés à ajuster les variables nécessaires à chaque fois que des changements majeurs, dus au comportement des usagers, à des ajouts de services ou à la variabilité du trafic, sont apportés. De plus, la prise en compte des interactions entre ces divers facteurs permettent d'aboutir à des solutions optimisées sous différents aspects. Ainsi, la méthodologie préconisée se veut évolutive et contribue à concevoir un système qui supporte le plus d'usagers possible et qui répond à leurs exigences, en termes de disponibilité du système (faible probabilité de blocage des appels), de la qualité de la communication (interférence négligeable par rapport au signal reçu), de délai de connexion (faible temps de réponse du système). Nous apporterons alors une solution économiquement viable et offrant une bonne qualité de service au problème de planification.

De manière plus spécifique, cette thèse vise les objectifs suivants :

- Identifier les divers facteurs qui interviennent dans la conception des systèmes mobiles de la prochaine génération et analyser les interactions entre ces facteurs. Cela permettra d'identifier les relations de dépendance qui existent entre ces facteurs;
- Étant donné que les réseaux de la prochaine génération intègreront plusieurs systèmes mobiles de générations différentes qui devront interopérer, concevoir un mécanisme efficace de gestion de mobilité globale. Ce mécanisme vise à réduire non seulement le trafic de signalisation, mais aussi les délais associés à la télé-recherche (recherche d'abonnés dans le réseau) et à la localisation des utilisateurs lorsque ces derniers changent de sous-systèmes;
- Entreprendre une analyse critique du trafic et développer des modèles efficaces de distribution des usagers mobiles dans le système. Ces modèles permettront, entre autres, de faire une bonne étude de performance, ainsi qu'un dimensionnement adéquat des composantes du réseau. Il en résultera

alors des garanties de performances robustes, en cas de variabilité importante du trafic.

1.4 Principales contributions

Dans cette thèse, nous apportons trois principales contributions. La première contribution est la mise au point d'une nouvelle méthodologie de planification des systèmes mobiles de la prochaine génération. Cette méthodologie identifie les facteurs clés de la planification et se base sur leurs interactions pour optimiser conjointement les coûts de planification et la capacité du réseau, en respectant les contraintes de qualité de service. Une telle méthodologie contribue ainsi à concevoir des systèmes qui, non seulement sont économiquement viables, mais surtout répondent aux nouveaux besoins des usagers mobiles.

La deuxième contribution est la mise en œuvre d'une approche efficace de gestion de mobilité globale qui réduit significativement le nombre d'opérations à exécuter lors du passage d'un sous-système à un autre. Cette approche définit, entre les cellules des systèmes adjacents, des zones d'échange de trafic inter-systèmes qui seront contrôlées par des bases de données spéciales que nous appelons des WINGs (*Wireless INterworking Gateway*), chargés de faciliter l'interopérabilité des sous-systèmes. Ces WINGs enregistrent le profil, ainsi que les informations relatives aux sessions des usagers qui traversent les frontières des systèmes adjacents pour accélérer la télé-recherche et l'acheminement des communications. Une telle approche contribue ainsi à améliorer significativement les performances du réseau, en termes de trafic de signalisation généré et de temps de réponse aux requêtes dans un contexte de mobilité globale.

La troisième contribution est la mise en œuvre de deux modèles de files d'attente pour évaluer la distribution du trafic dans les systèmes mobiles de la prochaine génération : le M/G/∞ et le G/G/c/c. Le premier modèle se base sur la capacité élastique de la technologie CDMA pour modéliser chaque cellule par un système n'ayant aucune limite théorique sur le nombre de communications à gérer. Ce modèle permet d'analyser,

entre autres, l'influence de la taille des cellules sur l'intensité du trafic. Quant au modèle G/G/c/c, son analyse étant complexe, nous appliquerons les principes d'entropie maximale pour le résoudre. Cela permettra alors d'évaluer de manière réaliste à la fois la distribution du trafic et la probabilité de blocage d'appels dans chaque cellule, en plus de permettre l'analyse du comportement de tout système qui peut être modélisé par une file d'attente de type G/G/c/c.

1.5 Plan de la thèse

La suite de la thèse est organisée de la façon suivante. Le chapitre 2 passe en revue les éléments essentiels de planification des réseaux mobiles. Nous y décrirons l'architecture de base de ces réseaux, présenterons la formulation mathématique du problème de planification, ainsi que l'estimation de la charge du système, la gestion de la mobilité, l'estimation des paramètres et des contraintes de qualité de service. Pour finir, nous décrirons les principales étapes de planification et ferons une synthèse des approches traditionnelles.

Le chapitre 3 présente la méthodologie de planification préconisée pour l'optimisation conjointe des coûts et de la capacité du système. Nous y montrerons que les approches traditionnelles ne peuvent efficacement s'appliquer à la planification des réseaux de la prochaine génération. Nous présenterons alors les fondements de l'approche proposée et décrirons les modules qui la constitue. Pour finir, nous présenterons quelques résultats qui évaluent la capacité réelle de ces réseaux, en tenant à la fois compte du niveau d'interférence et de la qualité de service requise.

Le chapitre 4 présente l'approche de gestion de mobilité globale préconisée dans cette thèse. Nous y expliquerons les principes de base de la gestion de mobilité globale, analyserons les méthodes traditionnelles présentées dans la littérature, présenterons la nouvelle approche. Nous évaluerons alors les performances des méthodes présentées, en termes de trafic de signalisation et de temps de réponse du système, et terminerons par une comparaison de résultats.

Le chapitre 5 présente deux modèles de files d'attente pour caractériser le trafic dans les systèmes mobiles de la prochaine génération : le M/G/∞ et le G/G/c/c. Nous y caractériserons le taux d'arrivée des appels et la durée d'occupation des canaux dans chaque cellule. À partir du premier modèle, nous évaluerons la distribution du trafic dans chaque cellule et comparerons la densité de trafic des pico-cellules à celle des micro-cellules et des macro-cellules. En ce qui a trait au second modèle, il permettra d'évaluer l'impact des coefficients de variation du taux d'arrivée des appels et du temps d'occupation des canaux sur la distribution du trafic et sur la qualité de service. Pour terminer le chapitre, nous montrerons comment adapter les modèles proposés au trafic multimédia.

Pour conclure, nous ferons au chapitre 6 la synthèse des résultats. Nous relèverons aussi certaines limitations des méthodes et modèles présentés dans cette thèse, et soulèverons quelques défis de recherche pour les travaux futurs.

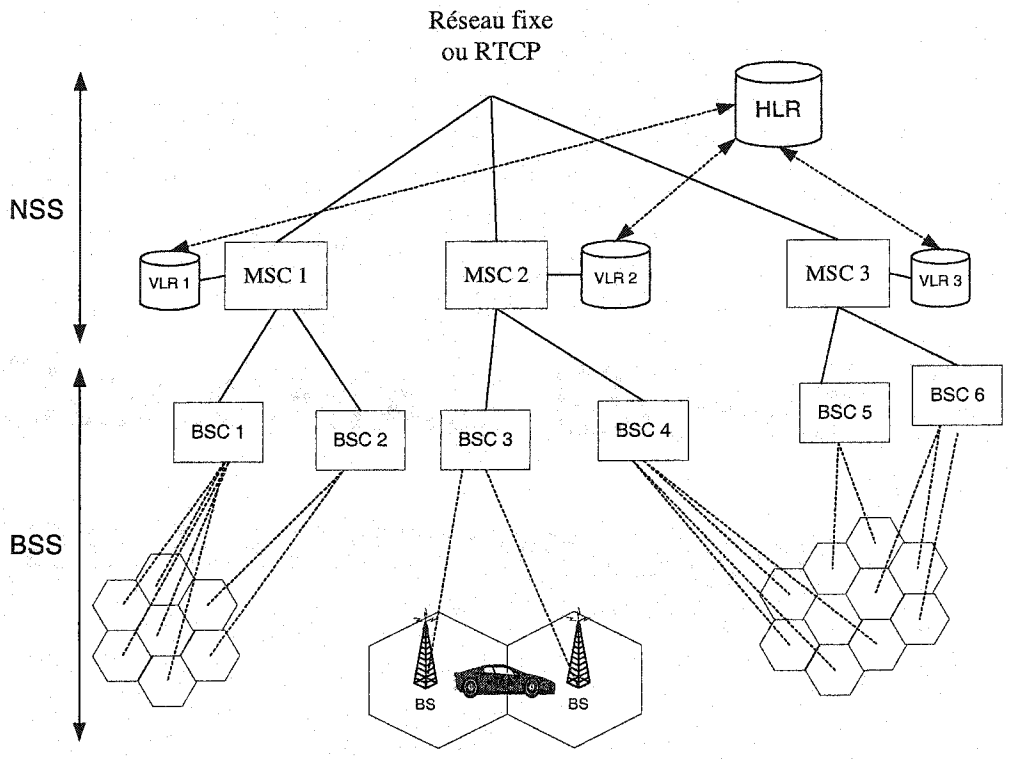
CHAPITRE 2

ÉLÉMENTS DE PLANIFICATION DES RÉSEAUX MOBILES

La planification des réseaux mobiles demeure une tâche complexe qui vise à déterminer l'ensemble des composantes matérielles et logicielles de ces systèmes, les positionner, les interconnecter et les utiliser de façon optimale, en respectant entre autres des contraintes de qualité de service. Ce processus qui peut être à la fois long et coûteux fait appel à une série de paramètres qu'il convient de spécifier. Ce chapitre passe en revue les éléments essentiels qui interviennent dans la planification des réseaux mobiles. Nous y décrirons d'abord l'architecture de base de ces réseaux, ainsi que les propositions d'architectures pour les systèmes mobiles de troisième génération. Nous caractériserons ensuite les principaux paramètres de planification, tels que la charge du système, la probabilité de blocage, les contraintes de qualité de service et la mobilité des usagers. Cela nous amènera à présenter la formulation mathématique du problème et à décrire les étapes importantes de la planification des réseaux mobiles. Nous terminerons le chapitre par une synthèse des approches traditionnelles de planification.

2.1 Architecture de base

Un réseau mobile est une infrastructure constituée essentiellement de deux parties (Lin et Chlamtac, 1996; Tabbane, 1997) : un sous-système réseau (*core* ou NSS : *Network Sub-System*) et un sous-système radio (*access network* ou BSS : *Base Station Sub-System*). La Figure 2.1 illustre une telle architecture.



Légende

BS	: Base Station
BSC	: Base Station Controller
BSS	: Base Station Subsystem
HLR	: Home Location Register
MSC	: Mobile Switching Center
NSS	: Networt Sub-System
RTCP	: Réseau Téléphonique Commuté Public
VLR	: Visitor Location Register

Figure 2.1 Architecture de base des réseaux mobiles

2.1.1 Sous-système radio

Le sous-système radio comprend tous les équipements nécessaires à la gestion de l'interface radio et des transmissions. Il est principalement constitué de deux entités : les contrôleurs de stations de base (BSCs : *Base Station Controllers*) et les stations de base (BSs). Les BSCs gèrent les ressources radio (affectation des canaux, aspects radio du

handoff), supervisent les accès et le contrôle de puissance, en plus de concentrer le trafic des BSs vers les commutateurs du service mobile (MSC : *Mobile Switching Center*). Quant aux stations de base, elles comprennent un certain nombre d'émetteurs/récepteurs (*transceivers* ou TRX) avec des codeurs/décodeurs de parole, ainsi que des modems et une ou plusieurs antennes. Cela leur permet de gérer les appels vers ou en provenance des terminaux mobiles situés dans leur zone de service (c'est-à-dire dans les différentes cellules), d'assurer le traitement physique (modulation, démodulation, égalisation, etc.), de mesurer les paramètres radio nécessaires à la supervision des communications en cours et de retransmettre ces mesures aux BSCs.

Par ailleurs, le terminal mobile (MT : *Mobile Terminal*), interface entre l'usager et le réseau, peut être considéré comme la dernière composante du sous-système radio. En plus de gérer les communications au niveau de l'interface radio du côté usager, un MT assure la gestion des communications et le contrôle des liens radio avec le système. Il est aussi en mesure de se caler sur un canal quelconque de la bande de fréquences allouée au système. Mentionnons que la couverture du sous-système radio peut s'étendre sur un pays entier, la gestion de la mobilité des usagers étant assurée par la définition de zones de localisation.

Les cellules sont organisées hiérarchiquement et ne peuvent communiquer que par l'intermédiaire des MSCs. L'opération qui consiste à noter le changement de cellule d'un utilisateur et à effectuer les mises à jour nécessaires constitue une relève *(handoff* ou *handover).* Quand la relève s'effectue entre deux cellules reliées à un même commutateur, on parle de *relève simple*, car les mises à jour à effectuer sont peu nombreuses. Quand cette relève se déroule entre deux cellules reliées à des commutateurs différents, on parle alors de *relève complexe* puisque les mises à jour consomment plus de ressources.

2.1.2 Sous-système réseau

Le NSS est constitué de commutateurs (MSCs) permettant de gérer les communications (établissement, acheminement, etc.) et de bases de données permettant de gérer le profil de chaque usager. Les MSCs sont généralement composés d'un commutateur classique, développé initialement pour le réseau fixe, auquel sont ajoutées des fonctions relatives à la gestion des ressources radio. Ils constituent l'interface entre les abonnés et le réseau téléphonique (RTCP : Réseau Téléphonique Commuté Public), gèrent les appels et assurent des fonctions telles que la signalisation, la commutation, la conversion analogique/numérique dans les systèmes à transmission radio analogique, la détection du décroché/raccroché du mobile, etc. Dans les systèmes de la seconde génération, de nouvelles fonctions, telles que la localisation des usagers, sont réalisées par des entités séparées fonctionnellement du MSC : il s'agit de la *base de données nominale* (HLR : Home Location Register) et des *bases de données visiteurs* (VLR : Visitor Location Register). En fait, dans tout système mobile, les informations concernant les abonnés sont centralisées au sein du HLR. Ce dernier, implanté au niveau de l'un des MSCs, contient les données d'abonnement des usagers, ainsi que leur localisation, c'est-à-dire l'adresse de leur zone d'itinérance. Pour sa part, le VLR contient une copie des données d'abonnement des usagers situés dans sa zone de contrôle. Ces informations y sont toutefois effacées une fois que l'abonné aura quitté cette zone de contrôle.

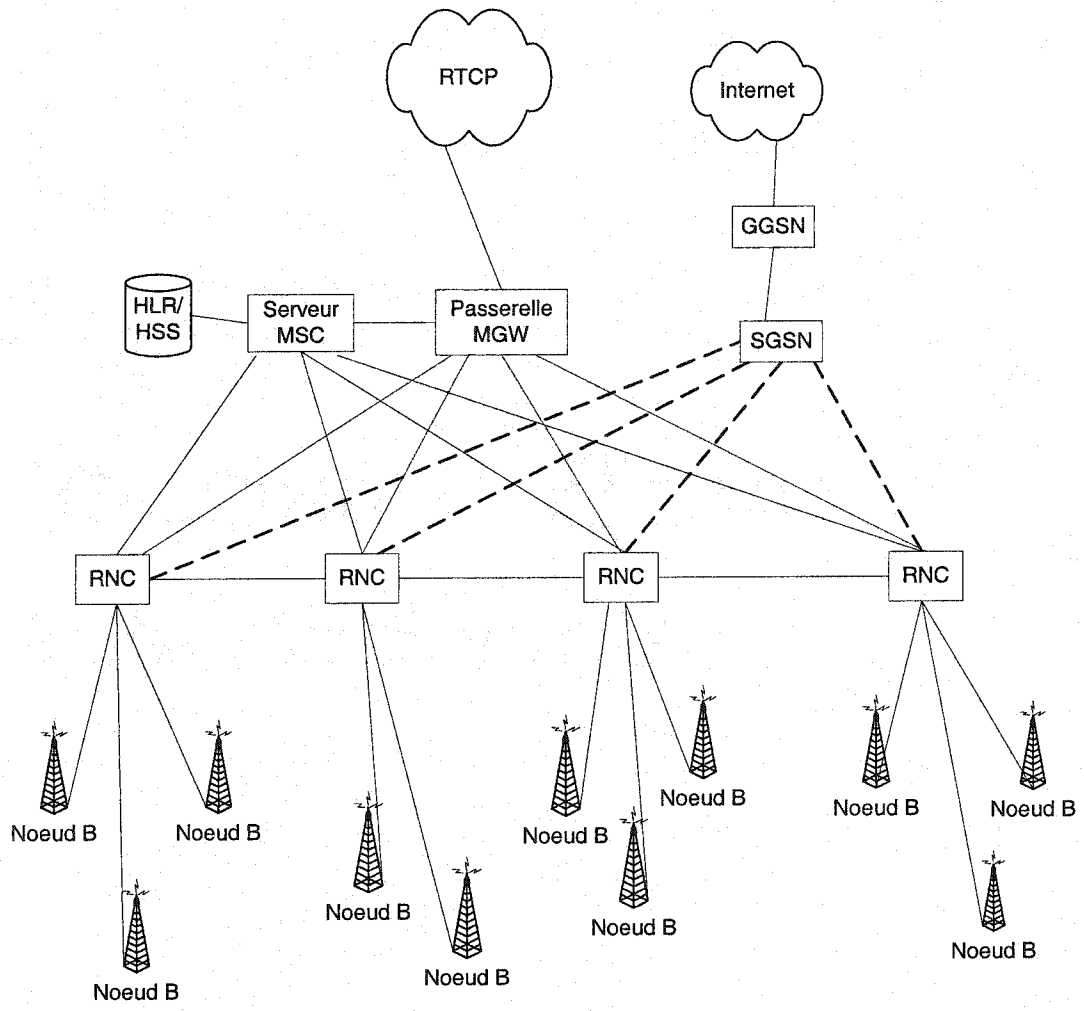
Par ailleurs, les communications entre le sous-système réseau et le sous-système radio sont réalisées à travers des lignes fixes qui fournissent les canaux de voix et de données. Ces deux sous-systèmes sont contrôlés et supervisés par un sous-système d'exploitation et de maintenance réseau (OMC-N : *Operating and Maintenance Center-Network*), et un sous-système d'exploitation et de maintenance radio (OMC-R : *Operating and Maintenance Center-Radio*). Ces entités permettent aux opérateurs d'intervenir au besoin, en fonction des statistiques observées. Elles peuvent correspondre à des équipements physiques distincts ou s'intégrer partiellement dans les mêmes équipements.

2.2 Systèmes mobiles de troisième génération

Lorsque l'Union Internationale des Télécommunications (UIT) a sollicité des solutions pour répondre aux nouveaux besoins des usagers mobiles, plusieurs technologies ont été proposées par divers groupes. Plus particulièrement, le groupe de travail 3GPP (*Third-Generation Partnership Project*), créé en 1998 par l'institut européen de normalisation ETSI (*European Telecommunications Standards Institute*), a proposé une architecture appelée UMTS (*Universal Mobile Telecommunications System*), compatible avec la norme GSM (*Global System for Mobile communications*) et ses possibles évolutions (Sanchez et Thioume, 2001). Parallèlement, l'institut américain de normalisation ANSI (*American National Standards Institute*) a créé un projet semblable au 3GPP appelé le 3GPP2 pour mettre au point les spécifications techniques d'une architecture appelée CDMA2000 (Smith et Collins, 2002). Présentons ces deux architectures.

2.2.1 Architecture de l'UMTS

L'architecture de l'UMTS, illustrée à la Figure 2.2, est essentiellement composée d'un réseau d'accès appelé UTRAN (*Universal Terrestrial Radio Access Network*) et d'un sous-système réseau (*core network*) dérivé de l'architecture GSM. L'UTRAN fournit aux terminaux mobiles ou *équipements usagers* (*UE : User Equipment*) les ressources radio et les mécanismes nécessaires pour accéder au sous-système réseau. Il est formé de stations de base appelées *Nœuds B* (*Nodes B*) et d'un ensemble de contrôleurs de stations de base appelés RNCs (*Radio Network Controllers*). Les nœuds B assurent la transmission et la réception radio, tout en appliquant des procédures telles que le codage/décodage pour la correction d'erreurs, l'adaptation de débit, la modulation, etc. L'ensemble constitué d'un RNC et des nœuds B qui y sont rattachés forme un sous-système radio (*RNS : Radio Network Subsystem*).



Légende

GGSN	: Gateway GPRS Support Node
GPRS	: General Packet Radio Systems
HLR	: Home Location Register
HSS	: Home Subscriber Server
MGW	: Media Gateway
MSC	: Mobile Switching Center
RNC	: Radio Network Controller
RTCP	: Réseau Téléphonique Commuté Public
SGSN	: Serving GPRS Support Node

Figure 2.2 Architecture de l'UMTS

Le sous-système réseau regroupe l'ensemble des équipements qui assurent le contrôle des appels, le contrôle de la sécurité, la gestion des interfaces avec les réseaux externes, etc. Dans un contexte d'architecture distribuée, on utilise un HSS (*Home Subscriber Server*) plutôt qu'un HLR. Les deux équipements sont fonctionnellement équivalents, avec la seule différence que le HSS utilise une interface basée sur la technologie des paquets (comme IP), alors qu'un HLR utilise le système de signalisation standard SS7 (Smith et Collins, 2002). Quant au MSC, il se divise en deux composantes distinctes : un serveur MSC et une passerelle appelée MGW (Media Gateway). Le serveur MSC assure toutes les fonctions liées à la mobilité et au contrôle logique, alors que la passerelle MGW, contrôlée par le serveur MSC, assure essentiellement les fonctions de signalisation entre les différents équipements. Plus précisément, le MGW accepte les appels provenant des RNCs et les achemine au destinataire en utilisant un réseau à commutation de paquets. Quant au SGSN (*Serving GPRS Support Node*), il remplit les mêmes fonctions qu'un MSC, mais dans un contexte de commutation de paquets. Ces fonctions comprennent, entre autres, la gestion de la mobilité, la sécurité, le contrôle d'accès. Le GGSN (Gateway GPRS Support Node) constitue l'interface entre le réseau mobile et les réseaux fixes à commutation de paquets (comme l'Internet). Au Japon, l'architecture de l'UMTS est commercialement connue sous le nom de WCDMA (*Wideband CDMA*) (Sanchez et Thioume, 2001).

2.2.2 L'architecture CDMA2000

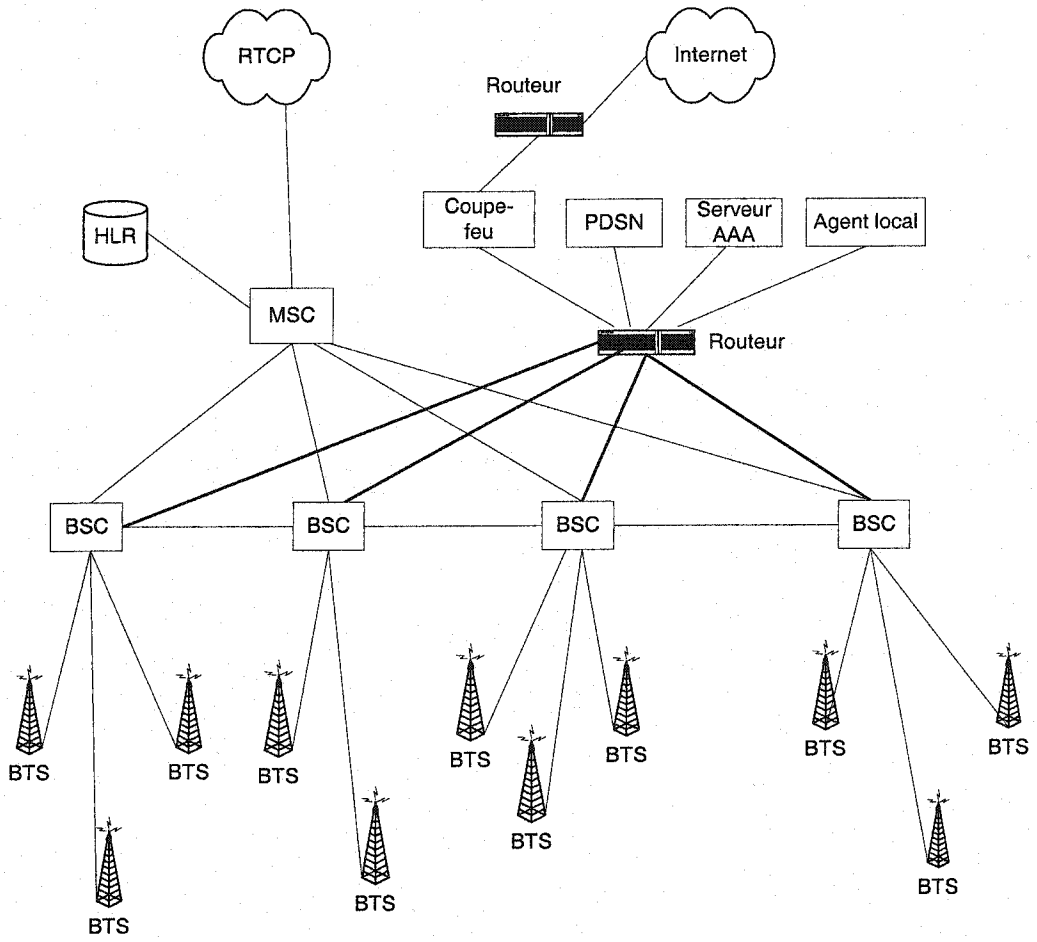
L'architecture CDMA2000 comporte tous les éléments qui font partie des réseaux de voix actuels, auxquels s'ajoutent plusieurs nouveaux équipements qui assurent l'interconnexion du réseau d'accès aux réseaux de données. Cette architecture, illustrée à la Figure 2.3, utilise une composante appelée PDSN (*Packet Data Serving Node*), essentielle au support des services de données (Smith et Collins, 2002). Une telle composante communique avec les nœuds du réseau de voix (HLR et VLRs) en utilisant un serveur d'authentification appelé AAA (*Authentication, Authorization Accounting*). Ce dernier identifie les abonnés, leur octroie des autorisations et aide à gérer leurs

comptes. Un agent local (*Home Agent*) garde la trace des usagers mobiles qui changent de zones de localisation de manière à s'assurer que les paquets sont bien acheminés à leurs destinataires. En outre, on utilise des routeurs pour faciliter l'échange de paquets entre les différentes entités du système, ainsi que l'échange entre le système et les autres réseaux fixes à commutation de paquets. Notons également qu'un coupe-feu (*firewall*) est également inséré dans l'architecture CDMA2000 pour en assurer la sécurité.

Au niveau du réseau d'accès, nous retrouvons, sous d'autres vocables, les mêmes éléments que dans l'UTRAN de l'UMTS. Plus spécifiquement, les nœuds B deviennent des BTSs (*Base Transceiver Stations*), alors que les RNCs deviennent des BSCs (*Base Station Controllers*). C'est pourquoi, dans la suite de la thèse, nous utiliserons indifféremment BTS, BS ou Nœud B pour désigner une station de base. Dans le même ordre d'idées, nous utiliserons RNC ou BSC pour désigner un contrôleur de stations de base et nous considérerons les termes HSS et HLR comme étant équivalents pour désigner une base de données nominale.

2.3 Paramètres de planification

Dans cette section, nous caractérisons les principaux paramètres de planification, tels que la charge du système, la probabilité de blocage, les contraintes de qualité de service, ainsi que la mobilité des usagers.



Légende:

AAA : Authentication, Authorization, Accounting
BSC : Base Station Controller
BTS : Base Transceiver Station
HLR : Home Location Register
MSC : Mobile Switching Center
PDSN : Packet Data Serving Node
RTCP : Réseau Téléphonique Commuté Public

Figure 2.3 Architecture CDMA2000

2.3.1 Charge du système

Le trafic est un facteur primordial de la planification des réseaux mobiles puisqu'il permet de déterminer le nombre de canaux nécessaires à la satisfaction des abonnés en fonction d'une qualité de service donnée. Ainsi, une mauvaise analyse de

trafic risque d'entraîner soit le déploiement d'un nouveau réseau, soit l'ajout d'équipements additionnels au système existant, ce qui peut faire l'objet de coûts significatifs. Il en résulte que la conception doit se baser sur une bonne analyse du trafic dans le système. Toutefois, une telle analyse demeure difficile en conception des réseaux mobiles, où le trafic dépend du comportement et du profil de chaque usager. En effet, la mobilité des usagers implique, d'une part, des communications qui débutent en un point, continuent en d'autres points, et s'achèvent en un point différent, après que le mobile eut effectué un certain nombre de handoffs. Autrement dit, il est difficile de connaître à tout instant le trafic exact d'une zone donnée, car les usagers peuvent s'y trouver de manière aléatoire.

Pour simplifier l'analyse, nous décrivons le comportement de chaque abonné selon deux niveaux : temporel et spatial. Les travaux réalisés sur la distribution temporelle du trafic ont montré que cette dernière peut être représentée par la loi *Normale*, pour signifier que les concentrations importantes sont rencontrées en ville pendant les heures de pointe et que, à la fin de la journée, cette concentration se modifie au fur et à mesure que les usagers regagnent leur domicile (Jabbari, 1996; Marchent *et al.*, 1999). Dans les réseaux mobiles traditionnels, un tel modèle s'obtient à partir des paramètres suivants : taux d'arrivée et durée moyenne des communications, durée d'occupation des ressources et taux de pénétration. Dans le cas des réseaux de la prochaine génération, des paramètres supplémentaires, tels que la nature et la quantité d'informations échangées, le taux de passage d'un réseau à un autre (résultant de la mobilité globale), doivent y être ajoutés.

Toutefois, dans cette thèse, nous allons plutôt nous attaquer à la distribution spatiale du trafic dans le réseau, ce qui permet d'évaluer le mieux possible le trafic que le réseau devra écouler en chaque point de sa couverture. Dans ce contexte, chaque cellule peut être modélisée par un système indépendant de file d'attente de type M/G/c/c, c'est-à-dire un système à perte de c canaux avec arrivées de Poisson des appels et une durée d'occupation des canaux de distribution générale (Viterbi et Viterbi, 1993). La durée d'occupation d'un canal est définie par le temps pendant lequel l'usager garde ce canal

occupé dans une cellule donnée (Orlik et Rappaport, 1998). Un tel modèle, illustré à la Figure 2.4, permet d'évaluer la distribution du trafic dans chaque cellule de la manière suivante (Kleinrock, 1975; Medhi, 1991) :

$$p(n) = \frac{\left(\lambda/\mu\right)^n / n!}{\sum_{k=0}^{c} \left(\lambda/\mu\right)^k / k!} \tag{2.1}$$

où n est le nombre de terminaux actifs dans la cellule, λ le taux moyen d'arrivée des appels, c le nombre total de canaux disponibles et $1/\mu$ le temps moyen d'occupation des canaux. Notons que la libération d'un canal a lieu après une déconnexion résultant, soit d'une relève (passage à une cellule adjacente), soit de la fin d'une communication.

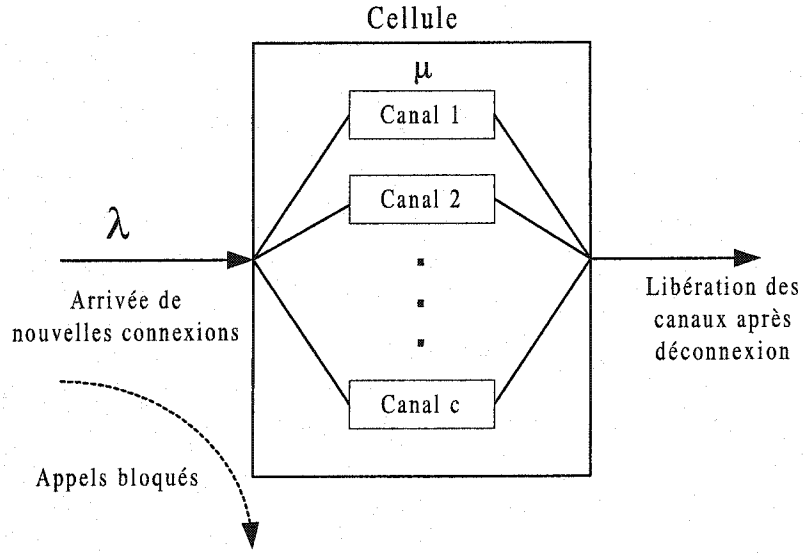


Figure 2.4 File d'attente modélisant une cellule

Mentionnons également que le trafic multimédia qui sera offert par les systèmes de la prochaine génération ressemble à celui de l'Internet (He *et al.*, 2000). Un tel trafic

se définit comme étant en rafales (*bursty*), asymétrique, et se caractérise par un taux d'erreur binaire, une taille de paquets et un taux d'arrivée aléatoires. Cela a d'ailleurs motivé certains chercheurs à utiliser des protocoles conçus principalement pour des réseaux câblés, tels que H.261, H.263, H.236, MPEG-1, MPEG-4, pour transporter le trafic de type vidéo dans les réseaux mobiles (Marchent *et al.*, 1999; Sevanto, 1999). Ces modèles sont, à toutes fins pratiques, inefficaces puisqu'ils ne tiennent pas compte de l'environnement de propagation, ni de la mobilité, ni de l'interopérabilité des réseaux.

2.3.2 Paramètres de qualité de service

Lors de la planification d'un réseau mobile, deux paramètres classiques de qualité de service peuvent être considérés : la probabilité de blocage et le temps d'attente d'un canal (Tabbane, 1997). Lorsqu'un usager veut générer ou recevoir un appel, le terminal essaie de lui octroyer un canal pour la connexion (Lin *et al.*, 1994). S'il n'existe pas de canaux disponibles pour cette connexion, l'appel est bloqué et effacé du système : on parle alors de blocage de nouvel appel (*new call blocking*). La probabilité de blocage est la probabilité qu'un abonné voit son appel bloqué. Par contre, si un canal est disponible lors de l'établisssement de la communication, il est utilisé pour la connexion, puis relâché soit à la fin de l'appel, soit lorsque l'usager quitte la cellule. Dans le même ordre d'idée, lorsque l'usager change de cellule (situation de relève) pendant une communication, un canal de communication doit lui être réservé dans la nouvelle cellule pour permettre à l'appel de se poursuivre. Si aucun canal n'est disponible dans la nouvelle cellule lors de la relève, l'appel est forcé de se terminer avant d'être complété. On parle alors de blocage d'appel de relève (*handoff call blocking*). Les probabilités de blocage de nouveaux appels et d'appels de relève sont considérées comme des paramètres importants d'évaluation de la qualité de service.

En modélisant chaque cellule par une file d'attente de type M/G/c/c, la probabilité de blocage peut être obtenue à partir de la relation (2.1) de la manière suivante :

$$p_B = \frac{\left(\lambda/\mu\right)^c / c!}{\sum_{k=0}^{c}\left(\lambda/\mu\right)^k / k!} \tag{2.2}$$

où c est le nombre de canaux disponibles dans la cellule considérée, λ le taux moyen d'arrivée des appels et $1/\mu$ le temps moyen d'occupation des canaux. Cette expression, connue sous le nom de la formule d'Erlang B, est indépendante de la distribution du temps d'occupation du canal (Medhi, 1991). Elle conditionne également la quantité de ressources à mettre à la disposition des abonnés pour satisfaire à leurs demandes.

D'autre part, pour éviter aux abonnés d'avoir à réitérer les appels infructueux, des files d'attente peuvent être mises en place lors d'une requête au niveau du système. L'opérateur, pour garantir une qualité de service acceptable, doit éviter qu'un usager attende trop longtemps pour voir sa demande satisfaite. La probabilité de cette attente est déterminée, à partir de la loi d'Erlang C, par la relation (Tabbane, 1997) :

$$P_D = \frac{\rho^c}{c!}\left[\frac{\rho^c}{c!} + \left(1 - \frac{\rho}{c}\right)\sum_{i=0}^{c-1}\frac{\rho^i}{i!}\right]^{-1} \tag{2.3}$$

où $\rho = \lambda/\mu$ est l'intensité de trafic dans la cellule considérée et c le nombre de canaux disponibles dans cette cellule. Dans le cas particulier d'une politique de type *premier arrivé premier servi* (*FIFO : First In First Out*), la probabilité que le délai d'attente respecte un certain seuil T_D suit une loi exponentielle dont la distribution est donnée par (Tabbane, 2000) :

$$P\left[t \leq T_D\right] = e^{-(c-\rho)t/h} \tag{2.4}$$

où h est la durée moyenne d'une connexion, c le nombre de canaux disponibles et ρ l'intensité de trafic du système. Dans le cas des systèmes de la prochaine génération, la relation (2.4) détermine la probabilité de respecter les contraintes de qualité de service imposées par les applications en temps réel.

2.3.3 Contraintes de qualité de service

La prochaine génération de réseaux mobiles offrira toute une série de services (l'accès rapide à l'Internet, le commerce électronique, la vidéoconférence, la télémédecine, l'apprentissage à distance, etc.) ayant chacun ses caractéristiques et contraintes particulières, en termes de taux d'erreur binaire (BER : *Bit Error Rate*), de délai maximal admissible, de variation de délai et de bande passante nécessaire. Il en résulte que chaque application doit spécifier au réseau ses exigences de qualité de service, en termes de type de trafic, de délai maximal de transfert, de variation de délai, de BER et de débit disponible. En particulier, pour les services mobiles de la voix, le BER doit être inférieur à 10^{-3}, alors que le délai admissible doit être inférieur à 300 ms. Pour les services de transfert de données (courrier électronique, transfert de fichiers), le BER doit être inférieur à 10^{-5} à travers le canal radio. Quant aux services multimédia (jeux interactifs, vidéoconférence), même s'ils peuvent tolérer un BER relativement élevé (entre 10^{-3} et 10^{-7}), ils ne supportent que des délais constants et inférieurs à 300 ms (Marchent *et al.*, 1999).

D'autre part, la capacité offerte par un système mobile dépend également de l'environnement radio utilisé. En effet, si cette capacité est limitée à 144 kbps dans les environnements desservis par des satellites ou dans les zones rurales, elle peut atteindre 384 kbps dans les zones urbaines (micro-cellules) et 2048 kbps à l'intérieur des édifices et dans les zones à faible portée ou pico-cellules (Marchent *et al.*, 1999). Le tableau 2.1 spécifie les contraintes de qualité de service des différents types d'applications des réseaux de la prochaine génération, en fonction de l'environnement d'exploitation.

2.3.4 Mobilité des usagers

Contrairement aux réseaux fixes où un abonné peut être rejoint à tout instant en un point fixe, les réseaux mobiles offrent à chaque usager la possibilité de se déplacer à son gré dans la zone desservie par son fournisseur. Il importe alors de déterminer la position de tout abonné avant de lui acheminer les communications qui lui sont dédiées. Cela se

réalise en subdivisant la zone de service en zones de localisation (LA : *Location Area*). Chaque LA regroupe un certain nombre de cellules et est contrôlée par un MSC (Lin et Chlamtac, 1996). De plus, chaque terminal doit mettre à jour ses informations de localisation à chaque fois qu'il entre dans une nouvelle LA. Cela nécessite l'utilisation des bases de données HLR et VLRs qui maintiennent le profil, ainsi que la zone de localisation de chaque abonné (Wang et Akyildiz, 2000).

Tableau 2.1 Contraintes de délais en fonction du type de service

	Services en temps réel (BER = 10^{-3} à 10^{-7})	Services non en temps réel (BER = 10^{-5} à 10^{-8})
Environnement d'exploitation (Vitesse maximale)	Délai maximal admis (ms)	Délai relatif (ms)
Satellite (1000 km/h)	400	Supérieur à 1200
Rural (500 km/h)	20 – 300	Supérieur à 150
Urbain (120 km/h)	20 – 300	Supérieur à 150
Intérieur/Extérieur à faible portée (10 km/h)	20 – 300	Supérieur à 150

La gestion de la mobilité des usagers comporte deux processus : la gestion de la localisation qui se réalise au niveau du sous-système réseau, et la gestion de la relève qui se passe au niveau du sous-système radio (Tabbane, 1997; Wang et Akyildiz, 2000). La gestion de la localisation est réalisée par des échanges de messages de signalisation à travers un réseau de signalisation, appelé réseau SS7 (*Signaling System* 7). Il existe

actuellement deux standards de gestion de localisation : la norme IS-41 (*Interim Standard* 41) proposée par EIA/TIA (*Electronic/Telephone Industry Associations*) et la norme GSM MAP (*Mobile Application Part*) (Akyildiz *et al.*, 1998). La norme IS-41 est souvent utilisée en Amérique du Nord (AMPS, IS-54, IS-136 et PACS), alors que la norme GSM MAP est principalement utilisée en Europe (GSM, DCS-1800, PCS-1900). Les deux standards se basent sur une hiérarchie de bases de données à deux niveaux (Akyildiz *et al.*, 1998).

Par ailleurs, la gestion de la localisation est un processus qui consiste à identifier le point d'attache de tout terminal en vue de l'acheminement des communications. Ce processus se déroule en deux phases. La première phase constitue l'enregistrement (ou la mise à jour) de la localisation, où la position courante du terminal est enregistrée (ou mise à jour) au niveau des bases de données appropriées (VLR, HLR) (Park et Lin, 1997). La deuxième phase est l'acheminement de la communication, ce qui est normalement réalisé en deux étapes : détermination du VLR qui dessert le terminal appelé et localisation de la cellule visitée par le terminal. Dans ce contexte, la plupart des méthodes de gestion de localisation visent à améliorer la norme IS-41, tout en maintenant inchangée l'architecture des bases de données (Jain *et al.*, 1994; Markoulidakis *et al.*, 1995; Safa *et al.*, 2000). À ce niveau, on relève plusieurs défis de recherche, dont la sécurité, la mise à jour dynamique des bases de données et l'optimisation du délai de recherche des usagers.

Par ailleurs, la gestion de la relève est un processus qui permet à une communication en cours de se poursuivre lorsqu'un usager mobile se déplace d'une cellule à une autre (McNair et *al.*, 2000). Ce processus se déroule en trois phases. La première phase se réfère à une certaine initialisation où, soit l'usager, soit un agent, soit l'état courant du réseau identifie le besoin de relève. La deuxième phase consiste à générer une nouvelle connexion par laquelle le réseau trouve de nouvelles ressources pour effectuer le routage de la communication. La dernière phase consiste en un contrôle du flux d'informations, en acheminant les données de l'ancien chemin au nouveau selon certaines contraintes de qualité de service.

Au cours de la dernière décennie, la gestion de la mobilité a été largement analysée pour les réseaux mobiles classiques (Jain *et al.*, 1994; Markoulidakis *et al.*, 1995; Safa *et al.*, 2000). En général, les méthodes de résolution se divisent en deux catégories (Tabbane, 1997). La première catégorie comprend toutes les approches qui utilisent des algorithmes statiques, basés sur l'architecture du réseau, alors que la deuxième catégorie englobe les méthodes dynamiques, basées sur le processus d'apprentissage et qui requièrent des statistiques sur la mobilité des usagers (Tabbane, 1997; Wong et Leung, 2000).

Toutefois, dans les réseaux mobiles de la prochaine génération, les usagers auront la possibilité de circuler au travers des réseaux de types différents, exploités par des opérateurs différents et situés dans des zones géographiques différentes. On parle alors de mobilité globale. Les méthodes de gestion de mobilité globale visent à trouver un bon compromis entre la localisation précise des usagers et les performances de la gestion de l'itinérance, en termes de rapidité (faible temps de réponse du réseau) et de trafic de signalisation généré dans le réseau (faible taux de mises à jour et d'enregistrements effectués dans les bases de données).

2.4 Formulation du problème de planification

La planification d'un système mobile peut avoir l'un des objectifs suivants : garantir une bonne qualité de service par l'utilisation d'un lien radio fiable et disponible dans la zone de service (faible interférence et faible probabilité de blocage), ou minimiser les coûts des équipements tout en maintenant une communication de qualité et une capacité élevée. Il s'agira alors de déterminer une configuration économique des sous-systèmes réseau et radio, tout en maintenant une bonne performance et une exploitation efficace du réseau. Cela peut se réaliser en utilisant essentiellement les informations sur la géographie et les caractéristiques de propagation des zones de couverture, ce qui permet de déterminer le nombre, la localisation et les capacités des différents commutateurs et bases de données, ainsi que le réseau d'interconnexion des différents

nœuds. Cette interconnexion doit se faire en minimisant les coûts d'investissement des nœuds, ainsi que les coûts des liaisons de transmission et les ressources mises en œuvre pour la gestion du trafic. Dans certains systèmes de la première génération, le coût d'implantation d'une station de base peut atteindre le million de dollars US (Tabbane, 1997). Il est donc clair qu'une bonne planification permettant d'optimiser l'utilisation des équipements revêt un intérêt fondamental pour les opérateurs.

Toutefois, il importe de tenir compte d'un certain nombre de contraintes. D'abord, le nombre limité de ressources disponibles (c'est-à-dire le nombre de canaux, la capacité des équipements) invite les concepteurs des réseaux mobiles à utiliser les ressources à bon escient et à s'assurer que de tels réseaux sont toujours en mesure de répondre à la demande des usagers sans pour autant affecter la qualité de service offerte. Ensuite, les concepteurs doivent s'assurer que la qualité et la fiabilité de la transmission radio soient optimales dans chaque zone de service. Cela se fait essentiellement en positionnant les stations de base (BSs, BTSs ou nœuds B) de telle manière que le signal reçu soit maximisé et l'interférence minimisé en tout point de la zone de service.

Les systèmes de la première génération étaient relativement simples à concevoir. Cette simplicité résulte de leur faible capacité et du nombre faible de fonctions qu'ils intégraient. Toutefois, en ce qui a trait aux systèmes de la deuxième génération, les concepteurs en avaient défini une série de nouvelles fonctions qui devaient être implantées dans des équipements physiques et qui, de ce fait, complexifiaient le processus de planification. Quant aux systèmes de la prochaine génération, ils doivent être conçus pour être à la fois flexibles et modulaires, c'est-à-dire extensibles en termes de services à offrir et du nombre d'usagers à supporter. Cette flexibilité peut également s'exprimer en termes d'adaptabilité à des environnements de propagation et de trafic différents et variables dans le temps, ou encore en termes de facilité de la gestion des ressources radio, de la capacité d'interopérer avec différents systèmes, et d'accommoder plusieurs types de cellules (pico, micro et macro cellules), ainsi que plusieurs opérateurs dans la même aire de services (partage efficace d'un spectre commun, partage d'infrastructures et possibilité de handoffs entre systèmes hétérogènes exploités par

différents opérateurs). Dans ce contexte, le problème de planification se formule de la manière suivante :

Étant donné :

- les caractéristiques de l'environnement à couvrir (caractéristiques géographiques et de propagation radio);
- le nombre de cellules ou de stations de base dans chaque sous-système;
- les caractéristiques des stations de base (comme la puissance et la hauteur) de chaque sous-système;
- les bandes de fréquences allouées à chaque sous-système;
- la technologie d'accès utilisée dans chaque sous-système;
- le profil des usagers à desservir (densité, comportement statistique);

Minimiser : le coût total de l'infrastructure radio et réseau

En fonction :

- de la distribution du trafic dans chaque cellule;
- du modèle de mobilité globale;
- de la définition de la couverture radio;
- de l'étendue de la zone à couvrir;
- de l'allocation des ressources;
- de l'architecture du réseau (localisation des stations de base, du réseau de signalisation et des bases de données);
- du plan de fréquences;

En respectant les contraintes :

- de délai de connexion, de qualité de la transmission (interférence négligeable par rapport au signal) et de probabilité de blocage.

Même si ce problème paraît simple, il peut résulter en un processus long qui fait appel à toute une série de sous-problèmes, incluant la détermination du modèle de mobilité et de trafic, la définition de la couverture radio, le dimensionnement des cellules, la planification des fréquences, la définition du réseau de commutation, ainsi que la planification du réseau de signalisation et des bases de données.

On se rend compte que le nombre de paramètres à prendre en compte est supérieur au cas des réseaux fixes. En outre, les conditions de trafic et de propagation, la mobilité des usagers et les services varient constamment dans le temps. Il en résulte que chaque opérateur doit être à l'écoute de ces changements pour déterminer les nouvelles valeurs des paramètres de fonctionnement et réajuster les paramètres liés aux procédures de recherche, de handoff, de contrôle de puissance, de gestion de la ressource et des algorithmes de mise à jour de localisation, de façon à optimiser l'utilisation des ressources radio. Cela permet d'offrir des communications de qualité et une capacité élevée.

2.5 Étapes de planification

Le processus de planification se déroule essentiellement en deux phases: la planification radio qui consiste à sélectionner les sites radio à l'aide d'outils de planification, et la phase d'ingénierie qui comprend le dimensionnement des différents nœuds du réseau.

2.5.1 Planification de la partie radio

La planification de la partie radio se déroule en plusieurs étapes qui aboutissent à la détermination des sites des stations de base et à l'affectation des fréquences aux cellules. L'étape de détermination des sites consiste à identifier, à partir d'une configuration théorique donnée, les sites pratiques en fonction de la géographie de la zone et des caractéristiques de propagation. Les points les plus élevés sont alors passés au crible pour trouver les sites idéaux (immeubles appartenant à l'opérateur, pylônes, etc.), alors que les paramètres récoltés au cours des campagnes de mesures sont analysés par un outil qui calcule les champs d'interférence, dessine la couverture et fixe les points d'implantation des stations de base. Dans le cas des systèmes micro-cellulaires, la planification est différente et nécessite le recours aux cartes à haute définition, où les détails des structures de terrains sont pris en compte (Tabbane, 2000). Les antennes peuvent alors se situer en dessous des toits, ce qui donne lieu à des modèles de propagation différents.

À la fin de la première étape de planification, le nombre, la capacité et le rayon des cellules sont connus. L'étape suivante, appelée affectation des fréquences, consiste à allouer à chaque site un certain nombre de canaux de façon à ce que le niveau d'interférence entre les cellules soit minimal et que le niveau de puissance en tout point de la couverture soit suffisant pour garantir une qualité de service acceptable. Les méthodes théoriques de résolution d'un tel problème reposent essentiellement sur des modèles de réseaux à structure régulière hexagonale, où chaque canal est réutilisé selon un motif régulier et fixe. L'affectation des fréquences étant un problème complexe, difficile à résoudre, sa planification ne peut être réalisée que par le recours à des heuristiques. Deux types d'heuristiques peuvent être alors appliqués : les méthodes déterministes qui conduisent à un plan de fréquences unique et les méthodes non déterministes qui peuvent conduire à plusieurs plans de fréquences (Tabbane, 1997).

2.5.2 Planification de la partie réseau

La planification de la partie réseau consiste à déterminer les capacités et emplacements des différents concentrateurs et bases de données, ainsi que les liaisons de transmission entre ces équipements en fonction de la distribution du trafic et des positions des sites radio. Cette phase comprend généralement la détermination du nombre et de la localisation des contrôleurs de stations de base (BSCs ou RNCs) et des commutateurs du service mobile (MSCs), la définition du réseau d'interconnexion des stations de base aux BSCs et des BSCs aux MSCs, la définition des connexions entre les MSCs, ainsi que la détermination de la capacité et de l'emplacement des bases de données (HLR/HSS et VLRs). Généralement, plusieurs paramètres sont fixés à l'avance pour faciliter la résolution de ce problème. Par exemple, la localisation des MSCs, ainsi que les capacités des liaisons entre MSCs et BSCs peuvent être fixées au départ.

L'objectif principal consiste à minimiser les coûts d'investissement des BSCs (ou RNCs) et des liaisons d'interconnexion des nœuds, ainsi que les coûts attribués aux ressources mises en œuvre pour la gestion du trafic. En effet, les moyens mis en œuvre pour l'échange des signaux entre les différents sites du réseau représentent jusqu'à 35% des coûts d'opération du réseau (Tabbane, 1997). Dans ce cas, il importe de tenir compte des échanges de signalisation engendrés par les handoffs, les mises à jour de localisation et les opérations relatives à l'itinérance, ce qui génère une quantité importante de trafic entre les MSCs. Dans certains systèmes, la surcharge des commutateurs peut atteindre 30% de la charge totale. Il importe alors de répartir les cellules entre les différents commutateurs de manière à distribuer uniformément cette charge dans le réseau, tout en minimisant la surcharge engendrée par les communications inter-commutateurs.

Mentionnons que le processus complet de planification peut être itératif, étant donné que les valeurs de certains paramètres ne peuvent être déterminées qu'une fois certains éléments du réseau sont fixés. De plus, la planification d'un réseau mobile est un processus continu. En fait, les opérateurs de certains réseaux (comme le GSM en Europe) ajoutent des stations de base toutes les semaines et sont obligés de replanifier leur réseau tous les mois dans les phases de forte croissance (Tabbane, 1997).

2.6 Synthèse des approches traditionnelles de planification

Les concepteurs des systèmes de la première génération adoptaient une approche plutôt pragmatique, nécessitant une implémentation par étape (Bahai et Aghvami, 2000). Il s'agissait tout simplement de localiser les sites d'installation des stations de base à des positions déterminées par des ingénieurs expérimentés, de prendre des mesures sur le champ (une fois le système est opérationnel) et d'apporter les ajustements appropriés. Une telle approche tendait généralement à surdimensionner le système. C'est pourquoi, dans la planification des réseaux mobiles de la deuxième génération, une approche ascendante (*bottom-to-top*) a été adoptée (Aghvami et Jafarian, 2000). Par cette approche, une interface radio est d'abord conçue et optimisée pour la transmission de la voix; ensuite, l'épine dorsale (*core*) du réseau est développée pour supporter ce service. Certains chercheurs adaptent cette méthodologie à la conception des réseaux mobiles de la prochaine génération, en choisissant d'abord les interfaces radio, puis en développant les réseaux supportant de telles interfaces et en identifiant les services qui correspondent aux besoins des utilisateurs. Cependant, une telle approche est inappropriée pour des systèmes offrant des services avec une qualité de service et des performances variables. Il en résulte plusieurs limitations, dont l'utilisation inefficace de la technologie d'accès.

Par ailleurs, Aghvami et Jafarian (2000) ont proposé une approche descendante (*top-to-bottom*) de planification qui se décompose en quatre étapes : définition des services et des applications, mise en place de l'épine dorsale du réseau, définition du modèle de mobilité et détermination de la technologie d'accès. La première étape consiste à spécifier et à modéliser les services et applications qui doivent être offerts, tant aux usagers fixes que mobiles. Ensuite, un ensemble de protocoles doivent être spécifiés et implantés au niveau de l'épine dorsale pour gérer les services à offrir. La prochaine étape consiste à concevoir les protocoles de gestion de la mobilité et à développer le réseau d'accès qui livre les services aux usagers. Nous pouvons toutefois remarquer que cette approche ne fait nullement mention de l'analyse du trafic et, de ce fait, ne peut s'appliquer efficacement à la planification des réseaux mobiles de la prochaine génération.

Une autre approche de planification souvent utilisée est illustrée à la Figure 2.5 (Tutschku et Tran-Gia, 1998). Il s'agit d'un processus itératif qui pose tout d'abord un certain nombre d'objectifs liés à la couverture du réseau. Ensuite, utilisant les données géographiques, démographiques, ainsi que l'analyse de la propagation, on détermine une architecture initiale du système. Cette architecture est optimisée, en intervenant itérativement au niveau des variables du problème. Cependant, cette approche met surtout l'emphase sur la planification de la partie radio du processus, c'est-à-dire la détermination des sites radio, l'affectation des fréquences et le dimensionnement des antennes. D'autres aspects, comme le comportement des usagers et le trafic, sont pris en compte séparément et un peu plus tard dans le processus de planification. On se rend alors compte du manque d'interaction entre les différents aspects de planification. De plus, une telle méthodologie ne tient nullement compte de l'évolutivité, ni de l'interopérabilité des sous-réseaux du système, ce qui peut aboutir à des résultats non réalistes. Ainsi, pour pallier les lacunes des différentes approches présentées, nous proposons au chapitre suivant une nouvelle approche de planification des systèmes mobiles de la prochaine génération.

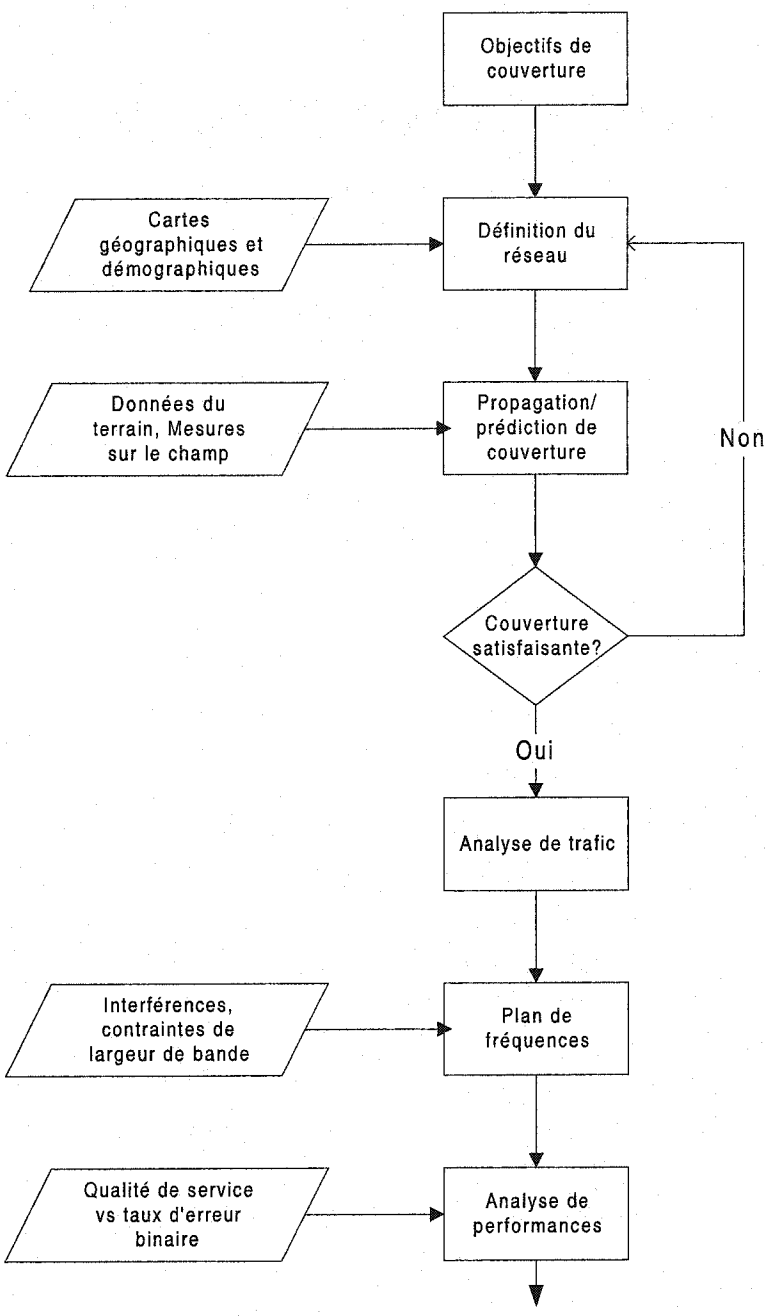


Figure 2.5 Approche traditionnelle de planification

CHAPITRE 3

MÉTHODOLOGIE

DE PLANIFICATION PROPOSÉE

Les concepteurs des systèmes mobiles de la prochaine génération doivent faire face à plusieurs nouveaux défis, dont l'augmentation considérable des demandes de services, la compatibilité avec les systèmes de première et de deuxième générations, l'avènement des nouvelles technologies et l'offre d'une qualité de service flexible. Tous ces facteurs ne font que compliquer le problème de planification qui, déjà, ne pouvait pas être directement résolu dans sa globalité. Dans ce contexte, nous proposons une approche modulaire qui subdivise le problème global en des sous-problèmes plus faciles à résoudre. Ce chapitre présente une vue d'ensemble d'une telle approche, ainsi que la description de chacun de ses modules. Plus spécifiquement, il décrit les principes d'opération de la méthodologie, expose les problèmes traités par chaque module et analyse les résultats obtenus eu égard au compromis entre la puissance à fournir et la capacité à offrir pour maintenir un certain niveau de qualité de service.

3.1 Caractérisation de la méthodologie

Pour résoudre efficacement le problème de planification des systèmes de la prochaine génération, nous proposons une approche qui identifie et fait interagir cinq facteurs fondamentaux d'ingénierie de systèmes pour contrôler le processus de planification. Ce sont : la couverture radio, l'architecture, l'allocation des ressources, l'itinérance (ou la mobilité) globale et l'ingénierie de trafic. Chaque facteur est implémenté dans un module qui traite d'un aspect particulier du problème de planification. Ces modules seront appelés à ajuster les variables appropriées, à chaque

fois que sont apportés des changements majeurs, dus au comportement des usagers, à des ajouts de services ou à la variabilité du trafic. La structure modulaire de la méthodologie est illustrée à la Figure 3.1. Pour chaque module, les entrées/sorties et les paramètres échangés avec les autres modules sont présentés aux figures 3.2, 3.4, 3.6, 3.7 et 3.8. À la section 3.2, nous justifierons le choix de ces paramètres. Notons que les modules *Couverture Radio*, *Architecture* et *Allocation des ressources* sont traités à la section 3.2, alors que les modules *Ingénierie du trafic* et *Itinérance globale* font respectivement l'objet des chapitres 4 et 5.

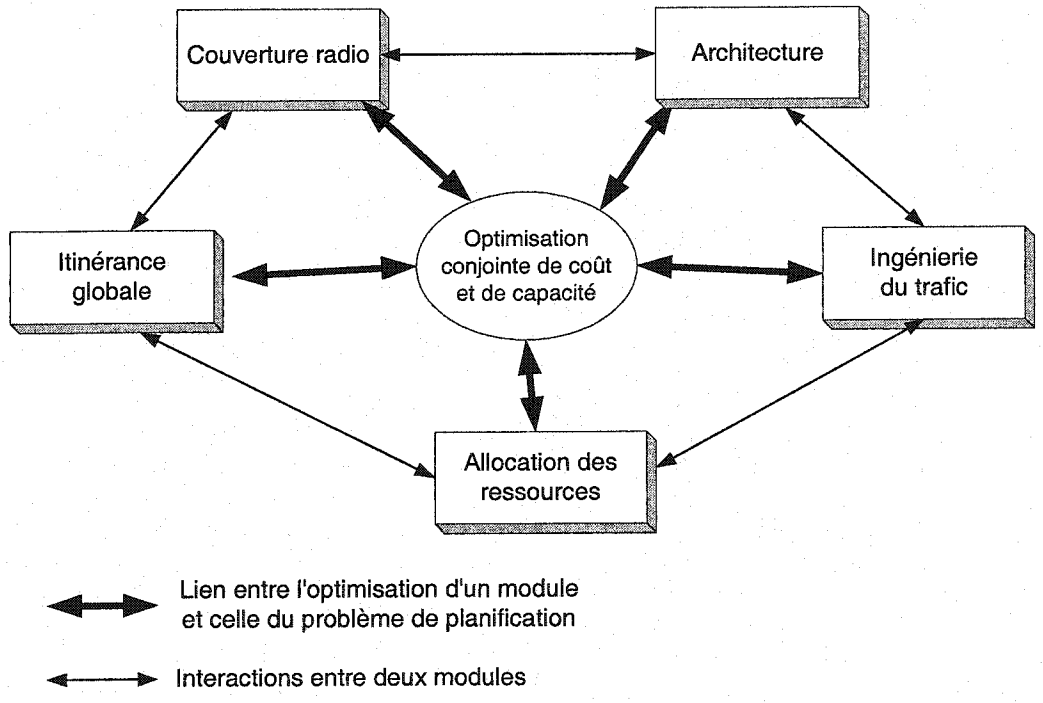


Figure 3.1 Méthodologie de planification proposée

Par ailleurs, la prise en compte des interactions entre les divers facteurs permet d'aboutir à des solutions optimisées sous différents aspects. Par exemple, les modules *Couverture radio* et *Architecture* doivent tenir compte des paramètres des modules *Ingénierie du trafic* et *Itinérance globale* pour ajuster et optimiser parallèlement leurs

paramètres. En même temps, ces modules doivent s'assurer qu'une telle optimisation n'affecte pas les paramètres du module *Allocation des ressources*. Dans le même ordre d'idées, s'il existe un échange entre les modules *Allocation des ressources* et *Couverture radio*, cet échange doit se faire de manière à ne pas affecter les paramètres du module *Architecture*. Il en résulte que la méthodologie préconisée se veut évolutive et contribue à concevoir un système qui supporte le plus d'usagers possible, tout en répondant à leurs besoins, en termes de disponibilité du système, de délai de connexion et de qualité de la communication. Nous apporterons ainsi une solution économiquement viable et offrant une bonne qualité de service au problème de planification.

3.2 Description des modules

Dans cette section, nous présentons le rôle et les fonctions des cinq modules faisant partie de la méthodologie préconisée, en l'occurrence la couverture radio, l'architecture, l'allocation des ressources, l'itinérance globale et l'ingénierie du trafic.

3.2.1 Couverture radio

La couverture radio se caractérise par la probabilité d'établir une communication de qualité dans la zone de service. Pour garantir cette qualité, une bonne méthode de conception doit à la fois s'assurer d'un niveau de signal suffisant en tout point de la zone de service et minimiser les perturbations provenant des interférences. Autrement dit, les sites des stations de base (BSs, BTSs ou nœuds B) doivent se situer à des positions qui maximisent en tout point du réseau le rapport *S/I*, où *S* désigne la puissance du signal reçu et *I* celle des interférences. Ce problème de couverture peut être mathématiquement formulé de la manière suivante (Tutschku, 1998) :

$$\text{Maximiser } Y = \sum_{i \in I} a_i \, y_i$$

Sujet à

$$\sum_{j \in N_i} x_j \geq y_i, \quad \forall i \in I \tag{3.1}$$

$$\sum_{j \in J} x_j = p, \quad \forall i \in I, \quad \forall j \in J \tag{3.2}$$

où J est l'ensemble des emplacements potentiels des BSs, I l'ensemble des nœuds de demandes (usagers mobiles), a_i la portion de trafic générée par l'abonné i, N_i l'ensemble des stations de base qui offrent une puissance suffisante au nœud i, p le nombre maximal de BSs à déployer. Les variables de décision x_j et y_i sont définies de la manière suivante :

$$x_j = \begin{cases} 1, \text{ s'il existe une station de base à la position } j \\ 0, \text{ sinon} \end{cases} \tag{3.3}$$

$$y_i = \begin{cases} 1, \text{ si un usager mobile } i \text{ est desservi par une station de base} \\ 0, \text{ sinon} \end{cases} \tag{3.4}$$

En zone urbaine, la fourniture d'une couverture adéquate ne pose généralement pas de problème d'investissement puisque le trafic écoulé dans ces zones permet d'obtenir un retour rapide sur investissement. Par contre, les zones rurales posent plus de problèmes, car l'opérateur doit y assurer une certaine couverture sans que le trafic généré dans ces zones ne soit suffisant pour lui assurer des revenus équivalents. Ainsi, la planification peut avoir des objectifs différents en fonction de la zone à planifier. En zone urbaine, l'objectif sera d'assurer une capacité suffisante en trafic, c'est-à-dire desservir un nombre élevé d'abonnés, alors qu'en zone rurale ou à faible densité d'abonnés, l'objectif sera plutôt d'assurer la couverture la plus complète possible sans pour autant nécessiter de capacité élevée.

Par ailleurs, pour les systèmes mobiles de la prochaine génération (basés sur la technologie CDMA), le niveau d'interférence est également sensible au nombre d'abonnés en communication dans le réseau (Tutschku et Tran-Gia, 1998). Dans ce contexte, la couverture radio est intimement liée à la capacité du système, c'est-à-dire chaque cellule peut supporter un certain nombre de terminaux actifs, tout en maintenant

une qualité de service acceptable. Toutefois, à partir du moment que l'on dépasse ce nombre, le réseau devient surchargé et le niveau d'interférence augmente très vite, résultant en une brusque détérioration de la qualité de la communication. Dans le but de préserver la qualité des connexions existantes, le système se doit de refuser les nouvelles demandes de connexions. Ce phénomène constitue un refus de service (*soft blocking* ou *outage*). Ici, nous parlons de refus de service et non de blocage pour éviter toute confusion avec le blocage traditionnel qui correspond à une indisponibilité de canal pour desservir les nouveaux appels.

Concrètement, le refus de service survient lorsque le nombre d'usagers en communication dans le réseau génère une quantité d'interférence I_0 qui dépasse le niveau de bruit N_0 d'une certaine quantité $1/\eta$ dont la valeur nominale est de 10 dB (Viterbi et Viterbi, 1993). Dans ce contexte, en tenant compte de l'atténuation du lien radio, de la sensibilité du récepteur et des gains des antennes, le rapport signal à bruit minimal requis par usager pour maintenir une communication de qualité peut être déterminée de la manière suivante (Viterbi et Viterbi, 1993) :

$$\left(\frac{S}{N_0 W}\right)_{\min} = \frac{\left(1/\eta\right) - 1}{\rho\left(\lambda/\mu\right)} \tag{3.5}$$

où S est la puissance du signal reçu, N_0 la puissance du bruit, W la bande passante disponible, ρ le facteur d'activité de la voix, $1/\eta = I_0/N_0$ le rapport interférence à bruit et λ/μ la charge du réseau pour laquelle la probabilité de refus de service P_{out} est de 0.01. Le facteur d'activité de la voix est un paramètre qui indique le pourcentage de temps pendant lequel intervient un abonné durant une communication. La relation (3.5) permet non seulement d'évaluer la capacité du système, mais aussi de spécifier les conditions qui permettent de maintenir une bonne qualité de service.

Ainsi, le module *Couverture radio* doit intégrer un algorithme efficace pour s'assurer que la qualité de la transmission radio est optimale dans la zone de service. Autrement dit, ce module doit garantir un niveau de signal suffisant dans la région de

planification, tout en minimisant les probabilités d'interférence et de refus de service. Ses opérations sont illustrées à la Figure 3.2. Mentionnons que l'interaction entre les modules *Couverture radio* et *Architecture* signifie que le processus de planification doit tenir compte des interférences et des caractéristiques de propagation pour la mise en place de l'infrastructure du réseau. Autrement dit, les BSs doivent être positionnés de manière telle que la puissance du signal reçu soit maximale et que celle de l'interférence provenant des autres sources soit minimale.

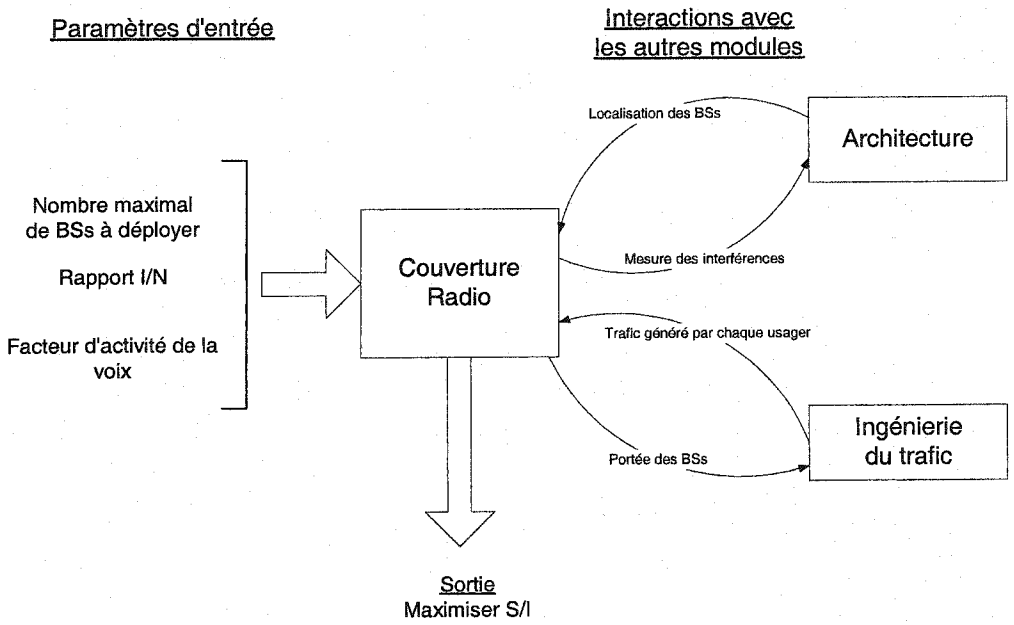


Figure 3.2 Opérations du module *Couverture radio*

3.2.2 Architecture

Pour offrir un degré élevé de flexibilité à chacune des composantes du réseau, il est souvent recommandé de concevoir indépendamment le sous-système réseau (*NSS : Network Sub-System*) du sous-système radio (*BSS : Base Station Sub-system*) (Bahai et Aghvami, 2000). Toutefois, il faut se rendre compte qu'en pratique, les deux sous-

systèmes sont interdépendants les uns des autres, car le traitement et la gestion du trafic provenant du BSS se font en grande partie au niveau du NSS. Dans ce contexte, le module *Architecture* se doit non seulement de concevoir le BSS et le NSS, mais aussi d'interconnecter ces sous-systèmes de façon optimale, en respectant une série de contraintes.

La planification du sous-système radio vise à sélectionner, localiser et configurer un ensemble de stations de base (en termes de puissance de transmission, hauteur et directionnalité des antennes) de manière à optimiser les coûts d'installation. Ces choix de paramètres déterminent les dimensions et la capacité des cellules. Actuellement, la tendance est à la réduction de la taille des cellules, ce qui conduit au déploiement intensif de micro-cellules, particulièrement dans les régions métropolitaines (Ganz *et al.*, 1997). En général, les micro-cellules offrent une capacité élevée, en plus de permettre à la fois de respecter les exigences de qualité de service et d'utiliser efficacement le spectre radio (Lagrange, 1997). Toutefois, le nombre élevé de stations de base déployées dans les micro-cellules peut conduire à une augmentation considérable des coûts de planification. Il importe alors de déployer dans la zone de service une combinaison de micro et de macro-cellules pour avoir un bon compromis entre le coût et la capacité. Ganz *et al.* (1997) ont d'ailleurs montré que, pour les mêmes contraintes de qualité de service, l'utilisation de cellules de tailles différentes aide à réduire considérablement les coûts de planification. C'est l'une des raisons pour laquelle les systèmes mobiles de la prochaine génération intégreront plusieurs types de cellules : les pico-cellules (pour les centres d'achat), les micro-cellules (pour le centre-ville), les macro-cellules (pour les zones à faible densité de trafic) et les satellites pour desservir les zones où il est difficile d'installer des stations de base.

Dans le même ordre d'idées, la planification du NSS vise à déterminer les capacités et emplacements optimaux des commutateurs du service mobile (MSC : Mobile Switching Center), des contrôleurs de stations de base (BSCs ou RNCs), des bases de données nominales (HLRs ou HSSs) et des bases de données visiteurs (VLRs), ainsi que les liaisons d'interconnexion de ces équipements, en fonction de la distribution du trafic

et des positions des sites radio (Tabbane, 2000). Pour faciliter la résolution d'un tel problème, plusieurs paramètres, tels que la localisation des MSCs, ainsi que les capacités des liaisons entre MSCs et BSCs, peuvent être fixés à l'avance. Il s'agit d'un problème de conception topologique qui peut être résolu soit par des techniques d'intelligence artificielle (Pierre, 1998), soit par des méthodes heuristiques de recherche (Pierre et Elgibaoui, 1997).

La dernière étape de la planification d'architecture consiste à interconnecter le sous-système radio au sous-système réseau. Cette phase est généralement laissée de côté par les concepteurs des réseaux mobiles traditionnels qui préfèrent garder l'indépendance de ces deux sous-systèmes (Tutschku, 1996). Une telle pratique peut conduire au surdimensionnement de l'un des sous-systèmes et à une surcharge en trafic de l'autre, d'où une mauvaise utilisation des équipements. Pour pallier cette carence, le module *Architecture* doit intégrer un algorithme efficace d'affectation des stations de bases (BTSs ou nœuds B) aux MSCs. Cette affectation suppose que les BTSs sont directement branchés aux MSCs, c'est-à-dire les contrôleurs de stations de base sont intégrés aux MSCs, comme dans le modèle CDMA2000 (Smith et Collins, 2002). Dans ce cas, le problème d'affectation consiste essentiellement à minimiser une fonction de coût composée du coût des liaisons et du coût des relèves sous des contraintes de capacités des MSCs, en interconnectant les BTSs aux MSCs de façon à régulariser le trafic émergeant de la partie radio vers la partie réseau, minimisant ainsi le trafic de signalisation. Nous allons formuler le problème général d'affectation tel que présenté par Merchant et Sengupta (1995), ainsi que par Pierre et Houéto (2002).

Considérons un ensemble de n BTSs à affecter à m MSCs. La localisation des BTSs et des MSCs est fixe et connue, résultant des opérations de planification du NSS et du BSS. Soit H_{ij} le coût par unité de temps d'une relève simple (impliquant un seul MSC) entre les cellules i et j, et H'_{ij} le coût par unité de temps d'une relève complexe (impliquant deux MSCs) entre les cellules i et j ($i, j = 1, ..., n$ *avec* $i \neq j$). Alors, les paramètres H_{ij} et H'_{ij} sont proportionnels à la fréquence des relèves entre les cellules i et j. Soit c_{ik} le coût de la liaison qui relie la cellule i (c'est-à-dire BTS i) au MSC k, λ_i le

nombre d'appels par unité de temps générés ou destinés à la cellule i et M_k la capacité d'un MSC en nombre d'appels par unité de temps.

Posons

$$x_{ik} = \begin{cases} 1, \text{ si la cellule } i \text{ est reliée au MSC } k \\ 0, \text{ sinon.} \end{cases} \tag{3.6}$$

L'affectation des cellules (ou BTSs) aux MSCs est sujette à un certain nombre de contraintes. En effet, chaque cellule doit être assignée à un et un seul MSC, ce qui se traduit par la relation suivante :

$$\sum_{k=1}^{m} x_{ik} = 1 \quad \text{pour} \quad i = 1, ..., n \ . \tag{3.7}$$

De plus, la capacité limitée des MSCs impose la contrainte suivante :

$$\sum_{i=1}^{n} \lambda_i x_{ik} \leq M_k \quad \text{pour } k = 1, ..., m \tag{3.8}$$

Cela implique que la charge totale de toutes les cellules affectées à un commutateur ne doit pas dépasser la capacité de ce commutateur.

Pour déterminer la fonction objectif, considérons deux variables binaires z_{ijk} et y_{ij} définies par :

$$z_{ijk} = x_{ik} x_{jk} \text{ pour } i, j = 1, ..., n \text{ et } k = 1, ..., m, \text{ avec } i \neq j. \tag{3.9}$$

$$y_{ij} = \sum_{k=1}^{m} z_{ijk} \text{ pour } i, j = 1, ..., n, \text{ et } i \neq j. \tag{3.10}$$

La variable z_{ijk} est égale à 1 si les cellules i et j (avec $i \neq j$) sont toutes deux connectées au même commutateur k, sinon elle est nulle. Il en résulte que la variable y_{ij} prend la valeur 1 si les cellules i et j sont toutes deux connectées au même commutateur, et la valeur 0 si les cellules i et j sont connectées à des commutateurs différents. Ainsi, le coût par unité de temps f s'exprime comme suit:

$$f = \sum_{i=1}^{n}\sum_{k=1}^{m} c_{ik}x_{ik} + \sum_{i=1}^{n}\sum_{j=1, j\neq i}^{n} H'_{ij}(1-y_{ij}) + \sum_{i=1}^{n}\sum_{j=1, j\neq i}^{n} H_{ij}y_{ij} \qquad (3.11)$$

Le premier terme est le coût de liaison, alors que le deuxième prend en compte le coût des relèves complexes et le troisième celui des relèves simples. Mentionnons que la fonction de coût est quadratique en x_{ik}, car y_{ij} est une fonction quadratique des x_{ik}.

À ce niveau, nous pouvons simplifier le problème en posant :

$$h_{ij} = H'_{ij} - H_{ij}$$

où h_{ij} est le coût réduit par unité de temps d'une relève complexe entre les cellules i et j. La relation (3.11) devient alors :

$$f = \sum_{i=1}^{n}\sum_{k=1}^{m} c_{ik}x_{ik} + \sum_{i=1}^{n}\sum_{j=1, j\neq i}^{n} h_{ij}(1-y_{ij}) + \underbrace{\sum_{i=1}^{n}\sum_{j=1, j\neq i}^{n} H_{ij}}_{cons\,\tan te}$$

Ce qui permet de formuler le problème d'affectation de la façon suivante :

Minimiser

$$f = \sum_{i=1}^{n}\sum_{k=1}^{m} c_{ik}x_{ik} + \sum_{i=1}^{n}\sum_{j=1, j\neq i}^{n} h_{ij}(1-y_{ij}) \qquad (3.12)$$

Sujet à :

$$x_{ik} = 0 \text{ ou } 1 \text{ pour } i = 1, \ldots, n \text{ et } k = 1, \ldots, m \qquad (3.13)$$

$$\sum_{k=1}^{m} x_{ik} = 1 \quad \text{pour} \quad i = 1, \ \ldots, \ n \qquad (3.7)$$

$$\sum_{i=1}^{n} \lambda_i x_{ik} \leq M_k \quad \text{pour} \quad k = 1, \ \ldots, \ m \qquad (3.8)$$

$$z_{ijk} = x_{ik}x_{jk} \text{ pour } i, j = 1, \ldots, n \text{ et } k = 1, \ldots, m, \text{ avec } i \neq j. \qquad (3.9)$$

$$y_{ij} = \sum_{k=1}^{m} z_{ijk} \text{ pour } i, j = 1, \ldots, n, \text{ et } i \neq j. \tag{3.10}$$

$$z_{ijk} \leq x_{ik} \tag{3.14}$$

$$z_{ijk} \leq x_{jk} \tag{3.15}$$

$$z_{ijk} \leq x_{ik} + x_{jk} \tag{3.16}$$

$$z_{ijk} \geq 0 \tag{3.17}$$

Pour résoudre ce problème, nous proposons une méthode basée sur le principe d'optimalité de Bellman (1962) pour diviser le problème général en n sous-problèmes d'affectation plus faciles à résoudre. Chaque sous-problème correspond à une étape i où l'on définit une variable de décision pour spécifier à quel MSC est affectée une certaine cellule. Ainsi, au lieu d'avoir à prendre simultanément des décisions multiples et interdépendantes, nous prenons une décision simple, mais optimale à chaque étape du processus d'affectation. Et, à la fin de ce processus, les solutions obtenues des sous-problèmes sont combinées pour former la solution du problème général.

Par ailleurs, selon Bellman (1962), une décision optimale a la propriété que, peu importe le nombre de cellules considérées, l'optimalité de l'ensemble des décisions dépend essentiellement de la décision initiale. Il s'en suit que la première affectation est primordiale. Dans Beaubrun *et al.* (1999), nous définissons une série de critères qui permettent de déterminer l'ordre d'affectation des cellules en classant ces dernières de 1 à n. Cette étape permet d'appliquer le principe d'optimalité de Bellman (1962) sur les cellules classées pour obtenir la meilleure affectation possible. Le principe général de cette phase est illustré à la Figure 3.3 et se résume de la manière suivante. Soit X_{i-1} le nombre de cellules à affecter à l'étape i, $i = 1, 2, ..., n$. Désignons par E_i un paramètre de décision et par $c_i(E_i)$ le coût d'affectation à l'étape i. Alors, $c_i(E_i)$ se compose du coût de câblage et du coût de relève généré par la cellule i. Il importe toutefois de noter que le coût de la première affectation $c_1(E_1)$ inclut seulement le coût de câblage de la cellule 1 à

son MSC. En outre, le coût total d'affectation à l'étape i, noté $D_i(X_{i-1}, E_i)$, dépend du coût total à l'étape i-1 et d'un coût qui dépend de la décision prise à l'étape i.

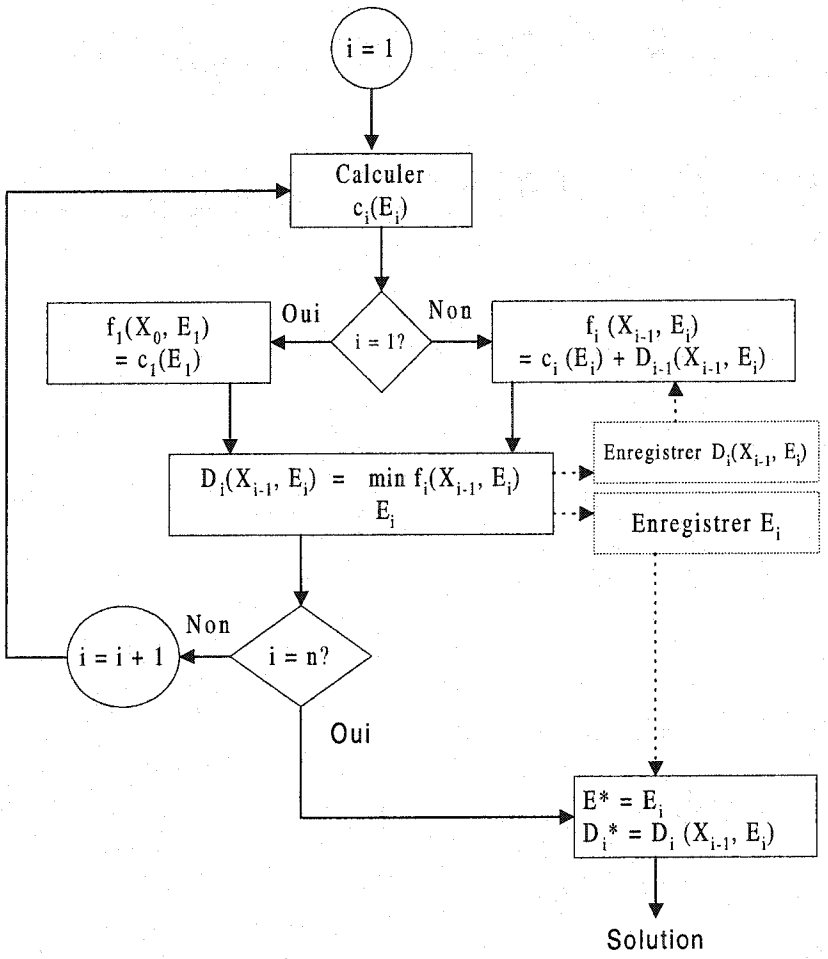


Figure 3.3 Séquence des opérations de la méthode d'affectation proposée

Nous pouvons également définir une fonction $f_i(X_{i-1}, E_i)$ qui constitue le coût total d'affectation avant de prendre une décision à l'étape i. Cette fonction peut être définie de la manière suivante :

$$f_i(X_{i-1}, E_i) = c_i(E_i) + D_{i-1}(X_{i-1}, E_i) \tag{3.18}$$

À ce niveau, la meilleure solution est obtenue en choisissant la meilleure affectation possible qui résulte de l'ensemble des décisions. Ainsi, à partir de (3.18), nous pouvons déduire le coût total de la meilleure affectation de la manière suivante :

$$D_i(X_{i-1}, E_i) = \min_{E_i} f_i(X_{i-1}, E_i) \tag{3.19}$$

Mentionnons également qu'à chaque étape, la décision courante et le coût total sont notés et gardés en mémoire. Évidemment, le processus s'arrête à l'étape n, d'où l'on obtient la solution du problème. Ainsi, par induction, cette méthode conduit toujours au minimum absolu à chaque étape, en fonction des contraintes données et de l'ordre d'affectation des cellules. Il en résulte que, dans le contexte de notre méthodologie de planification, le module *Architecture* détermine non seulement une configuration économique du réseau, mais aussi une intégration des sous-systèmes radio et réseau, ce qui permet de maintenir à la fois une performance et une exploitation efficaces de l'infrastructure. Les entrées/sorties de ce module, ainsi que les paramètres échangés avec les autres modules, sont illustrés à la Figure 3.4.

Toutefois, dans le cas où les stations de base ne sont pas directement reliées aux MSCs (comme pour l'UMTS), la méthode d'affectation proposée s'applique en deux étapes. Dans un premier temps, les stations de base (ou nœuds B) sont affectées aux contrôleurs de stations de base (RNCs). Ce problème se formule de la même façon que celui d'affectation des BTSs aux MSCs et, de ce fait, peut être résolu par l'utilisation de la méthode illustrée à la Figure 3.3, en considérant que la capacité d'un RNC est fixée par le nombre maximal de BTSs que ce RNC peut contrôler. Le résultat de cette affectation sert, dans un deuxième temps, à interconnecter les RNCs aux serveurs MSCs, ainsi qu'aux passerelles MGWs et aux SGSNs. Dans ce contexte, la capacité des serveurs MSCs est mesurée par le nombre de BHCAs (*Busy Hour Call Attempts*), alors que celle des SGSNs est mesurée par le nombre total d'abonnés à supporter et celle des passerelles MGWs par le nombre total d'Erlangs à supporter (Smith et Colins, 2002).

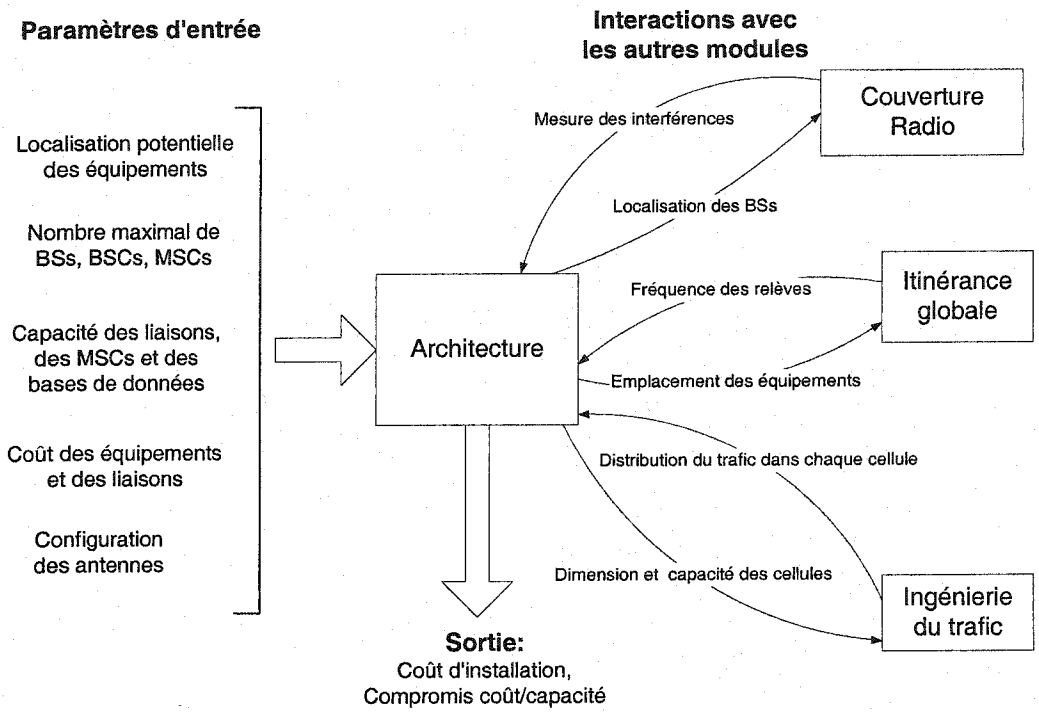


Figure 3.4 Opérations du module *Architecture*

3.2.3 Allocation des ressources

L'allocation des ressources consiste à partager efficacement la bande passante disponible pour répondre aux besoins de chaque catégorie d'usagers mobiles. En fait, lorsqu'un abonné désire établir une communication, le système doit s'assurer qu'il existe suffisamment de ressources disponibles pour lui permettre d'établir la connexion. Étant donné le caractère aléatoire des déplacements de cet abonné, le système se doit également de lui réserver dans les cellules adjacentes les ressources nécessaires pour garantir la continuité de la communication à la suite d'une relève. En effet, des ressources insuffisantes peuvent conduire soit au rejet (ou blocage) d'une demande de connexion, soit à l'arrêt prématuré d'une communication en cours. Il importe alors d'utiliser les ressources du réseau à bon escient de manière à s'assurer que la probabilité de blocage des nouvelles connexions et la probabilité d'interruption des communications de relève sont maintenues en dessous d'un certain seuil (Schwartz, 1995).

Pour illustrer notre propos, considérons la Figure 3.5 et supposons qu'un usager mobile veut établir une communication à partir de la cellule A. Alors, une portion suffisante de la bande passante disponible doit lui être octroyée dans la cellule A, tandis qu'une autre portion doit lui être réservée dans les cellules adjacentes à A, c'est-à-dire dans B, 3, 4, 5, 6 et 7. Si l'usager se déplace pour se rendre dans la cellule B, la portion réservée dans B sera alors utilisée, tandis que la portion réservée dans les cellules 4, 5 et 6 sera libérée pour devenir disponible à d'autres abonnés, car les cellules 4, 5 et 6 ne sont pas adjacentes à B. En même temps, une portion de la bande passante des cellules A, 1, 2 et 8 devient également réservée, de sorte que toutes les cellules adjacentes à B réservent à l'usager une portion de leur bande passante. Ainsi, une méthode efficace d'allocation de ressources se base à la fois sur des informations locales (comme la bande passante disponible dans la cellule où se trouve l'abonné) et sur des informations globales (comme la bande passante disponible dans les cellules adjacentes) pour octroyer ou réserver les ressources à ses usagers. Il faudrait toutefois noter que la portion de la bande passante à allouer ou à réserver dépend des exigences du trafic véhiculé dans le réseau.

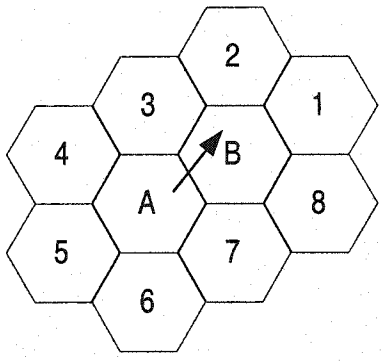


Figure 3.5 Prédiction de déplacement pour l'allocation des ressources

Pour les réseaux mobiles traditionnels, où toutes les connexions obéissent pratiquement aux mêmes contraintes de qualité de service, l'allocation des ressources se fait en réservant un certain nombre de canaux de garde (*guard channels*) aux nouveaux

appels et un pourcentage fixe de la capacité des stations de base aux appels de relève (Ramanathan *et al.*, 1999). On parle alors d'allocation fixe de ressources. Si le pourcentage réservé aux appels de relève est élevé, une capacité suffisante sera alors disponible pour satisfaire aux exigences des appels de relève au risque de bloquer les nouveaux appels. Dans ce contexte, une approche dynamique (ou adaptative) permettant à chaque station de base d'adapter dynamiquement la capacité réservée aux appels de relève devient plus efficace (Ramanathan *et al.*, 1999).

Les stratégies d'allocation dynamique de ressources sont nécessaires pour fournir la qualité de service requise et améliorer la capacité du système. Par ces méthodes, le MSC, au lieu de la station de base, se base sur le nombre de communications en cours dans les cellules adjacentes, sur les distances de réutilisation des fréquences, la probabilité de blocage et la distribution des usagers pour réserver les ressources nécessaires aux nouveaux appels et aux appels de relève (Hàc et Chen, 2000). En outre, si une cellule ne dispose pas de capacité suffisante pour satisfaire à une nouvelle demande de connexion, elle peut soit emprunter des canaux appartenant aux cellules adjacentes, soit autoriser une variation dans la qualité des images ou vidéos, acceptant ainsi une réduction de la qualité plutôt que l'interruption ou le rejet de la communication en cours.

De plus, dans le contexte des systèmes mobiles de la prochaine génération, le problème d'allocation de ressources paraît plus complexe, car en plus de la mobilité globale, il fait intervenir différents types de trafic ayant des contraintes spécifiques de débit et de délai. Par exemple, un amateur de jeux vidéos peut avoir souvent besoin d'une portion considérable de la bande passante disponible. Mais s'il a besoin de transférer un courrier électronique entre deux téléchargements de jeux vidéos, le système doit être en mesure de réajuster la bande passante requise en fonction de ce besoin. Dans ce contexte, l'allocation dynamique des canaux permet d'avoir des gains considérables en capacité.

Ainsi, pour être efficace, un algorithme d'allocation de ressources doit considérer au moins deux classes de trafic : classe 1 ou trafic en temps réel (comme la voix ou la vidéo), et classe 2 ou trafic non en temps réel (comme le courrier électronique). Lors d'une demande de connexion, cet algorithme doit déterminer deux paramètres en fonction

de la classe de trafic : la largeur de bande désirée pour la connexion et la largeur de bande minimale requise. Pour le trafic de classe 1, l'objectif est d'allouer à la communication la quantité désirée de la bande passante dans la cellule où la communication est générée et de réserver dans les cellules adjacentes la bande passante minimale requise. Cette portion réservée peut automatiquement se réajuster en fonction du déplacement de l'usager et de la disponibilité des ressources. Toutefois, pour le trafic de classe 2, on peut allouer par défaut la bande passante minimale requise à la connexion, sans être obligé de réserver des ressources dans les cellules adjacentes. À ce niveau, il importe de définir un temps de garde pendant lequel ce type de trafic peut être mis en attente, au cas où les ressources ne seraient pas disponibles. Si la largeur de bande minimale requise n'est pas disponible et que le temps de garde est dépassé, on peut alors décider de rejeter la connexion.

Ces principes constituent la base d'un algorithme efficace d'allocation de ressources. Toutefois, il faudrait admettre qu'un tel algorithme peut surestimer la largeur de bande à réserver, ce qui résulte en une faible utilisation de la bande passante. En essayant de prédire le mouvement des usagers (ce qui est difficile à implémenter), il est possible de réduire cette surestimation. Ainsi, la caractérisation du trafic et la détermination du modèle de mobilité sont essentielles à l'évaluation de la quantité de ressources à allouer à chaque usager pour répondre à ses besoins. Cela permet non seulement d'optimiser les ressources du réseau, mais aussi de garantir la qualité de service, en termes de probabilité de blocage de nouveaux appels et des appels de relève. Les entrées/sorties du module *Allocation des ressources*, ainsi que les paramètres échangés avec les modules *Itinérance globale* et *Ingénierie du trafic*, sont illustrés à la Figure 3.6.

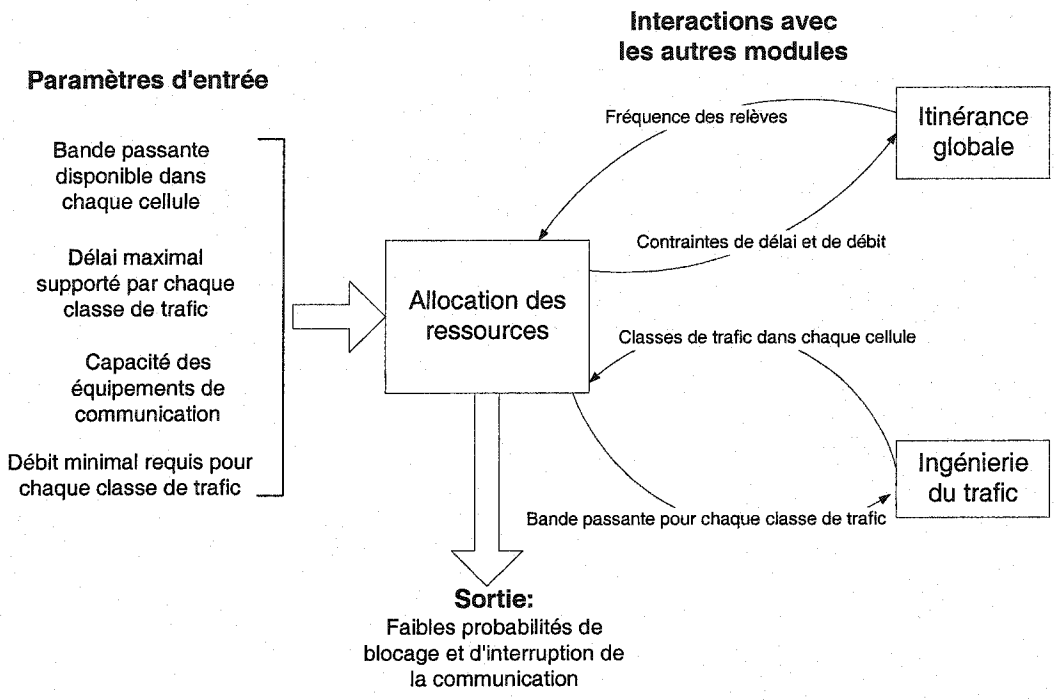


Figure 3.6 Opérations du module *Allocation des ressources*

3.2.4 Itinérance globale

Ce module permet de gérer efficacement la mobilité globale, d'autant plus que, dans les réseaux mobiles de la prochaine génération, les usagers auront la possibilité de circuler au travers des réseaux utilisant des technologies et protocoles différents, exploités par des opérateurs différents et situés dans des zones géographiques différentes. Dans ce contexte, le système doit leur permettre de pouvoir accéder à leurs services à partir d'un point géographique quelconque. Le problème consiste alors à trouver un bon compromis entre la localisation précise des usagers et les performances de la gestion de l'itinérance (ou la mobilité), en termes de rapidité (faible temps de réponse du réseau), de trafic de signalisation généré, de taux de requêtes et de mises à jour effectués au niveau des bases de données. Dans ce cas, le module *Itinérance globale* doit implémenter une approche efficace qui permet de réduire le trafic de signalisation au niveau des bases de données et qui facilite l'interopérabilité entre les sous-systèmes lors de l'itinérance

globale. Les entrées/sorties, ainsi que les paramètres échangés entre le module *Itinérance globale* et les modules *Allocation des ressources*, *Architecture* et *Ingénierie du trafic*, sont illustrés à la Figure 3.7.

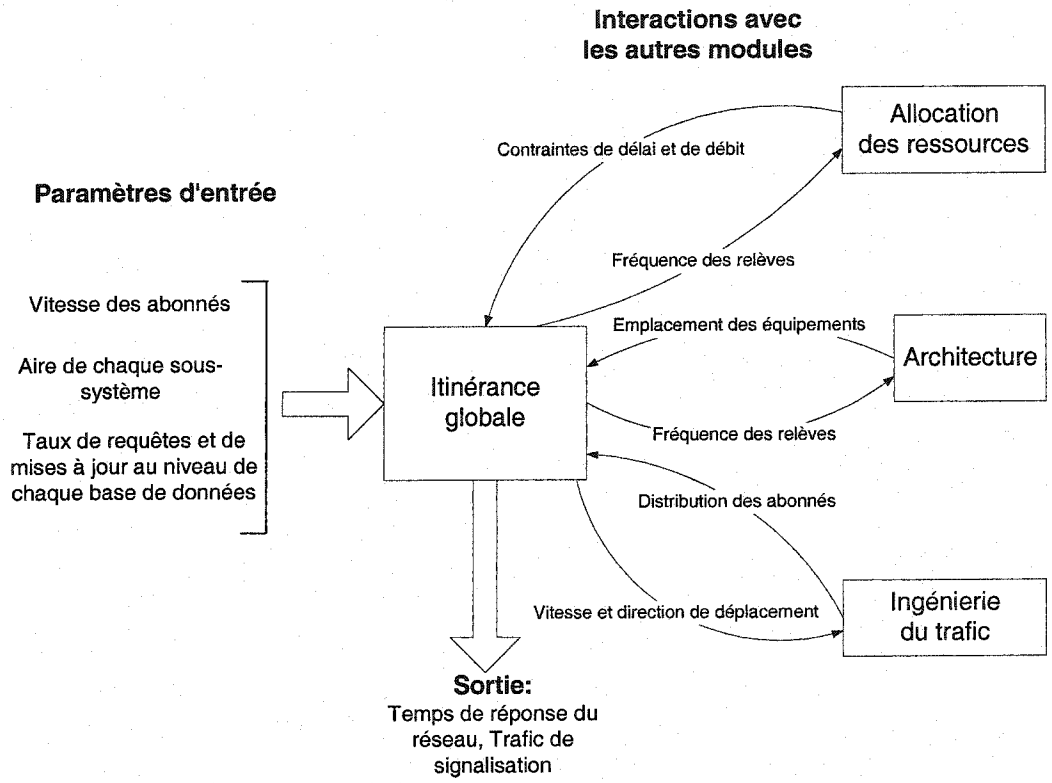


Figure 3.7 Opérations du module *Itinérance globale*

Mentionnons que l'itinérance globale n'avait jusque là jamais fait l'objet des approches de planification de réseaux mobiles proposées dans la littérature (Tutschku et Tran-Gia, 1998; Aghvami et Jafarian, 2000). En introduisant ce module dans notre méthodologie, nous reflétons le plus possible la réalité, puisqu'en pratique, l'allocation des ressources dépend en grande partie de la mobilité des usagers. Il faut toutefois comprendre que cet aspect ne fait que rendre plus complexe le problème de planification. Nous y reviendrons au chapitre 4.

3.2.5 Ingénierie du trafic

Le trafic est un facteur déterminant de la planification des réseaux de la prochaine génération. En effet, la localisation des stations de base, le dimensionnement des équipements fixes, l'interconnexion du sous-système radio au sous-système réseau, ainsi que l'allocation des ressources se font principalement en fonction de la distribution du trafic dans le système. Autrement dit, le module *Ingénierie du trafic* interagit avec tous les autres modules, ce qui est illustré à la Figure 3.8. Ainsi, une mauvaise analyse du trafic peut conduire soit au déploiement d'un nouveau réseau, soit à l'ajout d'équipements additionnels au système existant, ce qui peut faire l'objet de coûts significatifs. Toutefois, l'analyse du trafic demeure difficile en conception des réseaux mobiles, puisqu'elle doit tenir compte non seulement du comportement de chaque usager mobile dans le réseau, mais aussi de l'émergence des nouveaux services qui créent à toutes fins pratiques un trafic multimédia hétérogène. De ce fait, l'ingénierie du trafic dans les réseaux mobiles atteint une nouvelle dimension dans la théorie du télé-trafic et mérite d'être soigneusement investiguée. Cette investigation se fera au chapitre 5.

3.3 Résultats obtenus

Dans cette section, nous présentons une série de résultats analysant le compromis entre la capacité du système et la puissance minimale requise par utilisateur pour maintenir une communication de qualité. Ces résultats constituent une application de la relation (3.5). Dans cette optique, nous considérons que le refus de service survient lorsque le rapport interférence à bruit dépasse 10 dB, comme l'ont supposé Viterbi et Viterbi (1993). À partir de la relation (3.5) et en fixant $\rho = 0.4$, nous pouvons évaluer la puissance minimale requise par usager pour maintenir une communication de qualité, ce que nous illustrons à la Figure 3.9 (pour plusieurs valeurs de I_0/N_0). Nous nous rendons compte que, pour les valeurs faibles des capacités, toutes les courbes possèdent une décroissance rapide, alors que pour les capacités élevées, ces courbes ont une décroissance beaucoup plus lente. En général, les concepteurs préfèrent opérer dans la

partie de faible décroissance pour s'assurer d'une certaine stabilité dans la qualité de service à offrir.

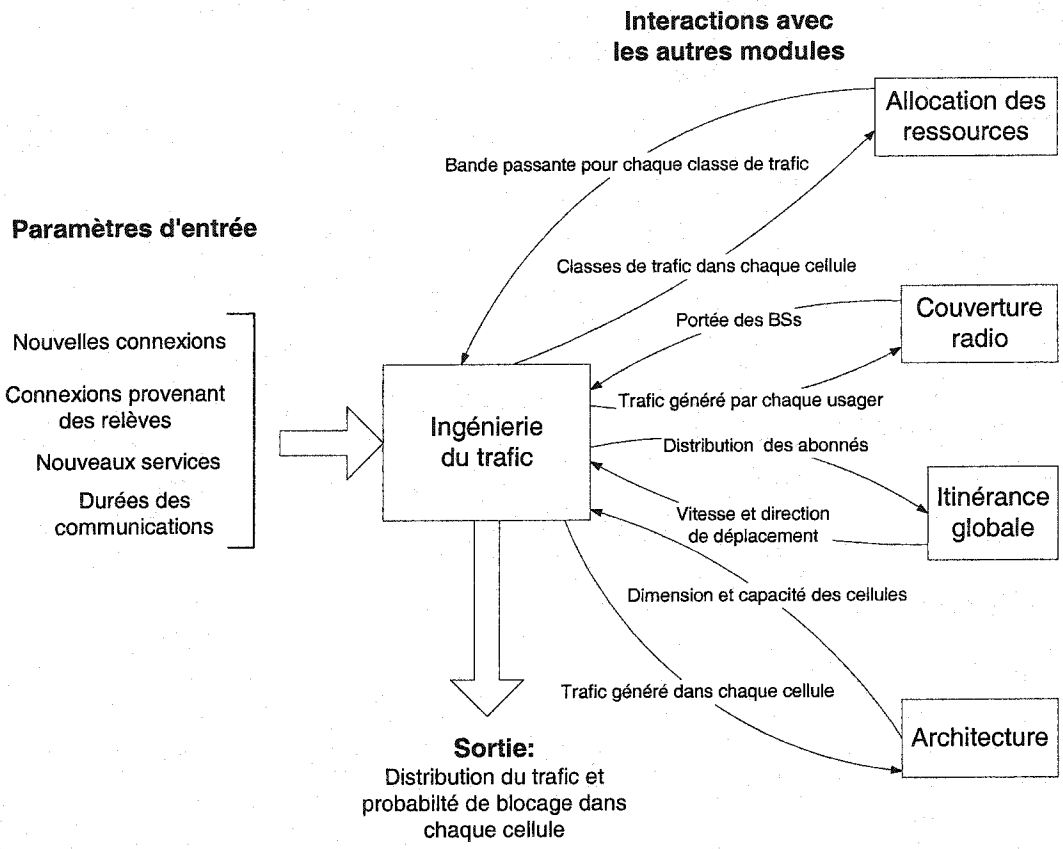


Figure 3.8 Opérations du module *Ingénierie du trafic*

Par ailleurs, les résultats ont montré que, pour une valeur donnée du rapport I_0/N_0, la puissance minimale requise par usager décroît lorsque la capacité de la cellule augmente. Par exemple, si l'on considère une cellule de 150 Erlangs, où le rapport I_0/N_0 est égal à 2 dB, on se rend compte que, pour maintenir une bonne qualité de service, la puissance requise par utilisateur doit être supérieure à –20 dB. Toutefois, si la capacité du système monte à 250 Erlangs, le rapport signal à bruit minimal requis par utilisateur devient aussi faible que –22.3 dB.

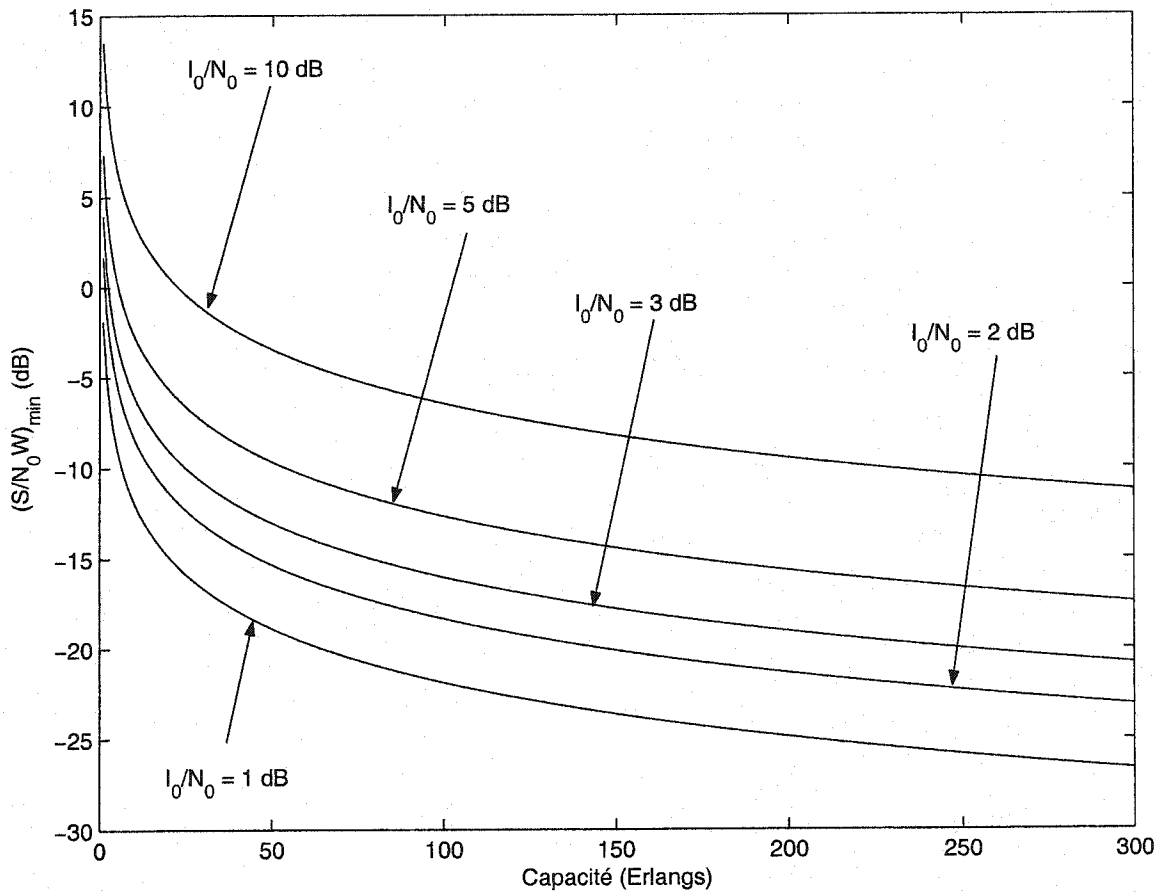


Figure 3.9 Rapport signal à bruit minimal par usager

D'autre part, nous avons analysé le comportement du rapport signal à bruit minimal reçu par utilisateur en fonction du nombre de terminaux actifs, c'est-à-dire du niveau d'interférence. Nous faisons alors varier I_0/N_0 et observons son effet sur $(S/N_0W)_{min}$ pour une valeur de capacité donnée. À partir de la Figure 3.10, nous observons que $(S/N_0W)_{min}$ augmente en fonction de I_0/N_0 (pour une capacité donnée). Autrement dit, plus le trafic est intense, plus le système doit fournir de la puissance pour maintenir une bonne qualité de service. Par exemple, pour une capacité de 250 Erlangs, le Tableau 3.1 donne, pour quelques valeurs de I_0/N_0, le rapport signal à bruit minimal

requis pour maintenir une qualité de service acceptable. Nous notons alors une différence de 15.4 dB sur la valeur de $(S/N_0W)_{min}$ lorsque I_0/N_0 passe de 1 dB à 10 dB.

Tableau 3.1 Rapport signal à bruit acceptable en fonction des interférences

I_0/N_0 (dB)	1.0	2.0	3.0	5.0	10.0
$(S/N_0W)_{min}$ (dB)	-25.7	-22.3	-20.0	-16.7	-10.4

La même analyse peut être entreprise pour comparer $(S/N_0W)_{min}$ dans les macro-cellules, les micro-cellules et les pico-cellules, en utilisant un taux moyen d'arrivée λ et une durée moyenne d'occupation des canaux $1/\mu$ dont les valeurs sont spécifiées au Tableau 3.2. Pour les macro-cellules, le paramètre $1/\mu$ est considéré comme la durée moyenne de séjour dans la cellule et correspond au temps que prend normalement un véhicule pour traverser une cellule d'environ 1 km de rayon, alors que pour les pico-cellules, ce paramètre représente la durée moyenne des appels. Cela signifie que, la plupart du temps, les appels débutent et s'achèvent à l'intérieur d'une pico-cellule, à cause de la vitesse relativement faible de ses occupants qui sont généralement des piétons. Pour les micro-cellules, la durée d'occupation du canal peut être soit le temps de séjour dans la cellule, soit la durée de l'appel, selon la vitesse et la direction des usagers.

La Figure 3.10 illustre et compare $(S/N_0W)_{min}$ pour les macro-cellules, les micro-cellules et pico-cellules. Nous en déduisons que le rapport signal à bruit minimal par usager est plus élevé pour les macro-cellules que pour les autres types de cellules. Autrement dit, les macro-cellules doivent développer plus de puissance que les micro-cellules et les pico-cellules pour maintenir le même niveau de qualité de service. Les résultats obtenus permettent également de vérifier que l'utilisation de micro-cellules ou de pico-cellules aide à augmenter la capacité du système.

Tableau 3.2 Paramètres d'évaluation du trafic

	Macro-cellule	Micro-cellule	Pico-cellule
λ (appels/sec)	1.0	1.1	1.15
$1/\mu$ (sec)	50.0	60.0	100.0

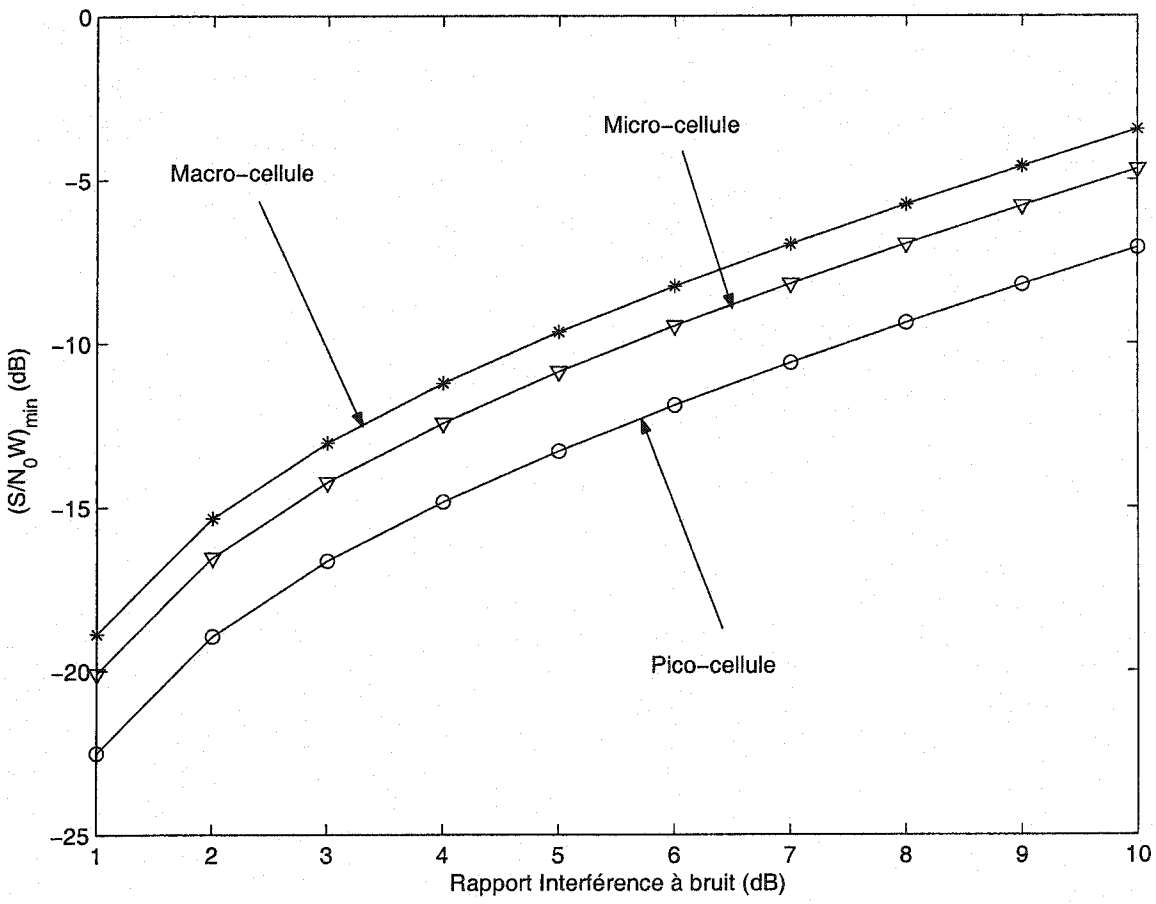


Figure 3.10 Rapport signal à bruit acceptable par usager selon le type de cellules

CHAPITRE 4

APPROCHE PROPOSÉE POUR LA GESTION DE LA MOBILITÉ GLOBALE

Les systèmes mobiles de la prochaine génération visent à offrir à tous les usagers fixes et mobiles la possibilité de se déplacer et d'être rejoints de partout à travers la planète, en profitant pleinement des services auxquels ils ont droit (Buchanan *et al.*, 1997; Pandya, 1999; Aghvami et Jafarian, 2000). On parle alors de mobilité (ou d'itinérance) globale et de portabilité de services. Dans ce contexte, la localisation d'un abonné ou l'accès à ses services exige l'interopérabilité de plusieurs sous-réseaux fixes et mobiles, ce qui augmente le trafic de signalisation et qui peut résulter en une diminution considérable des performances du réseau. Il s'avère alors impérieux de mettre au point une stratégie de gestion de la mobilité globale qui facilite l'interopérabilité des composantes hétérogènes des systèmes mobiles de la prochaine génération. Dans ce chapitre, nous présentons une nouvelle approche qui améliore les performances de tels systèmes, en termes de trafic de signalisation généré et de temps de réponse aux requêtes. Plus précisément, nous y passons en revue les principales approches existantes, décrivons les principes ainsi que le fonctionnement de l'approche proposée, définissons les paramètres d'évaluation de performance, ce qui permettra de comparer les résultats obtenus de l'approche proposée à ceux des approches présentées dans la littérature.

4.1 Gestion de la mobilité globale

L'objectif principal d'un mécanisme de gestion de mobilité globale est de permettre l'interopérabilité de sous-systèmes hétérogènes, c'est-à-dire des sous-systèmes qui utilisent des technologies, protocoles et spectres radio différents, en plus

d'implémenter chacun son propre mécanisme de gestion de mobilité. La Figure 4.1 donne un exemple de réseau global composé de quatre sous-réseaux ou sous-systèmes hétérogènes. Chaque sous-réseau implémente alors son propre mécanisme de gestion de localisation, tandis que le format d'informations véhiculées dans les bases de données (HLR/HSS et VLRs) est différent d'un sous-système à un autre, ce qui rend difficile la gestion de la mobilité globale. Garg et Wilkes (1996) ont d'ailleurs montré que l'interopérabilité de bases de données hétérogènes augmente significativement le trafic de signalisation dans le réseau. Cela a motivé plusieurs chercheurs (Lin et Chlamtac, 1996; Akyildiz et Wang, 2002) à proposer des stratégies pour réduire un tel trafic et améliorer ainsi les performances du réseau.

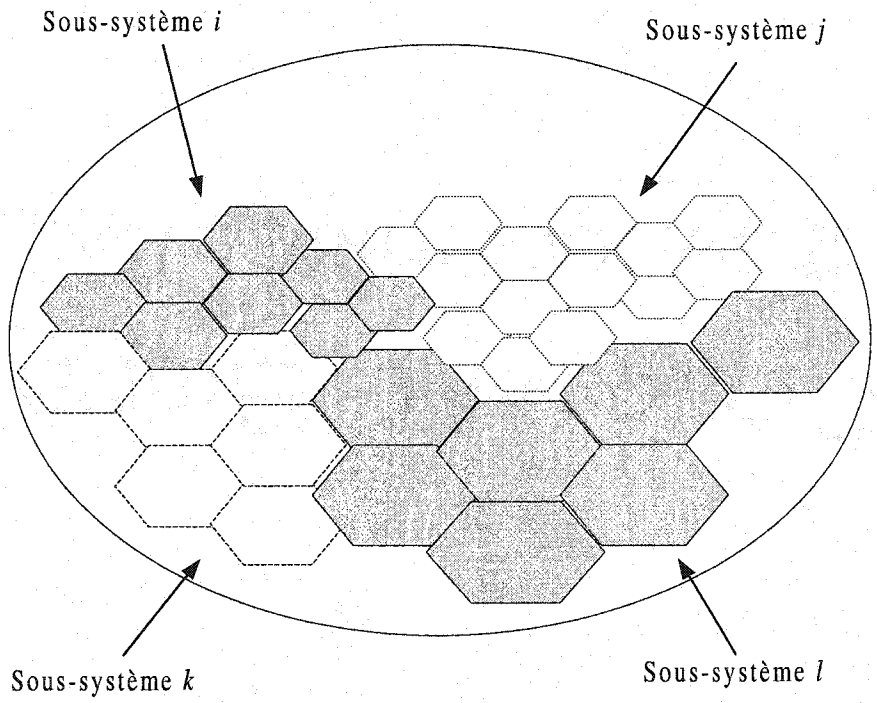


Figure 4.1 Réseau global constitué de sous-systèmes hétérogènes

Sans perte de généralité, considérons le réseau de la Figure 4.2 constitué de trois sous-systèmes hétérogènes. Chaque sous-système divise son aire de service en un certain nombre de zones de localisation (*LA* : *Location Area*) selon sa propre méthode (problème de partitionnement). Dans l'exemple de la Figure 4.2, le sous-système 1 contient cinq LAs, alors que le sous-système 2 en contient trois et le sous-système 3 en contient quatre. En termes de technologie et de protocoles utilisés, le sous-système 1 pourrait représenter le système nord-américain IS-95 (*Interim Standard 95*), tandis que le sous-système 2 pourrait être un réseau de type GSM et le sous-système 3 un réseau ATM sans fil (*wireless ATM*). Dans ce contexte, la mise à jour de localisation se réfère à la localisation d'un abonné qui se déplace à travers l'ensemble du système, alors que la télé-recherche vise à rechercher cet abonné dans le système en vue de lui acheminer une communication.

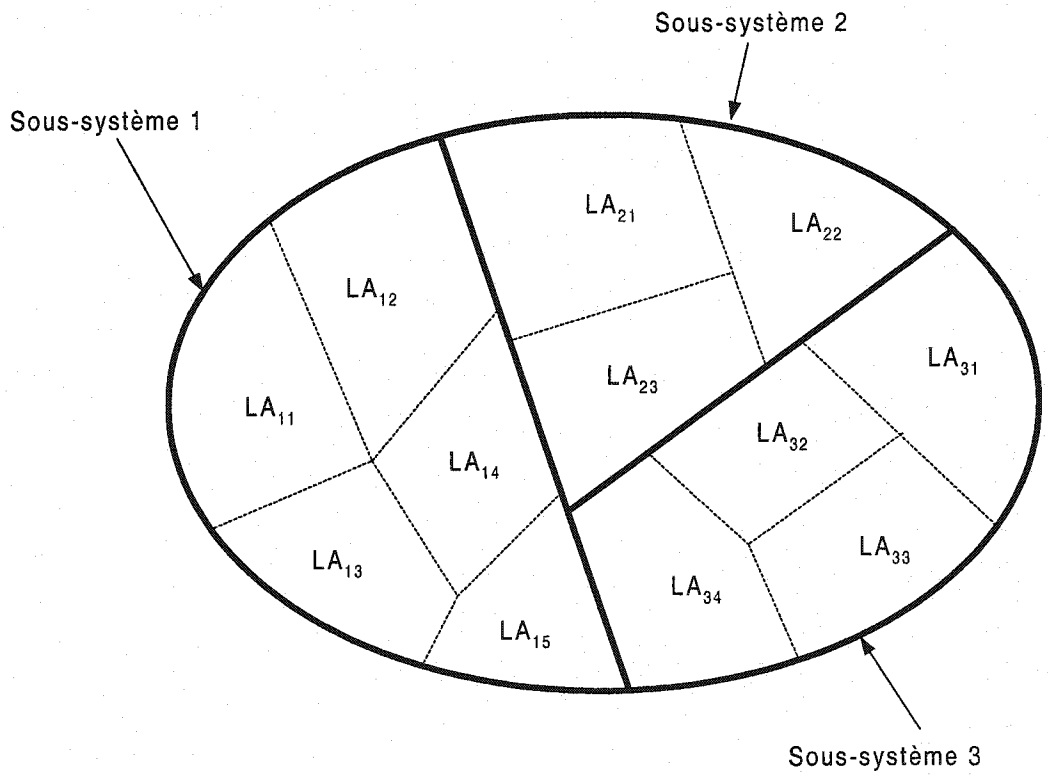


Figure 4.2 Subdivision de la zone de service de chaque sous-système en LAs

Comme les sous-systèmes utilisent des technologies, spectres radio et protocoles différents, il semble difficile d'interconnecter directement leurs HLRs (ou HSSs) respectifs sans l'aide d'un équipement qui assure leur interopérabilité. Dans ce contexte, Lin et Chlamtac (1996) ont proposé d'utiliser une base de données centralisée appelée *HLR multi-systèmes* ou *MHLR* (*Multi-tier Home Location Register*) chargée de permettre aux HLRs de systèmes hétérogènes de se communiquer et d'interopérer. L'architecture correspondant à cette approche est illustrée à la Figure 4.3 pour un réseau constitué de trois sous-systèmes hétérogènes.

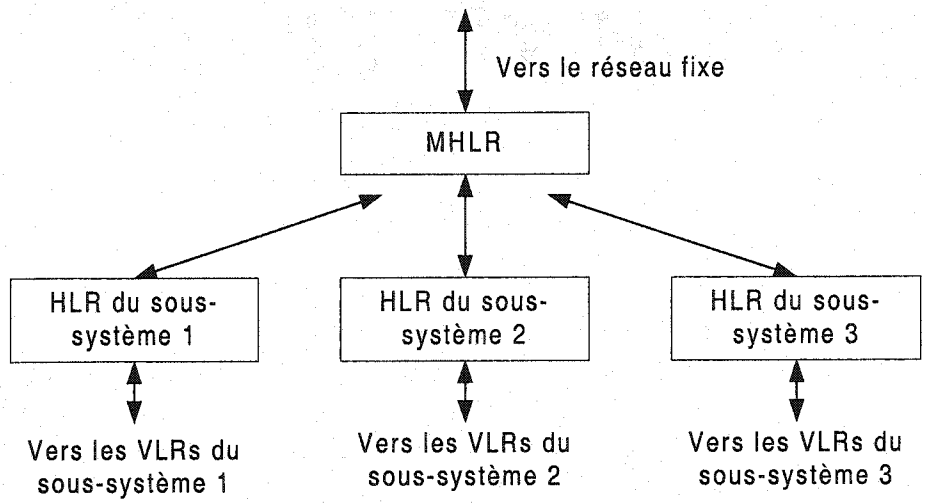


Figure 4.3 Architecture centralisée pour la gestion de la mobilité globale

Deux approches sont alors utilisées pour exploiter cette architecture : l'approche d'enregistrement simple (*SR* : *Single Registration*) et l'approche d'enregistrement multiple (*MR* : *Multiple Registration*). L'approche SR permet au terminal mobile de ne s'enregistrer que dans une seule zone de localisation auprès du MHLR pour lui indiquer sa position courante. Dans ce cas, chaque passage d'un sous-système à un autre génère une mise à jour de localisation au niveau du MHLR. Cette mise à jour se fait en deux étapes : enregistrement auprès du nouveau sous-réseau et désenregistrement auprès de

l'ancien. Les différentes opérations d'enregistrements/désenregistrements sont présentées par Lin et Chlamtac (1996) lorsque le mobile passe d'un sous-réseau *i* à un sous-réseau *j*. En ce qui concerne la méthode MR, elle permet au terminal mobile de s'enregistrer simultanément dans plusieurs sous-systèmes à partir du MHLR. Autrement dit, chaque sous-système peut procéder individuellement à la gestion de l'itinérance des abonnés comme s'il n'était intégré à aucun autre sous-système, ce qui se traduit par une mise en pratique plus simple de la gestion de mobilité que pour la méthode SR. Les résultats obtenus par Lin et Chlamtac (1996) montrent que cette méthode contribue à réduire le nombre d'opérations d'enregistrements.

Toutefois, il faut se rendre compte que l'architecture associée aux deux approches (SR et MR) fait l'objet de plusieurs lacunes. D'abord, la quantité de trafic de signalisation générée au niveau du MHLR peut être énorme, ce qui peut réduire la qualité de service attendue du réseau. Deuxièmement, le MHLR doit être doté d'une capacité très élevée pour pouvoir gérer l'ensemble de tous les sous-systèmes existants. Le coût d'un tel équipement pourrait être alors prohibitif. Ensuite, comme dans tout système centralisé, il se pose un problème de fiabilité : advenant une défaillance matérielle ou logicielle du MHLR, le système perd sa fonctionnalité de gestion de mobilité globale. Pour terminer, mentionnons un problème de nature plutôt politique : il s'agit de décider de la localisation du MHLR. Vu qu'il s'agit d'une passerelle par où passeraient toutes les communications, aussi bien X (par exemple les Américains) que Y (par exemple les Européens) auraient bien voulu l'installer chez eux de manière à y exercer un certain contrôle.

Par ailleurs, une autre approche a été récemment proposée par Akyildiz et Wang (2002) pour réduire le trafic de signalisation généré par l'itinérance globale. Cette approche suggère de relier les VLRs des sous-systèmes hétérogènes par des équipements appelés *BIUs* (*Boundary Interworking Units*). L'évaluation de performance de cette méthode se fait en déterminant les coûts de signalisation, les taux de perte d'appels et les délais de télé-recherche. Les résultats montrent que le coût total de signalisation C_T décroît lorsque le rapport *appel à mobilité* diminue, ou lorsque le rayon des cellules augmente. Toutefois, le paramètre C_T augmente lorsque la dimension des zones de

localisation augmente, c'est-à-dire pour un nombre élevé de cellules faisant partie d'une même zone de localisation. Quant au délai de télé-recherche, il croît linéairement avec la dimension des zones de localisation. Les résultats montrent également que l'utilisation des BIUs contribue à réduire le taux de perte d'appels et les délais de télé-recherche après le passage d'une cellule à une autre. Toutefois, la connexion des BIUs aux VLRs de chaque sous-système donne lieu à plusieurs inconvénients. D'abord, puisque le coût de câblage est proportionnel au nombre de VLRs, ce coût tend à augmenter lorsque le nombre de LAs faisant partie de chaque sous-système augmente. De plus, un délai supplémentaire peut être généré au HLR de chaque sous-système lors des opérations de mises à jour. Dans la prochaine section, nous proposons une nouvelle approche pour améliorer les performances du réseau, en termes de trafic de signalisation et de temps de réponse.

4.2 Approche proposée

L'approche proposée se base sur l'utilisation d'un équipement d'interconnexion appelé WING (*Wireless INterworking Gateway*) pour à la fois faciliter l'interopérabilité des sous-systèmes hétérogènes et réduire le trafic de signalisation dans un contexte d'itinérance globale. Une telle approche utilise l'architecture illustrée à la Figure 4.4. Énonçons-en les principes et décrivons la séquence d'opérations induites par son utilisation pour l'enregistrement et la mise à jour de localisation, ainsi que pour l'acheminement des communications.

4.2.1 Principes généraux

Le WING constitue une passerelle spécialement conçue pour faciliter l'échange d'informations entre deux sous-systèmes hétérogènes. Il est doté d'une interface sans fil qui lui permet de recueillir et d'enregistrer le profil, ainsi que les informations relatives aux sessions des usagers qui changent de sous-systèmes. De plus, il est muni de ports

permettant son interconnexion aux HLRs (ou HSSs) des sous-systèmes adjacents et est en mesure de supporter les fonctionnalités suivantes :

- Conversion de messages et de signalisation d'un certain format à un autre;
- Garantie de la compatibilité des interfaces radio et de l'authentification des abonnés;
- Maintien des informations d'itinérance des abonnés après les notifications d'enregistrements et de mises à jour;
- Garantie de l'acheminement fiable des appels et du trafic de signalisation (c'est-à-dire, le WING doit être en mesure de faire le contrôle d'erreur).

La Figure 4.4 illustre l'interconnexion des HLRs par des WINGs pour le réseau de la Figure 4.2.

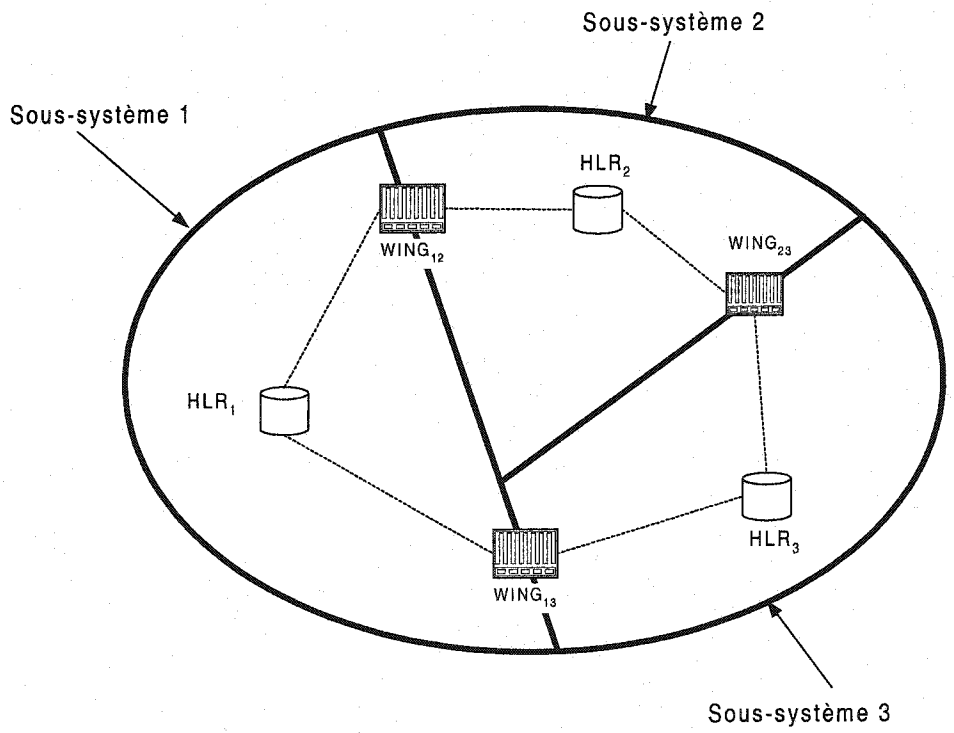


Figure 4.4 Schéma d'interconnexion des HLRs selon l'architecture proposée

Soit deux sous-systèmes adjacents *i* et *j* dont les HLRs respectifs (HLR_i et HLR_j) sont directement reliés par un WING (noté $WING_{ij}$). Considérons que le sous-système *i* est le sous-système de rattachement d'un certain abonné. Tant que cet abonné reste dans le sous-système *i*, le $WING_{ij}$ n'a pas à intervenir, laissant au sous-système le soin d'appliquer son propre mécanisme de gestion de mobilité. L'intervention du $WING_{ij}$ survient seulement lorsque la présence du terminal de l'abonné est détectée à la frontière des sous-systèmes *i* et *j*. À partir de cette détection, le profil de l'abonné est transféré au $WING_{ij}$ qui lit, interprète, traduit ces informations, avant de les transférer dans un format compréhensible par le HLR_j. Ce dernier communique alors au VLR_{jx}, chargé de gérer la zone de localisation où se trouve l'abonné au moment du changement de sous-systèmes. Dès lors, l'usager mobile est considéré comme un abonné à part entière du sous-système *j*. Les mises à jour de localisation se font alors comme dans le cas classique, avec la différence suivante : tant que l'abonné demeure dans le sous-système *j*, des notifications de mises à jour sont périodiquement transmises au $WING_{ij}$ qui les garde en mémoire et qui informe le sous-système *i* des nouvelles positions de l'abonné dans le sous-système *j*. Cela permet, entre autres, de rejoindre rapidement cet abonné lorsqu'un appel lui est destiné.

Par rapport à la méthode proposée par Akyildiz et Wang (2002), notre approche offre une solution plus économique, en terme de coût de câblage. En effet, l'architecture que nous proposons ne nécessite que deux liaisons pour interconnecter les sous-systèmes *i* et *j* : une liaison pour relier le $WING_{ij}$ au HLR_i et une autre pour relier le $WING_{ij}$ au HLR_j, tandis que l'architecture présentée par Akyildiz et Wang (2002) se propose de relier, d'une part, le BIU à chaque VLR du sous-système *i* et, d'autre part, le BIU à chaque VLR du sous-système *j*. Soit N_{VLR_i} le nombre de VLRs du sous-système *i* et N_{VLR_j} le nombre de VLRs du sous-système *j*. Alors, le coût de câblage de l'approche proposée par Akyildiz et Wang (2002) dépend du nombre de zones de localisation de chaque sous-système et est $(N_{VLR_i} + N_{VLR_j})/2$ fois plus élevé que le coût de notre solution.

De manière plus globale, notre approche peut contribuer à réduire les coûts d'exploitation du réseau. En effet, le WING est conçu pour être suffisamment flexible, de manière à supporter à la fois l'infrastructure et la technologie existantes, tout en étant capable de s'adapter aux nouveaux développements technologiques. Il en résulte que le WING n'imposera pas de modifications majeures aux composantes des sous-systèmes déjà en place, à la suite d'un changement technologique donné. Pour terminer, nous devons nous assurer que les fonctionnalités d'administration et de maintenance sont supportées par le WING, tout en étant accessibles à distance. Cela permettra une gestion plus facile et plus efficace d'un tel équipement.

4.2.2 Séquence des opérations

Pour évaluer la quantité de trafic de signalisation générée, nous devons définir la séquence d'opérations mises en œuvre par le réseau dans un contexte d'itinérance globale. En termes d'activités des bases de données, cette séquence d'opérations génère un trafic constitué de deux composantes : des requêtes et des mises à jour (Wey *et al.*, 1997). Nous allons définir la séquence de requêtes et de mises à jour induites par notre approche lors des enregistrements, mises à jour de localisation et acheminements des appels dans un contexte d'itinérance globale.

Considérons à nouveau deux sous-systèmes adjacents *i* et *j*. Lorsqu'un abonné du sous-système *i* pénètre pour la première fois au sous-système *j*, il doit s'y enregistrer, ce qui permet au sous-système *j* de l'authentifier. Cette authentification facilitera, d'une part, l'acheminement des appels destinés à cet abonné et, d'autre part, l'accessibilité (à partir du sous-système *j*) à tous les services auxquels un tel abonné a droit. Le processus d'enregistrement est déclenché par le $WING_{ij}$ lorsque ce dernier détecte la présence de l'abonné à la frontière des sous-systèmes *i* et *j*. Le $WING_{ij}$ annonce alors au HLR_j l'arrivée éventuelle d'un nouvel usager dans son aire de service et lui transfère le profil de cet usager. Le HLR_j procède à la mise à jour de ses informations et communique au VLR_{jx} qui vérifie que l'usager se trouve bel et bien dans sa zone de localisation et qui

procède, à son tour, à la mise à jour de ses informations. Pour terminer, le WING rapporte au sous-système *i* que l'abonné est passé au sous-système *j*, ce qui se traduit par des mises à jour à la fois au niveau du HLR_i et du VLR_{ix} (c'est-à-dire le VLR qui gérait la localisation de l'abonné juste avant le changement de sous-systèmes).

La séquence d'opérations qui caractérise le processus d'enregistrement dans un nouveau sous-système est illustrée à la Figure 4.5 et décomposée de la manière suivante :

a. Requête au $WING_{ij}$
b. Requête au HLR_j
c. Mise à jour du HLR_j
d. Requête au VLR_{jx}
e. Mise à jour du VLR_{jx}
f. Mise à jour du HLR_i
g. Mise à jour du VLR_{ix}

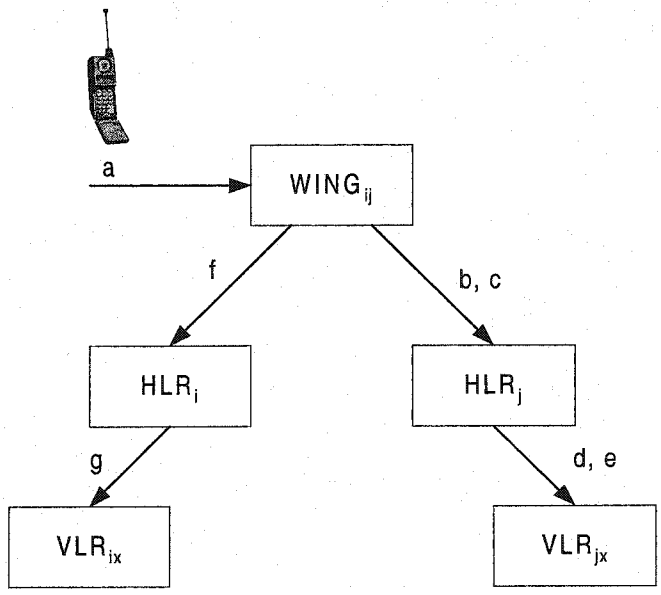


Figure 4.5 Processus d'enregistrement dans un nouveau sous-système

D'autre part, le terminal doit périodiquement mettre à jour sa localisation tant qu'il reste à l'intérieur du sous-système *j*. Cette mise à jour se fait au sein du sous-système *j*, mais les informations de localisation recueillies doivent être transférées au sous-système *i*, ce qui permet au sous-système *i* de pouvoir retracer son abonné à n'importe quel instant. La séquence d'opérations qui intervient dans un tel processus est illustrée à la Figure 4.6 et décomposée de la manière suivante :

a. Requête au VLR_{jx}
b. Requête au HLR_j
c. Mise à jour du HLR_j
d. Mise à jour du $WING_{ij}$
e. Mise à jour du VLR_{jx}
f. Mise à jour du HLR_i
g. Mise à jour du VLR_{ix}.

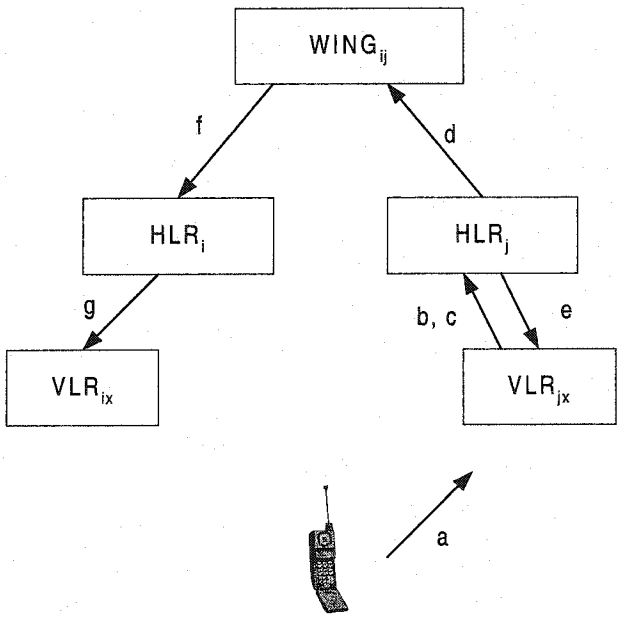


Figure 4.6 Processus de mise à jour de localisation

Dans le but d'évaluer la quantité de trafic de signalisation généré par le processus d'itinérance globale, il convient d'ajouter aux opérations précédentes celles générées par les appels émis (c'est-à-dire, le trafic sortant) et reçus (c'est-à-dire, le trafic entrant) par le terminal. En terme de trafic de signalisation, un appel émis par un abonné du sous-système *i* n'exécute qu'une requête au VLR_{ix}. Toutefois, un appel destiné à cet abonné requiert une télé-recherche (ou *paging*) dans la zone de localisation LA_{ix}. Il en résulte que l'acheminement d'un appel fait appel à la séquence d'opérations illustrées à la Figure 4.7 et décomposées de la manière suivante :

a. Requête au $WING_{ij}$
b. Requête au HLR_i
c. Requête au VLR_{ix}

Si le mobile ne se trouve pas dans le sous-système *i* :

d. Requête au HLR_j
e. Requête au VLR_{jx}
f. Recherche du MT.

4.3 Évaluation de performance

Pour évaluer les performances des approches présentées, nous allons définir un certain nombre de paramètres, puis décrire les modèles de mobilité sur lesquels nous nous basons pour évaluer les taux de requêtes (nombre de requêtes par unité de temps) et le taux de mises à jour (nombre de mises à jour par unité de temps) au niveau de chaque équipement. Nous terminerons la section par la détermination du temps de réponse du système lorsque ce dernier est sollicité par une requête dans un contexte de mobilité globale.

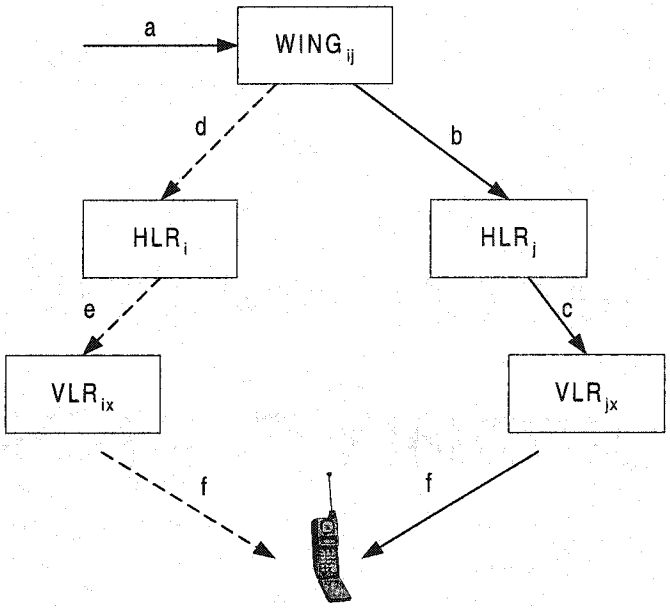


Figure 4.7 Processus d'acheminement d'une communication

4.3.1 Définition des paramètres

Pour évaluer et comparer les performances des approches de gestion de mobilité présentées, il convient de définir les paramètres suivants :

- N_{VLR_i} : nombre de VLRs du sous-système i;
- ρ_i : densité d'abonnés mobiles (ou de terminaux actifs) dans le sous-système i;
- v_i : vitesse moyenne dans la zone de localisation LA_{ix}, $x = 1, .., N_{VLR_i}$;
- A_i : aire totale du sous-système i;
- λ_{out_i} : taux moyen de génération d'appels dans le sous-système i (en appels/seconde/terminal);
- λ_{in_i} : taux moyen d'appels destinés aux usagers mobiles du sous-système i (en appels/seconde/terminal);

- $\tau^{M}_{req_e}$: taux de requêtes (nombre de requêtes par unité de temps) générées au niveau de l'entité *e* et calculées en utilisant la méthode M (*e* peut désigner aussi bien le VLR_{ix}, que le HLR_i, le MHLR ou un WING, tandis que M désigne soit l'approche SR, soit MR, soit notre approche);
- $\tau^{M}_{update_e}$: taux de mises à jour (nombre de mises à jour par unité de temps) générées au niveau de l'entité *e* et calculées en utilisant la méthode M.

4.3.2 Modèles de mobilité

L'évaluation des taux de requêtes et de mises à jour doit tenir compte du modèle de mobilité utilisé. Un tel modèle se base sur la vitesse, la direction et l'historique de déplacements de chaque abonné pour décrire ses mouvements. Lam *et al.* (1997) ont fait la description de plusieurs modèles de mobilité, dont le modèle fluide (*Fluid Flow Model*) et le modèle de gravité (*Gravity Model*). Dans le cadre de notre recherche, nous considérons des variations de ces deux modèles pour caractériser le comportement des usagers mobiles et évaluer les taux de requêtes et de mises à jour. Nous supposons alors que, dans un sous-système *i*, les utilisateurs sont uniformément distribués et leurs directions de mouvements sont uniformément réparties sur [0, 2π]. Dans ce cas, si nous considérons le modèle fluide et que nous désignons par L_i la longueur totale du sous-système *i*, le taux de traversée (en mobiles par seconde) du sous-système *i* au sous-système *j* est donné par la relation suivante :

$$r_{ij_F} = \rho_i v_i L_i/\pi \tag{4.1}$$

où ρ_i est la densité d'usagers mobiles dans le sous-système *i* et v_i la vitesse moyenne dans la zone de localisation LA_{ix}.

Pour analyser l'influence du modèle de mobilité utilisé sur les résultats, nous considérons également une variation du modèle de gravité. Un tel modèle est généralement utilisé dans la recherche sur le transport, mais peut tout aussi bien s'adapter

à des situations où des individus se déplacent d'une zone à une autre (Lam *et al.*, 1997). Dans le cadre de notre recherche, l'utilisation de ce modèle permet d'évaluer le taux d'abonnés (ou de terminaux actifs) qui se déplacent d'un sous-système i à un sous-système j de la manière suivante :

$$r_{ij_G} = K_{ij} * P_i * P_j \tag{4.2}$$

où P_i et P_j désignent le nombre de terminaux actifs respectivement dans les sous-systèmes i et j, alors que le paramètre K_{ij} peut être calculé de la manière suivante :

$$K_{ij} = \frac{m_i * m_j}{d_{ij}^{\gamma_i + \gamma_j}} \tag{4.3}$$

où d_{ij} est la distance entre les centres de gravité respectifs des sous-systèmes i et j, tandis que m_i, m_j, γ_i, γ_j sont des paramètres qui peuvent être calibrés. Lam *et al.* (1997) donnent quelques valeurs expérimentales de tels paramètres.

4.3.3 Taux de requêtes et de mises à jour

Nous nous basons sur les opérations exécutées lors des changements de sous-systèmes pour évaluer les taux de requêtes et de mises à jour exécutées au niveau de chaque entité (HLR ou HSS, VLR, WING ou MHLR). Cette évaluation se fait en modélisant chacune de ces entités par une file d'attente de type M/G/1. Ainsi, pour la méthode SR, nous utilisons la séquence d'opérations exécutée pendant le processus d'itinérance globale et présentée par Lin et Chlamtac (1996) pour évaluer, pour chaque sous-système i, les taux moyens d'arrivée des requêtes et des mises à jour au niveau du HLR, de chaque VLR, du WING et du MHLR. Nous en déduisons que les requêtes et les mises à jour forment des processus de Poisson dont les taux moyens d'arrivée à chaque entité sont donnés par les relations suivantes :

$$\tau^{SR}_{req_VLR_{ix}} = N_{VLR_i} [r_{ji} + (\lambda_{out_i} + \lambda_{in_i}) \rho_i A_i] \tag{4.4}$$

$$\tau^{SR}_{update_VLR_{ix}} = N_{VLR_i}(r_{ij} + r_{ji}) \quad (4.5)$$

$$\tau^{SR}_{req_VLR_{jx}} = N_{VLR_i} * r_{ij} + N_{VLR_j}[2 * r_{ji} + (\lambda_{out_j} + \lambda_{in_j})\rho_j * A_j] \quad (4.6)$$

$$\tau^{SR}_{update_VLR_{jx}} = N_{VLR_i} * (r_{ij} + r_{ji}) + 2 * N_{VLR_j} * r_{ji} \quad (4.7)$$

$$\tau^{SR}_{req_HLR_i} = N_{VLR_i}[r_{ji} + \lambda_{in_i} * \rho_i * A_i] \quad (4.8)$$

$$\tau^{SR}_{update_HLR_i} = N_{VLR_i}(r_{ij} + r_{ji}) \quad (4.9)$$

$$\tau^{SR}_{req_HLR_j} = N_{VLR_i} * r_{ij} + N_{VLR_j}[r_{ji} + \lambda_{in_j} * \rho_j * A_j] \quad (4.10)$$

$$\tau^{SR}_{update_HLR_j} = N_{VLR_i} * (r_{ij} + r_{ji}) + N_{VLRj} * r_{ji} \quad (4.11)$$

$$\tau^{SR}_{req_MHLR} = N_{VLR_j}(r_{ij} + \lambda_{in_j} * \rho_j * A_j) + N_{VLR_i}(r_{ji} + \lambda_{in_i} * \rho_i * A_i) \quad (4.12)$$

$$\tau^{SR}_{update_MHLR} = N_{VLR_i}(r_{ij} + r_{ji}) + N_{VLRj} * \rho_j * A_j(\lambda_{out_j} + \lambda_{in_j}) \quad (4.13)$$

où N_{VLR_i} et N_{VLRj} désignent le nombre respectif de VLRs des sous-systèmes i et j, alors que r_{ij} est équivalent à r_{ij_F} (relation 4.1) si nous considérons le modèle de mobilité fluide et égal à r_{ij_G} (relation 4.2) si nous considérons le modèle de gravité.

Dans le même ordre d'idées, pour la méthode MR, les taux de requêtes et de mises à jour au niveau de chaque entité (HLR ou HSS, VLR, WING ou MHLR) sont également déduits à partir des algorithmes présentés par Lin et Chlamtac (1996) et en modélisant chacune de ces entités par une file d'attente de type M/G/1. Nous en déduisons que les requêtes et les mises à jour forment des processus de Poisson dont les taux moyens d'arrivée à chaque entité sont exprimés par les relations suivantes :

$$\tau^{MR}_{req_VLR_{ix}} = N_{VLR_i} * [r_{ji} + (\lambda_{out_i} + \lambda_{in_i}) * \rho_i * A_i] + N_{VLR_j} * \lambda_{in_j} * \rho_j * A_j \quad (4.14)$$

$$\tau^{MR}_{update_VLR_{ix}} = N_{VLR_i} * r_{ji} \quad (4.15)$$

$$\tau^{MR}_{req_VLR_{jx}} = N_{VLR_j} * [2 * r_{ji} + (\lambda_{out_j} + \lambda_{in_j}) * \rho_j * A_j] \quad (4.16)$$

$$\tau^{MR}_{update_VLR_{jx}} = N_{VLR_i} * r_{ji} + 2 * N_{VLR_j} * r_{ji} \quad (4.17)$$

$$\tau^{MR}_{req_HLR_i} = N_{VLR_i} * [r_{ji} + \lambda_{in_i} * \rho_i * A_i] + N_{VLR_j} * \lambda_{in_j} * \rho_j * A_j \quad (4.18)$$

$$\tau^{MR}_{update_HLR_i} = N_{VLR_i} * r_{ji} \quad (4.19)$$

$$\tau^{MR}_{req_HLR_j} = N_{VLR_j} * [r_{ji} + \lambda_{in_j} * \rho_j * A_j] \quad (4.20)$$

$$\tau^{MR}_{update_HLR_j} = N_{VLR_i} * r_{ji} + N_{VLR_j} * r_{ji} \quad (4.21)$$

$$\tau^{MR}_{req_MHLR} = N_{VLR_j} * \lambda_{in_j} * \rho_j * A_j + N_{VLR_i} (r_{ji} + \lambda_{in_i} * \rho_i * A_i) \quad (4.22)$$

$$\tau^{MR}_{update_MHLR} = N_{VLR_i} * r_{ji} + N_{VLRj} * \rho_j * A_j \ (\lambda_{out_j} + \lambda_{in_j}) \quad (4.23)$$

En ce qui a trait à notre méthode, les taux de requêtes et de mises à jour au niveau des bases de données sont déduits à partir des opérations d'enregistrement, de mises à jours et d'acheminement présentées aux figures 4.5 à 4.7. En modélisant à nouveau chaque entité par une file d'attente de type M/G/1, les moyennes de ces taux s'expriment de la manière suivante :

$$\tau^{NM}_{req_VLR_{ix}} = 2 * r_{ji} + (\lambda_{out_i} + \lambda_{in_i}) * \rho_i * A_i \quad (4.24)$$

$$\tau^{NM}_{update_VLR_{ix}} = 2 * (r_{ij} + r_{ji}) \quad (4.25)$$

$$\tau^{NM}_{req_VLR_{jx}} = 2 * r_{ij} + (\lambda_{out_j} + \lambda_{in_j}) * \rho_j * A_j \quad (4.26)$$

$$\tau^{NM}_{update_VLR_{jx}} = 2 * (r_{ij} + r_{ji}) \quad (4.27)$$

$$\tau^{NM}_{req_HLR_i} = 2 * r_{ji} + \lambda_{in_i} * \rho_i * A_i \quad (4.28)$$

$$\tau^{NM}_{update_HLR_i} = 2 * (r_{ij} + r_{ji}) \quad (4.29)$$

$$\tau^{NM}_{req_HLRj} = 2 * r_{ij} + \lambda_{in_j} * \rho_j * A_j \quad (4.30)$$

$$\tau^{NM}_{update_HLRj} = 2 * (r_{ij} + r_{ji}) \quad (4.31)$$

$$\tau^{NM}_{req_WING_{ij}} = r_{ij} + r_{ji} + \lambda_{in_i} * \rho_i * A_i + \lambda_{in_j} * \rho_j * A_j \quad (4.32)$$

$$\tau^{NM}_{update_WING_{ij}} = 2 * (r_{ij} + r_{ji}) \quad (4.33)$$

4.3.4 Temps de réponse du réseau

Les taux de requêtes et de mises à jour exprimés dans les relations (4.4) à (4.33) permettent d'évaluer le taux moyen λ_e d'arrivée des messages au niveau de chaque entité e du système. Soit P_{update_e} la probabilité qu'un message destiné à l'entité e soit une mise à jour et P_{req_e} la probabilité que ce message soit une requête. Alors, P_{req_e} et P_{update_e} dépendent de la méthode utilisée. Pour en tenir compte, nous noterons respectivement ces probabilités $P^M_{req_e}$ et $P^M_{update_e}$ (M désignant soit la méthode SR, soit la méthode MR, soit notre approche). Dans ce cas, la probabilité $P^M_{req_e}$ s'exprime de la manière suivante :

$$P^M_{req_e} = 1 - P^M_{update_e} = \frac{\tau^M_{req_e}}{\tau^M_{req_e} + \tau^M_{update_e}} \quad (4.34)$$

où M représente l'une des méthodes présentées (SR, MR ou notre approche) et e une entité fonctionnelle quelconque du réseau (VLR, HLR, MHLR ou WING).

En supposant que les temps moyens de traitement d'une requête ou d'une mise à jour au niveau de chaque entité sont connus et notés respectivement $T_{pr_{req_e}}$ et $T_{pr_{update_e}}$, nous pouvons évaluer le temps moyen de traitement des messages au niveau d'une entité e de la manière suivante :

$$E(T_{pr_e}) = P_{req_e} * T_{pr_{req_e}} + P_{update_e} * T_{pr_{update_e}} \quad (4.35)$$

La relation (4.35) permet alors de calculer le moment de second ordre du temps de traitement de la manière suivante (Ross, 1997) :

$$E(T_{pr_e}^{2}) = P_{req_e} * T_{pr_{req_e}}^{2} + P_{update_e} * T_{pr_{update_e}}^{2} \qquad (4.36)$$

En modélisant chaque entité *e* par une file d'attente de type M/G/1, nous pouvons utiliser la formule de *Pollaczek-Kintchine* pour évaluer le délai moyen d'attente W_e à chaque entité *e* des sous-systèmes. Ce délai s'exprime de la manière suivante (Ross, 1997) :

$$W_e = \frac{\lambda_e * E(T_{pr_e}^2)}{2\left[1 - \lambda_e * E(T_{pr_e})\right]} \qquad (4.37)$$

où λ_e est le taux moyen d'arrivée des messages, c'est-à-dire la somme des requêtes et des mises à jour qui se présentent à l'entité *e*, alors que $E(T_{pr_e})$ et $E(T_{pr_e}^{2})$ sont respectivement donnés par (4.35) et (4.36).

Par ailleurs, l'amélioration d'une méthode de gestion de mobilité peut être obtenue en exécutant certains traitements au niveau d'un VLR plutôt qu'au niveau du HLR (Wey *et al.*, 1997). Dans ce cas, si l'on désigne par *p* la probabilité qu'un message soit traité au niveau d'un VLR, alors le temps de réponse du réseau aux requêtes provenant du sous-système *i* dépend du paramètre *p*. Désignons par T_{res_i} ce temps de réponse et par T_x le délai de transmission d'un message (supposé constant) entre le VLR_{ix} et le HLR_i. Posons $T_{pr_VLR_{ix}}$ et $T_{pr_HLR_i}$ les temps moyens de traitement respectivement au VLR_{ix} et au HLR_i. Notons $W_{VLR_{ix}}$ et W_{HLR_i} les temps moyens d'attente respectivement au VLR_{ix} et au HLR_i. En nous basant sur les activités des bases de données et sur le trafic de signalisation généré par la mobilité des usagers, nous pouvons définir le temps de réponse comme la somme pondérée des temps de réponse du VLR_{ix} et du HLR_i (Wey *et al.*, 1997). Plus précisément, le temps de réponse à partir du sous-système *i* est donné par la relation suivante :

$$T_{res_i} = p * (W_{VLRix} + T_{tr_VLRix}) + (1 - p) * (2 * T_x + W_{HLRi} + T_{tr_HLRi}) \qquad (4.38)$$

où T_{res_i} est le temps moyen de réponse du réseau lorsque le terminal mobile lance une requête à partir du sous-système *i*.

4.4 Résultats obtenus et analyse

Pour analyser le trafic de signalisation généré par l'itinérance globale, nous évaluons le nombre moyen de messages de signalisation échangés par seconde au niveau du MHLR et du WING, ainsi que le temps de réponse du système, en fonction du comportement des usagers. Cela permettra de comparer les performances de notre méthode à celles des méthodes existantes, en l'occurrence les méthodes SR et MR. Cette analyse se fait d'abord avec le modèle de mobilité fluide, ensuite avec le modèle de gravité.

4.4.1 Modèle de mobilité fluide

Pour ce modèle de mobilité, nous analysons dans un premier temps l'influence du comportement des abonnés sur les taux de requêtes et de mises à jour au niveau du MHLR et du WING. En général, pour l'itinérance classique (ou *intra-système*), le comportement des usagers se caractérise par le rapport *appels à mobilité* (*LCMR* : *Local Call-to-Mobility Ratio*), c'est-à-dire le rapport du taux moyen d'appels que reçoit un abonné au nombre de fois par seconde que cet abonné change de zones de localisation (Jain *et al.*, 1994). Toutefois, dans un contexte d'itinérance globale, nous caractérisons plutôt le comportement des usagers mobiles par le paramètre *GCMR* (*Global Call-to-Mobility Ratio*). Il s'agit du rapport entre le taux moyen d'appels émis ou reçus par un abonné au nombre de fois par seconde que cet abonné change de sous-systèmes (*intersystem handoff*).

La comparaison de performance des méthodes SR et MR avec notre approche se fait en considérant un bassin de deux (2) millions d'usagers mobiles qui utilisent les services de deux sous-systèmes hétérogènes *i* et *j*. Les paramètres de chaque sous-

système sont tirés de Lam *et al.* (1997) et présentés au Tableau 4.1. Ces paramètres permettent d'étudier, dans un premier temps, l'influence du comportement des usagers sur les taux de requêtes et de mises à jour. Cette étude se fait en traçant d'abord l'évolution du taux de requêtes, puis celle du taux de mises à jour en fonction du GCMR et ce, pour les méthodes SR, MR, ainsi que pour notre approche. Les figures 4.8 et 4.9 illustrent respectivement les taux de requêtes et de mises à jour pour chacune de ces méthodes. Nous nous rendons compte qu'en général, les taux de requêtes et de mises à jour au niveau du MHLR et du WING ont une décroissance exponentielle lorsque le paramètre GCMR augmente. Autrement dit, pour un taux fixé d'appels émis ou reçus, les taux de requêtes et de mises à jour exécutées au niveau des bases de données (MHLR ou WING) auront tendance à augmenter lorsque le degré de mobilité globale (changements de sous-systèmes) des abonnés augmente. Les résultats révèlent aussi que notre approche contribue à améliorer significativement les résultats obtenus à la fois par les méthodes SR et MR, en termes de taux de requêtes et de taux de mises à jour générées pendant le processus d'itinérance globale. Nous pouvons également mentionner que, pour $0.1 < GCMR < 0.42$, la méthode MR permet de réduire le taux de requêtes obtenu par la méthode SR. Toutefois, la méthode MR devient moins efficace que la méthode SR pour des valeurs plus élevées du GCMR. En fait, la méthode MR a été particulièrement conçue pour réduire surtout le taux de mises à jour au niveau du MHLR, ce qui est illustré à la Figure 4.9.

Les résultats précédents ont permis de comparer l'influence du comportement des usagers sur la quantité de trafic de signalisation généré au niveau du MHLR et du WING, ainsi que sur le temps de réponse du réseau. Analysons maintenant l'influence de la stratégie de stockage d'informations sur le temps de réponse du réseau, en faisant varier la probabilité p qu'une information requise lors d'une requête soit accessible au niveau d'un VLR. Une telle analyse vise à trouver un compromis entre la possibilité de rendre certaines informations disponibles au niveau des VLRs (ce qui accélère les temps de réponse, mais entraîne une augmentation du trafic dans les réseaux) et celle de concentrer cette information au niveau du HLR ou du HSS (ce qui permet de ne pas introduire de

trafic supplémentaire dans le réseau, mais engendre un temps de réponse plus long). Pour cette analyse, nous considérons les délais de traitement aux VLRs et au HLR indiqués au Tableau 4.2 et faisons varier le paramètre p de 0 à 1 pour simuler plusieurs patrons de trafic (Wey *et al.*, 1997). Nous illustrons à la Figure 4.11 l'évolution du temps de réponse aux requêtes en fonction du paramètre p. Cette figure montre que le temps de réponse décroît linéairement en fonction de p. Autrement dit, le temps de réponse du réseau tend à diminuer lorsque la probabilité que l'information requise soit directement accessible au niveau du VLR augmente. Les résultats confirment aussi que, peu importe la valeur de p ($0 < p < 1$), notre approche permet de réduire le temps de réponse du réseau.

Tableau 4.1 Paramètres d'analyse pour le modèle fluide

Paramètre	Sous-système i	Sous-système j
Nombre de LAs	6	4
Aire d'une cellule (km^2)	0.04	36.0
Vitesse moyenne (km/h)	5.0	20.0
Nombre de VLRs	6	4
Nombre de HLRs	1	1
λ_{in} (appels/seconde/terminal)	$8.333 * 10^{-4}$	$5.556 * 10^{-5}$
λ_{out} (appels/seconde/terminal)	$5.556 * 10^{-4}$	$2.7778 * 10^{-4}$

Nous avons également analysé, pour chaque approche, l'influence de la répartition des usagers sur les taux de requêtes et de mises à jour au niveau du MHLR et du WING, ainsi que sur le temps de réponse du système. Pour cette analyse, nous avons fait varier le pourcentage d'abonnés du sous-système i de 0.1% à 95% du bassin d'abonnés (évalué à 2

millions pour les deux sous-systèmes) et mesuré l'évolution des taux de requêtes, de mises à jour, ainsi que le temps de réponse, en utilisant à nouveau les paramètres du Tableau 4.1. Les résultats obtenus, illustrés aux figures 4.9, 4.10 et 4.11, montrent que, pour la série de paramètres choisis, les taux de requêtes et de mises à jour au niveau du MHLR et du WING, ainsi que le temps de réponse du système, diminuent lorsque le pourcentage d'abonnés faisant partie du sous-système *i* augmente. Autrement dit, le réseau offre une meilleure qualité de service lorsque les abonnés se concentrent davantage dans le sous-système *i*. Les résultats révèlent également que notre approche contribue à réduire significativement les taux de requêtes et de mises à jour au niveau du MHLR, ainsi que le temps de réponse du système, peu importe la répartition des abonnés.

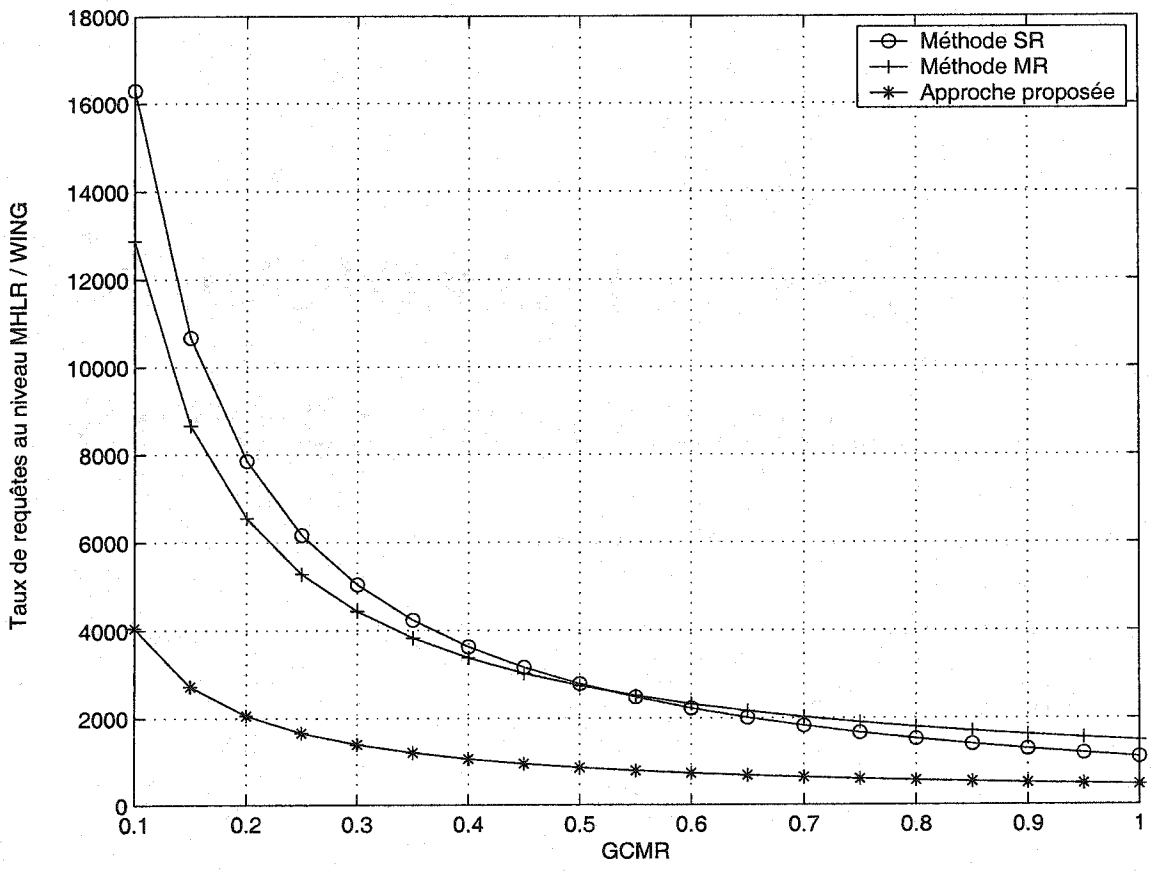


Figure 4.8 Influence du comportement des usagers sur le taux de requêtes

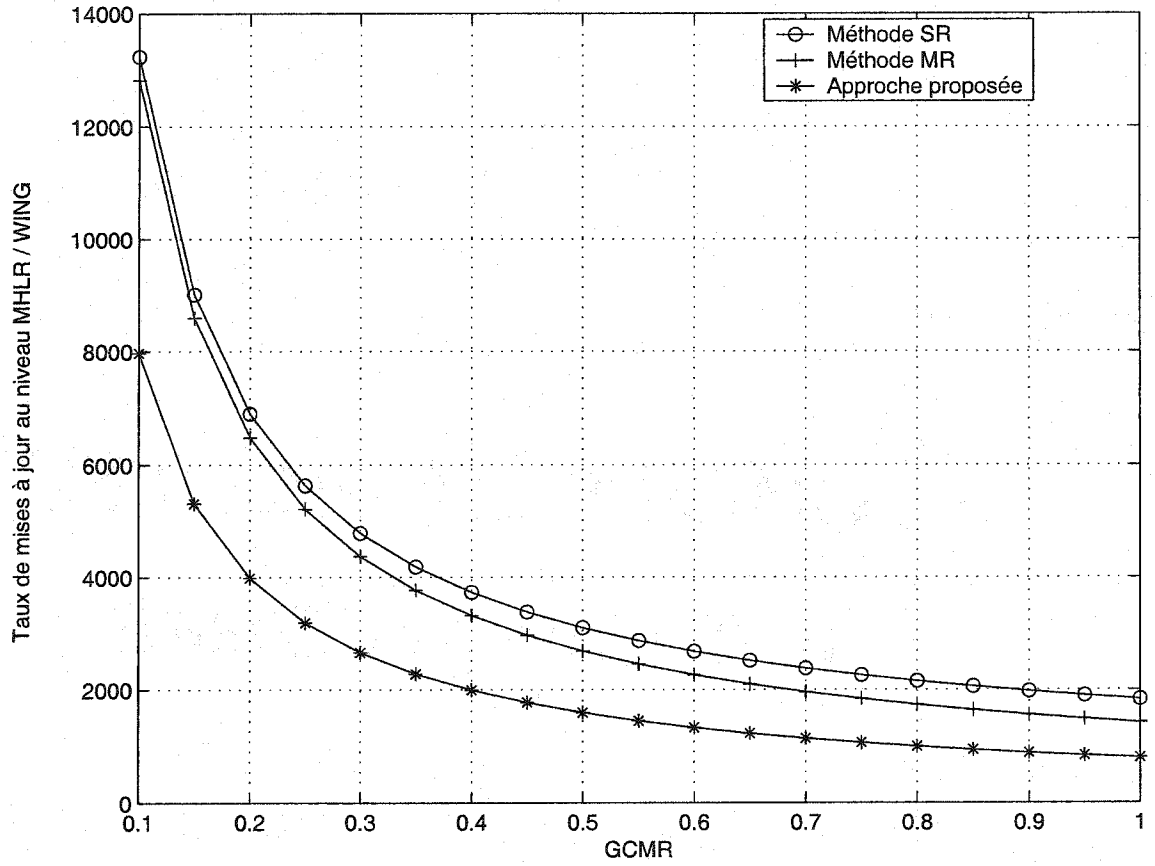


Figure 4.9 Influence du comportement des usagers sur le taux de mises à jour

Table 4.2 Temps moyen de traitement au niveau des bases de données

	Requête (ms)	Mise à jour (ms)
Au niveau du VLR	10	20
Au niveau du HLR/HSS	10	30

4.4.2 Modèle de gravité

Pour analyser l'influence du modèle de mobilité sur les résultats obtenus, nous allons étudier l'impact du comportement des abonnés à la fois sur le trafic de

signalisation et sur le temps de réponse, en considérant maintenant comme modèle de mobilité le modèle de gravité. Pour cette analyse, nous supposons une distance de 5.75 km entre les centres de gravité des sous-systèmes *i* et *j*, et utilisons les paramètres provenant de Lam *et al.* (1997), tels que représentés au Tableau 4.3, pour évaluer les taux de requêtes et de mises à jour au niveau du MHLR et du WING, ainsi que le temps de réponse du système. Ces paramètres permettent d'analyser l'influence du comportement des usagers mobiles sur le trafic de signalisation, puis celle de la répartition des usagers sur à la fois le taux de mises à jour et le temps de réponse du système.

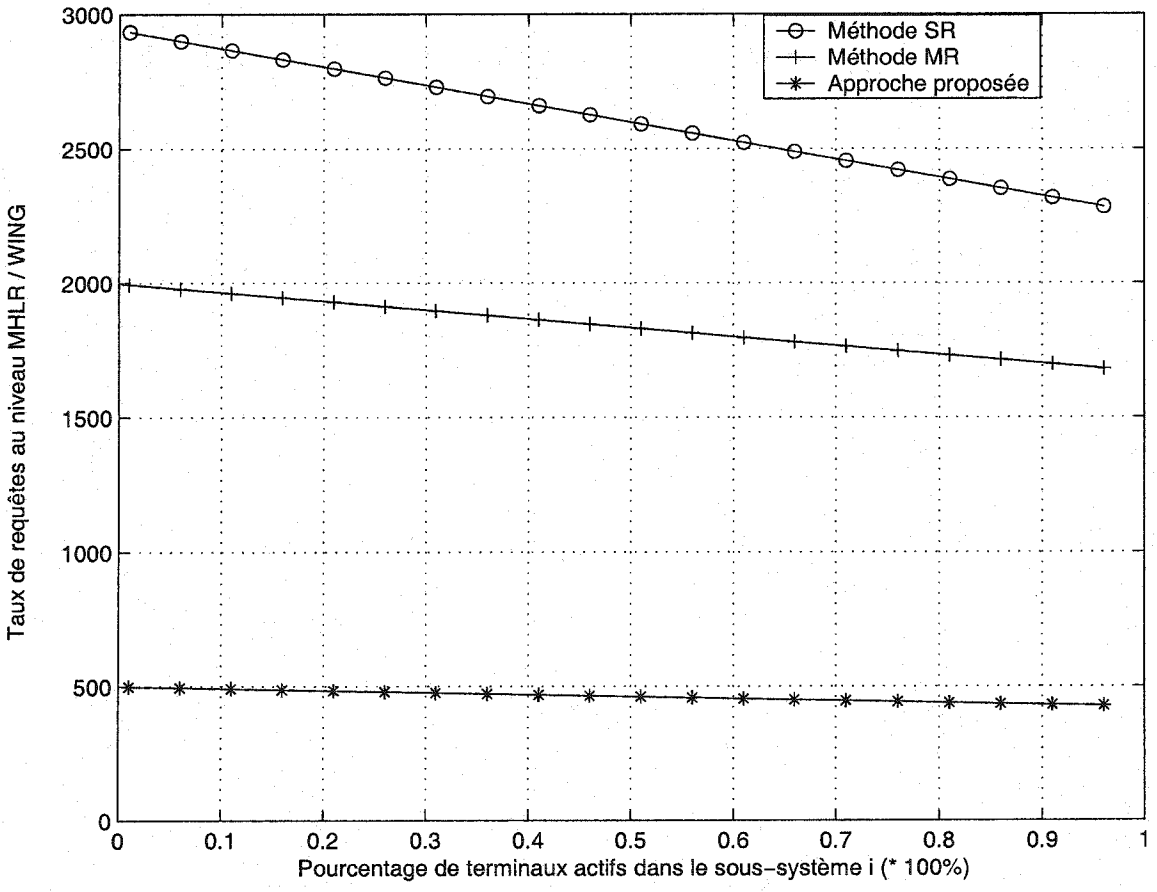


Figure 4.10 Influence de la distribution des usagers sur le taux de requêtes

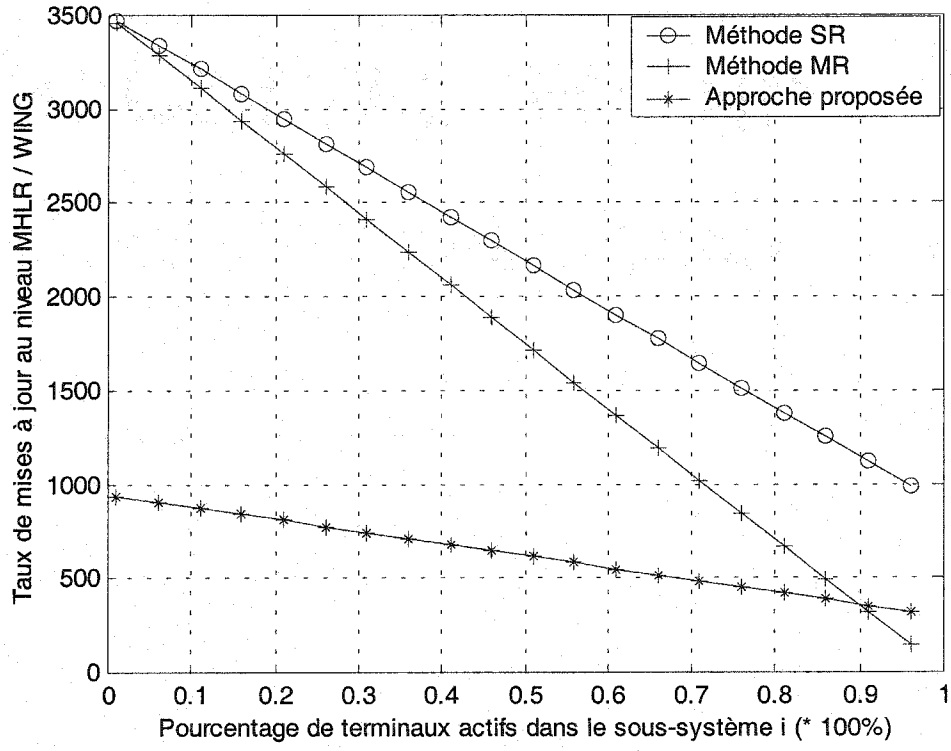


Figure 4.11 Influence de la distribution des usagers sur le taux de mises à jour

Tableau 4.3 Paramètres d'analyse pour le modèle de gravité

Paramètre	Sous-système *i*	Sous-système *j*
Nombre d'abonnés	75 000	425 000
Nombre de LAs	6	4
Aire d'une cellule (km^2)	0.04	56.25
γ	0.302	0.302
M	$4.118 * 10^{-4}$	$4.118 * 10^{-4}$
λ_{in} (appels/seconde/terminal)	$8.333 * 10^{-4}$	$5.556 * 10^{-5}$
λ_{out} (appels/seconde/terminal)	$5.556 * 10^{-4}$	$2.7778 * 10^{-4}$

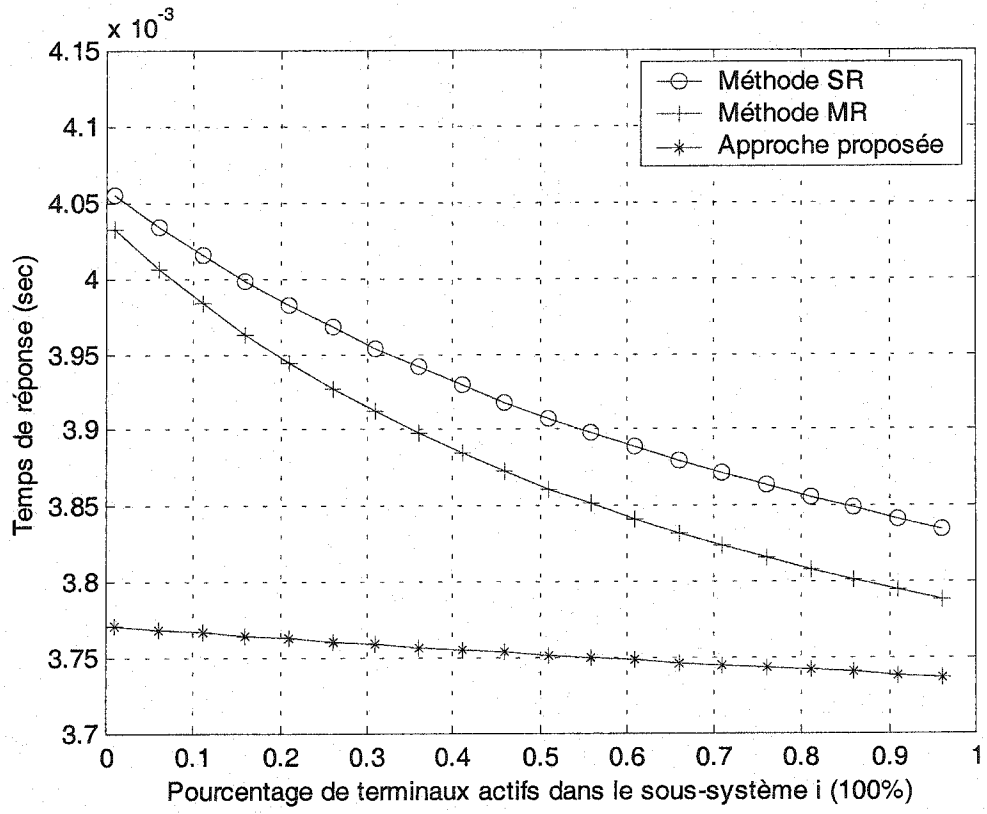


Figure 4.12 Influence de la distribution des usagers sur le temps de réponse

De manière plus particulière, nous illustrons respectivement aux figures 4.14 et 4.15 l'évolution du taux de requêtes et de mises à jour pour chacune des méthodes présentées. Nous nous rendons compte qu'à nouveau, les taux de requêtes et de mises à jour tant au niveau du MHLR qu'au niveau du WING ont une décroissance exponentielle en fonction du GCMR. Cela signifie que, pour un taux fixé d'appels émis ou reçus, le nombre de requêtes et de mises à jour exécutées par unité de temps au niveau des bases de données (MHLR ou WING) aura tendance à augmenter rapidement lorsque le degré de mobilité globale des abonnés augmente. Cela confirme que, peu importe le modèle de mobilité utilisé, le réseau offre une meilleure qualité de service lorsque le GCMR augmente. Les résultats révèlent également que, pour le modèle de gravité, notre

approche contribue à améliorer significativement les résultats obtenus à la fois par les méthodes SR et MR.

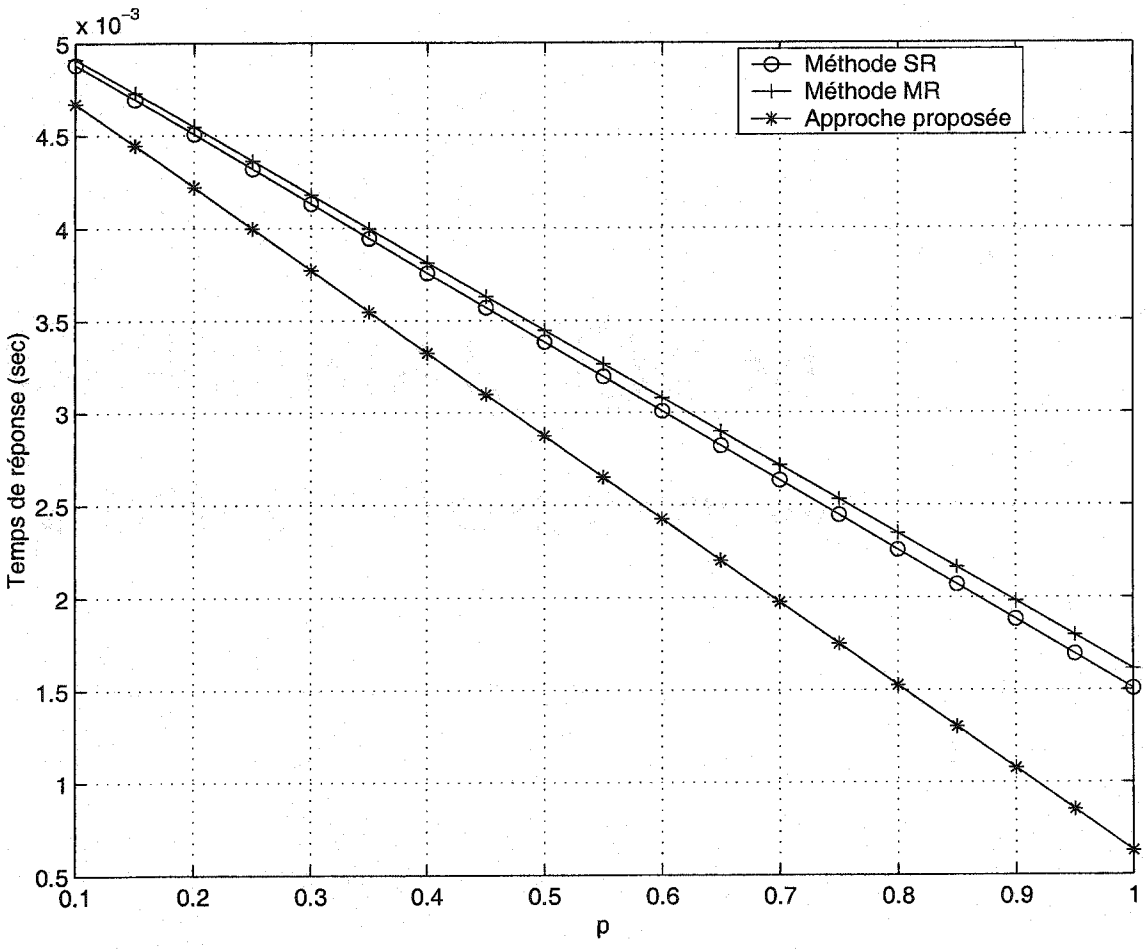


Figure 4.13 Influence du paramètre *p* sur le temps de réponse

Nous avons également analysé, pour le modèle de gravité, l'influence de la répartition des usagers respectivement sur le taux de requêtes et sur le temps de réponse du système, en faisant varier le pourcentage d'abonnés se trouvant dans le sous-système *i* de 0 à 100%. Les résultats obtenus de cette analyse sont illustrés aux figures 4.16 et 4.17. Ces résultats révèlent que le taux de requêtes, ainsi que le temps de réponse du système, augmentent linéairement lorsque le pourcentage d'abonnés du sous-système *i* augmente.

Autrement dit, le réseau offre une meilleure qualité de service lorsque les abonnés se concentrent davantage dans le sous-système j : il s'agit de l'effet inverse du modèle fluide. Toutefois, les résultats illustrés aux figures 4.16 et 4.17 montrent que notre approche contribue à réduire significativement à la fois le taux de requêtes et le temps de réponse obtenu par les méthodes SR et MR, peu importe la répartition des abonnés.

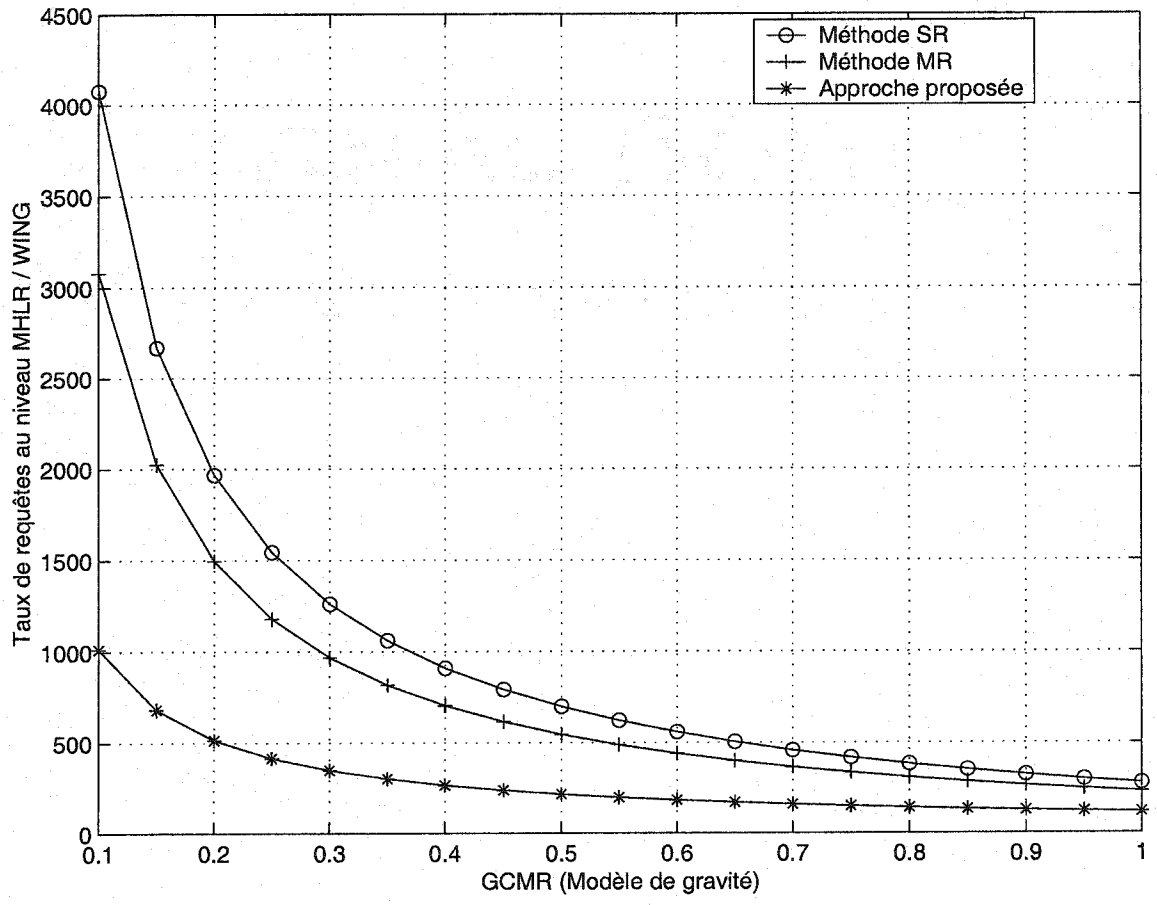


Figure 4.14 Influence du GCMR sur le taux de requêtes (modèle de gravité)

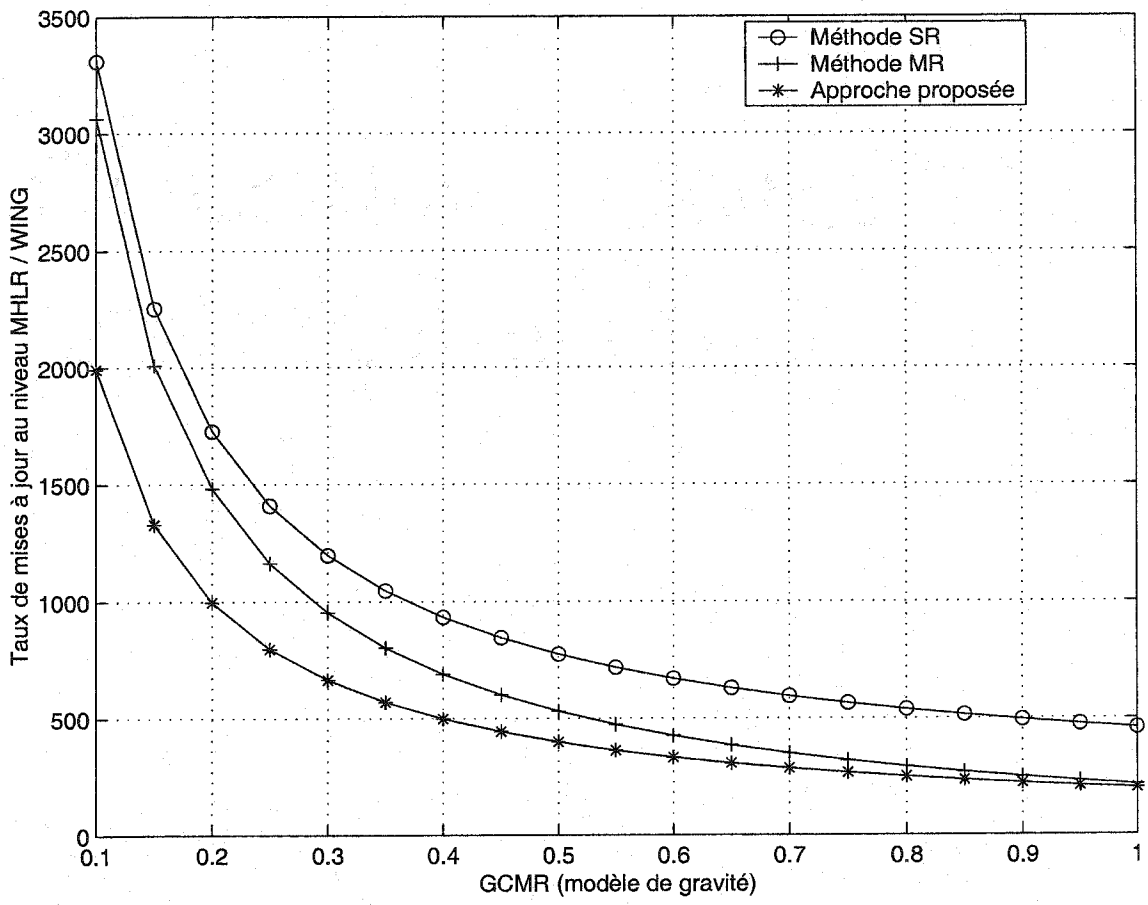


Figure 4.15 Influence du GCMR sur le taux de mises à jour (modèle de gravité)

Ainsi, l'analyse des résultats montre clairement les avantages offerts par notre approche de gestion de mobilité globale. Ce sont les suivants :

- Faibles taux de requêtes et de mises à jour au niveau des bases de données;
- Localisation rapide des usagers;
- Simplicité de mise en œuvre;
- Faible temps de réponse aux requêtes des usagers.

Cela contribuera à améliorer significativement les performances du réseau. Le chapitre suivant fera l'objet d'un autre facteur fondamental de la planification des systèmes mobiles de la prochaine génération : il s'agit de l'ingénierie du trafic.

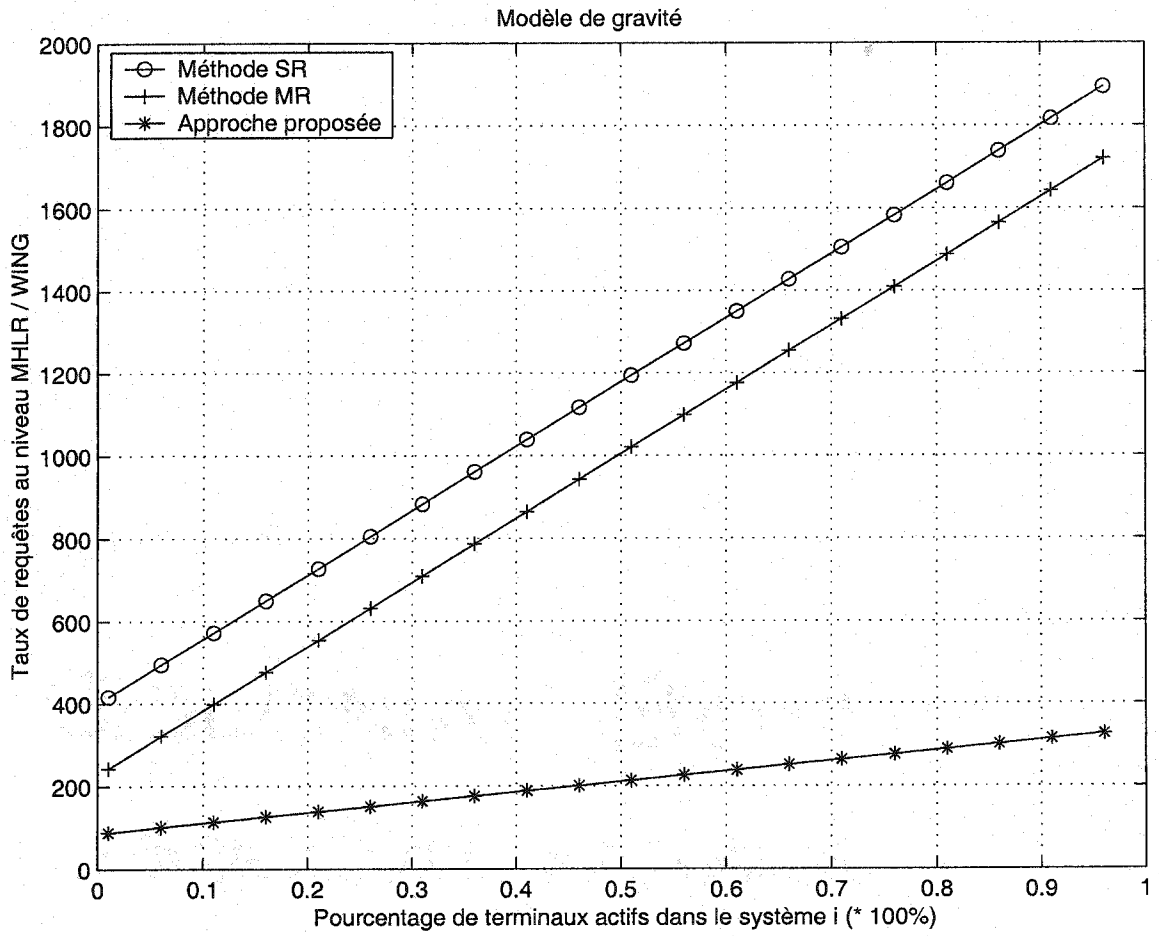


Figure 4.16 Influence de la distribution des abonnés sur le taux de requêtes (modèle de gravité)

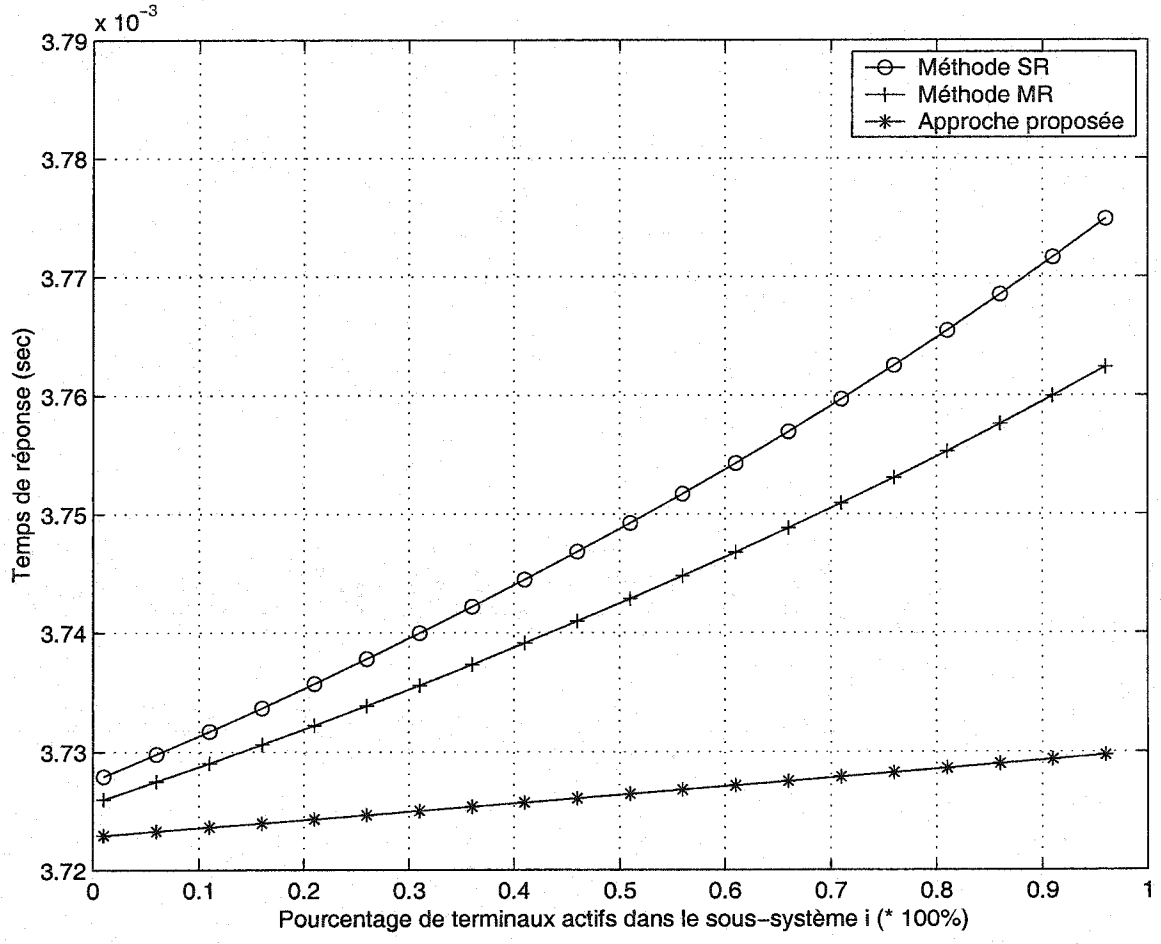


Figure 4.17 Influence de la distribution des abonnés sur le temps de réponse (modèle de gravité)

CHAPITRE 5

INGÉNIERIE DU TRAFIC

L'intégration des services, combinée au caractère aléatoire des déplacements des usagers et de la durée des communications, rend difficile l'analyse du trafic dans les systèmes mobiles de la prochaine génération (Lam *et al.*, 1997; Orlik et Rappaport, 1998; Fang *et al.*, 2000). Dans ce contexte, il est primordial de mettre au point des modèles efficaces qui permettront non seulement d'évaluer l'intensité de trafic dans chaque cellule, mais aussi de déterminer le nombre maximal d'usagers mobiles qui peuvent y être desservis avec une bonne qualité de service. Dans ce chapitre, nous présentons deux modèles de files d'attente pour caractériser le trafic dans chaque cellule : le M/G/∞ et le G/G/c/c. Le premier modèle se base sur la capacité élastique de la technologie CDMA (Viterbi et Viterbi, 1993) pour modéliser chaque cellule par une file d'attente avec un nombre infini de serveurs, c'est-à-dire un système n'ayant aucune limite théorique sur le nombre de communications à gérer. Quant au second modèle, il permettra d'évaluer de manière réaliste à la fois la distribution du trafic et la probabilité de blocage d'appels. Toutefois, avant la présentation de ces deux modèles, il convient de caractériser les principaux paramètres de modélisation. Cela permettra par la suite de comparer la distribution de trafic des pico-cellules à celle des micro-cellules et celle des macro-cellules. Pour finir, nous utiliserons le second modèle pour évaluer l'impact des coefficients de variation du trafic d'arrivée et de la durée d'occupation des canaux sur l'intensité du trafic et sur la qualité de service, et montrerons comment adapter les modèles proposés au trafic multimédia.

5.1 Paramètres de modélisation

La modélisation du trafic dans les systèmes mobiles de la prochaine génération fait appel à la détermination des distributions qui caractérisent à la fois le trafic d'arrivée

et la durée d'occupation des canaux. Nous allons passer en revue les distributions les plus couramment utilisées dans la littérature pour caractériser ces paramètres. Mais, présentons tout d'abord les principes de base de la modélisation des cellules.

5.1.1 Principes de modélisation des cellules

Dans le contexte des réseaux mobiles, chaque cellule peut être modélisée par une file d'attente où les arrivées correspondent à des demandes de connexions (demandes de ressources), les départs correspondent aux déconnexions engendrées par la fin des communications ou par des relèves vers les cellules adjacentes (libération des ressources). Les serveurs représentent alors les canaux disponibles, tandis que les clients représentent les terminaux actifs, c'est-à-dire les usagers mobiles en communication. Ainsi, à chaque fois qu'un usager mobile allume son terminal pour effectuer un appel (arrivée d'un nouvel appel), un canal de communication doit lui être affecté. Dans le même ordre d'idée, lorsque cet usager change de cellule durant une communication (arrivée d'un appel de relève), un canal doit lui être réservé dans la nouvelle cellule pour permettre la poursuite de la communication. Dans les deux cas, on parle de demande de service. Le temps de service correspond alors à la durée d'occupation d'un canal, c'est-à-dire l'intervalle de temps pendant lequel l'usager garde ce canal occupé pendant son séjour dans la cellule (Orlik et Rappaport, 1998). Le temps de service est ainsi directement lié à la durée de séjour dans une cellule, c'est-à-dire l'intervalle de temps pendant lequel un usager mobile demeure dans cette cellule durant une communication (Fang et Chlamtac, 1999). Il convient alors de mentionner que le temps de service est différent de la durée de la communication qui indique le temps écoulé entre le début et la fin d'une communication. Les durées d'occupation des canaux, les durées de séjour dans les cellules, les durées des communications et les arrivées des appels étant aléatoires, nous allons les caractériser par des fonctions de densité de probabilité qu'il faudra éventuellement déterminer pour évaluer la distribution du trafic dans la zone de service considérée.

5.1.2 Caractérisation du trafic d'arrivée

Le trafic d'arrivée dans une cellule se compose d'une combinaison de trafic provenant à la fois des nouveaux appels et des appels de relève. En général, le trafic provenant des nouveaux appels est considéré comme une variable de Poisson (Lin *et al.*, 1994; Jabbari, 1996; Fang *et al.*, 2000). Dans ce cas, si nous désignons par $X(t)$ le nombre de nouveaux appels générés dans une cellule pendant un intervalle t, alors la distribution en régime permanent de $X(t)$ est donnée par (Kleinrock, 1975; Leon-Garcia, 1994) :

$$P[X = i] = p(i) = e^{-\lambda} \frac{\lambda^i}{i!}, \qquad i = 0, 1, \ldots \tag{5.1}$$

où λ (> 0) désigne le paramètre de la distribution de Poisson. La moyenne de X est alors donnée par $E[X] = \lambda$. Autrement dit, les nouveaux appels sont générés dans chaque cellule selon un processus de Poisson à un taux moyen de λ par unité de temps.

D'autre part, si nous désignons par Y le nombre d'appels de relève qui se poursuivent dans une cellule donnée, alors la variable aléatoire Y forme un processus indépendant de X (Orlik et Rappaport, 1998). Autrement dit, l'événement de l'un n'altère pas la probabilité de l'autre. De plus, si la probabilité de blocage est nulle, la variable aléatoire Y forme également un processus de Poisson (Chlebus et Ludwing, 1995). Dans ce cas, le trafic d'arrivée est la variable aléatoire $(X+Y)$, c'est-à-dire la combinaison du trafic résultant des nouveaux appels et des appels de relève. Nous allons prouver que ce trafic forme également un processus de Poisson de moyenne $(\lambda_1+\lambda_2)$, où λ_1, $\lambda_2 > 0$ désignent les paramètres respectifs de X et de Y.

Considérons que X et Y ont respectivement comme moyennes λ_1 et λ_2. Alors, l'événement $\{X + Y = k\}$ peut être écrit comme l'union de deux événements disjoints $\{X = k, Y = n - k\}$, $0 \le k \le n$. Nous avons alors :

$$P[X + Y = n] = \sum_{k=0}^{n} P[X = k, Y = n - k]$$

Puisque X et Y sont indépendantes, nous pouvons écrire :

$$P[X+Y=n] = \sum_{k=0}^{n} P[X=k]*P[Y=n-k]$$

Si X et Y suivent une loi exponentielle, nous obtenons :

$$P[X+Y=n] = \sum_{k=0}^{n} e^{-\lambda_1}\frac{\lambda_1^k}{k!} \quad e^{-\lambda_2}\frac{\lambda_2^{n-k}}{k!}$$

$$P[X+Y=n] = e^{-(\lambda_1+\lambda_2)}\sum_{k=0}^{n}\frac{\lambda_1^k * \lambda_2^{n-k}}{k!(n-k)!}$$

$$P[X+Y=n] = \frac{e^{-(\lambda_1+\lambda_2)}}{n!}\sum_{k=0}^{n}\frac{n!}{k!(n-k)!}\lambda_1^k * \lambda_2^{n-k}$$

À ce niveau, il convient d'utiliser la relation suivante (Ross, 1997) :

$$(\lambda_1+\lambda_2)^n = \sum_{k=0}^{n}\frac{n!}{k!(n-k)!}\lambda_1^k * \lambda_2^{n-k}$$

Ce qui permet d'obtenir la distribution de la variable aléatoire $X+Y$ de la manière suivante :

$$P[X+Y=n] = \frac{e^{-(\lambda_1+\lambda_2)}}{n!}(\lambda_1+\lambda_2)^n \qquad (5.2)$$

Nous concluons que, lorsque X et Y sont indépendantes et suivent chacune une loi de Poisson, la variable aléatoire $(X+Y)$ suit également une loi de Poisson de moyenne $\lambda_1+\lambda_2$. Autrement dit, lorsque la probabilité de blocage d'appels dans les cellules est nulle, le nombre total de demandes de connexions dans chaque cellule suit une loi de Poisson et se fait à un taux moyen de $\lambda_1+\lambda_2$ par unité de temps.

5.1.3 Durée d'occupation des canaux

La durée d'occupation d'un canal est un paramètre primordial à la modélisation du trafic. Nous allons proposer une méthode pour l'évaluer et passer en revue une série de distributions qui peuvent la caractériser.

5.1.3.1 Évaluation

Sans perte de généralité, considérons les deux scénarios illustrés à la Figure 5.1. Pour chaque scénario, supposons qu'un usager mobile établit une connexion au temps t_0, suit la trajectoire indiquée et libère cette connexion au temps t_1. Pour les deux scénarios, cette connexion génère l'arrivée d'un nouvel appel à la cellule 1. Pour le scénario 1, puisque l'appel commence et se termine à l'intérieur de la même cellule, il n'y a pas de trafic de relève et la durée d'occupation du canal affecté à la communication est équivalente à la durée de séjour dans la cellule, ce qui correspond également à la durée de l'appel. Si nous désignons par T_c, T_s et T_r respectivement la durée d'occupation du canal, la durée de la communication et la durée du séjour dans la cellule, nous avons : $T_c = T_s = T_r$. Toutefois, pour le scénario 2, où la communication est générée à la cellule 1, traverse la cellule 2 et se termine à la cellule 3, nous avons une arrivée résultant du trafic de relève respectivement dans les cellules 2 et 3. La durée d'occupation d'un canal dans chaque cellule correspond à la durée du séjour à l'intérieur de cette cellule, ce qui est inférieur à la durée de la communication. Dans ce cas, nous avons : $T_c = T_r < T_s$. En général, nous avons :

$$T_c = \min\{T_s, T_r\} \tag{5.3}$$

Dans cette analyse, nous supposons que T_s et T_r sont des variables aléatoires dont les fonctions de densité de probabilité sont connues.

Nous nous rendons compte que la durée d'occupation d'un canal dépend à la fois de la mobilité des usagers et de la taille des cellules. En fait, lorsqu'une cellule est suffisamment grande par rapport à la distance qu'un usager mobile parcourt lors d'une communication, la durée d'occupation d'un canal peut être approximée par la durée de la communication, ce qui est illustré au scénario 1 de la Figure 5.1.

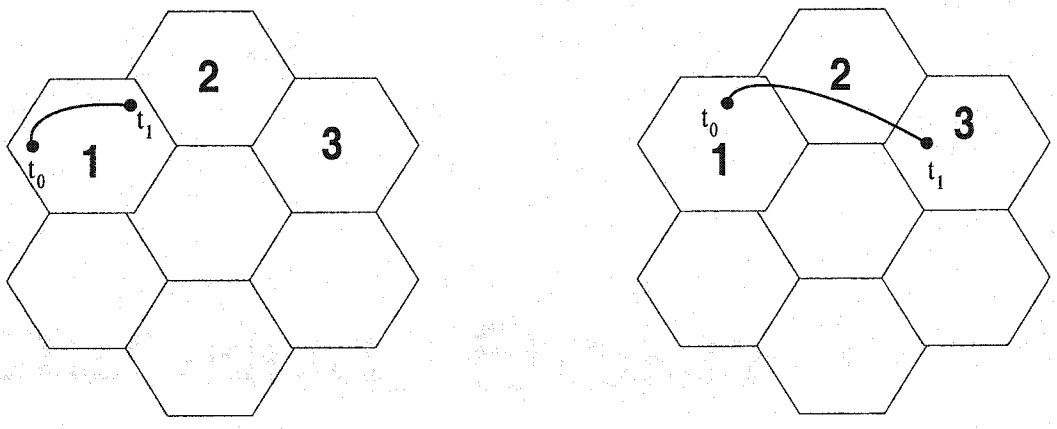


Figure 5.1 Détermination de la durée d'occupation d'un canal

5.1.3.2 Durée d'occupation exponentielle

Par souci de simplicité, plusieurs analyses de trafic supposent que les durées de séjour, ainsi que les durées des communications sont distribuées selon une loi exponentielle (Jabbari, 1996). Une telle hypothèse est justifiée dans deux situations : lorsque les appels sont de courte durée et lorsque la dimension d'une cellule est suffisamment grande pour permettre à un appel de commencer et de se terminer à l'intérieur de cette même cellule (Fang *et al.*, 1997; Fang et Chlamtac, 1999). Nous allons montrer que, lorsque la durée de séjour T_r et la durée totale des communications T_s sont indépendantes et exponentiellement distribuées avec des paramètres respectifs λ_1 et λ_2, la durée d'occupation des canaux T_c sera aussi exponentiellement distribuée.

Soit $f_{T_c}(t)$ la fonction de densité de probabilité de T_c. Nous avons (Ross, 1997) :

$$f_{T_c}(t) = \frac{d}{dt}\left[P(T_c \le t)\right]$$

$$f_{T_c}(t) = \frac{d}{dt}\left[1 - P(T_c \ge t)\right] \qquad (5.4)$$

À partir de (5.3), nous pouvons déterminer $P[T_c \ge t]$ de la manière suivante :

$$P[T_c \geq t] = P[\min(T_s, T_r) \geq t]$$

$$P[T_c \geq t] = P[T_s \geq t, T_r \geq t] \quad (5.5)$$

L'indépendance de T_s et de T_r permet d'écrire (5.5) de la manière suivante :

$$P[T_c \geq t] = P[T_s \geq t] * P[T_r \geq t]$$

$$P[T_c \geq t] = e^{-\lambda_1 t} * e^{-\lambda_2 t}$$

$$P[T_c \geq t] = e^{-(\lambda_1+\lambda_2)t} \quad (5.6)$$

En combinant les relations (5.4) et (5.6), nous obtenons :

$$f_{T_c}(t) = (\lambda_1 + \lambda_2)\, e^{-(\lambda_1+\lambda_2)t}, \quad t \geq 0 \quad (5.7)$$

En conséquence, lorsque la durée de séjour dans la cellule T_r et la durée des communications T_s sont indépendantes et exponentiellement distribuées de paramètres respectifs λ_1 et λ_2, la durée d'occupation des canaux T_c est aussi exponentiellement distribuée de paramètre $\lambda = (\lambda_1+\lambda_2)$. La fonction de densité de probabilité de T_c peut alors être réécrite de la manière suivante :

$$f_{T_c}(t) = \begin{cases} \lambda\, e^{-\lambda t}, & \text{si } t \geq 0 \\ 0, & \text{sinon} \end{cases} \quad (5.8)$$

où λ ($\lambda > 0$) représente le paramètre de la distribution. Il en résulte que la durée moyenne d'occupation des canaux est donnée par $1/\lambda = 1/(\lambda_1+\lambda_2)$. La Figure 5.2 illustre le comportement de T_c pour trois valeurs différentes de λ.

5.1.3.3 Durée d'occupation des canaux de loi générale

Pour les systèmes mobiles de la prochaine génération, il est plus réaliste de modéliser à la fois les durées de séjour et les durées des communications par des lois générales, ce qui permet de prendre en compte le degré de variabilité des types d'applications, la dimension ou la forme des cellules, ainsi que la mobilité des usagers. Il

en résulte que la durée d'occupation T_c des canaux suivra une loi générale (Fang et Chlamtac, 1999; Fang *et al.*, 2000). Spécifions les principales distributions générales qui sont les plus couramment utilisées dans la littérature pour caractériser T_c.

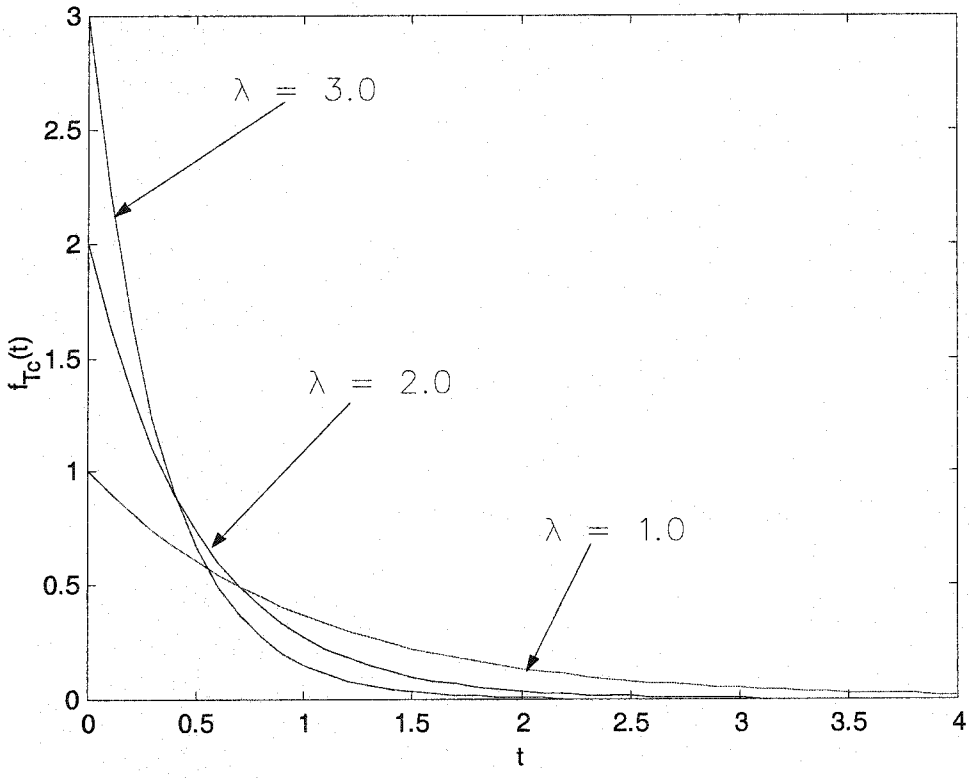


Figure 5.2 Distribution exponentielle de T_c

Durée d'occupation de loi hyperexponentielle

La variable aléatoire T_c est distribuée selon une *loi hyperexponentielle* si sa fonction de densité de probabilité est donnée par la relation suivante (Orlik et Rappaport, 1998) :

$$f_{T_c}(t) = \begin{cases} \sum_{i=1}^{k} \alpha_i \, \mu_i \, e^{-\mu_i t}, & \text{si } t \geq 0 \\ 0, & \text{sinon} \end{cases} \tag{5.9}$$

où $\sum_{i=1}^{k} \alpha_i = 1$, $\mu_i > 0$. La Figure 5.3 illustre la forme de plusieurs distributions hyperexponentielles pour $k = 2$, $\mu_1 = 1$, $\mu_2 = 4$ et plusieurs combinaisons des α_i.

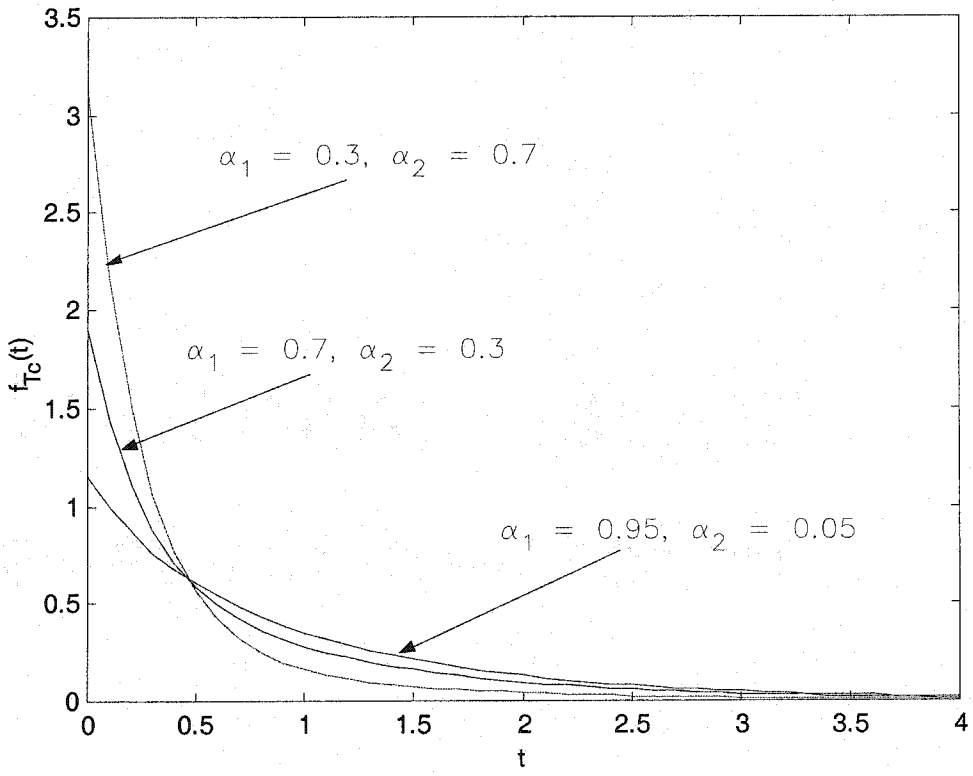


Figure 5.3 Distribution hyperexponentielle de T_c avec $k = 2$, $\mu_1 = 1$, $\mu_2 = 4$

Par ailleurs, Orlik et Rappaport (1998) ont défini une extension de la distribution hyperexponentielle appelée SOHYP (Sum Of HYPerexponentials) et qui, selon eux, permet de bien caractériser à la fois les durées des communications et les durées de séjour

dans chaque cellule. En plus de permettre la prédiction des performances de chaque cellule, la distribution SOHYP préserve les propriétés markoviennes des modèles de files d'attente qui modélisent les cellules, ce qui en facilite l'analyse et la résolution (Orlik et Rappaport, 1998). Toutefois, selon Fang et Chlamtac (1999), le modèle SOHYP n'est pas en mesure de tenir compte de tous les facteurs qui interviennent dans l'évaluation de la durée d'occupation des canaux, d'où l'importance de trouver d'autres distributions pour caractériser T_c.

Durée d'occupation de loi normale ou lognormale

La durée d'occupation des canaux T_c peut être aussi modélisée par une distribution *normale* (Jabbari, 1996), ou par une distribution *lognormale* (Jedrzycki et Leung, 1996; Barcelo et Jordan, 1997). On dit que la variable aléatoire T_c est distribuée selon une loi normale si sa fonction de densité de probabilité est de la forme (Crow et Shimizu, 1988) :

$$f_{T_c}(t) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(t-m)^2}{2\sigma^2}}, \quad -\infty \leq t \leq \infty \tag{5.10}$$

où m et σ sont des nombres réels non négatifs qui représentent respectivement la moyenne et l'écart-type de T_c. La Figure 5.4 montre que la distribution normale est une courbe en forme de cloche centrée et symétrique par rapport à la moyenne (m = 5), et dont la largeur augmente en fonction de σ.

Une extension de la distribution normale, souvent utilisée comme durée d'occupation des canaux, est la distribution lognormale (Jedrzycki et Leung, 1996; Barcelo et Jordan, 1997), c'est-à-dire la distribution d'une variable aléatoire dont le logarithme népérien est distribué selon une loi normale. Autrement dit, une variable aléatoire T est distribuée selon une loi lognormale de paramètres m et σ^2 si $T_c = ln\ T$ est distribuée selon une loi normale de moyenne m et de variance σ^2. La fonction de densité de probabilité de T est alors donnée par la relation (Crow et Shimizu, 1988) :

$$f_T(t) = \begin{cases} \dfrac{1}{\sigma t\sqrt{2\pi}} e^{\frac{-(\ln t - m)^2}{2\sigma^2}}, & \text{si } t > 0 \\ 0, & \text{sinon} \end{cases} \qquad (5.11)$$

La Figure 5.5 illustre la distribution de T pour $m = 0$, $\sigma =$ 0.1, 0.3 et 1.0.

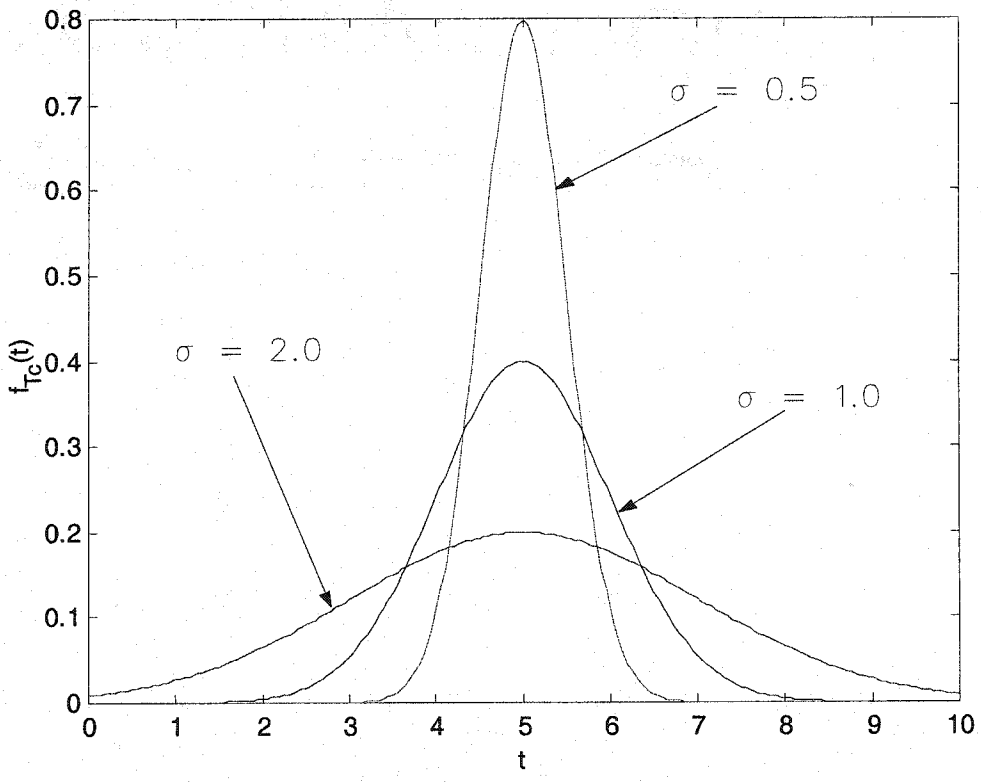


Figure 5.4. Distribution normale de T_c avec m = 5

Une série de mesures prises sur le champ par Jedrzycki et Leung (1996) ont statistiquement montré que la distribution lognormale donne une bonne approximation de la durée d'occupation des canaux. Pour leur part, Barcelo et Jordan (1997) ont prouvé que la distribution lognormale combinée à plusieurs autres distributions bien connues de la littérature offre une bonne cohérence statistique de la durée d'occupation des canaux.

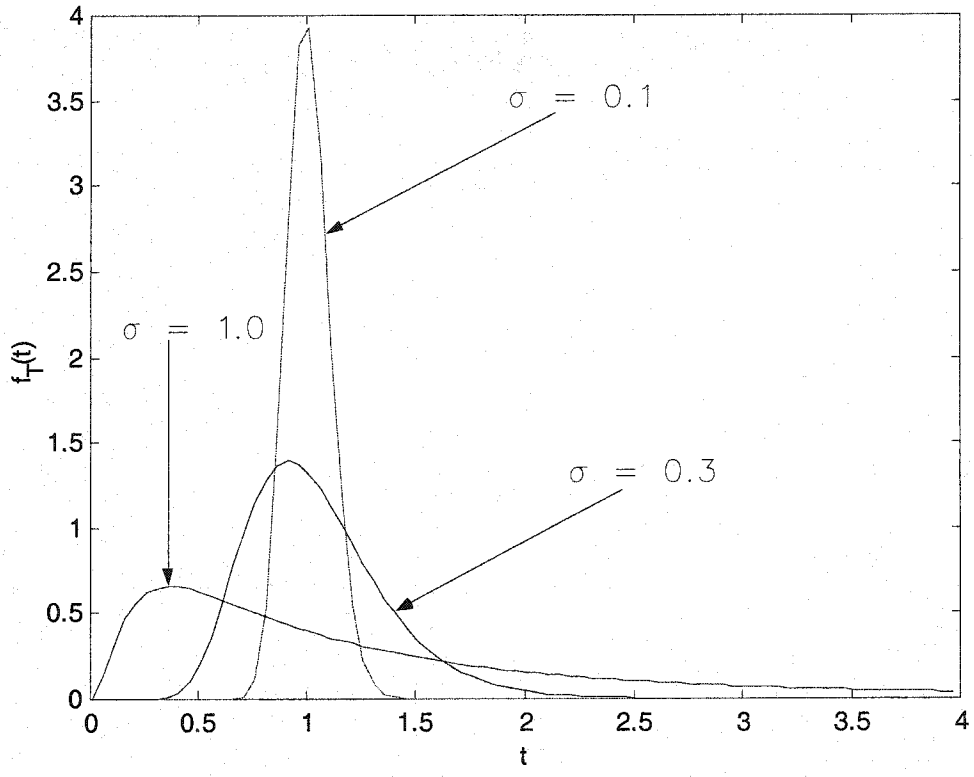


Figure 5.5 Distribution lognormale de *T* avec *m* = 0

Durée d'occupation de loi Gamma

La durée d'occupation des canaux T_c peut aussi être distribuée selon la loi *Gamma* (Fang *et al.*, 2000). Dans ce cas, sa fonction de densité de probabilité est donnée par la relation suivante (Leon-Garcia, 1994) :

$$f_{T_c}(t) = \frac{\alpha^{\gamma} (t)^{\gamma-1} e^{-\alpha t}}{\Gamma(\gamma)}, \qquad t > 0 \tag{5.12}$$

où γ et α désignent respectivement des paramètres de forme et d'échelle, alors que $\Gamma(\gamma)$ représente la fonction Gamma. Cette dernière est définie par l'intégrale :

$$\Gamma(\gamma) = \int_0^\infty x^{\gamma-1} e^{-x}\, dx, \qquad \gamma > 0$$

et possède les propriétés suivantes (Leon-Garcia, 1994) :

$$\Gamma(½) = \sqrt{\pi}$$

$$\Gamma(\gamma + 1) = \gamma\, \Gamma(\gamma)$$

$\Gamma(m + 1) = m!$, pour m entier non négatif.

La loi Gamma a l'avantage d'avoir une transformée de Laplace simple, ce qui favorise le développement de formules simples pour l'évaluation des performances (Fang *et al.*, 1997). De plus, comme illustré à la Figure 5.6, la distribution Gamma peut prendre plusieurs formes, en faisant varier ses paramètres. De cette façon, il est possible de faire correspondre la distribution Gamma à plusieurs séries de données expérimentales. Par exemple, lorsque $\gamma = 1$, la distribution Gamma est équivalente à la distribution exponentielle, alors que lorsque la valeur γ est élevée, elle devient équivalente à une normale de moyenne α (Fang *et al.*, 2000). De plus, lorsque $\gamma = k$ (entier positif), la loi Gamma devient équivalente à la loi d'*Erlang-k* dont la fonction de densité de probabilité est donnée par la relation (Fang *et al.*, 1997) :

$$f_{T_c}(t) = \begin{cases} \dfrac{\alpha^k (t)^{k-1} e^{-\mu t}}{(k-1)!}, & \text{si } t \geq 0 \\ 0, & \text{sinon} \end{cases} \tag{5.13}$$

Par ailleurs, une extension de la loi Gamma a été proposée par Fang et Chlamtac (1999) pour modéliser la durée d'occupation des canaux. Il s'agit de la loi *hyper-Erlang* dont la fonction de densité de probabilité est donnée par la relation suivante :

$$f_{T_c}(t) = \begin{cases} \displaystyle\sum_{i=1}^{M} \alpha_i \frac{(m_i \eta_i)^{m_i} (t)^{m_i - 1}}{(m_i - 1)!} e^{-m_i \eta_i t}, & \text{si } t \geq 0 \\ 0, & \text{sinon} \end{cases} \tag{5.14}$$

où $\alpha_i \geq 0$, $\sum_{i=1}^{M} \alpha_i = 1$, alors que M, m_1, m_2, ..., m_M sont des nombres entiers non négatifs et η_1, η_2, ..., η_M sont des nombres positifs. Il a été prouvé que la loi *hyper-Erlang* offre la possibilité de tenir compte de tous les facteurs qui influencent les durées d'occupation des canaux, en plus de préserver les propriétés markoviennes des modèles de files d'attente (Fang et Chlamtac, 1999).

Cette analyse montre combien il est difficile de caractériser la durée d'occupation des canaux T_c par une loi spécifique. Ainsi, pour la suite de notre étude, nous considérons que T_c suit une loi exponentielle sous certaines conditions, ou une loi générale correspondant à l'une des lois présentées dans cette section.

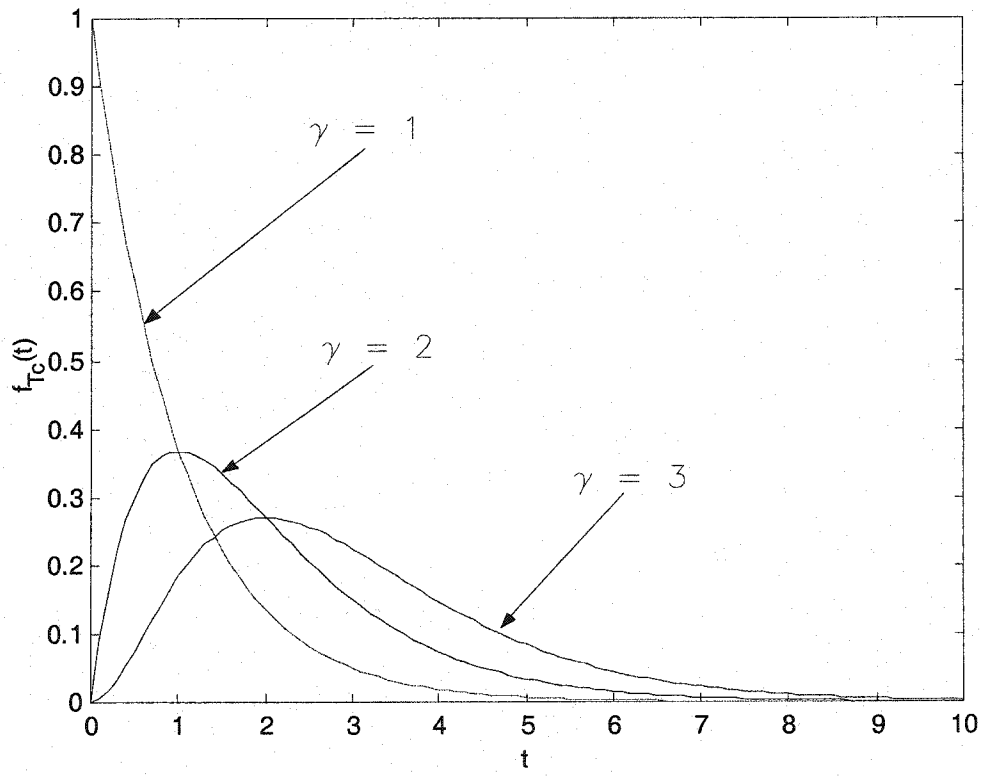


Figure 5.6 Durée d'occupation de loi Gamma avec $\alpha = 1$

5.2 Modèle théorique d'évaluation du trafic

La plupart des travaux d'analyse de trafic répertoriés jusqu'ici ont surtout mis l'accent sur le développement de modèles pour caractériser le trafic d'arrivée et la distribution du temps de service dans les cellules (Jabbari, 1996; Orlik et Rappaport, 1998; Fang et Chlamtac, 1999). Plus spécifiquement, sous certaines hypothèses, Jabbari (1996) utilise les durées d'occupation des canaux et les durées de séjour dans chaque cellule pour estimer à la fois la probabilité de blocage des nouveaux appels et la probabilité d'interruption forcée des communications. Quant à Orlik et Rappaport (1998), ils supposent que les durées des communications, ainsi que les durées de séjour dans chaque cellule suivent une loi générale pour évaluer la probabilité de blocage des appels ainsi que la distribution de la durée d'occupation des canaux. Sous les mêmes hypothèses, Fang et Chlamtac (1999) déterminent la probabilité de compléter une communication avec succès, ainsi que la distribution des durées effectives d'occupation des canaux, en plus d'analyser l'influence de la distribution des durées de séjour sur les durées d'occupation des canaux. Cependant, ces modèles ne permettent pas d'évaluer l'intensité de trafic dans les cellules, en tenant compte du niveau d'interférence générée par les usagers en communication. Nous nous basons sur l'analyse effectuée à la section précédente pour proposer un modèle théorique d'évaluation du trafic dans chaque cellule des systèmes mobiles de la prochaine génération.

Pour évaluer la distribution du nombre d'usagers mobiles en communication, nous modélisons chaque cellule par une file d'attente de type M/M/∞ sous certaines conditions, et plus généralement par une file d'attente de type M/G/∞. Ces modèles supposent que le trafic d'arrivée forme un processus de Poisson, la durée de séjour dans chaque cellule ainsi que la durée des communications suivent des lois exponentielles ou générales, et qu'il existe un nombre arbitrairement élevé de canaux disponibles dans chaque cellule. De tels modèles peuvent être interprétés comme des situations où il existe toujours au moins un canal disponible pour accommoder toute nouvelle connexion ou toute connexion résultant d'une relève. Cela est rendu possible grâce à la capacité élastique de la technologie CDMA. En effet, la technologie CDMA affecte un code

pseudo-orthogonal à toute nouvelle communication, alors que le nombre de codes générés par unité de temps est supérieur au taux de nouvelles connexions (Castro, 2001). Il en résulte que chaque cellule est toujours en mesure d'accepter toute communication qui provient d'un nouvel usager, peu importe le nombre d'usagers déjà en communication dans cette cellule, indépendamment des affectations de temps et de fréquences, moyennant toutefois une légère augmentation du niveau d'interférence sur les autres usagers (Viterbi et Viterbi, 1993). La Figure 5.7 illustre une file d'attente à capacité infinie (de type M/M/∞ ou M/G/∞) qui modélise une cellule, dont les demandes de connexions forment un processus de Poisson de moyenne λ et dont la durée moyenne d'occupation des canaux est égale à $1/\mu$.

Pour évaluer l'intensité de trafic à partir de ce modèle, désignons par $p(n)$ la distribution des terminaux actifs (ou nombre d'usagers mobiles en communication) en régime permanent dans une cellule. On peut montrer que, pour tout modèle M/M/∞ ou M/G/∞, $p(n)$ s'exprime par la relation suivante (Gross et Harris, 1974; Kleinrock, 1975) :

$$p(n) = e^{-\lambda/\mu} \frac{\left(\lambda/\mu\right)^n}{n!} \tag{5.15}$$

où $1/\mu$ représente la durée moyenne d'occupation des canaux, λ le taux moyen d'arrivée du trafic et $p(0)$ la probabilité de n'avoir aucune communication en cours.

Ainsi, la distribution $p(n)$ suit une loi de Poisson dont la moyenne λ/μ correspond au nombre de canaux et de codes affectés aux usagers mobiles en communication. Ce résultat est valide pour tout modèle M/G/∞ (Gross et Harris, 1974; Kleinrock, 1975), c'est-à-dire $p(n)$ dépend seulement de la durée moyenne d'occupation des canaux et non de la forme de sa distribution. Par exemple, si T_c suit une loi lognormale de paramètres m et σ, la moyenne de T_c est donnée par (Crow et Shimizu, 1988) :

$$E(T_c) = e^{m+\sigma^2/2} \tag{5.16}$$

Dans ce cas particulier, la distribution des terminaux actifs est donnée par la relation :

$$p(n) = e^{-\lambda e^{m+\sigma^2/2}} \frac{\left(\lambda e^{m+\sigma^2/2}\right)^n}{n!} \tag{5.17}$$

Il en résulte que le processus de détermination de $p(n)$ reste le même pour n'importe quelle distribution de T_c, pourvu que l'on soit en mesure de déterminer la moyenne de T_c.

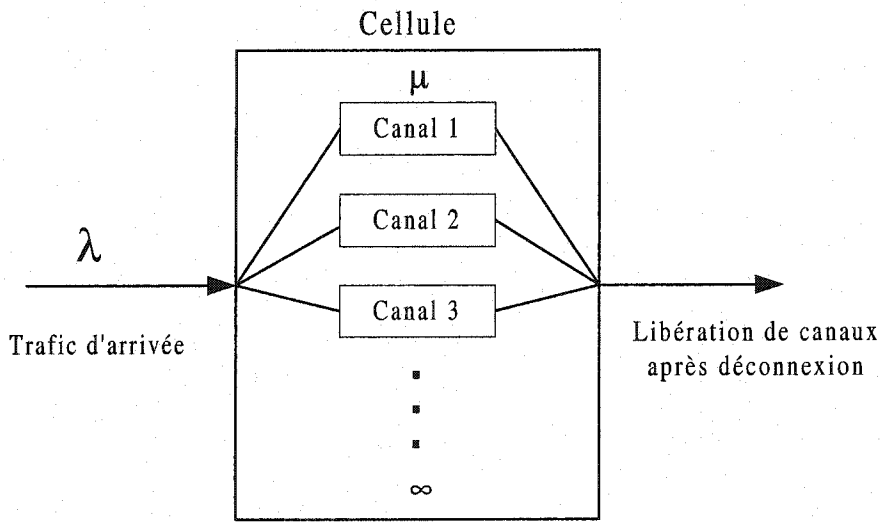


Figure 5.7 Modélisation d'une cellule par une file d'attente à capacité infinie

Toutefois, les modèles à capacité infinie, en l'occurrence M/M/∞ et M/G/∞, demeurent théoriques. En pratique, une cellule a une capacité finie et ne peut en aucun cas offrir une bonne qualité de service tout en accommodant un nombre infini d'usagers mobiles. Cela nous amène à définir la capacité d'une cellule par le nombre maximal d'usagers mobiles que peut supporter cette cellule, tout en maintenant un niveau acceptable de qualité de service en terme du niveau d'interférence. Nous nous basons sur cette définition pour proposer un modèle plus réaliste qui tient compte du nombre maximal d'usagers qui peuvent être desservis dans chaque cellule.

5.3 Proposition d'un modèle général

Dans cette section, nous allons, dans un premier temps, caractériser et justifier les paramètres du modèle proposé, soit le G/G/c/c. L'analyse d'un tel modèle étant complexe, nous allons par la suite appliquer les principes de l'entropie maximale, surtout utilisés en théorie de l'information, pour le résoudre. Nous terminerons alors la section en montrant la généricité de la solution obtenue.

5.3.1 Caractérisation du modèle

En termes de modélisation, nous considérons qu'une cellule peut simultanément supporter un nombre maximal de c communications, tout en maintenant un niveau acceptable de qualité de service. Autrement dit, même si une cellule a une capacité théorique infinie, un algorithme de contrôle d'admission rejète toute nouvelle communication qui y arrive lorsque c communications sont déjà en cours dans cette cellule et ce, pour maintenir la qualité de service des communications en cours. Dans ce contexte, le blocage correspond au refus de service. Par ailleurs, nous prenons pour acquis que le système ne tolère pas d'attente. Autrement dit, un usager mobile qui ne voit pas sa demande de connexion satisfaite doit réitérer sa demande pour l'établissement d'une nouvelle connexion. Une telle analyse a donné lieu au modèle M/G/c/c qui a été proposé dans la littérature et qui permet d'évaluer la distribution du trafic dans chaque cellule de la manière suivante (Viterbi et Viterbi, 1993) :

$$p(n) = \frac{\left(\lambda/\mu\right)^n / n!}{\sum_{k=0}^{c} \left(\lambda/\mu\right)^k / k!} \tag{5.18}$$

où n désigne le nombre de terminaux actifs, λ le taux moyen du trafic d'arrivée et $1/\mu$ la durée moyenne d'occupation des canaux. La relation (5.18) permet alors d'obtenir la probabilité de blocage d'appels de la manière suivante :

$$p_B = \frac{\left(\lambda/\mu\right)^c / c!}{\sum_{k=0}^{c} \left(\lambda/\mu\right)^k / k!} \tag{5.19}$$

où c est le nombre maximal de communications admissibles dans la cellule considérée, λ le taux moyen du trafic d'arrivée et $1/\mu$ la durée moyenne d'occupation des canaux. La relation (5.19), connue sous le nom de formule à perte d'Erlang (ou formule d'Erlang-B), est indépendante de la loi du temps de service (Medhi, 1991).

Cependant, le modèle M/G/c/c ne reflète pas tout à fait la situation réelle. En effet, on peut montrer qu'en général, le trafic résultant des relèves entre cellules adjacentes n'est pas un processus aléatoire sans mémoire et, de ce fait, ne peut être modélisé par un processus de Poisson que dans certaines conditions (Rajaratnam et Takawira, 1997). Plus particulièrement, Chlebus et Ludwin (1995) ont prouvé que le trafic résultant des relèves n'est poissonien que si la probabilité de blocage dans le réseau est nulle. Il en résulte que, pour des modèles à capacité finie, le trafic d'arrivée doit suivre une loi générale. C'est la raison pour laquelle nous proposons de modéliser chaque cellule par un système de file d'attente de type G/G/c/c. Un tel modèle peut être interprété comme une file d'attente ayant un trafic d'arrivée dont le taux moyen est égal à λ et le carré du coefficient de variation est donné par C_a. Dans ce cas, le coefficient de variation est défini par le rapport de l'écart-type à la moyenne des arrivées. Quant à la distribution de service d'un tel modèle, elle est caractérisée par sa moyenne μ et par le carré de son coefficient de variation C_s.

5.3.2 Principes de résolution du modèle

Plusieurs approches ont été proposées pour analyser et résoudre certains systèmes complexes de files d'attente (Hoorn et Seelen, 1986; Simonot, 1998). Cependant, ces approches sont limitées soit à l'analyse de modèles avec des distributions de service

exponentielles comme le G/M/c (Simonot, 1998), soit à celle de modèles à capacité infinie comme le G/M/∞ (Hoorn et Seelen, 1986). Aucune de ces approches n'a cependant entrepris l'analyse des systèmes à capacité finie et dont les lois du trafic d'arrivée et du temps de service sont générales. Nous allons appliquer les principes de l'entropie maximale (Walstra, 1985) pour analyser le comportement d'un tel modèle (c'est-à-dire le G/G/c/c) et évaluer la distribution du trafic, ainsi que la probabilité de blocage dans chaque cellule.

La fonction d'entropie caractérise l'incertitude qui existe avant qu'un système occupe un certain état. Cette fonction atteint son minimum (c'est-à-dire 0) lorsque les résultats d'un événement particulier sont certains, alors qu'elle atteint son maximum lorsque tous les résultats de cet événement sont équiprobables (Walstra, 1985). Dans un contexte d'analyse de files d'attente, les principes de maximisation d'entropie ont été d'abord appliqués pour évaluer la distribution stationnaire d'un système de type M/M/1/c, en faisant l'analogie avec la mécanique statistique (Ferdinand, 1970). De tels principes ont également été appliqués par Shore (1982) pour déterminer les distributions stationnaires et les approximations des files d'attente de types M/G/1 et G/G/1. Par ailleurs, El-Affendi et Kouvatsos (1983) ont appliqué les mêmes principes pour analyser les files d'attente de type G/M/1, alors que Kouvatsos (1986) a réussi à obtenir des approximations de type *exponentiel général* (GE : General Exponential) pour la distribution des files de types G/G/1 et G/G/1/c respectivement. Plus récemment, Miller et Horn (1998) ont utilisé les principes d'entropie maximale pour estimer des densités de probabilité conditionnelles. Mais, de tels principes n'ont jamais été appliqués dans un contexte d'analyse de file d'attente de type G/G/c/c. Nous proposons donc d'utiliser les principes d'entropie maximale pour résoudre un tel modèle.

Étant donné qu'une distribution générale est caractérisée par sa moyenne et son coefficient de variation, nous allons utiliser les notations suivantes pour l'étude du modèle G/G/c/c :

λ : le taux moyen d'arrivée du trafic dans chaque cellule ;

C_a : le carré du coefficient de variation de la distribution du trafic d'arrivée ;

μ : le taux moyen de service, c'est-à-dire le nombre moyen d'usagers desservis dans chaque cellule par unité de temps ;

C_s : le carré du coefficient de variation de la distribution du temps de service ;

$p(n)$: la distribution du nombre d'usagers en communication (ou terminaux actifs) dans chaque cellule, $n = 0, 1, \ldots, c$.

L'objectif est de déterminer $\{p(n), n = 0, 1, \ldots, c\}$ en fonction des paramètres λ, C_a, μ, C_s et c. Un tel objectif peut être atteint en maximisant la fonction suivante d'entropie (Walstra, 1985) :

$$H(p) = -\sum_{n=0}^{c} p(n) \ln [p(n)] \tag{5.20}$$

sujette à une série de contraintes linéaires qui dépendent de la nature du problème. Dans le contexte d'un système de file d'attente de type G/G/c/c, de telles contraintes peuvent être formulées de la manière suivante :

- la normalisation, c'est-à-dire :

$$\sum_{n=0}^{c} p(n) = 1 \tag{5.21}$$

- l'utilisation $u(l)$, c'est-à-dire :

$$\sum_{n \geq l} p(n) = u(l),\ l = 1, \ldots, c \tag{5.22}$$

où $u(l) \in (0, 1)$

- la longueur moyenne des files (c'est-à-dire le nombre moyen de communications en attente de la disponibilité d'un canal) qui dépend d'une fonction $f(n)$ à évaluer. Dans le cas d'un système sans attente, la longueur moyenne des files s'exprime par la relation suivante :

$$\sum_{n=0}^{c} f(n)\, p(n) \;=\; 0 \tag{5.23}$$

- la probabilité de blocage $p(c)$, c'est-à-dire :

$$\sum_{n=0}^{c} g(n)\, p(n) \;=\; \alpha \tag{5.24}$$

où $\alpha \in (0, 1)$ et $g(n) = \max(0, n\text{-}c+1)$, c'est-à-dire :

$$g(n) \;=\; \begin{cases} 0, & \text{si } n = 0, 1, 2, \cdots, c-1 \\ 1, & \text{si } n = c \end{cases} \tag{5.25}$$

La maximisation de $H(p)$, sujette aux contraintes (5.21) à (5.24), peut être alors obtenue par l'utilisation de la méthode des multiplicateurs de Lagrange (Kouvatsos et Xenios, 1989). Cela conduit à une solution de la forme :

$$p(n) \;=\; K\left[\prod_{l=1}^{n} h(l)\right] x^{f(n)}\, y^{g(n)}, \qquad n = 1, \ldots, c \tag{5.26}$$

où :

- $K = p(0)$ est la constante de normalisation ;
- $h(l) = \exp(-\beta_{1l})$, $l = 1, .., c$;
- $x = \exp(-\beta_2)$;
- $y = \exp(-\beta_3)$.

Dans ce cas, les paramètres β_{1l} (avec $l = 1, .., c$), β_2 et β_3 représentent les multiplicateurs de Lagrange qui correspondent aux contraintes (5.21) à (5.24). De plus, la contrainte (5.23) permet d'obtenir directement $f(n) = 0$, pour $n = 0, 1, \ldots, c$. Il en résulte une nouvelle forme plus simple de la relation (5.26) qui devient :

$$p(n) \;=\; K\left[\prod_{l=1}^{n} h(l)\right] y^{g(n)}, \qquad n = 1, \ldots, c \tag{5.27}$$

5.3.3 Évaluation de la distribution du trafic et de la probabilité de blocage

La relation (5.27) caractérise l'intensité de trafic dans chaque cellule modélisée par une file d'attente de type G/G/c/c. Nous nous proposons maintenant d'évaluer les diverses inconnues de cette relation. Dans cette optique, la méthode d'entropie maximale peut être implémentée en utilisant la distribution exponentielle générale (ou GE) pour caractériser les distributions générales (Kouvatsos et Xenios, 1989). Il convient alors de caractériser la cumulative de la distribution GE dont la forme pour une variable aléatoire T est la suivante (Kouvatsos, 1986) :

$$F_T(t) = 1-\frac{2}{C+1}\exp\left(-\frac{2vt}{C+1}\right), \qquad t \geq 0 \tag{5.28}$$

où $1/v$ est la moyenne de T, alors que C est le carré de son coefficient de variation. Il a été prouvé que, dans un contexte d'analyse de files d'attente, la loi GE est suffisamment robuste pour remplacer toute distribution générale dont les coefficients de variation du trafic d'arrivée et du temps de service sont supérieurs à 1 (Kouvatsos, 1986). Il en résulte que, dans ces conditions, tout modèle G/G/c/c devient analytiquement équivalent à GE/GE/c/c.

La distribution du trafic dans un système de file d'attente de type GE/GE/c/c, sujette aux contraintes (5.21) à (5.24), est alors déduite de (5.27) et exprimée de la manière suivante :

$$p(n) = p(0)\, H_n\, y^{g(n)}, \qquad n = 1, \ldots, c \tag{5.29}$$

avec

$$H_n = \prod_{l=1}^{n} h(l)$$

La détermination de $p(0)$ fait appel à la contrainte (5.21) que nous récrivons de la manière suivante :

$$p(0) + \sum_{n=1}^{c} p(n) = 1 \tag{5.30}$$

En combinant les relations (5.29) et (5.30), nous obtenons :

$$p(0) + \sum_{n=1}^{c} p(0)\, H_n\, y^{g(n)} = 1$$

$$p(0)\left[1 + \sum_{n=1}^{c} H_n\, y^{g(n)}\right] = 1 \qquad (5.31)$$

En combinant les relations (5.25) et (5.31), nous obtenons :

$$p(0) = \left[1 + \sum_{i=1}^{c-1} H_i + y\, H_c\right]^{-1} \qquad (5.32)$$

où :

$$H_i = \prod_{l=1}^{i} h(l) = h(1) * h(2) * \ldots * h(i) \qquad (5.33)$$

De plus, à partir des relations (5.25), (5.29), (5.32) et (5.33), nous pouvons établir les équations de récurrence suivantes :

$$p(1) = p(0) * h(1) \qquad (5.34)$$

$$p(n) = p(n\text{-}1) * h(n), \qquad n = 2, 3, \ldots, c\text{-}1 \qquad (5.35)$$

$$p(c) = y\, h(c) * p(c\text{-}1) \qquad (5.36)$$

De telles équations permettent d'étudier des systèmes de files d'attente générales pour lesquelles les méthodes classiques de résolution ne peuvent être efficacement appliquées. Toutefois, à partir de ces équations de récurrence, nous nous rendons compte que la solution à la relation (5.29) fait essentiellement appel à l'évaluation des multiplicateurs de Lagrange $h(l)$, $l = 1, 2, \ldots, c$, et y. Pour déterminer ces multiplicateurs, définissons les paramètres suivants :

$$a = \frac{2}{C_a + 1} \qquad (5.37)$$

$$b = \frac{2}{C_s + 1} \tag{5.38}$$

$$\rho = \lambda / c\mu \tag{5.39}$$

où λ et C_a désignent respectivement la moyenne et le carré du coefficient de variation du trafic d'arrivée, alors que μ et C_s sont respectivement la moyenne et le carré du coefficient de variation de la distribution de service. À ce niveau, nous pouvons appliquer la procédure présentée par Kesavan et Kapur (1989) pour déterminer $\{h(l),\ l = 1, 2, \ldots, c\}$ et y en fonction de a, b, ρ et c. Nous obtenons alors :

$$h(1) = \begin{cases} \dfrac{ac\rho}{b(1-a)+a}, & \text{si } c > 1 \\ \dfrac{a\rho b}{a\rho(1-b)+b}, & \text{si } c = 1 \end{cases} \tag{5.40}$$

$$h(l) = \begin{cases} \dfrac{ac\rho + (l-1)b(1-a)}{l[b(1-a)+a]}, & l = 2, \ldots, c-1 \\ \dfrac{b[ac\rho + (c-1)b(1-a)]}{c[a\rho(1-b)+b]}, & l = c \end{cases} \tag{5.41}$$

$$y = \frac{1}{1-(1-b)z} \tag{5.42}$$

avec

$$z = \frac{a\rho + b(1-a)}{a\rho(1-b)+b} \tag{5.43}$$

Ainsi, les relations (5.40) à (5.43) permettent de déterminer respectivement H_n (n = 1, 2, ..., c) et $p(0)$. La distribution du nombre de terminaux actifs $p(n)$ s'obtient alors en remplaçant $p(0)$, H_n, y et $g(n)$ par leur valeur dans la relation (5.29). Il convient toutefois de mentionner que cette solution est obtenue en supposant que le trafic d'arrivée, ainsi que la durée d'occupation des canaux suivent une loi de type GE. Une telle solution est donc exacte lorsque les paramètres C_a et C_s sont supérieurs à 1. Toutefois, si $C_a < 1$ ou $C_s < 1$, la solution conduira à une approximation de la distribution du trafic.

Par ailleurs, la probabilité de blocage p_B peut être obtenue en posant $n = c$ dans la relation (5.29). Cela conduit à la relation suivante :

$$p_B = p(c) = p(0)\, H_c\, y \tag{5.44}$$

où $p(0)$ et y sont respectivement donnés par (5.32) et (5.42), alors que H_c est donné par la relation :

$$H_c = h(1) * h(2) * \ldots * h(c-1) * h(c) \tag{5.45}$$

Mentionnons que $\{h(l)$, avec $l = 1, 2, \ldots, c\}$, z et y demeurent les mêmes que dans la relations (5.40) à (5.43).

5.3.4 Résolution du modèle M/M/c/c à partir du modèle présenté

Pour montrer le caractère générique de notre solution, nous allons résoudre les modèles de files de type M/M/c/c à partir de la relation (5.29). En effet, pour les modèles M/M/c/c, le temps d'interarrivées T_a suit une loi exponentielle de paramètre λ_a (> 0) et dont la fonction de densité de probabilité est donnée par la relation suivante :

$$f_{T_a}(t) = \begin{cases} \lambda_a e^{-\lambda_a t}, & \text{si } t \geq 0 \\ 0, & \text{sinon} \end{cases} \tag{5.46}$$

La moyenne et la variance de T_a sont respectivement données par :

$$E(T_a) = 1/\lambda_a \tag{5.47}$$

$$V(T_a) = \sigma_a^2 = 1/\lambda_a^2 \tag{5.48}$$

tandis que son coefficient de variation est donné par :

$$Cv_a = \sigma_a / E(T_a) = 1$$

Il en résulte que le carré du coefficient de variation des interarrivées C_a est donné par :

$$C_a = Cv_a^2 = 1 \tag{5.49}$$

Dans le même ordre d'idées, puisque le temps de service est aussi exponentiellement distribué, le carré de son coefficient de variation C_s est donné par :

$$C_s = Cv_s^2 = 1 \tag{5.50}$$

En remplaçant C_a et C_s par leur valeur dans les relations (5.37) à (5.43) et en utilisant l'indice p pour signifier qu'il s'agit de valeurs particulières propres au modèle M/M/c/c, nous obtenons :

$$a_p = 1 \tag{5.51}$$

$$b_p = 1 \tag{5.52}$$

$$h_p(l) = \left(\frac{c\rho}{l}\right), \quad l = 1, 2, \cdots, c \tag{5.53}$$

$$z_p = \rho \tag{5.54}$$

$$y_p = 1 \tag{5.55}$$

Dans ce cas particulier, $p(n)$ devient :

$$p_p(n) = p_p(0)\, H_{p_n} \tag{5.56}$$

ce qui est une forme simplifiée de la relation (5.29).

Pour trouver $p_p(n)$, nous allons déterminer dans un premier temps H_{p_n}, puis $p_p(0)$. Le paramètre H_{p_n} peut être obtenu en combinant les relations (5.33) et (5.53) de la manière suivante :

$$H_{p_n} = \prod_{l=1}^{n}\left(\frac{c\,\rho}{l}\right)$$

$$H_{p_n} = (c\,\rho)*\left(\frac{c\,\rho}{2}\right)*\ldots*\left(\frac{c\,\rho}{n}\right)$$

$$H_{p_n} = \frac{(c\,\rho)^n}{n!} \tag{5.57}$$

Dans le même ordre d'idées, en combinant (5.32), (5.33), (5.53) et (5.55), nous pouvons déterminer $p_p(0)$ de la manière suivante :

$$p_p(0) = \left[1 + \sum_{n=1}^{c-1} \frac{(c\,\rho)^n}{n!} + \frac{(c\,\rho)^c}{c!}\right]^{-1}$$

$$p_p(0) = \left[1 + \sum_{n=1}^{c} \frac{(c\,\rho)^n}{n!}\right]^{-1}$$

En posant $\rho = \lambda/c\mu$, comme dans la relation (5.39), nous obtenons :

$$p_p(0) = \left[1 + \sum_{n=1}^{c} \frac{\left(\lambda/\mu\right)^n}{n!}\right]^{-1} \tag{5.58}$$

En combinant les relations (5.56), (5.57) et (5.58), nous obtenons :

$$p_p(n) = \frac{\left(\lambda/\mu\right)^n / n!}{1 + \lambda/\mu + \frac{\left(\lambda/\mu\right)^2}{2!} + \cdots + \frac{\left(\lambda/\mu\right)^n}{n!}} \tag{5.59}$$

La relation (5.59) caractérise la distribution des terminaux actifs lorsque l'on considère un trafic d'arrivée poissonien et un temps de service exponentiel. Une telle distribution est équivalente à la relation (5.18), ce qui signifie que (5.18) est un cas particulier de la formule générale exprimée par la relation (5.29) et déduite de l'analyse du modèle G/G/c/c.

Dans le même ordre d'idées, nous pouvons utiliser la forme générale de la probabilité de blocage exprimée par la relation (5.44) pour évaluer directement la probabilité de blocage dans un système de type M/M/c/c. Nous obtenons alors :

$$p_{p_B} = p_p(0)\, H_{p_c}\, y_p \tag{5.60}$$

où les paramètres y_p et $p_p(0)$ sont respectivement donnés par (5.55) et (5.58), tandis que H_{p_c} est donné par :

$$H_{p_c} = \frac{(c\,\rho)^c}{c!} \tag{5.61}$$

En combinant (5.55), (5.58), (5.60) et (5.61), nous obtenons :

$$p_{p_B} = \frac{\left(\lambda/\mu\right)^c \Big/ c!}{1 + \lambda/\mu + \frac{\left(\lambda/\mu\right)^2}{2!} + \cdots + \frac{\left(\lambda/\mu\right)^c}{c!}} \tag{5.62}$$

Ainsi, la relation (5.62) obtenue à partir de notre analyse du modèle G/G/c/c est équivalente à la relation (5.19) qui représente la formule d'Erlang-B. Il en résulte que cette formule d'Erlang-B constitue un cas particulier de notre formule générale exprimée par la relation (5.44).

5.4 Résultats obtenus et analyse

Dans cette section, nous allons utiliser les deux modèles pour déterminer la distribution du nombre d'usagers en communication, en plus d'évaluer, à partir du second modèle, la probabilité de blocage en fonction de l'intensité de trafic. Plus spécifiquement, nous allons faire une analyse de résultats en trois étapes : l'étude de l'influence de la taille des cellules sur l'intensité de trafic, celle de l'influence des coefficients de variation sur la distribution du trafic et celle de l'influence des coefficients de variation sur la probabilité de blocage.

5.4.1 Influence de la taille des cellules sur l'intensité de trafic

Cette analyse se fait avec le modèle à capacité infinie (M/M/∞ ou M/G/∞) dont la distribution du nombre de terminaux actifs est exprimée par la relation (5.15). Étant donné que l'évaluation d'une telle distribution nécessite la détermination du trafic moyen d'arrivée et celle de la durée moyenne des communications dans chaque cellule, nous allons faire quelques considérations particulières pour justifier le choix de ces paramètres.

Les pico-cellules étant des zones à forte concentration d'abonnés, nous avons choisi de leur affecter le taux moyen d'arrivée de trafic le plus élevé, soit 1.15 appel par seconde. Dans le même ordre d'idées, nous avons décidé d'affecter le taux le plus faible aux macro-cellules, car ces dernières sont en général des zones de service peu denses ayant toutefois une grande superficie, ce qui justifie un trafic moyen d'arrivée de 1.0 appel par seconde. Pour les micro-cellules qui, en général, ont un trafic d'arrivée compris entre celui des pico-cellules et celui des macro-cellules (Tabbane, 2000), nous avons choisi un trafic moyen d'arrivée de 1.1 appel par seconde.

D'autre part, pour les macro-cellules, le temps moyen d'occupation des canaux $1/\mu$ correspond au temps moyen de séjour des automobiles qui voyagent dans des cellules d'environ 1 km de rayon, soit environ 50.0 secondes (Orlik et Rappaport, 1998). Cependant, pour les pico-cellules qui sont généralement utilisés par les piétons, le paramètre $1/\mu$ représente la durée moyenne des appels, soit environ 100.0 secondes (Wey *et al.*, 1997). Autrement dit, pour ce type de cellule, les communications commencent et s'achèvent généralement à l'intérieur de la même cellule, ce qui s'explique par la faible vitesse des usagers qui y circulent. Toutefois, pour les micro-cellules, le temps d'occupation des canaux peut être, soit le temps de séjour dans la cellule, soit la durée de l'appel, dépendant de la vitesse et de la direction des abonnés. Les durées moyennes d'occupation des canaux, ainsi que les taux moyens du trafic d'arrivée utilisés dans cette analyse, sont résumés au Tableau 5.1 pour les trois types de cellules.

Tableau 5.1 Paramètres d'analyse de trafic

	Macro-cellule	Micro-cellule	Pico-cellule
λ (appels/sec)	1.0	1.1	1.15
$1/\mu$ (sec)	50.0	60.0	100.0

Ces paramètres permettent de représenter et de comparer la distribution $p(n)$ des terminaux actifs dans les macro-cellules, ainsi que dans les micro-cellules et les pico-cellules. La Figure 5.8 illustre une telle comparaison pour les valeurs spécifiées de λ et de μ du Tableau 5.1, en mettant en évidence la sensibilité de $p(n)$ par rapport à la catégorie des cellules. Nous en déduisons que, pour les macro-cellules, l'intensité moyenne du trafic est plus faible que pour les micro-cellules ou les pico-cellules qui, quant à elles, sont généralement considérées comme des zones plus denses. Il en résulte que le risque d'interférence est plus élevé dans les pico-cellules que dans les micro-cellules ou les macro-cellules.

5.4.2 Influence des coefficients de variation sur la distribution du trafic

Pour l'analyse de l'impact des coefficients de variation des distributions d'arrivée et de service C_a et C_s sur la distribution du trafic $p(n)$, nous reprenons les valeurs utilisées par Kouvatsos (1986), ainsi que par Orlik et Rappaport (1998), c'est-à-dire une cellule de c = 100 canaux, un taux moyen d'arrivée de λ = 0.75 appels/sec et un temps moyen d'occupation des canaux de $1/\mu$ = 100 s. Cette analyse comporte trois parties : d'abord C_a constant et C_s variable, ensuite C_a variable et C_s constant, à la fois C_a et C_s variable. Pour la première partie, nous suivons le même principe que Kouvatsos (1986), en fixant C_a à 1.5 et en affectant à C_s les valeurs suivantes : 1.0, 2.0, 3.0, 5.0 et 10.0. Ainsi, pour les valeurs fixées de λ, μ, C_a et une intensité de trafic $\rho = \lambda/c\mu = 0.75$, nous caractérisons l'évolution de la distribution du trafic $\{p(n), n = 1, 2, \ldots c\}$ pour chaque valeur de C_s, ce

qui est illustré à la Figure 5.9. Nous trouvons que, pour $66 \leq n \leq 85$, $p(n)$ augmente en fonction de C_s, alors que pour $n \leq 66$ ou $85 \leq n \leq 100$, $p(n)$ décroît lorsque C_s augmente. Autrement dit, pour $85 \leq n \leq 100$, le degré de certitude sur le nombre d'usagers en communication augmente en fonction de C_s.

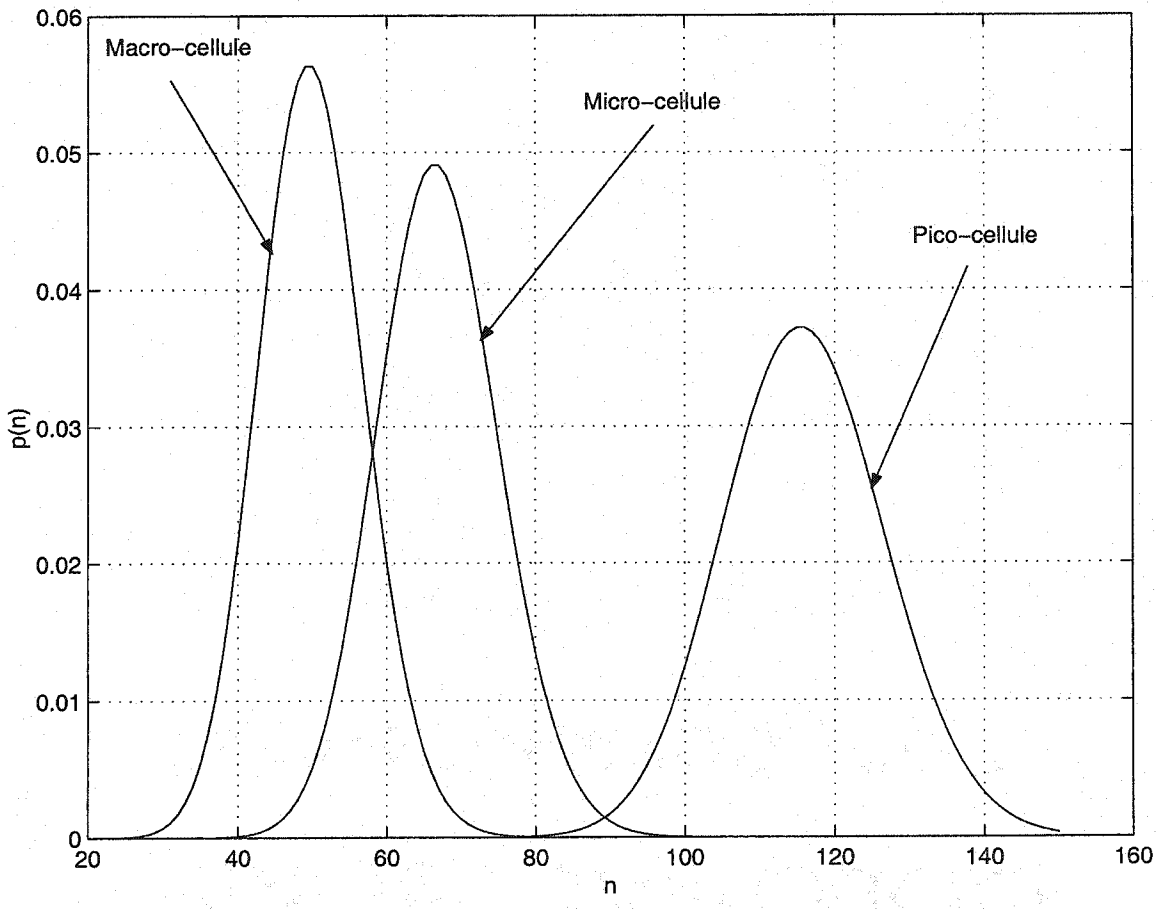


Figure 5.8 Distribution des terminaux actifs pour plusieurs types de cellules

De plus, la valeur maximale de $p(n)$, soit la probabilité de trafic la plus élevée, est atteinte pour $n = 74$ appels/sec quelle que soit la valeur de C_s. Il convient également de mentionner que, pour $n = c = 100$, $p(n)$ est supérieur à 0, c'est-à-dire la probabilité de blocage p_B est non nulle pour $C_s = 1.0$, 2.0, 3.0, 5.0 et 10.0. Le Tableau 5.2 donne la

valeur maximale de $p(n)$, ainsi que la probabilité de blocage pour les valeurs choisies de C_s. Nous nous rendons compte que $p_{max}(n)$ augmente, alors que p_B décroît lorsque nous faisons augmenter C_s. Toutefois, il faudrait noter que la variabilité de $p_{max}(n)$ et de p_B demeure relativement faible lorsque C_s varie. D'ailleurs, nous avons vérifié que, lorsque $C_a = 1$, $p(n)$ devient insensible à la valeur de C_s (≥ 1). Autrement dit, lorsque la distribution du temps d'interarrivées est exponentielle et que le temps de service est général, $p(n)$ demeure invariant pour tout $C_s \geq 1$. Ainsi, les résultats obtenus permettent de valider que la distribution du trafic est insensible au temps de service de tout modèle de file d'attente de type M/G/c/c.

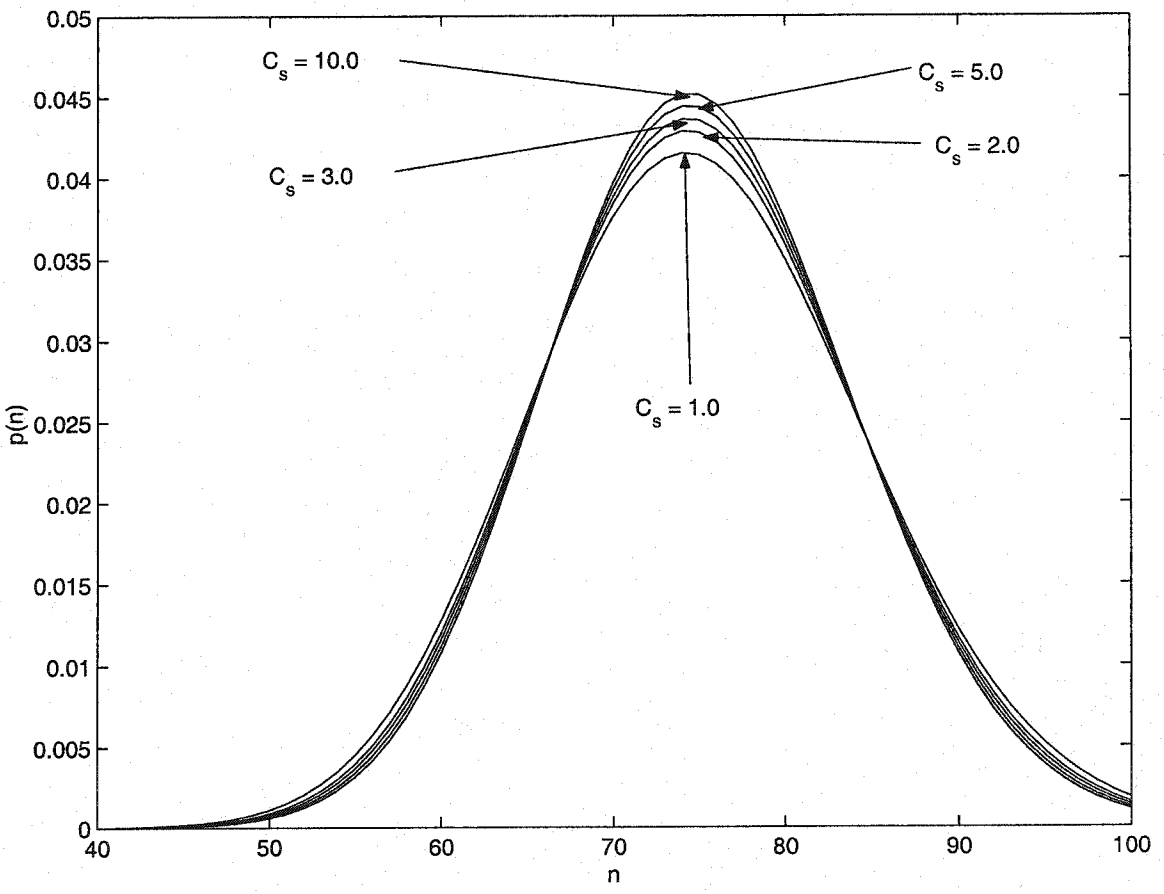


Figure 5.9 Distribution des terminaux actifs avec $C_a = 1.5$

Tableau 5.2 Valeur maximale de $p(n)$ et de p_B en fonction de C_s

C_s	1.0	2.0	3.0	5.0	10.0
$p_{max}(n)$	0.0415	0.0429	0.0436	0.0444	0.0452
p_B	0.0018	0.0015	0.0014	0.0012	0.0011

Pour la seconde partie de cette analyse, nous fixons C_s à 1.5 et donnons à C_a les valeurs suivantes : 1.0, 2.0, 3.0, 5.0 et 10.0, comme l'a fait Kouvatsos (1986) pour l'analyse de son modèle. En utilisant les mêmes valeurs que précédemment pour λ, μ, C_a et ρ, nous traçons la courbe caractéristique de la distribution du nombre d'usagers mobiles pour chaque valeur de C_a, ce qui est illustré à la Figure 5.10. Nous pouvons alors observer une variabilité importante de la distribution du trafic lorsque C_a varie. En outre, nous trouvons que, pour $66 \leq n \leq 85$, $p_{max}(n)$ augmente lorsque C_a décroît, alors que pour $n \leq 62$ ou $91 \leq n \leq 100$, $p_{max}(n)$ augmente en fonction de C_a. Dans le même ordre d'idées, la valeur maximale de $p(n)$ est atteinte à $n = 74$ appels/sec pour $C_a = 1.0$, 2.0 et 3.0, alors qu'elle est atteinte à $n = 73$ appels/sec pour $C_a = 5.0$ et à $n = 71$ appels/sec pour $C_a = 10.0$. Le Tableau 5.3 récapitule les valeurs maximales de $p(n)$, ainsi que les probabilités de blocage relatives aux valeurs sélectionnées de C_a. Nous en déduisons que, lorsque C_a augmente, $p_{max}(n)$ décroît, alors que p_B augmente.

Les résultats précédents montrent que $p_{max}(n)$ et p_B ont des comportements opposés, dépendant duquel des paramètres C_a ou C_s changent. Nous allons faire varier à la fois C_a et C_s au même taux et analyser l'effet d'une telle variation sur la distribution du trafic. Dans cette optique, nous allons donner des valeurs spécifiques à $C_a = C_s$ et caractériser l'évolution de $p(n)$ pour chaque $C_a = C_s$. La Figure 5.11 illustre la distribution du trafic $p(n)$ pour $C_a = C_s = 1.0$, 2.0, 3.0, 5.0 et 10.0. Nous trouvons que, pour les mêmes valeurs de λ, μ, ρ, et pour $66 \leq n \leq 85$, $p(n)$ augmente lorsque $C_a = C_s$ décroît, alors que pour $n \leq 65$ ou $87 \leq n \leq 100$, $p(n)$ augmente en fonction de $C_a = C_s$. De plus, la valeur maximale de $p(n)$ est atteinte à $n = 74$ appels/sec pour les valeurs

sélectionnées de $C_a = C_s$. Le Tableau 5.4 donne les valeurs maximales de $p(n)$, ainsi que les probabilités de blocage pour les valeurs choisies de $C_a = C_s$. Nous nous rendons compte que $p_{max}(n)$ et p_B ont le même comportement que dans le cas où C_s est constant et C_a est variable. Ainsi, pour les paramètres utilisés dans notre analyse, le coefficient de variation de la distribution du trafic d'arrivée a plus d'influence sur la distribution du trafic que celui de la distribution du temps d'occupation des canaux.

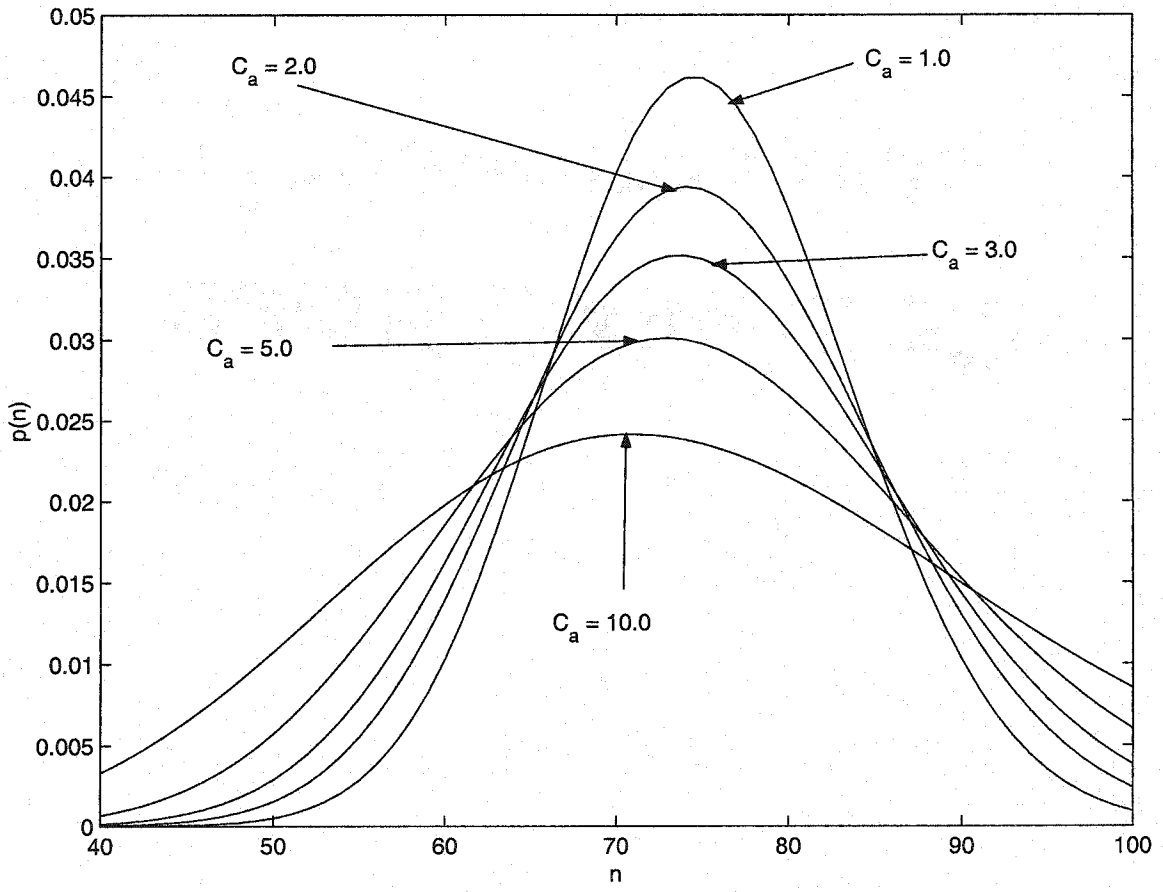


Figure 5.10 Distribution des terminaux actifs pour $C_s = 1.5$

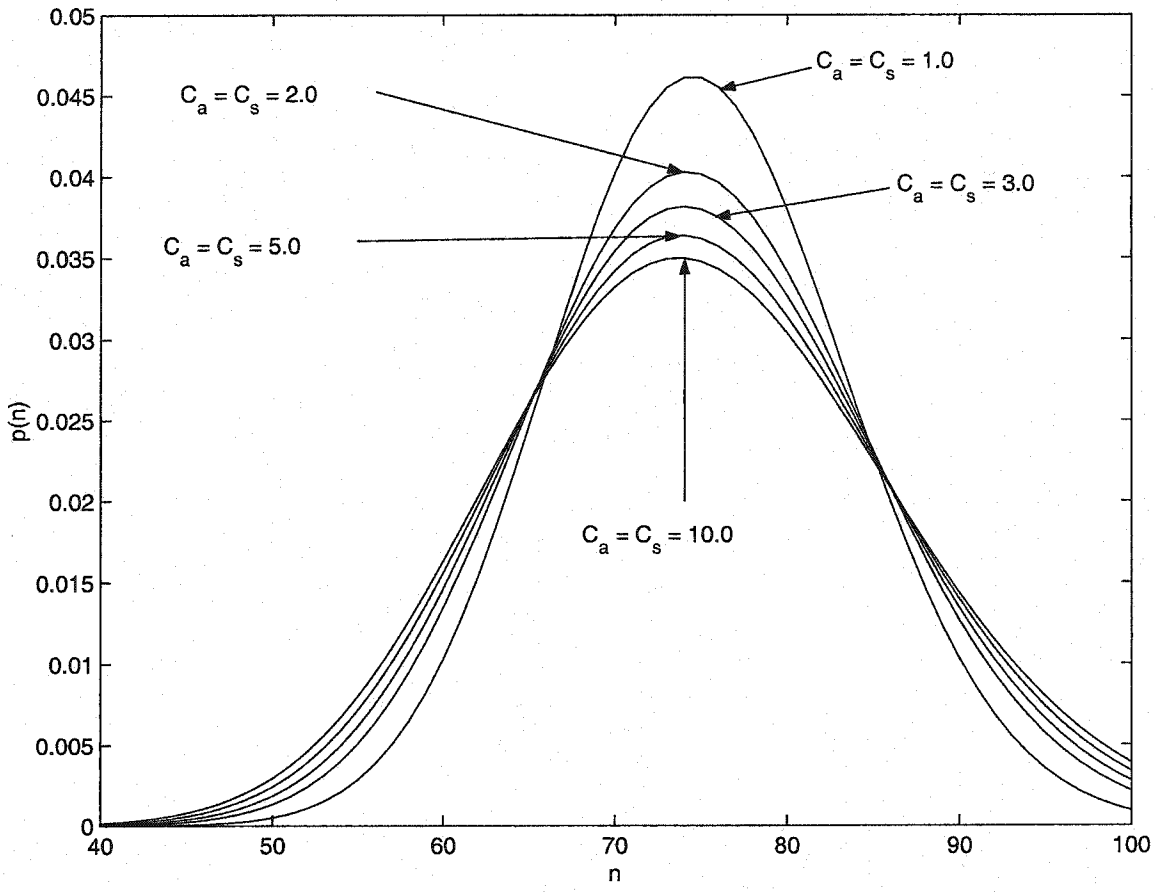


Figure 5.11 Distribution des terminaux actifs avec C_a et C_s variables

Tableau 5.3 Valeur maximale de $p(n)$ et de p_B en fonction de C_a

C_a	1.0	2.0	3.0	5.0	10.0
$p_{max}(n)$	0.0461	0.0394	0.0352	0.0301	0.0242
p_B	0.0009	0.0024	0.0038	0.0060	0.0085

Tableau 5.4 Valeur maximale de $p(n)$ et de p_B en fonction de $C_a = C_s$

$C_a = C_s$	1.0	2.0	3.0	5.0	10.0
$p_{max}(n)$	0.0461	0.0403	0.0382	0.0364	0.0350
p_B	0.0009	0.0021	0.0028	0.0034	0.0039

5.4.3 Influence de C_a et de C_s sur la probabilité de blocage

Les tableaux 5.2 à 5.4 ont montré l'évolution de la probabilité de blocage p_B pour une intensité de trafic bien spécifique ρ de 0.75. Nous allons généraliser ces résultats en évaluant p_B en fonction de ρ, avec $0 \leq \rho \leq 1$, pour différentes valeurs des coefficients de variation des distributions d'arrivée et de service. Cela permettra de déterminer, en fonction de C_a et de C_s, les conditions selon lesquelles le système offre de faibles probabilités de blocage, c'est-à-dire une meilleure qualité de service. Comme dans la section précédente, nous divisons cette analyse en trois parties : C_a constant et C_s variable, C_a variable et C_s constant, à la fois C_a et C_s variables.

Pour la première partie de l'analyse, nous suivons à nouveau le même principe que Kouvatsos (1986), en fixant C_a à 1.5 et en affectant à C_s les valeurs suivantes : 1.0, 2.0, 3.0, 5.0 et 10.0. Pour chacune des valeurs de C_s, nous déterminons la courbe caractéristique de la probabilité de blocage p_B en fonction de l'intensité de trafic ρ, ce qui est illustré à la Figure 5.12 en prenant une échelle logarithmique pour l'axe des ordonnées. Nous observons de faibles changements dans les mesures de performance (c'est-à-dire p_B) lorsque le paramètre C_s change. En outre, nous trouvons que, pour $\rho \leq 0.86$, p_B augmente lorsque C_s décroît, alors que pour $0.90 \leq \rho \leq 1.0$, p_B augmente en fonction de C_s.

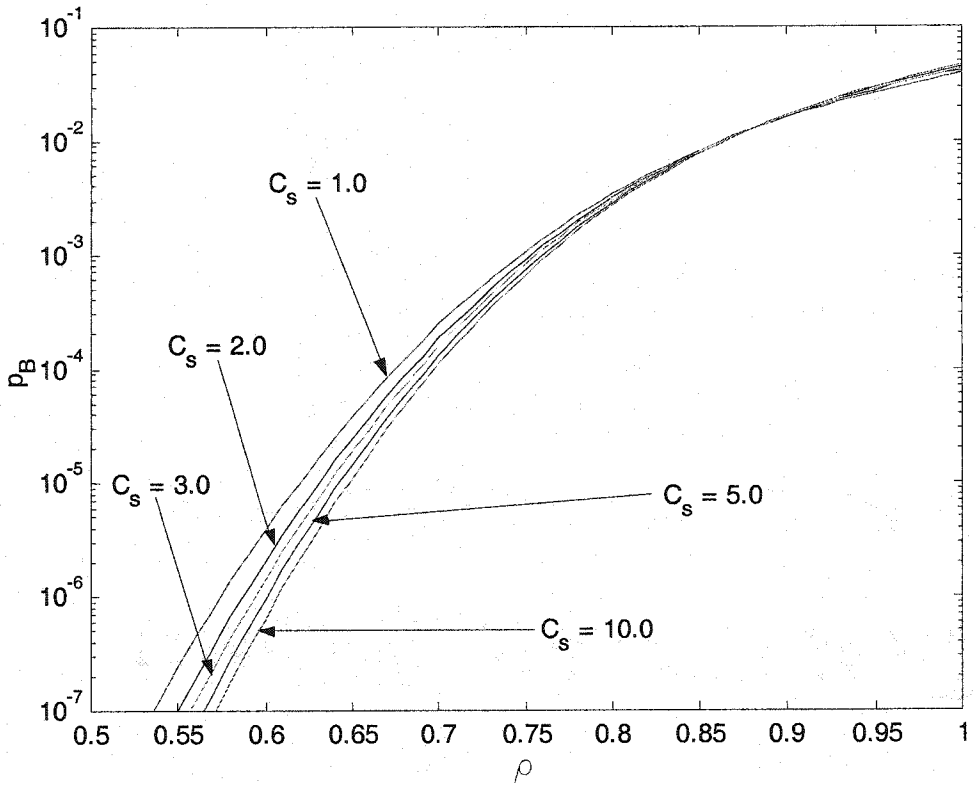


Figure 5.12 Évolution de la probabilité de blocage avec C_a = 1.5

En considérant qu'une faible intensité de trafic favorise la qualité de service (c'est-à-dire offre une probabilité de blocage plus faible), nous avons réalisé que ρ doit être maintenue inférieure à 0.87 pour avoir une probabilité de blocage inférieure à 0.01 ($p_B \leq 0.01$). Dans cette circonstance, une meilleure qualité de service peut être attendue pour des valeurs élevées de C_s. De plus, à partir de la relation (5.44), nous avons vérifié que, lorsque C_a = 1 (trafic d'arrivée poissonien), p_B reste invariante pour tout $C_s \geq 1$, c'est-à-dire la probabilité de blocage est insensible à la distribution du temps de service dans un modèle de file d'attente de type M/G/c/c.

Pour la deuxième partie de cette analyse, nous fixons C_s à 1.5 et donnons à C_a les valeurs suivantes : 1.0, 2.0, 3.0, 5.0 et 10.0 (Kouvatsos, 1986). À partir de ces paramètres,

nous caractérisons l'évolution de la probabilité de blocage p_B en fonction de l'intensité de trafic ρ, avec $0 \leq \rho \leq 1.0$. La Figure 5.13 illustre le comportement de p_B pour les valeurs sélectionnées de C_a. Nous nous rendons compte que p_B est plus sensible aux variations de C_a qu'à celles de C_s (situation précédente). En outre, pour $\rho \leq 0.77$, p_B augmente en fonction de C_a, alors que pour $0.87 \leq \rho \leq 1.0$, p_B augmente lorsque C_a décroît. Il en résulte qu'une meilleure qualité de service (faible p_B) est offerte dans la cellule pour des valeurs faibles de C_a.

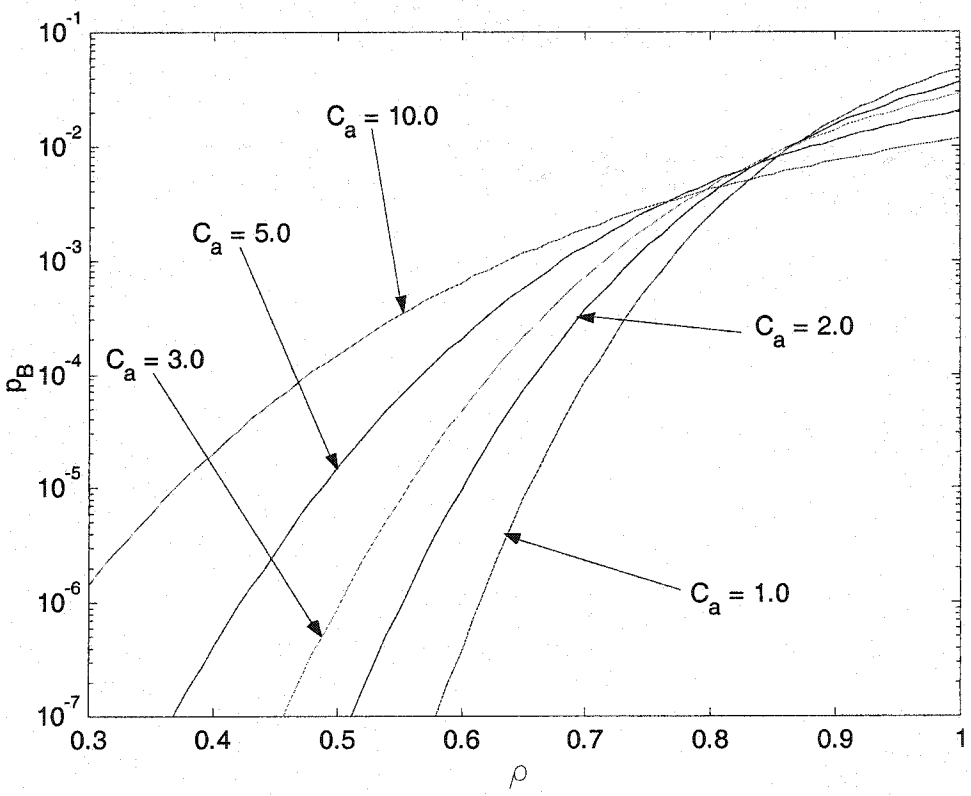


Figure 5.13 Évolution de la probabilité de blocage avec C_s = 1.5

L'analyse précédente a montré que C_a et C_s ont des effets opposés sur la probabilité de blocage p_B. Nous allons faire varier simultanément ces deux paramètres

(C_a et C_s) au même taux et analyser l'effet d'une telle variation sur p_B. Nous déterminons alors, en utilisant une échelle logarithmique, l'évolution de p_B en fonction de ρ pour chaque valeur de $C_a = C_s$ = 1.0, 2.0, 3.0, 5.0 et 10.0, ce qui est illustré à la Figure 5.14. Nous trouvons que, pour $\rho \leq 0.86$, p_B augmente en fonction de $C_a = C_s$, alors que pour $0.88 \leq \rho \leq 1.0$, p_B augmente lorsque $C_a = C_s$ diminue. Ainsi, une meilleure qualité de service est offerte dans chaque cellule pour des valeurs faibles de $C_a = C_s$. Il en résulte que p_B a le même comportement que dans la situation où C_s est constant et C_a variable (situation précédente). Nous concluons que le coefficient de variation du trafic d'arrivée a plus d'impact sur p_B que celui du temps d'occupation des canaux.

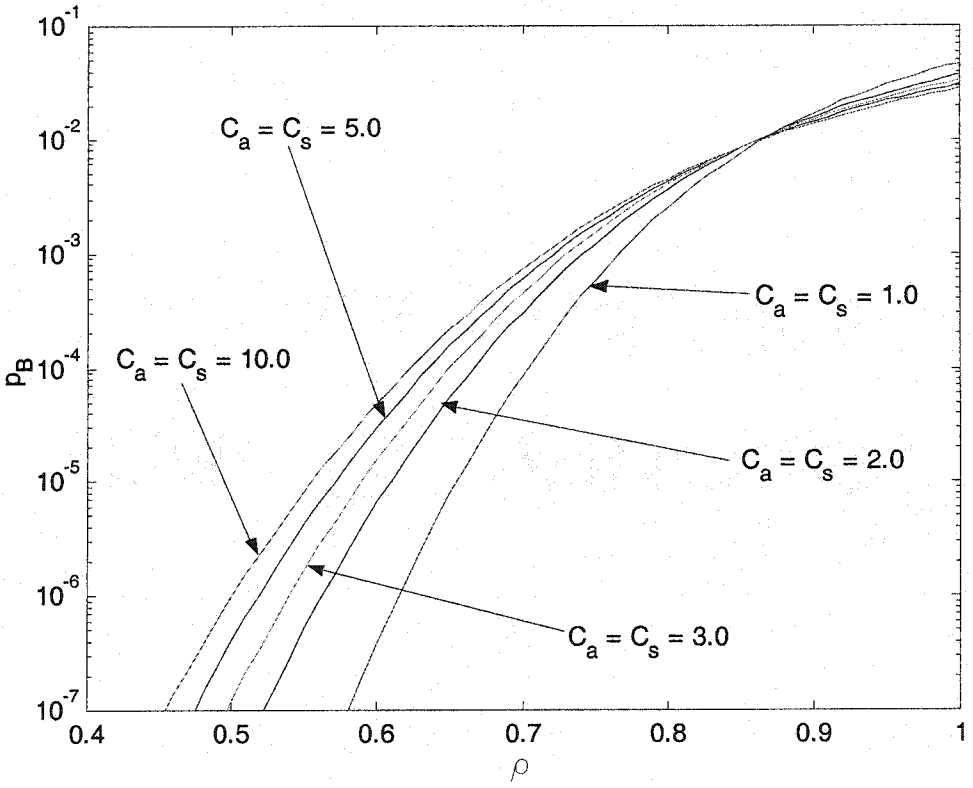


Figure 5.14 Évolution de la probabilité de blocage avec C_a et C_s variables

5.5 Adaptabilité des modèles proposés au trafic multimédia

La plupart des études modélisant le trafic des réseaux mobiles de la prochaine génération considèrent un flux unique regroupant à la fois la voix, les données et la vidéo (Jabbari, 1996; Lam *et al.*, 1997; Fang et Chlamtac, 1999). Ces études supposent également que les canaux sont en mesure d'accommoder indistinctement l'un ou l'autre de ces médias, le type de média étant pris en compte par le processus d'arrivée et par la discipline de service. C'est d'ailleurs pourquoi nous avons insisté, à la section 5.1, sur les principaux processus d'arrivée et les disciplines de service qui sont proposés dans la littérature pour caractériser le trafic multimédia. Dans cette section, nous nous proposons de décomposer un tel trafic en R composantes et d'évaluer, à partir des modèles proposés aux sections 5.2 et 5.3, la distribution de chacune d'entre elles.

Soit $\{n_r,\ r = 1, 2, \ldots, R\}$ le nombre d'usagers mobiles qui utilisent le trafic de classe r dans chaque cellule. Il s'agit alors de déterminer la distribution marginale $p_r(n_r)$, chaque classe r de trafic étant caractérisée par son processus d'arrivée et sa distribution de service. À partir du modèle M/G/∞, l'évaluation d'une telle distribution en régime permanent se fait en utilisant les résultats présentés par Baynat (2000) pour exprimer $p_r(n_r)$, ce qui conduit à la relation suivante :

$$p_r(n_r) = e^{-\lambda_r/\mu_r} \frac{\left(\lambda_r/\mu_r\right)^{n_r}}{n_r!} \tag{5.63}$$

où $1/\mu_r$ représente la durée moyenne d'occupation des canaux par les utilisateurs du trafic de classe r, λ_r le taux moyen d'arrivée du trafic de classe r et n_r le nombre d'utilisateurs du trafic de classe r, $r = 1, 2, \ldots, R$.

Toutefois, dans le cas du modèle général à capacité finie, nous devons définir un paramètre c_r pour caractériser le nombre maximal d'utilisateurs de classe r qui peuvent être supportés dans chaque cellule. Le modèle à résoudre devient alors une file d'attente de type $G/G/c_r/c_r$. Dans ce contexte, nous pouvons nous inspirer de la relation (5.29) pour exprimer la distribution marginale $p_r(n_r)$ de la manière suivante :

$$p_r(n_r) = p_r(0)\, H_{n_r}\, y_r^{\,g_r(n_r)}, \qquad n_r = 1, \ldots, c_r \tag{5.64}$$

où :

$$g_r(n_r) = \begin{cases} 0, & \text{si } n_r = 0,\ 1,\ 2, \cdots,\ c_r - 1 \\ 1, & \text{si } n_r = c_r \end{cases} \tag{5.65}$$

et

$$H_{n_r} = \prod_{l=1}^{n_r} h_r(l)$$

Le paramètre $p_r(0)$ peut être obtenu à partir de la relation (5.32) de la manière suivante :

$$p_r(0) = \left[1 + \sum_{i=1}^{c_r - 1} H_i + y_r\, H_{c_r} \right]^{-1} \tag{5.66}$$

avec :

$$H_i = \prod_{l=1}^{i} h_r(l) = h(1) * h(2) * \ldots * h(i) \tag{5.67}$$

Pour déterminer $\{h_r(l),\ l = 1,\ 2,\ \ldots,\ c_r\}$ et y_r, définissons les paramètres suivants :

$$a_r = \frac{2}{C_{a_r} + 1} \tag{5.68}$$

$$b_r = \frac{2}{C_{s_r} + 1} \tag{5.69}$$

$$\rho_r = \lambda_r / c_r\, \mu_r \tag{5.70}$$

où λ_r et C_{a_r} désignent respectivement la moyenne et le carré du coefficient de variation du trafic d'arrivée de classe r, alors que μ_r et C_{s_r} sont respectivement la moyenne et le carré du coefficient de variation de la distribution de service du trafic de classe r. À ce niveau, nous nous inspirons des relations (5.40) à (5.43) pour déterminer $\{h_r(l),\ l = 1,\ 2,\ \ldots,\ c_r\}$ et y_r en fonction de a_r, b_r, ρ_r et c_r de la manière suivante :

$$h_r(1) = \begin{cases} \dfrac{a_r c_r \rho_r}{b_r(1-a_r)+a_r}, & \text{si } c_r > 1 \\ \dfrac{a_r \rho_r b_r}{a_r \rho_r (1-b_r)+b_r}, & \text{si } c_r = 1 \end{cases} \tag{5.71}$$

$$h_r(l) = \begin{cases} \dfrac{a_r c_r \rho_r + (l-1)b_r(1-a_r)}{l\left[b_r(1-a_r)+a_r\right]}, & l = 2, \ldots, c_r - 1 \\ \dfrac{b_r\left[a_r c_r \rho_r + (c_r-1)b_r(1-a_r)\right]}{c_r\left[a_r \rho_r (1-b_r)+b_r\right]}, & l = c_r \end{cases} \tag{5.72}$$

$$y_r = \frac{1}{1-(1-b_r)z_r} \tag{5.73}$$

avec

$$z_r = \frac{a_r \rho_r + b_r(1-a_r)}{a_r \rho_r (1-b_r)+b_r} \tag{5.74}$$

Ainsi, les relations (5.71) à (5.74) permettent de déterminer respectivement $\{H_{n_r}, n_r = 1, 2, \ldots, c_r\}$ et $p_r(0)$. La distribution marginale $p_r(n_r)$ s'obtient alors en remplaçant $p_r(0)$, H_{n_r}, y_r et $g_r(n_r)$ par leur valeur dans la relation (5.64).

Ainsi, dans ce chapitre, nous avons présenté et justifié deux modèles d'analyse de trafic qui permettent de caractériser le comportement des usagers dans les systèmes mobiles de la prochaine génération : le M/G/∞ (ou M/M/∞ sous certaines conditions) et le G/G/c/c. Même si le premier modèle reste théorique, il nous a permis d'évaluer et de comparer la distribution du trafic dans les pico-cellules, micro-cellules et macro-cellules. Quant au second modèle qui considère un trafic d'arrivée de loi générale, il nous a permis d'évaluer de manière plus réaliste et dans plusieurs conditions d'analyse la distribution du nombre d'usagers en communication, ainsi que la probabilité de blocage d'appels dans chaque cellule. En plus de permettre de spécifier les conditions dans lesquelles chaque cellule offre une bonne qualité de service, ce modèle permet, entre autres, d'analyser le comportement de tout système qui peut être modélisé par une file d'attente de type G/G/c/c.

CHAPITRE 6

CONCLUSION

Il est clair que le processus de planification constitue une tâche complexe pour tout concepteur de réseaux mobiles. Pour les systèmes de la prochaine génération, ce processus devient encore plus complexe car il fait appel à une série de paramètres difficiles à évaluer et à contrôler. C'est dans ce contexte que nous avons présenté une approche modulaire de planification et élaboré sur deux aspects fondamentaux de ce processus : la gestion de la mobilité globale et l'ingénierie de trafic. Il nous revient maintenant de synthétiser les principaux résultats obtenus et de proposer quelques orientations de recherche.

6.1 Synthèse des résultats

La méthodologie préconisée se base sur l'identification de cinq facteurs fondamentaux d'ingénierie de systèmes pour contrôler le processus de planification et optimiser conjointement les coûts de planification et la capacité du système, en respectant les contraintes de qualité de service. Chaque facteur est pris en compte par un module qui traite d'un aspect particulier du problème de planification. Ces modules interviennent de façon parallèle dans le processus de planification, et la prise en compte des interactions entre eux permet une gestion simultanée de plusieurs sous-problèmes ayant des objectifs éventuellement conflictuels les uns par rapport aux autres.

Étant donné que tous les modules ne sont pas encore implémentés, nous n'avons pas présenté de résultats sur les coûts de planification. Par contre, nous avons présenté et analysé des résultats qui portent sur le compromis existant entre la capacité du réseau et la puissance minimale requise par utilisateur pour maintenir une communication de qualité. Ces résultats ont montré que, pour une valeur donnée du rapport signal à bruit, la

puissance minimale requise par usager décroît lorsque la capacité de la cellule augmente. En outre, puisque l'interférence constitue une mesure caractéristique de la qualité de service, nous avons analysé le comportement du rapport signal à bruit minimal reçu par utilisateur en fonction du nombre de terminaux actifs, c'est-à-dire du niveau d'interférence. Nous avons alors trouvé que plus le trafic dans le système est intense, plus le système doit fournir de la puissance pour maintenir une bonne qualité de service. Nous avons également comparé le rapport signal à bruit minimal par usager pour les macro-cellules, micro-cellules et pico-cellules. Nous avons alors trouvé qu'un tel rapport est plus élevé pour les macro-cellules que pour les autres types de cellules, c'est-à-dire les macro-cellules doivent développer plus de puissance que les micro-cellules ou les pico-cellules pour maintenir le même niveau de qualité de service. Les résultats obtenus ont aussi permis de vérifier que l'utilisation de micro-cellules ou de pico-cellules aide à augmenter la capacité du système.

Par ailleurs, nous avons vu que, dans les systèmes mobiles de la prochaine génération, la localisation d'un abonné ou l'accès à ses services repose sur l'interopérabilité de plusieurs sous-réseaux fixes et mobiles. Dans ce contexte, nous avons conçu et mis en œuvre une approche de gestion de mobilité globale basée sur l'utilisation d'un équipement d'interconnexion appelé WING (*Wireless INterworking Gateway*) qui facilite l'interopérabilité des composantes hétérogènes du réseau. De plus, pour évaluer la quantité de trafic de signalisation généré, nous avons défini les séquences d'opérations mises en œuvre au niveau du WING lors des changements de sous-systèmes, et modélisé chaque WING par une file d'attente de type M/G/1. Les performances de l'approche proposée ont été évaluées pour deux modèles de mobilité : le modèle fluide et le modèle de gravité.

L'évaluation de performance nous a permis de tirer plusieurs conclusions. D'abord, pour un taux fixé d'appels émis ou reçus, le trafic de signalisation généré au niveau des bases de données tend à augmenter lorsque le degré de mobilité globale des abonnés augmente, mais ne dépend pas du modèle de mobilité utilisé. Les résultats ont

aussi révélé que le temps de réponse aux requêtes décroît linéairement en fonction d'un paramètre p qui caractérise la stratégie de stockage d'informations et qui indique la probabilité qu'une information requise lors d'une requête soit accessible au niveau d'un VLR. Dans le même ordre d'idées, nous avons considéré deux sous-systèmes i et j, et analysé, pour chaque modèle de mobilité, l'influence de la répartition des usagers sur le trafic de signalisation, ainsi que sur le temps de réponse du réseau. Les résultats obtenus de cette analyse ont montré que l'influence de la répartition des usagers sur la qualité de service dépend du modèle de mobilité considéré : alors que pour le modèle fluide, le réseau offre une meilleure qualité de service lorsque les abonnés se concentrent davantage dans le sous-système i, ce même réseau offre une meilleure qualité de service lorsque les abonnés se concentrent davantage dans le sous-système j lorsque l'on considère le modèle de gravité. Dans tous les cas, nous avons trouvé que, peu importe le comportement ou la répartition des abonnés, la stratégie de stockage ou le modèle de mobilité considéré, l'approche préconisée dans cette thèse contribue à améliorer significativement les performances du réseau, en termes de trafic de signalisation et de délai associé à la télé-recherche et à la localisation des utilisateurs.

D'autre part, puisque les systèmes mobiles de la prochaine génération sont basés sur la technologie CDMA, la dimension et la capacité de leurs cellules seront essentiellement déterminées par la distribution spatiale des usagers mobiles (Tutschku et Tran-Gia, 1998). Dans ce contexte, nous avons proposé deux modèles de files d'attente pour caractériser le trafic dans chaque cellule : le M/G/∞ (ou M/M/∞ sous certaines conditions) et le G/G/c/c. Le premier modèle se base sur la capacité élastique de la technologie CDMA pour modéliser chaque cellule par un système n'ayant aucune limite théorique sur le nombre de communications à gérer. Ce modèle nous a permis d'analyser, entre autres, l'effet de la taille des cellules sur l'intensité de trafic. Les résultats de cette analyse ont montré que, pour les macro-cellules, l'intensité moyenne du trafic est plus faible que pour les micro-cellules ou les pico-cellules qui, quant à elles, sont généralement considérées comme des zones plus denses. Il en résulte que le risque

d'interférence est plus élevé dans les pico-cellules que dans les micro-cellules ou les macro-cellules.

Quant au modèle G/G/c/c, son analyse étant complexe, nous avons appliqué les principes d'entropie maximale (surtout utilisés en théorie de l'information) pour le résoudre. Cela nous a alors permis d'évaluer de manière réaliste à la fois la distribution du trafic et la probabilité de blocage d'appels dans chaque cellule, en plus de permettre l'analyse du comportement de tout système qui peut être modélisé par une file d'attente de type G/G/c/c. D'ailleurs, nous en avons directement déduit les résultats des modèles de type M/M/c/c, dont la formule d'Erlang-B. En termes de résultats, le modèle G/G/c/c nous a permis d'analyser, d'une part, l'effet des coefficients de variation des distributions d'arrivée et de service C_a et C_s sur la distribution du trafic $p(n)$ et, d'autre part, celui des coefficients de variation sur la probabilité de blocage p_B. Pour l'analyse de l'effet de C_a et de C_s sur $p(n)$, nous avons trouvé que la valeur maximale de $p(n)$, notée $p_{max}(n)$, augmente, alors que p_B décroît lorsque nous faisons augmenter C_s. Toutefois, nous avons noté que la variabilité de $p_{max}(n)$ et de p_B demeure relativement faible lorsque C_s varie. D'ailleurs, lorsque $C_a = 1$, nous avons vérifié que $p(n)$ devient insensible à la valeur de C_s (≥ 1). Autrement dit, lorsque la distribution du temps d'interarrivées est exponentielle et que le temps de service est général, $p(n)$ demeure invariant pour tout $C_s \geq 1$. Ces résultats ont ainsi permis de valider que la distribution du trafic est insensible au temps de service pour tout modèle de file d'attente de type M/G/c/c.

En outre, lorsque C_s est fixe et C_a variable, nous avons obtenu une variabilité importante de la distribution du trafic. Plus spécifiquement, nous avons trouvé que, lorsque C_a augmente, le paramètre $p_{max}(n)$ diminue, alors que la probabilité de blocage p_B augmente. Nous nous sommes rendus compte que $p_{max}(n)$ et p_B ont des comportements opposés selon que l'on fait varier C_a ou C_s. C'est pourquoi nous avons entrepris de faire varier simultanément C_a et C_s au même taux, et d'analyser l'effet d'une telle variation sur la distribution du trafic dans les cellules. Nous avons alors trouvé que le coefficient de variation de la distribution du trafic d'arrivée a plus d'influence sur la distribution du

nombre de terminaux actifs que celui de la distribution du temps d'occupation des canaux.

Nous avons enfin entrepris d'analyser l'effet de C_a et de C_s sur p_B de manière à déterminer, en fonction de C_a et de C_s, les conditions selon lesquelles le réseau offre de faibles probabilités de blocage, c'est-à-dire une meilleure qualité de service. Nous avons trouvé que, pour des valeurs fixées de C_a, une meilleure qualité de service peut être offerte pour des valeurs élevées de C_s. De plus, nous avons vérifié que, pour un trafic d'arrivée poissonien (C_a = 1), p_B reste invariante pour tout $C_s \geq 1$, ce qui valide que la probabilité de blocage est insensible à la distribution du temps de service dans un modèle de type M/G/c/c. Dans le même ordre d'idées, nous avons trouvé qu'une meilleure qualité de service est offerte dans chaque cellule pour des valeurs faibles de C_a lorsque C_s est fixe. Nous avons toutefois constaté que p_B est plus sensible aux variations de C_a qu'à celles de C_s. C'est pourquoi nous avons fait varier C_a et C_s au même taux et analyser l'effet de cette variation sur p_B. Nous avons alors trouvé qu'une meilleure qualité de service est offerte dans chaque cellule pour des valeurs faibles de $C_a = C_s$. Il en résulte que le coefficient de variation du trafic d'arrivée a plus d'impact sur la qualité de service que celui du temps d'occupation des canaux.

6.2 Limitations et orientations de recherche

L'implémentation de la méthodologie préconisée doit demeurer au cœur de toute activité de recherche future relative à cette thèse. Toutefois, une telle méthodologie nécessite non seulement une analyse plus approfondie des aspects de couverture de l'environnement radio, d'architecture du réseau et d'allocation de ressources, mais aussi un raffinement des aspects de gestion de mobilité globale et d'ingénierie de trafic. Cela conduirait évidemment à l'implémentation de l'ensemble du processus, en intégrant tous les modules et en définissant les paramètres d'échanges qui facilitent les interactions entre ces modules. Nous aurions ainsi la preuve que l'utilisation de cette méthodologie

permet de répondre aux objectifs de coût et de capacité, en plus d'améliorer les résultats obtenus par d'autres méthodologies proposées dans la littérature.

Par ailleurs, les recherches futures relatives à l'approche de gestion de la mobilité globale devraient se concentrer sur l'implémentation, au niveau des WINGs, d'algorithmes performants qui permettent de tenir compte du profil des usagers. Cela donnerait lieu à des équipements intelligents capables de prédire le comportement de tout usager mobile susceptible de changer de sous-systèmes. On aurait ainsi davantage amélioré les performances du réseau. Toutefois, une question demeure sans réponse : combien coûterait l'insertion de tels équipements (WINGs intelligents) aux fournisseurs de service ? Une telle question invite à entreprendre une étude de coût de manière à évaluer la rentabilité d'une telle proposition.

Dans un autre ordre d'idée, les travaux futurs relatifs au modèle de trafic M/G/∞ devraient mettre l'accent sur l'évaluation de la qualité de service lorsque le système fonctionne avec un niveau minimal de rapport signal à interférence par usager. Cela pourrait faciliter l'analyse d'un autre paramètre important de la qualité de service, en l'occurrence le taux d'erreur binaire. Une autre extension possible de ce modèle de trafic serait de considérer un trafic d'arrivée non poissonien et de résoudre un système de type G/G/∞ pour l'évaluation de la distribution des terminaux actifs. Il s'agirait alors d'une autre contribution d'envergure à la modélisation, car selon nos recherches, il n'existe pas encore de méthode permettant de résoudre un tel problème.

Pour finir, les recherches futures pourraient également porter sur d'autres particularités pour modéliser le trafic dans les systèmes mobiles de la prochaine génération. Parmi lesquelles, on pourrait entreprendre une analyse de résultats en considérant différentes classes d'usagers qui utilisent des patrons de trafic différents. Cette analyse permettrait d'étudier l'influence de la taille des cellules sur chaque classe de trafic, en plus d'étudier l'évolution du niveau de qualité de service en fonction du type de média. Cependant, une telle analyse est loin d'être triviale.

BIBLIOGRAPHIE

AGHVAMI, H. et JAFARIAN, B. (2000). A Vision of UMTS/IMT-2000 Evolution. Electronics and Communication Engineering Journal, 12, 148-152.

AGRAWAL, P. et FAMOLARI, D. (1999). Mobile Computing in the Next-Generation Wireless Networks. Proceedings of the 3rd International Workshop on Discrete Algorithms and Methods for Mobile Computing and Communications 1999, Seattle, USA, 32-39.

AKYILDIZ, I. F., McNAIR, J., HO, J.S.M., UZUNALIOGLU, H. et WANG, W. (1998). Mobility Management in Current and Future Communication Networks, IEEE Network, 12, 39-49.

AKYILDIZ, I. F. et WANG, W. (2002). A Dynamic Location Management Scheme for Next-Generation Multitier PCS Systems, IEEE Transactions on Wireless Communications, 1, 178-189.

BAHAI, A. R. S. et AGHVAMI, H. (2000). Network Planning and Optimization in the Third-Generation Wireless Networks, First International Conference on 3G Mobile Communication Technologies, 27-29 Mars 2000, London, UK, 441-445.

BAR-NOY, A., KESSLER, I. et NAGHSHINEH, M. (1996). Topology-Based Tracking Strategies for Personal Communication Networks, Mobile Networks and Applications, 1, 49-56.

BARCELO, F. et JORDAN, J. (1997). Channel Holding Time Distribution in Cellular Telephony, Proceeding of 9th International Conference on Wireless Communications (Wireless '97), Alta, Canada, 9-11 Juillet 1997, 1, 125-134.

BAYNAT, B. (2000). Théorie des files d'attente, HERMES Science, Paris, France.

BEAUBRUN, R., PIERRE, S. et CONAN, J. (1999). An Efficient Method for Optimizing the Assignment of Cells to MSCs in PCS Networks, Proceeding of

11th International Conference on Wireless Communications (Wireless '99), Calgary, Canada, 12-14 Juillet 1999, 1, 259-265.

BEAUBRUN, R., PIERRE, S., FLOCCHINI, P. et CONAN, J. (2002). Global Roaming Management in the Next-Generation Wireless Systems, IEEE International Conference on Communications 2002 (ICC 2002), New York, USA, 28 Avril - 2 Mai 2002, 4, 2070-2074.

BELLMAN, R. E. et DREYFUS, S. (1962). Applied Dynamic Programming, Princeton University Press, Princeton, N. J., USA.

BRIA, A., GESSLER, F., QUESETH, O., STRIDH, R., UNBEHAUN, M., WU, J., ZANDER, J. et FLAMENT M. (2001). 4th-Generation Wireless Infrastructures: Scenarios and Research Challenges, IEEE Personal Communications, 8, 25-31.

BROWN, T. X. et MOHAN, S. (1997). Mobility Management for Personal Communications Systems. IEEE Transactions on Vehicular Technology, 46, 269-278.

BUCHANAN, K., FUDGE, R., McFARLANE, D., PHILLIPS, T., SASAKI, A. et XIA, H. (1997). IMT-2000: Service Provider's Perspective, IEEE Personal Communications, 4, 8-13.

CASTRO, J. P. (2001). The UMTS Network and Radio Access Technology, Wiley & Sons Ltd., West Sussex, England.

CHLEBUS, E. et LUDWIN, W. (1995). Is Handoff Traffic Really Poissonian?, IEEE International Conference on Universal Personal Communications (ICUPC '95), Tokyo, Japon, 6-10 Novembre 1995, pp. 348-353.

COOMBS, R. et STEELE, R. (1999). Introducing Microcells into Macrocellular Networks: A case study, IEEE Transactions on Communications, 47, 568-576.

COX, D. C. (1990). Personal Communications: A Viewpoint, IEEE Communications Magazine, 28, 8-20.

CROW, E. L. et SHIMIZU, K. (1988). Lognormal Distribution: Theory and Applications, Marcel Dekker Inc., New York, USA.

DAYEM, R. A. (1997). PCS & Digital Cellular Technologies : Assessing your options, Prentice Hall, Upper Saddle River, NJ, 93-141.

EL-AFFENDI, M. A. et KOUVATSOS, D. D. (1983). A Maximum Entropy Analysis of the M/G/1 and G/M/1 Queueing Systems at Equilibrium, Acta Informatica, 19, 339-355.

ESCALLE, P. G., GINER, V. C. et OLTRA, J. M. (2002). Reducing Location Update and Paging Costs in a PCS Network, IEEE Transactions on Wireless Communications, 1, pp. 200-209.

EVANS, J. R. et MINIEKA, E. (1992). Optimization Algorithms for Networks and Graphs, Marcel Dekker, New York, USA.

FANG, Y. (2000). Registration Traffic and Service Availability for Two-tier Wireless Networks, IEEE Wireless Communications and Networking Conference 2000 (WCNC 2000), 23-28 Septembre 2000, Chicago, IL, USA, 3, 1090-1095.

FANG, Y. et CHLAMTAC, I. (1999). Teletraffic Analysis and Mobility Modelling of PCS Networks, IEEE Transactions on Communications, 47, 1062-1072.

FANG, Y., CHLAMTAC, I. et LIN, Y.-B. (1997). Call Performance for PCS Network, IEEE Journal on Selected Areas in Communications, 15, 1568-1581.

FANG, Y., CHLAMTAC, I. et LIN, Y.-B. (2000). Portable Movement Modeling for PCS Networks, IEEE Transactions on Vehicular Technology, 49, 1356-1363.

FERDINAND, A. E. (1970). A Statistical Mechanical Approach to Systems Analysis, IBM Journal of Research and Development, pp. 539-547.

FRULLONE, M., RIVA, G., GRAZIOSO, P. et FALCIASECCA G. (1996). Advanced Planning Criteria for Cellular Systems, IEEE Personal Communications, 3, 10-15.

FRUSCIO, G. et PETRONE, V. (1999). Evolving towards Universal Mobile Telecommunication Systems: Development of a new System Based on Emerging Technologies, IEEE Wireless Communications and Networking Conference (WCNC 1999), Septembre 1999, 403-407.

GANZ, A., KRISHNA, C. M., TANG, D. et HAAS Z. J. (1997). On Optimal Design of Multitier Wireless Cellular Systems, IEEE Communications Magazine, 35, 88-93.

GARG, V. K., HALPERN, S. et SMOLIK, K. S. (1999). Third Generation (3G) Mobile Communications Systems, IEEE International Conference on Personal Wireless Communications (ICPWC 99), 17-19 Février 1999, Jaipur, Inde, 39-43.

GARG, V. K. et WILKES, J. E. (1996). Interworking and Interoperability Issues for North American PCS, IEEE Communications Magazine, 34, 94-99.

GELENBE, E. et PUJOLLE, G. (1982). Introduction aux réseaux de files d'attente, Eyrolles, Paris, France.

GIBSON, J. D. (1999). The Mobile Communications Handbook, CRC Press, Miami, USA, 20-1 – 24-12.

GILL, D., COSMAS, J. P. et PEARMAIN A. (2000). Mobile Audio-Visual Terminal: System Design and Subjective Testing in DECT and UMTS Networks, IEEE Transactions on Vehicular Technology, 49, 1378-1391.

GÖRG, C., GUNTERMANN, M. et KLEIER, S. (1997). Future Systems for Personal Mobility Services: Design, Performance Evaluation and Implementation, IEEE Journal on Selected Areas in Communications, 15, 1672-1683.

GROSS, D. et HARRIS, C. M. (1974). Fundamentals of Queueing Theory, John Wiley & Sons, New York, USA.

HAC, A. et CHEN, Z. (2000). A Hybrid Channel Allocation Method for Wireless Communication Network, International Journal of Network Management, 10, 59-74.

HE, E., DU, F., DONG X., NI, L. M. et HUGHES H. D. (2000). Video Traffic Modeling Over Wireless Networks, IEEE International Conference on Communications 2000 (ICC 2000), Juin 2000, 536-542.

HO, J. S. M., AKYILDIZ, I. F. (1997). Dynamic Hierarchical Database Architecture for Location Management in PCS Networks, IEEE/ACM Transactions on Networking, 5, 646-660.

HONG, D. et RAPPAPORT, S. S. (1986). Traffic Model and Performance Analysis for Cellular Mobile Radio Telephone Systems with Prioritized and Nonprioritized Handoff Procedures, IEEE Transactions on Vehicular Technology, 35, 77-92.

HOORN, M. H. V. et SEELEN L. P. (1986). Approximations for the GI/G/c queue, Journal of Applied Probability, 23, 484-494.

IERA, A., MOLINARO, A. et MARANO S. (2000). Managing Symmetric and Asymmetric Data Traffic in Integrated Terrestrial-Cellular and Satellite Systems, IEEE Personal Communications, 7, 56-64.

JABBARI, B. (1996). Teletraffic Aspects of Evolving and Next-Generation Wireless Communication Networks, IEEE Personal Communications, 3, 4-9.

JAIN, R., LIN, Y.-B., LO, C. et MOHAN S. (1994). A Caching Strategy to Reduce Network Impacts of PCS, IEEE Journal on Selected Areas in Communications, 12, 1434-1444.

JEDRZYCKI, C. et LEUNG, V. C. M. (1996). Probability Distributions of Channel Holding Time in Cellular Telephone Systems, IEEE Vehicular Technology Conference (VTC '96), Atlanta, GA, Mai 1996, 247-251.

JORGUSESKI, L., FLEDDERUS, E., FARSEROTU J. et PRASAD R. (2001). Radio Resource Allocation in Third-Generation Mobile Communication Systems, IEEE Communications Magazine, 39, 117-123.

KESAVAN, H. K. et KAPUR, J. N. (1989). The Generalized Maximum Entropy Principle, IEEE Transactions on Systems, Man, and Cybernetics, 19, 1042-1052.

KIM, M.-J. (2000). Traffic Engineering for Next-Generation Mobile Communication Systems, IEEE Wireless Communications and Networking Conference 2000 (WCNC 2000), 23-28 Septembre 2000, Chicago, IL, USA, 3, 1453-1456.

KLEINROCK L. (1975). Queueing Systems: Vol. I: Theory, John Wiley & Sons, Inc., New York, USA.

KOUVATSOS, D. D. (1986). Maximum Entropy and the G/G/1/N Queue, Acta Informatica, 23, 545-565.

KOUVATSOS, D. D., XENIOS, N. P. (1989). MEM for Arbitrary Queueing Networks with Multiple General Servers and Repetitive-service Blocking, Performance Evaluation, 10, 169-195.

LAGRANGE, X. (1997). Multitier Cell Design, IEEE Communications Magazine, 35, 60-64.

LAM, D., COX, D. C. et WIDOM, J. (1997), Teletrafic Modeling for Personal Communications Services, IEEE Communications Magazine, 35, 79-87.

LAM, D., JANNINK, J., COX, D. C. et WIDOM, J. (1996). Modeling Location Management in Personal Communications Services, International Conference on Universal Personal Communications, 29 Septembre - 2 Octobre 1996, Cambridge, MA, USA, 2, 596-601.

LEON-GARCIA, A. (1994). Probability and Random Process for Electrical Engineering, Addison-Wesley, Reading, Massachusetts, USA.

LEUNG, K. K. et LEVY, Y. (1997). Global Mobility Management by Replicated Databases in Personal Communication Networks, IEEE Journal on Selected Areas in Communications, 15, 1582-1596.

LEUNG, K. K., MASSEY, W. A. et WHITT, W. (1994). Traffic Models for Wireless Communication Networks, IEEE Journal on Selected Areas in Communications, 12, 1353-1364.

LIN, Y.-B. (1997). Modeling Techniques for Large-Scale PCS Networks, IEEE Communications Magazine, 35, 102-107.

LIN, Y.-B., CHANG, L. F. et NOERPEL, A. (1996). Performance Modeling of Multi-Tier PCS System, International Journal of Wireless Information Networks, 3, 67-78.

LIN, Y.-B. et CHLAMTAC, I. (1996). Heterogeneous Personal Communications Services : Integration of PCS Systems, IEEE Communications Magazine, 34, 106-113.

LIN, Y.-B., MOHAN S. et NOERPEL A. (1994). Queueing Priority Channel Assignment Strategies for PCS Hand-Off and Initial Access, IEEE Transactions on Vehicular Technology, 43, 704-712.

LISTER, D., DEHGHAN, S., OWEN, R. et JONES, P. (2000). UMTS Capacity and Planning Issues, First International Conference on 3G Mobile Communication Technologies, 27-29 Mars 2000, London, UK, 218-223.

MARCHENT, B. G., WILSON, M. J. et ROUZ, A. (1999). Support of Mobile Multimedia over Radio for a Wide of QoS and Traffic Profiles, IEEE International Conference on Personal Wireless Communications (ICPWC 99), Février 1999, 145-149.

MARKOULIDAKIS, J. G., LYBEROPOULOS, G. L., TSIRKAS, D. F. et SYKAS, E. D. (1995). Evaluation of Location Area Planning Scenarios in Future Mobile Telecommunication Systems, Wireless Networks, 1, 17-28.

MARSAN, M. A., MARANO, S., MASTROIANNI, C. et MEO, M. (2000). Performance Analysis of Cellular Mobile Communication Networks Supporting Multimedia Services, Mobile Networks and Applications, 5, 167-177.

MASSEY, W. A. et SRINIVASAN, R. (1997). A Packet Delay Analysis for Cellular Digital Packet Data, IEEE Journal on Selected Areas in Communications, 15, 1364-1372.

MATHAR, R. et NIESSEN T. (2000). Optimum Positioning of Base Stations for Cellular Radio Networks, Wireless Networks, 6, 421-428.

MATSUYA, Y., IZUMIYA, S. et MURATA, J. (2000). Summary of IMT-2000 Experiments, Vehicular Technology Conference Proceedings, 2000 (VTC2000), 15-18 Mai 2000, Tokyo, Japon, 1, 123-127.

MEDHI, J. (1991). *Stochastic Models in Queueing Theory*, Academic Press, San Diego, USA.

McNAIR, J., AKYLDIZ, I. F. et BENDER, M.D. (2000). An inter-system Handoff Technique for the IMT-2000 System, IEEE Infocom '00, Mars 2000, 208-216.

MERCHANT, A., SEGUPTA, B. (1995). Assignment of Cells to Switches in PCS Networks, IEEE/ACM Transactions on Networking, 3, 521-526.

MILLER, G., HORN, D. (1998). Maximum Entropy Approach to Probability Density Estimation, Second International Conference on Knowledge-Based Intelligent Electronic Systems, 21-23 April 1998, Adelaide, Australia, 225-230.

MILSTEIN, L. B. (2000). Wideband Code Division Multiple Access, IEEE Journal on Selected Areas in Communications, 18, 1344-1354.

MOHAN, S., JAIN, R. (1994). Two User Location Strategies for Personal Communications Services, IEEE Personal Communications, 1, 42-50.

MOMTAHAN, O. et HASHEMI, H. (2001). A Comparative Evaluation of DECT, PACS and PHS Standards for Wireless Local Loop Applications, IEEE Communications Magazine, 39, 156-163.

OJANPERÄ, T. et PRASAD, R. (1998). An overview of Third-Generation Wireless Personal Communications : A European Perspective, IEEE Personal Communications, 5, 59-65.

OLIVEIRA, C., KIM, J. B. et SUDA, T. (1998). An Adaptive Bandwidth Reservation Scheme for High-Speed Multimedia Wireless Networks, IEEE Journal on Selected Areas in Communications, 16, 858-874.

ORLIK, P. V. et RAPPAPORT, S. S. (1998). A Model for Teletraffic Performance and Channel Holding Time Characterization in Wireless Cellular Communication with General Session and Dwell Time Distributions, IEEE Journal on Selected Areas in Communications, 16, 788-803.

PAHLAVAN, K., KRISHNAMURTHY, P., HATAMI, A., YLIANTTILA, M., MAKELA J.-P., PICHNA R. et VALLSTRÖM J. (2000). Handoff in Hybrid Mobile Data Networks, IEEE Personal Communications, 7, 34-47.

PANDYA R. (1999). Mobile and Personal Communication Systems and Services, IEEE Press, New York, N. Y., USA.

PANDYA, R., GRILLO, D., LYCKSELL, E., MIEYBÉGUÉ, P., OKINAKA, H., YABUSAKI, M. (1997). IMT-2000 : Network Aspects, IEEE Personal Communications, 4, 20-29.

PARK, K. I. et LIN, Y.-B. (1997). Reducing Registration Traffic for Multitier Personal Communications Services, IEEE Transactions on Vehicular Technology, 46, 597-602.

PIERRE, S. (1998). Inferring New Design Rules by Machine Learning: A Case Study of Topological Optimization, IEEE Transactions on Systems, Man and Cybernetics, 25, 575-585.

PIERRE, S. et ELGIBAOUI, A. (1997). A Tabu-Search Approach for Designing Computer Network Topologies with Unreliable Components, IEEE Transactions on Reliability, 46, 350-359.

PIERRE, S. et HOUÉTO, F. (2002). A Tabu Search Approach for Assigning Cells to Switches in Cellular Mobile Networks, Computer Communications, 25, 464-477, 2002.

PRASAD, N. R. (1999). GSM Evolution towards Third Generation UMTS/IMT-2000, IEEE International Conference on Personal Wireless Communications (ICPWC 99), Février 1999, 50-54.

PRISCOLI, F. D. (1999). Interworking of a Satellite System for Mobile Multimedia Applications with the Terrestrial Networks, IEEE Journal on Selected Areas in Communications, 17, 385-394.

PÜTZ, S. et SCHMITZ, R. (2000). Secure Interoperation Between 2G and 3G Mobile Radio Networks, First International Conference on 3G Mobile Communication Technologies, 27-29 Mars 2000, London, UK, 28-32.

RAJARATNAM, M. et TAKAWIRA, F. (1997). Hand-off Traffic Modeling in Cellular Networks, IEEE Global Telecommunications Conference (Globecom '97), 3-8 Novembre 1997, Phoenix, AZ, 131-137.

RAMANATHAN, P., SIVALINGAM, K. M., AGRAWAL, P. et KISHORE S. (1999). Dynamic Resource Allocation Schemes During Handoff for Mobile Multimedia Wireless Networks, IEEE Journal on Selected Areas in Communications, 17, 1270-1283.

RAPPAPORT, T. S. (1996). Wireless Communications: Principles and Practice, Prentice Hall, Upper Saddle River, N. J., USA.

RICHARDSON, K. W. (2000). UMTS Overview, Electronics and Communication Engineering Journal, 12, 93-100.

ROSS, S. M. (1997). Introduction to Probability Model, Academic Press, San Diego, USA.

ROUZ, A., WILSON, M. J. et MARCHENT, B. G. (1999). Broadband Interworking Architecture (BRAIN) for Future Mobile Multimedia Systems, IEEE International Conference on Personal Wireless Communications (ICPWC 99), Février 1999, 279-283.

RYAN, R. R. (1999). Roaming Between Heterogeneous 3rd Generation Wireless Networks, IEEE Wireless Communications and Networking Conference (WCNC) 1999, 21-24 Septembre 1999, New Orleans, USA, 1, 321-323.

SAATY, T. L. (1961). Elements of Queueing Theory with Applications, McGraw-Hill Inc., New York, USA.

SAFA, H. (2001). Modèles et algorithmes pour la gestion de la localisation dans les réseaux à composantes mobiles multiservices, Thèse de doctorat, École Polytechnique de Montréal, Canada.

SAFA, H., PIERRE, S. et CONAN, J. (2000). A New Location Management Strategy for IS-41-Based Mobile Networks, Proceedings of the Twelfth Annual International Conference on Wireless Communications, 'Wireless 2000', Juillet 2000, Calgary, Canada, 295-301.

SANCHEZ, J. et THIOUME, M. (2001). UMTS : Services, Architecture et WCDMA, Hermès-Science Publications, Paris, France.

SCHWARTZ, M. (1995). Network Management and Control Issues in Multimedia Wireless Networks, IEEE Personal Communications, 2, 8-16.

SEVANTO, J. (1999). Multimedia Messaging Service for GPRS and UMTS, IEEE Wireless Communications and Networking Conference (WCNC 1999), Septembre 1999, 1422-1426.

SHORE, J. E. (1982). Information Theoretic Approximations for M/G/1 and G/G/1 Queueing Systems, Acta Informatica, 17, 43-61.

SIMONOT, F. (1998). A Comparison of the Stationary Distributions of GI/M/c/k and GI/M/c, Journal of Applied Probability, 35, 510-515.

SMITH, C. et COLLINS, D. (2002). 3G Wireless Networks, McGraw-Hill, New York, USA.

SOLLENBERGER, N. R., SESHADRI, N. et COX, R. (1999), The Evolution of IS-136 TDMA for Third-Generation Wireless Services, IEEE Personal Communications Magazine, 6, 8-18.

SPILLING, A. G., NIX, A. R., BEACH M. A. et HARROLD T. J. (2000). Self-Organisation in Future Mobile Communications, Electronics and Communication Engineering Journal, 12, 133-147.

STEEL, R. (1990). Deploying Personal Communication Networks, IEEE Communications Magazine, 28, 12-15.

TABBANE, S. (1997). Location Management Methods for Third-Generation Mobile Systems, IEEE Communications Magazine, 35, 72-84.

TABBANE, S. (1997). Réseaux mobiles, Éditions HERMES, Paris, France.

TABBANE, S. (2000). Handbook of Mobile Radio Networks, Artech House, Boston, MA, USA.

TIAN, Q. et COX, D. C. (2000). Location Management in a Heterogeneous Network Environment, IEEE Wireless Communications and Networking Conference 2000 (WCNC 2000), 23-28 Septembre 2000, Chicago, IL, USA, 2, 753-758.

TUTSCHKU, K. (1998). Interference Minimization Using Automatic Design of Cellular Communication Networks, IEEE 48th Vehicular Technology Conference 1998 (VTC 98), 18-21 Mai 1998, Ottawa, Canada, 634-638.

TUTSCHKU, K. et TRAN-GIA, P. (1998). Spatial Traffic Estimation and Characterization for Mobile Communication Network Design, IEEE Journal on Selected Areas in Communication, 16, 807-811.

VARSHNEY, U., SNOW, A. P. et MALLOY, A. D. (1999). Designing Survivable Wireless and Mobile Networks, IEEE Wireless Communications and Networking Conference 1999 (WCNC 99), 21-24 Septembre 1999, New Orleans, USA, 30-34.

VITERBI, A. M., VITERBI, A. J. (1993). Erlang Capacity of a Power Controlled CDMA System, IEEE Journal on Selected Areas in Communications, 11, 892-900.

WALSTRA, R. J. (1985). Nonexponential Networks of Queues: A Maximum Entropy Analysis, ACM SIGMETRICS Conference on Measurement and Modeling of Computer Systems, Austin, Texas, 26-29 Août 1985, 27-37.

WANG, S., ZHENG, H. et COPELAND, J. A. (2000). An Error Control Design for Multimedia Wireless Networks, Vehicular Technology Conference Proceedings 2000 (VTC 2000), Mai 2000, 795-799.

WANG, W. et AKYILDIZ, I. F. (2000). Intersystem Location Update and Paging Schemes for Multitier Wireless Networks, Proceedings of the Sixth Annual International Conference on Mobile Computing and Networking, 6-11 Août 2000, Boston MA, USA, pp. 99-109.

WEY, J.-K., YANG, W.-P. et SUN, L.-F. (1997). Traffic Impacts of International Roaming on Mobile and Personal Communications with Distributed Data Management, Mobile Networks and Applications, 2, 345-356.

WONG, V. W.-S. et LEUNG, V. C. M. (2000). Location Management for Next-Generation Personal Communications Networks, IEEE Network, 14, 18-24.