Titre: Title:	Mise en correspondance de la qualité de service UMTS avec celle d'une dorsale IP pour les services de téléphonie multimédia
Auteur: Author:	Racha Ben Ali
Date:	2003
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Ben Ali, R. (2003). Mise en correspondance de la qualité de service UMTS avec celle d'une dorsale IP pour les services de téléphonie multimédia [Mémoire de maîtrise, École Polytechnique de Montréal]. PolyPublie. https://publications.polymtl.ca/7222/

URL de PolyPublie: PolyPublie URL:	https://publications.polymtl.ca/7222/
Directeurs de recherche: Advisors:	Samuel Pierre
Programme: Program:	Génie informatique

UNIVERSITÉ DE MONTRÉAL

ÉCOLE POLYTECHNIQUE DE MONTRÉAL

MISE EN CORRESPONDANCE DE LA QUALITÉ DE SERVICE UMTS AVEC CELLE D'UNE DORSALE IP POUR LES SERVICES DE TÉLÉPHONIE MULTIMÉDIA

RACHA BEN ALI
DÉPARTEMENT DE GÉNIE INFORMATIQUE
ÉCOLE POLYTECHNIQUE DE MONTRÉAL

MÉMOIRE PRÉSENTÉ EN VUE DE L'OBTENTION
DU DIPLÔME DE MAÎTRISE ÈS SCIENCES APPLIQUÉES (M.Sc.A.)
(GÉNIE INFORMATIQUE)
SEPTEMBRE 2003

Your file Votre référence
ISBN: 0-612-89177-1
Our file Notre référence
ISBN: 0-612-89177-1

UNIVERSITÉ DE MONTRÉAL

ÉCOLE POLYTECHNIQUE DE MONTRÉAL

Ce mémoire intitulé :

MISE EN CORRESPONDANCE DE LA QUALITÉ DE SERVICE UMTS AVEC CELLE D'UNE DORSALE IP POUR LES SERVICES DE TÉLÉPHONIE MULTIMÉDIA

Présenté par : BEN ALI Racha

En vue de l'obtention du diplôme de : Maîtrise ès sciences appliquées

A été dûment accepté par le jury d'examen composé de :

M. CHAMBERLAND Steven, Ph.D., Président

M. PIERRE Samuel, Ph.D., Membre et Directeur de Recherche

M. CONAN Jean, Ph.D., Membre

REMERCIEMENTS

Je souhaite remercier chaleureusement mon directeur de recherche, M. Samuel PIERRE, pour ses conseils et son appui constant tout au long de mon travail de recherche.

Mes remerciements vont aussi à M. Yves LEMIEUX de Ericsson Recherche Canada pour avoir supervisé mon stage dans le cadre de la Chaire CRSNG/ERICSSON en Systèmes Réseautiques Mobiles de Prochaines Générations.

Je tiens à remercier également les membres du LARIM et plus particulièrement M. Fabien HOUÉTO pour leur soutien et leur aide précieuse.

Finalement, je tiens à exprimer ma gratitude aux membres de ma famille et à mes amis pour leur soutien inconditionnel et leur encouragement.

RÉSUMÉ

L'interconnexion des réseaux d'accès UMTS à travers des dorsales IP constitue une solution potentielle pour transporter les nouveaux services de communications multimédia sur de larges échelles géographiques à des coûts compétitifs. Toutefois, pour permettre le bon fonctionnement de ces services, il est primordial d'assurer une qualité de service (QoS) de bout en bout. D'une part, 3GPP a bien spécifié l'architecture de QoS que doit suivre un réseau d'accès UMTS. D'autre part, l'IETF a défini des standards que doit implanter un réseau IP pour supporter les architectures de QoS telles que IntServ, DiffServ et MPLS. L'expérience a montré que l'architecture DiffServ combinée avec MPLS offre une différenciation de services tout en ayant une bonne évolutivité, primordiale pour un réseau dorsal. Ainsi, l'interconnexion d'un domaine UMTS et d'un domaine dorsal IP révèle un problème de correspondance de QoS entre les deux architectures. Cela est généralement résolu par un mécanisme de mise en correspondance entre classes de service UMTS et classes DiffServ. Ce mécanisme doit être bien défini et particulièrement efficace pour les services de téléphonie multimédia émergents qui sont les plus critiques en terme de QoS.

En partant d'une caractérisation du trafic conversationnel de l'UMTS, nous avons constaté qu'il existe de grandes divergences stochastiques entre trafic de voix et trafic de vidéo-téléphonie, ce qui nous a amené, dans notre travail, à considérer un raffinement de la mise en correspondance de QoS en différenciant entre ces deux types de trafic. En effet, le trafic de voix est caractérisé par des petits paquets de taille constante alors que le trafic de vidéo-téléphonie est caractérisé par de longs paquets de taille plus ou moins variable. En outre, nous avons effectué une analyse statistique qui nous a permis de déduire, à partir du calcul du paramètre de Hurst, qu'une source de vidéo-téléphonie, contrairement à une source de voix, exprime des dépendances à longs intervalles LRD et que sa modélisation est nécessairement différente de celle de la voix, ce qui nous a conduit à introduire un modèle élaboré de trafic de vidéo-téléphonie se basant sur un processus stochastique multi-fractal. Donc, pour pouvoir différencier entre

les deux types de trafic, nous avons considéré les algorithmes d'ordonnancement les plus courants et les plus implantés sur les routeurs dorsaux, afin d'évaluer leur performance à servir le trafic conversationnel. La discipline de partage équitable pondéré WFQ étant plus appropriée, elle a fait l'objet d'un intérêt particulier dans notre étude. En effet, pour trouver les poids WFQ les plus adéquats pour les classes DiffServ de la voix et de la vidéo-téléphonie dans une configuration réelle, nous avons conduit une étude analytique qui a permis d'optimiser les performances globales du trafic conversationnel en terme de délai.

Ensuite, nous avons implanté des simulations avec différentes configurations d'algorithmes d'ordonnancement. En analysant les résultats de ces simulations, nous avons déduit que le partage équitable pondéré *WFQ*, paramétré avec les poids adéquats trouvés dans l'étude analytique, permet d'améliorer les performances globales du trafic conversationnel en termes de délai et de gigue sans affecter la synchronisation entre voix et vidéo. En outre, nous avons remarqué un fort impact du réseau d'accès radio terrestre de l'UMTS (UTRAN) sur les délais subis par les deux types de trafic. En effet, une grande portion du délai de bout en bout vient du réseau UTRAN qui est considéré par conséquent comme le goulot d'étranglement de tout le réseau étendu. De plus, nous avons noté que la contribution du réseau d'accès UTRAN au délai de bout en bout est beaucoup plus importante pour le trafic de vidéo-téléphonie que pour celui de la voix. Enfin, nous avons validé le nouveau modèle de vidéo-téléphonie que nous avons introduit dans nos simulations.

ABSTRACT

UMTS access networks interconnection through IP backbones form a potential solution for transporting new multimedia communication services over large geographic scales with competitive costs. However, so that these services work well, it is crucial to provide an end-to-end quality of service (QoS). On one hand, 3GPP has a well-specified QoS architecture that must be followed by a UMTS access network. On the other hand, IETF has defined standards that an IP network must implement to support QoS architectures such as IntServ, DiffServ and MPLS. The experience has shown that the DiffServ over MPLS architecture provides service differentiation while giving a good scalability, which is essential for a backbone network. Thus, UMTS domain and IP backbone domain interconnection reveals a problem of interoperability between the two correspondent QoS architectures. In general, this is resolved by a mapping mechanism between UMTS service classes and DiffServ classes. This mechanism must be well-defined and particularly efficient for multimedia telephony services which are the most crucial in terms of QoS.

While performing a UMTS conversational traffic characterization, we have noticed important stochastic dissimilarities between voice and video-telephony traffics, which led us to consider a QoS mapping refinement in our work. In fact, voice traffic is characterized by short and constant size packets while video-telephony traffic is characterized by long and variable size packets. Moreover, a statistical analysis that we have done shows us that, from the computing of the Hurst parameter, a video-telephony source, opposed to a voice one, expresses long range dependencies LRD and its modeling is absolutely different from the voice one. Consequently, we have introduced an elaborated model for the video-telephony traffic which is based on a multi-fractal stochastic process. Hence, in order to differentiate between the two kinds of traffic, we have considered the most common and most implemented scheduling algorithms inside backbone routers to evaluate their performance in serving conversational traffic. The

weighted fair queuing algorithm was the most appropriate and we have taken special interest in it for our study. In fact, to find the most adequate WFQ weights for voice and video-telephony DiffServ classes in a real configuration, we have conducted an analytic study for optimizing overall performance of conversational traffic in terms of delays.

Then, we have implemented simulations with scheduling algorithms under different configurations. While analysing simulation results, we have deduced that WFQ configured with the proper weights found in analytic study, improve the overall performance of conversational traffic in terms of delays and jitters without affecting the synchronisation between voice and video. Moreover, we have noticed a big impact of UMTS terrestrial radio network (UTRAN) on the delay of the two kinds of traffic. In fact, an important portion of end-to-end delay comes from the UTRAN which is consequently considered as the bottleneck of the whole extended network. Furthermore, we have noticed that the contribution of the UTRAN in the end-to-end delay is more important for video-telephony traffic than for voice traffic. Last but not least, we have validated the new video-telephony model that we have introduced in our simulations.

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

LISTE DES SIGLES ET ABRÉVIATIONS

3GPP	Third Generation Partnership Project
ATM	Asynchronous Transfer Mode
BER	Bit Error Rate
CN	Core Network
Codec	Coder/Decoder
DiffServ	Differentiated Services
EF	Expedited Forwarding
FEC	Forwarding Equivalence Class
FIFO	First In First Out
ICI	Interface Command Information
ITU	International Telecommunication Union
IETF	Internet Engineering Task Force
IntServ	Integrated Services
IP	Internet Protocol
LER	Label Edge Router
LRD	Long Range Dependencies
LSP	Label Switched Path
LSR	Label Switched Router
MPEG	Motion Picture Experts Group
MPLS	Multi-Protocol Label Switching
OSI	Open System Interconnection
PCF	Policy Control Function
PDU	Protocol Data Unit
PHB	Per Hop Behavior
PSTN	Public Switched Telephone Network
PQ	Priority Queuing
QoS	Quality of Service

RFC	Request For Comments
RSVP-TE	Resource ReserVation Protocol Traffic Engineering
SDP	Session Description Protocol
SDU	Service Data Unit
SIP	Session Initiation Protocol
SRD	Short Range Dependencies
TCP	Transmission Control Protocol
TE	Traffic Engineering
ToS	Type of Service
UDP	User Datagram Protocol
UTRAN	UMTS Terrestrial Radio Access Network
UMTS	Universal Mobile Telecommunications System
VIP	Video over IP
VoIP	Voice over IP
WFQ	Weighted Fair queuing

CHAPITRE 1
INTRODUCTION

Les progrès technologiques récents en télécommunications mobiles permettent non seulement d'offrir des hauts débits aux usagers des réseaux cellulaires de nouvelle génération mais aussi de les équiper de périphériques mobiles (*PDA*, *smartphone*, *terminaux UMTS*) assez évolués en capacité de traitement et d'affichage pour commencer à supporter les nouvelles applications de communications multimédia telles que la vidéo téléphonie, la vidéo à la demande ou encore les jeux en réseaux. L'Internet, de son coté, est en train de devenir la plate-forme de communications universelle par la convergence des différents services de communications hétérogènes faisant intervenir voix, vidéo et données sur un même réseau global qui relie les quatre coins de la planète. Les nouveaux réseaux cellulaires suivant la norme UMTS visent à offrir une panoplie de services multimédia et se veulent accessibles de n'importe où de la planète. Ainsi, ils ont à première vue tous les éléments nécessaires pour bien satisfaire ces attentes. En effet, l'utilisation de l'Internet comme plate-forme d'interconnexion de ces réseaux constitue une solution potentielle pour transporter ces services multimédia émergents à l'échelle mondiale à des coûts compétitifs. D'où le présent mémoire qui s'intéresse à des aspects bien particuliers de cette solution d'interconnexion. Dans ce chapitre d'introduction, nous présenterons les définitions et les concepts de base, nous préciserons les éléments de la problématique ainsi que les objectifs de recherche, pour finir avec une esquisse du plan du mémoire.

1.1 Définitions et concepts de base

Un *réseau cellulaire de troisième génération (3G)* est un réseau sans fil desservant des terminaux mobiles sur une zone de couverture décomposée en surfaces géographiques de différentes tailles appelées *cellules* (Figure 1.1). Le qualificatif de

troisième génération indique que c'est le successeur du réseau cellulaire de deuxième génération tel que le réseau GSM. Cette troisième génération vise à offrir une mobilité globale à l'échelle de la planète et une bande passante étendue facilitant l'introduction de nouveaux services telles que les communications multimédia. L'Union Internationale des Télécommunications (UIT) a proposé la norme IMT-2000 pour ces systèmes 3G en définissant le type de l'interface air, la bande de fréquences utilisée, la bande passante utile et les services offerts. On cite deux variantes distinctes de cette norme : l'UMTS et le CDMA2000. Bien que tous les deux définissent intégralement et distinctement l'architecture d'un accès 3G de l'interface air jusqu'aux services offerts, l'UMTS est promu pour être la technologie d'accès la plus adaptée pour une couverture internationale du fait qu'elle constitue l'évolution progressive des anciens systèmes GSM bien présents mondialement.

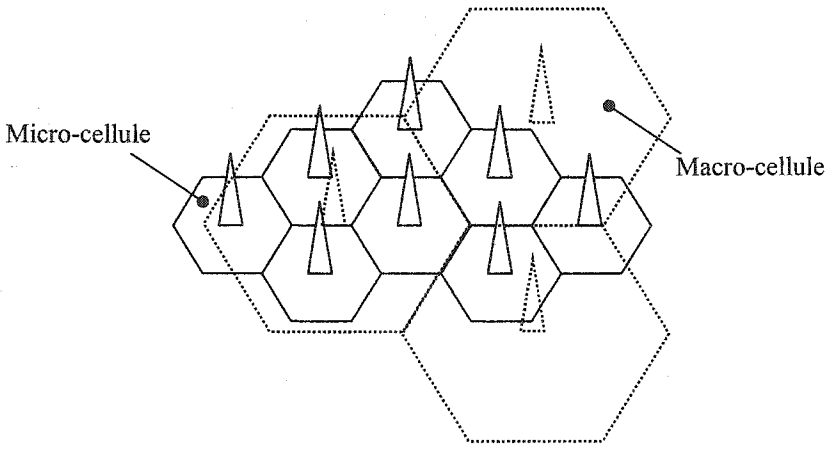


Figure 1.1 Décomposition de la couverture en cellules de différentes tailles

L'*UMTS* ou le *Système de Télécommunication Mobile Universel* est un réseau 3G qui se décompose en deux sous-réseaux : le réseau d'accès radio (RAN) et le réseau de cœur (CN) (Figure 1.2).

- Le *Réseau d'accès radio* (RAN) : il contient d'une part, l'*équipement usager* (UE) appelé aussi *station mobile* (MS) qui inclut l'*équipement terminal* (TE) et le *terminal mobile* (MT), et d'autre part, le *réseau d'accès radio terrestre* (UTRAN) qui inclut les

nœud-B (node-B) et le *contrôleur du réseau radio* (RNC). Terminal mobile et nœud-B communiquent en utilisant la technologie d'accès multiple par répartition de code à bande étendue (W-CDMA) définie un peu plus loin dans cette section.

- Le *Réseau de cœur (CN) :* il est essentiellement un réseau à commutation de paquets incluant deux types de nœuds : le *nœud de support de desserte GPRS* (SGSN) et le *nœud de support GPRS de transit* (GGSN). Ces nœuds de support incluent toutes les fonctionnalités nécessaires pour le support des services orientés paquets de UMTS. Le SGSN contrôle la localisation des usagers et accomplit les fonctions de sécurité et de contrôle d'accès. Le GGSN contient les informations de routage pour les usagers attachés au réseau en mode paquets et assure l'interopérabilité avec des réseaux à commutation de paquets externes tel que l'Internet.

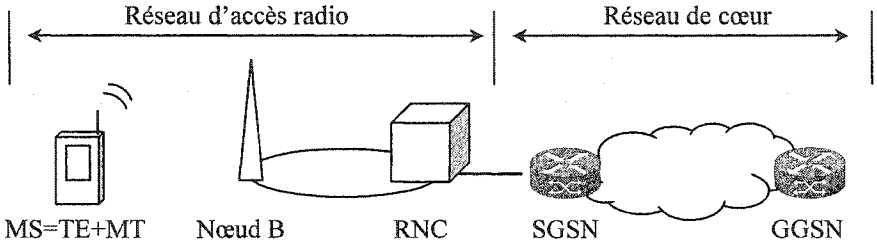


Figure 1.2 Architecture d'un réseau d'accès UMTS

W-CDMA est une technique d'accès qui permet aux unités mobiles d'une même cellule de partager un même canal radio en utilisant une grande partie de la bande passante globale. Un étalement de spectre est effectué en affectant à chaque usager un code unique qui détermine les fréquences et la puissance utilisées pour la transmission radio. Ce code d'étalement est utilisé pour permettre aux différents usagers de partager la même bande de fréquences. Dans UMTS, l'étalement de spectre se fait avec un taux de 3.84 millions de jetons (*chips*) par seconde sur une bande nominale de 5 MHz. Avec cette technique d'accès, on peut assurer une mobilité complète avec une couverture totale permettant des débits allant jusqu'à 144 Kbps voire 384 Kbps ainsi qu'une

mobilité restreinte avec une couverture plus limitée conditionnant des débits pouvant atteindre les 2 Mbps. Avec de tels débits, l'horizon s'ouvre aux communications multimédia pour les usagers mobiles.

Les *communications multimédia* sont des services de communications qui se basent généralement sur différents types de média tels que la voix, l'audio, la vidéo et les données. Ces services permettent de construire des applications telles que la vidéo-téléphonie et la vidéo-conférence qui ont été déjà introduites dans les réseaux numériques d'intégration de services (RNIS), ou encore des applications de vidéo à la demande (Video On Demand ou VoD) qui sont considérées comme les nouveaux venus dans les réseaux de télécommunications. La particularité de ces services multimédia par rapport aux autres services de données tels que le courriel ou la navigation Web, apparaît essentiellement dans leurs exigences beaucoup plus strictes en matière de qualité de service de bout en bout.

La *qualité de service (QoS) de bout en bout* est un ensemble de caractéristiques ou contraintes qui doivent être satisfaites par un réseau de communications tout au long du chemin de transport du flux de trafic de la source à la destination. Généralement, les attributs de QoS sont souvent définis en termes de débit garanti, taux de perte, délai et gigue.

La *mise en correspondance de QoS ou QoS (Quality of Service) mapping* est une fonction permettant l'inter-fonctionnement de la QoS de différents types de réseaux. Elle est généralement définie par une table d'association entre les attributs de QoS issus de deux domaines réseaux hétérogènes et adjacents.

1.2 Éléments de la problématique

Les services multimédia émergents proposés aux usagers UMTS ont des exigences très strictes en matière de qualité de service (QoS) de bout en bout (bande passante, délai, etc.). Dans un contexte d'interconnexion des réseaux d'accès UMTS à travers une dorsale IP, ces services se retrouvent transportés à travers des réseaux

hétérogènes implémentant chacun leurs propres mécanismes de gestion de QoS. De ce fait, cette QoS de bout en bout ne peut être assurée que si une interopérabilité efficace entre ces mécanismes est implantée. D'une part, les mécanismes de gestion de QoS dans les réseaux d'accès UMTS ont été bien spécifiés par le projet de partenariat de troisième génération (3GPP). D'autre part, beaucoup de travaux de recherche se sont intéressés à améliorer l'approvisionnement de la QoS dans les réseaux dorsaux IP en proposant plusieurs modèles se basant sur les mécanismes de différenciation de services. Toutefois, rares sont les travaux qui se sont intéressés à l'interopérabilité entre mécanismes de gestion de QoS issus d'un accès UMTS d'un coté et d'une dorsale IP de l'autre coté. D'où le présent mémoire qui s'intéresse aux moyens permettant de réaliser efficacement l'inter-fonctionnement entre ces mécanismes dans le but d'assurer la QoS de bout en bout.

Généralement, un réseau d'accès UMTS offre un service de transport beaucoup plus onéreux que celui offert par un réseau dorsal IP, du fait du coût élevé de l'infrastructure et des liens radios souvent limités en bande passante par rapport aux coûts très bas des liens filaires. En conséquence, des mécanismes de gestion de QoS assez complexes et très stricts pouvant se baser sur un transport ATM ont été déployés dans les réseaux d'accès UMTS. Par contre, des mécanismes de gestion de QoS par différenciation de services simples et évolutifs pouvant se baser sur un surdimensionnement des liens filaires ont été déployés dans les réseaux dorsaux IP. Si on peut être sûr qu'un service aura la QoS qu'il souhaite au niveau de l'accès UMTS, on est souvent moins sûr qu'il gardera cette même QoS en traversant une dorsale IP. Ainsi, la QoS qui doit toujours être maintenue de bout en bout peut être détériorée au moment où le flux correspondant au service passe du domaine d'accès UMTS au domaine dorsal IP. Et cela du fait que les deux domaines n'ont pas les mêmes définitions des attributs de QoS. En effet, UMTS définit un large éventail de paramètres de QoS avec une très haute granularité, alors qu'une dorsale IP définit un nombre très restreint de classes de services dont les paramètres de QoS ne peuvent pas être tous garantis.

Plusieurs architectures de ce qu'on appelle l'Internet de prochaine génération basées sur la différenciation de services ont été proposées pour offrir une bonne décomposition en classes de services adaptées le plus possible aux exigences des services multimédia en émergence. De ce fait et pour assurer la QoS de bout en bout, une mise en correspondance doit être maintenue entre les paramètres de QoS les plus pertinents du domaine UMTS avec les classes de service de la dorsale IP. Toutefois, la norme UMTS telle que spécifiée par 3GPP définit une mise en correspondance se basant simplement sur les quatre classes de services UMTS, ce qui nous parait peu efficace du fait que généralement la dorsale IP offre un nombre plus important de classes de services. De plus, pour garantir une interopérabilité de la QoS, il ne suffit pas de spécifier la mise en correspondance mais aussi de définir les mécanismes de gestion des classes de services utilisées dans la dorsale IP tels que les algorithmes d'ordonnancement et leurs différents paramètres, ce qui n'a pas été présenté dans les dernières spécifications techniques de la norme.

De toute évidence, la question de la qualité de service de bout en bout des interconnexions de ces accès UMTS à travers des réseaux IP soulève un éventail de problèmes dont la résolution conditionne l'interopérabilité et l'adoption des réseaux mobiles de prochaine génération.

1.3 Objectifs de la recherche

L'objectif principal de ce mémoire est de proposer un certain nombre d'algorithmes et de modèles pour améliorer la performance des mécanismes existants d'interopérabilité entre QoS d'un accès UMTS et QoS d'une dorsale IP, en prenant en compte les caractéristiques spécifiques du trafic multimédia dans les réseaux UMTS. D'une manière plus spécifique, nous visons les objectifs suivants :

- Analyser les solutions déjà proposées dans la littérature en matière d'inter-fonctionnement de QoS, en vue d'identifier leurs faiblesses éventuelles eu égard à la manipulation du trafic multimédia ;

- Analyser les différents choix possibles parmi les attributs de QoS UMTS (classe UMTS, délai, taux de perte, débit) dans leur mise en correspondance avec les attributs de différenciation de service d'une dorsale IP ;
- Concevoir un algorithme de mise en correspondance optimisé pour le trafic multimédia en se basant sur le fait que des média de types différents peuvent générer des trafics de comportements distincts qui doivent être différenciés de manière particulière ;
- Proposer un modèle de processus modulaire et réutilisable pour la gestion du trafic vidéo qui est généralement considéré comme compliqué à modéliser ;
- Évaluer la performance des algorithmes et du modèle proposés au moyen de simulations en les comparant aux solutions existantes afin de mesurer la qualité de l'amélioration apportée par nos solutions.

1.4 Plan du mémoire

Juste après ce chapitre d'introduction, ce mémoire se poursuit avec le deuxième chapitre qui dresse un état de l'art sur ce qui a été fait concernant la QoS UMTS et la QoS IP, ainsi que les travaux antérieurs qui ont traité de l'interopérabilité entre ces deux réseaux. Dans le troisième chapitre, nous introduisons notre solution de mise en correspondance raffinée pour le trafic de téléphonie multimédia avec une étude analytique sur les performances d'ordonnancement de ce type de trafic critique, et nous présenterons par la suite les modèles de trafic de voix et de vidéo sur lesquels se base notre solution. Dans le quatrième chapitre, nous présenterons en détails l'implantation de notre algorithme de mise en correspondance ainsi que l'implantation de notre modèle de génération de trafic vidéo téléphonie qui sera utilisée pour évaluer la performance de nos algorithmes et modèle. Enfin, le chapitre de conclusion servira à faire une synthèse globale du travail de recherche réalisé dans le cadre de cette maîtrise, en mettant l'accent sur les principales contributions apportées par notre travail et des directions de recherche possibles pour des travaux futurs.

CHAPITRE 2
ARCHITECTURES ET MÉCANISMES DE QOS DE BOUT EN BOUT DANS LES RÉSEAUX 3G

Durant ces dernières décennies, on assistait à une croissance très importante de deux secteurs technologiques : l'Internet et les communications mobiles sans fil. Selon l'Union Internationale des Télécommunications (ITU), il y a approximativement 600 millions d'utilisateurs d'Internet à la fin de l'année 2002. Une croissance de même ampleur a touché les réseaux cellulaires mobiles. Bien qu'ils aient tous les deux un succès remarquable auprès des usagers, ces deux technologies ont été souvent vues séparément l'une de l'autre. Ceci est dû principalement à leur nature différente : l'Internet a été conçu pour transporter du trafic de données alors que les réseaux cellulaires sans fil ont été initialement conçus pour transporter la voix. Ces dernières années, la frontière entre ces deux technologies commence à s'estomper, surtout avec l'introduction de plusieurs propositions pour les télécommunications mobiles internationales IMT-2000 dans l'ITU [3GPP, 2002a]. Cela a conduit à l'apparition des systèmes réseautiques mobiles des prochaines générations qui sont supposés être des plates-formes multiservices supportant voix, vidéo et services de données à des hauts débits. L'évolution des réseaux mobiles dans le sens d'offrir cette multitude de services hybrides aux usagers mobiles n'a pas pu réussir sans l'introduction du support de la qualité de service (QoS). Ainsi, dans ce chapitre, nous commencerons par donner un aperçu sur les spécifications de l'architecture de la QoS des accès de troisième génération selon la norme UMTS. Ensuite, nous décrirons les différentes architectures de QoS utilisées dans les réseaux IP en discutant le meilleur choix pour un réseau dorsal basé IP vu comme une plate-forme d'interconnexion globale des réseaux d'accès

UMTS. Enfin, nous présenterons le modèle d'inter-fonctionnement de la QoS d'un réseau d'accès UMTS et celle d'un réseau dorsal IP selon les spécifications techniques de 3GPP en mettant l'accent sur la fonction de translation et de mise en correspondance des paramètres de QoS de chacun de ces deux domaines réseau.

2.1 QoS dans les réseaux UMTS

L'introduction d'un nouveau réseau de cœur à commutation de paquets dans le système cellulaire assure, en plus de l'utilisation optimale des ressources, le transport des nouveaux services multimédia à faible coût. Mais, pour assurer le bon fonctionnement de ces services émergents, il est indispensable de fournir une plate-forme de support de QoS dont l'architecture sera détaillée dans ce qui suit.

2.1.1 Architecture de la QoS dans les réseaux UMTS

La QoS n'est rien d'autre qu'un ensemble de besoins de service qui doivent être satisfaits par le réseau tout au long du transport des flux de trafic de la source à la destination. Généralement, les attributs de QoS sont définis en termes de débit binaire garanti, taux d'erreur binaire, délai et gigue. Souvent, la QoS s'effectue par allocation de ressources, ce qui va introduire la notion de gestion de ressources en utilisant des services support, des protocoles de réservation ainsi que des mécanismes de différenciation.

Contrairement aux services de liaison de données qui sont généralement considérés point à point ou point à multipoint, les services réseau sont établis de bout en bout, c'est-à-dire, d'un équipement terminal (*TE*) à un autre. Le flot de communication de bout en bout de ces services réseau traversent différents types de réseaux offrant différents QoS à l'usager.

Pour assurer un certain niveau de QoS, des *Bearer Service* (*BS*) ou *services support* avec des paramètres et des fonctionnalités bien définis, doivent être établis entre la source et la destination d'un service réseau donné. Un service support doit avoir la capacité de fournir la QoS contractée entre un usager et un réseau ou entre un domaine

réseau et un autre domaine réseau adjacent. Ceci est essentiellement réalisé par des mécanismes tels que le contrôle de signalisation, le transport sur le plan usager, et les différentes fonctionnalités de gestion de QoS. Une architecture en couches des services support est illustrée à la Figure 2.1, chaque service support d'une couche spécifique offre ces services particuliers en utilisant les services fournis par les couches inférieures.

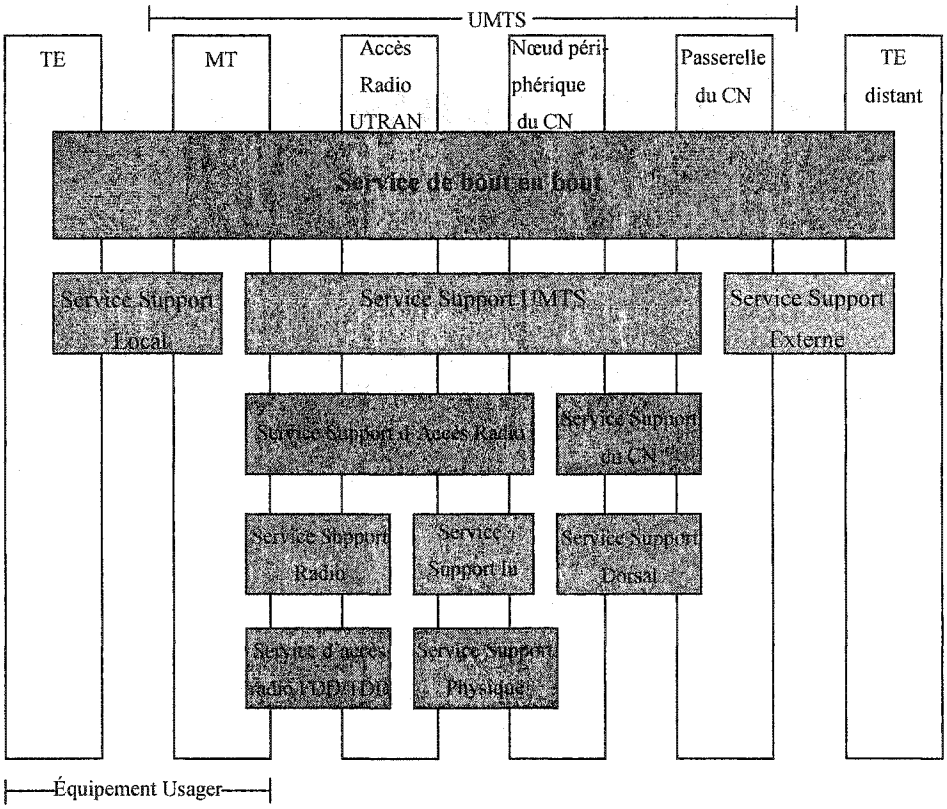


Figure 2.1 Architecture fonctionnelle de la QoS UMTS

2.1.2 Le service support UMTS

Le trafic d'un équipement terminal TE à un autre TE doit passer par différents supports de service du réseau. Le service de bout en bout de la couche application utilise une combinaison d'un ensemble de trois services support des couches inférieures : un

service support local qui réside dans le TE/MT et qui définit ses capacités locales de QoS, un service support UMTS et un service support externe. Dans cette section, nous nous intéressons au service support UMTS vu que c'est la composante qui fournit la QoS UMTS. Le service support UMTS est constitué de deux sous-composantes : le service support d'accès radio ou *Radio Access Bearer* (RAB) et le service support du réseau de cœur ou *Core Network Bearer* (CNB). Les deux services utilisent des méthodes optimisées pour fournir le service support UMTS au dessus des topologies cellulaires respectives, en prenant en considération des aspects tel que *la mobilité* et *le profil des usagers mobiles*.

La QoS côté radio assurée par le service support d'accès radio (RAB)

Le RAB assure le transport confidentiel de la signalisation et des données usager entre le terminal mobile MT et le nœud périphérique du CN (interface Iu ou point d'interconnexion entre RNC et CN) avec une QoS conforme au service support UMTS négocié. Pour cela, il utilise des techniques spécifiques comme le contrôle de puissance ou le contrôle d'admission radio qui tiennent compte des différents profils de QoS (ensemble des attributs du service support UMTS). En plus, le service RAB est basé sur les caractéristiques spécifiques de l'interface radio et doit être maintenu tout au long du mouvement d'un terminal mobile. Pour supporter différents niveaux de protection contre les erreurs, le réseau d'accès terrestre UTRAN et le terminal mobile MT ont la capacité de segmenter et de réassembler des flots d'usagers en différents sous-flots à la demande du service support radio. Ce service support radio traite différemment les sous-flots d'un même flot usager, de manière à assurer les exigences en *fiabilité* spécifiques à chaque sous-flot. Ces exigences en fiabilité, tel que le format exact d'une unité de données de service (SDU), sont sujettes à une signalisation préalable avec l'UTRAN à la phase d'établissement du RAB en utilisant des attributs standardisés. Le service support Iu, en utilisant le service support physique, offre un transport entre UTRAN et CN avec différents autres services assurant une variété de niveaux de QoS.

La QoS côté réseau assuré par le service support du réseau de cœur (CNB)

Le service support du réseau de cœur connecte le nœud périphérique du CN avec la passerelle du CN jusqu'au réseau externe. Le rôle de ce service est de contrôler et d'utiliser efficacement le réseau de cœur UMTS afin d'offrir le service support UMTS. De plus, une intégration suffisamment efficace est effectuée au niveau de toutes les couches existantes en chaque point de multiplexage traversé, c'est-à-dire dans chaque nœud UMTS du réseau de cœur. Le transfert asynchrone propre aux réseaux en mode paquet actuels perd la structure temporelle du flux et introduit un délai et une gigue aléatoires. Pour éviter les engorgements, problème important dans les réseaux à commutation de paquets, on augmente souvent la capacité des mémoires tampon au niveau des files des routeurs. Cette technique, utilisée dans les nœuds UMTS, peut toutefois introduire des retards inacceptables pour les applications en temps réel tel que la téléphonie IP. Cela nous conduit à la définition d'un système de gestion de la QoS spécifique à un réseau UMTS et dont les différentes fonctions sont détaillées dans ce qui suit.

2.1.3 Les fonctions de gestion de QoS dans un réseau UMTS

Pour assurer la prestation du service négocié au service support UMTS avec une QoS bien spécifique entre les points d'accès, le réseau UMTS offre différentes fonctionnalités de gestion classifiées en deux plans : le plan contrôle et le plan usager.

Les fonctions de gestion de QoS dans le plan de contrôle

Ces fonctions assurent l'établissement et la modification d'un service support UMTS en utilisant la signalisation et la négociation avec les services UMTS externes ainsi que l'établissement et la modification de tous les services internes avec les caractéristiques requises. Les fonctions de gestion de QoS du plan de contrôle englobent:

- *Le gestionnaire de service* qui coordonne les fonctions du plan de contrôle (établissement, modification et maintenance du service). Il fournit toutes les

fonctions de gestion de QoS du plan usager avec les attributs demandés. De plus, il peut interroger d'autres fonctions de contrôle pour recevoir la permission de fournir le service ;

- *La fonction de translation* qui effectue la conversion entre les attributs de QoS du service support UMTS et ceux des protocoles de contrôle des services du réseau externe (exemple : entre les attributs de service UMTS et les TSPEC de l'IETF). Elle peut aussi se charger de la conversion entre ses propres attributs de service et les attributs d'un service de couche inférieure qu'elle utilise (exemple : entre les attributs des services UMTS et attributs des classes ATM) ;
- *Le contrôle d'admission et de capacité* qui maintient les informations concernant toutes les ressources disponibles d'une entité réseau ainsi que toutes les ressources allouées au service support UMTS. Il détermine pour chaque requête ou modification d'un service support UMTS si les ressources demandées peuvent être fournies par l'entité. Si c'est le cas, il les alloue au service support UMTS et assure leur maintien. De plus, cette fonction vérifie la capacité de l'entité réseau à fournir le service demandé. Le contrôle des ressources effectué par le contrôle d'admission supporte également la rétention de service ;
- *Le contrôle de souscription* qui vérifie les droits administratifs de l'usager d'un service support UMTS pour l'utilisation du service demandé avec les attributs de QoS spécifiées.

Les gestionnaires du service support UMTS dans le terminal mobile MT, dans le nœud périphérique du CN ainsi que dans la passerelle du CN s'échangent des données de signalisation à travers *la fonction de translation* et avec les instances du réseau externe afin d'établir ou de modifier un support de service UMTS. Par conséquent, la QoS contractée peut être fournie au service de bout en bout dans le réseau UMTS. Dans le cadre de notre travail, nous exprimerons un intérêt particulier pour cette fonction de translation dans son utilisation par le gestionnaire du service support UMTS *au niveau du nœud passerelle* GGSN pour effectuer la traduction entre les attributs du service

support UMTS et ceux du service support externe. Dans notre cas, le réseau externe sera la dorsale IP multiservices utilisant des protocoles de support de QoS qui seront présentés plus loin dans ce chapitre.

Les fonctions de gestion de QoS dans le plan usager

Ces fonctions assurent la prestation de la QoS négociée pour un service support UMTS en maintenant le trafic des données usager dans les limites définies par des attributs de QoS signalées à l'avance. Les fonctions de gestion de QoS du plan usager englobent:

- *La fonction d'association* qui fournit à chaque unité de données un marquage spécifique au moment de son transfert par le service de support et lui permettant de recevoir la QoS contractée ;
- *La fonction de classification* qui attribue les unités de données aux différents services établies pour un terminal mobile MT selon leurs attributs de QoS relatifs. Le service support UMTS approprié est dérivé à partir de l'entête de l'unité de données ou à partir des caractéristiques du trafic des données ;
- *Le gestionnaire de ressources* qui partage et distribue les ressources disponibles aux différents services suivant leurs besoins en QoS. L'ordonnancement, la gestion de bande passante et le contrôle de puissance pour le support radio sont des exemples de gestion de ressources ;
- *Le conditionneur de trafic* qui assure la conformité entre la QoS négociée pour un service et le trafic des unités de données correspondant. Le conditionnement du trafic est réalisé par des mécanismes de réglementation et/ou de mise en forme du trafic (*policing* et/ou *shaping*). La réglementation du trafic se fait en marquant les unités de données qui ne correspondent pas avec les attributs de QoS appropriés, et en les rejetant dans le cas de congestion. La mise en forme du trafic se fait en accord avec les attributs de la QoS contractée.

2.1.4 Les classes de QoS de l'UMTS

Le projet de partenariat de troisième génération définit quatre classes de QoS [3GPP, 2002b] pour l'UMTS : la classe conversationnelle, la classe d'écoulement ou à flux continu, la classe interactive et la classe d'arrière-plan. Le facteur distinctif principal de ces classes est la sensibilité du trafic aux délais. Les caractéristiques de ces classes sont définies au Tableau 2.1. La classe conversationnelle est conçue pour les trafics les plus sensibles aux délais, alors que la classe d'arrière-plan est la classe de trafic la moins sensible aux délais. La classe conversationnelle et la classe à flux continu sont généralement prévues pour acheminer un flux temps réel (dit non élastique) et le niveau de sensibilité au délai distingue entre les deux. En effet, les services conversationnels tel que la voix et la vidéo téléphonie sont les applications les plus sensibles aux délais et doivent être acheminés dans la classe conversationnelle. La classe interactive et la classe d'arrière-plan sont généralement utilisées pour les applications traditionnelles, dites élastiques, comme le WWW, le courriel, Telnet, FTP et les News. En raison des contraintes de délais moins strictes comparées aux classes conversationnelles et à flux continu, ces deux classes offrent un meilleur taux d'erreur en utilisant des mécanismes avancés de retransmissions et de codage de canal.

La distinction entre classe interactive et classe d'arrière-plan assure un temps de réponse plus court pour les applications interactives tel que la navigation web. De plus, le trafic interactif a une plus haute priorité que celle du trafic d'arrière-plan au niveau des mécanismes d'ordonnancement, et les applications générant un trafic d'arrière-plan utilisent les ressources de transmission seulement si les applications interactives n'en ont pas besoin. Ceci est très important dans un environnement sans fil où la bande passante est très limitée.

Tableau 2.1 Classes de trafic UMTS

Type de trafic	*Délai de transmiss-ion*	*Variation du délai*	*Faible taux d'erreurs sur les bits*	*Débit binaire garanti*	*Exemple*
Conversationnel	Stricte	Stricte	Non	Oui	VoIP, visioconférence, audioconférence
À flux continu	Limité	Limitée	Non	Oui	Services de diffusion (audio, vidéo), actualités, sports
Interactif	Limité	Non	Oui	Non	Navigation sur le web, cyber-bavardage, jeux, commerce mobile
D'arrière-plan	Non	Non	Oui	Non	Courriel, SMS, téléchargements de BDs, transfert de mesures

La classe conversationnelle

La voix téléphonique comme celle du GSM est le service d'utilisation le plus connu pour cette classe. Toutefois, avec l'émergence des services multimédia sur Internet, beaucoup d'autres nouvelles applications peuvent profiter de cette classe, comme la voix sur IP et la vidéo-conférence. La conversation temps réel est souvent effectuée entre des paires de terminaux humains. C'est le seul schéma pour lequel les caractéristiques requises sont données exclusivement par la perception humaine. En effet, le délai maximal de transfert est sujet à la perception humaine de la conversation vidéo et audio. Par conséquent, la limite du délai de transfert pour cette classe est non seulement significativement basse mais aussi plus stricte que le délai d'aller-retour du trafic de la classe interactive.

La classe à flux continu ou à écoulement

Cette classe est prévue pour les flux temps réel audio ou vidéo. Généralement, un flux temps réel est transféré à l'intention d'une destination ayant une présence humaine et, contrairement à la classe conversationnelle dont le flux de données est bidirectionnel, le flux de données de la classe d'écoulement est unidirectionnel. Ce schéma de trafic est

l'un des nouveaux venus dans les réseaux de communications soulevant un certain nombre de nouvelles exigences non seulement dans les réseaux de communication mais aussi dans les systèmes de télécommunications. La variation de délai d'un flux de bout en bout doit être limitée, afin de préserver la relation temporelle (variation) entre les entités de données du flux, malgré qu'il n'y ait aucune exigence sur le niveau exact du délai de transfert. Toutefois, comme le flux est temporellement aligné au bout récepteur (par des techniques de mise en mémoire tampon dans l'équipement usager), la plus haute variation de délai acceptable à travers le médium de transmission est donnée par la capacité de la fonction d'alignement temporel au niveau de l'application. Ainsi, la variation de délai acceptable est beaucoup plus importante que celle exigée par les limites de la perception humaine.

La classe interactive

Le schéma de cette classe s'applique lorsqu'une machine ou un usager humain lance une requête de données vers un équipement tel qu'un serveur web. Le trafic interactif est un autre schéma classique des communications de données caractérisé essentiellement par le délai d'aller-retour d'une requête réponse. Une autre caractéristique de ce genre de trafic est la transparence dans le transfert du contenu des paquets de données en assurant un taux d'erreur binaire (*BER*) très faible.

La classe d'arrière-plan (background)

Ce schéma s'applique quand l'usager ou encore une machine envoie et reçoit des fichiers de données en arrière-plan. Courriels, SMS, téléchargement de base de données et réception d'enregistrements de mesures sont quelques-uns des différents services qui peuvent être délivrés par la classe d'arrière-plan. Ce genre de trafic est caractérisé essentiellement par le fait que la destination n'est pas en attente d'une réponse jusqu'à un certain temps. Ce qui fait que le trafic de cette classe est le moins sensible aux délais. Toutefois, la transparence dans le transfert du contenu des paquets doit être assurée par des mécanismes de contrôle d'erreurs.

2.1.5 Les attributs de QoS du service support UMTS

Ces attributs décrivent le service fourni par le réseau UMTS à l'usager d'un service support UMTS. Un ensemble d'attributs de QoS ou encore un profil de QoS définit ce service. En effet, à l'établissement ou à la modification d'un service support, on doit tenir compte de la disponibilité de plusieurs profils de QoS:

- *La classe de trafic* (conversationnelle, à écoulement, interactive ou d'arrière plan) : le type d'application pour lequel le service support UMTS est optimisé ;
- *Le débit binaire maximal* (en Kbps) : le nombre maximal de bits fournis par le réseau UMTS en un point du réseau tout au long d'une période de temps donnée, divisé par la durée de cette période ;
- *Le débit binaire garanti* (en Kbps) : le nombre garanti de bits fournis par le réseau UMTS en un point du réseau tout au long d'une période de temps donnée, divisé par la durée de cette période. Les attributs de délai et de fiabilité discutés plus loin ne sont garantis que si le trafic n'excède pas le débit binaire garanti ;
- *L'ordre de livraison* (oui/non) : indique si le support UMTS doit livrer les SDU dans le bon ordre de séquence ou non ;
- *La taille maximale du SDU* (octets) : la taille maximale permise d'une unité de données de service ;
- *Le taux d'erreur des SDU:* indique la fraction de SDUs perdues ou erronées. Cet attribut n'est défini que pour le trafic conforme ;
- *Le taux d'erreur binaire résiduel:* indique le taux d'erreur binaire indétectable dans les SDUs livrées ;
- *Livraison des SDUs erronées* (oui/non/-) : indique si les SDUs erronées sont livrées ou rejetées ;
- *Le délai de transfert* (msec) : indique le délai maximal du $95^{ème}$% de la distribution du délai pour toutes les SDUs livrées durant la durée de vie du support de service (5% des SDUs livrées pendant cette durée peuvent ne pas satisfaire cette contrainte de délai maximal). Le délai d'une SDU est défini comme le temps écoulé entre la requête de transfert du SDU en un bout du réseau et sa livraison à l'autre bout du réseau ;

- *La priorité de traitement du trafic:* spécifie l'importance relative du traitement de toutes les SDUs appartenant à un support UMTS par rapport à ceux appartenant à d'autres supports ;

- *La priorité d'allocation et de rétention:* spécifie l'importance relative de l'allocation et de la rétention d'un support UMTS par rapport aux autres ;

-*Le descripteur de statistiques sur la source* (voix/inconnu) : spécifie les caractéristiques de la source des SDUs.

Les attributs du support UMTS définis précédemment ainsi que leur pertinence pour chaque classe de trafic sont résumés au Tableau 2.2. Pour la classe conversationnelle et la classe à flux continu, malgré que le débit binaire de la source puisse varier, le trafic est supposé être relativement dépourvu de rafales excessives. De ce fait, il est significatif de garantir un délai de transfert pour chaque SDU. De plus, pour ces deux classes, l'information de format de SDU est utilisée dans le cas où le mode RLC (Radio Link Control) transparent est activé, ce qui permet de réduire les PDUs de leur entête. Pour la classe interactive, la priorité de traitement du trafic permet de différencier entre plusieurs qualités de service pour les services support de cette classe.

Tableau 2.2 Attributs de QoS du service support UMTS

Classe de trafic	*Trafic Conversationnel*	*Trafic À flux continu*	*Trafic Interactif*	*Trafic D'arrière-plan*
Débit maximal	X	X	X	X
Livraison dans le bon ordre	X	X	X	X
Taille maximale d'une SDU	X	X	X	X
Information de format de SDU	X	X		
Taux résiduel d'erreur sur les bits	X	X	X	X
Remise de SDUs erronées	X	X	X	X
Délai de transfert	X	X		
Débit binaire garanti	X	X		
Priorité de traitement du trafic			X	
Priorité d'allocation/rétention	X	X	X	X

2.2 La QoS dans les réseaux dorsaux IP

À sa naissance, le protocole IP a été conçu pour n'offrir aucune garantie de QoS, du fait qu'on attendait seulement qu'il donne un service « au mieux » ou *best effort*. Quand un lien est congestionné dans un réseau IP, les paquets qui font déborder les files des routeurs sont souvent rejetés. Comme le réseau traite les paquets de la même façon, n'importe quel flux peut être affecté par une congestion. Le service « au mieux » avec sa simplicité et sa gratuité a facilité la croissance exponentielle de l'Internet vers un système global planétaire en popularisant des applications tel que le courriel, la navigation Web et le transfert de fichiers. Toutefois, contrairement aux applications qui tolèrent une grande variation de délais ou des pertes de paquets, les nouvelles applications multimédia en récente émergence ont des besoins beaucoup plus stricts qui ne peuvent pas être satisfaits par ce service basé sur le concept «même service pour tous». De ce fait, une solution communément utilisée consiste à étendre les potentialités d'Internet avec des mécanismes de différenciation de services, dans le but d'offrir un niveau de service plus élevé aux applications qui en ont besoin en acceptant différents accords avec l'opérateur du réseau à des coûts plus ou moins élevés. Ces considérations ont conduit au développement rapide par l'IETF de deux standards pour assurer la QoS: le premier étant le modèle de services intégrés associé au protocole de réservation de ressources (IntServ/RSVP) et le deuxième le modèle des services différenciés (DiffServ). Dans ce qui suit, nous présenterons l'architecture détaillée de ces deux modèles et nous discuterons la faisabilité de leur déploiement sur une dorsale IP multiservices.

2.2.1 L'architecture d'intégration de services IntServ et ses limitations

Le modèle IntServ [Braden et al., 1994] définit des mécanismes qui contrôlent le niveau de QoS fourni par le réseau à des applications nécessitant une garantie de service beaucoup plus stricte que celle fournie par le service « au mieux » traditionnel. Le

contrôle de QoS de bout en bout dans ce modèle est basé sur une approche « par micro flux ». En effet, chaque micro flux est traité séparément des autres dans chaque routeur le long du chemin de communication. L'architecture IntServ suppose que des mécanismes d'établissement de services sont utilisés explicitement pour transmettre les informations de QoS aux routeurs impliqués dans un chemin origine/destination. Ces mécanismes permettent à chaque flux de demander un niveau de QoS particulier, le protocole de réservation de ressources RSVP [Wroclawski, 1997a] étant le plus utilisé de ces mécanismes.

Les niveaux de QoS offerts

L'approche IntServ spécifie en plus du service « au mieux » deux services de contrôle de QoS : le service garanti (*GS*) [Shenker et al., 1997] et la charge contrôlée (*CL*) [Wroclawski, 1997b].

- *Le service garanti*: assure un niveau garanti de bande passante, fixe des limites strictes aux délais de bout en bout et ne rejette pas les paquets dans le cas de débordement des files. Ce service est conçu pour les applications très sensibles aux délais tel que la vidéo conférence ;
- *Le service à charge contrôlée* : sous ce service, les applications sont traitées de la même façon que dans le cas du service « au mieux » avec une charge réduite. Ce service est conçu pour les applications temps réel adaptatives qui peuvent tolérer une variation de délai tel que l'écoulement vidéo ;

Le modèle architectural

Dans cette architecture, RSVP est utilisé pour réserver les ressources dans les routeurs qui sont sur le chemin du flux de trafic. Pour maintenir cette réservation tout au long du chemin, un « état logiciel » est utilisé en opposition avec un « état matériel » fourni par les circuits virtuels tel que ceux dans ATM. Quand cet état logiciel est établi, les nœuds s'envoient des messages « PATH » et « RESV » périodiquement (toutes les 30 secondes) pour rafraîchir l'état du chemin et de la réservation (Figure 2.2). Si aucun

message de rafraîchissement n'est reçu avant l'écoulement d'une certaine période, l'état de réservation est supprimé. L'état peut être aussi supprimé en émettant un message explicite « RESVTEAR » ou « PATHTEAR » de fin de session. Les états manipulés par RSVP sont définis par les services GS et CL.

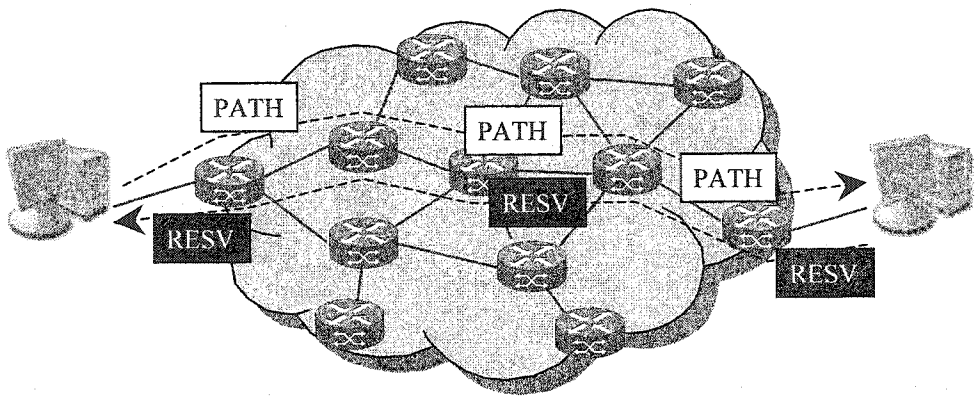


Figure 2.2 Modèle architectural de IntServ/RSVP

Un message « PATH » est initié par l'expéditeur et transporte trois types d'informations :

- *Les spécifications du saut RSVP précédent* (adresse IP et numéro d'interface logique) : utilisées pour router les messages « RESV » en amont du saut précédent ;
- *Un TSpec* : décrit le trafic de l'expéditeur, utilisé pour prévenir la sur réservation sur les liens à proximité de l'expéditeur ;
- *Un AdSpec* : utilisé pour mesurer les propriétés du chemin de données.

Un message « RESV » est périodiquement initié par le destinataire pour demander la réservation et il est acheminé en amont du flux. Une demande de réservation est référée par un descripteur de flux et comprend :

- *Un FlowSpec* : spécifie la QoS désirée et utilisé pour configurer les paramètres au niveau de l'ordonnanceur de paquets d'un nœud ;

- *Un FilterSpec* : définit l'ensemble des paquets pour lesquels une réservation a été demandée.

Description des composants IntServ

Comme décrit à la Figure 2.3, l'implémentation de IntServ exige que les éléments réseaux contiennent les modules suivants :

- *Un protocole de routage* : maintient une table de routage utilisée pour déterminer le saut suivant à prendre pour se rendre à une certaine destination ;
- *Un processus RSVP* : traite les spécificités de la signalisation RSVP et maintient les informations d'états des micro flux ;
- *Des états RSVP* ;
- *L'agent de gestion* : modifie la base de données de contrôle de trafic et dirige le module de contrôle d'admission afin d'établir les règles de contrôle d'admission ;
- *Le contrôle d'admission* : détermine s'il y a suffisamment de ressources disponibles pour assurer la QoS demandée à partir d'informations récoltées sur la charge du réseau ;
- *Le classificateur* : classifie les paquets IP en flux ayant des caractéristiques de QoS définies et identifiées par adresse source, adresse destination, protocole, port source, port destination ;
- *Le gestionnaire de réglementation* : applique des règles selon les caractéristiques du trafic fourni. Il détermine quel flux excède ses attributs de QoS et décide quel traitement il doit appliquer au trafic hors profil (application du rejet par défaut) ;
- *L'ordonnanceur* : gère une ou plusieurs files pour chaque port de sortie et effectue des décisions de mise en file suivant la classe du paquet, le contenu de la base de données de contrôle de trafic et l'activité courante et antérieure du port de sortie.

Discussion sur les limitations et le choix de IntServ

Plusieurs facteurs ont empêché le déploiement de IntServ/RSVP dans les dorsales de l'Internet [Iera et al., 2002]:

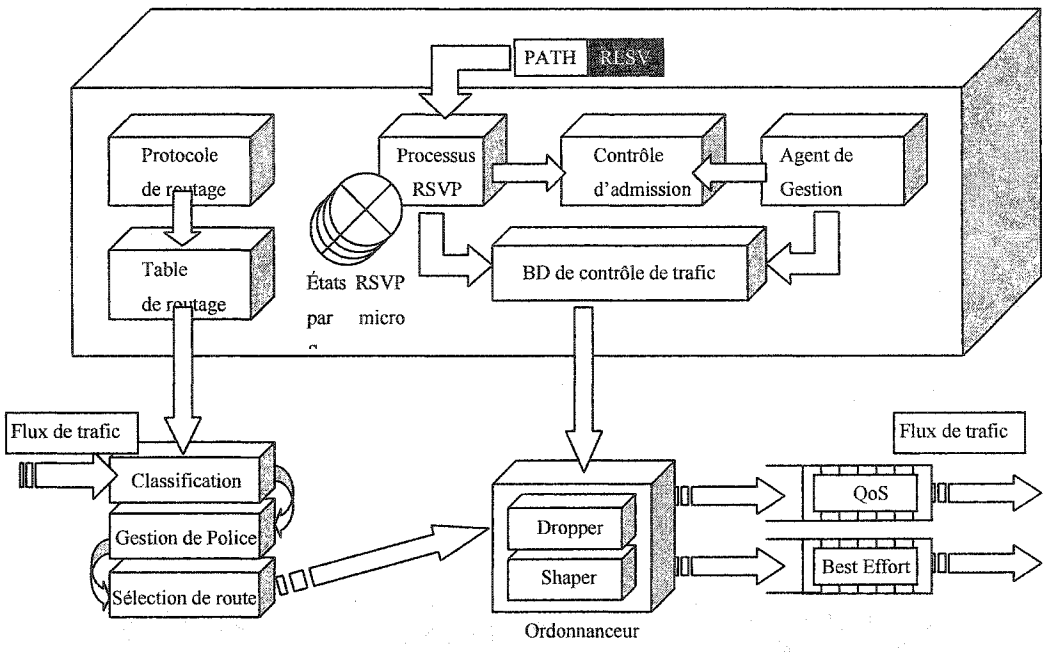


Figure 2.3 Modules d'un routeur IntServ/RSVP

- L'utilisation d'un état par micro flux et d'un traitement granulaire sur ces micros flux soulèvent des problèmes d'évolutivité pour les réseaux de grande taille. En effet, d'après Mankin et al. [1997], les ressources requises par RSVP pour le traitement et le stockage dans un routeur augmentent proportionnellement avec le nombre de sessions IntServ. Par conséquent, maintenir un très grand nombre de réservations pour des micros flux sur un lien haut débit peut s'avérer extrêmement coûteux en temps de traitement et en espace mémoire ;
- Dans ce modèle, tous les équipements du réseau doivent garder un état par micro flux réservé. Il suffit qu'un nœud dans la route n'implémente pas les fonctionnalités IntServ pour que la QoS ne puisse plus être strictement garantie ;
- Actuellement, un nombre très restreint de stations usagers génèrent une signalisation RSVP. Malgré que ce nombre soit prévu à la hausse, surtout avec la disponibilité de Generic QoS Winsock API 2 de Microsoft, plusieurs applications peuvent ne jamais implanter cette API et ne jamais générer cette signalisation ;

- Un autre reproche fait au modèle IntServ est la complexité du protocole de signalisation RSVP. Une grande partie de la lourdeur du protocole est due à la gestion des flux multicast et des routes symétriques. La réservation de ressources pour des flux multicast exige la définition de règles d'agrégation et de désagrégation dans les noeuds intermédiaires.

Cependant, même si RSVP, dans l'approche IntServ, recèle des limitations dans son déploiement sur les dorsales IP à large échelle, il constitue une solution qui a bien des avantages pour contrôler la QoS, surtout si elle est appliquée à des réseaux de petites tailles tel que les réseaux d'accès sans fil où la bande passante est très onéreuse. En effet, dans un réseau d'accès UMTS, il est possible d'utiliser la signalisation RSVP (dans le cas où l'équipement terminal implante le mécanisme) pour permettre à l'application de mieux exprimer ses besoins en QoS aux couches inférieures (notamment la couche d'accès radio). De plus, RSVP permet d'assurer un contrôle de la QoS à travers la dorsale d'interconnexion par des mécanismes d'agrégation et de mise en correspondance IntServ/DiffServ au niveau du nœud passerelle GGSN.

2.2.2 Architecture de différenciation de services DiffServ

L'architecture des services différenciés DiffServ [Blake et al., 1998] contourne les problèmes d'évolutivité de RSVP/IntServ en réduisant le trafic à un nombre très restreint d'agrégats, chacun avec un ensemble de besoins de QoS différents. Contrairement à RSVP/IntServ, dans le modèle DiffServ, il n'existe pas de signalisation dans le cœur du réseau : seule une étiquette dans l'en-tête, le point de code DiffServ ou *DSCP*, est utilisée pour assurer le traitement différencié. Cette capacité permet aux routeurs de garder un mode de fonctionnement relativement simple qui n'affecte pas considérablement leur capacité d'acheminement. La complexité est reportée dans les routeurs de frontière, se trouvant à l'entrée ou à la sortie d'un domaine DiffServ [Blake et al., 1998]. Ces équipements sont chargés de déterminer la valeur de l'étiquette de chaque paquet en fonction d'un contrat, du service demandé et des mécanismes de

contrôle de trafic. Le DSCP occupe les six bits de poids fort du champ DS (DiffServ) de l'en-tête IP. La Figure 2.4 montre l'emplacement de ce champ. En IPv4, il se substitue à l'ancien champ *ToS*; en IPv6, il se situe à la place du champ *Traffic Class*.

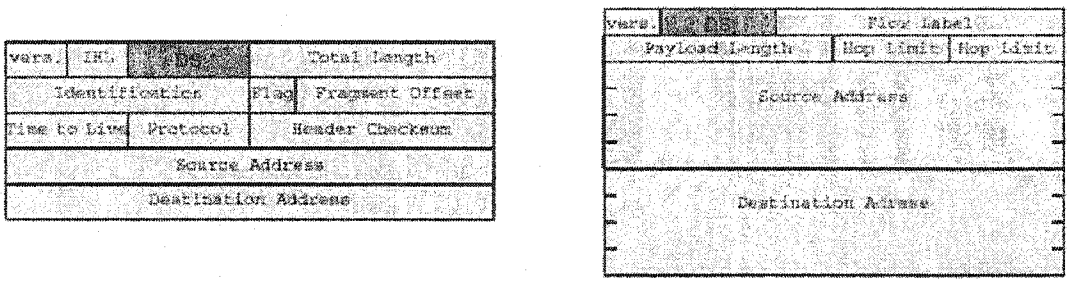


Figure 2.4 Champ DiffServ dans l'entête IP

Les niveaux de QoS offerts

En utilisant le champ DSCP, DiffServ sélectionne le comportement par nœud ou *per-hop behaviour* (*PHB*) qu'un paquet subit à chaque nœud. Un PHB est un traitement d'acheminement de paquet qui est généralement défini avec un poids relatif de bande passante et/ou une priorité relative de rejet de paquet. Ainsi, on peut utiliser les PHBs pour construire des services de différente qualité. De plus, du service « au mieux » et du service de contrôle du réseau, l'IETF a conçu deux grands modèles de service (Tableau 2.3): le service *Premium* basé sur le PHB de l'acheminement expédié et le service *Assuré* basé sur le PHB de l'acheminement assuré.

- *Le service Premium* : le PHB de l'acheminement expédié (*EF*) [Jacobson et al., 1999] spécifie la classe de paquets que le routeur doit acheminer suivant un taux de transfert particulier, indépendant du taux de transfert de tous les autres PHBs implémentés dans ce routeur. En général, l'implémentation doit être aussi simple qu'une file prioritaire qui est toujours servie avant les autres, ce qui permet d'offrir un service de ligne spécialisée virtuelle louée. Une source de trafic négocie avec un ou plusieurs réseaux DiffServ le débit binaire dont il a besoin. Si la négociation est réussie, on établit un contrat SLA (*Service Level Agreement*) qui permet de donner à la source la permission de transmettre

jusqu'au débit binaire requis qui doit être garanti à travers tous les routeurs entre les deux extrémités du domaine DiffServ. Pour garantir que le taux de l'agrégat des paquets EF acheminés est supérieur au taux de l'agrégat arrivé au routeur, un certain nombre de mécanismes d'ingénierie de trafic doivent être mis en place. En particulier, (1) le réseau DiffServ doit être sur-provisionné en accord avec le trafic EF admis, (2) un contrôle d'admission doit être effectué aux extrémités du réseau DiffServ, (3) les routeurs d'extrémités doivent mettre en forme le trafic par limitation de débit binaire pour que le débit binaire maximal négocié ne soit jamais dépassé. Des résultats expérimentaux montrent que EF offre un service de bout en bout avec faibles pertes, faibles délais et faibles gigues, ce qui est adéquat pour des applications tel que la téléphonie IP.

Tableau 2.3 Codes DSCP de DiffServ

PHB	DSCP	RFC
Contrôle Réseau	110000 111000	2474
EF	101110	2598
AF1[1..3]	001010 001100 001110	2597
AF2[1..3]	010010 010100 010110	
AF3[1..3]	011010 011100 011110	
AF4[1..3]	100010 100100 100110	
BE	000000	2474

- *Le service Assuré* : le groupe du PHB de l'acheminement assuré (*AF*) [Heinanen et al., 1999] définit 12 points de code (*DSCP*): 4 classes d'acheminement, chacune avec 3 classes de précédence de rejet (Tableau 2.3). Dans les conditions de fonctionnement normal, un routeur implantant le mécanisme doit acheminer un paquet de la classe i avant un paquet de la classe j, si $i<j$. De plus, dans le cas de congestion, un routeur doit rejeter un paquet avec précédence m avant un paquet avec précédence n, si $m<n$. Plusieurs mécanismes d'ordonnancement pour la distribution des ressources aux

différentes classes ainsi que des mécanismes de gestion de files d'attente pour le traitement (tel que le rejet de paquet) au sein d'une même file ont été développés et peuvent être implantés pour les services AF. Comme exemples de mécanismes d'ordonnancement, on peut citer les algorithmes PQ [Zhang et al., 1993], WFQ [Jamaloddin, 1996], CBQ [Floyd et al., 1995], et comme exemples de mécanismes de gestion de files d'attentes on peut citer le traditionnel FIFO, RED [Floyd et al., 1993], WRED [Cisco, 1999]. Pour le service AF, les fournisseurs implantent souvent des classes de différents niveaux relatifs de QoS en terme de délai moyen et de taux de rejet de paquets, sans autant offrir une garantie absolue sur cette QoS.

- *Le service de contrôle du réseau* : le PHB correspondant est conçu afin de fournir la QoS nécessaire aux messages critiques du contrôle du réseau tel que les messages de contrôle ICMP ;

- *Le service « au mieux »* : le service par défaut fourni par Internet.

Le modèle architectural

Avant qu'un paquet entre dans un domaine DiffServ (Figure 2.5), son champ DSCP est marqué par le routeur d'entrée suivant la QoS qui va lui être fournie et qui a été éventuellement négociée dans le cadre d'un contrat SLA. Dans le domaine DiffServ, chaque routeur a besoin de regarder seulement le DSCP pour décider quel traitement il doit effectuer sur le paquet. Aucune classification complexe d'un état par micro flux n'est nécessaire.

Description des composants DiffServ (Figure 2.6)

- *Le module de classification* : dans les routeurs d'entrée, cette opération est similaire à celle définie dans le modèle IntServ. En effet, elle utilise l'identification du flux (adresse source, adresse destination, protocole, port source, port destination) pour le classifier. Pour les routeurs du cœur du domaine DiffServ, une classification plus simple se fait seulement à partir du champ DSCP ;

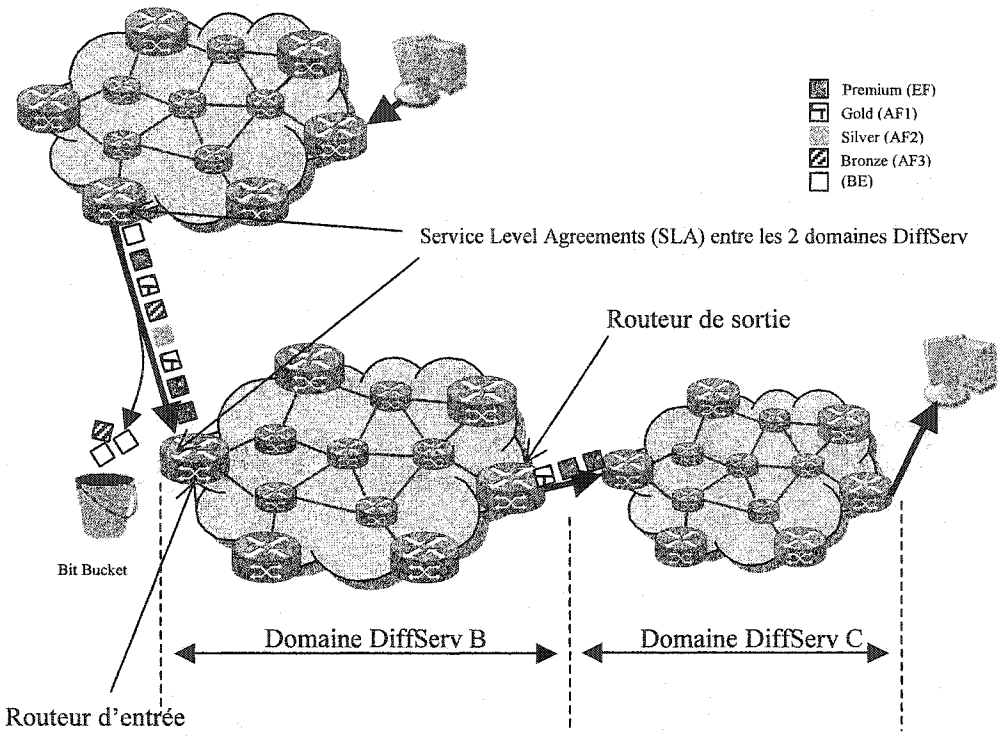


Figure 2.5 Modèle architectural de DiffServ

- *Le module de vérification* : un vérificateur est chargé de déterminer le niveau de conformité pour chaque paquet du flux arrivant dans le routeur. Cette valeur dépend du comportement instantané du flux et des caractéristiques du contrat ;
- *Le module d'action* : l'élimination (*drop*), la mise en forme (*shape*) et le marquage (*mark*) sont les principales actions du module. L'élimination agit pour contrôler sévèrement le débit d'émission en éliminant les paquets non conformes au contrat de trafic. La mise en forme essaye de rendre le trafic conforme au contrat en retardant l'acheminement des paquets non conformes. Le marquage est l'action qui attribue une précédence ou priorité aux paquets en fonction du résultat de la vérification ;
- *Le module d'étiquetage* : l'étiquetage permet de mettre à jour le champ DSCP de tous les paquets qui entrent dans le domaine DiffServ au niveau du routeur d'entrée. Ce module est absent pour un routeur du cœur du domaine vu qu'il ne doit pas changer le DSCP.

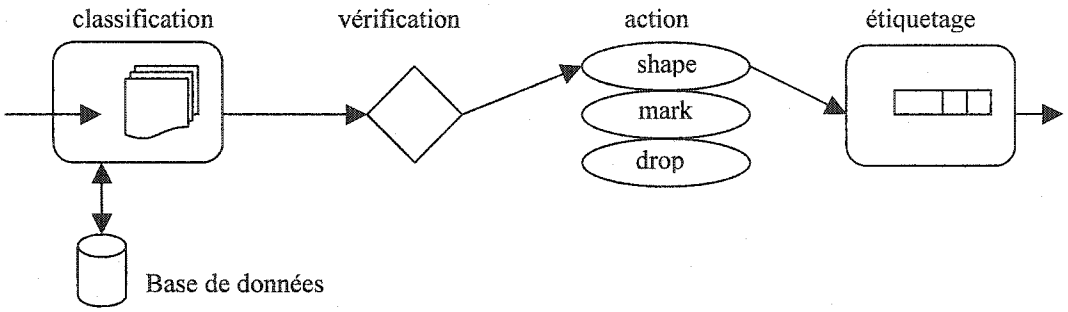


Figure 2.6 Modules d'un routeur d'entrée DiffServ

DiffServ sur MPLS comme solution de transport dans les dorsales IP

MPLS ou la commutation d'étiquettes [Rosen et al., 2001] est une technologie de commutation de couche 3 qui a été conçue dans le but d'améliorer considérablement la performance des routeurs dans l'acheminement des paquets au niveau des réseaux dorsaux de l'Internet ou de tout autre réseau à large échelle. L'idée de base est d'acheminer les paquets en se basant sur un court identificateur de taille fixe appelé *étiquette*, au lieu de l'adresse de la couche réseau qui induit une taille variable du champ en concordance avec les entrées de la table de routage. Les étiquettes sont assignées aux paquets au *nœud d'entrée (ingress)* de la dorsale MPLS. À l'intérieur de la dorsale MPLS, les étiquettes attachées aux paquets sont utilisées pour prendre les décisions d'acheminement. Ces étiquettes sont enlevées des paquets quand ces derniers sortent de la dorsale MPLS aux *nœuds de sortie* (*egress*). Les routeurs qui supportent MPLS sont appelés *routeurs à commutation d'étiquettes* ou *LSR,* et le chemin entre un nœud d'entrée et un nœud de sortie d'une dorsale MPLS qui est suivi par les paquets portant la même étiquette est appelé *un chemin à commutation d'étiquettes* ou *LSP*. Malgré que l'idée générale derrière le développement de MPLS soit de réaliser une commutation de paquets *rapide*, actuellement son but principal est de supporter *l'ingénierie de trafic* et de fournir la QoS dans les réseaux dorsaux.

Le but de l'ingénierie de trafic est de permettre le fonctionnement du réseau avec efficacité et fiabilité et en même temps d'optimiser l'utilisation des ressources du réseau.

Un autre avantage important de MPLS est l'introduction de la différenciation des services [Le Faucheur et al., 2002] avec les mêmes principes et mécanismes de DiffServ discutés dans les sections précédentes. De plus, MPLS introduit la notion de *jonction de trafic* (*trafic trunk*) qui n'est autre qu'une agrégation de flux de trafic d'une même classe, flux qui sont placés dans un même LSP. Ceci permet aux différents types de flux de recevoir des acheminements et des traitements distincts qui tiennent compte de différents niveaux de QoS. Tous les paquets d'une même jonction de trafic ont la même étiquette et le même champ classe de service *EXP* (*expérimental*) de l'entête MPLS. L'utilisation de ce champ EXP pour la différenciation de service a été introduite par l'IETF [Le Faucheur et al., 2002] pour pallier le problème de l'interdiction d'accès au champ DSCP par la couche MPLS à cause de l'encapsulation de IP dans MPLS. Toutefois, limité à une taille de 3 bits, le champ EXP (Figure 2.7) ne permet de représenter que 8 PHBs distincts.

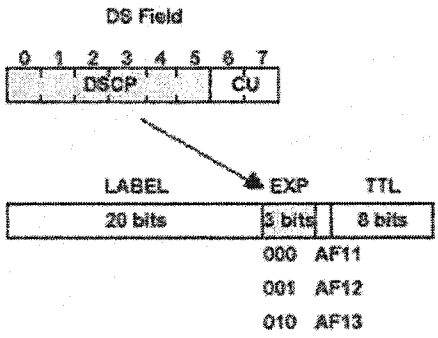


Figure 2.7 Utilisation du champ EXP pour déterminer le PHB

La mise en correspondance entre champ EXP de l'entête MPLS et champ DSCP de l'entête IP peut être faite par pré-configuration des routeurs de la dorsale ou bien par l'utilisation de l'objet DiffServ dans le message PATH du protocole de signalisation *RSVP-TE*, le protocole RSVP étendu avec des fonctionnalités permettant l'ingénierie de trafic dans un réseau MPLS.

2.2.3 Les algorithmes d'ordonnancement et leur impact sur les délais

Pour la transmission des services de téléphonie multimédia sur des dorsales IP, il est nécessaire de partager les liens dorsaux entre différents types de trafic. Pour cela, comme on l'a vu précédemment, on a le choix parmi plusieurs algorithmes d'ordonnancement et de partage des liens dorsaux. Beaucoup de travaux ont été réalisés dans le domaine du calcul des délais des paquets pour les différentes disciplines d'ordonnancement (FIFO [Elwalid, 1995], PQ [Elwalid, 1995], GPS [Parekh et Gallager, 1994], WFQ [Parekh et Gallager, 1994], EDF [Firoiu, 1997]). Souvent, en comparant les disciplines d'ordonnancement, il est nécessaire d'évaluer les aspects suivants [Elsayed, 2000]:

- capacité de l'ordonnancement: combien de connexions de chaque classe peuvent être admises pour l'ordonnancement sans violation de leurs contraintes de délai maximal toléré ;
- l'isolation et l'équité des connexions ;
- la facilité de l'implémentation et la complexité du calcul nécessaire pour effectuer l'ordonnancement.

On suppose que les connexions (sessions) sont contraintes par un seau à jetons « *leaky bucket* » (Figure I.1 en annexe) comme filtre de trafic servant à contrôler le débit et la sporadicité des sources de trafic. Toutefois, puisque le contrôle de ces sources par ce type de seau à jetons est déjà effectué au niveau du contrôle d'admission *CAC* en amont du réseau UMTS (au niveau RNC et même au niveau réseau de cœur [Throeung, 2001]), notre utilisation des seaux à jetons se limitera à la caractérisation des sources de trafic entrant la dorsale IP. Ainsi, chaque connexion *i* (session ou source de trafic) sera caractérisée par le descripteur (R_i, ρ_i, b_i), où R_i est le taux crête, ρ_i le taux moyen, et b_i la taille de la rafale la plus longue. Avec ce modèle de trafic, il est possible de calculer le délai de bout en bout au pire cas pour la majorité des disciplines d'ordonnancement et de partage de liens dorsaux.

GPS et WFQ

Le partage pondéré équitable (*WFQ*) et le processeur partagé généralisé paquet par paquet (*PGPS*) sont des approximations de la discipline de service du processeur partagé généralisé (*GPS*). Le principe du GPS est le suivant : les paquets sont servis comme s'ils sont dans des files logiques séparées, le serveur visite chaque file non vide à tour de rôle et sert une petite quantité infinitésimale de données de chaque file. Des coefficients de pondération (Φ_1, Φ_2, ..., Φ_N) peuvent être associés aux N connexions présentes et ces dernières perçoivent alors les services proportionnellement à la valeur relative de leurs coefficients tant que des données sont présentes dans leurs files d'attente associées. Dans le cas d'une file d'attente vide, le serveur assure le service de la file d'attente non vide suivante. Soit S(i,τ,t) la quantité de trafic de la session *i* (connexion *i*) servie dans l'intervalle (τ,t], A(i,τ,t) la quantité de trafic de la session *i* issue du filtre du seau à jetons (Figure 2.8) de paramètres (ρ_i, b_i) qui entre dans le réseau dans le même intervalle (τ,t]. Une session *i* est en attente de service (*backlogged*) à l'instant *t* si une quantité strictement positive du trafic qu'elle génère est placée dans sa file d'attente à cet instant. En supposant que chaque session *i* est continuellement en attente de service dans l'intervalle (τ,t], alors on a :

$$\frac{S(i,\tau,t)}{S(j,\tau,t)} \geq \frac{\Phi_i}{\Phi_j}, \qquad j = 1,\ 2,\ ...,\ N \tag{2.1}$$

Avec un serveur GPS de taux de service *r,* chaque session reçoit un taux de service minimal garanti de:

$$g_i = \frac{\Phi_i}{\sum_j \Phi_j} r \tag{2.2}$$

Ainsi, GPS fournit une équité parfaite dans l'allocation de la bande passante. Quand une source de trafic est contrainte par un seau à jetons de taille de rafale b_i et de taux de jetons $\rho_i \leq g_i$, GPS a la possibilité de garantir une borne supérieure pour le délai :

$$d_i \leq \frac{b_i}{\rho_i} \tag{2.3}$$

Pour l'ordonnanceur WFQ, son fonctionnement est comme suit : il calcule le temps pour lequel un paquet finit son service sous l'ordonnancement GPS, puis il sert les paquets dans l'ordre de leur temps de fin de service. Le calcul de ce temps est illustré dans l'article de Keshav [1997]. Pour déterminer le délai de bout en bout au pire cas, on considère une connexion contrainte par un seau à jetons (ρ_i, b_i) passant à travers L ordonnanceurs WFQ, avec le $l^{ième}$ ordonnanceur partageant un lien de capacité C_l. Notons par : $g_i = \min_l g_{i,l}$ avec $\rho_i \leq g_i$ pour la stabilité des files (sinon les files peuvent se construire à l'infini). Et par $P_{max,i}$ le paquet le plus long de la connexion i, et supposons que P_{max} soit la plus grande taille de paquet permis dans le réseau. Alors, le délai de bout en bout d_i pour un paquet de la connexion i satisfait la relation 2.4 établie par Parekh et Gallager [1994]:

$$d_i \leq \frac{b_i}{g_i} + \sum_{l=1}^{L-1} \frac{P_{\max,i}}{g_{i,l}} + \sum_{l=1}^{L} \frac{P_{\max}}{C_l} \tag{2.4}$$

et cela *indépendamment du comportement des autres sessions. Dans cette relation, le* premier terme b_i/g_i représente le temps pour servir la rafale la plus longue. Le deuxième terme est le temps de service du plus long paquet de la session quand il arrive après son tour aux L-1 serveurs WFQ. Le troisième terme est la latence de transmission sur les L liens du chemin. Il est très important de noter que, quand la vitesse des liens est très élevée par rapport à P_{max}, la borne supérieure de d_i ci-dessus se réduit à b_i/g_i, du fait que la mise en paquets est très importante pour fournir un délai de bout en bout assez faible. De plus, dans la formule, on ne tient pas compte du délai de propagation du lien puisqu'il peut être ou bien négligé pour des réseaux limités géographiquement ou bien inclus dans la bande passante C.

PQ

Un ordonnanceur par priorité statique assigne à chaque connexion (session) un niveau de priorité p, avec $1 \leq p \leq P$, P étant le nombre de niveaux de priorité. Toutes les sessions ayant le niveau de priorité p auront la même borne de délai d_p, avec $d_p < d_q$ pour $p < q$. Autrement dit, une session avec haute priorité aura une faible borne de délai. Il est très simple d'implémenter cet algorithme d'ordonnancement puisqu'il utilise un nombre fixe de files FIFO, une pour chaque niveau de priorité.

Supposons qu'on a une seule connexion à chaque niveau de priorité et que la taille minimale d'un paquet est égale à zéro. Soit P le nombre de sessions, (ρ_i, b_i, d_i) le descripteur du trafic et la borne supérieure du délai pour la connexion i, alors l'ensemble des sessions sont ordonnancables au lien l si :

$$d_i \geq \frac{\sum_{q=1}^{i} b_q + \max_{r>i} P_{\max,r}}{C_l - \sum_{q=1}^{i-1} \rho_q} \tag{2.5}$$

EDF

Le principe de l'ordonnanceur EDF est d'assigner à chaque paquet un échéancier et de servir les paquets suivant leurs échéances. Au moment de la connexion, la source déclare son taux crête (maximal) et une borne supérieure pour le délai désiré. L'ordonnanceur effectue un test d'ordonnancement pour s'assurer que le flux de chaque source rencontre sa borne supérieure de délai même si elle transmet à son taux maximal. Considérons des sources contraintes à des seaux percés (ρ_i, b_i) et à une borne supérieure de délai d_i à l'ordonnanceur l. Liebherr et al. [1996] ont montré la condition d'ordonnancement suivante :

$$d_j^l \geq \frac{b_j + \sum_{i=1}^{j-1} (b_i - \rho_i d_i^l) + \max_{k>j} P_{\max,k}}{C_l - \sum_{i=1}^{j-1} \rho_i} \tag{2.6}$$

La complexité du test d'ordonnacement de EDF est très élevée puisque la vérification de la condition de la relation (2.6) est très coûteuse en temps de calcul.

Analyse comparative des disciplines d'ordonnancement

Dans [Elsayed, 2000], on a effectué une analyse comparative de performance entre les disciplines d'ordonnancement présentées précédemment, et on a trouvé que WFQ donne les meilleures performances (pour la totalité des sessions) dans le cas d'un réseau étendu avec plusieurs sauts (*hops*). Il rivalise même avec EDF qui a été toujours considéré comme optimal localement. En effet, dans WFQ la bande passante allouée est calculée en utilisant une méthodologie qui tient compte du réseau dans sa globalité, alors que pour EDF les délais locaux sont cumulés dans chaque nœud sans tenir compte du fait que le trafic est distribué sur des chemins à sauts multiples. De plus, la complexité de EDF introduit des latences dans l'ordonnancement. Pour PQ, on a souvent constaté qu'il souffre du *problème de famine*, dans le sens que les sessions de haute priorité monopolisent le service en affectant ainsi l'équité avec les autres sessions moins prioritaires.

À l'issue de cette analyse comparative sommaire, notre choix se fixe sur la discipline d'ordonnancement par partage équitable WFQ qui est d'ailleurs largement utilisée et implantée sur la plupart des routeurs dorsaux.

2.3 Inter-fonctionnement de la QoS d'un réseau UMTS avec une dorsale IP

Dans ce qui précède, nous avons décrit les architectures et les mécanismes de QoS utilisées dans un réseau d'accès UMTS et dans un réseau dorsal IP. Toutefois, dans un contexte d'interconnexion à l'échelle de l'Internet global des différents réseaux d'accès UMTS, pour fournir une QoS de bout en bout aux usagers, il n'est pas seulement nécessaire de gérer les mécanismes de QoS spécifiques à chaque domaine, mais il est aussi primordial de fournir un inter-fonctionnement efficace des QoS des domaines

réseaux adjacents. Ceci peut être assuré en concevant et implantant une fonction de translation de QoS réalisant une mise en correspondance « optimale » entre les différents attributs de QoS issues des deux domaines réseau, UMTS et dorsal IP. Nous décrirons l'architecture d'inter-fonctionnement de la QoS et nous présenterons la mise en correspondance des attributs de QoS telle qu'elle a été définie par le projet de partenariat de troisième génération 3GPP.

2.3.1 Architecture d'inter-fonctionnement de la QoS

Vu que les ressources dans la dorsale IP ne sont pas contrôlées par le réseau d'accès UMTS et qu'ils sont indispensables pour fournir la QoS de bout en bout, il est nécessaire d'inter-opérer avec les réseaux IP externes pour contrôler ces ressources. Plusieurs mécanismes permettent de réaliser cet inter-fonctionnement:

- *Le marquage DiffServ* ou *l'étiquetage MPLS* des paquets avec sur approvisionnement ou *approvisionnement statique* de QoS en utilisant *les contrats de niveau de service SLA* qui sont appliqués par le routeur de frontière entre les deux réseaux ;
- *La signalisation de bout en bout* tout au long du chemin de trafic pour *l'approvisionnement dynamique* de QoS en utilisant des protocoles de signalisation tel que RSVP et LDP ;
- L'interaction entre *le contrôle de police et/ou les éléments de gestion de ressources* tel que les courtiers de bande passante [Reichmeyer et al., 1998] pour *un approvisionnement dynamique* de QoS.

Les éléments d'inter-fonctionnement et leur fonction de gestion de QoS

- Le gestionnaire de service support IP: Il utilise des mécanismes standards pour gérer les supports de service IP. Ces mécanismes peuvent être différents des mécanismes utilisés dans les services support (*SS*) UMTS et peuvent avoir différents paramètres de contrôle de service (DiffServ, IntServ/RSVP, ...). Au cas où il est implémenté, le gestionnaire de SS IP doit supporter la fonction de différenciation de service DiffServ au

nœud frontière et la fonction RSVP et, dans ce cas, la même fonction n'est pas nécessaire au niveau de l'équipement usager. Si des gestionnaires de SS IP existent dans l'équipement usager ainsi que dans le noeud passerelle GGSN, ces gestionnaires peuvent communiquer entre eux directement en utilisant des mécanismes de signalisation. Le Tableau 2.4 définit les fonctionnalités minimales qui doivent être supportées par l'équipement afin de permettre aux opérateurs de réseaux de fournir un inter-fonctionnement entre leurs réseaux pour une QoS de bout en bout.

Tableau 2.4 Fonctionnalités du gestionnaire du SS IP

Fonctionnalité	***Équipement Usager***	***Passerelle GGSN***
Fonction de différenciation de services DiffServ	Optionnelle	Obligatoire
Fonction RSVP /IntServ	Optionnelle	Optionnelle
Fonction d'application de la politique de service IP	Optionnelle	Obligatoire

- *La fonction de translation et de mise en correspondance:* Elle fournit l'inter-fonctionnement entre les mécanismes et les paramètres utilisés dans chacun des services support UMTS et IP en interagissant avec le gestionnaire du SS IP. Dans le GGSN, les paramètres de QoS UMTS sont mis en correspondance avec les paramètres de QoS IP et vice versa. Dans l'équipement terminal, les paramètres de QoS exigés de chaque service ou application sont déterminés à partir de la couche applicative tel qu'une description de session SIP avec SDP. Ensuite, ils sont mis en correspondance directement avec les paramètres du contexte PDP ou bien éventuellement avec les paramètres de QoS de la couche IP tel que les TSpecs de RSVP.

- *La fonction de contrôle de politique (PCF):* C'est l'élément qui permet de prendre des décisions sur la politique de contrôle de QoS en utilisant des mécanismes standards au

niveau de la couche IP. Ces mécanismes sont conformes aux spécifications de l'IETF [Yavatkar et al., 2000] où la PCF est un point de décision sur la politique de service. La PCF établit des décisions en accord avec les politiques du service IP en utilisant des règles de politique, et communique les décisions au gestionnaire du service support IP dans le GGSN à travers *le point d'application de politique PEP*. L'interface entre PCF et GGSN s'appelle l'*interface Go* [3GPP, 2002a] et elle utilise le protocole client/serveur de demande de politique de service COPS [Durham et al., 2000]. La PCF établit ses décisions de politique en se basant sur les informations obtenues à partir de la fonction de contrôle du serveur d'appels au niveau *du serveur mandataire P-CSCF*. La Figure 2.8 illustre l'inter-fonctionnement des différents éléments pour l'établissement d'une session multimédia SIP.

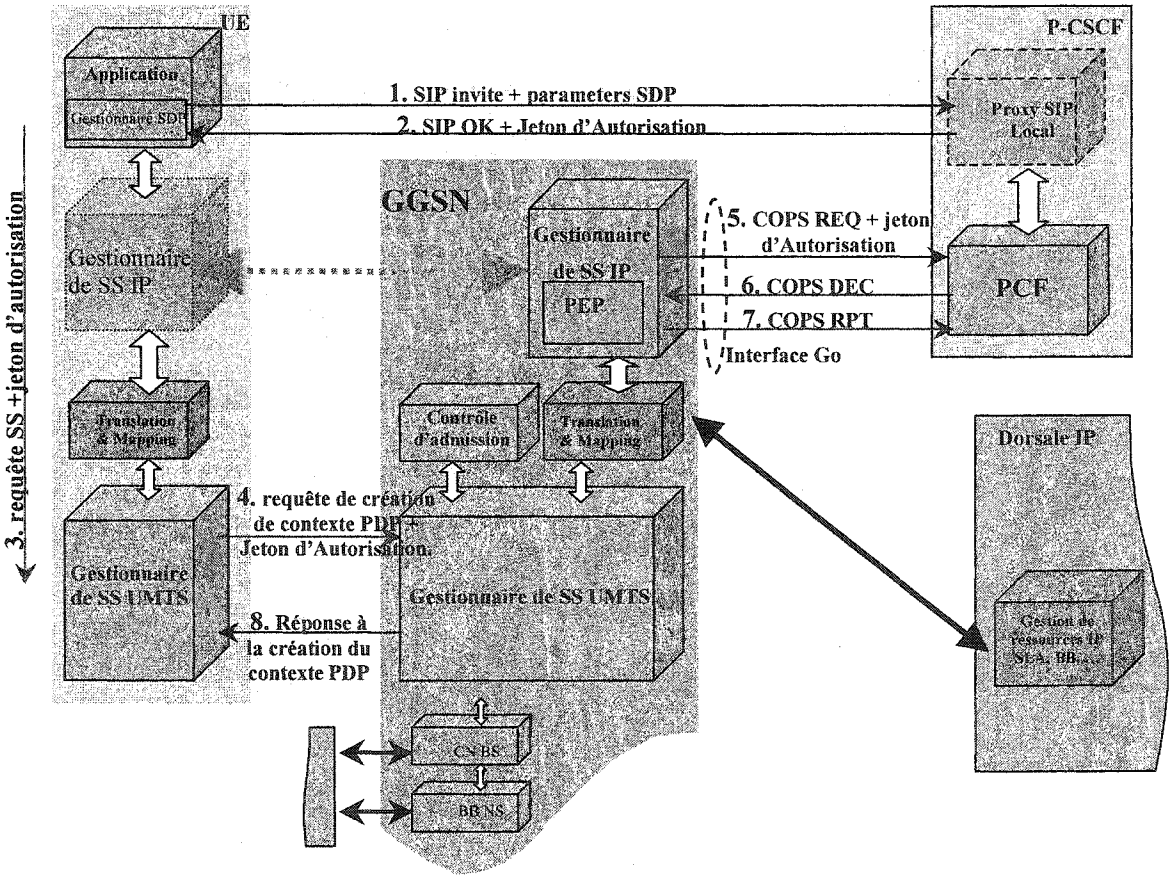


Figure 2.8 Architecture d'inter-fonctionnement du contrôle de QoS de bout en bout

2.3.2 La mise en correspondance des classes UMTS avec les classes DiffServ

Comme l'information de la QoS est issue de la description du type d'application et de ses besoins, nous considérons présentement la mise en correspondance de cette information de QoS dans les différents niveaux, commençant par le couche applicative et aboutissant à la couche d'accès UMTS. L'établissement de session et sa modification dans un sous-système multimédia IP dans UMTS nécessite un échange de messages de bout en bout entre l'application de l'usager et le P-CSCF en utilisant SIP/SDP. Le P-CSCF envoie les informations de description SDP appropriées au PCF. La PCF met en correspondance les paramètres SDP avec les paramètres de la QoS IP autorisée et les transfèrent au GGSN via l'interface Go. À son tour, le GGSN met en correspondance ces paramètres de QoS IP autorisée avec les paramètres de QoS UMTS autorisée. Le message SIP/SDP va aussi être traité par l'équipement usager qui réalise sa propre mise en correspondance pour fournir les paramètres de QoS UMTS afin de remplir les champs QoS dans les requêtes d'activation ou de modification de contexte PDP.

Un contexte PDP est une session d'un service de données UMTS dont le transport est orienté paquets. Dès la réception d'une requête de création ou de modification d'un contexte PDP, le GGSN doit comparer les paramètres de QoS UMTS demandées dans la requête avec les paramètres de QoS UMTS autorisée. Si la requête de QoS est dans les limites autorisées par le PCF, la création ou la modification de PDP doit être acceptée.

La mise en correspondance entre SDP et QoS IP autorisée

- Les *paramètres SDP* maintenues par la PCF sont :

- l'adresse IP de la destination ;
- le numéro de port de la destination ;
- le protocole de transport ;
- l'information sur la direction du média (sendrecv, sendonly, recvonly) ;

- la direction de la source (d'origine ou de destination) ;
- le groupe d'appartenance du composant média ;
- l'information sur le type de média (audio, vidéo, application ou contrôle) ;
- le paramètre de bande passante.

- Les *paramètres de QoS IP autorisée transmis au GGSN via l'interface Go*

- La classe DiffServ : détermine la plus haute PHB (offrant la meilleure QoS) qui peut être utilisée pour le composant média. Elle est dérivée de l'information sur le type de média dans SDP ;
- Le débit binaire des données : il est extrait du paramètre de bande passante dans SDP. Le débit binaire doit inclure tout l'overhead venant de la couche IP et les couches supérieures tel que UDP et RTP. Il doit aussi inclure l'overhead provenant de l'utilisation de RTCP.

- La *mise en correspondance:* Selon 3GPP [2002a], la PCF doit utiliser les règles de mise en correspondance exprimées selon l'algorithme de la Figure 2.9 pour dériver la plus haute PHB DiffServ autorisée à partir des paramètres SDP de description de session.

```
SI  (média = « audio » OU « vidéo » )  ET ( direction_média = « sendrecv »)
                ALORS
                Max_DiffServPHB_Autorisée = « EF » ;
SINON SI (média = « audio » OU « vidéo »)
        ET (direction_média = « sendonly » OU « recvonly »)
                ALORS
                Max_DiffServPHB_Autorisée = « AF4 » ;
SINON SI (média = « application » OU « contrôle »)
                ALORS
                Max_DiffServPHB_Autorisée = « AF3 » ;
SINON
                Max_DiffServPHB_Autorisée = « BE » ;
FIN SI;
```

Figure 2.9 Algorithme de mise en correspondance SDP/Max PHB DS autorisée

La mise en correspondance entre QoS IP et QoS UMTS

Le nœud passerelle GGSN invoque la mise en correspondance entre les informations de QoS IP reçues de la fonction de contrôle de politique de service PCF (la classe DiffServ et le débit binaire) et quelques paramètres de QoS UMTS parmi les attributs du service support UMTS décrits précédemment dans ce chapitre. La mise en correspondance est faite par la fonction de translation qui traduit l'information de QoS IP autorisée en information de QoS UMTS autorisée. Le projet 3GPP [2002a] recommande que le GGSN tire la plus haute classe de trafic UMTS autorisée pour le contexte PDP à partir de la PHB DiffServ maximale autorisée et cela suivant le Tableau 2.5.

Tableau 2.5 Mise en correspondance entre PHB et classe UMTS autorisée

PHB DiffServ	***Classe de trafic UMTS***	***Priorité de traitement***
EF	Conversationnelle	-
AF41	À flux continu	-
AF31	Interactive	1
AF21		2
AF11		3
BE	D'arrière plan	-

2.4 Synthèse des problèmes ouverts

En parcourant les spécifications techniques de l'architecture de QoS définie par 3GPP [3GPP, 2002a] dans leurs dernières versions de juin 2002, on remarque que l'éditeur remet les détails de l'étude de la mise en correspondance des paramètres de QoS à un travail futur. Toutefois, une mise en correspondance sommaire entre QoS UMTS d'une part et QoS IP d'autre part a bien été spécifiée par 3GPP, mais l'efficacité de son implémentation n'a pas été encore prouvée, surtout pour les applications les plus critiques en QoS tel que la téléphonie IP et la vidéo téléphonie. En effet, dans l'état actuel de la recherche, nous n'avons pas trouvé de travaux qui ont fait une étude de performance assez poussée pour montrer que le choix de l'ensemble des paramètres de

QoS à faire correspondre ainsi que la granularité de l'association entre ces différents paramètres tel que présentés par 3GPP constitue la meilleure façon pour garantir l'inter-fonctionnement d'un domaine d'accès UMTS avec un domaine dorsal IP.

Dans le chapitre qui suit, nous allons montrer qu'il serait possible de rencontrer des problèmes de performance pour la prestation d'une QoS de bout en bout aux services multimédia (tel que la vidéo téléphonie) en implémentant la mise en correspondance définie par 3GPP. En effet, la granularité de cette mise en correspondance ne permet pas de distinguer entre flux de trafic de natures différentes : audio et vidéo inclus dans la classe conversationnelle sont mis en correspondance avec le même PHB DiffServ au niveau de la dorsale IP, ce qui à notre avis va engendrer une perte de performances au niveau de la QoS de bout en bout fournie à chacun des deux types de flux. Ainsi, nous allons montrer que voix et vidéo exhibent des comportements différents et qu'il serait judicieux de les traiter séparément dans les mécanismes de différenciation de service.

CHAPITRE 3
DIFFÉRENCIATION DES SERVICES DE TÉLÉPHONIE MUTLIMÉDIA

Comme nous l'avons déjà présenté dans le chapitre précédent, le mécanisme couramment utilisé pour faire interopérer la QoS d'un réseau d'accès UMTS avec celle d'un réseau dorsal IP d'interconnexion est *la mise en correspondance* ou *QoS mapping* entre classes UMTS et classes DiffServ. Toutefois, compte tenu du fait que la téléphonie multimédia fait partie des services les plus exigeants en QoS et les plus critiques pour les opérateurs de télécommunications, la mise en correspondance de la QoS de ce type de services entre domaines UMTS et domaines DiffServ est un problème crucial pour leur bon fonctionnement. En effet, comme la faiblesse d'une chaîne est celle de son maillon le plus faible, la garantie d'une QoS de bout en bout pour un service de téléphonie est conditionnée par l'efficacité de l'algorithme de mise en correspondance qui peut être considéré comme le maillon faible de la chaîne. Ainsi, dans le présent chapitre, nous aborderons en premier lieu les faiblesses de l'algorithme de mise en correspondance pour la classe conversationnelle en nous basant sur la caractérisation du trafic de la téléphonie multimédia montrant son hétérogénéité. En second lieu, nous proposerons un algorithme de mise en correspondance raffinée permettant de différencier entre trafic de voix et trafic de vidéo téléphonie. Par la suite, une analyse théorique sera effectuée pour déterminer les conditions auxquelles doit se plier l'algorithme d'ordonnancement et de partage des liens afin de réaliser une différenciation efficace entre ces deux types de trafic en offrant à chacun son niveau de QoS requis. Enfin, des modèles de trafic de voix et de vidéo seront présentés.

3.1 Problématique de la mise en correspondance de QoS

Cette mise en correspondance a été définie par le projet de partenariat de troisième génération 3GPP [3GPP, 2002a] ainsi que dans plusieurs autres travaux de recherche [Chaskar et al., 2001; Maniatis et al., 2002]. On a remarqué que, dans toutes ces définitions, tous les services de la classe conversationnelle (services de téléphonie multimédia), qui possèdent les exigences de QoS les plus strictes, sont toujours associés à la même classe DiffServ 'EF'. Cette association provient du fait que la classe 'EF' transporte le seul service 'PREMIUM' [Jacobson et al., 1999] offert par DiffServ qui garantit la meilleure QoS : des délais extrêmement bas et des taux de pertes nuls obtenus par une émulation d'un service de ligne spécialisée virtuelle [Jacobson et al., 1999]. Toutefois, la classe conversationnelle englobe deux types de trafic distincts : la voix téléphonique et la vidéo téléphonie [3GPP, 2002b]. En effet, malgré que les trafics de voix et de vidéo téléphonie aient quelques caractéristiques communes (des latences très faibles et des contraintes de temps réel très strictes) qui ont poussé 3GPP à les regrouper dans une même classe UMTS, ils présentent des divergences stochastiques importantes, comme nous allons le montrer par la suite.

3.1.1 Caractérisation du trafic hétérogène dans la téléphonie multimédia

Nous nous intéressons à analyser le trafic de téléphonie multimédia en faisant intervenir simultanément des flux bidirectionnels de voix et de vidéo, constituant ainsi les deux types de trafics possibles véhiculés par la classe conversationnelle. Cette analyse, permettra essentiellement d'extraire les caractéristiques intrinsèques du trafic de voix sur IP (*VoIP*) d'une part et celles du trafic de vidéo téléphonie sur IP (*VIP*) d'autre part. Nous procéderons en capturant les paquets IP générés par une session réelle de vidéo conférence (voix et vidéo) se basant sur la téléphonie SIP. Après séparation des traces récoltées sur les paquets de voix de celles récoltées sur les paquets de vidéo, on effectuera une analyse statistique exhaustive sur chacun d'eux afin de caractériser le comportement des deux types de sources de trafic. Un attribut distinctif appelé

paramètre de Hurst sera utilisé pour mesurer une caractéristique essentielle du trafic qui est son degré d'auto-similarité.

Méthode proposée pour la capture du trafic de téléphonie multimédia

La méthode tente de capturer (ou 'sniffer') tous les paquets IP qui vont être transmis sur le réseau et qui ont été générés par une application de téléphonie multimédia lancée en temps réel. Le logiciel de capture *Ethereal* [Ethereal] installé du coté de la source de trafic garde trace des instants de capture (coïncidant avec les instants de génération à quelques microsecondes près) et des détails sur les différents champs des entêtes des paquets transmis sur le réseau. Ces informations nous seront utiles pour établir les distributions stochastiques des tailles et des temps d'arrivées des paquets IP de voix et de vidéo. Une autre application utile pour cette capture de paquets est le filtrage des champs intéressants dans les fichiers traces et leur utilisation pour alimenter un modèle de génération de trafic. Cela permet de piloter un processus de simulation servant à valider les modèles que nous implanterons dans le prochain chapitre. Comme nous sommes intéressés par des statistiques à la sortie de l'application source (les codecs de voix et de vidéo), nous avons pris garde de ne pas faire intervenir des facteurs externes qui peuvent biaiser les informations de traces de paquets telles que la charge réseau ou encore l'utilisation CPU. De ce fait, nous avons capturé le trafic sur la même station qui le génère (Figure 3.1).

Ainsi, nous avons lancé une session de *Microsoft Voice.NET* implémentant la téléphonie SIP en utilisant l'application de messagerie multimédia *Windows Messenger* tel que spécifié dans [Cisco, 2001]. Malgré son faible impact sur nos traces de trafic, nous avons rejeté le choix d'utiliser *Microsoft NetMeeting* implémentant la téléphonie H.323 puisque la téléphonie SIP constitue le standard recommandé par 3GPP pour les services multimédia de l'UMTS [3GPP, 2002a]. De plus, SIP [Handley et al., 1999] présente une signalisation textuelle certes plus importante et plus compréhensible que H.323, ce qui nous aidera à identifier facilement les flux et leurs paramètres négociés à l'initiation de la session (Figure 3.2).

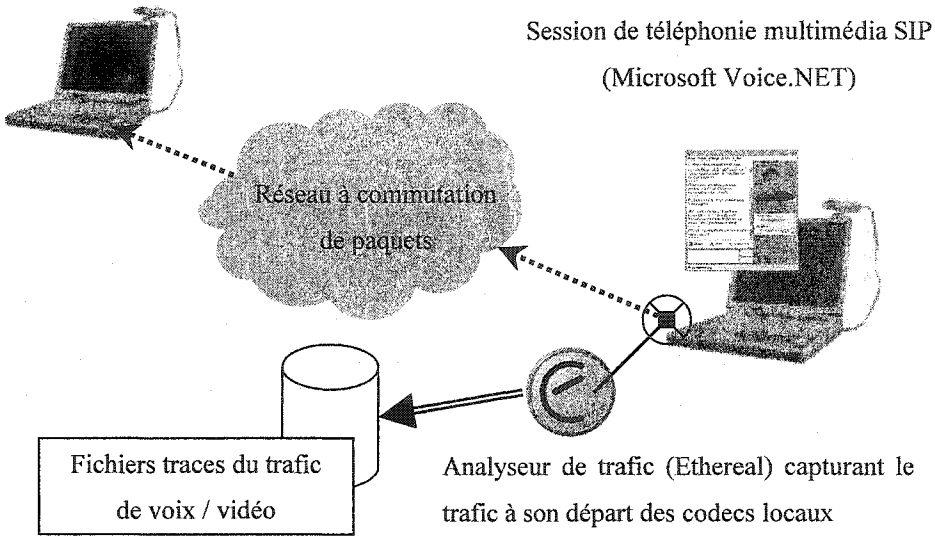


Figure 3.1 Capture des paquets IP d'une session de téléphonie multimédia

```
Session Initiation Protocol
  Request line: INVITE sip:65.92.215.154:9578 SIP/2.0
Message Header
    Via: SIP/2.0/UDP 81.48.225.149:13694
    From: "KAMOU" <sip:rachabenali@hotmail.com>;tag=8ccba2b9-0a23-439f-97b1-787d33f8c0ba
    To: <sip:65.92.215.154:9578>
    Content-Type: application/sdp
    Content-Length: 528
  Session Description Protocol
    Media Description, name and address (m): audio 18414 RTP/AVP 97 111 112 6 0 8 4 5 3 101
    Media Attribute (a): rtpmap:97 red/8000
    Media Attribute (a): rtpmap:111 SIREN/16000
    Media Attribute (a): fmtp:111 bitrate=16000
  Media Description, name and address (m): video 36026 RTP/AVP 34 31
    Media Attribute (a): rtpmap:34 H263/90000
```

Figure 3.2 Codecs de voix et vidéo dans une session de téléphonie multimédia

À partir des paramètres de description de session SDP extraits du trafic capturé et des capacités de décodage et d'analyse de trames offertes par le logiciel Ethereal, nous pouvons déduire que le codec utilisé pour la voix est *'SIREN'* à 16 Kbps, qui n'est autre qu'une extension du codec standard G.722.2. Pour la vidéo, le codec négocié et sélectionné est celui pour les communications à faible taux binaire [ITU-T, 1998] avec un encodage à 90 Kbps. D'après les spécifications techniques de l'ITU-T, H.263 se base

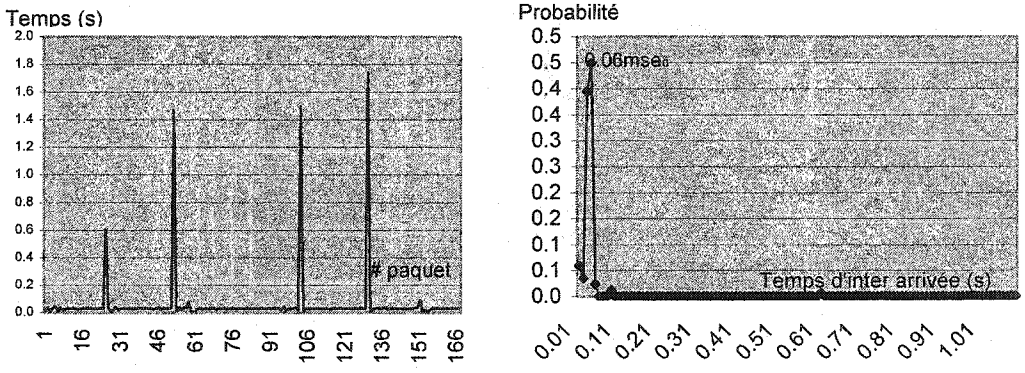
sur la compression MPEG 2. Les codecs se basant sur la compression MPEG 4 pour la vidéo téléphonie sont en cours de standardisation au sein de l'ITU-T. Selon ce même organisme [ITU-T, 1998], G.722 et H.263 sont considérés parmi les plus utilisés dans les communications mobiles à faible bande passante.

Tableau 3.1 Court extrait des traces de paquets filtrées et formatées

Temps d'arrivée (msec)	Longueur (octets)
483.809167	1212
523.858024	455
573.929206	573
664.075584	282
764.202152	788
824.308708	640
964.586853	644
974.655742	1079
1044.729633	813

Analyse statistique du trafic généré par une source de voix

La fonction de densité de probabilité (PDF) de l'inter arrivée des paquets de voix (Figure 3.3b que nous avons construit à partir de la Figure 3.3a) présente un pic correspondant à une valeur d'inter-arrivée de 60 msec qui coïncide avec la moyenne calculée à partir des statistiques récoltées dans les traces de trafic. Le taux moyen d'arrivée des paquets de voix a une valeur de 1/60 paquets/msec ≈16.7 paquets/sec.



a) Inter-arrivée des paquets de voix **b) PDF correspondante**

Figure 3.3 Inter-arrivée des paquets de voix et PDF correspondante

Afin de tester l'auto-similarité de la distribution des inter-arrivées des paquets de voix, il faut calculer le paramètre de Hurst. En général, pour estimer ce paramètre, on dresse la courbe de statistique de l'intervalle réajusté et redimensionné R/S(n) [Annexe B] en fonction de *n* sur un repère logarithmique et on calcule sa pente [Rose, 1996]. Toutefois, il est possible d'utiliser des outils logiciels qui implémentent des algorithmes optimisés [Hagiwara et al., 1999] pour une estimation plus robuste et efficace du paramètre de Hurst. Dans notre cas, nous avons utilisé l'outil *Hurst ESTimator* [hest]. D'après la Figure 3.4, les paquets de voix ont des tailles constantes de 86 octets au total, en omettant l'*overhead* des entêtes (Ethernet + PPP + IP + UDP = 14 + 8 +20 +8 =50), on trouve une très faible charge utile de 36 octets par paquet. Toutefois, le taux binaire moyen d'une source de voix transmettant sur le réseau doit tenir compte de l'*overhead* et sa valeur peut être calculée : 16.7 paquets/sec * (86 * 8 bits/paquet) ≈ 11 kbps.

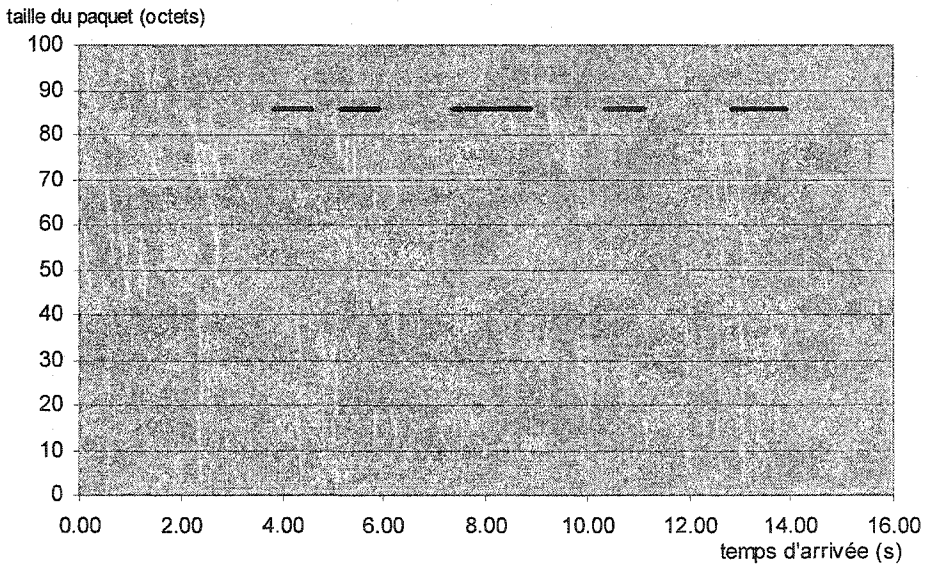


Figure 3.4 Tailles des paquets de voix

D'après l'effet de Hurst [Rose, 1996], le paramètre de Hurst estimé à **H≈0.5** (Figure 3.5) nous montre que la distribution des inter-arrivées des paquets générés par une source de voix n'est pas auto-similaire et qu'elle est caractérisée par des

dépendances à courts intervalles. Cette constatation, en plus de la nature constante des tailles des paquets de voix (86 octets chacun), nous permettra de présenter un modèle assez simple d'une source de voix en utilisant des processus markoviens largement utilisés en téléphonie.

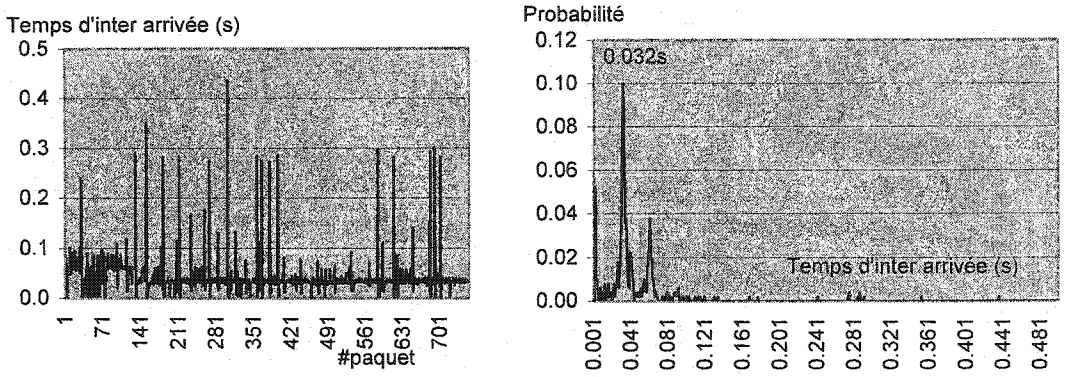
```
D:\>hest –A -t 0 –k 0 VoiceInterarrivals.txt
        hest:     moyenne = 0.0635857284, variance = 0.0737763643
        rs():     f(x)=-0.011641 +0.484413x    →    H=0.484413 ≈ 0.5
```

Figure 3.5 Sortie de l'outil Hest pour les inter-arrivées des paquets de voix

De plus, à la Figure 3.4, on note des discontinuités fréquentes dans la génération des paquets de voix. En effet, comme nous allons le montrer plus tard dans la modélisation d'une source de voix, le codage de la voix tient compte des périodes de silence très fréquentes dans la parole humaine : il les détecte et les supprime. En se basant sur le fait que le codec de voix G.722.2 que nous avons utilisé dans nos captures génère des paquets contenant une seule trame de 20 msec chacun, le taux moyen de génération des paquets sera de l'ordre de 1/20 paquets/msec = 50 paquets/sec si la voix est transmise en continuité dans le temps. Toutefois, les discontinuités causées par les silences ont abaissé le taux moyen de génération à 15.87 paquets/sec. Cela nous permet d'estimer la durée moyenne de la période d'activité vocale à 16/50 = 32% de celle de la période de silence. Cette valeur sera évoquée dans la modélisation de la voix.

Analyse statistique du trafic généré par une source de vidéo

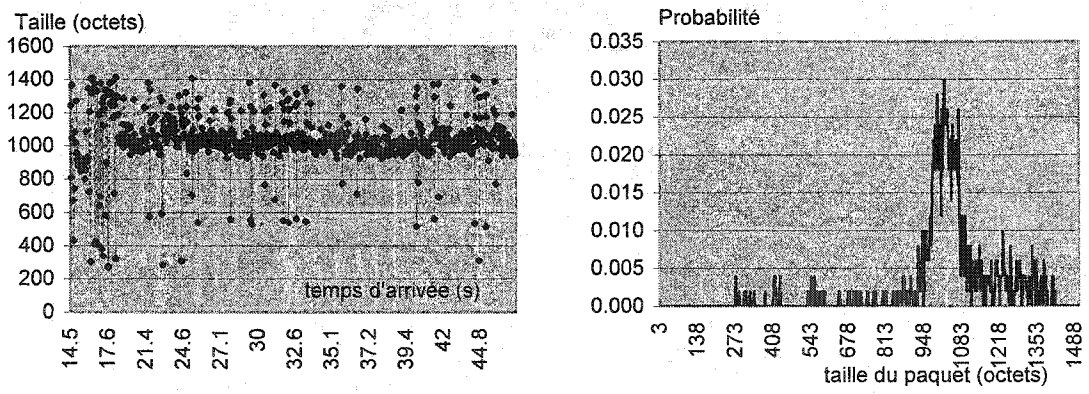
La fonction de densité de probabilité de l'inter-arrivée des paquets de vidéo (Figure 3.6b que nous avons construit à partir de la Figure 3.6a) présente plusieurs pics, le plus élevé correspond à une valeur d'inter-arrivée de 32 msec. La moyenne calculée à partir des statistiques récoltées dans les traces de trafic est de 43 msec. Le taux moyen d'arrivée des paquets de vidéo a une valeur de 1/43 paquets/msec = 23 paquets/sec.



a) Inter-arrivée des paquets de vidéo **b) PDF correspondante**

Figure 3.6 Inter-arrivée des paquets vidéo et PDF correspondante

D'après la Figure 3.7a, les paquets de vidéo ont des tailles très variables dans le temps. La Figure 3.7b, nous montre que la taille d'un paquet de vidéo peut varier de 921 à 1074 octets avec la taille la plus probable qui est de 1028 octets et qui correspond à la moyenne calculée sur les tailles de tous les paquets capturés.



a) Tailles des paquets de vidéo **b) PDF correspondante**

Figure 3.7 Tailles des paquets de vidéo et PDF correspondante

En se basant sur l'effet de Hurst, des paramètres de Hurst, estimés à **H≈0.87** pour la distribution des inter-arrivées des paquets vidéo (Figure 3.8) et à **H≈0.64** pour la distribution de leurs tailles (Figure 3.9), nous montrent que *contrairement au trafic généré par une source de voix, celui généré par une source vidéo est auto-similaire*, puisque *H* pour une source vidéo est loin de la valeur ½ de la non auto-similarité.

```
D:\>hest -t 0 -k 0 VideoInterarrivals.txt
        hest:       moyenne = 0.0431258753, variance = 0.0018659993
        rs():       f(x)=-0.398290 +0.867350x  →  H=0.867350
```

Figure 3.8 Sortie de l'outil Hest pour les inter-arrivées des paquets de vidéo

```
D:\>hest -t 0 -k 0 VideoSizes.txt
        hest:       moyenne = 1028.3787841797, variance = 27278.2578125000
        rs():       f(x)=-0.165592 +0.643706x  →  H=0.643706
```

Figure 3.9 Sortie de l'outil Hest pour les tailles des paquets de vidéo

De telles valeurs du paramètre de Hurst nous prouvent que le trafic généré par une source vidéo présente des dépendances sur de longs intervalles (LRD) et que sa modélisation n'est pas aussi simple qu'une source de voix, puisque les processus markoviens et poissoniens classiques ne peuvent capturer le comportement à long terme du trafic de vidéo téléphonie, comme il a été démontré dans la littérature [Krunz et al., 2001]. De ce fait, un modèle se basant sur un processus multi-fractal sera présenté plus loin. Le Tableau 3.2 présente une synthèse des caractéristiques d'une source de voix/vidéo.

Tableau 3.2 Caractéristiques d'une source de voix/vidéo

Source de trafic	Voix (ex : G.722.1)	Vidéo (ex : H.263)
Taille moyenne des paquets IP (octets)	86	1028
Ecart type de la taille des paquets (octets)	0 (tailles constantes)	165 (tailles variables)
Taux (paquets par seconde)	50	23
Paramètre de Hurst	0.5	0.87
Taux de rafales du trafic généré	Très faible/Nul	Important
Auto-similarité du trafic généré	NON	OUI
Modèle	Markovien classique	Multi-fractal

3.1.2 Problématique du trafic hétérogène de la téléphonie multimédia

D'après l'étude statistique que nous avons effectué précédemment, contrairement au trafic de voix ayant des paquets IP de taille réduite, ceux de la vidéo ont des tailles importantes. De plus, nous avons déduit que la voix est codée à un taux constant et sa transmission est souvent dépourvue de rafales, contrairement à la vidéo qui est codée généralement à un taux variable, ce qui génère un trafic auto-similaire.

Grossglauser et Keshav [1996] ont étudié les performances d'un trafic à taux constant (CBR) tel que celui de la voix dans un réseau à grande échelle, en considérant un nombre très important de sessions et de commutateurs ATM. Ils ont conclu que le délai encouru par un flux à taux constant (multiplexé avec des flux de même type) ne dépasse pas le temps de transmission de quelques cellules ATM, et cela, même sous une charge très élevée. De plus, l'allocation de ressources et le contrôle d'admission sont assez simples, vu qu'il n'y a pas de variation dans les besoins en ressources. Toutefois, le trafic vidéo est souvent caractérisé par un taux variable naturel, dû aux technologies de codage et de compression utilisées.

Ainsi, le fait de négliger la différenciation entre ces deux types de trafic dans la mise en correspondance de QoS existante, telle que définie par 3GPP, peut poser un certain nombre de problèmes. À partir de ces constations sur le trafic de la classe

conversationnelle, l'agrégation du trafic de voix et celui de la vidéo dans la même classe DiffServ EF au niveau des routeurs dorsaux peut affecter les performances d'acheminement des deux types de trafic. Plus particulièrement, une étude sur le trafic multimédia sur Internet [Tobagi et al., 2001] a montré que le fait d'augmenter le nombre de flux de vidéo multiplexés avec des flux de voix conduit à l'augmentation significative du délai des deux types de trafic. Cela a été prouvé par les longues rafales vidéo qui s'injectent dans les mêmes files contenant le trafic de voix.

Dans notre cas, ce problème sera accentué par le fait d'avoir un trafic conversationnel issu d'un accès radio UMTS. En effet, il est bien connu que l'accès radio favorise la fragmentation des longs paquets afin de réduire les erreurs de transmission radio. Ainsi, les longs paquets vidéo se trouvent divisés en plusieurs fragments qui se regroupent dans les différentes files de transmissions pendant un intervalle de temps très réduit. Cela a pour effet d'augmenter le taux de rafales du trafic vidéo, qui à son tour, peut provoquer une augmentation du taux de pertes de paquets et du délai, et certainement une forte gigue non souhaitable.

De ce fait, la solution immédiate à ces problèmes est de pouvoir contrôler la QoS fournie aux deux types de trafic. Et cela ne se fait qu'en séparant et en différenciant les services de voix et de vidéo de la classe conversationnelle. D'où l'algorithme de mise en correspondance raffinée de QoS proposé dans ce qui suit.

3.2 Algorithme de différenciation voix/vidéo

À la section 3 du chapitre 2, nous avons mentionné que la mise en correspondance entre classes UMTS et classes DiffServ est effectuée au nœud GGSN et vérifiée au nœud CSCF. Au niveau de la couche réseau du GGSN, une traduction directe est effectuée par la fonction de translation entre la classe UMTS contenue dans le message d'activation de contexte PDP et la classe DiffServ correspondante. Au niveau de la couche applicative du CSCF, une interrogation de la fonction de contrôle de politique PCF est effectuée pour déterminer à partir des paramètres de description SDP

établis à l'initiation de la session si le service est autorisé à utiliser la classe DiffServ demandée par la fonction de translation. Dans la fonction de translation, le raffinement de la mise en correspondance de QoS que nous avons proposé pour les services de téléphonie multimédia est présenté au Tableau 3.3.

Tableau 3.3 Séparation entre voix et vidéo dans la mise en correspondance

PHB DiffServ	Classe de trafic UMTS	Priorité de traitement
EF	Conversationnelle (voix)	-
AF41	Conversationnelle (vidéo)	
AF31	À flux continu	-
AF21	Interactive	1
AF22		2
AF23		3
BE	D'arrière plan	-

Dans la fonction PCF que nous avons présenté au chapitre 2, ce même raffinement de mise en correspondance est décrit à la Figure 3.10.

```
SI  (média = « audio » )  ET ( direction_média = « sendrecv »)
            ALORS
            Max_DiffServPHB_Autorisée = « EF » ;
SINON SI (média = « vidéo » )  ET ( direction_média = « sendrecv »)
            ALORS
            Max_DiffServPHB_Autorisée = « AF4 » ;
SINON SI (média = « audio » OU « vidéo »)
        ET (direction_média = « sendonly » OU « recvonly »)
            ALORS
            Max_DiffServPHB_Autorisée = « AF3 » ;
SINON SI (média = « application » OU « contrôle »)
            ALORS
            Max_DiffServPHB_Autorisée = « AF2 » ;
SINON
            Max_DiffServPHB_Autorisée = « BE » ;
FIN SI;
```

Figure 3.10 Algorithme d'autorisation de la mise en correspondance raffinée

L'algorithme de mise en correspondance entre les classes UMTS et les classes DiffServ distinguant entre voix et vidéo téléphonie permet de séparer les deux types de trafic dans deux files d'ordonnancement différentes dans chaque routeur dorsal du fait qu'ils auront nécessairement deux marquages différents, soit deux DSCP distincts du champ EXP de l'entête MPLS : la voix est marquée 'EF' et la vidéo est marquée 'AF4'. Toutefois, cette séparation n'indique rien sur la discipline selon laquelle les deux types de trafic vont être ordonnancés et transmis sur les liens dorsaux. En effet, une fois les paquets de voix et de vidéo marqués au niveau du GGSN, ils doivent être ordonnancés pour être transmis sur les liens dorsaux. Plusieurs schémas d'ordonnancement se présentent, nous avons retenu ceux qui sont les plus importants et les plus adaptés à notre cas qui sont PQ (Figure 3.11) et WFQ (Figure 3.12).

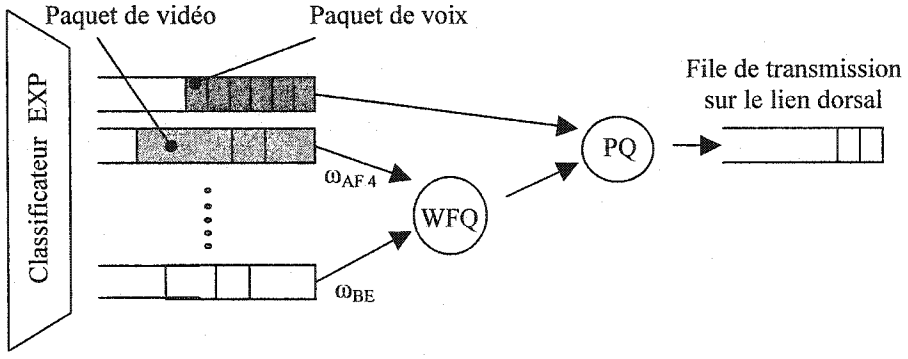


Figure 3.11 Ordonnancement par priorité donnée au trafic de voix

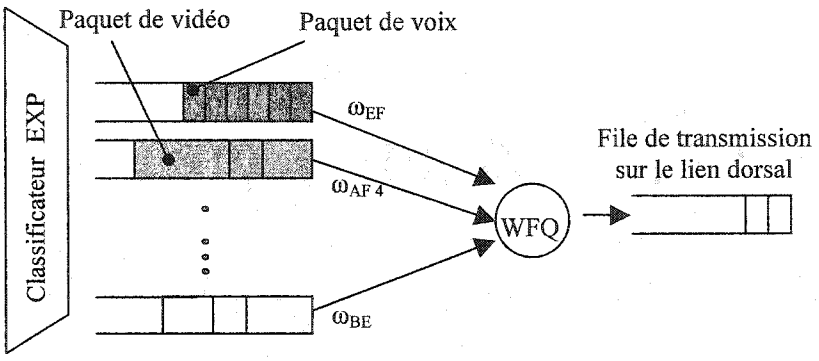


Figure 3.12 Ordonnancement par partage équitable pondéré entre voix et vidéo

Afin d'effectuer un choix judicieux des paramètres de l'algorithme d'ordonnancement mis en jeu dans la mise en correspondance, nous allons analyser, dans la prochaine section, l'ordonnancement voix/vidéo. L'ordonnancement par priorité ne sera pas considéré vu que le schéma donnant priorité à la voix par rapport à la vidéo ne fait intervenir aucun autre paramètre, alors que le schéma de partage équitable WFQ fait intervenir les poids à donner à chaque type de trafic. Toutefois, l'ordonnancement par priorité fera l'objet du prochain chapitre.

3.3 Optimisation des délais de la téléphonie multimédia

Dans le chapitre précédent, la QoS a été définie simplement comme la capacité du réseau à fournir un service avec un niveau de qualité donné, mesuré suivant différents paramètres plus ou moins importants selon le type de l'application. Les paramètres de QoS de bout en bout qui reviennent le plus souvent sont : *le délai, la variation du délai ou gigue, le débit et le taux de pertes de paquets*. En outre, Schwartz [1996] a défini la QoS comme une fonction objective des différentes classes de service et faisant intervenir un ou plusieurs paramètres de QoS parmi ceux que nous venons d'énoncer. Ainsi, le but de notre étude de performance est d'évaluer cette fonction objective en fonction des paramètres d'ordonnancement, en estimant les valeurs de l'indice de performance le plus discriminant et pertinent pour les services de téléphonie multimédia qui se trouve être le délai réseau de bout en bout. Le principal objectif de l'ordonnancement dans la gestion de la QoS est de distribuer des priorités quant à la consommation de la bande passante, au contrôle de la gigue et du délai, et de minimiser le taux d'erreurs résiduelles. Vu la difficulté à construire un modèle analytique exact pour un trafic sporadique ainsi que la complexité de l'algorithme d'ordonnancement WFQ, la formulation mathématique de tous les paramètres de QoS en fonction du trafic s'avère très complexe pour notre étude de performance. De ce fait, nous nous limiterons à optimiser les performances de *l'indice de délai de bout en bout au pire cas*. Les autres indices de QoS ainsi que l'étude du cas moyen (au lieu du pire cas) sont impossibles à évaluer avec notre formulation

mathématique et feront l'objet de simulations au prochain chapitre. En outre, il est très important de s'assurer que le fait d'accorder des priorités (des poids) à un ou plusieurs flux dans l'algorithme d'ordonnancement, ne va pas entraîner des problèmes de dégradation de QoS pour les autres flux.

3.3.1 Délai WFQ pour les agrégats de sessions

La discipline de service équitable pondéré (WFQ) tient compte de la longueur variable des paquets et constitue une excellente approximation de la discipline de service équitable théorique GPS dans le cadre des réseaux à commutation de paquets de longueur variable [Keshav, 1997], ce qui la rend appropriée dans notre cas, vu la longueur variable des paquets vidéos.

La formule donnant la borne supérieure du délai de bout en bout pour la discipline WFQ est basée sur l'hypothèse des sessions individuelles indépendantes [Parekh et Gallager, 1994]. En effet, WFQ a été initialement utilisé pour les services intégrés IntServ pour garantir le service aux différents micro flots formant chacun une session indépendante. Toutefois, avec l'apparition des services différenciés DiffServ, l'utilisation de WFQ est maintenant basée sur des macro flots qui ne sont que des agrégations des micro flots. Ces macro flots ou agrégats constituent les OAs (*Ordered Aggregate*) de DiffServ, qui sont souvent classifiés selon les DSCP ou EXP DiffServ présentés au chapitre 2.

On définit un *agrégat de sessions A* comme l'agrégation d'un nombre n de sessions individuelles indépendantes ayant un ensemble de propriétés communes (par exemple : sessions transmettant des paquets de voix, de vidéo, de données) et partageant un même chemin réseau de longueur L.

Supposant que la discipline utilisée pour ordonnancer la classe de service DiffServ associée à cet agrégat A est l'ordonnancement de partage équitable pondéré WFQ avec un poids Φ_A et supposant les simplifications suivantes :

- tous les L liens sur le chemin du trafic ont la même capacité C ;
- la somme des poids WFQ de tous les agrégats est égale à 1 ;

- $P_{\max,A} = \max_{i \in A} P_{\max,i}$.

La borne supérieure du délai au pire cas pour l'agrégat A doit être celle du cas des sessions gloutonnes étagées.

Nous rappelons qu'une *session gloutonne (greedy session)* [Parekh et Gallager, 1994] est une session contrainte à un seau à jetons (ρ_i, b_i) qui, à partir de l'instant τ, utilise autant de jetons que possible pour la transmission de ses paquets à tout instant $t \geq \tau$. Pour ces instants, la session i transmet alors ses paquets au débit le plus élevé possible.

En outre, *les sessions gloutonnes étagées* (*« staggered » greedy sessions*) [Parekh et Gallager, 1994] partageant un chemin réseau $P = 1, 2, \ldots, n$ et formant un agrégat de sessions sont définies par un vecteur ($T_1, T_2, \ldots, T_n$), $T_1 \leq T_2 \leq \ldots \leq T_n$, tel que toutes les sessions au nœud 1 sont gloutonnes simultanément à l'instant $t = T_1$ et que les sessions indépendantes au nœud j n'envoient pas de trafic dans l'intervalle $[T_1, T_j]$, mais sont gloutonnes simultanément à l'instant $t = T_j$. Autrement dit, un agrégat de sessions gloutonnes forme un agrégat glouton pour lequel toutes les sessions envoient à leur maximum en même temps. Cela revient à considérer le trafic dans "*le pire des pires cas*" : les rafales maximales de trafic, pouvant être transmises par toutes les sessions de l'agrégat, arrivent simultanément au premier nœud du chemin. Donc, l'obtention des valeurs maximales de délai en attente de service pour un agrégat A est subordonnée à l'étude du cas des sessions gloutonnes étagées. Ainsi la formule de la borne supérieure du délai au pire cas, présenté au chapitre 2 (relation 2.4) :

$$d_A \leq \frac{\sum_{i \in A} b_i}{g_A} + \sum_{l=1}^{L-1} \frac{P_{\max,A}}{g_A} + \sum_{l=1}^{L} \frac{P_{\max}}{C} \tag{3.1}$$

Avec :

$$g_A = \Phi_A C \tag{3.2}$$

D'où :

$$d_A \leq \frac{\sum_{i \in A} b_i + (L-1)P_{\max,A}}{\Phi_A C} + \frac{LP_{\max}}{C} \qquad (3.3)$$

3.3.2 Séparation Voix/Vidéo et son effet sur le délai de bout en bout sous WFQ

Notations et hypothèses

Supposons qu'on ait les trois agrégats suivants :

- A : agrégat de n_A sessions de voix (n_A flux identiques de VoIP partageant un même chemin LSP sur la dorsale entre l'*Ingress* raccordé au réseau UMTS source et l'*Egress* raccordé au réseau UMTS destination) ;
- B : agrégat de n_B sessions de vidéo téléphonie (n_B flux identiques de VIP partageant un même chemin LSP sur la dorsale entre l'*Ingress* raccordé au réseau UMTS source et l'*Egress* raccordé au réseau UMTS destination) ;
- C : agrégat de n_C sessions conversationnelles (n_C flux conversationnels partageant un même chemin LSP sur la dorsale entre l'*Ingress* raccordé au réseau UMTS source et l'*Egress* raccordé au réseau UMTS destination).

Pour comparer les délais WFQ relatifs aux deux algorithmes de mise en correspondance celui sans différenciation voix/vidéo et celui avec différenciation voix/vidéo, il est nécessaire de faire l'étude comparative avec la même configuration du trafic d'entrée (même demande, même charge, même nombre de sessions, etc…). Donc, nous supposons que $n_C = n_A + n_B$, avec $n_A \geq n_B$ puisqu'il est très rare de trouver des sessions de vidéo téléphonie qui n'ont pas de sessions de voix qui leurs sont associées. En outre, pour des raisons de simplification de l'étude, nous avons supposé que les sessions d'un même agrégat sont issues de sources parfaitement identiques (les mêmes types de codecs). Ainsi, nous caractériserons un agrégat par le nombre de sessions et les paramètres d'une source de trafic suivant un seau à jetons (n_A, ρ_A, b_A). De plus, puisque la dorsale d'interconnexion des réseaux d'accès UMTS a souvent des tailles géographiques importantes, cela peut induire des temps de propagation T_p sur les liens que nous devons ajouter au délai global.

Si on note D_A^*, D_B^* et D_C^* les bornes supérieures du délai dorsal (le délai encouru par le paquet le plus long de l'agrégat au pire cas) respectivement pour l'agrégat de voix, l'agrégat de vidéo et l'agrégat conversationnel, alors on aura :

$$D_A^* = \frac{n_A b_A + (L-1)P_{\max,A}}{\Phi_A C} + \frac{LP_{\max}}{C} + T_p \tag{3.4}$$

$$D_B^* = \frac{n_B b_B + (L-1)P_{\max,B}}{\Phi_B C} + \frac{LP_{\max}}{C} + T_p \tag{3.5}$$

$$D_C^* = \frac{n_A b_A + n_B b_B + (L-1)P_{\max,C}}{\Phi_C C} + \frac{LP_{\max}}{C} + T_p \tag{3.6}$$

En fixant le poids : $\Phi_C = \Phi_A + \Phi_B = 1$, on pose : $\alpha = \Phi_A = 1 - \Phi_B$.

Nous avons déjà établi que les tailles des paquets vidéo sont toujours beaucoup plus importantes que ceux de la voix, alors on a :

$$P_{\max,C} = P_{\max,B} \tag{3.7}$$

Notons aussi les différents types de délais :

- $D^{max} = D^{bout\text{-}en\text{-}bout} - D^{UMTS}$ désigne le délai maximal absolu permis encouru par le trafic conversationnel sur un chemin dorsal IP de longueur L et sous la discipline WFQ ;
- $D^{bout\text{-}en\text{-}bout} = 150$ msec, tel que défini par l'ITU-T [ITU-T, 2000] pour les services de voix et de vidéo, et par 3GPP [3GPP, 2002b] pour les services conversationnels ;
- D^{UMTS} est le délai encouru par le réseau d'accès UMTS et qui sera évalué par simulations au prochain chapitre.

Analyse comparative des délais

L'analyse analytique des performances (en terme de la borne supérieure du délai au pire cas) de l'approche de différenciation voix/vidéo par rapport à l'approche

classique sous la discipline WFQ, revient à poser le problème d'optimisation suivant (décomposé en deux sous-problèmes):

Sous-problème d'optimisation pour la voix:

$$\begin{cases} \textit{Minimiser la fonction objective :} \\ \textbf{min}\ \ D_A^* - D_C^* = \dfrac{n_A b_A + (L-1)P_{\max,A}}{\alpha C} - \dfrac{n_A b_A + n_B b_B + (L-1)P_{\max,B}}{C} \qquad (3.8) \\ \textit{Sujet à :} \\ D_A^* \le D_C^* \le D^{\max} \text{ et } \alpha \le 1 \\ \textit{Étant donné} : \text{le trafic de voix } (n_A, \rho_A, b_A) \text{ et le trafic de vidéo téléphonie } (n_B, \rho_B, b_B), \text{ les } L \text{ liens de capacité } C \text{ chacun ;} \end{cases}$$

Sous-problème d'optimisation pour la vidéo téléphonie:

$$\begin{cases} \textit{Minimiser la fonction objective :} \\ \textbf{min}\ \ D_B^* - D_C^* = \dfrac{n_B b_B + (L-1)P_{\max,B}}{(1-\alpha)C} - \dfrac{n_A b_A + n_B b_B + (L-1)P_{\max,B}}{C} \qquad (3.9) \\ \textit{Sujet à :} \\ D_B^* \le D_C^* \le D^{\max} \text{ et } \alpha \le 1 \\ \textit{Étant donné} : \text{le trafic de voix } (n_A, \rho_A, b_A) \text{ et le trafic de vidéo téléphonie } (n_B, \rho_B, b_B), \text{ les } L \text{ liens de capacité } C \text{ chacun.} \end{cases}$$

Pour chercher une solution réalisable au premier sous-problème, la relation (3.8) est soumise aux contraintes, et nous donne :

$$D_A^* - D_C^* = \frac{n_A b_A + (L-1)P_{\max,A}}{\alpha C} - \frac{n_A b_A + n_B b_B + (L-1)P_{\max,B}}{C} \le 0 \qquad (3.10)$$

Ce qui équivaut à :

$$n_A b_A + (L-1)P_{\max,A} - \alpha(n_A b_A + n_B b_B + (L-1)P_{\max,B}) \le 0 \qquad (3.11)$$

D'où :

$$\alpha \geq \frac{n_A b_A + (L-1)P_{\max,A}}{n_A b_A + n_B b_B + (L-1)P_{\max,B}} \tag{3.12}$$

avec : $\alpha \leq 1$

Donc, pour réduire la borne supérieure du délai de la voix en utilisant l'approche de différenciation WFQ, il faut que le poids α donné à la classe DiffServ mise en correspondance avec l'agrégat de voix vérifie la relation (3.12) donnant un seuil minimal à ce paramètre. Toutefois, on remarque de la relation (3.8) que l'optimisation du premier sous-problème seul revient à maximiser α, ce qui revient aussi à faire tendre α vers 1. En d'autres termes, donner à la voix le maximum de poids WFQ possible, ce qui évidemment affectera considérablement les performances de la vidéo (relation 3.9) qui voit son poids tendre vers 0. Donc, il est nécessaire de considérer le problème en entier (les deux sous-problèmes) pour dégager une solution globale satisfiable et optimale.

Dans le deuxième sous-problème (relatif à la vidéo téléphonie), la relation (3.9) soumise aux contraintes nous amène à écrire :

$$D_B^* - D_C^* = \frac{n_B b_B + (L-1)P_{\max,B}}{(1-\alpha)C} - \frac{n_A b_A + n_B b_B + (L-1)P_{\max,B}}{C} \leq 0 \tag{3.13}$$

Ce qui équivaut à:

$$\alpha(n_A b_A + n_B b_B + (L-1)P_{\max,B}) - n_A b_A \leq 0 \tag{3.14}$$

D'où :

$$\alpha \leq \frac{n_A b_A}{n_A b_A + n_B b_B + (L-1)P_{\max,B}} \tag{3.15}$$

Donc, pour réduire la borne supérieure du délai de la vidéo téléphonie en utilisant l'approche de différenciation WFQ, il faut que le poids α donné à la classe DiffServ mise en correspondance avec l'agrégat de la vidéo téléphonie vérifie la relation (3.15) donnant un seuil maximal à ce paramètre. Notons qu'une solution qui satisfait le problème en entier (les deux sous-problèmes en même temps) doit vérifier simultanément les deux relations (3.12) et (3.15), ce qui est impossible puisque :

$$\frac{n_A b_A + (L-1)P_{\max,A}}{n_A b_A + n_B b_B + (L-1)P_{\max,B}} \geq \frac{n_A b_A}{n_A b_A + n_B b_B + (L-1)P_{\max,B}} \tag{3.16}$$

Par conséquent, l'amélioration du délai d'un type de trafic engendre la dégradation de l'autre. Ainsi, le problème de départ ne présente pas de solutions satisfiables, ce qui nous amène à considérer le bon compromis qui privilégiera un sous-problème à un autre. En effet, dépendamment des choix stratégiques de l'opérateur, il pourra établir des préférences à optimiser les performances d'un type de trafic au prix de celles d'un autre. Souvent, les opérateurs choisissent de privilégier la voix sur IP par rapport à la vidéo sur IP puisque celle-là constitue le service critique qui est largement utilisé par beaucoup de clients et que sa QoS doit être absolument maintenue au même niveau que la voix téléphonique à commutation de circuits classique. Toutefois, si l'approche de mise en correspondance classique satisfait la contrainte de la borne supérieure du délai absolu pour le trafic de voix $D_A^* \leq D^{\max}$ et que celle de la vidéo téléphonie $D_B^* \leq D^{\max}$ n'est pas satisfaite, il est nécessaire d'améliorer la vidéo sur IP afin d'essayer de satisfaire cette dernière, sans pour autant relâcher la première contrainte ($D_A^* \leq D^{\max}$).

Ainsi, il est possible de reconsidérer tout le problème de deux manières :

- Optimiser le service de téléphonie multimédia en cherchant une amélioration du délai au pire cas pour la voix qui est supérieure à sa dégradation pour la vidéo. Cela revient à trouver les valeurs de α qui vérifient la relation (3.12) pour la condition de diminution du délai pour la voix et la condition suivante :

$$D_B^* - D_C^* < D_C^* - D_A^* \tag{3.17}$$

- Optimiser le service de téléphonie multimédia en cherchant une amélioration du délai au pire cas pour la vidéo qui est supérieure à sa dégradation pour la voix. Cela revient à trouver les valeurs de α qui vérifient la relation (3.15) pour la condition de diminution du délai pour la vidéo et la condition suivante :

$$D_A^* - D_C^* < D_C^* - D_B^* \tag{3.18}$$

On remarque que les relations (3.17) et (3.18) sont équivalentes et reviennent à écrire :

$$\frac{\alpha(n_A b_A + n_B b_B + (L-1)P_{\max,B}) - n_A b_A}{(1-\alpha)C} + \frac{n_A b_A + (L-1)P_{\max,A} - \alpha(n_A b_A + n_B b_B + (L-1)P_{\max,B})}{\alpha C} < 0$$

En posant :

$$\omega = n_A b_A + n_B b_B + (L-1)P_{\max,B} \tag{3.19}$$

l'inéquation devient :

$$Q(\alpha) = 2\omega\alpha^2 - (\omega + (L-1)P_{\max,A})\alpha + n_A b_A + (L-1)P_{\max,A} < 0 \tag{3.20}$$

avec :

$$1 \geq \alpha \geq \frac{n_A b_A + (L-1)P_{\max,A}}{\omega} \tag{3.21}$$

pour l'amélioration du délai de la voix et :

$$\alpha \leq \frac{n_A b_A}{\omega} \tag{3.22}$$

pour l'amélioration du délai de la vidéo.

Cherchons la valeur de α qui minimise le facteur de dégradation de performance Q(α) :

$$\frac{\partial Q(\alpha)}{\partial \alpha} = 4\omega\alpha - \omega + (L-1)P_{\max,A} = 0 \Leftrightarrow \alpha = \frac{\omega - (L-1)P_{\max,A}}{4\omega} \tag{3.23}$$

La dérivée seconde est égale à 4ω qui est positive, la courbe Q(α) est par conséquent une parabole concave qui décroît puis croît. Donc, elle doit s'annuler si on veut qu'elle ait des valeurs négatives. Si nous désignons par Δ le discriminant, nous avons :

$$Q(\alpha) = 0 \Rightarrow \Delta = (\omega + (L-1)P_{\max,A})^2 - 8\omega(n_A b_A + (L-1)P_{\max,A}) \geq 0 \tag{3.24}$$

Donc, pour avoir une solution à l'inéquation (3.20), il faut que α soit dans l'intervalle défini par les deux racines éventuelles de l'équation Q(α)=0. Compte tenu de la complexité du calcul à cause du nombre de paramètres mis en jeu, il est judicieux de s'orienter vers une étude de cas réel pour fixer des valeurs à certains paramètres.

Étude de cas

Nous rappelons que le trafic issu d'une session UMTS est soumis à un contrôle de flux (*policing*) par un seau à jetons pour limiter l'utilisation excessive des ressources non autorisées. Les paramètres du seau à jeton sont établis par la fonction d'autorisation

dans le CSCF et dépendent des informations sur les codecs voix/vidéo signalées par SDP. Selon l'annexe C de 3GPP [2002c], on assigne à chaque session un seau à jetons avec un accumulateur de capacité égale à la taille maximale d'un paquet. D'où : $b_A = P_{\max,A}$ et $b_B = P_{\max,B}$. En outre, on peut supposer que les flux de voix et de vidéo traversent un chemin à commutation d'étiquettes (LSP) sur une dorsale IP de longueur L=7 sauts. Les tailles maximales des paquets de voix et de vidéo sont prises à partir de l'étude statistique faite dans la première section de ce chapitre: $P_{\max,A} = 72$ *octets* et $P_{\max,B} = 1000$ *octets*. De plus, nous posons n le ratio du nombre de sessions de voix par rapport au nombre de sessions de vidéo. Avec toutes ces hypothèses, la relation (3.20) devient :

$$Q(\alpha) = (144n + 14000)\alpha^2 - (72n + 7432)\alpha + 72n + 432 \tag{3.25}$$

Pour optimiser les performances de notre approche de séparation voix/vidéo (maximiser l'amélioration du délai pour la voix tout, en minimisant la dégradation du délai pour la vidéo ou vice versa), il faut chercher à minimiser le facteur de dégradation de performance Q. Cela se traduit à la Figure 3.13 par les minima de la surface parabole z=Q(α,n) qui sont au-dessous du plan z=0.

L'amélioration des performances n'est possible que si Q(α,n)<0, ce qui donne la condition suivante :

$$0 < n < 8.79 \qquad \Rightarrow n_{\max} = 8 \tag{3.26}$$

Avec :

$$\frac{9n + 929 - \sqrt{-567n^2 - 50166n + 485041}}{9n + 875} < \alpha < \frac{9n + 929 - \sqrt{-567n^2 - 50166n + 485041}}{9n + 875}$$

Notons que cette restriction sur α vérifie la relation (3.22) sur l'amélioration du délai de la vidéo mais pas la relation (3.21) de l'amélioration du délai de la voix.

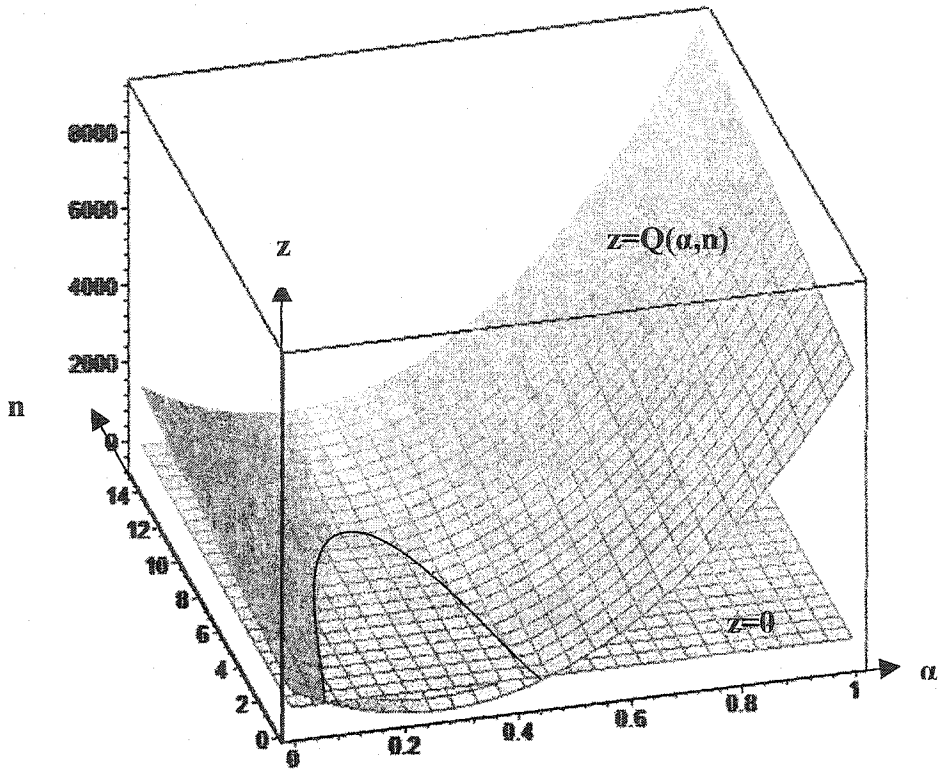


Figure 3.13 Facteur de dégradation de performance Q à minimiser

Ainsi, on peut avoir une diminution du délai de la vidéo qui est supérieure à l'augmentation du délai de la voix mais pas le contraire. Toutefois, il est possible de déterminer le poids optimal $\alpha_{optimal}$ qui doit être sur le minimum de la surface parabole (Figure 3.13).

$$\alpha_{optimal}(n) = 0.2656 - 2.10^{-4} n \tag{3.27}$$

Par conséquent, pour optimiser les performances du service de téléphonie multimédia, il faut que le poids optimal $\alpha_{optimal}$ donné à la classe de service DiffServ transportant le trafic de voix suive la courbe de la Figure 3.14. Il faut noter aussi que ce paramètre du poids α ne varie pas beaucoup en fonction du ratio n (nombre de sessions de voix / nombre de sessions de vidéo) et qu'à partir d'un nombre de sessions de voix 9 fois supérieur au nombre de sessions de vidéo, le délai de la voix se dégrade beaucoup

plus que l'amélioration du délai de la vidéo. Pour remédier à cet effet, nous avons établi le mécanisme suivant.

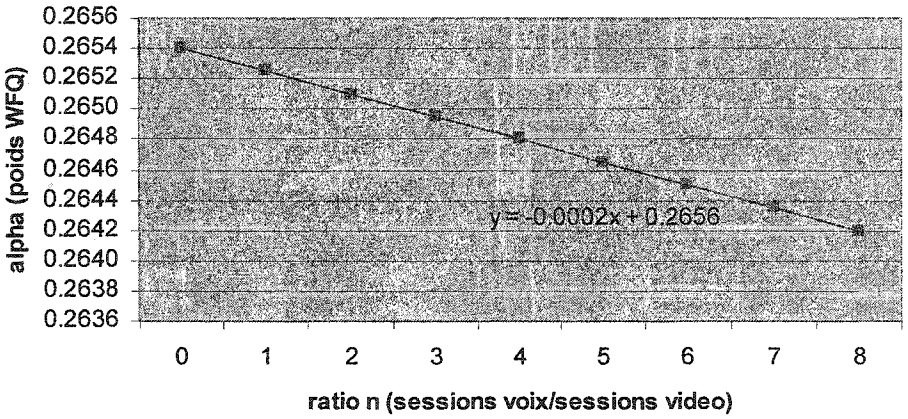


Figure 3.14 Courbe $\alpha_{optimal}=f(n)$

En supposant que l'établissement d'un E-LSP se fait en réservant une bande passante sur la dorsale MPLS, le partage équitable se fera de manière hiérarchique : chaque E-LSP aura le poids WFQ qui lui permettra de garantir sa bande passante réservée, et chaque type de trafic au sein d'un même E-LSP aura une fraction prédéterminée du poids WFQ. Ainsi, au cas où la demande en sessions de voix serait 9 fois supérieure à celle de la vidéo, on peut transporter l'excédent de sessions de voix sur un autre E-LSP de façon à offrir toujours la possibilité d'optimiser les performances de l'acheminement du service de téléphonie multimédia. Cela se fera en choisissant $\alpha_{optimal}$ comme fraction du poids WFQ pour la voix.

3.4 Indices de performance pour la téléphonie multimédia

Bien que dans l'étude précédente nous nous sommes intéressé exclusivement à l'évaluation du délai réseau, la QoS peut être évaluée par d'autres paramètres qui peuvent être critiques s'ils ne vérifient pas certaines contraintes. En effet, les autre

paramètres de QoS de bout en bout qui reviennent le plus souvent sont : *la variation du délai ou gigue, le débit et le taux de pertes de paquets*. Dans ce qui suit, on va étudier qualitativement ces paramètres, dans le contexte d'un service de téléphonie multimédia intégrant voix et vidéo, en discutant les contraintes auxquelles ils doivent se plier pour assurer un bon niveau de QoS.

3.4.1 Paramètres et contraintes de QoS pour la Voix

Le délai

Comme l'information vocale possède une caractéristique temporelle, le délai est le paramètre de QoS le plus critique pour la voix. En effet, les syllabes et les mots doivent être prononcés en introduisant un intervalle de temps de repos entre eux. Cette courte pause fait partie intégrante de la parole et elle a la même importance que les périodes d'activité vocale. Ainsi, une temporisation bien particulière doit être préservée afin de garder le bon enchaînement de la parole. Par conséquent, le fait de minimiser le délai unidirectionnel de bout en bout doit être considéré avec une grande importance lors de la conception et de l'implémentation des mécanismes de codage et de transmission de la voix sur le réseau. La recommandation G.114 de ITU-T [ITU-T, 2000] spécifie que, pour une bonne qualité de la voix, pas plus que 150 msec de délai de bout en bout unidirectionnel ne doit être toléré. Cette contrainte coïncide bien avec les spécifications du 3GPP pour la limitation de délai pour les services de la classe conversationnelle. Selon la même recommandation G.114 de l'ITU-T, un délai entre 150 msec et 400 msec peut être considéré comme acceptable dans certains cas, malgré la perception d'une dégradation de la qualité. Toutefois, un délai qui dépasse les 400 msec est toujours considéré comme intolérable.

La variance du delai (gigue)

Le second paramètre important pour la QoS de la voix est la variation de délai (appelée aussi gigue). En effet, la gigue peut causer des pauses imprévues entre les

prononciations qui peuvent affecter la continuité et par suite la compréhension de la parole. Elle doit souvent être maintenue à une valeur inférieure à 10 msec.

Le débit

Comme la parole est un signal sonore analogique qui varie lentement et très légèrement dans le temps avec une bande passante qui ne dépasse pas les 4kHz, sa numérisation, sa compression et son codage se font avec un taux constant. De ce fait, le débit moyen de transmission de la voix est constant pendant l'activité vocale. En effet, des systèmes de détection de silence ont été introduits dans les codecs (codeurs/décodeurs) de voix, et peuvent donc faire varier légèrement le débit de manière statistiquement prévisible. Généralement, les débits fixés pour les codecs de voix s'étalent entre 5.8 kbps et 64 kbps. Mais, les codecs les plus utilisés pour les communications mobiles ne dépassent pas les 16 kbps.

Le taux de perte de paquets

À cause de la nécessité d'avoir un délai très strict pour la voix, des protocoles de transport fiables telles que TCP ne peuvent être utilisés. D'autres protocoles tel que UDP sont utilisés pour satisfaire les contraintes de délai au coût de pertes de paquets. Les pertes causées par les congestions réseau sont inévitables et sont généralement de l'ordre de 5 à 20% sur l'Internet public. Toutefois, ils peuvent être compensées par les mécanismes de recouvrement de pertes des codeurs/décodeurs (codecs). Par exemple, un codec G.723.1 interpole une trame perdue en simulant les caractéristiques vocales de la trame précédente et en atténuant lentement le signal. Jusqu'à 10% de taux de perte de paquets semble n'avoir aucune influence notable sur la qualité de la voix, tel que reporté dans [ITU-T, 2001].

3.4.2 Paramètres et contraintes de QoS pour la Vidéo

Les exigences en matière de QoS pour les applications de vidéo interactive de bonne qualité sont presque similaires à ceux des applications de voix : un délai très bas

de l'ordre de 150 msec, une gigue minimale ne dépassant pas les 10 msec, ainsi qu'un taux de perte tolérable. Toutefois, il existe certaines différences au niveau des exigences de la vidéo par rapport à ceux de la voix. En effet, nous avons déjà vu que le trafic vidéo est différent du trafic de voix et que le taux binaire moyen pour la vidéo est généralement variable, alors que celui de la voix est toujours constant. En général, l'ITU-T laisse le choix dans l'utilisation d'un encodage à taux variable (VBR) ou à taux constant (CBR) pour la vidéo à l'application en question. Toutefois, la plupart des codecs vidéo pour la téléphonie mobile sont à taux variables, leur permettant ainsi de s'adapter aux ressources dynamiques de l'environnement sans fil. De plus, même avec un encodage à taux constant, les tailles des paquets de vidéo présentent une certaine variabilité due, non seulement à la fragmentation, mais aussi au temps de réponse du mécanisme rétroactif de quantification dynamique aux variations rapides des scènes vidéo [ITU-T, 2000]. En effet, pour éviter l'augmentation du débit d'encodage lors de l'accroissement de la complexité de la scène, le codeur effectue une quantification plus grossière qui a pour effet d'abaisser considérablement la qualité de visualisation. Ce mécanisme présente une utilisation intensive de la CPU qui est souvent très limitée dans les mobiles, ce qui décourage son implantation.

3.4.3 La contrainte de synchronisation voix/vidéo

Parfois, des sessions de vidéo conférence présentent des erreurs de synchronisation entre voix et vidéo. Ceci est connu sous le nom du « lipsync » ou *synchronisation de lèvres*. Étant donné que la lumière voyage plus vite que le son, la perception humaine est plus tolérante à un retard de traînée de l'audio par rapport à la vidéo dans une session de téléphonie multimédia, plutôt que le contraire. En effet, Holier et al. [1999] ont étudié l'effet de l'asynchronisme et ont trouvé que, lors de la visualisation d'un conférencier en train de parler, une avance de 150 msec de la voix par rapport à la vidéo est beaucoup plus perceptible qu'un retard de 300 msec de la voix par rapport à la vidéo. Comme il est souvent plus rapide de coder l'audio que la vidéo pour le même intervalle de temps, les trames de voix sont généralement transmises en tête des

trames vidéo, provoquant ainsi cet asynchronisme entre voix et vidéo. Cet effet peut s'accentuer surtout que les délais de transmission retardent beaucoup plus les longs paquets vidéo que les paquets de voix.

Généralement, cette contrainte de synchronisation ne fait pas partie des contraintes de QoS fournie par le réseau, du fait qu'elle s'effectue au niveau des couches supérieures. Souvent, c'est les couches applicatives terminales qui s'occupent de régler ce problème de synchronisation entre services distincts aux moyens d'estampillage et de mise en mémoire tampon. Cependant, l'effet de la différenciation voix/vidéo sur la synchronisation sera considéré dans notre évaluation de performance.

3.5 Modèles de sources de trafic de téléphonie multimédia

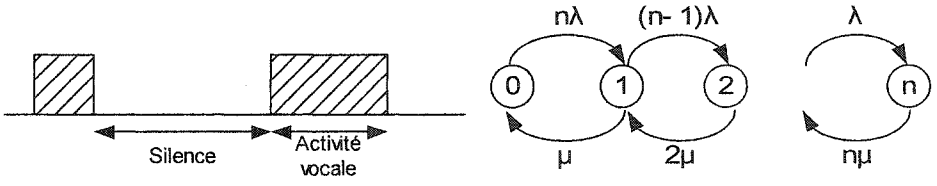
Les divergences retrouvées dans la caractérisation du trafic de voix et de vidéo nous conduisent à présenter les modèles de source de trafic qui vont être utilisés par la suite.

3.5.1 Modèle d'une source de trafic de Voix

La voix, analogique dans les réseaux 1G et numérique dans les réseaux 2G, est généralement transportée sur un circuit dédié établi au moment de la connexion. Ce circuit qui se présente sous forme d'une sous-bande de fréquences dans FDMA ou *slot de temps* dans TDMA, est construit à l'établissement de l'appel téléphonique et maintenu par une signalisation pendant toute la durée de la communication. Dans ce type de réseaux à commutation de circuits par multiplexage temporel ou fréquentiel, on a longtemps utilisé des modèles de trafic de voix qui font intervenir des lois d'Erlang et de Poisson. Toutefois, avec l'apparition des réseaux 3G faisant intervenir la commutation de paquets, la voix est remodelée pour être véhiculée dans des paquets plutôt que dans des circuits, afin de profiter de l'efficacité du multiplexage statistique de la commutation de paquets. De plus, le mécanisme de mise en paquets de la voix peut profiter des périodes de silence fréquentes dans la parole, pour ne transmettre que l'information de voix utile générée pendant les périodes d'activité vocale. D'où la

technologie de détection et de suppression de silence qui a été largement implémentée dans presque tous les codecs de voix. De ce fait, la modélisation d'une source de trafic de voix numérisée doit tenir compte des deux états : l'état « *talkspurt* » ou activité vocale (état ON) pour lequel des paquets de voix sont transmis et l'état *silence* (état OFF) pour lequel aucun paquet de voix n'est transmis sur le réseau. Des études menées par Brady [1969] ont montré que les périodes relatives à ces deux états sont distribuées suivant une loi exponentielle.

Ainsi, le modèle qui est communément accepté pour une source de voix est celle d'une chaîne de Markov à temps continu et à deux états discrets. Le temps de maintien de chacun des deux états suit une loi exponentielle de moyenne $1/\lambda$ pour l'état ON et $1/\mu$ pour l'état OFF (Figure 3.15a). Les valeurs couramment utilisées pour ces moyennes sont : $1/\lambda = 650$ msec et $1/\mu = 325$ msec. Pour modéliser plusieurs appels en cours, il est possible d'utiliser un simple processus de naissance et de mort, avec comme état le nombre d'appels en activité vocale (ON) (Figure 3.15b).



a) Modèle ON/OFF d'un appel simple **b) Modèle d'appels multiples**

Figure 3.15 Modèles de voix d'un simple appel et d'appels multiples

En pratique, pour modéliser une source de voix suivant un codec particulier reviens à configurer un modèle de voix générique avec les paramètres définies au Tableau 3.4.

Tableau 3.4 Paramètres de modélisation des différents codecs de voix

Codec de voix	G.729	G.722.2	G.723.1
Durée du silence (msec)	exp(0.650)	exp(0.650)	exp(0.650)
Durée de l'activité vocale (msec)	exp(0.325)	exp(0.325)	exp(0.325)
Nombre de trames par paquet	1	1	1
Longueur de la trame (sec)	10	20	20
Taille du "Lookahead" (sec)	5	7.5	7.5
Taux de codage (kbps)	8	40	5.3
Détection et suppression de silence	Actif	Actif	Actif

3.5.2 Modèle d'une source de trafic de Vidéo

Nous avons recensé un nombre très abondant de travaux qui se sont intéressés à la modélisation du trafic vidéo [Sen et al., 1989; Yegengolu et al., 1993; Hughes et al., 1995]. Toutefois, nous nous baserons sur l'article de Hughes et al. [1995] pour modéliser le trafic H.263 d'une source de vidéo téléphonie puisqu'il se base sur la vidéo compressée MPEG. Le caractère stochastique fractal d'une source de vidéo MPEG se présente sous la forme d'une décomposition en multi échelles (Figure 3.16). Le flux vidéo se décompose en plusieurs scènes, une scène se décompose en plusieurs GOP (Group Of Pictures) ou groupe d'images, un GOP se décompose en plusieurs trames vidéo de différents types (Intra codées *I*, Prédictives *P*, Bidirectionnelle prédictives *B*). On a étudié les distributions des tailles des différents éléments (trames, GOP, scènes) et montré que la longueur de la scène (d*N) suit une loi géométrique, que les trames *I*, *P* et *B* suivent des lois log-normales, et que la taille de la trame *I* de chaque début de GOP subit une auto régression dans une même scène.

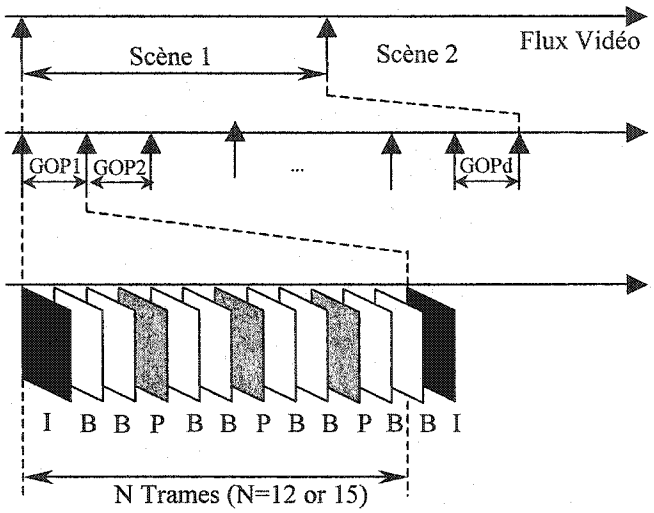


Figure 3.16 Modèle vidéo MPEG

Les trames *I* sont modélisées comme suit :

Posons I_0 la taille de la première trame *I* (ou le premier GOP) pour la scène en cours, *n* étant l'index de la GOP dans une même scène et AR_ε le coefficient d'auto-régression. On a:

- $\delta I_0 = \delta L_{-1} = 0$
- $I_n = I_{n-1} + \delta I_n$
- $\delta I_n = a_1 * \delta I_{n-1} + a_2 * \delta I_{n-2} + AR_\varepsilon$

Nous aborderons dans le prochain chapitre les détails d'implémentation de ce modèle en l'intégrant dans le modèle d'une station UMTS.

CHAPITRE 4
ÉVALUATION DE PERFORMANCE DES SERVICES DE TÉLÉPHONIE MULTIMÉDIA

Notre étude [Ben Ali et al., 2003] consiste à évaluer les performances du trafic de téléphonie multimédia soumis à notre algorithme raffiné de mise en correspondance de QoS entre domaine UMTS et domaine dorsal DiffServ, de les comparer avec les performances obtenues avec l'algorithme standard, ainsi qu'à dégager l'influence des choix des différents paramètres d'ordonnancement sur la QoS fournie à ce trafic critique. Notre objectif principal est de montrer l'impact réel de l'idée de séparation entre trafic de voix et trafic de vidéo sur les performances et le contrôle de la QoS des services conversationnels de l'UMTS. En outre, nous vérifierons les déductions que nous avons établies dans l'analyse théorique sur les conditions auxquelles doivent se plier les paramètres de différenciation pour optimiser les performances. Nous commencerons par un bref aperçu sur l'environnement de simulation OPNET. Ensuite, nous procèderons à une étude des différentes contraintes et hypothèses à émettre pour une modélisation efficace du trafic conversationnel de l'UMTS. Puis, nous exposerons l'implémentation et la configuration sous OPNET des sources de trafic utilisées pour piloter les simulations. Nous présenterons par la suite l'implémentation des indices de performance les plus pertinents qui serviront à récolter les différents résultats de simulation sous forme de statistiques. Nous exposerons ensuite les résultats des scénarios simulés accompagnés d'interprétations détaillées. Enfin, nous procéderons à la validation du modèle générique et modulaire que nous avons développé pour la génération du trafic de vidéo téléphonie.

4.1 Choix de l'outil *OPNET modeler* et modèles utilisés

Dans le chapitre précédent, nous avons abordé une étude analytique qui se base sur un modèle mathématique simplifié et très approximatif du trafic de voix et de vidéo (agrégat de sessions contraintes par le modèle du seau à jetons). Toutefois, malgré que cette étude a permis de dégager les conditions nécessaires à appliquer aux paramètres de différenciation de trafic afin d'assurer l'optimisation des performances, elle ne peut donner une quantification réaliste du gain de performance. En effet, d'un coté l'étude analytique n'a pu être effectuée que pour évaluer des délais réseau au pires cas. Et d'un autre coté, le comportement réel d'un agrégat de sessions de trafic de type bien spécifique (voix ou vidéo téléphonie) traversant des accès radio W-CDMA ne peut être étudié par des résolutions analytiques qui s'avèrent insuffisantes dans ce cas, vu la complexité de la modélisation mathématique du système (trafic et réseau). En outre, une évaluation de performance d'un système réel en utilisant des prises de mesures est impraticable dans ce contexte, puisqu'il nous est impossible de mettre en œuvre un réseau UMTS et un réseau dorsal exclusivement pour des tests. Ainsi, la simulation est l'approche la plus adaptée à notre étude de performance. Pour cela, nous n'avons pas eu beaucoup de choix sur l'outil utilisé puisqu'on s'est orienté directement vers l'environnement de simulation OPNET modeler qui, à notre connaissance, est le seul simulateur sur le marché à fournir une implémentation quasi complète du modèle UMTS.

OPNET modeler© [OPNET] développé par *OPNET technologies Inc*, est un logiciel commercial offrant un environnement de modélisation, de simulation et d'analyse pour les protocoles de communications, les équipements réseaux et les systèmes réseautiques de bout en bout. Il intègre un noyau de simulation à évènements discrets qui, rappelons-le, est un ordonnanceur et exécuteur d'évènements réseau successifs dans le temps. Toutefois, les simulations à évènements discrets sont souvent très coûteuses en temps CPU surtout avec des modèles assez complexes tel que ceux d'un transport ATM ou encore ceux des liens radio W-CDMA qui font intervenir les phénomènes d'interférences, de bruits, de chemins multiples, etc. Pour réduire le temps

de simulation, OPNET combine l'approche analytique avec l'approche par évènements discrets dans ce qui est appelé la *micro simulation* ou *simulation hybride* présentée plus loin. En outre, *OPNET modeler* offre un outil puissant de modélisation permettant d'intégrer des modèles très détaillés pour représenter efficacement la plupart des technologies de communications.

L'environnement de modélisation de OPNET se compose de projets et de scénarios. Un projet regroupe un ou plusieurs scénarios de simulation. Dans un scénario, on trouve différents éléments tels que des modèles de réseau, des fichiers de sondage de résultats, des séquences de simulation, des analyses de configuration et des fichiers de vecteurs de résultats.

Dans le cadre de notre travail, l'architecture d'un réseau 3G/UMTS étendu est modélisée en totalité, en englobant un certain nombre de modèles fournis avec OPNET:

- *Modèle UMTS* (Figure 4.1) qui intègre les éléments d'un réseau d'accès UMTS qui, en plus des modèles spécifiques à UMTS tel que RLC et MAC, utilise le modèle radio pour la transmission W-CDMA entre UE et UTRAN, ainsi que le modèle ATM pour le transport à l'intérieur de l'UTRAN et entre UTRAN et CN.

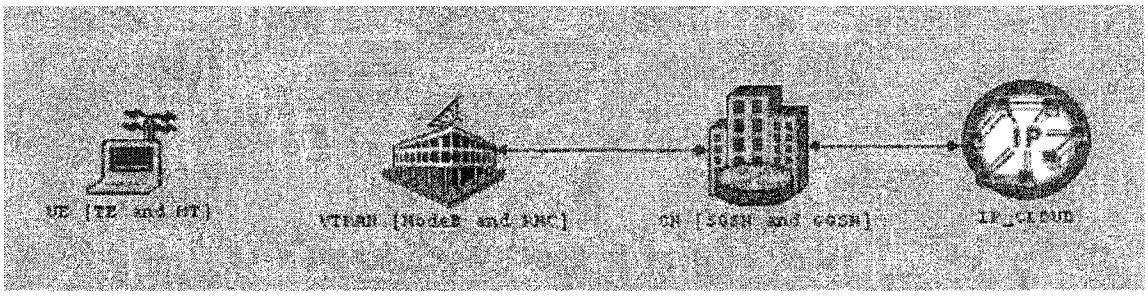


Figure 4.1 Une représentation du modèle UMTS sous OPNET (domaine paquets)

- *Modèle MPLS* (Figure 4.2) qui intègre les éléments réseau et de configuration d'un réseau dorsal à commutation d'étiquettes en se basant sur le protocole de routage IP sous-jacent.

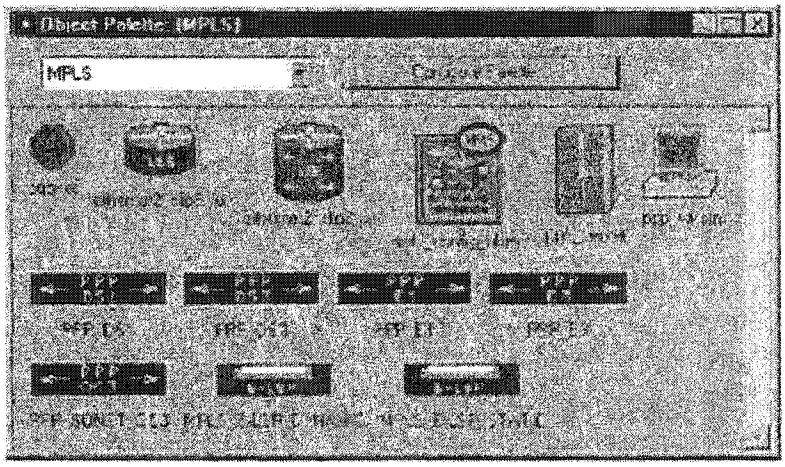


Figure 4.2 Éléments réseau du modèle MPLS sous OPNET

- *Modèle DiffServ sur MPLS* (Figure 4.3) qui intègre les algorithmes de classification, de marquage EXP, et d'ordonnancement (PQ, WFQ, etc.) ainsi que les modèles de jonction LSP (LSP trunks) pour la différenciation de services.

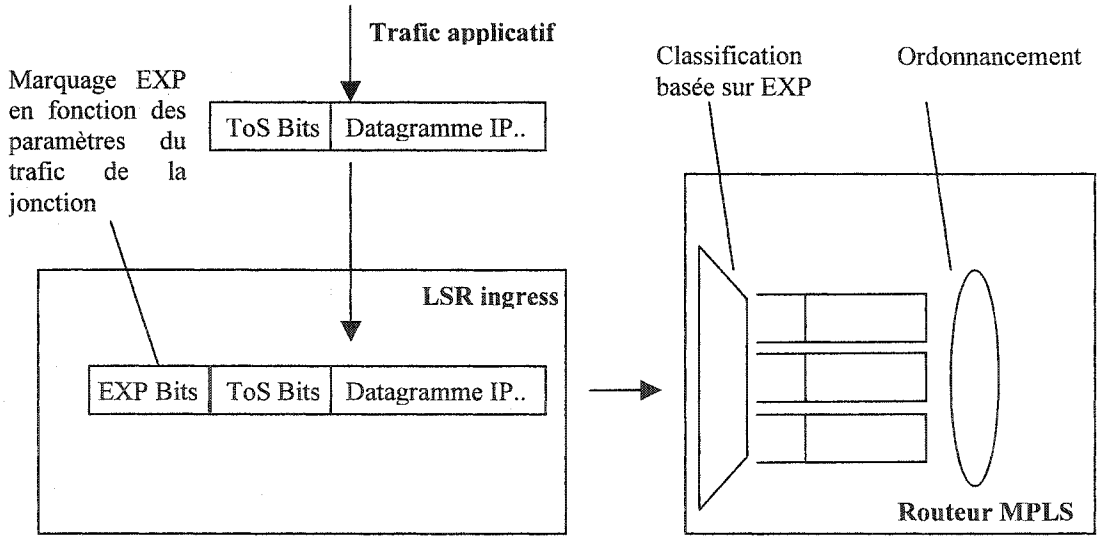


Figure 4.3 Modèle DiffServ sur MPLS

OPNET modeler, par son architecture logicielle modulaire, ouverte et extensible, offre aussi un environnement complet de développement pour concevoir et implanter de

nouveaux modèles de processus qui sont à la base de tout autre type de modèle. En effet, un modèle de réseau (ex : WLAN, UMTS, LAN, WAN, etc.) se décompose en différents modèles de nœuds (ex : Commutateur, concentrateur, routeur, station de travail, unité mobile, etc.) et de liens (ex : PPP, SONET, ATM, FR, Ethernet, etc.). Un modèle de nœud se décompose en différents modules en couche (ex : PHY, MAC, IP, TCP, UDP, etc.) qui eux-mêmes intègrent des modèles de processus reproduisant leur fonctionnement. Les modèles de processus sont souvent représentés sous OPNET par des machines à états finis (MEF) appelées aussi *diagrammes état-transition*. Comme nous allons le voir plus loin, nous serons intéressés par l'implémentation d'une MEF dans un module de couche applicative d'une station UMTS pour la génération d'un trafic de vidéo téléphonie.

4.2 Contraintes et hypothèses de modélisation du trafic

Une étape cruciale dans toute évaluation de performance par simulation est la modélisation du trafic. En effet, c'est souvent le trafic qui est à la base de tout processus de simulation et on doit toujours commencer par poser les bonnes contraintes pour modéliser le bon type de trafic. De plus, vu que le trafic est influencé par le comportement de l'usager qui peut s'avérer complexe et très sporadique, il est primordial de spécifier des hypothèses simplificatrices.

4.2.1 Trafic sans fil vs. trafic filaire

Les services de téléphonie multimédia basés IP peuvent être initiés à partir de terminaux fixes avec accès à large bande (LAN, ADSL, Câble) ou à partir des terminaux mobiles avec accès sans fil (GPRS, UMTS). L'étude de l'effet du multiplexage stochastique sur les caractéristiques du trafic agrégé montre que le type du réseau (filaire ou sans fil) a un grand impact sur ces caractéristiques. En effet, partant des mêmes applications générant les mêmes flux conversationnels à un taux d'inter-arrivée constant (voix et vidéo), on remarque que leur agrégation dans un réseau filaire produit un trafic total qui est également à inter-arrivée constante, alors que dans un réseau sans fil le

trafic total est à inter-arrivée variable (Figure 4.5). La variabilité du trafic est beaucoup plus remarquable pour la vidéo (Figure 4.4) que pour la voix. Ici on défini la variabilité par le rapport écart-type/moyenne. Elle est égale à 0,0029 pour la voix et 0,1246 pour la vidéo. Puisque notre réseau d'accès UMTS fait intervenir plusieurs technologies de transmission (accès radio W-CDMA, RLC, ATM), il nous est difficile de déterminer la cause exacte qui explique ces constatations. Toutefois, nous savons que la transmission radio des paquets IP introduit des retards plus ou moins aléatoires dus essentiellement aux interférences, chemins multiples, etc. Ceci est d'autant plus important pour les gros paquets vidéo qui subissent des erreurs de transmissions plus fréquentes, ce qui explique l'allure de la courbe du trafic de vidéo qui est plus fluctuante que celle du trafic de voix.

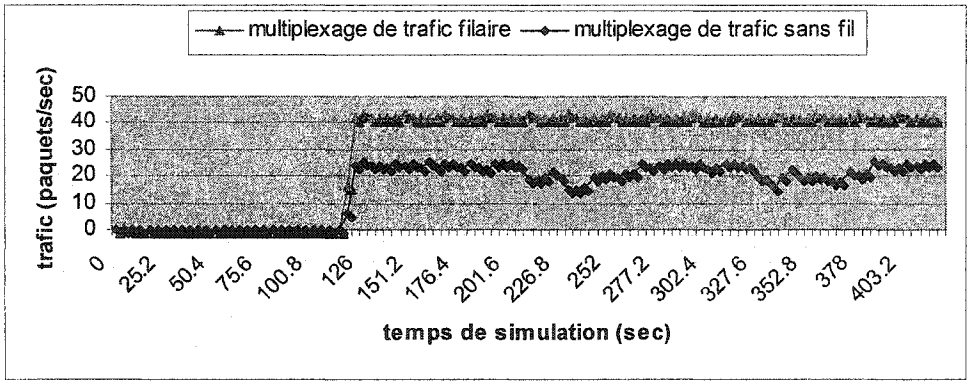


Figure 4.4 Impact de l'accès radio de l'UMTS sur l'agrégation du trafic de vidéo

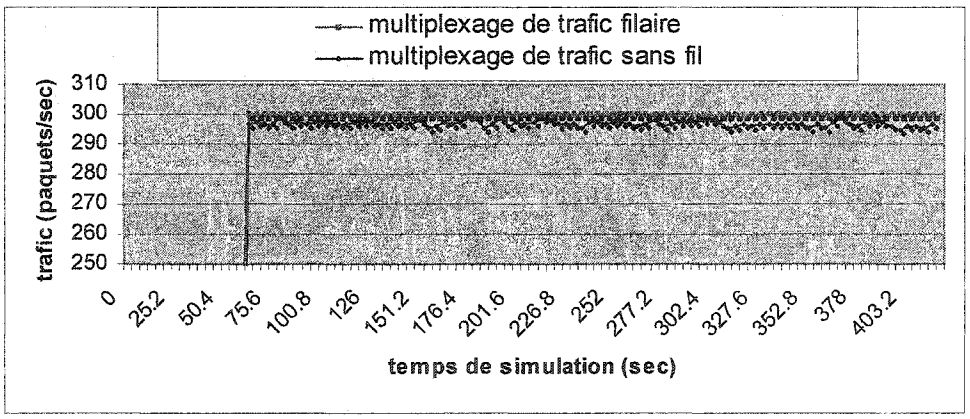


Figure 4.5 Impact de l'accès radio de l'UMTS sur l'agrégation du trafic de voix

4.2.2 Trafic dorsal

Vu que notre algorithme de mise en correspondance raffinée est implémenté dans le routeur GGSN qui se trouve à la frontière du réseau dorsal, nous sommes contraints à modéliser un trafic dorsal *très sporadique* et *de volume très important*. Plusieurs solutions ayant leurs avantages et leurs inconvénients se présentent :

- *Modèle de trafic auto similaire (module RPG d'OPNET)* Nous savons que l'agrégation d'un nombre important de sources ON/OFF telles que celles de la voix produit un trafic auto similaire. De plus, l'agrégation de plusieurs sources auto-similaires telle que la vidéo produit un trafic qui a les mêmes propriétés. Ainsi, à première vue, il nous parait judicieux d'utiliser le modèle de génération de trafic multi-fractal d'OPNET pour générer et simuler le trafic de voix et de vidéo à l'entrée de la dorsale. Toutefois, malgré que plusieurs processus multi-fractals sont implémentés avec des caractéristiques distinctives telles que le paramètre de Hurst, il nous est difficile d'identifier le modèle multi-fractal le plus adapté pour un agrégat de trafic de voix ou de vidéo téléphonie.

- *Modèles de trafic d'arrière-plan (micro-simulation)* Les simulations à évènements discrets sont souvent très coûteuses en temps de simulation pour des trafics de très grand volume tel que le trafic dorsal. De ce fait, on a souvent tendance à combiner la rapidité des simulations analytiques avec l'exactitude des simulations à évènements discrets dans ce qu'on appelle la *simulation hybride*. Dans une micro-simulation, il est possible de modéliser un trafic implicite et statique d'arrière-plan en tant que des demandes avec différents paramètres tels que l'origine, la destination, le type de service, la classe de service, la PDF des tailles des paquets, la PDF des inter-arrivées, etc. Ainsi, pour éviter la grande lenteur de la simulation à évènements discrets d'un trafic dorsal, il nous est possible avec cette technique de modéliser un grand nombre de flots de manière statique comme trafic d'arrière-plan en donnant les propriétés spécifiques et les classes de services appropriées pour voix et vidéo. Toutefois,

l'utilisation de ce type de simulation dans notre cas particulier d'un trafic provenant d'un accès radio UMTS engendre une perte dans l'exactitude du modèle de trafic. En effet, le trafic d'arrière-plan est modélisé d'une manière statique sur la couche IP seulement, ce qui rend le modèle de trafic assez inexact puisque nous voulons modéliser le comportement d'un agrégat de sources de trafic sans fil UMTS qui, comme nous l'avons déjà vu, est différent de celui d'un trafic filaire. De plus, la simulation hybride ne permet pas de représenter la dynamique des algorithmes d'ordonnancement.

- *Modèle de trafic explicite (simulation à évènements discrets)* Avec ce modèle de simulation à évènements discrets, malgré que nous aurons des temps de simulation assez élevés, nous sommes plus confiant à l'égard de l'exactitude de notre trafic dorsal et les caractéristiques qu'il s'est forgées en transitant par différents éléments réseau du domaine UMTS avec la QoS conversationnelle. En effet, le modèle UMTS d'OPNET, constitué des différents modules faisant intervenir machines à états finis et processus évènementiels, nous aide à produire un trafic à la frontière du réseau très proche de la réalité. En outre, cela nous permettra de récolter des statistiques sur les délais partiels encourus en transitant par une partie spécifique du réseau et non seulement les délais de bout en bout comme c'est le cas pour un trafic d'arrière-plan. Ainsi, nous pourrons avoir une idée sur la contribution du réseau d'accès UMTS et du réseau dorsal au délai de bout en bout.

4.2.3 Hypothèses sur le trafic

Mobilité des usagers

L'étude de la différenciation et de l'ordonnancement de la voix et de la vidéo sur la dorsale ne peut être affectée par la mobilité des usagers qui génèrent le trafic en question. En effet, en s'intéressant au trafic agrégé au niveau du GGSN et en supposant une mobilité totalement transparente, la micro-mobilité (d'un RNC à un autre) et la

macro-mobilité (d'un SGSN à un autre) n'introduisent aucun effet sur notre algorithme de séparation et d'ordonnancement des services conversationnels. Ainsi, une de nos hypothèses est de considérer un trafic de stations UMTS qui sont immobiles tout au long de la simulation.

Types de trafic impliqués

Dans presque toutes les études de mise en correspondance de QoS, on fait intervenir des trafics de différents types. Toutefois, notre étude comparative s'intéresse essentiellement aux services conversationnels de l'UMTS qui englobent les deux types de trafic : la voix et la vidéo. Les autres types de trafic peuvent être considérés, mais vu que l'ordonnancement WFQ est caractérisé par *son équité*, l'impact des classes de trafic entre elles est négligeable. En effet, WFQ a le pouvoir de séparer les besoins en bande passante garantie des différents classes, comme nous l'avons vu au chapitre 2.

4.3 Configuration des sources de trafic conversationnel UMTS

Comme on l'a déjà vu précédemment, qu'ils s'agissent de codecs G.722 ou G.723 (resp. de codecs MPEG2/H263 ou MPEG4), ils ont le même comportement dans la génération du trafic de voix (resp. de vidéo). Ainsi, il est possible de créer un modèle générique pour chacun des deux types de trafic de voix et de vidéo, en définissant les caractéristiques des tailles des trames et des temps d'inter-arrivée comme des attributs de configuration.

4.3.1 Source de trafic de vidéo téléphonie (Générateur de flux MPEGx)

On a vu que les caractéristiques d'un trafic vidéo ne peuvent être représentées par de simples distributions sur les inter-arrivées et les tailles des paquets. Le caractère multi-fractal de ce trafic peut être reconstitué par un modèle de processus OPNET. Trois interruptions OPNET auto-programmées (suivant le motif de génération IBBPBBPBBP...) sont définis, un pour chaque type de trame vidéo I, P et B. Ainsi,

l'occurrence des GOPs, des changements de scènes et leurs différentes distributions sont implémentées suivant le modèle présenté par Hughes [1995] et résumé au chapitre 3.

La Figure 4.6 montre la machine à états finis développée pour modéliser un module applicatif générique d'un service de vidéo-téléphonie émetteur et récepteur. Le récepteur est un puits de trafic permettant d'analyser les paquets vidéo reçus (leur date d'arrivée et leur numéro de séquence) pour construire des statistiques sur le délai, la gigue et la synchronisation. L'émetteur est une source de trafic multi-fractal MPEG générique et paramétrable avec différents attributs. Pour nos simulations, on a configuré ce modèle générique pour générer un trafic MPEG4. Ainsi, les valeurs de ces attributs qui sont données au Tableau 4.1 ont été extraites en analysant un fichier trace [Trace-MPEG4] d'une session vidéo codée et compressée en MPEG4 et intégrant des scènes de type « *head and shoulders* » qui sont assez courantes dans les vidéo-conférences.

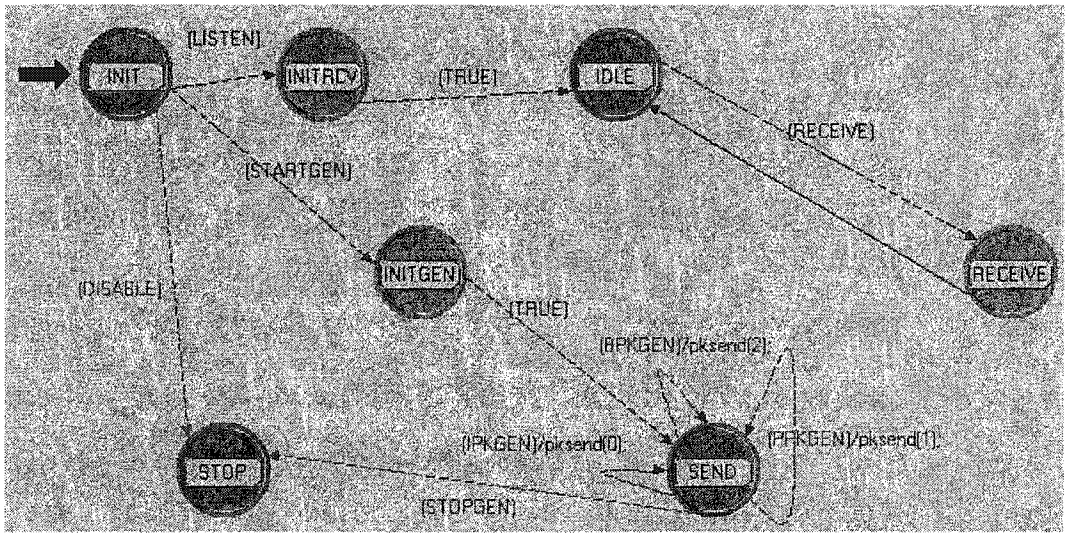


Figure 4.6 MEF d'une source/puits de trafic vidéo

Tableau 4.1 Configuration du modèle de video-téléphonie suivant MPEG4

Attribut	Unité	Distribution	Paramètre1	Paramètre2
Taille de I	Bits	Lognormale	μ=2549	ν=141932
Taille de P	Bits	Lognormale	μ=187	ν=25593
Taille de B	Bits	Lognormale	μ=62	ν=2991
Taille de la scène	-	Géométrique	p=0.1	-
M (occurrence de P)	-	Constante	3	-
N (taille du GOP)	-	Constante	12	-
a_1 (autoreg1)	-	Constante	0.53	-
a_2 (autoreg2)	-	Constante	0.15	-
ARε	-	Normale	μ=0	ν=392
Taux de génération	Trame/sec	Constante	25	-
Début de génération	Sec	Constante	120	-
Fin de génération	Sec	Constante	300	-
Direction	-	-	Source/puits	-
Adresse destination	-	-	# Station	-
Port source	-	-	VideoConf	-
Port destination	-	-	VideoConf	-
ToS	-	-	5	-

4.3.2 Source de trafic de voix (Générateur de flux G.72x)

Le modèle générique d'une source de trafic de voix a été paramétré pour générer un trafic conforme à celui issu d'un codec G.723 (Tableau 4.2).

Tableau 4.2 Configuration du modèle de voix suivant G.723

Attribut	Unité	Distribution	Paramètre1	Paramètre2
Taux de génération	Trame/sec	Constante	50	-
Taille d'une trame	Bits	Constante	160	-
$1/\lambda$ (*silence)*	Msec	Exponentiel	650ms	-
$1/\mu$ (*talkspurt*)	Msec	Exponentiel	352ms	-
Port source	-	-	Voix	-
Port destination	-	-	Voix	-
ToS	-	-	6	-

4.3.3 Intégration des sources de trafic dans une station UMTS

Afin de pouvoir router le trafic et différencier entre les types de paquets, il est nécessaire d'intégrer le générateur de trafic dans le modèle en couches d'une station UMTS (Figure 4.7). Généralement, les applications de téléphonie IP telles que SIP et H.323 utilisent RTP sur UDP comme transport afin de bénéficier de sa grande vitesse de transmission, contrairement à TCP qui se caractérise par des connexions et des transmissions relativement lentes dues principalement aux attentes d'acquittements et aux algorithmes de contrôle de congestion. Ainsi, notre module de génération de flux de téléphonie multimédia invoquera directement les services de la couche UDP de la station UMTS. Ceci se fera par la configuration et l'installation de l'ICI OPNET udp_command_v3. Rappelons que dans le modèle OSI, une ICI est une Information de Commande de l'Interface qui doit être ajoutée à l'unité de données de service SDU (trame de voix ou de vidéo dans notre cas) pour former l'unité de données d'interface (IDU) qui sera transformée en unité de données de protocole PDU (paquet UDP dans notre cas) : SDU + ICI = IDU => PDU.

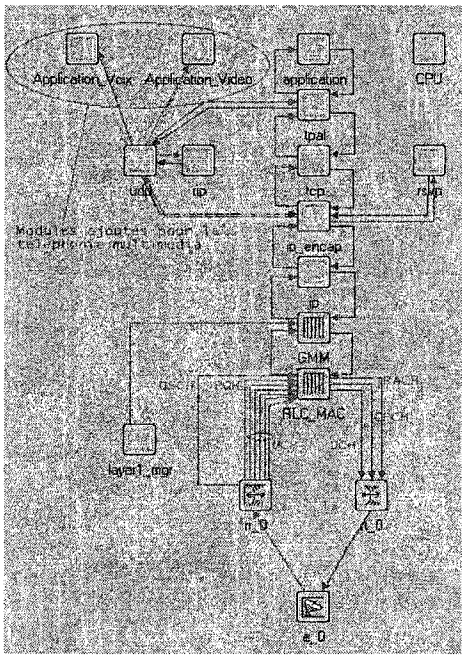


Figure 4.7 Intégration de la téléphonie multimédia dans une station UMTS

4.3.4 Mise en correspondance des services voix et vidéo téléphonie avec la classe conversationnelle UMTS

Afin de s'assurer que les paquets générés par les sources de trafic de voix et de vidéo sont transportés à travers l'accès UMTS dans des contextes PDP de type conversationnel, il faut mettre en correspondance le champ ToS (*Type of Service)* du paquet de voix ou de vidéo avec la classe conversationnelle au niveau de la station UMTS et plus exactement entre les modules des couches IP et MAC. Ainsi, cette mise en correspondance de QoS locale se fait comme présenté au Tableau 4.3.

Tableau 4.3 Mise en correspondance de QoS locale à la station UMTS

Flux	Type de Service (ToS)	Classe UMTS (code QoS)
Voix Téléphonique	*Interactive Voice* (ToS=6)	Conversationnel (0)
Vidéo Téléphonie	*Interactive Multimedia* (ToS=5)	Conversationnel (0)

4.4 Plan d'expérience

Après avoir défini les modèles de trafic, nous allons présenter dans ce qui suit le plan d'expérience que nous avons suivi pour construire les différents scénarios de simulation. D'abord, nous décrivons les hypothèses que nous avons fixées pour l'ensemble des expériences et scénarios :

- Topologie de deux réseaux d'accès UMTS interconnectés avec un réseau dorsal IP (Figure 4.8) ;
- $n_A = 50$ clients de voix et $n_B = 10$ clients de vidéo (ratio = 5 conforme au taux de pénétration du service vidéo-conférence selon l'UMTS forum [UMTS-Forum, 1999a]) ;
- Facteur d'utilisation des liens dorsaux = $\rho = 0.9$;
- Client de voix G.723 = (ρ_A = 5.3 kbps , b_A = 160 bits) ;
- Client de vidéo MPEG4 = (ρ_A = 51 kbps , b_A = 8500 bits) ;

- Temps de simulation = 5 minutes ;
- Temps de génération de trafic voix/vidéo = 180 sec (durée moyenne d'un appel HIMM/conversationnel selon le forum UMTS [UMTS-Forum, 1999a]).

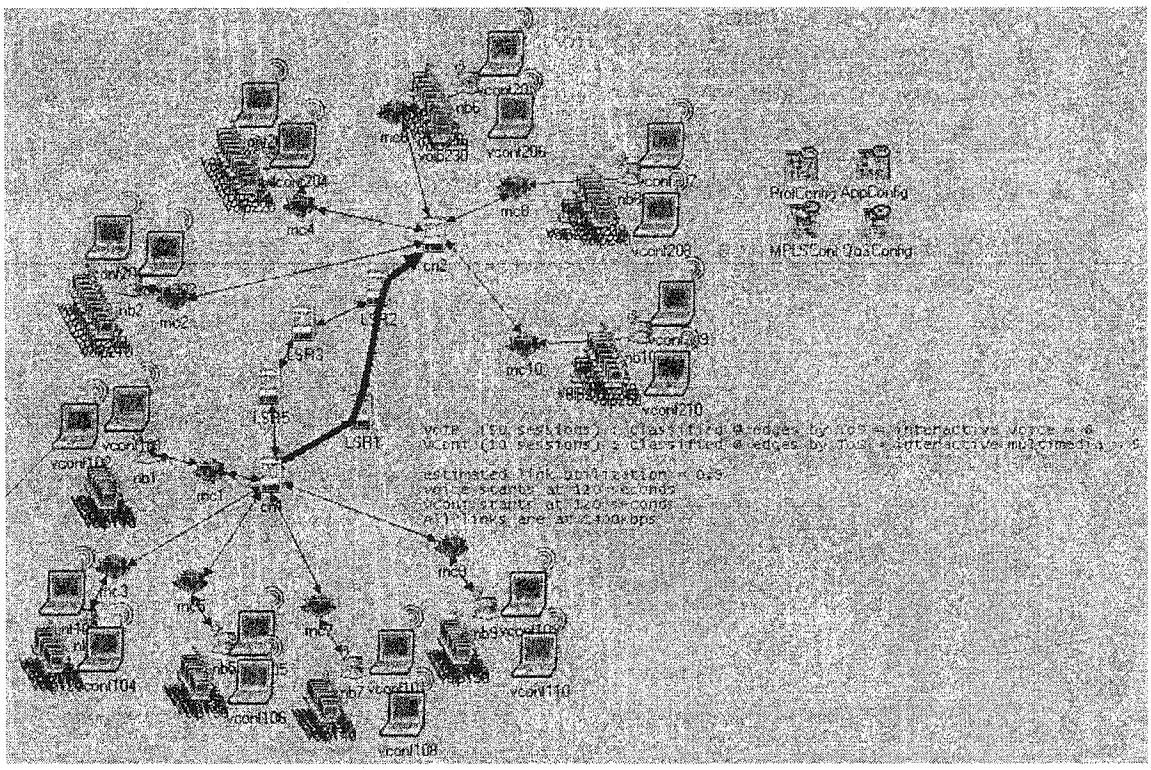


Figure 4.8 Topologie du modèle de réseau commune à tous les scénarios

Essentiellement, notre plan d'expérience (Tableau 4.4) se base sur la variation des disciplines d'ordonnancement retenues et sur leurs paramètres de configurations éventuelles. Ainsi, la discipline FIFO représente l'approche de l'algorithme de mise en correspondance standard qui ne différencie pas entre voix et vidéo-téléphonie mais plutôt sert les deux types de trafic indistinctement à la manière du premier-arrivé premier-transmis. La discipline PQ est une configuration sous laquelle le trafic de voix a

la priorité exclusive de transmission sur le trafic de vidéo-téléphonie. La discipline WFQ, quant à elle, est configurée avec différentes valeurs du poids α, $0 < \alpha < 1$ pour la classe DiffServ véhiculant le trafic de voix et $1-\alpha$ pour celle véhiculant le trafic de vidéo-téléphonie.

Pour simplifier notre étude de performance, nous avons jugé légitime de ne conduire des expériences que pour les configurations intéressantes. Ainsi, pour déterminer le délai dorsal, nous avons été obligé d'essayer un nombre important de paramètres de poids α dans le but de déterminer celui qui optimise les performances et par conséquent vérifier les résultats que nous avons établis dans la partie théorique au chapitre 3. Pour les résultats sur les délais de bout en bout, le but n'est pas de vérifier l'impact de la variation de la configuration de l'ordonnancement puisque cela affecte le délai dorsal, mais plutôt pour évaluer la contribution de la dorsale et du réseau d'accès UMTS au délai global (de bout en bout).

Dans ce plan d'expérience, nous nous sommes limités à récolter les statistiques les plus importantes représentant les trois indices de performances les plus critiques pour les services de téléphonie multimédia qui sont : le délai, la gigue et la synchronisation. Nous avons écarté les statistiques sur le débit parce qu'avec 5.3 Kbps pour le codec de voix et 51 Kbps pour le codec de vidéo, il ne représente qu'une petite fraction de la bande passante disponible pour un usager. De plus, le débit ne peut être affecté que s'il y a des pertes de paquets, puisque le transport utilisé est UDP qui contrairement à TCP, ne possède aucun contrôle sur la congestion ni sur la vitesse de transmission. Il faut noter aussi que le réseau dorsal a été dimensionné pour avoir un taux d'utilisation $\rho = 0.9$ de telle sorte qu'il n'y ait pas de pertes de paquets. Donc s'il y a des pertes, ça sera essentiellement dû au réseau d'accès radio, ce dernier n'entre pas dans nos différents choix de configuration de l'ordonnancement.

Tableau 4.4 Configuration des différents scénarios et statistiques récoltées

Scénario	1	2	3	4	5	6	7	8	9	10	11
Ordonnancement	FIFO	PQ	WFQ	WFQ	WFQ	WFQ	WFQ	WFQ	WFQ	WFQ	WFQ
α (poids de la voix)	-	-	0.15	0.20	0.25	0.33	0.50	0.66	0.75	0.80	0.85
Délai dorsal voix/vidéo	X	X	X	X	X	X	X	X	X	X	X
Délai de bout en bout voix/vidéo	X	X		X			X			X	
Gigue voix / vidéo	X	X		X			X			X	
Synchronisation	X	X					X				

4.5 Statistiques sur les indices de performance

Comme il peut y avoir plusieurs définitions pour les mêmes indices de performance, nous présenterons dans ce paragraphe la façon avec laquelle les indices que nous avons retenus lors de l'établissement du plan d'expérience vont être calculés :

- *Le délai de bout en bout*

Cette statistique est recueillie à chaque réception par le destinataire d'un nouveau paquet d'un flux de téléphonie (voix ou vidéo) bien particulier et représente le délai écoulé entre la date de création d'un paquet par l'application source et la date de son arrivée à l'application cible. Sous OPNET, l'implémentation de cette statistique se fait dans la routine déclenchée par l'interruption OPC_INTRPT_STRM.

- *La gigue*

D'une façon similaire au délai, les statistiques sur la gigue sont récoltées à chaque réception d'un nouveau paquet et se calculent comme la valeur absolue de la différence entre les délais encourus par deux paquets successifs.

- *La synchronisation*

Cet indice de performance est évalué en suivant l'évolution dans le temps du numéro de séquence des paquets de voix et de vidéo reçus par l'application cible. Pour implémenter ces statistiques, on a ajouté un champ dans le format d'un paquet de voix/vidéo représentant un numéro de séquence avec une incrémentation automatique à chaque création. Avec une taille de 16 bits qu'on a fixée pour ce nouveau champ, il faut plus que 43 minutes de flux continu à 25 paquets/sec pour épuiser tout l'espace de numérotation sans reboucler, ce qui est largement suffisant puisque nos simulations ne dépassent pas les dizaines de minutes.

4.6 Résultats et interprétation

4.6.1 Impact de l'algorithme d'ordonnancement et de ses paramètres sur les performances de séparation voix/vidéo

La première chose qu'on remarque à partir des premiers résultats sur les performances obtenues lors de la séparation entre voix et vidéo (Figure 4.9 et 4.10) est que l'amélioration des délais de l'un engendre la dégradation des délais de l'autre dans tous les cas de figure. Ce qui est parfaitement prévisible puisque la ressource partagée par les deux types de trafic est toujours la même (la même bande passante des liens dorsaux ayant la même utilisation dans tous les cas de figure) et que la seule différence est la manière d'accorder les priorités et les poids de partage de cette bande passante à l'aide de l'algorithme d'ordonnancement. Ainsi, il est tout-à-fait évident d'augmenter le poids d'une classe DiffServ correspondante au trafic (voix ou vidéo) que nous voulons améliorer. Toutefois, si nous nous intéressons à un trafic de vidéo-téléphonie, il se trouve qu'il est souvent associé au niveau d'une session applicative à un trafic de voix. Donc, l'amélioration de la QoS fournie à l'application de téléphonie multimédia doit tenir compte de la QoS fournie aux deux types de trafic conjointement (voix et vidéo). L'étude théorique que nous avons faite au chapitre précédent avait comme but de déterminer les conditions auxquelles l'algorithme d'ordonnancement doit se plier pour

améliorer le délai de la vidéo sans trop dégrader le délai de la voix ou vice versa. La Figure 4.10 montre que le facteur de dégradation de performances $Q(\alpha)$ est négatif (interprété par le fait que la diminution du délai de la vidéo est supérieure à l'augmentation du délai de la voix) pour les valeurs de α = 0.2, 0.25 et 0.33, ce qui confirme les résultats théoriques présentés sur la courbe parabole (cf. Figure 3.11). De plus, la valeur minimale de $Q(\alpha)$ est obtenue avec α = 0.25 (Figure 4.10) qui se rapproche du $\alpha_{optimal}$=0.26 de la théorie (cf. Figure 3.12). Comme nous l'avons déjà vu au chapitre 3, il est impossible d'avoir une diminution du délai de la voix qui est supérieure à une augmentation du délai de la vidéo. Toutefois, l'amélioration du délai de la voix qui minimise le facteur $Q(\alpha)$ correspond à un α = 0.85 ainsi que pour l'ordonnancement PQ selon la Figure 4.10. Ce qui est vérifié dans la théorie par un α (poids de la classe relative au trafic de voix) qui est très grand.

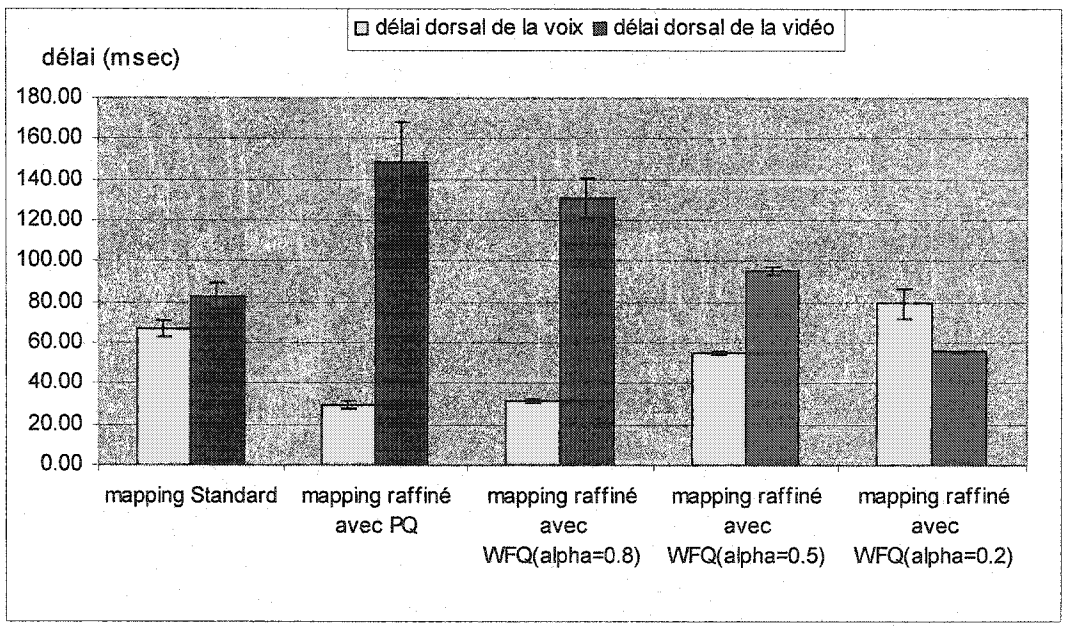


Figure 4.9 Délai dorsal de la voix et de la vidéo téléphonie

Analysons avec plus de détails la Figure 4.9. On remarque que, sous la discipline d'ordonnancement par priorités PQ, notre algorithme de mise en correspondance raffinée abaisse le délai dorsal du trafic de voix de 35 msec avec toutefois une augmentation de celui du trafic de vidéo de 70 msec. WFQ, avec un α=0.8 (un poids pour la voix qui est

4 fois plus important que celui de la vidéo), donne presque les mêmes performances, ce qui peut être expliqué par le fait que WFQ a tendance à émuler PQ en configurant un poids très élevé pour la classe priorisée. En fixant un même poids pour la voix et la vidéo (α=0.5), on remarque que la voix bénéficie d'une diminution de délai (10 msec), alors que la vidéo encourt une augmentation de délai beaucoup plus importante (de 15 msec). Pour WFQ avec un α=0.2 (poids de la vidéo = 4 x poids de la voix), le mécanisme de différenciation voix/vidéo fait diminuer le délai dorsal de la vidéo de 20 msec au coût d'une plus petite augmentation de 15 msec pour la voix. Ce poids peut être utilisé pour améliorer la QoS de la vidéo sans trop dégrader celle de la voix.

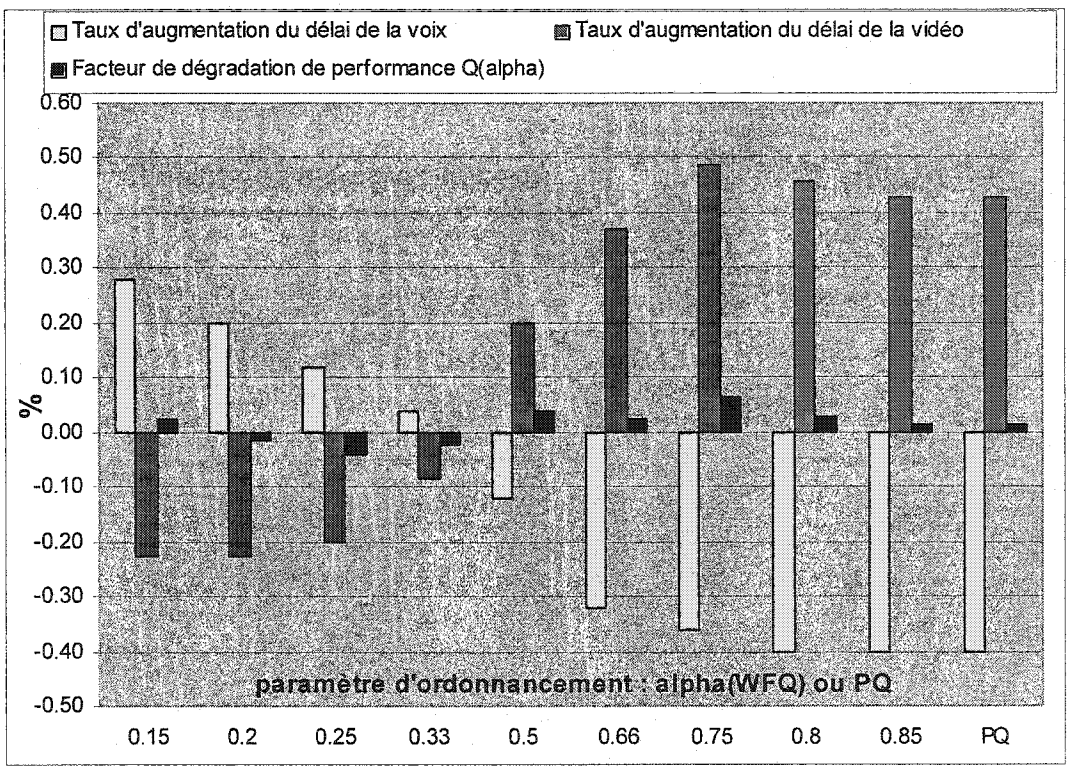
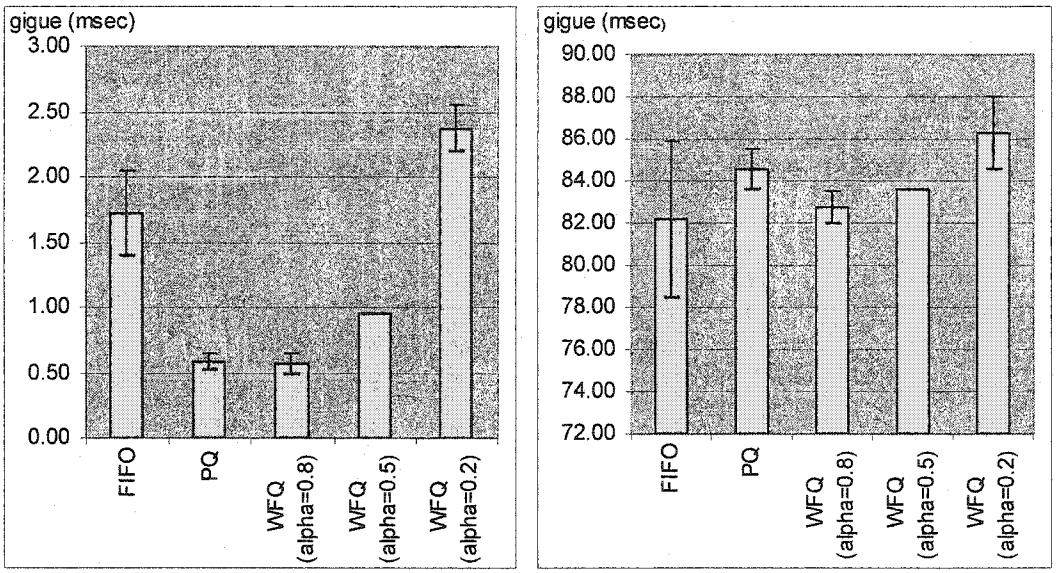


Figure 4.10 Performances attendues en délais en fonction de l'ordonnancement

En s'intéressant aux effets de la différenciation voix/vidéo sur la gigue de la voix (Figure 4.11a), on remarque que la voix subit une nette réduction (de plus que la moitié) de cet indice de performance dans le cas d'un ordonnancement PQ ou un WFQ équivalent avec un α=0.8. Un α=0.5 améliore aussi la gigue de la voix alors qu'un α très

petit (de 0.2) la détériore. Pour la vidéo, la variance des résultats de simulations d'un même scénario (avec différentes valeurs de semences pour le générateur de nombre aléatoires) est assez importante pour rendre impossible la comparaison des résultats des différents scénarios (Figure 4.11b). Toutefois, on peut dire que la gigue du trafic vidéo varie légèrement en adoptant l'approche de différenciation, vu que la variance est moins importante que la valeur absolue de la gigue.



a) Gigue de la voix **b) Gigue de la vidéo**

Figure 4.11 Impact de l'ordonnancement sur la gigue

Comme on l'a vu dans le chapitre précédent, la séparation entre trafic de voix et trafic de vidéo téléphonie se fait en mettant les paquets des deux types de trafic dans deux files d'attente distinctes dans les routeurs dorsaux et en les servant selon une discipline PQ ou WFQ (FIFO étant la discipline qui sert les paquets selon l'ordre du premier-arrivé-premier-transmis, indifféremment de leur type voix ou vidéo, ce qui correspond à l'algorithme du *mapping* standard). Des statistiques sur l'occupation de ces deux files d'attente sous différents paramètres d'ordonnancement (Figure 4.12) montrent que la file contenant les paquets de voix est toujours plus longue que celle des paquets

de vidéo. Ceci est dû non seulement au fait qu'on a 5 fois plus de clients de voix que de vidéo (avec un partage WFQ équitable de poids 0.5 pour la voix et 0.5 pour la vidéo, la file EF est occupée par 10 paquets de voix et la file AF4 par 1 paquet de vidéo en moyenne) mais aussi aux propriétés dissemblables des deux types de trafic. En effet, les petits paquets de voix sont transmis plus rapidement que les gros paquets de vidéo (le temps de transmission d'un paquet sur un lien est inversement proportionnel à sa taille) et s'accumulent ainsi plus vite dans leurs files réservées.

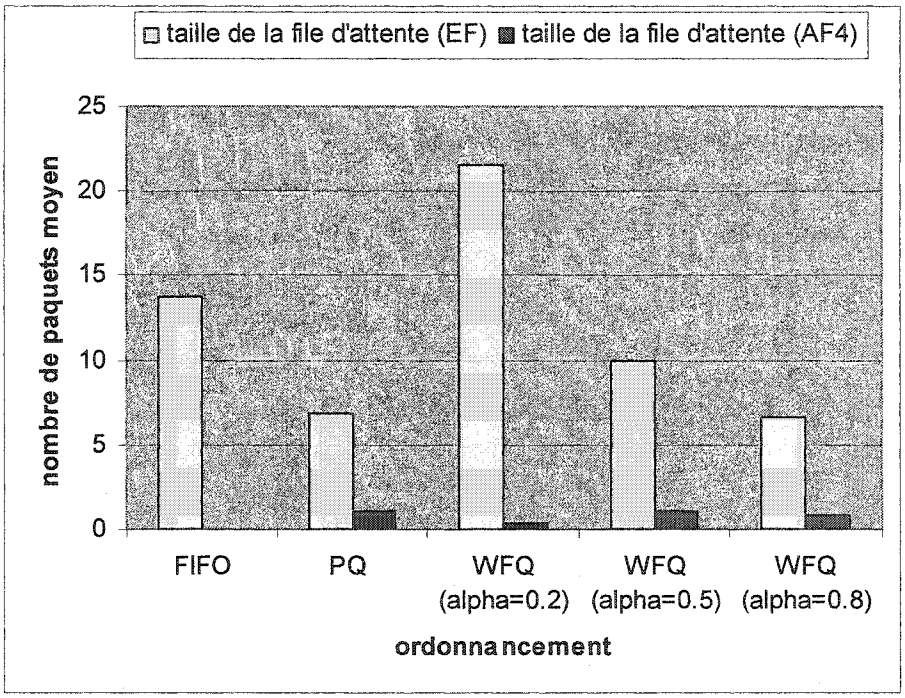


Figure 4.12 Occupation des files d'attente EF et AF4

D'après la même figure, on remarque que la taille moyenne de la file EF dans le cas d'un mapping standard (voix et vidéo dans la même classe EF avec un ordonnancement FIFO) reste toujours supérieure à celle obtenue dans le cas d'un mapping raffiné avec un poids WFQ de 0.5 ou de 0.8 (trafic de voix dans la classe EF et trafic de vidéo dans la classe AF4). Dans ce dernier cas, on peut inclure

l'ordonnancement PQ, qui comme nous l'avons cité auparavant, donne presque les mêmes résultats que WFQ avec un poids élevée (de 0.8 par exemple).

4.6.2 Délai dorsal vs. Délai UMTS

Le délai de bout en bout du trafic conversationnel est souvent décomposé en délai UMTS et en délai dorsal. En effet, le service de téléphonie multimédia de bout en bout est transporté par un réseau d'accès UMTS source puis par un réseau dorsal d'interconnexion et finalement par un autre réseau d'accès UMTS destination. En analysant la contribution de chaque domaine réseau au délai de bout en bout, le réseau dorsal contribue en moyenne avec le quart du délai global alors que les deux réseaux d'accès UMTS contribuent avec les ¾ restants. Ceci étant pour la voix (Figure 4.13). Pour la vidéo (Figure 4.14), on remarque que la contribution du réseau UMTS au délai global est très élevée (une moyenne de 9/10 du délai global) alors que la dorsale contribue par le 1/10 seulement.

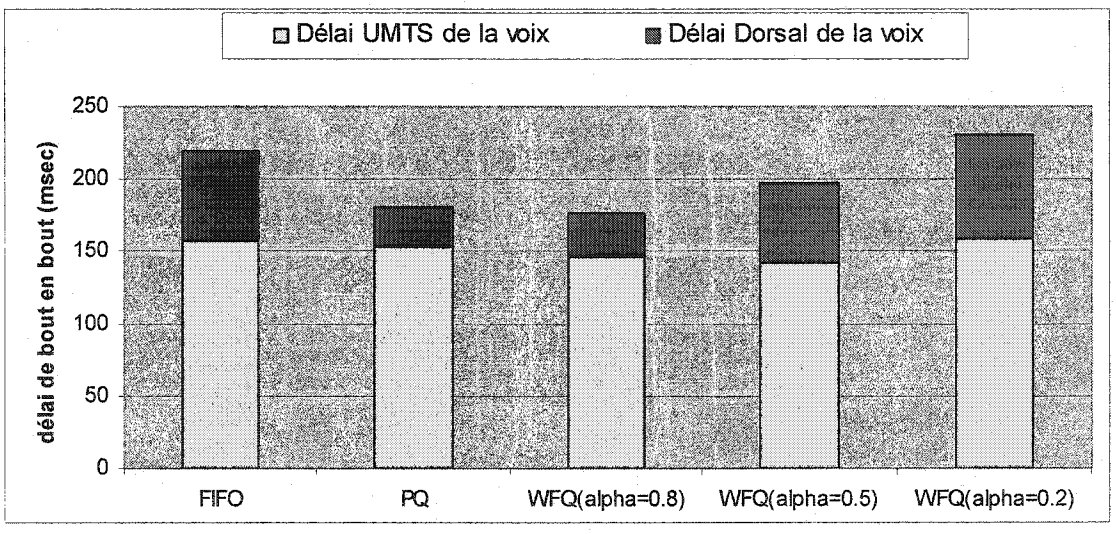


Figure 4.13 Délai Dorsal vs. Délai UMTS du trafic de voix

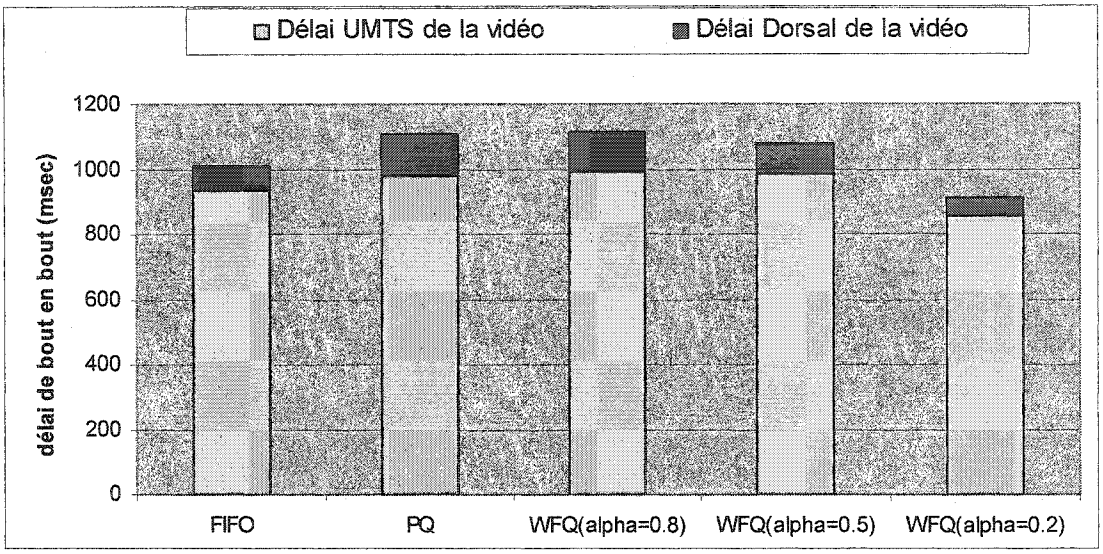


Figure 4.14 Délai Dorsal vs. Délai UMTS du trafic de vidéo téléphonie

De ces résultats, nous déduisons un certain nombre de conclusions intéressantes, la première étant que le réseau d'accès radio terrestre UTRAN constitue dans tous les cas le goulot d'étrangement du système global malgré sa bande passante d'accès assez élevée (jusqu'à 2Mbps). Notons que les liens dorsaux sont largement utilisés avec un facteur d'utilisation moyen de 0.9 qui engendre des congestions occasionnelles. Le délai élevé encouru par le trafic en traversant l'accès UMTS est dû aux phénomènes d'interférences radio et aux latences causées par les mécanismes de contrôle d'admission. Ceci étant pour les petits paquets tels que ceux de la voix. La deuxième conclusion est que les gros paquets de vidéo encourent un délai très élevé par rapport à ceux de la voix en traversant un accès UMTS. Ceci est dû essentiellement à la fragmentation des longs paquets de vidéo en un nombre élevé de cellules ATM en passant par l'UTRAN. En effet, le routeur SGSN (premier nœud à commutation de paquets sur le chemin du trafic ascendant) n'est capable d'acheminer un paquet IP de 1000 octets de vidéo que s'il rassemble dans sa mémoire tampon les différents fragments transmis chacun dans une cellule ATM de 48 octets de charge utile chacune. Cela créée une latence qui est relativement importante par rapport à celle des petits paquets de voix qui n'ont souvent besoin que d'une seule cellule ATM pour être acheminés à travers

l'UTRAN. De plus, les gros paquets de vidéo sont beaucoup plus vulnérables aux erreurs de transmission radio que les petits paquets de voix. En effet :

Probabilité (paquet de taille t subit une erreur de transmission)

$= Probabilité\ (au\ moins\ un\ bit\ erroné) = 1- (1-BER)^t$

le BER étant le taux d'erreur binaire qui est paramétré à 10^{-5} dans le modèle OPNET du lien radio W-CDMA.

4.6.3 Effet de la séparation voix/vidéo sur la synchronisation

La différence entre le temps d'arrivée d'un paquet de vidéo et celui d'un paquet de voix ayant le même numéro de séquence (notons que deux paquets de voix et de vidéo ayant le même numéro de séquence sont transmis par l'application émettrice au même instant) constitue le décalage de synchronisation instantané. À la Figure 4.15, nous avons représenté le décalage de synchronisation moyen pour différents schémas d'ordonnancement et différents facteurs d'utilisation des liens. On remarque que la mise en correspondance standard (i.e. avec l'ordonnancement de la voix et de la vidéo dans une même file sous la discipline FIFO) donne des décalages de synchronisation qui sont plus élevées que ceux fournis par une séparation voix/vidéo avec un partage équitable pondérée avec 50% de poids pour la voix et 50% de poids pour la vidéo.

Pour des facteurs d'utilisation de liens assez faibles (0.3), ceci n'est plus valide. Pour un facteur d'utilisation de 0.9, le décalage entre flux de voix et flux de vidéo sous un mapping standard est de 35 msec, avec la vidéo qui devance la voix (valeurs instantanées des arrivées des paquets de vidéo soustraites de ceux de la voix sont négatives). Sous un mapping raffiné (voix et vidéo différentiées par WFQ), ce même décalage devient plus faible (25 msec). Ceci est causé essentiellement par la diminution du retard induit par la mise en file des paquets de voix attendant la transmission des gros paquets de vidéo, vu que l'ordonnancement régule un service équitable pondéré entre les deux types de trafic. Dans notre cas, un poids de 0.5 pour chacun des deux types de trafic et avec des paquets de vidéo 10 fois plus gros que les paquets de voix,

l'ordonnancement WFQ transmet à la file de sortie (de transmission) alternativement 10 paquets de voix puis un paquet de vidéo, et ainsi de suite tant qu'il y a des paquets en attente. Ceci a pour effet d'accélérer la transmission des paquets de voix et de minimiser ainsi leur retard par rapport aux paquets de vidéo. Une étude peut être faite pour déterminer les valeurs des poids WFQ qui permettent d'annuler complètement ce décalage pour avoir une synchronisation parfaite entre voix et vidéo. À première vue, un poids de 10/11 pour la voix permet d'avoir un ordonnancement qui transmet alternativement 1 paquet de chaque classe (voix et vidéo), ce qui diminue considérablement le décalage entre les deux types de flux. Bien sûr, d'autres paramètres peuvent entrer en jeu tels que la variation de la taille des paquets de vidéo ou encore le nombre de flux multiplexés.

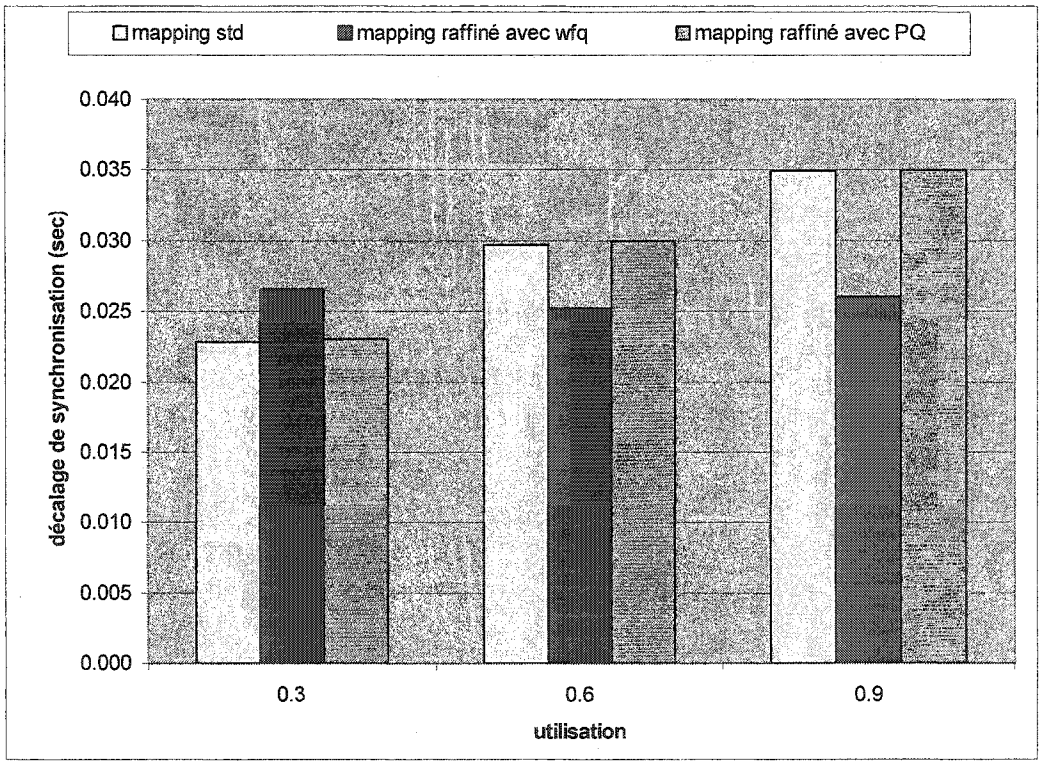


Figure 4.15 Décalage de synchronisation entre voix et vidéo

4.7 Validation du modèle de trafic de vidéo-téléphonie implanté

La Figure 4.16 montre bien que le trafic généré par notre modèle générique MPEG a une allure et un comportement qui s'approche beaucoup de la réalité représentée par les valeurs réelles d'une trace d'une session de vidéo conférence [Trace-MPEG4]. En effet, on remarque bien la même régularité dans les hauts pics dans les deux courbes (trace et modèle) qui représentent les trames *I* (intra codées) qui ont une taille très importante et qui ont une occurrence à chaque début de groupe d'images GOP. On remarque aussi les petites fluctuations entre deux pics qui représentent les faibles tailles des trames *P* et *B*.

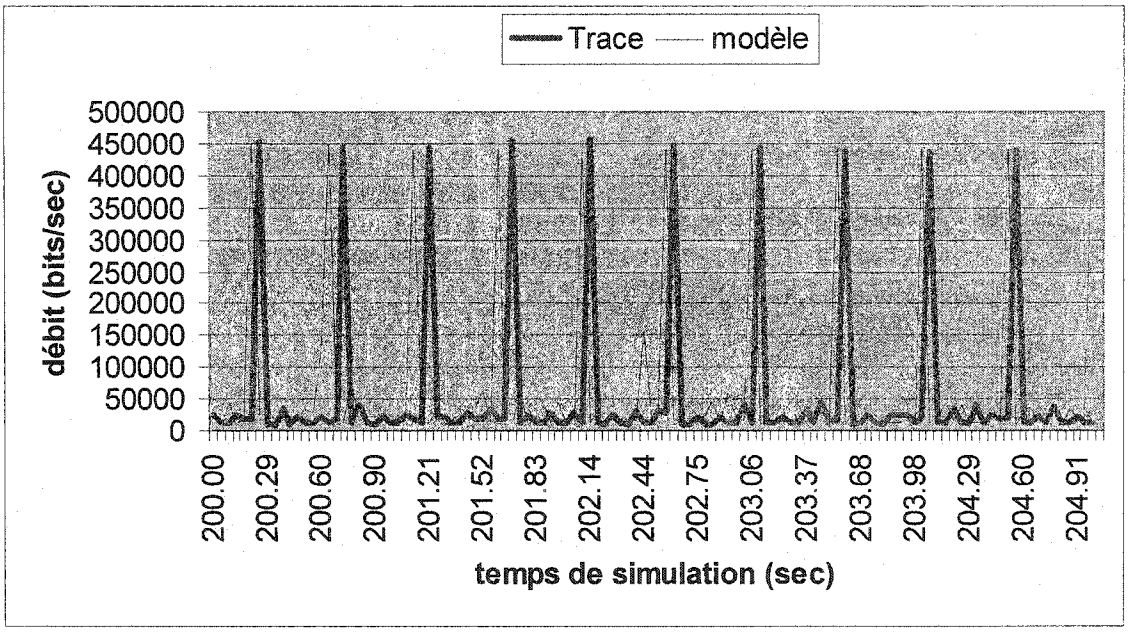


Figure 4.16 Validation du modèle implanté avec une trace

CHAPITRE 5
CONCLUSION

L'étude de performance des réseaux mobiles de troisième génération à transporter efficacement les nouveaux services de téléphonie multimédia fait partie des travaux les plus récents. De plus, vu que ces types de réseaux sont dans leur première phase de déploiement, la simulation reste le moyen le plus approprié pour évaluer leur performance. Ce mémoire s'est concentré particulièrement sur l'étude de l'impact du mécanisme assurant la continuité de la QoS tout au long d'un réseau dorsal IP sur les performances d'acheminement du trafic conversationnel. L'optimisation de ce mécanisme de mise en correspondance de QoS pour la téléphonie multimédia a été faite en cherchant analytiquement et par simulations les paramètres des disciplines d'ordonnancement faisant améliorer les performances. Nous avons montré que du à leur caractéristiques divergentes, il est préférable de différencier entre trafic de voix et trafic de vidéo téléphonie. Pour cela, il apparaît qu'il faut paramétrer un ordonnancement par partage équitable pondéré avec les bonnes valeurs de poids qui permettrait de mieux contrôler la QoS donnée à chacun de ces types de trafic sans affecter une synchronisation éventuelle entre les deux.

5.1 Synthèse des travaux et principales contributions

Ce mémoire repose sur l'idée principale de séparation du trafic de vidéo téléphonie du trafic de voix dans les files d'ordonnancements des routeurs dorsaux afin d'améliorer l'efficacité de l'acheminement du trafic conversationnel. En partant d'une analyse stochastique approfondie de ce type de trafic, nous avons constaté des divergences remarquables entre les caractéristiques des flux de voix et ceux de la vidéo téléphonie. En plus de l'idée de différenciation voix/vidéo, ces constatations nous ont

permis de proposer un modèle de simulation générique pour supporter l'application de vidéo téléphonie au niveau des stations UMTS.

Plus particulièrement, dans ce mémoire nous nous sommes intéressés à comparer les performances des algorithmes les plus utilisés pour ordonnancer le trafic conversationnel qui est généralement considéré comme le trafic le plus critique pour les opérateurs de téléphonie. De cette comparaison, nous avons retenu que l'algorithme de partage équitable pondéré WFQ est le meilleur ordonnancement pour assurer la QoS de la téléphonie multimédia en termes de délais et de gigues. En outre, pour des hypothèses réalistes sur le trafic UMTS, un paramètre de poids WFQ pour la voix qui est de l'ordre du tiers de celui de la vidéo donne les meilleures performances globales. En d'autres termes, si l'opérateur veut améliorer le délai du trafic de vidéo téléphonie sans trop dégrader celui de la voix, il doit pondérer le service de la voix dans l'ordonnancement WFQ par un poids optimal qui est égal à $\alpha \approx 0.25$. Cette valeur a été établie par la résolution d'un problème d'optimisation au chapitre 3 et a été vérifiée par simulations au chapitre 4. D'un autre coté les simulations que nous avons implémentées nous ont servi à déduire l'impact des différents choix des paramètres de l'algorithme d'ordonnancement sur les délais, les gigues et la synchronisation entre flux de voix et de vidéo d'une même session de téléphonie multimédia. Les résultats trouvés ont permis de montrer que l'algorithme WFQ paramétré avec le poids optimal ne dégrade ni la gigue ni la synchronisation voix/vidéo.

De plus, les simulations nous ont permis de déceler le goulot d'étranglement dans un réseau UMTS étendu qui s'avère être le réseau d'accès radio terrestre UTRAN malgré sa bande passante abondante pouvant aller jusqu'à 2 Mbps et un réseau à commutation de paquets utilisé à 90%. Nous avons déduit aussi que le réseau d'accès UTRAN affecte beaucoup plus le trafic de vidéo téléphonie que celui de la voix. Sa contribution au délai de bout en bout du trafic de vidéo téléphonie est 3 fois plus importante que celle de la voix.

En outre, dans ce mémoire, nous avons contribué à la construction d'un modèle de simulation générique et modulaire pour l'application de vidéo conférence. La

généricité vient du fait du nombre des paramètres permettant de configurer plusieurs codecs de vidéo se basant sur MPEG2 ou MPEG4. La modularité vient du fait de pouvoir installer le modèle non seulement sur une station UMTS mais aussi sur n'importe quelle station munie de la pile protocolaire TCP/IP en permettant une bonne utilisation dans l'environnement *OPNET modeler*. La validation de ce modèle a montré qu'une configuration adéquate des paramètres donne un trafic avec un profil qui s'approche de très près de celui d'un trafic réel.

5.2 Limitations des travaux

Dans les spécifications techniques de 3GPP, la QoS UMTS a été bien définie dans le domaine d'accès UMTS. Cependant dans la mise en correspondance de cette QoS avec celle d'une dorsale IP d'interconnexion, il n'a été spécifié que la correspondance entre classes UMTS et classes DiffServ alors que les mécanismes d'ordonnancement qui doivent être utilisés sur le réseau dorsal n'ont pas été définis. De ce fait, nos travaux se sont focalisés sur l'étude du réseau dorsal dans son support d'une QoS de bout en bout pour les services UMTS. Notre intérêt qui s'est porté à la dorsale IP a permis d'écarter les aspects de mobilité des usagers qui se fait en général au niveau du réseau radio. Ainsi dans notre cas, nous avons fixé des localisations uniformes d'un nombre assez élevé de stations UMTS de telle sorte que nous avons eu une charge assez importante, régulière et prévisible.

En outre, vu que les réseaux UMTS sont supposés supporter un nombre important de services de toutes sortes, nous nous somme intéressé à l'étude des services les plus critiques en matière de QoS et les plus importants pour un opérateur de téléphonie 3G. Toutes ces suppositions ont permis de réduire considérablement les temps d'exécution des simulations qui sont extrêmement élevés pour des réseaux complexes véhiculant un trafic important de voix et de vidéo.

5.3 Recommandations pour des travaux futurs

Notre travail a montré que la séparation entre trafic de voix et de vidéo téléphonie dans les files d'ordonnancement des routeurs dorsaux avec une discipline de partage équitable pondérée avec les poids adéquats permet d'améliorer la QoS des services de téléphonie multimédia d'un réseau UMTS. Toutefois, nous avons remarqué que dans certains cas malgré la diminution du délai de bout en bout avec notre approche de différenciation voix/vidéo, la contrainte sur le délai maximal n'est pas vérifiée. En effet, comme nous l'avons vu, l'utilisation du mécanisme de différenciation de service DiffServ dans la dorsale IP ne permet pas d'assurer les délais absolus mais plutôt une bonne évolutivité en supportant un nombre très important de sessions UMTS sans dégradation de performance. Toutefois, avec l'apparition des mécanismes avancés d'ingénierie de trafic utilisant RSVP-TE, il est possible de considérer des réservations dynamiques à la manière de IntServ/RSVP pour des chemins à commutation d'étiquettes LSP véhiculant un agrégat de flots plutôt que pour des micro flots séparés et permettant en conséquence d'améliorer l'évolutivité de la dorsale tout en garantissant une QoS minimale.

Ainsi, une direction de recherche future serait de concevoir un mécanisme de contrôle d'admission à la frontière de la dorsale permettant la mise à jour d'une réservation dynamique (en fonction de la demande de trafic entrant, de ses paramètres et contraintes de QoS) d'un ensemble de LSP chacun transportant un type de service UMTS particulier entre routeur frontière origine et routeur frontière destination. Cela peut être réalisé en faisant inter opérer la signalisation de la QoS UMTS du contexte PDP avec les extensions de RSVP-TE permettant par exemple de mettre à jour la bande passante disponible pour un LSP de sorte à assurer la contrainte du délai maximal.

BIBLIOGRAPHIE

Articles de revues et de conférences, RFCs

[Ben Ali et al., 2003] R. Ben Ali, Y. Lemieux, S. Pierre, "UMTS to IP Backbone QoS Mapping Refinement for Multimedia Telephony Services", Accepted to appear in Proceedings of OPNETWORK 2003 with Distinguished Paper Award, August 25-29, 2003, Washington D.C, USA.

[Blake et al., 1998] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss, "RFC2475: An Architecture for Differentiated Services", December 1998.

[Brady, 1969] Brady P.T, "A model for Generating ON_OFF Speech Patterns in Two-Way Conversations", Bell System Technology Journal, vol.48, September 1969, pp. 2445-2472.

[Braden et al., 1994] R. Braden, D. Clark, S. Shenker, "RFC1633: Integrated Services in the Internet Architecture: an Overview", June 1994.

[Chaskar et al., 2001] H. Chaskar and R. Koodli, "MPLS and DiffServ for UMTS QoS in GPRS Core Network Architecture", ISOC Conference Proceedings, Stockholm, Sweden, June 2001 "www.isoc.org/isoc/conferences/inet/01/CD_proceedings/T56/MPLSDiff.htm"

[Cruz, 1991] R.L.Cruz, "A calculus for Network Delay, Part I : Network Elements in Isolation", IEEE Transactions on Information Theory, vol. 37, pp. 114-131, January 1991.

[Durham et al., 2000] D. Durham, Ed., J. Boyle, R. Cohen, S. Herzog, R. Rajan, A. Sastry, "RFC2748: The COPS (Common Open Policy Service) Protocol", January 2000.

[Elsayed, 2000] K. Elsayed, "Comparative Performance Analysis of Call Admission Control Schemes in ATM Networks", Journal of Computer Networks and ISDN Systems pp. 56-124, 2000.

[Elwalid, 1995] A. Elwalid, "Fundamental Bounds and Approximations for ATM Multiplexers with Applications to Video Teleconferencing", IEEE Journal on Selected Areas in Communications, 1995, Vol. 13, pp. 1004-1016.

[Firoiu, 1997]V. Firoiu, "Efficient admission control for EDF Schedulers", IEEE Infocom'97, 7-11 April 1997, Vol. 1, pp. 310-317.

[Floyd et al., 1993] S. Floyd, V. Jacobson, "Random Early Detection Gateways for Congestion Avoidance", IEEE/ACM Transactions on Networking, August 1993, Vol. 1, N°4, pp. 397-413.

[Floyd et al., 1995] S. Floyd and V. Jacobson, "Link-Sharing and Resource Management Models for Packet Networks", IEEE/ACM Transactions on Networking, August 1995, Vol. 3 N°4 , pp. 365 –386.

[Grossglauser et al., 1996] M. Grossglauser and S. Keshav, "On CBR service", IEEE INFOCOM, March 1996,Vol. 1 , pp.: 129 –137.

[Hagiwara et al., 1999] T. Hagiwara, H. Doi, H. Tode, and H. Ikeda "High-Speed Calculation Method of the Hurst Parameter Based on Real Traffic", Proceedings 25th Annual IEEE Conference on , November 2000, pp.: 662 –669.

[Handley et al., 1999] M. Handley, H. Schulzrinne, E. Schooler and J. Rosenberg "RFC 2543: SIP: Session Initiation Protocol", March 1999.

[Heinanen et al., 1999] J. Heinanen, F. Baker, W. Weiss, J. Wroclawski, "RFC2597 : Assured Forwarding PHB Group ", June 1999.

[Holier et al., 1999] Holier and al., "Multi-Modal Perception", BTTJ, January 1999, Vol.17.

[Hughes et al., 1995] H. Hughes and M. Krunz, "A Traffic Model for MPEG-Coded VBR Streams", ACM Sigmetrics, 1995, Vol. 23 N°1, pp. 47-55.

[Iera et al., 2002] A. Iera, A. Molinaro, "Designing the Interworking of Terrestrial and Satellite IP-Based Networks", IEEE Communications Magazine, February 2002, Vol. 40 N°2, pp. 136-144.

[Jacobson et al., 1999] V. Jacobson, K. Nichols and K. Podurial., "RFC2598: An Expedited Forwarding PHB", June 1999.

[Jamaloddin, 1996] S. Jamaloddin, "Fair Queueing Algorithms for Packet Scheduling in BISDN", International Zurich Seminar on Digital Comunications, 1996, pp. 39-51.

[Keshav, 1997] S. Keshav, "An engineering approach to computer networking: ATM Networks, the Internet, and the Telephone Network", Professional Computing Series, 1997, pp. 253-258.

[Krunz et al., 2001] M. Krunz and A. Makowski, "Modeling Video Traffic Using M/G/ Input Processes: A Compromise between Markovian and LRD Models", IEEE JSAC, Special Issue on Computational Aspects of Teletraffic Models, 2001, Vol. 16 N°5, pp. 733-748.

[Le Faucheur et al., 2002] F. Le Faucheur, L. Wu, B. Davie, S. Davari, P. Vaananen, R. Krishnan, P. Cheval, J. Heinanen, "RC3270 : Multi-Protocol Label Switching (MPLS) Support of Differentiated Service", May 2002.

[Liebherr et al., 1996] J. Liebherr, D. Werge et D. Ferrari, "Exact admission control for networks with a bounded delay service", IEEE Trans. on Networking, 1996, vol.4, pp. 885-901.

[Maniatis et al., 2002] I. Maniatis, G. Nikolouzou and S. Venieris., "QoS Issues in the Converged 3G Wireless and Wired Networks", IEEE Communications magazine, August 2002, pp. 44-53.

[Mankin et al., 1997] A. Mankin, Ed., F. Baker, B. Braden, S. Bradner, M. O`Dell, A. Romanow, A. Weinrib, L. Zhang, "RFC2208: Resource ReSerVation Protocol (RSVP) -- Version 1 Applicability Statement Some Guidelines on Deployment", September 1997.

[Parekh et Gallager, 1994] A.K. Parekh and R.G Gallager, "A Generalized Processor Sharing Approach to Flow Control in Integrated Services Networks: The

multiple node case", IEEE transactions On Networking, 1994, vol. 2, pp. 137-150.

[Sen et al., 1989] P.Sen, Endre I. Sára and Wei Sun., "Models for packet switching of variable-bit-rate video sources", IEEE Journal on Selected Areas in Communications, June 1989, Vol.7, pp. 865-869.

[Schwartz, 1996] M. Schwartz, "Broadband Integrated Networks", 1996, chapter 1, page 3.

[Shenker et al., 1997] S. Shenker, C. Partridge, R. Guerin, "RFC2212: Specification of Guaranteed Quality of Service", September 1997.

[Reichmeyer et al., 1998] F. Reichmeyer, L. Ong, A. Terzis, L. Zhang, and R. Yavatkar, "A Two-Tier Resource Management Model for Differentiated Services Networks", IETF draft-rotzytwo-tier-01.txt, November 1998.

[Rose, 1996] O. Rose, "Estimation of the Hurst Parameter of Long-Range Dependent Time Series", Report N°137, February 1996..

[Rosen et al., 2001] E. Rosen, A. Viswanathan, R. Callon, "RFC3031: Multiprotocol Label Switching Architecture", January 2001.

[Throeung, 2001] O. Throeung, "Un algorithme de contrôle d'admission pour la garantie de qualité de service dans le réseau de cœur des systèmes de troisième génération", Mémoire de maîtrise ès sciences, Ecole Polytechnique de Montréal, 2001.

[Tobagi et al., 2001] F.Tobagi, W.Noureddine, M.Karam, A.Markopoulou, C.Fraleigh, J.M.Pulido, J.I.Kimura et B.Chen, "Service Differentiation in the Internet to support Multimedia Traffic", IWDC '01 Proceedings , Taormina, Italy, September 17-20, 2001, pp. 279-294.

[Wroclawski, 1997a] J. Wroclawski, "RFC2210: The Use of RSVP with IETF Integrated Services", September 1997.

[Wroclawski, 1997b] J. Wroclawski, "RFC2211: Specification of the Controlled-Load Network Element Service", September 1997.

[Yavatkar et al., 2000] R. Yavatkar, D. Pendarakis, R. Guerin, "RFC2753: A Framework for Policy-Based Admission Control". January 2000.

[Yegengolu et al., 1993] F. Yegenoglu, B. Jabbari and Ya-Qin Zhang, "Motion-classified autoregressive modelling of variable bit rate video", IEEE Trans. on Circuits and Systems for Video Technology, Feb. 1993, vol. 3, pp. 42-53.

[Zhang et al., 1993] H. Zhang, D. Ferrari, "Rate-Controlled Static-Priority queueing", Proceedings INFOCOM'93, April 1993, pp. 227-236.

Rapports techniques

[3GPP, 2002a] 3GPP, "TS29.207V5: Policy Control over Go Interface (Release 5)", June 2002.

[3GPP, 2002b] 3GPP, "TS23.107V5: Universal Mobile Telecommunication System (UMTS); QoS Concepts and Architecture", June 2002.

[3GPP, 2002c] 3GPP, "TS23.207V5: End-to-End QoS Concept and Architecture", June 2002.

[Cisco, 2001] Cisco White Paper, "How to Add MSN Messenger Services for PC-to-Phone Functionality to Cisco Packet Voice Networks", 2001.

[ITU-T, 1998] ITU-T, "Recommendation H.263: Video coding for low bit rate communication", February1998.

[ITU-T, 2000] ITU-T, "Recommendation G.114: One way transmission", May 2000.

[ITU-T, 2001] ITU-T, "Recommandation G.113: transmission impairments due to speech processing", February 2001.

[UMTS-Forum, 1999a] UMTS-Forum, "Report N6 : UMTS/IMT-2000 Spectrum", June 1999.

Outils logiciels et pages Web de référence

[Cisco, 1999] "Online Manual: Cisco IOS Quality of Service Solutions Configuration Guide", 1999, http://www.cisco.com/univercd/cc/td/doc/product/software/ios120/12cgcr/qos_c/

[Ethereal] G. Combs, "The Ethereal Network Analyzer Software Tool Version 0.9.11", www.ethereal.com

[Hest] Ilias Stergiou, The Hurst ESTimator (hest), http://www.dcs.shef.ac.uk/~ilias/software.html

[OPNET] http://www.opnet.com

[Trace-MPEG4]http://www-tkn.ee.tu-berlin.de/research/trace/pics/FrameTrace/mp4/Verbose_MobilkomBoche_10_14_18.dat

ANNEXE I

Seau à jetons et seau percé

Le seau à jetons et le seau percé sont des mécanismes utilisés pour réguler ou caractériser des sources de trafic. La première évaluation des propriétés du filtre en "seau percé" fut donnée par Cruz [20] pour une caractérisation simple des sources dans le cadre des réseaux à intégration de services. Toute connexion active dans un réseau à intégration de services dispose d'un filtre en "seau percé" assurant la régulation en entrée du réseau du trafic transmis par sa source. Ce mécanisme de régulation du trafic met en oeuvre un accumulateur de jetons appelé "seau" de capacité maximale b et de taux de remplissage constant ρ. Les paquets sont admis dans le réseau uniquement après avoir retiré le nombre de jetons requis du "seau".

Souvent la régulation de trafic se fait dans le but de le mettre en forme (shaping) de telle sorte qu'il soit conforme au contrat de service. Il y a deux mécanismes différents pour cette mise en forme l'un basé sur un seau à jetons (Figure 5.1) et l'autre sur un seau percé (Figure 5.2). Les deux partagent des caractéristiques communes telles que le fait d'utiliser des jetons dans un seau pour réguler un flux de trafic entrant à travers une file. La différence maîtresse réside dans le fait que les jetons sont accumulés à un taux constant s'il s'agit d'un seau à jetons, alors que dans le seau à jetons, les jetons sont retirés à un taux constant. Les seaux percés sont généralement dédiés aux cellules ATM, alors que les seaux à jetons sont dédiés aux paquets. Comme les cellules ont des tailles fixes, leur transmission peut se faire sans avoir à attendre toutes les cellules dans le PDU. D'un autre coté, les paquets sont transmis lorsqu'il y a suffisamment de jetons disponibles dans le seau pour tous les segments de paquets et en attendant le nombre de jetons requis pour tous les segments, le paquet est retenu dans une file. Dans ce sens les seaux à jetons sont des mécanismes basés sur le crédit de jetons, alors que les seaux percés sont basés sur le déficit de jetons

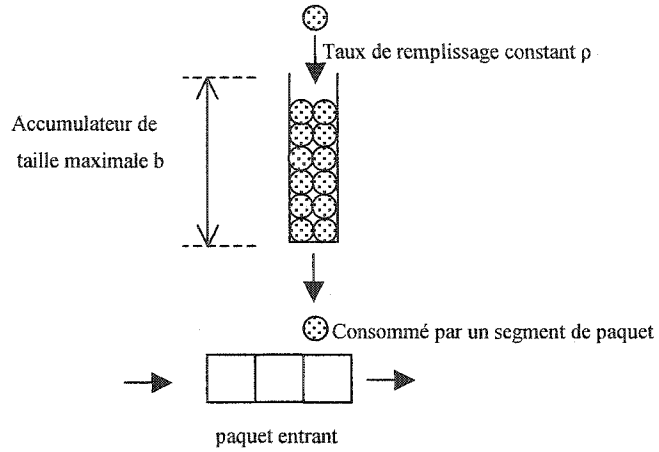


Figure I.1 Filtre en seau à jetons pour les paquets

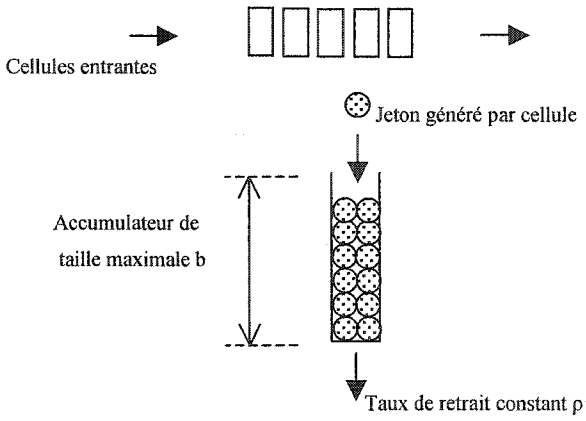


Figure I.2 Filtre en seau percé pour les cellules

Un régulateur en seau percé donne un flux fluide et lisse avec un taux constant. Tandis qu'un régulateur à seau à jetons limite le taux du flux sortant à un certain taux moyen, mais permet un trafic en rafale de se produire tant que la taille de la rafale la plus longue n'a pas dépassé la profondeur du seau à jetons. D'un autre coté, comme les cellules sont de tailles égales, le taux moyen est égale au taux maximal et la taille de la rafale est constante.

ANNEXE II

Le paramètre de Hurst et l'autosimilarité

Considérée comme l'avancement le plus important dans la modélisation et l'évaluation de performance durant les cinq dernières années, la notion d'autosimilarité permet de modéliser des processus qui sont très variables sur toutes les échelles temporelles. Étant donnée une série temporelle $\{X_i\}$, i=1..n, on la regarde sous différents niveaux d'agrégations : X(m) est une nouvelle série temporelle qui est définie en sommant m élément de la série originale en un élément de la nouvelle. Par exemple, $X(3) = \{ X_1+X_2+X_3, X_4+X_5+X_6, \ldots \}$. La fonction d'auto corrélation r(k) d'une série temporelle donne une indication sur le degré de similarité de la série par rapport a elle même translaté de *k* points temporels. Une série qui est auto similaire est celle qui est « similaire a elle même » sur différentes échelles temporelles. Généralement, une série autosimilaire se caractérise par un affaiblissement très lent de sa variance (suivant une loi hyperbolique plutôt qu'exponentielle comme c'est le cas des autres distributions). Ceci veut dire aussi que la série auto similaire originale a une variance infinie. La fonction d'auto corrélation r(k) diminue très lentement ce qui se manifeste par ce qu'on appelle des dépendances sur de longs intervalles (Long Range Dependencies ou LRD).

$$r(k) \approx k\text{-}\beta L(t) \quad \text{for some } 0 < \beta < 2 \text{ as } k \to \infty$$

Ou L(t) est une fonction qui change lentement (constante asymptotiquement). Ce qui implique que chaque X_i affecte les futures X_i sur un long période de temps. Les dépendances sur de longs intervalles LRD caractérisent des distributions de séries temporelles qui sont a queues lourdes : les valeurs qui sont très loin de la moyenne apparaissent souvent. Dans l'équation précédente on a introduit le paramètre β. Ce β mesure le degré de similarité dans la série temporelle originale. Toutefois, pour des raisons historiques, on n'utilise pas le paramètre β mais plutôt le paramètre de Hurst, qui est défini comme suit :

$$H = 1 - \frac{\beta}{2}$$

Si H=½ alors le processus n'est pas auto-similaire, et quand ½ < H < 1 il est auto-similaire avec une dérivée positive (la majorité des données auto-similaires sont positives). Hurst, le chercheur qui a donné son nom à ce paramètre, a travaillé sur beaucoup de séries temporelles observées dans la nature et il s'est intéressé à vérifier l'auto similarité en calculant des statistiques R/S sur des intervalles échelonnés et ajustées. Pour une série de temps X ayant une moyenne et une variance échantillonnée $S_2(n)$, cette statistique est donnée par :

$$R(n) / S(n) = [\ 1 - S(n)\ ] \times [\ \max(0, W_1, W_2, \ldots, W_k) - \min(0, W_1, W_2, \ldots, W_k)\ ]$$

Ou: $W_k = (X_1 + X_2 + \ldots + X_k) - k\,\overline{X}(n) \qquad (k \geq 1)$

L'observation de dépendances sur des courts intervalles s'avèrent satisfaire la relation :

$$E[\ R(n) / S(n)\ ] \approx c_0 n^{0.5}$$

Toutefois les données avec des dépendances sur de longs intervalles, tels que les processus auto-similaires, satisfont la relation :

$$E\ [\ R(n) / S(n)\ ] \approx c_0 n^{H} \quad (0 < H < 1)$$

Ceci est connu sous le nom de l'effet de Hurst et peut être utilisé pour distinguer un processus auto-similaire d'un autre qui ne l'est pas. Pour calculer le paramètre de Hurst H, on dresse souvent la courbe suivante sur une échelle logarithmique sur les deux axes et on calcule la pente de la droite qui sera une estimation de H:

$$\log [\ R(n) / S(n)\ ] = c_2 + H \log(n)$$