Allocation de ressources à prise de décisions distribuées dans des réseaux mobiles hétérogènes

Christophe Gaie

N° D'ORDRE

THÈSE DE DOCTORAT

SPECIALITE : PHYSIQUE

Ecole Doctorale « Sciences et Technologies de l'Information des Télécommunications et des Systèmes »

Présentée par :

Christophe Gaie

Sujet :

Allocation de Ressources à Prise de Décisions Distribuées dans des Réseaux Mobiles Hétérogènes

Soutenue le 7 avril 2010 devant les membres du jury :

M. ASSAAD Mohamad	SUPELEC (encadrant)
M. DUHAMEL Pierre	LSS Paris-Sud 11 (directeur de thèse)
M. MUCK Markus	Infineon Technologies (examinateur)
M. TERRE Michel	CNAM Paris (rapporteur)
Mme VEQUE Veronique	IEF Paris-Sud 11 (examinateur)
M. ZEGHLACHE Djamal	INT-TELECOM SudParis (rapporteur)

Contents

Remerciements

Je remercie particulièrement Mohamad Assaad et Pierre Duhamel qui ont encadré ma thèse à Supélec. Leur expertise du domaine des Télécommunications, leurs conseils et directives m'ont permis de réaliser des travaux de recherche enrichissants. Je remercie également Hikmet Sari de m'avoir permis d'effectuer ma thèse dans un cadre propice à mes travaux de recherche au sein du Département Télécommunications de Supélec. Je remercie l'école doctorale STITS et ses membres pour la confiance qui m'a été accordée au cours de ces trois années de recherche.

J'exprime ma reconnaissance à mes collègues de Supélec qui m'ont accueilli chaleureusement et ont montré un grand intérêt pour mes travaux de recherche : Merouane Debbah, Jocelyn Fiorina, Pablo Piantanida, Mari Kobayashi, Sheng Yang, Thierry Letertre, Raul Lacerda.

Je témoigne également de ma gratitude envers les personnels techniques de Supélec qui m'ont fourni les moyens nécessaires à la réalisation de cette thèse : Huu Hung Wong, José Fonseca, Catherine Magnet.

J'adresse également mes remerciements à mes collègues de Motorola, avec qui j'ai interagi de manière constructive et dans un cadre industriel dynamique. J'ai notamment pu bénéficier de l'aide technique de mes encadrants industriels : Markus Muck, David Grandblaise, Didier Bourse, Guillaume Vivier, Jean Noel Patillon et Nicolas Demassieux.

De nombreux collègues ont également facilité mes travaux de recherche au sein de l'équipe radio : Stephanie Rouquette-Leveil, Mohamed Kamoun, Laurent Mazet, Patrick Labbé, Sebastien Simoens, Anaid Robert, Sandro Sital, Salim Belgroune, Rémy Pintenet, Vincent Merat, Roberta Fracchia, Soodesh Buljore, David Bateman, Marco Fratti, Marc de Courville, Karine Gosse. J'ai également pu bénéficier du soutien technique et informatique notamment de la part de Fabrice Barbarin, Vivien Venerosy et Olivier Lahaye. En outre, je remercie Nathalie Croci, Bahia Mokedem, Sylvain Gaillard, Julien Richard, Johanne Maleville et Béatrice Lemoinne pour m'avoir offert les meilleures conditions de travail.

Je témoigne de la plus grande gratitude envers mes amis doctorants qui n'ont cessé de m'aider et me soutenir : Naveed Ul Hassan, Ayaz Ahmad, Lina Mroueh, Gaonig He, Leonardo Cardoso, Najett Neji, ...

Je tiens également à faire part de ma plus sincère reconnaissance à tous les enseignants qui ont participé à ma scolarité, tant pendant ma thèse qu'auparavant. De plus, j'adresse mes remerciements à tous ceux que j'aurais pu oublier mais qui ont joué un rôle, même minime dans mes travaux.

Enfin, je remercie du fond du coeur tous ceux que j'aime et sans qui rien de tout cela n'aurait été possible.

Abréviations

Abréviations Françaises

ARR : Allocation de Ressources Radios
GCRR : Gestion Commune des Ressources Radios
MCA : Modulation et Codage Adaptatifs
QdS : Qualité de Service
RSIB : Rapport Signal à Interférence plus Bruit
TAR : Technologie d'Accès Radio

Abréviations Anglo-Saxonnes

AMC : Adaptive Modulation and Coding
BSC : Base Station Controller
CDMA : Code Division Multiple Access
CRRM : Common Radio Resource Management
DSA : Dynamic Spectrum Access
EPF : Extended Proportional Fair Resource Allocation
FDMA : Frequency Division Multiple Access
GSM : Global System for Mobile Communications
HSPA : High Speed Packet Access
LTE : Long Term Evolution
MLIP : Mixed Linear Integer Programming
OFDMA : Orthogonal Frequency Division Multiple Access
RAT : Radio Access Technology
RNC : Radio Network Controller
SDR : Software Defined Radio
SINR : Signal to Interference plus Noise Ratio
TCP : Transport Control Protocol
TDMA : Code Division Multiple Access
UMTS : Universal Mobile Telecommunications System
WCDMA : Wideband Code Division Multiple Access
WLAN : Wireless Local Area Network
WMAN : Wireless Metropolitan Area Network
WV : Wireless Vegas Resource Allocation
WWAN : Wireless Wide Area Network

Résumé Général

Récemment, le domaine des Télécommunications a bénéficié du développement et de l'essor de nombreuses technologies (2G avec le TDMA/FDMA, 3G avec le CDMA, 3G+ avec le TDMA/CDMA, 4G avec l'OFDMA, ...). Ces technologies se sont concrétisées par le déploiement de réseaux de communication dont l'utilisation n'est pas pleinement efficace. En effet, de nombreux réseaux n'utilisent pas la totalité de leurs ressources tandis que d'autres réseaux sont surchargés.

Par conséquent, les recherches effectuées au cours de cette thèse et présentées au sein de ce document, poursuivent l'objectif d'améliorer l'utilisation des ressources radios dans des systèmes hétérogènes. Cela peut permettre, dans les zones où coexistent plusieurs technologies, d'offrir à chaque utilisateur une meilleur Qualité de Service (QdS). Les travaux développés dans cette thèse se focalisent sur l'allocation discrète de ressources puisque les technologies existantes utilisent majoritairement des systèmes MCA (allocation conjointe d'une modulation et d'un système de codage). En outre, cette allocation de ressources sera obtenue à l'aide d'un processus distribué.

Ainsi les recherches effectuées diffèrent sensiblement de l'etat de l'art puisqu'ils traitent simultanément des *réseaux hétérogènes*, d'une *allocation de ressources discrètes* et de *solutions distribuées*.

Ces travaux ont abouti à une modélisation du problème de l'allocation de ressources à prise de décision distribuée dans un système hétérogène, à sa résolution théorique ainsi qu'à la proposition et la mise en oeuvre de solutions pouvant être implémentées en pratique. Le contenu de cette thèse est brièvement écrit au sein des paragraphes suivants.

Le Chapitre 1 transcrit nos recherches préalables concernant l'allocation de ressources dans un systèmes hétérogène est effectué. Ce chapitre met en exergue l'intérêt des travaux développés dans cette thèse vis-à-vis de l'état de l'art.

Dans le Chapitre 2, des rappels concernant la théorie de l'optimisation convexe sont effectués. Différents algorithmes d'optimisation utilisés dans la suite du document (sous-gradient, algorithme d'optimisation discrète et convexe) sont explicités.

Au sein du Chapitre 3, le problème de l'allocation de ressources à prise de décision distribuée dans des réseaux mobiles hétérogènes est clairement posé. Ce problème peut être résolu de diverses manières et deux approches sont étudiées au cours de cette thèse (respectivement au Chapitre 4 et au Chapitre 5).

Les élements développés au Chapitre 4, nous permettent de résoudre, de manière distribuée, le problème de la minimisation de la puissance utilisée dans l'ensemble du système hétérogène, sous contrainte de débit minimal pour les utilisateurs et de puissance instantanée maximale pour les stations de base.

Les réflexions apportées au Chapitre 5, aboutissent à la résolution du problème de l'allocation de ressources lorsque l'objectif consiste à maximiser le débit moyen alloué, tout en assurant l'équité entre les utilisateurs et en respectant les contraintes de puissance instantanée aux stations de base.

Enfin une conclusion, concernant les travaux effectués ainsi que leurs perspectives ultérieures, est proposée dans le Chapitre 6.

Chapter 1

Introduction

1.1 Motivations

Au cours des dernières années, le domaine des télécommunications a considérablement évolué. En effet, nous avons assisté à l'émergence et au développement de nombreuses technologies radios. Les technologies nouvellement proposées ont permis d'améliorer les services offerts aux utilisateurs en termes de débit, de mobilité, de couverture, de Qualité de Service (QdS), etc. Un aspect intéressant des nouveaux systèmes de transmission est leur tendance à offrir des services convergents. Effectivement, de nombreuses technologies permettent désormais de transmettre simultanément ou alternativement des services vocaux, des services vidéos ainsi que des données. Pour bien comprendre les évolutions des technologies de télécommunication, les paragraphes suivant viseront à effectuer un récapitulatif des avancées récentes dans ce domaine.

Récemment, les évolutions dans le domaine des télécommunications ont été particulièrement importantes, notamment en ce qui concerne les réseaux mobiles cellulaires (WWAN pour Wireless Wide Area Network). En effet, la mise en place de la téléphonie de troisième génération (WCDMA) a permis de répondre à l'accroissement des exigences en terme de débit, sans altérer la qualité du service conversationnel héritée des technologies précédentes (GSM fondé sur le TDMA/FDMA). Toutefois, les limites en terme de débit et de synchronisation ont mené à l'élaboration de nouvelles technologies fondées sur une meilleure utilisation des ressources radios dans chaque cellule (CDMA/TDMA avec HSPA, OFDMA avec LTE).

Parallèlement, le domaine des réseaux cellulaires locaux (WLAN pour Wireless Local Area Network) a nettement évolué. Par exemple, le WiFi bénéficie d'une large couverture en Europe et offre à ses utilisateurs des services attrayants à usage domestique (suppression de la connectique au domicile, délivrance de débits élevés, mise à disposition d'un accès partagé) ou itinérant (extension des couvertures internet et intranet).

Enfin, le domaine des réseaux cellulaires métropolitains (WMAN pour Wireless Metropolitan Area Network) s'est développé sans toutefois atteindre son paroxysme. En effet, l'élaboration de technologies fondées sur l'OFDMA ne s'est pas encore concrétisée par une mise en oeuvre complète. Par exemple, le déploiement du WiMax reste à ef-

fectuer. Cette réalisation permettra d'une part l'accomplissement d'une Boucle Locale Radio en vue de combler les zones blanches (zones où la densité d'utilisateurs est faible) et d'autre part l'accroissement des débits fournis aux utilisateurs dans les zones déjà couvertes par d'autres technologies.

Cette multitude de technologies amène à s'interroger sur la rationalisation de l'utilisation des ressources radios. En effet, de nombreuses études ont montré que l'utilisation du spectre radio était inefficace puisque de nombreux réseaux n'utilisent pas la totalité de leurs ressources tandis que d'autres réseaux sont surchargés. Pour résoudre ce problème, deux approches sont généralement envisagées. La première approche consiste à proposer un nouveau standard permettant d'unifier l'ensemble des communications et fondé sur une technologie efficace. Cette approche est celle qui guide les recherches concernant IMT-Advanced. Cette approche présente l'inconvénient de nécessiter la mise en oeuvre d'une nouvelle technologie et/ou la modification des technologies existantes, ce qui se révèle très coûteux. A l'inverse, une seconde approche moins onéreuse, consiste à créer un mécanisme permettant de mieux allouer les ressources radios disponibles sans modifier les technologies actuellement déployées. C'est dans cette optique que s'intègre le travail effectué. Ainsi, l'objectif recherché est de proposer des solutions permettant d'allouer efficacement les ressources dans des réseaux mobiles hétérogènes (i.e. composé de réseaux non-interférants et utilisant des technologies distinctes).

L'allocation de ressources radios (ARR) dans des réseaux mobiles hétérogènes, peut reposer sur l'association de chaque utilisateur avec une unique station de base. Cette approche peut être aisément mise en oeuvre mais se révèle très rigide. En effet, un utilisateur pourra être rejeté parce que le débit qu'il requiert ne peut être atteint sur aucun réseau, alors qu'il pourrait être accepté en partageant sa demande sur plusieurs réseaux. De plus, la mobilité des utilisateurs peut se traduire par la dégradation de leurs conditions radios. Ce phénomène peut alors déclencher, pour un utilisateur, la nécessité d'un basculement ("handover") vers un réseau avec lequel il possède de meilleures conditions radios.

A l'inverse, une utilisation efficace des ressources radios incite à laisser aux utilisateurs la possibilité de se connecter simultanément à plusieurs réseaux. En effet, tandis que les terminaux multi-modes permettaient initialement de se connecter alternativement à différents réseaux, les nouveaux prototypes offrent désormais la possibilité de se connecter simultanément à différents réseaux. Ainsi, les nouveaux mobiles permettent de transmettre des informations simultanément sur plusieurs réseaux. De plus, ces terminaux sont capables de recouvrer l'information transmise lorsqu'ils sont en phase de réception.

Les paragprahes précédents nous ont permis de décrire les motivations concernant le positionnement de notre problème d'allocation de ressources. Le paragraphe suivant justifie l'approche utilisée pour résoudre le problème évoqué précédemment.

La résolution du problème de l'allocation de ressources dans des réseaux mobiles hétérogènes peut s'effectuer de manière centralisée ou distribuée. L'intérêt des méthodes distribuées est de limiter les échanges entre les différentes entités (utilisateurs et stations de base) du système hétérogène, de s'adapter rapidement aux variations des paramètres du scénario et d'éviter l'ajout d'une structure centralisée devant être dotée d'une capacité de calcul élevée. Par conséquent, le travail proposé s'est articulé autour de cette orientation.

Dans cette section, nous avons cherché à mettre en exergue les aspects attractifs des

travaux de recherche développés dans cet thèse. Ainsi, l'étude de l'ARR soumise à des contraintes de QdS se révèle être un problème attrayant dans un contexte hétérogène. En effet, déterminer la manière dont s'effectue l'allocation des ressources entre les stations de bases et les utilisateurs s'avère séduisant pour les fournisseurs d'accès qui cherchent à améliorer l'utilisation de leurs infrastructures existantes. En outre, cet aspect est bénéfique aux utilisateurs qui pourront bénéficier d'une meilleure QdS ainsi que de nouveaux services.

1.2 Etat de l'art

L'objectif de cette thèse est de proposer des mécanismes distribués permettant d'allouer efficacement les ressources radios dans un système hétérogène, pour les communications descendantes ("downlink", i.e. des stations de base vers les utilisateurs). Afin de situer le contexte dans lequel s'inscrivent nos recherches, il y a lieu d'effectuer l'état de l'art concernant les systèmes hétérogènes, puis de poursuivre par l'analyse des propositions existantes dans le domaine de l'allocation de ressources radios.

1.2.1 Etat de l'art concernant les systèmes hétérogènes

Parmi les publications traitant des systèmes hétérogènes, divers aspects sont abordés, notamment la planification du réseau hétérogène ainsi que les architectures et protocoles de gestion du réseau hétérogène.

La planification de réseaux hétérogènes consiste à définir mathématiquement les lieux les plus propices à l'installation de nouvelles stations de base en tenant compte des paramètres existants. Ce domaine de recherche a été peu étudié dans le cadre des réseaux hétérogènes. Toutefois, dans [Joha 05], Johansson et Zander proposent un modèle permettant d'effectuer la planification d'un réseau hétérogène. Ce modèle se fonde sur une métrique permettant d'évaluer le coût d'utilisation d'infrastructures hétérogènes (HSDPA + 802.11 par exemple) ou hiérarchiques (en utilisant une seule technologie mais en utilisant différents modes d'accès : Macro/Micro/Pico). Le modèle proposé permet également d'évaluer l'impact d'une modification de capacité ou de portée sur le coût d'utilisation. Une des limites de cette modélisation provient du fait que le coût d'un point d'accès prend uniquement en compte sa capacité et sa portée. En outre, la demande de débit n'est pas uniforme. Par contre, elle est générée à l'aide d'un modèle stochastique. Enfin, il est intéressant de noter que l'introduction d'une diversité de points d'accès permet de réduire le coût d'utilisation de l'infrastructure. En outre, le modèle proposé peut servir aux opérateurs dans le cadre de la planification de leurs réseaux bien que la qualité de service du système ne soit pas étudiée dans la publication.

Si la planification n'a que peu été étudiée dans le cadre des réseaux hétérogènes, de nombreuses publications ont traité de l'architecture et des protocoles nécessaires au fonctionnement simultané et efficace de technologies différentes. Parmi les nombreuses publications, deux grands types se distinguent. Le premier type de recherche consiste à proposer une architecture protocolaire générique pouvant inclure un vaste ensemble de

technologies. Une seconde approche consiste à étudier une architecture plus spécifique utilisant par exemple un réseau cellulaire et un réseau libre. Les approches fondées sur la proposition d'une architecture protocolaire sont présentées ci-dessous.

Dans [Gela 05b], Gelabert et al. justifient l'intérêt de la Gestion Commune des Ressources Radios (GCRR ou CRRM) dans la mise en oeuvre de technologies ultérieures à la 3ème génération (B3G : Beyond 3rd Generation). En effet, l'architecture qu'ils proposent pour effectuer la GCRR se révèle utile dans le cadre des réseaux mobiles hétérogènes puisqu'elle permet de mettre en commun l'ensemble des ressources des différents réseaux (constitution d'une réserve de ressources) et de les attribuer en fonction des nécessités des différents scénarios. Adopter une telle stratégie accroît la flexibilité et la granularité du système hétérogène. En outre, dans ce papier, les auteurs définissent les fonctions devant être implémentées par un système GCRR et la répartition de ces fonctions entre le processus global de GCRR et les processus locaux de GRR. Pour ce faire, une décomposition du problème d'allocation de ressources fondée sur l'utilisation de politiques/stratégies est préconisé (exemple : "favoriser tel type d'utilisateur", "diriger tel type de trafic sur tel réseau", etc.). Cette décomposition s'articule autour d'un couplage GCRR/GRR plus ou moins serré. La perspective de la transmission simultanée d'informations sur plusieurs technologies ("multi-homing") est évoquée sans toutefois être réellement traitée.

Dans [Sall 08], Sallent et al. proposent une architecture permettant de résoudre le problème de la GCRR inter- et intra-opérateurs. L'architecture proposée se révèle être hiérarchique et s'articule autour de protocoles à portée croissante (d'une reconfiguration locale à une reconfiguration globale). L'approche proposée peut être implémentée de manière distribuée à l'aide d'un canal cognitif dont l'utilisation à la demande permet un gain de signalisation significatif en comparaison avec son utilisation en diffusion globale (" broadcast ").

Dans [Havi 01], Havinga et al. proposent une architecture permettant de transmettre des informations sur un réseau hétérogène à un coût peu élevé. Leur architecture repose sur l'utilisation de terminaux multi-modes pouvant communiquer sur plusieurs réseaux simultanément. Cependant, leur proposition repose sur l'hypothèse qu'un flux de données ne peut être décomposé sur plusieurs réseaux car ceux-ci se révèlent spécialisés pour un certain type de communication (voix, vidéo, données, etc.). Ainsi, un changement de conditions radios ne peut engendrer qu'un basculement vers un autre réseau. En outre, l'architecture proposée suppose la mise en place d'un canal réservé à la signalisation ("Basic Access Network") et d'un réseau coeur commun (fonctionnant sur IP). Enfin, il est intéressant de noter que la mise en oeuvre de cette structure se révèle transparente pour l'utilisateur et gère la QdS ainsi que la mobilité.

Dans [Budd 05], Buddhikot et al. proposent une architecture permettant d'améliorer l'utilisation des ressources dans un système hétérogène. Dans ce système hétérogène, les utilisateurs accèdent de manière coordonnée à des ressources qui évoluent dynamiquement ("Coordinated Dynamic Spectrum Access"). Cette architecture repose sur l'émergence de terminaux mobiles reconfigurables (technologie "Software Defined Radio") et sur la mise en oeuvre d'un gestionnaire de spectre régional (GSR en anglais Regional Spectrum Broker) qui met en commun les ressources de différents réseaux. En outre, l'architecture repose sur des gestionnaires de ressources radios locaux (en anglais RAN MANager) qui négocient l'accès au spectre.

Dans [Rayc 03], Raychaudhuri et Jing proposent un protocole permettant de coordonner l'utilisation dans un système hétérogène composé de multiples réseaux libres de droit ("unlicensed"). Le protocole (nommé CSCC) repose sur la réservation d'une bande de fréquence restreinte et dédiée à l'envoi d'informations de signalisation. En effet, les utilisateurs transmettent périodiquement des informations concernant leurs communications à l'entité responsable de la coordination. Cette dernière agence l'utilisation des ressources en fonction des paramètres reçus (débit demandé, priorité, etc.). Ainsi, pour tirer parti du canal de signalisation, les utilisateurs sont supposés disposer de terminaux bimodes (au minimum). Cela leur permet en effet de communiquer sur une autre bande de fréquence. De plus, différentes politiques de prioritisation ("etiquette protocol") peuvent être sélectionnées car le protocole proposé est suffisamment générique pour cela (il permet d'assurer la compatibilité entre différents standards existants et futurs).

Dans [Vuce 08], Vucevic et al. proposent des mécanismes de négociation de la QdS dans le cadre de la connexion à une station d'un système hétérogène. L'architecture proposée permet également d'assurer la QdS de bout-en-bout quand un utilisateur effectue un basculement entre deux réseaux. L'architecture proposée s'appuie sur plusieurs entités : un gestionnaire commun de ressources radios (CRRM), un gestionnaire de bande passante (BB) et un coeur de réseau entièrement IP (all IP Core Network). La solution est testée sur la plateforme AROMA à laquelle les auteurs ont ajoutés plusieurs fonctionnalités (gestion des modules temps réels, algorithmes de sélection de station de base, etc.).

Dans [Carn 05], Carneiro et al. proposent une architecture intégrée permettant l'allocation coordonnée de ressources dans un système hétérogène composé de technologies diverses. L'architecture proposée permet d'allouer des ressources aux mobiles en leur assurant une QdS ininterrompue grâce à des couches d'abstractions (Abstraction Layers). En effet, un module permet de gérer la QdS de chaque connexion, tandis que des modules spécifiques à une technologie gèrent l'allocation de ressources dans le réseau correspondant. La QdS dans le système hétérogène est assurée par des mécanismes de réservation de ressources ainsi que par la gestion des basculements entre les réseaux lorsque les conditions le justifient.

Cette présentation des approches génériques permet de bien cerner leur portée théorique. Quoique ces approches se révèlent indispensables à la mise en place de nouveaux moyens de communication, elles se révèlent difficilement opérationnelles en l'absence de création d'un standard normalisé. En effet, il est particulièrement risqué pour un opérateur de déployer une technologie qui ne serait pas acceptée par ses concurents, puisque le déploiement d'un coût très élevé pourrait se révéler inutile. Par conséquent, des solutions intermédiaires pouvant être mises en oeuvre immédiatement, ont été proposées par de nombreux chercheurs et sont présentées ci-dessous.

Dans [Ferr 05], Ferrus et al. étudient l'allocation de ressources dans un réseau mobile hétérogène, composé de plusieurs technologies (GSM, UMTS, WLAN). Les auteurs utilisent un réseau opérateur IP-UMTS supportant la différenciation de services. Ainsi, l'intégration s'effectue au niveau réseau. Les auteurs proposent ici une architecture permettant d'effectuer des basculements ("handovers") entre les réseaux, c'est-à-dire fondée sur la connaissance globale du système et non sa connaissance uniquement locale. En outre, l'architecture proposée tient compte du type de trafic et s'inscrit dans le cadre du projet européen EVEREST.

Dans [Vuli 05], Vulic et al. proposent d'implanter des stations WLAN dans un réseau UMTS afin de bénéficier, d'une part des avantages en terme de couverture fournis par le réseau cellulaire, et d'autre part des avantages de débit offerts par le réseau local. L'intégration précitée s'effectue au niveau radio et l'architecture permettant de parvenir à la mise en commun des ressources radios est mise en exergue dans cette publication. Deux types d'architectures sont proposés. La première consiste à faire transiter l'ensemble de la signalisation sur le réseau UMTS, tandis que la seconde permet à chaque réseau de fonctionner de manière indépendante d'un point de vue radio. Dans ce papier, la gestion des ressources radios s'effectue grâce à un protocole de supervision (CRRM) qui peut être centralisé au niveau du RNC ou distribué entre les différents réseaux (RNC et IWU). Ce protocole permet de déclencher le basculement entre les deux réseaux, selon différents paramètres (charge, conditions radios, etc.).

Dans [Pere 05], Perez-Romero et al. proposent une architecture permettant d'effectuer la GCRR (CRRM) dans des réseaux mobiles hétérogènes. Cette architecture est hiérarchique et repose sur la mise en commun de ressources dans une réserve commune gérée par les stations de base GSM (BSC) ou UMTS (RNC). Les ressources précédentes sont gérées par un algorithme RRM. En outre une entité assure la gestion des ressources issues de plusieurs RRM à l'aide d'un algorithme de CRRM et coordonne l'utilisation des réseaux en communiquant avec d'autres entités utilisant également l'algorithme de CRRM. Les interactions entre le CRRM et le RRM permet d'échanger des informations de contexte ainsi que des informations concernant les politiques de décision à adopter. Différents types d'implémentation sont envisagés et dépendent de la répartition des rôles entre CRRM et RRM. La solution proposée suppose l'existence de terminaux multi-modes.

Dans [Cris 03], Cristache et al. étudient des systèmes hétérogènes UMTS/WLAN. Les auteurs présentent tout d'abord les différentes architectures pouvant être mises en place lorsque le WLAN est implémenté sous forme d'infrastructure composée de différents points d'accès ("access points"). Les auteurs montrent que la complexité de l'implémentation de la coopération entre les systèmes croît avec le couplage entre les technologies. Cependant, le coeur de l'étude correspond à la mise en oeuvre d'une architecture utilisant un réseau UMTS et des systèmes WLAN ad-hoc. L'étude montre ainsi que cette architecture améliore le système UMTS en réduisant la puissance de transmission des utilisateurs éloignés puisque ceux-ci passent désormais par des relais. En outre, l'architecture proposée accroît la capacité du système grâce à l'utilisation d'un spectre plus large (qui se révèle en outre libre de droits).

Dans [Pere 03], Jasemuddin propose une architecture complète permettant de faire fonctionner simultanément UMTS et WLAN. Cette architecture permet de transmettre des informations concomitamment sur les deux réseaux. L'architecture décrit également les procédures à effectuer pour assurer la connexion aux réseaux, le basculement entre les réseaux ainsi que le fonctionnement des services de voix et de données sur un réseau coeur commun (architecture tout-IP). L'auteur fait le choix de transmettre la voix uniquement sur le réseau UMTS, tandis que les données sont transmises sur le réseau WLAN si la connexion est possible et sur le réseau UMTS sinon. L'architecture permet de transmettre des données tout en ayant une conversation téléphonique.

Dans [Luo 03], Luo et al. proposent un système de GCRR adapté aux réseaux hétérogènes composés d'un réseau cellulaire 3G (UMTS) et d'un réseau libre WLAN (HY-

PERLAN 2). Leur proposition vise à tirer avantage de la connectivité élevée des systèmes cellulaires et de l'importante capacité de transmission de données des systèmes à portée réduite. En outre, l'objectif poursuivi par les auteurs consiste à permettre une utilisation efficace des deux réseaux en les faisant collaborer dans les domaines de la synchronisation, du contrôle d'admission, de l'ordonnancement ("scheduling"), de la répartition de charge, de l'utilisation de leurs ressources radios, etc. De plus, les auteurs supposent que les terminaux ont la possibilité d'utiliser simultanément plusieurs interfaces radios pour transmettre ou récupérer leurs données ("multihoming"). Cette capacité est très intéressante car elle permet, par exemple, de conserver une communication continue sur le réseau cellulaire tout en bénéficiant des débits élevés des systèmes ad-hoc. Toutefois, pour permettre le multihoming, il est nécessaire de contrôler la latence sur les deux réseaux (une bonne synchronisation est nécessaire sinon il est impossible de reconstituer les données dans un temps acceptable). Les auteurs étudient également la transmission de différents types de services (http et vidéo).

Dans [Magn 05], Magnusson et al. proposent une architecture complète permettant d'effectuer la GCRR dans le cadre du projet Réseaux Ambiants (projet cherchant à exploiter la diversité radio existante pour un même opérateur ou pour plusieurs). Les auteurs décrivent les processus nécessaires à la mise en commun et la gestion des ressources radios dans ce contexte hétérogène et la communication entre les différentes entités de gestion des ressources radios. La solution proposée peut être implémentée de diverses manières : en se fondant sur la sélection de station de base ("MRAS" dans la publication) ou en utilisant simultanément plusieurs stations ("parallel MRTD" dans la publication). Les auteurs préconisent l'utilisation de la première solution dans le cas de réseaux relativement chargés (pour assurer la QdS) et de la seconde dans le cas de réseaux relativement peu utilisés (pour améliorer les débits). Bien que les utilisateurs proposent de transmettre des données sur plusieurs réseaux simultanément, ils n'étudient pas les difficultés engendrées par cette solution.

Les précédents paragraphes ont permis d'effectuer l'état de l'art concernant les systèmes hétérogènes. Cette analyse des solutions existantes dans des domaines proches de notre sujet de recherche se poursuit, dans la section suivante, par l'étude des propositions effectuées dans le cadre de l'allocation de ressources radios.

1.2.2 Etat de l'art concernant l'allocation de ressources radios (ARR)

Dans ce qui suit, nous étudions les propositions réalisées dans le domaine de l'ARR et les comparons à notre sujet de recherche. Rappelons que l'objectif poursuivi dans le cadre de nos travaux consiste à proposer des mécanismes devant être facilement implémentés de manière distribuée. L'étude réalisée doit donc permettre d'identifier des solutions efficaces pour résoudre le problème de l'ARR. Au cours des dernières années, les propositions effectuées par les scientifiques pour résoudre le problème de l'ARR se sont révélées nombreuses, tant pour le contexte mono-technologique que pour celui multi-technologique. Par conséquent, nous étudierons successivement les deux approches retenues.

ARR dans un réseau mono-technologique

Le problème de l'ARR dans des réseaux mono-technologiques a été étudié à la lumière de nombreuses méthodes : centralisées ou non, fondées sur l'optimisation ou la théorie des jeux et suivant différents objectifs (minimisation de puissance ou maximisation de débit). L'ensemble des méthodes évoquées seront analysées dans les paragraphes suivants.

Avant de poursuivre, il est pertinent de remarquer que la plupart des technologies existantes utilisent des systèmes de Modulation et Codage Adaptatifs (MCA). Cela signifie que les technologies s'adaptent aux conditions radios en allouant simultanément une modulation et un codage choisis parmi un ensemble de couples de valeurs disponibles. Par conséquent, les mécanismes que nous proposerons viseront à résoudre le problème de l'ARR dans des réseaux mobiles hétérogènes composés de technologies utilisant les systèmes MCA.

De nombreuses recherches proposent des solutions centralisées fondées sur l'optimisation convexe [Boyd 04], [Yu 04] et [Rhee 00a] qui utilisent la théorie de Lagrange et l'algorithme de remplissage d'eau ("waterfilling"). Ces solutions supposent l'existence d'une entité principale devant connaître l'ensemble des paramètres spécifiques à chaque utilisateur (conditions radios, type de terminal, QdS requise , etc.). L'allocation est ainsi calculée par l'entité principale à chaque instant et transmise à chaque utilisateur ce qui implique la mise en oeuvre d'une signalisation élevée.

Pour pallier ces inconvénients, Chiang et Sutivong ont proposé une méthode permettant de résoudre le problème de l'ARR sous contrainte [Chia 03]. Cette méthode consiste à résoudre le problème évoqué en combinant la théorie de Lagrange et la Programmation Géométrique. En effet, de nombreux problèmes d'optimisation convexe peuvent être résolus en utilisant la Programmation Géométrique (contrôle d'admission, allocation de ressources, contrôle de puissance, etc.), après avoir été transformés à l'aide de la fonction logarithme (le logarithme d'une somme d'exponentielles est une fonction convexe). Toutefois, la solution proposée requiert une importante signalisation, se fonde sur l'optimiation continue et admet une complexité de convergence polynomiale.

Afin de limiter la signalisation dédiée à l'allocation de ressources dans les réseaux sans fils, la théorie des jeux non-coopératifs a été proposée dans [Sara 02], [Han 05b] et [Nash 53]. Cependant, les solutions distribuées que cette théorie permet d'obtenir, aboutissent à une allocation de ressources sous-optimale, malgré l'utilisation de méthodes incitant les utilisateurs à adopter un comportement socialement responsable (utilisation de fonctions d'utilité).

Ainsi, dans [Sara 02] Saraydar et al. utilisent un système CDMA sans MCA en liaison acendante (uplink). Le jeu non-coopératif initial aboutit à un équilibre de Nash qui consiste à faire fonctionner l'ensemble des utilisateurs au même Rapport Signal à Interférence plus Bruit (RSIB en français et SINR en anglais). L'équilibre obtenu n'étant pas efficace, un mécanisme de tarification linéaire des resources ("linear pricing") est ajouté au système et permet d'obtenir un équilibre Pareto-supérieur qui n'atteint cependant pas l'optimalité. Le mécanisme de tarification peut être interprété comme une forme de coopération implicite, guidée par les stations de base. La tarification ne remet pas en cause la compétition entre les utilisateurs en ce qui concerne l'utilisation des ressources

radios (en l'occurence leur puissance d'émission).

Dans [Han 05b], Han et Liu cherchent à résoudre le problème de l'allocation de ressources d'un système MCA en liaison acendante, via un contrôle de puissance associé à une gestion du débit système. Ce problème peut s'écrire sous une forme matricielle bilinéaire reliant la puissance utilisée et le débit système. Pour résoudre le problème étudié, les auteurs proposent d'implémenter: 1- un jeu non coopératif consitant à déterminer la puissance à émettre côté utilisateurs; et 2- un jeu non coopératif consistant à déterminer le débit alloué côté stations de base. Cette solution est ensuite comparée à un algorithme centralisé utilisant l'optimisation convexe et se révèle sous-optimale.

Par conséquent, des solutions fondées sur des modèles coopératifs ont été développées, notamment dans le cadre de l'analyse des réseaux et du partage de spectre [Pete 92], [Zhou 97] et [Han 05a]. Ces solutions permettent d'obtenir une solution bénéfique pour l'ensemble des joueurs en résolvant un critère commun. Toutefois, la théorie des jeux coopératifs requiert une signalisation conséquente, puisque les entités doivent se coordonner entre elles. Cela résulte en un accroissement de la signalisation et est donc contradictoire avec l'objectif affiché.

Dans [Han 05a], Han et al. proposent une solution permettant d'allouer les porteuses, les débits et la puissance dans un système OFDMA. Cette solution est fondée sur la Négociation de Nash (Nash Bargaining Solution). Cette approche repose sur la théorie des jeux coopératifs et se révèle être une généralisation de l'allocation de ressources proportionnelle ("Generalized Proportionnal Fair Allocation"). La solution proposée consiste à résoudre des problèmes de Négociation de Nash à deux utilisateurs après avoir constitué des coalitions de deux utilisateurs. Obtenir une telle coalition est possible en utilisant un algorithme aléatoire ou bien à l'aide de l'algorithme Hongrois [Kuhn 55]. Ainsi, l'algorithme proposé nécessite une signalisation importante.

Dans la littérature scientifique, le problème de l'allocation de ressources dans des réseaux mobiles reposant sur une unique technologie, a été étudié suivant l'un des deux aspects suivants : 1- Adaptation du Débit (Rate Adaptive), qui consiste à maximiser la capacité totale du système (se référer à [Jang 03] et [Rhee 00b] pour obtenir davantage d'informations à ce propos), et 2- Adaptation de la Puissance (Margin Adaptive), qui consiste à minimiser la puissance totale utilisée au sein du système sous contrainte d'une QdS minimale pour chaque utilisateur (se référer à [Wong 99] et [Kiva 03]). Minimiser la consommation de puissance est très attrayant pour les opérateurs de téléphonie car cela permet de réduire le coût de fonctionnement du réseau, de réduire l'interférence co-canal (interférence vis-à-vis des autres cellules utilisant la même fréquence) et d'accroître le nombre d'utilisateurs pouvant bénéficier d'une prestation de qualité (ce qui résulte en des gains économiques potentiels).

ARR dans un réseau multi-technologique (hétérogène)

Nous avons observé que le problème de l'ARR dans les réseaux mono-technologiques avait été traité de manière exhaustive. Les paragraphes suivant visent à détailler les propositions concernant l'ARR dans les réseaux multi-technologiques. Quoique plus proche de notre sujet d'étude, cette approche se révèle moins aboutie en termes d'implémentation.

En effet, de nombreuses publications se contentent de proposer des heuristiques permettant d'améliorer quelque peu l'ARR sans chercher à la rendre optimale sous un certain critère. En outre, nous observerons que les publications étudiées abordent principalement l'architecture protocolaire et l'allocation de ressources par sélection de réseau, ce qui ne permet pas de bénéficier complètement des capacités multi-technologiques des nouveaux systèmes de communication (smartphones).

L'étude bibliographique permet de répertorier les publications selon le type d'allocation (sélection de réseau avec ou sans basculement, transmission simultanée sur plusieurs réseaux) et le type de décision prise (utilisation de politiques préétablies, de fonctions d'utilité ou de critères radios plus théoriques).

De nombreux articles ont proposé de résoudre le problème de l'ARR en utilisant la sélection de station de base sans basculement. Ces articles s'articulent autour de plusieurs méthodes de résolution du problème de l'ARR. Commençons par la présentation des travaux concernant les méthodes fondées sur des modèles économiques (fonctions d'utilité ou tarification).

Dans [Gela 05a], Gelabert et al. étudient l'allocation de ressources dans des réseaux hétérogènes en se concentrant sur la sélection de la technologie utilisée lors de l'initialisation de la communication. Aucun basculement ("handover") entre les technologies n'est envisagé. Deux politiques d'allocation sont étudiées : une sélection se fondant uniquement sur le service requis et une sélection cherchant à répartir la charge sur les différents réseaux. Les résultats de simulation montrent que la répartition de charge accroît le débit perçu au détriment du délai de transmission.

Dans [Badi 05b], Badia et al. étudient l'allocation de canaux dans un réseau mobile hétérogène composé de deux technologies distinctes (n'interférant pas entre elles). Les auteurs utilisent une métrique probabiliste. Cela leur permet de calculer la probabilité d'acceptation d'une requête d'un utilisateur et d'en déduire le revenu total du système ainsi que l'utilité totale perçue par les utilisateurs. Les auteurs comparent les deux stratégies pouvant être adoptées par les réseaux : coopération ou indépendance. Les résultats sont conformes aux attentes puisqu'ils mettent en évidence la supériorité de la stratégie coopérative. En outre, quelque soit la stratégie, les résultats montrent que la fonction de revenu est concave et admet un maximum (atteint au prix optimal).

Dans [Badi 05a], Badia et al. étudient l'allocation de ressources dans des réseaux hétérogènes en cherchant une solution qui satisfasse les utilisateurs et les gestionnaires du réseau en ce qui concerne les points de vue technique et économique. Les solutions envisagées consistent à sélectionner la Technologie d'Accès Radio (TAR ou RAT en anglais) à utiliser. L'utilisation des ressources radios est plus efficace quand une fonction d'utilité est utilisée pour effectuer cette décision. Enfin quand une estimation du taux d'acceptation est effectuée, le fonctionnement du système est meilleur qu'avec l'utilisation brute de la fonction d'utilité.

Dans [Gozl 08], Gozalvez et al. proposent des méthodes pour la Gestion Commune des Ressources Radios (GCRR ou CRRM). Ces méthodes permettent de sélectionner la Technologie d'Accès Radio (TAR ou RAT en anglais) ainsi que le nombre de ressources à utiliser afin de satisfaire la QdS requise. Ces méthodes sont fondées sur l'outil d'optimisation qu'est la programmation linéaire (utilisation de la méthode du simplex dans ce

papier). En outre, l'optimisation proposée s'appuie sur des fonctions d'utilité propres à chaque type de service tandis que les technologies radios utilisent de techniques MCA (AMC en anglais).

Dans [Luca 08], Lucas-Estan et al. proposent une solution permettant de résoudre le problème de la GCRR en sélectionnant simultanément la technologie à utiliser et la quantité de ressources à allouer à l'intérieur de celle-ci afin d'assurer le service demandé. Les auteurs utilisent pour cela des fonctions d'utilité pour chaque type de service pouvant être utilisé sur les différents réseaux. En outre, ils évaluent la quantité de ressources disponibles dans chaque réseau afin de déterminer l'utilité pouvant être obtenue, en remarquant que chaque technologie utilise des techniques MCA (AMC en anglais). La politique d'allocation de ressources retenue vise à maximiser l'efficacité spectrale tout en garantissant la QdS perçue par les utilisateurs. Cela revient à résoudre un problème d'optimisation dont la solution est obtenue via un algorithme de programmation linéaire (linear programming). L'allocation globale est calculée à chaque nouvelle connexion ou à chaque fin de connexion.

Les paragraphes précédents témoignent de l'ampleur des travaux fondés sur un modèle économique. Ces modèles permettent la sélection d'un réseau au sein d'un système hétérogène. Le paragraphe suivant montre qu'il est également possible de réaliser cette sélection à l'aide de l'optimisation de critères radios théoriques.

Dans [Chen 08], Chen et al. proposent une solution permettant d'effectuer la GCRR dans des systèmes hétérogènes et a pour objectif de minimiser le coût d'utilisation des ressources tout en respectant les contraintes de délais des services requis. Les auteurs proposent de résoudre le problème d'optimisation sous-jacent en utilisant des méthodes d'optimisation linéaire (algorithme du simplex) et en faisant l'hypothèse que tous les réseaux seront utilisés successivement (dans le temps) pour chaque communication. Cette hypothèse permet de réduire le coût calculatoire qui reste cependant élevé. L'allocation doit être recalculée à chaque changement significatif (lorsqu'un seuil est atteint) des demandes ou des conditions radios.

De nombreuses méthodes permettant de résoudre le problème de l'ARR sans mettre en oeuvre de basculement en oeuvre ont été décrites ci-dessus. L'absence de basculement entre les technologies implique que ces méthodes présentent l'inconvénient de nécessiter une terminaison d'une ou plusieurs connexions quand un utilisateur se déplace. Par conséquent, nous étudions ci-dessous les solutions permettant d'effectuer le basculement entre deux stations de base. Comme précédemment, il est possible de résoudre ce problème via différentes méthodes.

La méthode la plus directe qui permet d'implémenter ce basculement consiste à utiliser des politiques préétablies. En effet, dans ce cas il n'est pas nécessaire de prévoir des mécanismes complexes pour effectuer le basculement. Celui-ci est en effet systématiquement déclenché lorsque des seuils ("triggers" ou "thresholds") sont atteints.

Dans [Shen 07], Shen et al. proposent deux algorithmes permettant d'allouer des ressources à des utilisateurs pouvant utiliser des services différents, dans des réseaux hétérogènes. Les auteurs partent de l'hypothèse que les terminaux utilisent une seule technologie pour communiquer et basculent entre les stations de base selon leurs conditions radios et le type de service qu'ils choisissent. Ainsi, les auteurs se focalisent sur l'étude des

basculements intra-technologies ("horizontal handoffs") et inter-technologies ("vertical handoffs") à l'aide de politiques établies au préalable. En outre, ils améliorent les solutions existantes en proposant un système de préemption des ressources, de sorte qu'une communication avec une priorité faible pourra se voir retirer l'utilisation d'une partie des ressources au profit d'une communication de priorité supérieure.

Dans [Murr 03], Murray et Pesch proposent une gestion des ressources radios dans un système hétérogène (UMTS/EDGE) fondé sur l'utilisation d'une entité de répartition des utilisateurs en fonction des politiques de gestion de la QdS. Les politiques de gestion reposent sur la définition de surfaces de capacité propres à chaque type de réseau (UMTS ou EDGE). Ces surfaces de capacité reflètent les demandes de services (voix, vidéo, données) pouvant être traitées par le réseau et sont obtenues par simulation. L'entité de répartition intervient à différents moments : connexion d'un utilisateur, modification de ses conditions radios ... Les résultats de simulation montrent que cette approche est plus efficace qu'une répartition aléatoire entre les réseaux.

Dans [Mott 02], Motte et al. proposent un module permettant d'effectuer la GCRR (CRRM) dans des réseaux cellulaires (FDMA, TDMA, CDMA) et dans le cadre du projet TRUST. Les auteurs considèrent des mobiles reconfigurables ("SDR context"). Leur solution permet d'étudier l'opportunité d'un basculement d'un réseau vers un autre, de le déclencher puis de l'effectuer en fonction des conditions perçues par les mobiles (canaux de propagation notamment) et par les stations de base (charge de celles-ci notamment). Les simulations effectuées montrent les gains obtenus grâce à la gestion conjointe des réseaux mais reposent sur des heuristiques peu élaborées.

Une méthode plus théorique pour déclencher un basculement entre deux stations consiste à optimiser les critères radios. Les papiers suivants proposent différentes approches s'inscrivant dans cette démarche.

Dans [Pere 07], Romero et al. se placent dans le cadre de l'utilisation de réseaux TDMA et CDMA. Pour obtenir leur débit, les utilisateurs ne peuvent se connecter qu'à un des deux réseaux. Il s'agit donc d'un processus de sélection de réseau ("RAT-selection") qui s'effectue grâce à des basculements inter-technologies ("vertical handovers"). Les auteurs calculent l'allocation de ressources optimale en ce qui concerne la minimisation de la probabilité de coupure sur la liaison ascendante ("uplink"). Ainsi, les auteurs cherchent à déterminer de quelle manière les utilisateurs se connectent aux différentes technologies de manière à assurer le fonctionnement d'un service identique (voix ou données) et en minimisant la probabilité de coupure sur la liaison ascendante

Dans [Papa 06], Papadaki et Friderikos proposent une solution au problème de la GCRR (CRRM) fondée sur l'équilibre de charge ("load balancing"). Cet équilibre est obtenu en effectuant un calcul différé des basculements verticaux à réaliser. Chaque traitement par lot ("batch") est effectué à l'aide d'un algorithme d'optimisation mixte entière-linéaire (MLIP). Les simulations montrent qu'il existe un compromis entre le délai induit par le processus différé et l'efficacité en termes de ressources de la solution GCRR. L'introduction du processus différé permet en outre de diminuer la probabilité de rejet d'un basculement.

Dans [Blau 07], Blau et al. proposent un algorithme de sélection de réseau permettant d'améliorer l'allocation de ressources dans un réseau hétérogène (comparaison avec

la solution d'équilibre de charge -load balancing strategy-). Les auteurs expriment initialement le problème de la GCRR sous la forme d'un problème d'assignation généralisée ("Generalized Assignment Problem") de complexité non-polynomiale. Ce problème fait intervenir notamment la qualité des canaux de propagation et le type de service requis. Par contre, les auteurs n'envisagent pas la possibilité qu'un utilisateur puisse communiquer simultanément sur plusieurs réseaux. Pour résoudre efficacement le problème posé (dans un temps raisonnable), les auteurs proposent une solution sous-optimale en relâchant une des contraintes de ce problème. Ainsi, le nouveau problème se résout via un algorithme d'optimisation linéaire (simplex, ellipsoïde, intersection de plans) et est ainsi d'une complexité calculatoire réduite (mais toujours élevée). La résolution du problème nécessite toutefois une signalisation significative.

Enfin, des méthodes originales ont été étudiées dans la littérature scientifique et présentent diverses caractéristiques présentées ci-dessous.

Dans [Coup 08], Coupechoux et al. proposent une solution au problème de l'allocation initiale de ressources dans une cellule composée de deux stations (UMTS et WLAN) colocalisées. La solution proposée est uniquement fondée sur la sélection de l'une ou l'autre des stations et se fait grâce à un processus de décision semi-markovien ("Semi-Markovian Decision Process"). La solution proposée dans le papier est optimale vis-à-vis d'un critère de satisfaction des utilisateurs prenant en compte le débit qui leur est fourni ainsi que la probabilité de rejet de leurs communications.

Dans [Wang 07], Wang et al. proposent une solution permettant d'effectuer la GCRR dans des réseaux hétérogènes (CDMA et TDMA/FDMA). Partant de la constatation que les réseaux CDMA sont plus sujets aux interférences, les auteurs proposent de réduire la couverture du CDMA (en instaurant un seuil d'atténuation -path loss-) tout en aiguillant les autres utilisateurs vers le TDMA/FDMA. Cette solution permet de diminuer la probabilité de rejet des utilisateurs et accroît le débit offert par comparaison avec la solution standard d'équilibre de charge ("load balancing"). La valeur d'atténuation de canal à partir de laquelle un basculement est effectué entre les réseaux doit être sélectionné précautionneusement afin d'assurer un bon fonctionnement du système hétérogène.

Dans [Murr 04], Murray et Pesch complètent la publication [Murr 03] en introduisant un système dynamique de réservation de ressources dédié à la gestion des basculements entre les réseaux. Ce système fonctionne à l'aide d'un réseau de neurones. En effet, les basculements suivent un schéma prédictif (heure, jour, localisation dans la cellule, mobilité) ce qui permet de prédire la quantité de ressources à réserver pour permettre leur mise en oeuvre. Dans [Murr 07], les auteurs procèdent de manière similaire, hormis le fait qu'ils utilisent la logique floue pour évaluer la nécessité d'effectuer un basculement.

Dans les paragraphes précédents, l'allocation de ressources dans les réseaux hétérogènes a fait l'objet d'une bibliographie étendue. Jusqu'ici, seules les approches fondées sur la sélection d'un réseau ont été évoquées. Nous allons présenter quelques publications dont l'approche permet d'obtenir simultanément des ressources radios de plusieurs réseaux. Ces publications tirent parti des nouveaux moyens de communications qui peuvent utiliser simultanément plusieurs interfaces radios.

Dans [Suli 04], Suliman et al. étudient l'allocation de ressources dans des réseaux hétérogènes, pour des utilisateurs dont les applications peuvent être transmises de manière

partagée entre les différents réseaux. Les auteurs proposent une solution fondée sur la théorie des jeux coopératifs, où l'allocation de ressources est obtenue grâce à des jeux de Stackelberg multiniveaux effectués par les différents réseaux.

Dans [Chen 06], Chen et al. étudient les mécanismes d'allocation de ressources dans des réseaux mobiles hétérogènes avec réservation de ressources (transmission de contenu vocal ou vidéo mais pas de trafic élastique). Les auteurs considèrent des terminaux multi-modes. En outre, ils formalisent le problème de l'allocation de ressources commune (CRRM) sous la forme d'un problème d'équilibre de charge ("load balancing"). L'équilibre de charge est calculé pour un ensemble de tâches temporaires réalisées par des processeurs indépendants. Le problème est résolu en utilisant un graphe pondéré Mobiles-Stations. Les auteurs étudient la complexité des différents algorithmes pouvant être mis en oeuvre. Une solution sous-optimale est proposée pour pallier les difficultés liées à l'implémentation de la solution optimale.

1.3 Contributions

Dans cette thèse, nous cherchons à améliorer l'utilisation des ressources radios pour la transmission de données, lorsque des utilisateurs se situent dans des zones où coexistent plusieurs technologies. Puisque la plupart des technologies existantes utilisent des systèmes MCA (allocation conjointe d'une modulation et d'un système de codage), nous restreindrons l'étude à l'allocation discrète de ressources. En outre, cette allocation de ressources sera obtenue à l'aide d'un processus distribué.

Ainsi le travail réalisé diffère de l'etat de l'art du fait qu'il traite à la fois des *réseaux hétérogènes*, d'une *allocation de ressources discrètes* et de *solutions distribuées*. La nouveauté consiste à traiter ces aspects simultanément.

Les travaux réalisés au cours de cette thèse ont visé à modéliser le problème évoqué au paragraphe précédent, puis à le résoudre de manière théorique et enfin à proposer des solutions implémentables en pratique. Ainsi, les contributions de cette thèse sont les suivantes:

- Le problème de l'allocation de ressources à prise de décision distribuée dans des réseaux mobiles hétérogènes est clairement posé dans le Chapitre 3. Il se révèle être un problème d'optimisation discrète sous contrainte tel que présenté au Chapitre 2. Ce problème peut être traité de nombreuses manières suivant que l'on souhaite résoudre le problème de manière instantanée, de manière moyenne, avec des contraintes instantanées, des contraintes moyennes ou en modifiant le critère à optimiser. Chaque problème induit des contraintes dans l'élaboration de sa solution qui possède alors certaines caractéristiques spécifiques (convergence, complexité, équité, débit total et puissance totale dans le système hétérogène, ...). Le choix d'un problème à résoudre plutôt qu'un autre dépend clairement des besoins de l'opérateur chargé de réaliser le système d'allocation de ressources hétérogènes. Dans cette thèse, nous résolverons plusieurs problèmes d'allocation de ressources et montrerons leurs avantages et inconvénients respectifs dans les Chapitres 4 et 5.

- Plus spécifiquement, dans le Chapitre 4, nous nous proposons de résoudre, de manière distribuée, le problème de la minimisation de la puissance utilisée dans l'ensemble du système hétérogène sous contrainte de débit minimal pour les utilisateurs et de puissance instantanée maximale pour les stations de base. La solution optimale du problème instantané est obtenue en transformant le problème d'optimisation primal en problème dual, puis en recherchant itérativement l'allocation qui minimise le lagrangien obtenu précédemment. Pour ce faire, le problème dual est décomposé en sous-problèmes pouvant être résolus indépendamment par les utilisateurs ou les stations de base (selon le cas). Ainsi, chaque utilisateur résout un problème d'optimisation discrète à l'aide d'un algorithme s'inspirant de celui développé par Shoham et Gersho [Shoh 88] dans le cadre du codage source. La solution de ce sous-problème constitue la requête que l'utilisateur adressera aux stations de base. Cette solution dépend du paramètre de charge de chaque réseau qui est intialement considéré nul. Une fois les requêtes transmises aux stations de base, celles-ci calculent le paramètre de charge associé à l'utilisation de chaque réseau. Ce dernier est mis à jour à l'aide de l'algorithme du sous-gradient. Ainsi, un réseau surchargé verra son paramètre de charge augmenter, tandis qu'un réseau dont les demandes sont trop faibles verra son paramètre décroître. Le processus d'optimisation décrit ci-dessus est répété jusqu'à ce que la convergence soit atteinte. La solution proposée est très communément utilisée pour la résolution de problèmes d'optimisation. La difficulté majeure dans la résolution de ce problème réside dans le choix du pas d'itération. En effet, pour assurer la convergence du problème d'optimisation ce pas doit être suffisamment faible, tandis que pour limiter le nombre d'itérations, le pas doit être suffisamment élevé. En outre, le pas d'itération dépend de l'ensemble des conditions de la simulation (valeurs des canaux de propagation, position des utilisateurs, etc.).

  Pour améliorer la vitesse de convergence du problème d'allocation instantanée, un nouvel algorithme est proposé. Cet algorithme adopte la même décomposition du problème que précédemment, et les requêtes des utilisateurs sont également identiques. Par contre, le sous-problème devant être résolu par les stations de base est différent. En effet, toute station suchargée cherche désormais à déterminer immédiatement la meilleure allocation disponible. Cette station est ensuite retirée du processus d'optimisation. Par conséquent, la vitesse de convergence de l'algorithme est, dans le pire des cas, égale au nombre de réseaux considérés. Par conséquent, la convergence est plus rapide et l'implémentation est facilitée car la structure de trame peut être fixée préalablement à toute mise en oeuvre. En outre, la solution proposée se révèle quasi-optimale comme en témoignent les résultats décrits dans le Chapitre 4 et les simulations effectuées dans le cadre de la publication disponible en Annexe 7.2.

- Dans le Chapitre 5, nous résolvons le problème de l'allocation de ressources lorsque l'objectif consiste à maximiser le débit moyen alloué tout en assurant l'équité entre les utilisateurs et en respectant les contraintes de puissance instantanée aux stations de base. La solution proposée s'articule en deux sous-problèmes résolus respectivement par les mobiles et les stations de base. Plus précisément, l'algorithme permet de résoudre : 1- un sous-problème permettant de déterminer les requêtes

des mobiles, consistant à minimiser la congestion dans le réseau, en supposant que la capacité à l'instant considéré est égale à la capacité mesurée précédemment et 2- un sous-problème d'optimisation visant à maximiser la capacité de chaque station de base. Il est primordial de bien comprendre que le sous-problème 1 est traité indépendamment par les mobiles et que le sous-problème 2 est traité indépendamment par les stations de base.

Le sous-problème traité par les utilisateurs, repose sur une mesure de la congestion (égale à la différence entre les demandes et les débits obtenus) et à la maîtrise de cette congestion. En effet, pour obtenir une allocation stable en moyenne, il y a lieu d'assurer que les demandes des utilisateurs correspondent aux capacités des stations de base. Ainsi, afin d'assurer que la congestion soit nulle, la solution développée vise à adapter les demandes des utilisateurs à chaque instant, à l'aide d'une méthode proche de celle utilisée dans TCP Vegas [Low 01].

Le sous-problème traité par les stations de base est un problème d'optimisation convexe et discrète qui permet de maximiser la capacité du réseau. Cette optimisation s'inspire de l'algorithme proposé par Shoham et Gersho dans [Shoh 88] et se révèle très similaire aux optimisations effectuées au Chapitre 4.

Chapter 2

Rappels sur l'Optimisation Convexe

2.1 Résumé du Chapitre

Dans ce chapitre, nous effectuons des rappels concernant l'optimisation convexe et son application à notre problème d'allocation de ressources dans un système composé de stations hétérogènes.

La première section démontre que le problème étudié correspond bel et bien à un problème d'optimisation convexe. Pour ce faire, l'écriture traditionnelle d'un problème d'optimisation est proposée, puis le lien avec le problème traité est réalisé.

La seconde section présente les algorithmes généralement utilisés dans le cadre de l'optimisation convexe. Etant donné que le problème étudié est discret (les mobiles choisissent un mode de transmission MCA), seule les méthodes d'optimisation convexe utilisant des fonctions discrètes sont développées dans les sections suivantes. Ainsi dans la troisième section l'algorithme du sous-gradient est présenté, tandis que la quatrième section décrit l'algorithme proposé par Shoham et Gersho ([Shoh 88]) dont des adaptations seront proposées dans les chapitres subséquents.

2.2 Expression des Problèmes d'Optimisation

L'optimisation convexe est une méthode mathématique très puissante puisqu'elle permet de résoudre des problèmes complexes de manière rapide et efficace. En effet, les principes sur lesquels repose l'optimisation convexe sont très simples : toute fonction convexe admet un unique minimum (qui résout le problème en l'absence de contrainte[1]) et toute fonction convexe est de dérivée seconde positive (ce qui indique la direction de convergence quand on utilise un algorithme de gradient).

En outre, l'optimisation convexe est une méthode très largement utilisée dans le cadre de la résolution de problèmes d'ingénierie ([Boyd 04], [Bert 99], etc.) et recèle donc de nombreuses implémentations directement applicables. Toutefois, certains aspects de l'optimisation convexe appliquée à des problèmes discrets ne sont pas aussi célèbres qu'en ce qui concerne les problèmes continues et méritent d'être présentés dans le cadre de cette thèse.

2.2.1 Formulation du Problème Primal dans le Cas Continu

La littérature scientifique traitant de l'optimisation convexe ([Boyd 04], [Bert 99]) décrit les problèmes d'optimisation d'une manière précise que nous allons décrire ci-dessous. Il est important de préciser qu'il s'agit ici d'un travail bibliographique primordial à la bonne compréhension des travaux présentés par la suite.

Soit $\mathbf{x} \in \mathcal{R}^n$ le vecteur à optimiser et f_0 le critère suivant lequel le vecteur est minimisé. La fonction objectif s'exprime sous la forme $f_0 : \mathcal{R}^n \; to \mathcal{R}$. Les contraintes d'inégalité s'écrivent $f_i : \mathcal{R}^n \rightarrow \mathcal{R}$ tandis que les contraintes d'égalité sont notées $h_i : \mathcal{R}^n \rightarrow \mathcal{R}$.

Pour être un problème d'optimisation convexe, le problème doit également être tel que la fonction f_0 est convexe, les fonctions f_i sont convexes et les fonctions h_i sont linéaires.

Ainsi, le problème d'optimisation défini ci-dessus s'écrit de la manière suivante :

$$\begin{array}{rll} \mathbf{x}^* = & \arg\min_{\mathbf{x}} & f_0(\mathbf{x}) \\ & \text{contraintes} & f_i(\mathbf{x}) \leq 0, \quad i = 1 \ldots m \\ & & h_i(\mathbf{x}) = 0, \quad i = 1 \ldots p \end{array} \tag{2.1}$$

L'optimisation est dite continue lorsque les fonctions utilisées (f_0, f_i, h_i) et le vecteur à optimiser sont continus.

Revenons au problème (2.1) et notons que l'ensemble des solutions possibles est délimité par les contraintes d'égalité et d'inégalité. Mathématiquement parlant, l'ensemble des solutions possibles est l'intersection entre l'ensemble des contraintes d'inégalité

[1] en effet, tout minimum local est également global

et de l'ensemble des contraintes d'égalité et s'intitule domaine d'optimisation $\mathcal{D}$:

$$\mathcal{D} = \bigcap_{i=0}^{m} \mathbf{dom} f_i \quad \cap \quad \bigcap_{i=0}^{p} \mathbf{dom} h_i \tag{2.2}$$

Notons que toute solution du problème d'optimisation appartient à cet ensemble $\mathcal{D}$. Observons de plus que le domaine d'optimisation est convexe d'après la convexité de chaque fonction f_i et la linéarité de chaque fonction h_i. Par conséquent tout problème d'optimisation convexe consiste à minimiser une fonction objectif convexe sur un ensemble convexe.

De nombreux problèmes d'optimisation appartiennent à la théorie de l'optimisation convexe. Nous pouvons citer par exemple l'optimisation linéaire, l'optimisation quadratique ou la programmation géométrique. Le lecteur avide d'explications complémentaires pourra se référer aux travaux de Boyd et Vandenberghe [Boyd 04] ou de Bertsekas [Bert 99].

2.2.2 Formulation du Problème Primal dans le Cas Discret

La formulation d'un problème d'optimisation dans le cas discret est très analogue à celle proposée dans le cas continu. Toutefois, certaines différences seront détaillées dans cette partie.

Soit $\mathbf{x}$ le vecteur à optimiser. Soit k, l'indice associé à la kème valeur de $\mathbf{x}$ et n le nombre de paramètres de x. Chaque paramètre $\mathbf{x_k}$ du vecteur $\mathbf{x}$ peut prendre des valeur parmi un ensemble de valeurs discrètes $\mathcal{E}_k$. Le produit cardinal de ces ensembles forme $\mathcal{E}$ l'ensemble des valeurs possibles pour le vecteur $\mathbf{x}$. Cet ensemble est de dimension n. Soit $\breve{f}_0$ la fonction objectif transformant le vecteur $\mathbf{x}$ en une valeur $\breve{f}_0(\mathbf{x})$. Bien qu'il soit possible de définir des ensembles discrets pour la fonction objectif, pour les contraintes d'égalité ainsi que pour les contraintes d'inégalité, la fonction $\breve{f}_0$ sera supposée à valeurs dans $\mathcal{R}$.

La fonction objectif est donc telle que $\breve{f}_0 : \mathcal{E} \rightarrow \mathcal{R}$. Les contraintes d'inégalité s'écrivent $\breve{f}_i : \mathcal{E} \rightarrow \mathcal{R}$ tandis que les contraintes d'égalité sont notées $\breve{h}_i : \mathcal{E} \rightarrow \mathcal{R}$.

Pour être un problème d'optimisation convexe, le problème doit également être tel que, pour l'ensemble des vecteurs disponibles, la fonction $\breve{f}_0$ est convexe, les fonctions $\breve{f}_i$ sont convexes et les fonctions $\breve{h}_i$ sont linéaires.

Ainsi, le problème d'optimisation défini ci-dessus s'écrit de la manière suivante :

$$\begin{array}{rlll} \mathbf{x}^* = & \arg\min_{\mathbf{x}} & \breve{f}_0(\mathbf{x}) & \\ & \text{contraintes} & \breve{f}_i(\mathbf{x}) \leq 0, & i = 1 \ldots m \\ & & \breve{h}_i(\mathbf{x}) = 0, & i = 1 \ldots p \end{array} \tag{2.3}$$

Cette forme est quasiment identique à celle définie dans le cas continu. Pour obtenir

aisément un problème d'optimisation discrète, il est possible d'échantillonner les fonctions continues définies à l'équation (2.3). Ainsi, nous nous contenterons dans la partie suivante de formuler le problème dual pour les fonctions continues. En effet, le problème dual peut être formulé de manière identique, dans le cas discret, en remplaçant les fonctions continues par leur fonctions discrètes associées.

2.2.3 Formulation du Problème Dual

Résoudre un problème d'optimisation convexe sous contrainte sous sa forme primale peut se révéler peu approprié. En effet, lorsque la solution du problème non contraint n'est pas faisable, il est nécessaire de recommencer la recherche en sachant qu'au moins une contrainte est atteinte. Cette recherche peut nécessiter de nombreuses itérations et se révéler compliqué puisqu'il peut exister de nombreuses contraintes d'égalité ou d'inégnalité(se référer à (2.1)). Dans ce cas, la transformation du problème primal en problème dual permet généralement d'obtenir des contraintes plus simples (multiplicateurs de Lagrange positifs). En outre, la transformation du problème primal en un problème dual est souvent appropriée à l'établissement de solutions distribuées car elle permet fréquemment de séparer le problème en sous-problèmes pouvant être résolu de manière indépendante.

Pour écrire le problème dual, il est nécessaire d'introduire des multiplicateurs de Lagrange $\lambda = [\lambda_0 \dots \lambda_m]$, qui correspondent aux contraintes d'inégalité et $\mu = [\mu_0 \dots \mu_p]$, qui correspondent aux contraintes d'égalité. De plus, le problème dual requiert la définition préalable du lagrangien exprimé ci-dessous :

$$\mathcal{L}(\mathbf{x}, \lambda, \mu) = f_0(\mathbf{x}) + \sum_{i=0}^{m} \lambda_i f_i(\mathbf{x}) + \sum_{i=0}^{p} \mu_i h_i(\mathbf{x}) \tag{2.4}$$

En outre, de ce lagrangien il est possible de définir la fonction duale de lagrange $q(\lambda, \mu)$, telle que :

$$q(\lambda, \mu) = \inf_{\mathbf{x}} \ \mathcal{L}(\mathbf{x}, \lambda, \mu) \tag{2.5}$$

Finalement, le problème dual s'écrit :

$$\begin{aligned} (\lambda^*, \mu^*) = \quad & \arg\max_{(\lambda,\mu)} \quad q(\lambda, \mu) \\ & \text{contraintes} \quad \lambda \geq 0 \end{aligned} \tag{2.6}$$

Remarquons que la continuité des multiplicateurs de Lagrange, implique la continuité du problème dual, que le problème initial soit continu ou discret. De plus, d'après [Boyd 04], résoudre le problème primal et le problème dual est équivalent dans le cas d'un problème convexe continu. Dans le cas d'un problème discret, l'utilisation d'un algorithme d'Optimisation Discrète et Convexe (ODC) assure la convergence de l'algorithme sur l'enveloppe convexe [Shoh 88].

Toutefois, la résolution du problème dual dans le cas discret ne permet pas nécessairement d'obtenir l'optimum global. En effet, il est possible qu'une solution à valeurs continues soit meilleure que la solution obtenue en résolvant le problème discret. La Figure 2.1 représente un problème discret pour lequel il existe un écart de dualité. Il est intéressant de noter que la solution optimale du problème discret peut être obtenue en effectuant une recherche exhaustive (pour des valeurs continues) dans la zone délimitée par la meilleure solution appartenant à l'enveloppe convexe (qui respecte donc les contraintes) et la meilleure configuration de l'enveloppe convexe ne respectant pas les contraintes. En outre, la différence entre la solution optimale et la solution optimale sur l'enveloppe convexe peut être choisie aussi faible que souhaitée en sélectionnant des valeurs de x suffisamment échantillonnées (puisque la zone d'optimalité diminue proportionnellement au pas d'échantillonnage entre deux valeurs de x).

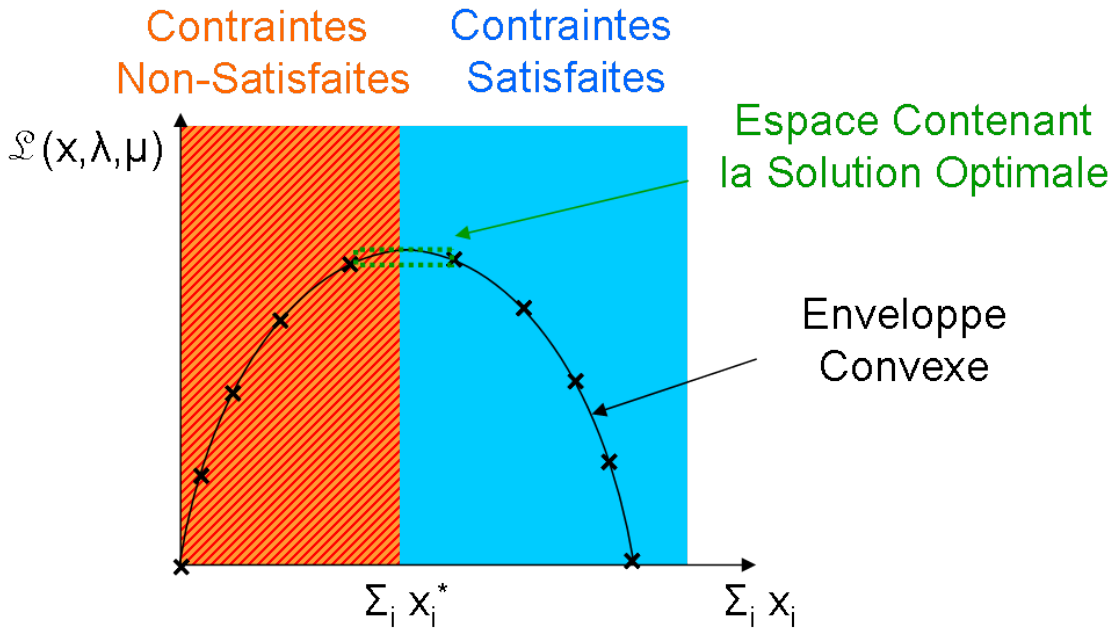

Figure 2.1: Résolution d'un problème d'optimisation.

2.3 Algorithmes d'Optimisation Convexe

Cette section vise à présenter brièvement les algorithmes convexes d'optimisation existant. Il ne s'agit pas ici de traiter de l'ensemble des algorithmes mais plutôt de donner un aperçu des méthodes disponibles et de laisser au lecteur le soin de se les approprier si ses recherches l'y invitent. Seules les algorithmes effectivement utilisés dans le cadre de nos travaux, à savoir l'algorithme du sous-gradient et l'algorithme de Shoham-Gersho (proposé dans l'article [Shoh 88]) seront présentés plus précisément dans cette section

2.3.1 Algorithmes Disponibles dans le Cas Continu

Le tableau qui suit décrit les algorithmes d'optimisation généralement utilisés dans le cadre de l'optimisation de fonctions convexes et continues. Ces méthodes sont très répandues et de nombreuses références proposent des algorithmes permettant de les implémenter. Les livres de Boyd [Boyd 04] et Bertsekas [Bert 99] sont particulièrement utiles car il permettent au lecteur d'aborder le domaine de l'optimisation dans sa globalité.

	Problème Contraint	**Type de Méthode**	**Références**
Méthode de descente de gradient	Non	Primale	[Boyd 04], [Bert 99]
Méthode de projection du gradient	Oui	Primale	[Boyd 04], [Bert 99]
Méthode des barrières et du point intérieur	Oui	Duale	[Boyd 04], [Bert 99]
Méthode des pénalités	Oui	Duale	[Boyd 04], [Bert 99]
Méthode de l'ascension duale	Oui	Duale	[Bert 99]
Méthode primale/duale du point intérieur	Oui	Primale et Duale	[Boyd 04], [Bert 99]

Table 2.1: Algorithmes d'Optimisation Convexe et Continue

2.3.2 Algorithmes Disponibles dans le Cas Discret

Le tableau ci-dessous répertorie quelques méthodes utilisées dans le cadre de l'optimisation de fonctions discrètes et convexes. Ce type de méthode est peu utilisé dans le domaine de l'optimisation car de nombreux outils ont été développés dans le cadre de l'optimisation continue.

	Problème Contraint	**Type de Méthode**	**Références**
Sous-gradient	Oui	Duale	[Boyd 04], [Bert 99]
Intersection de plans	Oui	Duale	[Boyd 04], [Bert 99]
Algorithme de Shoham-Gersho	Non	Primale	[Shoh 88]

Table 2.2: Algorithmes d'Optimisation Convexe et Discrète

2.3.3 Algorithme du Sous-Gradient

Afin de faciliter la compréhension de l'algorithme, celui-ci sera développé dans le contexte d'un problème d'optimisation ne comprenant pas de contrainte d'égalité. Il est néanmoins tout à fait possible d'étendre l'algorithme décrit à des problèmes plus complexes ainsi que nous le verrons dans le Chapitre 4. L'objectif est donc de résoudre :

$$\begin{aligned} \mathbf{x}^* = \quad & \underset{\mathbf{x}}{\arg\min} \quad f_0(\mathbf{x}) \\ & \text{contraintes} \quad f_i(\mathbf{x}) \leq 0, \quad i = 1 \dots m \end{aligned} \tag{2.7}$$

Puisque le problème étudié est convexe, résoudre le problème primal de l'équation (2.7) est équivalent à résoudre le problème dual dont le Lagrangien s'écrit :

$$\mathcal{L}(\mathbf{x}, \lambda, \mu) = f_0(\mathbf{x}) + \sum_{i=0}^{m} \lambda_i f_i(\mathbf{x}) \tag{2.8}$$

La solution de ce problème est à la fois minimale suivant le vecteur des ressources $\mathbf{x}$ et minimal suivant le vecteur des paramètres λ. Pour obtenir cette solution, commençons par écrire la fonction duale de lagrange $q(\lambda)$:

$$q(\lambda) = \inf_{\mathbf{x}} \ \mathcal{L}(\mathbf{x}, \lambda) \tag{2.9}$$

Pour résoudre le problème dual, il s'agit alors de maximiser la fonction duale vis-à-vis du vecteur des paramètres λ :

$$\begin{aligned} \lambda^* = \quad & \underset{(\lambda)}{\arg\max} \quad q(\lambda) \\ & \text{contraintes} \quad \lambda \geq 0 \end{aligned} \tag{2.10}$$

Comme la solution du problème d'optimisation s'écrit sous la forme d'un couple $(\mathbf{x}^*, \lambda^*)$, la solution du problème d'optimisation dépend des deux variables et optimiser successivement sur $\mathbf{x}$ puis sur λ, n'assure pas nécessairement l'obtention de la solution optimale. La méthode des sous-gradients offre un moyen de garantir cette convergence.

Avant de décrire la méthode évoquée, rappelons que le sous-gradient $\mathbf{g}$ d'une fonction convexe f au point x_0 s'exprime de la manière suivante :

$$f(x) - f(x_0) \leq \mathbf{g}.(x - x_0) \tag{2.11}$$

La méthode mentionnée repose sur le fait que le paramètre λ est incrémenté suivant la direction du sous-gradient. En outre, la méthode permet d'assurer que le paramètre λ s'approche de sa valeur optimale à chaque itération, ce qui assure la convergence de l'algorithme.

Pour ce faire, la solution retenue consiste à incrémenter la valeur du vecteur de paramètre λ en fonction du sous-gradient calculé sur chaque réseau. La valeur obtenue est ensuite projetée sur l'ensemble de définition du vecteur de paramètre λ (la projection est notée $[\cdot]^+$). L'incrémentation est effectuée à l'aide d'un pas d'itération ϵ qui doit être

suffisamment faible pour assurer la convergence. Ainsi, à chaque itération, le nouveau vecteur de paramètre est obtenu en effectuant l'opération suivante :

$$\lambda^{\mathbf{k+1}} = [\lambda^{\mathbf{k}} + \epsilon^k \mathbf{g}^{\mathbf{k}}]^+ \tag{2.12}$$

où $\lambda^{\mathbf{k}}$ est la valeur du vecteur de paramètres à l'itération k, ϵ^k représente la valeur du pas d'itération à l'itération k et $\mathbf{g}^{\mathbf{k}}$ dénote la valeur du sous-gradient à l'itération k.

Cette méthode converge sous réserve que le paramètre λ s'approche de sa valeur optimale à chaque itération :

$$\|\lambda^{\mathbf{k+1}} - \lambda^*\| < \|\lambda^{\mathbf{k}} - \lambda^*\| \tag{2.13}$$

Il est possible de montrer [Bert 99] (Chapitre 6, Section 3) que la condition de l'équation précédente est équivalente à la condition suivante sur le pas d'itération :

$$0 < \epsilon^k < \frac{2(q(\lambda^*) - q(\lambda^{\mathbf{k}}))}{\|\lambda^{\mathbf{k}}\|^2} \tag{2.14}$$

Ainsi, l'algorithme du sous-gradient converge, si le pas d'itération est suffisamment petit.

2.3.4 Algorithme de Shoham et Gersho

L'algorithme de Shoham et Gersho [Shoh 88] a été développé dans le cadre du codage source de l'information et permet d'allouer efficacement les bits à des échantilloneurs. L'efficacité prend deux formes : 1- la minimisation de la distorsion due au codage et 2- la complexité calculatoire de l'algorithme permettant d'effectuer l'allocation.

L'algorithme proposé dans [Shoh 88] repose sur l'utilisation de fonctions de quantification convexes définies pour des valeurs discrètes et propres à chaque échantillonneur. Il est intéressant de remarquer que les fonctions de quantification établissent le lien entre le débit et le nombre de bits alloués. Rappelons ici le fonctionnement de l'algorithme proposé dans [Shoh 88].

Soit S l'ensemble des vecteurs d'allocations possibles. Soit B un sous-ensemble de S. Soit $H(\cdot)$ la fonction objectif définie pour tout vecteur de B et à valeurs réelles. Soit $R(\cdot)$ une fonction de contrainte définie pour tout vecteur de B et à valeurs réelles et R_c la valeur maximale de la contrainte. L'algorithme vise à résoudre le problème suivant :

$$\begin{array}{ll} \min_{B \in S} & H(B) \\ \text{contraintes} & R(B) \leq R_c \end{array} \tag{2.15}$$

La résolution de ce problème s'effectue à l'aide du Lagrangien suivant :

$$\mathcal{L}(B, \lambda) = H(B) + \lambda R(B) \tag{2.16}$$

En effet, un théorème démontré dans [Shoh 88] assure que la résolution du problème non contraint à la valeur optimale du multiplicateur de Lagrange donne la solution du problème intial (2.15), sur l'enveloppe convexe. L'algorithme proposé consiste donc à rechercher la valeur optimale du multiplicateur de Lagrange puis à en déduire l'allocation associée (la solution du problème primal). Pour ce faire, l'algorithme fait varier le multiplicateur de Lagrange de telle sorte que la solution obtenue s'approche itérativement de la contrainte. Dans cette optique, l'algorithme accroît le multiplicateur de Lagrange de telle sorte que la nouvelle solution appartienne à l'enveloppe convexe et qu'elle soit la plus proche possible de la solution actuelle (plus petite augmentation du multiplicateur de Lagrange). Ce processus est répété, jusqu'à ce que la solution atteigne la contrainte. La solution du problème (2.15), appartenant à l'enveloppe convexe, est alors obtenue.

Voici l'algorithme expliqué en détails (directement issu de [Shoh 88]) :

1. Obtenir une valeur initiale pour le multiplicateur de Lagrange λ (Initialisation à 0 ou utilisation de méthodes plus évoluées proposées dans [Shoh 88])

2. Résoudre le problème d'optimisation non contraint en utilisant sa séparabilité (équations 7 à 9 de [Shoh 88]) Si la valeur de λ n'est pas singulière, il n'y a qu'une solution et une contrainte $R^*(\lambda)$. Si λ est une valeur singulière, alors il y a au moins deux solutions. Dans ce cas, il y a lieu de sélectionner la solution dont la contrainte est la plus grande $R_g^*(\lambda)$ et celle dont la contrainte est la plus petite $R_p^*(\lambda)$.

3. Si la contrainte recherchée est telle que :
   $R_p^*(\lambda) \leq R_c \leq R_g^*(\lambda)$
   alors il s'agit de rechercher la solution dont la contrainte R_a^* est la plus proche de R_c par valeur inférieure. Si $R_a^* = R_c$, une solution optimale a été trouvée. Sinon, une solution approchante a été trouvée. L'algorithme prend alors fin.

4. Si $R_c < R*(\lambda)$ (ou $R_c < R_p^*(\lambda)$) alors il y a lieu de chercher la nouvelle valeur de λ en utilisant l'équation qui s'approche de λ^* par valeur inférieure.

5. Si $R_c > R*(\lambda)$ (ou $R_c > R_g^*(\lambda)$) alors il y a lieu de chercher la nouvelle valeur de λ en utilisant l'équation qui s'approche de λ^* par valeur supérieure.

Cet algorithme sera adapté dans les chapitres suivants à nos problèmes d'allocation de ressources. En effet, un parallèle peut être effectué entre le codage de l'information (à l'aide de relations convexes entre le débit et le nombre de bits utilisé) et l'allocation de puissance (qui utilise des relations convexes entre la puissance et le débit).

Chapter 3

Position du Problème

3.1 Résumé du Chapitre

Dans ce chapitre, nous décrivons mathématiquement le problème considéré dans nos travaux. Afin de faciliter la compréhension des travaux par lecteur, nous commençons par la description des hypothèses d'étude. Puis, nous poursuivons ce chapitre par la formulation du problème d'optimisation générique étudié et montrons que ce problème est un problème d'optimisation convexe et discrète sous contraintes.

La première section de ce chapitre met en évidence les paramètres techniques retenus dans le cadre de notre étude. Ainsi de nombreuses notions utilisées dans la suite de cette thèse sont définies. En effet, le caractère hétérogène du système est explicité ainsi que les hypothèses retenues (telles les informations connues par chaque entité du système) et les contraintes liées à l'architecture hétérogène.

La seconde section développée explicite le problème générique étudié au cours de nos travaux sous la forme d'un problème d'optimisation convexe. Elle permet d'exprimer le problème sous sa forme primale, puis sous sa forme duale et montre l'équivalence entre les deux expressions.

L'introduction nous a permis de décrire le contexte général des travaux de recherche effectués. Ceux-ci visent à résoudre le problème de l'allocation de ressource dans des réseaux mobiles hétérogènes lorsque les ressources à disposition sont de nature discrète. Dans le chapitre que nous initions, l'objectif entrepris est d'exprimer mathématiquement le problème de l'allocation de ressource effectuée conjointement dans des réseaux hétérogènes. En outre, l'objectif de ce chapitre consiste à démontrer que le problème étudié s'inscrit bien dans le cadre de l'optimisation convexe et discrète.

3.2 Spécificités techniques du problème générique étudié

L'introduction a mis en exergue l'aspect novateur des recherches présentées dans cette thèse à savoir l'allocation de ressources *discrètes* dans un système *hétérogène* à l'aide d'une prise de décision *distribuée*. Afin de limiter l'étendue de l'étude, seule l'allocation descendante a été étudiée mais une extension aux allocations sur le lien ascendant se révèle pleinement envisageable.

Le positionnement du problème induit certaines difficultés dans l'élaboration de la solution distribuée. En effet, pour bénéficier pleinement de l'hétérogénéité du système, les mobiles doivent avoir la possibilité de communiquer *simultanément* sur plusieurs réseaux. Ainsi, les mobiles recherchent l'allocation de ressources optimale sur l'ensemble des réseaux, ce qui nécessite une coordination efficace des prises de décisions dans le système hétérogène.

De plus, comme toute allocation de ressources, le problème étudié ici cherche à allouer les ressources suivant une approche satisfaisante à la fois du point de vue des utilisateurs (caratérisés par les mobiles) et du point de vue des opérateurs (caractérisés par les différents réseaux). Deux approches sont généralement proposées dans le cadre de l'allocation de ressources dans un réseau monosystème : la maximisation du débit dans l'ensemble du système (Rate Adaptive approach) et la minimisation de la puissance sous contrainte de débit minimal pour chaque utilisateur (Margin Adaptive approach). La seconde approche a été étendue au cadre des réseaux multisystèmes et a ensuite été utilisée dans cette thèse. Cette approche assure une équité entre les utilisateurs, limite la puissance utilisée par les stations de base et permet de provisionner des ressources pour de nouveaux utilisateurs.

L'extension de l'approche visant à minimiser la puissance provient de la limite de puissance existant à chaque station de base. En effet, cette contrainte n'est pas explicitement prise en compte dans le cadre d'une allocation monosystème puisqu'une allocation n'est faisable que si la puissance demandée est inférieure à la puissance disponible dans le réseau étudié. Ainsi, comme seules des allocations faisables sont considérées, il n'est nul besoin d'expliciter la contrainte de puissance. Par contre, dans le cadre des systèmes hétérogènes, si un utilisateur n'obtient pas satifaction dans un réseau, il peut chercher à obtenir des ressources provenant d'un autre réseau. Par conséquent, la contrainte de puissance qui est généralement implicite pour les systèmes monotechnologiques devient explicite dans le cadre des systèmes multitechnologiques.

Une autre contrainte technique provenant du caractère distribué de l'allocation se

révèle être la connaissance partielle du contexte radio par chaque entité du système. En effet, pour éviter des signalisations coûteuses en terme de complexité, chaque utilisateur ne connaît que la qualité des canaux associés aux liens qu'il maintient sur chaque technologie. De manière analogue, chaque station de base ne connaît que la qualité des canaux associés aux utilisateurs avec lesquels elle maintient un lien de communication.

Par simplification, nous supposerons par la suite que nous effectuons l'allocation de ressources dans un système ou chaque réseau est constitué d'une station de base unique. De plus, dans le travail effectué, chaque utilisateur maintient un lien avec l'ensemble des réseaux (avec chaque station de base d'après l'hypothèse précédente).

3.3 Formulation mathématique du problème générique

Considérons un système hétérogène composé de stations de base n'interférant pas entre elles (chaque station de base fonctionne sur une fréquence spécifique). Ces stations sont référencées par les indices $j = [1 \dots N_s]$ et disposent d'une puissance maximale $P_{max,j}$. Supposons également que les mobiles ont pour indice $k = [1 \dots N_u]$ et cherchent à obtenir un débit minimal descendant $R_{min,k}$. Rappelons que les utilisateurs peuvent obtenir simultanément des ressources de plusieurs sources et les reconstituer. Le lien entre le mobile k et la station j est représenté par (k, j).

De plus, les stations de base utilisent un système de MCA et peuvent donc allouer aux mobiles une configuration physique combinant simultanément une modulation (par exemple BPSK, QPSK, etc.) et un codage (Code Correcteur d'Erreur). Cette configuration correspond à un débit donné et est associée à une puissance qui dépend de la qualité du lien de transmission via une relation convexe. Ainsi, les mobiles k peuvent communiquer avec les stations de base j uniquement en utilisant une valeur discrète de débit associée à un "mode de transmission". Ce mode est référencé par l'indice $i_{k,j} = [1 \dots I_{k,j}]$.

Dans l'environnement décrit, l'objectif recherché est de minimiser la puissance utilisée dans l'ensemble des réseaux, sous réserve d'assurer le débit minimum requis par les mobiles. Au cours de l'optimisation, les variables utilisées sur le lien (k, j) seront notées $R_{k,j}(i_{k,j})$ en ce qui concerne le débit, $g_{k,j}$ en ce qui concerne la qualité du canal de transmission et $P_{k,j}(g_{k,j}, i_{k,j})$ en ce qui concerne la puissance. La relation entre le débit et la puissance est concave et donnée par $R_{k,j} = B \log(1 + SINR_{k,j}) = B \log(1 + \alpha_{k,j} P_{k,j})$, où $SINR_{k,j}$ correspond au Rapport Signal à Interférence plus Bruit (RSIB en français et SINR en anglais) sur le lien (k, j) et où $\alpha_{k,j}$ permet d'exprimer le SINR en fonction de la puissance (dénominateur supposé constant). On peut également écrire $R_{k,j} = f_{k,j}(P_{k,j})$ où $f_{k,j}$ est une relation concave. Par conséquent, le problème étudié vise à rechercher la matrice optimale des modes de transmission $\mathbf{I}^* = \{i^*_{k,j}\}$, telle que :

$$\mathbf{I}^* = \underset{\substack{\{i_{k,j}\} \\ i_{k,j}\in[1...I_{k,j}]}}{\arg\min} \sum_{k=1}^{N_u}\sum_{j=1}^{N_s} P_{k,j}(g_{k,j}, i_{k,j})$$

$$\text{contraintes} \quad \forall k, \quad \sum_{j=1}^{N_s} R_{k,j}(i_{k,j}) \geq R_{min,k}$$

$$\forall j, \quad \sum_{k=1}^{N_u} P_{k,j}(g_{k,j}, i_{k,j}) \leq P_{max,j} \tag{3.1}$$

Afin d'améliorer la lisibilité du problème, la dépendance vis-à-vis du canal de transmission $g_{k,j}$ sera omise dans l'expression de $P_{k,j}$ par la suite (tout en étant considérée dans le cadre des simulations).

3.3.1 Vérification de la Convexité

Formulation du problème primal

Afin de bien mettre en exergue le caractère convexe de l'optimisation effectuée, transformons le problème décrit par l'équation (3.1) de manière à le faire correspondre à la formulation proposée par Boyd [Boyd 04]. Pour cela, introduisons la fonction convexe $f_{k,j}^{-1}$ qui est la fonction réciproque de $f_{k,j}$.

$$\mathbf{R}^* = \underset{\substack{\{R_{k,j}\}_{k=1..N_u} \\ j=1..N_s}}{\arg\min} \sum_{k=1}^{N_u}\sum_{j=1}^{N_s} f_{k,j}^{-1}(R_{k,j})$$

$$\text{contraintes} \quad \forall k, \quad R_{min,k} - \sum_{j=1}^{N_s} R_{k,j} \leq 0$$

$$\forall j, \quad \sum_{k=1}^{N_u} f_{k,j}^{-1}(R_{k,j}) - P_{max,j} \leq 0 \tag{3.2}$$

A ce stade, il est nécessaire d'identifier les différents éléments constituant le problème d'optimisation. Voici une description des composantes du problème :

- $f_0(\mathbf{R}) = \sum_{k=1}^{N_u}\sum_{j=1}^{N_s} f_{k,j}^{-1}(R_{k,j})$ correspond au critère à optimiser.
- $u_k(\mathbf{R}) = R_{min,k} - \sum_{j=1}^{N_s} R_{k,j}$ est la condition de débit à respecter pour chaque mobile.

- $v_j(\mathbf{R}) = \sum_{k=1}^{N_u} f_{k,j}^{-1}(R_{k,j}) - P_{max,j}$ est la condition de puissance à respecter pour chaque station de base.

Ainsi le problème d'optimisation s'exprime sous la forme suivante (forme primale) :

$$\begin{aligned} \mathbf{R}^* = \quad & \underset{\substack{\{R_{k,j}\}_{k=1..N_u} \\ j=1..N_s}}{\arg\min} \quad f_0(\mathbf{R}) \\ & \text{contraintes} \quad \forall k, \quad u_k(\mathbf{R}) \leq 0 \\ & \qquad\qquad\quad \forall j, \quad v_j(\mathbf{R}) \leq 0 \end{aligned} \tag{3.3}$$

Ce problème est bien un problème d'optimisation convexe car la fonction f_0 est convexe, les fonctions u_k sont linéaires et les fonctions v_j sont convexes.

Ce problème d'optimisation peut aisément être résolu de manière centralisée pour des fonctions continues. En effet, l'utilisation de méthodes de points intérieurs (méthode des barrières ou du point interieur primal-dual) permet de résoudre ce problème d'optimisation convexe [Boyd 04]. Par contre, si le problème d'optimisation est discret, l'utilisation d'autres algorithmes est nécessaire. Ainsi, l'algorithme du sous-gradient peut être sélectionné. En plus d'offrir la possibilité de résoudre le problème d'optimisation pour le cas discret, cet algorithme présente l'avantage d'admettre une implémentation centralisée ou distribuée [Bert 99].

Utiliser une approche centralisée nécessite de communiquer à une entité dédiée l'ensemble des paramètres du problème, c'est-à-dire les paramètres de chaque connexion entre les utilisateurs et les stations de base. Une telle architecture se révèle à la fois très coûteuse en terme de signalisation, difficile à mettre en oeuvre en termes de synchronisation et requiert une puissance de calcul élevée.

Par conséquent, il est primordial de réfléchir à des solutions distribuées permettant de résoudre le problème d'optimisation étudié. La transformation du problème primal en un problème dual à l'aide de multiplicateurs de Lagrange est une méthode très utilisée dans la littérature scientifique pour décomposer les problèmes de ce type et cette méthode sera donc étudiée ci-dessous.

Expression du problème dual

Résoudre directement un problème d'optimisation sous contrainte n'est pas une tâche aisée. En effet, la résolution du problème primal se révèle souvent périlleuse comme nous le verrons dans la description de l'algorithme proposé ci-dessous. Le problème primal peut être résolu en divisant la résolution du problème en deux phases. Dans une première phase, le problème primal est résolu sans considérer les contraintes, puis, dans une seconde phase, on vérifie si la solution potentielle est réalisable. Si tel est le cas la solution optimale est obtenue (et les contraintes sont inutiles), sinon il y a lieu de

chercher la solution optimale en considérant qu'au moins une des contraintes est atteinte (en utilisant l'alogorithme du simplex par exemple).

Par conséquant, l'utilisation d'une méthode permettant de relâcher les contraintes peut faciliter la résolution du problème d'optimisation sous contrainte. En effet, cette technique permet de s'affranchir de la vérification des contraintes et de se concentrer sur l'optimisation du critère recherché. En outre, l'écriture du problème dual permet souvent de décomposer le problème global en sous-problèmes indépendants. C'est dans cette optique que la relaxation lagrangienne a été utilisée dans nos travaux.

La relaxation lagrangienne consiste à ajouter la contrainte à la fonction objectif, a près pondération par un multiplicateur de Lagrange, en s'assurant que sa contribution au critère est nulle. En effet, afin d'assurer que le même problème d'optimisation est résolu, il y a lieu de garantir que soit la contrainte est atteinte, soit le multiplicateur associé est nul, mais dans tous les cas la contrainte pondérée ne doit pas modifier le critère recherché. L'utilisation de la théorie de Lagrange, dans le contexte de l'optimisation convexe est très efficace car il n'existe pas d'écart de dualité ("duality gap") entre la solution du primal et celle du dual [Boyd 04].

Ecrivons ci-dessous le problème dual associé au problème (3.1):

$$\mathcal{L}(\mathbf{P},\lambda,\mu)=\sum_{k=1}^{N_u}\sum_{j=1}^{N_s}P_{k,j}+\sum_{k=1}^{N_u}\lambda_k(R_{min,k}-\sum_{j=1}^{N_s}f_{k,j}(P_{k,j}))+\sum_{j=1}^{N_s}\mu_j(\sum_{k=1}^{N_u}P_{k,j}-P_{max,j}) \tag{3.4}$$

Ce problème peut également être écrit de la manière suivante, afin de faciliter la comparaison avec le problème primal rédigé en (3.2):

$$\mathcal{L}(\mathbf{R},\lambda,\mu)=\sum_{k=1}^{N_u}\sum_{j=1}^{N_s}f_{k,j}^{-1}(R_{k,j})+\sum_{k=1}^{N_u}\lambda_k(R_{min,k}-\sum_{j=1}^{N_s}R_{k,j}+\sum_{j=1}^{N_s}\mu_j(\sum_{k=1}^{N_u}f_{k,j}^{-1}(R_{k,j})-P_{max,j}) \tag{3.5}$$

En tout état de cause, la solution du problème dual est obtenue en résolvant :

$$\mathcal{L}(\mathbf{R}^*,\lambda^*,\mu^*)=\max_{\lambda,\mu}\min_{\mathbf{R}}\mathcal{L}(\mathbf{R},\lambda,\mu) \tag{3.6}$$

Ce chapitre a donc permis de démontrer que le problème étudié était un problème d'optimisation convexe et discrète sous contraintes. Ce problème sera résolu dans les chapitres suivants, en utilisant les résultats issus de la théorie de l'optimisation présentés au chapitre précédent.

Chapter 4

Minimisation de la Puissance Instantanée (Contraintes de Débit Instantané)

4.1 Résumé du Chapitre

Dans ce chapitre, nous cherchons à résoudre le problème de l'allocation de ressources discrètes de manière distribuée avec pour objectif de minimiser la puissance totale utilisée dans le réseau tout en respectant un débit minimum pour chaque utilisateur.

Dans la première section, l'approche considérée vise à résoudre le problème d'optimisation de manière instantanée en prenant la décision optimale sur l'enveloppe convexe (cf. publication en Annexe 7.2). Cette approche nécessite un nombre d'itérations élevé ce qui explique qu'une approche quasi-optimale est proposée dans la seconde section. Cette approche permet d'obtenir une solution par projections successives sur le bord du domaine d'optimisation.

Dans la seconde section, l'approche proposée vise à rechercher une allocation quasi-optimale convergeant rapidement. La solution retenue permet de résoudre le problème de requêtes excessives des utilisateurs envers une certaine station de base. En effet, l'algorithme proposé projète les requêtes des utilisateurs sur l'ensemble des solutions possibles (calculées par les stations de base). La projection s'accompagne en outre du retrait de la station de base du processus d'allocation ce qui assure une convergence rapide (le nombre d'itérations est au plus égal au nombre de stations de base). Davantage d'explications sont fournies dans ce Chapitre et les résultats de simulation sont disponibles dans le papier inséré en Annexe 7.2.

4.2 Approche Optimale

Dans cette section nous cherchons à résoudre le problème énoncé au chapitre précédent (équation (3.1)) en recherchant à chaque instant l'allocation optimale sur l'enveloppe convexe. L'approche proposée s'appuie sur une décomposition du problème dual et une résolution du problème en deux parties: 1- des demandes de débit côté utilisateur, de manière à obtenir leur débit minimal, en fonction de leurs connaissances restreintes des conditions radios et 2- une recherche d'un meilleur multiplicateur de Lagrange caractérisant les demandes dans le réseau à l'aide de l'algorithme du sous-gradient. Ces opérations sont répétées jusqu'à l'obtention de la convergence.

Afin de faciliter la lecture, rappelons l'expression du problème étudié :

$$\begin{aligned}
\mathbf{I}^* = \underset{\substack{\{i_{k,j}\} \\ i_{k,j}\in[1...I_{k,j}]}}{\arg\min} & \sum_{k=1}^{N_u}\sum_{j=1}^{N_s} P_{k,j}(g_{k,j}, i_{k,j}) \\
\text{contraintes} \quad & \forall k, \quad \sum_{j=1}^{N_s} R_{k,j}(i_{k,j}) \geq R_{min,k} \\
& \forall j, \quad \sum_{k=1}^{N_u} P_{k,j}(g_{k,j}, i_{k,j}) \leq P_{max,j}
\end{aligned}$$

Comme énoncé précédemment, ce problème primal a été réécrit sous sa forme duale. En effet, la forme duale permet d'aboutir à une solution distribuée optimale sur l'enveloppe convexe qui sera détaillée ci-dessous. Le lagrangien permettant de résoudre le problème dual s'écrit à partir du problème primal en y ajoutant les contraintes pondérées par des multiplicateurs de Lagrange :

$$\begin{aligned}
\mathcal{L}(\mathbf{P}, \lambda, \mu) &= \sum_{k=1}^{N_u}\sum_{j=1}^{N_s} P_{k,j} \\
&+ \sum_{k=1}^{N_u} \lambda_k (R_{min,k} - \sum_{j=1}^{N_s} f_{k,j}(P_{k,j})) \\
&+ \sum_{j=1}^{N_s} \mu_j (\sum_{k=1}^{N_u} P_{k,j} - P_{max,j})
\end{aligned} \tag{4.1}$$

Dans (4.1), les vecteurs $\lambda = [\lambda_1 ... \lambda_{N_u}]$ et $\mu = [\mu_1 ... \mu_{N_s}]$ correspondent aux multiplicateurs de Lagrange. Il est utile de rappeler au lecteur que les fonctions $f_{k,j}$ sont concaves et définies uniquement pour des valeurs discrètes correspondant aux "modes de transmission". Par conséquent, le lagrangien défini à l'équation précédente est convexe.

Ainsi, résoudre le problème dual (c'est-à-dire minimiser le lagrangien de l'équation (4.1)) permet d'obtenir la solution optimale sur l'enveloppe convexe. Lorsque les modes de transmission sont choisis suffisamment proches, cette solution est excessivement proche de la solution du problème primal. En conclusion, l'allocation optimale sur l'enveloppe convexe est obtenue lorsque le lagrangien atteint la valeur suivante :

$$\mathcal{L}(\mathbf{P}^*, \lambda^*, \mu^*) = \min_{\mathbf{P}} \max_{\lambda,\mu} \mathcal{L}(\mathbf{P}, \lambda, \mu) \tag{4.2}$$

Pour obtenir le minimum global d'un problème d'allocation de ressources à valeurs continues, l'utilisation de solutions centralisées a été fréquemment utilisé dans la littérature scientifique [Boyd 04], [Yu 04] and [Rhee 00a]. Toutefois, les solutions centralisées ne sont pas adaptées aux systèmes hétérogènes puisqu'elles nécessitent de rassembler l'ensemble des éléments d'information à une entité de décision unique. En outre, l'entité évoquée calcule l'allocation optimale complète, ce qui présente l'inconvénient de requérir une capacité calculatoire très importante. Enfin, les décisions concernant l'allocation sont transmises aux différentes stations de base qui attribuent les ressources radios aux utilisateurs. Cette approche implique une signalisation élevée puisqu'il faut transmettre l'ensemble des paramètres à l'entité, puis l'ensemble des décisions aux stations de base.

A l'inverse, il est particulièrement judicieux d'utiliser la théorie de la décomposition (pour davantage de détails se référer à [Boyd 04]), de manière à décomposer le lagrangien précédent. Cela permet de diviser le problème d'optimisation complet relativement complexe en sous-problème aisément résolus. Lorsque la théorie de la décomposition est utilisée, des lagrangiens partiels sont introduits. Ceux-ci peuvent être résolus de manière indépendante par chaque station de base. Ainsi, l'utilisation d'optimisation successives effectuées alternativement par les utilisateurs et les stations de base permet d'obtenir l'allocation optimale caractérisée par les multiplicateurs de Lagrange optimaux (μ^*) et par les débits à allouer à chaque utilisateur k à l'instant considéré ($\mathbf{R_k^{(n)}}$).

Les lignes qui suivent seront dédiées à la résolution mathématique du problème d'optimisation défini précédemment. En particulier, les opérations effectuées alternativement par les mobiles et les stations de base seront décrites. Ainsi, le lagrangien introduit à l'équation (4.1) est réécrit de manière à permettre la séparation du problème global d'optimisation en sous-problèmes plus simples résolus de manière indépendante par les mobiles :

$$\mathcal{L}(\mathbf{P}, \lambda, \mu) = \sum_{k=1}^{N_u} \mathcal{L}_\mu(\mathbf{P_k}, \lambda_k) - \sum_{j=1}^{N_s} \mu_j P_{max,j}$$

avec :

$$\mathcal{L}_\mu(\mathbf{P_k}, \lambda_k) = \sum_{j=1}^{N_s} (1+\mu_j) P_{k,j} + \lambda_k (R_{min,k} - \sum_{j=1}^{N_s} f_{k,j}(P_{k,j})) \tag{4.3}$$

Trouver l'allocation optimale qui résout le problème (4.1) revient à minimiser, au niveau de chaque mobile, le lagrangien partiel $\mathcal{L}_{\mu^*}(\mathbf{P_k}, \lambda_k)$, le vecteur μ^* étant connu. Le lecteur notera que le vecteur μ^* correspond au poids des requêtes sur chaque réseau. Conséquemment aux éléments évoqués ci-dessus, la décomposition du problème global

sur l'ensemble des utilisateurs aboutit à des sous-problèmes explicites. En outre, la résolution du problème dual (minmisation de $\mathcal{L}_{\mu}(\mathbf{P_k}, \lambda_k)$) est équivalente à celle du primal suivant :

$$\begin{aligned} \mathbf{P_k} = \quad & \underset{\{P_{k,j}\}_{j=1..N_s}}{\arg\min} \quad \sum_{j=1}^{N_s}(1+\mu_j)P_{k,j} \\ & \text{subject to :} \quad \sum_{j=1}^{N_s} f_{k,j}(P_{k,j}) \geq R_{min,k} \end{aligned} \tag{4.4}$$

Ce problème est un problème d'optimisation convexe de ressources *discrètes*. En effet, les fonctions $f_{k,j}$ sont à la fois définies pour des valeurs discrètes (puisque les mobiles utilisent des technologies MCA). Le problème précédent peut être résolu en utilisant la programmation entière non linéaire (Non Linear Integer Programming) [Li 06]. Cependant, cette méthode présente un coût calculatoire élevé. Ainsi, la mise en oeuvre d'un algorithme d'optimisation discrète et convexe inspiré des travaux de Shoham et Gersho [Shoh 88] sera préconisée car cette méthode se révèle plus efficace. L'algorithme proposé par Shoham et Gersho l'a été dans le cadre du codage source de l'information et permet de réduire la complexité calculatoire d'une optimisation en restreignant la recherche de solution à un nombre très limité de configurations. En effet, l'algorithme se concentre sur les configurations situées sur l'enveloppe convexe et accroît ainsi considérablement la vitesse de convergence de l'optimisation.

Dans les paragraphes précédents, nous avons décrit la méthode permettant de résoudre les problèmes d'optimisation partiels au niveau de chaque mobile k lorsque la valeur du paramètre μ_j est connue. Pour obtenir la solution globale du problème d'optimisation, il est nécessaire de proposer un moyen permettant de calculer le vecteur optimal des poids associés aux réseaux μ^*. Pour atteindre cet objectif, la méthode des sous-gradients est utilisée. Cette méthode consiste à ajuster itérativement chaque paramètre μ_j en fonction des demandes des mobiles, de la qualité des canaux de transmissions sur les différents liens et de la quantité de puissance disponible dans chaque réseau.

Ainsi, pour obtenir la solution optimale sur l'enveloppe convexe du problème global d'allocation de ressources, la solution proposée ici consiste premièrement à obtenir les requêtes de chaque mobile k. Plus précisément, les requêtes sont obtenues en résolvant l'équation (4.4) en considérant $\mu_j^{(n)}$ constant. Secondement, le sous-gradient δ_j associé à l'ensemble des demandes sur le réseau j est calculé et permet de déduire le nouveau paramètre réseau $\mu_j^{(n+1)}$. Le processus décrit ci-dessus est répété jusqu'à ce que l'algorithme converge c'est-à-dire jusqu'à ce que tous les utilisateurs soient satisfaits. Il est bien connu que l'utilisation de la méthode des sous-gradients permet d'aboutir à la solution globale optimale pour des fonctions continues [Chia 05], [Chia 07], l'extension aux problèmes discret est proposée dans [Bert 99].

Un algorithme du sous-gradient permettant de résoudre le problème d'optimisation est présenté ci-dessous. La démonstration complète de la convergence de cet algorithme est détaillée à l'Appendice B de l'Annexe 7.2. Puisque l'algorithme du sous-gradient consiste

à faire varier un des paramètres en considérant les autres optimaux, le vecteur $\mu_{\mathbf{j_0}}^{\check{}}$ tel que tous les μ_j sont optimaux sauf μ_{j_0} présente un intérêt particulier. Ce vecteur est tel que :

$$\begin{aligned} \forall j \neq j_0, \qquad & \mu_{\mathbf{j_0}}^{\check{}}(j) = \mu_j^* \\ \text{for } j = j_0, \qquad & \mu_{\mathbf{j_0}}^{\check{}}(j_0) = \mu_{j_0} \end{aligned} \tag{4.5}$$

Ainsi l'obtention du sous-gradient sur le réseau j_0 requiert l'étude de la différence entre le vecteur défini précédemment au cours de deux itérations successives :

$$\Delta\mathcal{L}_{j_0} = \mathcal{L}(\mathbf{P}^*, \lambda^*, \mu_{\mathbf{j_0}}^{\check{}\,(\mathbf{1})}) - \mathcal{L}(\mathbf{P}^*, \lambda^*, \mu_{\mathbf{j_0}}^{\check{}\,(\mathbf{0})}) \tag{4.6}$$

Le résultat obtenu ne peut pas être utilisé immédiatement. Toutefois, après quelques manipulations mathématiques aisées, l'équation précédente peut s'écrire sous la forme suivante :

$$\Delta\mathcal{L}_{j_0} \leq (\mu_{j_0}^{(1)} - \mu_{j_0}^{(0)})(\sum_{k=1}^{N_u} P_{k,j_0} - P_{max,j_0}) \tag{4.7}$$

D'après la définition du sous-gradient, le lecteur remarquera que $(\sum_{k=1}^{N_u} P_{k,j_0} - P_{max,j_0})$ constitue un sous-gradient pour la station de base j_0. En outre, l'ensemble des sous-gradient s'écrit comme suit :

$$\delta_{j_0} = \epsilon_{j_0}(\sum_{k=1}^{N_u} P_{k,j_0} - P_{max,j_0}) \tag{4.8}$$

où ϵ_{j_0} désigne le pas du sous-gradient et définit la vitesse de convergence de l'algorithme. Définissons $[x]^+ = \max(0, x)$, il est connu [Bert 99] que la méthode de mise à jour proposée ci-dessous converge sous réserve que le pas d'itération ϵ_j soit suffisamment petit :

$$\mu_j^{(n+1)} = [\mu_j^{(n+1)} + \epsilon_j(\sum_{k=1}^{N_u} P_{k,j} - P_{max,j})]^+ \tag{4.9}$$

Le choix du pas de mise à jour n'est pas simple. En effet, la démonstration proposée en annexe définit les conditions que doivent respecter les différents pas ϵ_j permettant d'assurer la convergence de l'algorithme. Si le pas d'itération est trop petit, l'algorithme convergera mais de manière très lente. Au contraire, si le pas est trop grand, l'algorithme convergera rapidement rapidement pour certains scenarios mais pas du tout dans d'autres cas. Pour obtenir davantage de détails, le lecteur se référera aux simulations du papier joint en Annexe 7.2.

On remarquera aisément que la valeur des sous-gradients dépend directement de l'écart entre la somme des demandes des utilisateurs et la puissance disponible à la station

considérée. Cela s'explique par le fait que plus une station reçoit de demandes, plus ses ressources sont rares et donc coûteuses.

Les résultats de simulations (disponibles dans la Section Simulations du papier accessible en Annexe 7.2) montrent que cette solution, quoiqu'optimale sur l'enveloppe convexe, requiert de nombreuses itérations et se révèle donc inaplicable en pratique. Par conséquent, l'approche développée à la section suivante vise à obtenir une solution quasi-optimale après un faible nombre d'itérations.

4.3 Allocation Sous-Optimale / QdS Instantanée : Approche à Convergence Rapide

Dans cette section, nous développons une solution permettant d'obtenir une allocation quasi-optimale après un nombre restreint d'itérations. Cette solution permet ainsi de réduire la signalisation dans le système hétérogène et se révèle bien adaptée à l'implémentation pratique. La solution proposée sera désignée par l'expression "Approche à Convergence Rapide".

L'Approche à Convergence Rapide repose sur les mêmes principes mathématiques que l'Approche Optimale hormis l'optimisation effectuée au niveau des stations de base. Ainsi, comme précédemment, le problème primal décrit par l'équation (3.1) est transformé en problème dual exprimé dans l'équation (4.1), c'est-à-dire :

$$\begin{aligned}\mathcal{L}(\mathbf{P},\lambda,\mu) &= \sum_{k=1}^{N_u}\sum_{j=1}^{N_s} P_{k,j} \\ &+ \sum_{k=1}^{N_u}\lambda_k(R_{min,k}-\sum_{j=1}^{N_s} f_{k,j}(P_{k,j})) \\ &+ \sum_{j=1}^{N_s}\mu_j(\sum_{k=1}^{N_u} P_{k,j}-P_{max,j})\end{aligned}$$

A l'instar de ce qui a été réalisé précédemment, la première partie de l'implémentation concernant l'Approche à Convergence Rapide consiste à décomposer le problème précédent en des sous-problèmes pouvant être résolus de manière indépendante par les mobiles :

$$\mathcal{L}(\mathbf{P},\lambda,\mu)=\sum_{k=1}^{N_u}\mathcal{L}_\mu(\mathbf{P_k},\lambda_k)-\sum_{j=1}^{N_s}\mu_j P_{max,j} \tag{4.10}$$

avec :

$$\mathcal{L}_\mu(\mathbf{P_k},\lambda_k)=\sum_{j=1}^{N_s}(1+\mu_j)P_{k,j}+\lambda_k(R_{min,k}-\sum_{j=1}^{N_s} f_{k,j}(P_{k,j})) \tag{4.11}$$

Comme auparavant, chaque utilisateur minimise le problème dual partiel qui le concerne ($\max_{\lambda_k} \min_{\mathbf{P_k}} \mathcal{L}_\mu(\mathbf{P_k}, \lambda_k)$) en utilisant un algorithme d'optimisation discrète et convexe inspiré de [Shoh 88]. Ces minimisations partielles aboutissent à l'obtention d'un vecteur de requêtes pour chaque utilisateur. Le vecteur $\widehat{\mathbf{R}}_\mathbf{k} = [\widehat{R}_{k,1} \dots \widehat{R}_{k,N_s}]$ contient les requêtes effectuées par l'utilisateur k. Il est intéressant de remarquer qu'à aucun moment ces sous-problèmes ne tiennent compte des contraintes de puissance au niveau des stations de base. Par conséquent, il est nécessaire de vérifier au niveau de chaque station de base si la somme des demandes reçues respecte la contrainte de puissance de la-dite station. Si tel est le cas, la solution obtenue est réalisable et la solution obtenue à la première itération est optimale sur l'enveloppe convexe. Les contraintes de puissance n'interviennent dans ce cas tout à fait spécifique.

Supposons désormais que la solution obtenue ne soit pas réalisable, c'est-à-dire qu'il existe une station de base qui ne peut pas traiter les demandes qu'elle a reçu à l'aide de la puissance dont elle dispose. Dans ce cas, une optimisation doit être réalisée par les stations de base.

Afin d'analyser avec précision l'optimisation devant être réalisée par les stations de base et d'obtenir *in fine* une solution réalisable, l'ensemble relatif aux contraintes de débit des mobiles est défini de la manière suivante :

$$S_R^j = \{R_{k,j} \quad / \quad \text{for} \quad k = 1 \dots N_u, \quad R_{k,j} \leq \widehat{R}_{k,j}\}$$

Comme l'optimisation effectuée vise à obtenir le débit demandé, la solution de l'équation (4.11) se situe sur l'un des bords de l'ensemble S_R^j.

La décomposition précédente a permis d'obtenir la solution optimale sur l'enveloppe convexe et d'un point de vue des mobiles. Le lecteur a pu noter que cette méthode aboutit fréquemment à des demandes excessives pour certaines stations de base. Pour résoudre ce problème, nous proposons d'étudier le problème d'optimisation du point de vue des stations de base. Ainsi, le problème original (4.1) est décomposé de manière à obtenir un sous-problème d'optimisation pouvant être résolu par les-dites stations :

$$\mathcal{L}(\mathbf{P}, \lambda, \mu) = \sum_{j=1}^{N_s} \mathcal{L}_\lambda(\mathbf{P_j}, \mu_j) + \sum_{k=1}^{N_u} \lambda_k R_{min,k} \tag{4.12}$$

où :

$$\mathcal{L}_\lambda(\mathbf{P_j}, \mu_j) = \sum_{k=1}^{N_u} (P_{k,j} - \lambda_k R_{k,j}) + \mu_j (\sum_{k=1}^{N_u} P_{k,j} - P_{max,j}) \tag{4.13}$$

Chaque station de base doit donc désormais résoudre le problème partiel d'optimisation caractérisé par le problème dual dont l'équation est donnée par l'équation (4.13). La station cherche donc à résoudre le problème suivant :

$$\max_{\mu_j} \min_{\mathbf{P_j}} \mathcal{L}_\lambda(\mathbf{P_j}, \mu_j) \tag{4.14}$$

Soit $\check{\mathbf{P}}_\mathbf{j} = [\check{P}_{1,j} \dots \check{P}_{N_u,j}]$ le vecteur de puissance qui minimise (4.13). Il est possible

de définir le vecteur de débits associé au précédent vecteur par : $\check{\mathbf{R}}_\mathbf{j} = [f_{1,j}(\check{P}_{1,j}) \dots f_{N_u,j}(\check{P}_{N_u,j})]$. D'après les définitions précédentes, l'ensemble caractérisant la contrainte de puissance de la station de base j s'écrit :

$$S_P^j = \{R_{k,j} \quad / \quad \forall k, R_{k,j} \leq \check{R}_{k,j} \quad \text{and} \quad \sum_k P_{k,j} \leq P_{max,j}\}$$

La solution de (4.14) se situe sur l'un des bords de l'ensemble S_P^j.

Par conséquent, la meilleure solution au problème d'optimisation (4.1) *pouvant être obtenue se situe à l'intersection des ensembles S_R^j et S_P^j. Cette intersection caractérise les meilleures allocations possibles vues par les mobiles et par les stations de base. L'intersection peut être interprétée de la manière suivante :*

Tout d'abord, le lecteur remarquera que l'intersection des deux ensembles n'est pas vide ($S_R^j \cap S_P^j \neq \emptyset$). En effet, l'allocation nulle (telle que $\forall k, \forall j, R_{k,j} = 0$) appartient nécessairement à S_R^j et S_P^j. De plus, il est utile d'observer que trois situations peuvent se produire :

- Les requêtes sont exactes ($S_R^j \subset S_P^j$)
- Les requêtes sont excessives ($S_P^j \subset S_R^j$)
- Les requêtes sont possibles ($S_R^j \cap S_P^j = S^j$)

Dans le premier cas, les requêtes peuvent immédiatement être traitées par les stations de base car l'ensemble des contraintes de débit S_R^j est inclus dans l'ensemble des contraintes de puissance S_P^j.

En ce qui concerne la seconde situation, celle-ci correspond au cas où l'ensemble des contraintes de puissance S_P^j est inclus dans l'ensemble des contraintes de débit S_R^j. Ainsi, les requêtes ne peuvent pas être respectées et l'allocation s'approchant au mieux de ces requêtes appartient à l'un des bords de S_P^j. En effet, le bord de cet ensemble correspond aux allocations utilisant toute la puissance disponible.

Enfin, la dernière situation possible reflète le cas où les deux ensembles ont une intersection commune. Deux possibilités se présentent alors. Soit les requêtes des mobiles sont situées sur le bord de S_R^j et à l'intérieur de S_P^j, auquel cas elles peuvent être réalisées. Soit les requêtes sont situées sur le bord de S_R^j mais à l'extérieur de S_P^j, auquel cas elles ne peuvent l'être. Nous proposons alors que la station de base sélectionne une allocation sur le bord de S_P^j et à l'intérieur de S_R^j. L'allocation sélectionnée viole les contraintes de débit des mobiles mais respecte les contraintes de puissance des stations de base. L'objectif est de proposer une méthode permettant de limiter l'impact de cette entorse aux contraintes sur le débit des utilisateurs.

Il ressort de cette analyse la nécessité d'étudier l'intersection $S_R^j \cap S_P^j$ lorsque le problème d'optimisation atteint au moins une des contraintes réseau et que l'on souhaite obtenir la meilleure allocation disponible. Plus précisément, il y a lieu de rechercher la solution de (4.13) sous contrainte que cette solution appartienne à l'ensemble des contraintes utilisateurs S_R^j. L'allocation précitée résout le problème d'optimisation primal décrit ci-dessous :

$$\mathbf{P_j} = \underset{\{P_{k,j}\}_{k\in E_j}}{\arg\min} \sum_{k\in E_j} [P_{k,j} - \lambda_k R_{k,j}]$$
$$\text{subject to : } \begin{cases} \sum_{k\in E_j} P_{k,j} \leq P_{max,j} \\ \forall k, \quad R_{k,j} \leq \widehat{R}_{k,j} \end{cases} \tag{4.15}$$

Toute station ayant reçu davantage de requête que de ressources disponibles utilise nécessairement l'ensemble de la puissance dont elle dispose pour résoudre le problème (4.15). Mathématiquement parlant, la solution recherchée se situe sur le bord de S_P^j et à l'intérieur de S_R^j. Ainsi, la contrainte d'inégalité sur la puissance devient une contrainte d'égalité :

$$\sum_{k\in E_j} P_{k,j}(\widehat{i}_{k,j}^{(n)}) = P_{max,j} \tag{4.16}$$

Il s'ensuit que l'allocation de ressource au niveau d'une station surchargée se simplifie de la manière suivante :

$$\mathbf{P_j} = \underset{\{P_{k,j}\}_{k\in E_j}}{\arg\max} \sum_{k\in E_j} \lambda_k R_{k,j}$$
$$\text{subject to : } \begin{cases} \forall k, \quad R_{k,j} \leq \widehat{R}_{k,j} \\ \sum_{k\in E_j} P_{k,j} = P_{max,j} \end{cases} \tag{4.17}$$

Ce problème peut être résolu en utilisant un algorithme d'optimisation discète et convexe inspiré des travaux de Shoham et Gersho [Shoh 88]. En effet, une méthode permettant de résoudre le problème développé à l'équation (4.4) a été proposée à la section 2.3.4 (la proposition avait déjà été effectuée en Annexe 7.2). Cette méthode utilisée afin d'obtenir les requêtes des utilisateurs, peut également être employée pour résoudre le problème (4.17) de l'allocation au niveau des stations de base. En effet, la première contrainte de l'équation (4.17) revient à ne pas considérer les modes de transmission correspondant à des débits trop élévés (i.e. tels que $R_{k,j} > \widehat{R}_{k,j}$). En outre, la puissance dans l'équation (4.17) joue un rôle identique au débit dans l'équation (4.4) et inversement.

Une fois que l'optimisation a été réalisée à la station de base surchargée à l'aide de l'équation (4.17), toute la puissance de la station est utilisée et celle-ci est retirée du processus d'optimisation. Ainsi, les utilisateurs n'ayant pas satisfait leur contrainte de débit à l'issue de cette itération, demanderont le reste de leur débit sur les autres stations disponibles. L'itération suivante se déroulera de manière similaire, à la différence près que l'allocation sur la station retirée sera fixe. Ainsi, les utilisateurs détermineront leurs requêtes en minimisant le lagrangien partiel (4.11) mais uniquement en ce qui concerne le

débit restant. Si des stations de base étaient alors surchargées, elles résoudraient alors le problème (4.13), et ainsi de suite jusqu'à ce que la convergence soit atteinte (i.e. jusqu'à ce que tous les utilisateurs soient satisfaits ou que toute la puissance des stations de base soit attribuée). Le lecteur averti remarquera immédiatement que, dans le pire des cas, cette approche converge en N_s itérations.

Chapter 5

Maximisation du Débit Moyen avec Équité

5.1 Résumé du Chapitre

Dans ce chapitre, nous introduisons un nouveau problème d'allocation de ressources dans un contexte discret, distribué et hétérogène. Ce problème est destiné à maximiser le débit moyen obtenu par les utilisateurs et doit garantir une équité dans l'attirbution des ressources entre les utilisateurs. Il est important de noter également que la solution proposée tient compte des conditions radios, en particulier les contraintes de puissance instantanée existant au niveau des stations de base.

Dans une première section, le problème sera formulé de manière mathématique. Cette formulation diffère de ce qui a été effectué dans les chapitres précédents puisque le critère est désormais optimisé en moyenne.

Dans une seconde section, le problème est résolu à l'aide de deux algorithmes différents. Le premier algorithme consitue une extension du célèbre algorithme "Proportional Fair" dans le cas de réseaux hétérogènes. Cet algorithme se révèle très simple à implémenter mais souffre de l'absence de négociation entre les différentes entités du réseau. Ainsi, un nouvel algorithme visant à améliorer l'utilisation des ressources radios est proposé. Les résultats de simulation ont été publiés dans le papier en Annex 7.3.

5.2 Formulation du problème

L'objectif de cette section est de proposer une allocation de ressources permettant de maximiser le débit offert utilisateurs tout en respectant une certaine équité entre ceux-ci. En outre, l'allocation proposée doit s'inscrire dans le contexte discret, distribué et hétérogène de nos travaux. L'équité étudiée dans cette section est obtenue en moyenne et non à chaque instant.

Ainsi, il est nécessaire de distinguer dans ce qui suit le débit instantané pendant l'intervalle de transmission t $R^t_{k,j} = R_{k,j}(i_{k,j})$ du débit moyen observé à l'issue de cet intervalle : $T^t_{k,j}$. La relation entre le débit et la puissance instantanée s'écrit : $R^t_{k,j} = f^t_{k,j}(P^t_{k,j})$ ($f_{k,j}$ étant définies de la même manière que dans les chapitres précédents). De plus, le débit moyen s'exprime à l'aide d'un filtre passe bas exponentiellement pondéré ayant pour variables les débits mesurés aux instants précédents :

$$\forall t, \quad T^t_{k,j} = (1-\beta)T^{t-1}_{k,j} + \beta R^t_{k,j} \tag{5.1}$$

Il est intéressant de remarquer que le débit moyen à l'issue de l'intervalle t s'exprime en fonction du débit moyen à l'instant précédent (i.e. $T^{t-1}_{k,j}$), du débit instantané utilisé pendant l'intervalle étudié (i.e. $R^t_{k,j}$) et de la taille de la fenêtre de transmission (i.e. $\frac{1}{\beta}$). L'ensemble des définitions nouvellement introduites permet de définir un nouveau problème d'optimisation qui cherche à maximiser l'utilisation des ressources dans l'ensemble des réseaux tout en respectant une certaine équité entre les utilisateurs :

$$\begin{aligned} \mathbf{I^{t,*}} = \underset{\substack{\{i_{k,j}\} \\ i_{k,j}\in[1\ldots I_{k,j}]}}{\arg\max}_{\substack{k=1..N_u \\ j=1..N_s}} & \sum_{k=1}^{N_u} \log\Big(\sum_{j=1}^{N_s} T^t_{k,j}\Big) \\ \text{contraintes} \quad & \forall j, \quad \sum_{k=1}^{N_u} P_{k,j}(g^t_{k,j}, i^t_{k,j}) \leq P_{max,j} \end{aligned} \tag{5.2}$$

5.3 Résolution du problème

Dans le cadre de réseaux monotechnologiques, l'algorithme à équité proportionnelle ("Proportional Fair algorithm" ou plus simplement PF) propose un compromis tout à fait attractif entre l'obtention d'un débit élevé dans le réseau et la garantie d'une véritable équité entre les utilisateurs. De plus, un avantage remarquable de cet algorithme est son implémentation totalement distribuée. Toutefois, la résolution du problème décrit à l'équation (5.2) requiert l'utilisation d'un algorithme prenant en compte le contexte hétérogène de l'allocation de ressources. Par conséquent, une extenstion de l'algorithme intitulée "Extended Proportional Fair Resource Allocation" (EPF) sera proposée dans ce chapitre.

Notons toutefois que cette approche ne permettra pas de résoudre le problème de l'équation (5.2) puisqu'elle offre un débit moins important que la solution optimale :

$$\sum_{k=1}^{N_u}\sum_{j=1}^{N_s}\log\left(T_{k,j}^{t}\right) \leq \sum_{k=1}^{N_u}\log\left(\sum_{j=1}^{N_s}T_{k,j}^{t}\right) \tag{5.3}$$

En d'autres termes, la résolution de l'équation (5.2) nécessitera la mise en oeuvre d'un nouvel algorithme dont la performance excedera celle de l'algorithme PFE. Par conséquent, l'algorithme PFE sera décrite dans la suite de ce chapitre mais ne sera utilisée qu'à titre de comparaison avec la solution proposée. Cette solution sera d'ailleurs appelée "Vegas-Discrete Convex Optimization" (V-DCO) car elle s'inspire de l'algorithme développé Vegas et pouvant être consulté à la référence [Low 01].

5.3.1 Extended Proportional Fair Resource Allocation (EPF)

Une manière directe d'allouer des ressources de manière équitable dans un système hétérogène consiste à implémenter dans chaque station de base l'algorithme PF. Dans un système TDMA, l'algorithme PF nécessite le calcul pour chaque utilisateur du rapport entre le débit pouvant être obtenu et le débit moyen obtenu pendants les intervalles précédents. Cela résulte en une utilisation du mode de transmission le plus élevé par l'utilisateur dont le rapport est le plus élevé. Cette attitude peut aboutir à une mauvaise utilisation des ressources radios disponibles. Dans ce qui suit, nous proposons une extension à l'algorithme PF de manière à prendre en compte le contexte hétérogène de notre étude. Le nouvel algorithme (EPF) doit permettre d'allouer simultanément une puissance et un débit sur l'ensemble des liens du réseau. Cela signifie que chaque station de base doit être capable de transmettre des données simultanément à plusieurs utilisateurs et que chaque utilisateur doit être capable de recevoir simultanément des informations de plusieurs stations de base. Dans ce cas, les optimisations effectuées par les stations de base peuvent s'écrire de la manière suivante :

$$\begin{aligned} \tilde{\mathbf{i}}_{\mathbf{j}}^{\mathbf{t}} = \quad & \underset{\substack{\{i_{k,j}^{t}\} \quad k=1..N_u \\ j=1..N_s \\ i_{k,j}^{t}\in[1...I_{k,j}]}}{\arg\max} \quad \sum_{k=1}^{N_u}\frac{R_{k,j}(i_{k,j}^{t})}{T_{k,j}^{t-1}} \\ & \text{contraintes} \quad \forall j, \quad \sum_{k=1}^{N_u}P_{k,j}(g_{k,j}^{t}, i_{k,j}^{t}) \leq P_{max,j} \end{aligned} \tag{5.4}$$

Ce problème est un problème d'optimisation discrète et convexe. Nous avons vu dans les chapitres précédents qu'il était possible de résoudre ce problème en utilisant un algorithme issu des travaux de Shoham et Gersho [Shoh 88] effectués dans le cadre du codage source de l'information.

5.3.2 Vegas-Discrete Convex Optimization (WV)

Résoudre le problème de l'allocation de ressources proposé à l'équation (5.2) se révèle relativement difficile puisque ce problème vise à assurer des débits moyens tout en respectant les contraintes de puissance instantanée. Quiconque rédige se problème, entrevoit la nécessité de s'affranchir d'une des deux variables mises en oeuvre (la puissance et le débit).

Afin de résoudre le problème étudié, nous proposons de transformer les contraintes de puissance en terme de débit puis de diviser le problème en deux parties dont la première est effectuée par les utilisateurs tandis que la seconde l'est par les stations de base.

Pour transformer les contraintes de puissance en conditions sur le débit, nous supposons la capacité connue dans l'équation (5.2) (cette supposition fera l'objet de justification dans les paragraphes suivants). Puis, nous séparons le problème en deux sous-prolèmes : 1- un sous problème devant être résolu par les utilisateurs et consistant à résoudre un problème d'optimisation dont la variable est le débit de l'utilisateur 2- un sous problème devant être traité par les stations de base et qui se limite à la maximisation de la capacité du réseau. Les paragraphes subséquents permettront de détailler la solution présentée ici dans son principe.

Optimisation réalisée au niveau des utilisateurs

Comme indiqué brièvement dans la présentation de la méthode, l'idée retenue consiste à remplacer la contrainte de puissance instantanée de l'équation (5.2) par une capacité ergodique (i.e. un débit moyen limite) puisque la contrainte de puissance détermine nécessairement la capacité du réseau. Comme en outre, le canal varie rapidement (évanouissement rapide), la capacité instantanée du système est variable. Toutefois, les utilisateurs sont supposés capables de mesurer ou estimer le débit qu'ils ont obtenu au cours des intervalles de temps précédents ce qui permet de d'évaluer la capacité moyenne du système. La valeur moyenne de la capacité est la conséquence des allocations successives effectuées par les stations de base (i.e. résulte du deuxième sous-problème). Comme en outre, l'équité entre les utilisateurs est assurée sur le long terme, il est possible d'utiliser cette valeur dans le processus d'optimisation.

L'importance des explications concernant l'utilisation de la capacité provient du fait que c'est le paramètre qui lie les deux parties de l'optimisation. En effet, le premier sous-problème consiste à maximiser le critère de l'équation (5.2) *sous contrainte de la capacité moyenne générée par les optimisations successives. De plus le second sous problème consiste à trouver une allocation instantanée qui maximise la capacité instantanée.*

Passons désormais à la description détaillée de la décomposition effectuée. Pour ce faire, introduisons $\widehat{R}^t_{k,j}$, la variable associée au débit requis par l'utilisateur k envers la station j à l'instant t. Considérons également $\tilde{R}^t_{k,j}$, le débit alloué par la station j à l'utilisateur k pendant l'intervalle t. Définissons $\bar{C}^t_j$, la capacité moyenne allouée par la station j et évaluée à l'instant t. Le sous-problème devant être résolu par l'utilisateur k peut s'écrire sous la forme suivante :

$$\begin{aligned} \mathbf{I}^* = \quad & \underset{\substack{\{i_{k,j}\} \\ i_{k,j}\in[1\ldots I_{k,j}]}}{\arg\max} \quad \sum_{k=1}^{N_u} \log\Big(\sum_{j=1}^{N_s} T_{k,j}^t\Big) \\ & \text{contraintes} \qquad \forall j, \quad \sum_{k=1}^{N_u} T_{k,j}^t \leq \bar{C}_j^t \end{aligned} \tag{5.5}$$

Il est judicieux de remarquer qu'un algorithme permettant de résoudre un problème tout à fait similaire a été proposé par Low et al. [Low 01]. Dans cet article concernant les réseaux TCP , les auteurs ont amélioré l'utilisation des ressources en diminuant la congestion dans le réseau à l'aide d'un algorithme nommé TCP Vegas. Tandis que cet algorithme vise à adapter l'émission de nouveau paquet sur le réseau en fonction de la taille des files d'attente de manière à éviter la congestion, la solution que nous proposons pour résoudre le problème (5.5) a pour objectif d'ajuster le débit des utilisateurs à la capacité radio perçue par les stations de base. Par conséquent, dans les explications qui suivent, le terme congestion référera à la situation pour laquelle les stations de base reçoivent des demandes excédant leur capacité radio.

Afin de respecter la paternité de l'algorithme développé dans [Low 01] et sa similarité avec le problème (5.5) traité dans ce chapitre, l'algorithme développé ici s'intitulera "Vegas-Discrete Convex Optimization" (V-DCO). Ainsi, l'algorithme développé ici a pour objectif d'adapter les demandes des mobiles à la congestion des stations de base perçue par les utilisateurs.

Une propriété attractive de l'algorithme est qu'il ne requiert aucune transmission d'information entre les utilisateurs et les stations de base. En effet, si un ordonnanceur visant à respecter les demandes des utilisateurs est implémenté du côté des stations de base, alors chaque utilisateur est capable de vérifier s'il y a congestion au niveau de la station de base. Cette capacité provient du fait qu'un utilisateur sait si ses demandes ont été respectées ou non pour chaque station de base. Si la demande sur une station de base a été respectée, alors il n'y a pas de congestion à cette station. A l'inverse, si la demande n'a pas été respectée, alors il y a congestion à cette station de base. En outre, les utilisateurs sont capables d'estimer quel est le niveau de congestion en comparant l'écart entre leurs demandes et leurs allocations.

Pour résumer, l'algorithme V-DCO fonctionne de la manière suivante : 1- les utilisateurs se configurent itérativement en tenant compte de la congestion mesurée dans chaque réseau à l'instant précédent et 2- les stations de base allouent leurs ressources de manière à respecter au mieux les demandes puis à maximiser leur capacité

Expliquons désormais de manière théorique le processus de résolution de l'équation (5.5). Tout d'abord, le lagrangien est exprimé de manière à mettre en exergue le problème devant être traité par les utilisateurs :

$$\begin{aligned}\mathcal{L}(\mathbf{R},\lambda) &= -\sum_{k=1}^{N_u}\log\big(\sum_{j=1}^{N_s}T_{k,j}^t\big)\\ &+ \sum_{j=1}^{N_s}\lambda_j(\sum_{k=1}^{N_u}T_{k,j}^t-\bar{C}_j^t)\end{aligned} \tag{5.6}$$

La valeur optimale du précédent lagrangien s'obtient de la manière suivant :

$$\mathcal{L}(\mathbf{T}^*,\lambda^*)=\min_{\mathbf{T}}\max_{\lambda}\mathcal{L}(\mathbf{T},\lambda) \tag{5.7}$$

Pour résoudre ce problème, l'utilisation de la méthode des sous-gradients fondée sur des modifications successives des paramètres réseaux λ_j est retenue. Le pas d'itération ϵ est choisi suffisamment petit pour assurer la convergence de l'algorithme. La différentiation du lagrangien défini à l'équation (5.6) mène à l'obtention de l'équation suivante :

$$\lambda_j^{t+1}=[\lambda_j^{t+1}+\epsilon\underbrace{(\sum_{k=1}^{N_u}T_{k,j}^t-\bar{C}_j^t)}_{\delta_j}]^+ \tag{5.8}$$

Avant de poursuivre les explications concernant l'algorithme, introduisons les variables associées à la capacité instantanée ainsi qu'à la capacité moyenne. La capacité instantanée de la station de base j peut s'écrire sous la forme $C_j^t=\sum_{k=1}^{N_u}\tilde{R}_{k,j}^t$. La capacité moyenne de cette même station peut s'écrire quant à elle $\bar{C}_j^t=(1-\beta)\bar{C}_j^{t-1}+\beta C_j^t$.

Poursuivons par l'expression, à l'équilibre, des contraintes KKT (Karush-Kuhn-Tucker) qui nous intéressent [Boyd 04] :

$$\frac{\partial\mathcal{L}}{\partial T_{k,j}}(\mathbf{T},\lambda)=\frac{1}{T_{k,j}}-\lambda_j=0 \tag{5.9}$$

En remplaçant λ_j par sa valeur à l'équilibre dans l'équation (5.8), on obtient après quelques manipulations basiques :

$$T_{k,j}^{t+1}=\frac{T_{k,j}^t}{1+\epsilon T_{k,j}^t(\sum_{k=1}^{N_u}T_{k,j}^t-\bar{C}_j^t)} \tag{5.10}$$

Comme ϵ a été choisi très petit, un développement de Taylor permet d'obtenir :

$$T_{k,j}^{t+1} = T_{k,j}^{t}(1 - \epsilon T_{k,j}^{t}(\sum_{k=1}^{N_u} T_{k,j}^{t} - \bar{C}_{j}^{t})) \tag{5.11}$$

Lorsqu'on recherche une approximation de la capacité, initialement sous la forme $\bar{C}_{j}^{t} = \sum_k \bar{C}_{k,j}^{t}$, on obtient $\tilde{T}_{j}^{t} = \sum_k \tilde{T}_{k,j}^{t}$. Cette approximation dépend des débits instantanés $\bar{R}_{k,j}^{t}$. Désignons la mesure de la congestion entre l'utilisateur k et la station j par $c_{k,j}^{t} = \widehat{T}_{k,j}^{t} - \tilde{T}_{k,j}^{t}$. Fixons le pas d'itération de telle sorte que $\epsilon = \frac{1}{T_{k,j}^{t} \| \sum_j c_{k,j}^{t} \|}$. Il vient l'équation suivante :

$$T_{k,j}^{t+1} = T_{k,j}^{t}(1 - \frac{(\sum_{k=1}^{N_u}(T_{k,j}^{t} - \bar{C}_{k,j}^{t}))}{\| \sum_j c_{k,j}^{t} \|}) \tag{5.12}$$

Etant donné que l'utilisateur k ne peut savoir quelles sont les demandes des autres utilisateurs, sa décision se résume à :

$$T_{k,j}^{t+1} = T_{k,j}^{t}(1 - \frac{c_{k,j}^{t}}{c_{k}^{t}}) \tag{5.13}$$

Ainsi, l'algorithme V-DCO simule la transmission du paramètre λ_j provenant de chaque réseau en estimant la congestion et son évolution au fil des itérations. Il est utile de noter que la valeur de congestion sur un lien donné est nécessairement positive lorsque l'allocation du réseau est inférieure à la requête de l'utilisateur concerné (il y a trop de demande sur le lien). De plus, la valeur de congestion est nulle si l'allocation du réseau est égale à la demande de l'utilisateur (allocation parfaite sur le lien considéré). Enfin, la valeur est négative si l'allocation dépasse la demande (il y a alors trop peu de demandes sur le lien). Après l'introduction d'un paramètre α qui reflète la tolérance entre l'allocation actuelle et l'allocation cible, nous proposons l'implémentation de l'algorithme suivant :

$$\forall k, \forall j, \quad \widehat{T}_{k,j}^{t+1} = \begin{cases} \widehat{T}_{k,j}^{t}(1 + \frac{c_{k,j}^{t}}{c_{k}^{t}}\epsilon) & \text{if} \quad \frac{c_{k}^{t}}{\tilde{T}_{k,j}^{t}} < 0 \\ \widehat{T}_{k,j}^{t}(1 - \frac{c_{k,j}^{t}}{c_{k}^{t}}\epsilon) & \text{if} \quad \frac{c_{k}^{t}}{\tilde{T}_{k,j}^{t}} > \alpha \\ \widehat{T}_{k,j}^{t} & \text{otherwise} \end{cases} \tag{5.14}$$

Base Station Optimization Algorithm

Décrivons maintenantle sous-problème traité par les stations de base dans le cadre de l'algorithme V-DCO. L'algorithme implémenté dans chaque station de base vise à assurer les requêtes des utilisateurs tout en maximisant le débit alloué avec la puissance restante.

Toutefois, certaines stations de base peuvent recevoir des requêtes excessives et ainsi être surchargées. Par conséquent, il est nécessaire de proposer des mécanismes permettant d'allouer les ressources au plus proche des requêtes des utilisateurs et assurant l'équité entre ceux-ci.

L'algorithme V-DCO nécessite donc que chaque station de base résolve le problème décrit ci-dessous et inspiré de [Shoh 88] :

$$\begin{aligned} \tilde{\mathbf{i}}_{\mathbf{j}}^{\mathbf{t}} = \quad & \underset{\{i_{k,j}^t\}}{\arg\max} \quad \sum_k R_{k,j}(i_{k,j}^t) \\ & \text{contraintes} \quad \forall k, \quad R_{k,j}(i_{k,j}^t) \geq \widehat{R}_{k,j}^t \\ & \qquad\qquad \sum_{k=1}^{N_u} P_{k,j}(i_{k,j}^t) \leq P_{max,j} \end{aligned} \tag{5.15}$$

Si une solution à ce problème est trouvée, alors les ressources sont allouées conformément à cette solution $\tilde{R}_{k,j}^t = R_{k,j}(\tilde{i}_{k,j}^t)$. Sinon, la station de base diminue les requêtes de chaque utilisateur d'un mode de transmission jusqu'à ce qu'une solution soit trouvée au problème de l'équation (5.15).

Chapter 6

Conclusion

6.1 Contributions de la thèse

Dans cette thèse, nous avons mis en évidence la nécessité de développer des approches permettant d'assurer l'allocation de ressources dans un système hétérogène. En effet, le Chapitre 1 constitue l'introduction de la thèse qui met en exergue l'absence d'approche générique dans le domaine de l'allocation de ressources dans des réseaux hétérogènes. Ainsi, les travaux présentés dans le reste de la thèse veillent à combler ce manque en décrivant le problème de manière mathématique puis en le résolvant à l'aide de l'optimisation discrète et convexe.

De plus, pour effectuer l'allocation de ressources dans des systèmes variés, nous avons montré l'intérêt d'utiliser des méthodes distribuées. En effet, ces méthodes se révèlent à la fois plus réactives aux variations des conditions du scénario (valeurs des canaux de transmission, déplacement des mobiles, ...) et moins coûteuses en terme de calcul car celui-ci est réparti parmi les différentes entités du système.

En outre, l'optimisation des ressources peut être obtenue à l'aide de nombreuses méthodes. Les travaux mathématiques effectués depuis de nombreuses années ont largement démontré l'intérêt d'utiliser la théorie de l'optimisation convexe qui se révèle être très efficace. Cette théorie, développée au Chapitre 2, permet de réduire amplement la complexité calculatoire par l'utilisation des propriétés d'une fonction convexe, à savoir le fait qu'un minimum local est également minimum global, que le problème dual peut aisément être décomposé permettant ainsi l'utilisation de méthodes distribuées.

Une originalité supplémentaire transcrite dans ce manuscrit, se révèle être la considération de modes de transmission discrets entre les mobiles et les stations de base. Cette hypothèse repose sur la constatation de l'utilisation, dans les technologies de communication existantes, de schémas de transmission prédéfinis et disponibles en nombre limité. En effet, les technologies actuelles reposent sur l'association d'une modulation et d'un système de codage (système de Modulation et Codage Adaptatifs), de telle sorte que ceux-ci soient adaptés aux conditions radios. Ainsi, il est possible d'exprimer la puissance d'émission à partir du débit utilisé, de la qualité du lien de transmission et d'une relation convexe.

Par conséquent, l'étude proposée dans ce document s'inscrit dans le cadre de l'optimisation discrète et convexe appliquée à des systèmes hétérogènes.

Une fois le cadre technique de l'étude défini, il convenait d'aborder le cadre mathématique dans lequel s'inscrivent les travaux présentés dans ce manuscrit. Pour ce faire, le Chapitre 2 a permis de présenter les différents algorithmes d'optimisation disponibles et de les classer suivant leur caractère continu ou discret. Ainsi, ce chapitre définit le cadre mathématique de l'étude réalisée et présente les outils qui sont utilisés dans le reste du document. Deux outils utilisés par la suite sont décrits dans ce chapitre : l'algorithme du sous-gradient et l'algorithme de Shoham et Gersho [Shoh 88] (dont une variante est développée par la suite).

Le Chapitre 3 a permis d'effectuer le lien entre les cadres technique et mathématique de l'étude. En effet, ce chapitre a démontré que le problème technique étudié s'exprime dans le cadre de l'optimisation discrète et convexe. C'est dans ce chapitre que l'ensemble des éléments mathématiques relatifs au problème technique ont été définis (variables, notations, hypothèses, etc.).

Par la suite, différentes approches concernant l'allocation de ressources ont été développées. Ces méthodes diffèrent de par leurs variables, leurs critères, leurs contraintes, etc. Dans cette thèse, nous avons choisi de distinguer les méthodes en fonction de leur critère d'optimisation.

Ainsi, dans le Chapitre 4, des méthodes visant à minimiser la puissance totale instantanée allouée dans l'ensemble des réseaux sont présentées. Ces deux méthodes reposent sur la décomposition du problème d'optimisation décrit au chapitre précédent en deux sous-problèmes. Cette décomposition est obtenue à partir du problème dual. Le premier sous-problème est résolu par les mobiles en utilisant un algorithme proche de celui de Shoham et Gersho [Shoh 88], tandis que le second sous-problème est résolu par les stations de base et diffère selon la méthode utilisée. La première méthode proposée est optimale, dans le sens où elle permet de résoudre le problème de minimisation de puissance instantanée à chaque instant. Cette méthode se fonde sur l'utilisation de sous-gradient au niveau des stations de base. Comme cette méthode nécessite de nombreuses itérations et se révèle ainsi inefficace en pratique, une nouvelle méthode fondée sur des projections sur l'espace des solutions permet d'aboutir rapidement à une solution approchant efficacement la solution optimale (comme en témoignent les simulations pouvant être consultées en Annexe 7.2).

Le Chapitre 5 traite des méthodes ayant pour but d'allouer les ressources de manière à maximiser le débit moyen en observant une équité entre les utilisateurs. La solution proposée divise le problème de l'allocation de ressources en deux sous-problèmes. Le premier sous-problème consiste à adapter le débit des mobiles aux capacités des différents réseaux tandis que le second sous-problème vise à maximiser à chaque instant la capacité de chaque réseau. Le premier sous-problème s'inspire de l'algorithme TCP Vegas développé par Low dans [Low 01]. Cet algorithme vise à améliorer l'utilisation des ressources dans le réseau en minimisant la congestion au sein de celui-ci. Une approche similaire a été développée ici et s'appuie sur la minimisation de l'erreur entre le débit demandé par les mobiles et le débit alloué par les stations de base. Ainsi, les utilisateurs adaptent leurs requêtes en fonction de l'erreur perçue, tandis que les stations de

base cherchent à maximiser le débit alloué à chaque instant. Le second sous-problème consiste à effectuer une optimisation convexe et discrète similaire à celles réalisées dans les chapitres précédents.

Par conséquent, les travaux présentés dans cette thèse ont permis de développer une approche cohérente permettant de résoudre le problème de l'allocation de ressources dans des systèmes hétérogènes. Ces travaux sont fondés sur l'optimisation discrète et convexe. Le problème évoqué a tout d'abord été clairement posé d'un point de vue mathématique, puis résolu suivant différents critères. Enfin, les méthodes proposées ont cherché, à l'aide de méthodes originales, à résoudre des problèmes d'implémentation concrets (durée de convergence, complexité calculatoire, etc.) qui n'étaient pas résolus à ce jour.

6.2 Pistes d'étude

Si le travail présenté dans ce manuscrit contribue à l'élaboration d'une approche mathématique permettant de résoudre le problème de l'allocation de ressources dans des systèmes hétérogènes, la portée de ce sujet ouvre de nombreuses perspectives de recherche. Les paragraphes suivants permettront donc de présenter certaines problématiques présentant, de notre point de vue, un intérêt particulier. Certaines de ces problématiques s'inscrivent dans la continuité directe des travaux effectués, d'autres élargissent plus largement la problématique étudiée.

Parmi les études pouvant être effectuées dans le prolongement de nos travaux, l'utilisation de nouveaux critères d'optimisation, peuvent permettre aux opérateurs de développer de nouvelles stratégies d'allocation de ressources. Il est également envisageable d'étendre, de modifier ou de spécifier la formulation mathématique du problème d'optimisation de manière à l'adapter à de nouvelles problématiques concrètes.

De plus, l'étude de l'allocation de ressources à l'aide de méthodes stochastiques peut se révéler très intéressante. En effet, une telle approche peut permettre d'améliorer l'allocation de ressources dans le système hétérogène puisqu'elle tire parti de la diversité temporelle. En effet, dans le cas évoqué, il est judicieux d'allouer les ressources aux utilisateurs bénéficiant d'un canal instantané de qualité élevée, plutôt qu'à ceux dont le canal est de qualité médiocre.

En outre, l'allocation de ressources peut refléter de manière plus précise un problème réel en tenant compte de la possibilité ou non pour chaque utilisateur de se connecter à un réseau particulier. En effet, il est possible qu'un mobile puisse disposer d'une seule technologie, de quelques technologies ou de toutes les technologies. D'ailleurs, il est envisageable qu'un mobile soit dans l'impossibilité momentanée de se connecter simultanément à plusieurs technologies. Ces différentes hypothèses peuvent être envisagées dans des recherches ultérieures de manière à mieux modéliser des problèmes concrets.

En ce qui concerne les recherches nécessitant un élargissement plus vaste du domaine d'étude, nous pensons en premier lieu à l'allocation de ressources sur le lien ascendant. Si ce problème est semblable à celui étudié dans notre travaux, son implémentation n'est toutefois pas immédiate. Cette étude s'avère nécessaire pour un opérateur qui voudrait implémenter une gestion des ressources radios complètes dans un système hétérogène

réel.

De plus, il serait particulièrement intéressant d'étudier le problème en considérant les interférences entre les utilisateurs au sein d'une même technologie ou entre les stations de base d'une même technologie (mais d'une autre cellule). Il s'avère tout à fait probable que ces phénomènes influencent l'allocation de ressources. En outre, il serait intéressant d'étudier l'influence de la mobilité des utilisateurs sur l'allocation de ressources ou l'utilisation de fonction non-convexes pour effectuer celle-ci.

Pour conclure, de nombreuses pistes peuvent permettre d'élargir le spectre de nos recherches. L'étude de l'allocation de ressources dans des réseaux hétérogènes, à l'aide de méthodes distribuées, se révèle donc être un sujet d'étude particulièrement ouvert. L'objet de cette thèse n'est pas de résoudre l'intégralité des problèmes particuliers pouvant être définis mais de proposer une méthode de résolution du problème générique.

Chapter 7

Annexes

7.1 Résumé

Dans ce chapitre, les différents articles publiés ou soumis au cours de la thèse sont répertoriés.

Le premier article, intitulé “Distributed Discrete Resource Optimization in Heterogeneous Networks” (32 pages), a été soumis à la revue internationale *IEEE Transactions on Vehicular Technology*. Cet article correspond contient notamment les idées développées dans le Chapitre 4 de cette thèse. En outre, cet article propose une implémentation pratique pour les algorithmes proposés, une évaluation des performances de ces algorithmes ainsi qu’une interprétation des résultats obtenus.

Le second article, intitulé “Distributed Optimization in Heterogeneous Networks with Proportional Fairness” (6 pages), a été soumis à la conférence internationale *WCNC 2010*. Cet article aborde notamment les idées développées dans le Chapitre 5. De plus, cet article détaille les performances de l’algorithme proposé ainsi qu’une interprétation des résultats obtenus.

Le troisième article, intitulé “Distributed Discrete Resource Optimization in Heterogeneous Networks” (5 pages), a été publié à la conférence internationale *SPAWC 2008*. Cet article a constitué une première approche du problème de l’allocation de ressources dans des réseaux mobiles hétérogènes. Les travaux de cet article ont permis d’initier nos travaux concernant ce domaine.

7.2 Annexe 1

Distributed Discrete Resource Optimization in Heterogeneous Networks

Christophe Gaie, *Student Member, IEEE,* Mohamad Assaad, *Member, IEEE,* and Pierre Duhamel, *Fellow, IEEE*

Abstract

This paper deals with distributed discrete resource allocation adapted to heterogeneous systems. Indeed, distributed optimization schemes are better suited to wireless optimization problems than centralized solutions as they require lower signaling. Moreover, discrete optimization is retained as existing technologies usually rely on Adaptive Modulation and Coding (AMC) schemes.

Moreover, the objective pursued in this paper is to minimize total power within the heterogeneous system, while ensuring a minimum Quality of Service (QoS) to each user and fulfilling maximum power constraint at each base station. The proposed framework aims to solve the Radio Resource Management (RRM) problem by performing successive negotiations between the users and the base stations, using an algorithm based on Discrete Convex Optimization (DCO).

Two approaches are proposed to implement negotiations between the users and the base stations. The first approach lies on successive subgradient updates, which requires a high and unknown number of iterations to converge. To overcome this problem, we propose a second approach, which is quasi-optimal and converges quickly. This proposal lies on successive projections of requirements on the space of existing solutions.

Then, the implementations of the two proposed approaches are discussed in this paper. Their required frame structure, as well as, their signaling overhead are provided. Finally, simulation results are given and assess the advantage of the second approach.

Index Terms

Heterogeneous Systems, Discrete Optimization, Convex Optimization, Distributed Algorithm

1

I. INTRODUCTION

Nowadays the number of existing technologies has highly increased (GSM, GPRS, EDGE, WCDMA, HSPA, WiMAX, LTE and so on) while many radio frequencies reveal under-utilized. This spectrum scarcity was underlined in [1], [2], [3], [4] and raised the interest towards heterogeneous system solutions.

In heterogeneous networks, the problem of resource allocation subject to Quality of Service (QoS) constraints reveals to be a significant problem. The problem studied in this paper consists in determining how to allocate resources to the users, in order to ensure them a given QoS while fulfilling the network constraints. The problem also consists in providing a solution, whose practical implementation is feasible. This means that the solution should describe precisely the information exchange and the resource allocated at each timeslot. Furthermore, current technologies rely on Adaptive Modulation and Coding (AMC) schemes. This means that they jointly allocate power to users and select a modulation and coding scheme among a discrete set of values. Consequently, the objective of this paper is to propose a method to solve the Radio Resource Management (RRM) problem of a heterogeneous system using discrete AMC schemes. This solution should ensure optimality, low computational cost and low signaling overhead.

In literature, the RRM problem in single and multi-radio context has attracted much research interest. Many studies propose a centralized approach based on *continuous* convex optimization which can be solved via Lagrange Theory and water-filling algorithm [5], [6] and [7]. In order to apply water-filling, they assume the existence of a network entity which has a complete knowledge of all the systems. As this method requires a huge signaling and computational cost, it is not adapted to the problem studied.

To improve the allocation efficiency, much work was performed to propose distributed solutions based on continuous Lagrange Theory. In [8], Chiang and Sutivong proposed a method to solve the problem of resource allocation under nonlinear QoS constraints. This problem was solved combining Lagrange Theory and Geometric Programming. However, the proposed solution is not suited to the problem described here. Indeed, the solution requires high signaling overhead, is based on continuous optimization and converges in polynomial time.

To overcome this problem, non-cooperative game theory has been used to analyze wireless networks [9], [10] and [11]. However, distributed solutions based on non-cooperative game models

2

result in a suboptimal solution of the resource optimization problem. Therefore, cooperative game models, where players coordinate to achieve a mutually desirable solution, have been studied in [12], [13] and [14] for analysis of networks and spectrum sharing. However, cooperative game theory requires much signaling overhead due to the coordination between the players which is contradictory with the initial objective.

In literature, any of these solutions to the RRM problem of a single system may be invesitgated under one of the two following approaches: 1- Rate Adaptive, which consists in maximizing the total capacity (see [15] and [16]), 2- Margin Adaptive, which aims to minimize the overall power usage under minimum user QoS (see [17] and [18]). Minimizing power consumption is more attractive for operators since it enables them to minimize network usage cost, reduce interference in other cells and therefore increase the number of satisfied users within the heterogeneous system. For the remainder of the paper, we consider that the users can use and maintain simultaneously multiple connections with the different base stations of non-interfering systems (i.e. multi-radio capability) and can re-assemble data from different sources (i.e. multihoming capability).

The problem we study here differs from previous literature in many aspects. First, it deals with *heterogeneous systems* and aims to benefit from this diversity. Then, a major difference relies in the *distributed* and *discrete* approach to solve the resource allocation problem.

Indeed, in *heterogeneous systems* the users can take advantage of multiple networks to obtain their required data rate. Solving the RRM problem in a single system using a Margin Adaptive approach requires to minimize power under a minimum data rate for each user. In this case, the allocation is feasible if it does not require more power than available on the single base station. Therefore, the constraint of maximum total power per base station is usually omitted in literature, since authors consider feasible problems. In heterogeneous systems, the optimization is quite different, as the users may obtain data rate on multiple networks. Thus a problem, which is not feasible in a single network may be feasible in a heterogeneous system, e.g. a system composed of multiple networks. Consequently, we propose to use a Modified Margin Adaptive approach, which consists in minimizing power in all networks under the two following constraint sets: 1- minimum aggregated data rate per user 2- maximum power on each base station. By maintaining simultaneously multiple connections, the users may be able to achieve a higher data rate in a heterogeneous system compared to a single one. This justifies the interest towards

3

heterogeneous systems.

The use of a *distributed* approach to solve the RRM problem in a heterogeneous contex may be understood easily. Indeed, each agent of the system (a base station or a user) has a partial knowledge of the radio conditions (for instance, a user only knows radio conditions on his links to the base stations). Furthermore, it is not efficient to introduce a network controller to gaher all pieces of information within the system, as the radio conditions change rapidly. Consequently, the global optimization problem can not be solved straightforwardly and mechanisms have to be proposed to exchange information enabling to solve the problem into a distributed manner.

The problem studied in this paper is also attractive since it deals with *discrete* resource allocation. This presents the advantage of being more adapted to practical implementation, as available technologies transmit data using a finite set of Modulation Coding Scheme. To our knowledge, the use of Shoham-Gersho algorithm [19] to solve the Discrete Convex Optimization problem, is also novel in the context of wireless communications.

The previous pragraphs explained why the problem studied in this paper was different from existing litterature. The next ones provide a brief description of the algorithms proposed in this paper. It enables to explain the major ideas enabling to solve the problem previously set.

In this paper, two approaches are proposed to solve the whole resource allocation problem. The first approach consists in finding the optimal price of resources at the base stations through negotiations and then deducing the optimal allocation. This is done via a subgradient method which relies on a small modification of the resource cost at each iteration. However, the convergence of this method depends on the subgradient stepsize, whose optimal value reveals to vary according to scenarios. Simulation results show that the number of iterations required by the subgradient method to converge can reveal too large. Therefore, a second approach is proposed. This solution is based on successive removal of the base stations which granted all of their resources. It is quasi-optimal and converges quickly.

Furthermore, we propose a frame structure to implement the proposed algorithms. This frame structure is dedicated to practical implementation. Finally, signaling overhead is studied to ensure the proposal feasibility.

The remainder of this paper is organized as follows. In Section II, the context of the study is described and the problem is explicitly stated. Section III provides the description of the solution suggested in this paper, while Section IV discusses the implementation of proposed algorithms.

4

Simulation parameters and results are provided in Section V and Section VI concludes the paper.

II. Problem Statement

We consider a heterogeneous system composed of non interfering base stations (each base station functions on a different bandwidth) with indices $j = [1 \dots N_s]$ and maximal available power $P_{max,j}$. We also assume that the users have indices $k = [1 \dots N_u]$ and seek to obtain a minimum data rate $R_{min,k}$ on the downlink. The users can also use and maintain multiple connections to the different base stations. The connection between the user k and the network j is denoted by (k, j).

Moreover, the base stations use power control and Adaptive Modulation and Coding. Therefore, the user k can only select discrete values of data rate to communicate with the base station j. These discrete values are called "modes" and are of indices $i_{k,j} = [1 \dots I_{k,j}]$. Each mode characterizes a physical configuration consisting in a suitable modulation type (e.g., BPSK, QPSK, etc.) and a suitable FEC (Forward Error Correction) encoding scheme (e.g. a convolutional code with a given code rate). We also suppose that data rates can be expressed from powers using convex relationships. This hypothesis is equivalent to the one made when dealing with continuous optimization. Moreover, the convex hypothesis holds in practice as some publications showed that the relationships between BLER and SINR are quasi-linear for real devices. To conclude, a possible relationship based on the sampling of the Shannon Capacity is detailed in the Simulation part.

In this context, the objective is to minimize the overall power usage, while ensuring a minimum mean data rate to every user and respecting maximum available power at the base stations. On link (k, j), the data rate is denoted as $R_{k,j}(i_{k,j})$, the channel quality as $g_{k,j}$ and the power as $P_{k,j}(g_{k,j}, i_{k,j})$. Consequently, the problem consists in obtaining $\mathbf{I}^* = \{i^*_{k,j}\}$, the matrix of modes to use, such that:

$$\mathbf{I}^* = \underset{\{i_{k,j}\}}{\arg\min} \sum_{\substack{k=1..N_u \\ j=1..N_s \\ i_{k,j} \in [1..I_{k,j}]}}^{N_u} \sum_{j=1}^{N_s} P_{k,j}(g_{k,j}, i_{k,j})$$

subject to:

$$\begin{cases} \forall k, & \sum_{j=1}^{N_s} R_{k,j}(i_{k,j}) \geq R_{min,k} \\ \forall j, & \sum_{k=1}^{N_u} P_{k,j}(g_{k,j}, i_{k,j}) \leq P_{max,j} \end{cases} \quad (1)$$

5

For convenience, the subscript $g_{k,j}$ will be omitted when expressing $P_{k,j}$. Note that $P_{k,j}$ and $R_{k,j}$ are related by the following relationship $R_{k,j} = f_{k,j}(P_{k,j})$ (with functions $f_{k,j}$ concave and only defined for discrete modes).

III. PROPOSAL

A. Principles

In this section, a distributed approach is proposed to solve the resource optimization problem in (1). The objective pursued is to provide an answer to the following problem: how base stations of different systems will optimize resources of all systems if they do not have knowledge of users' channel conditions in other systems?

The main idea of our proposal consists in using the fact that the users know their channel conditions in different systems. Consequently, the users and the base stations can proceed to successive negotiations in order to determine the minimum bit rate that each base station should allocate to each user.

It is important to mention that the proposed framework is used by the users and the base stations at each allocation in order to reconfigure the overall network (i.e. to determine how the users will connect to systems in the next TTI).

B. Optimal Approach

Solving the constrained optimization problem in (1) in a straightforward manner, can reveal rather difficult and not appropriate to distributed schemes. Therefore, the primal problem in (1) was converted into an unconstrained optimization problem by adding appropriate lagrange multipliers (associated to constraints) to the initial criterion. This new problem is the dual problem (see [5]) and expresses as:

$$\begin{aligned} \mathcal{L}(\mathbf{P}, \lambda, \mu) &= \sum_{k=1}^{N_u} \sum_{j=1}^{N_s} P_{k,j} \\ &+ \sum_{k=1}^{N_u} \lambda_k (R_{min,k} - \sum_{j=1}^{N_s} f_{k,j}(P_{k,j})) \\ &+ \sum_{j=1}^{N_s} \mu_j (\sum_{k=1}^{N_u} P_{k,j} - P_{max,j}) \end{aligned} \tag{2}$$

6

In (2), lagrange multipliers are the vectors $\lambda = [\lambda_1 ... \lambda_{N_u}]$ and $\mu = [\mu_1 ... \mu_{N_s}]$. Recall that the functions $f_{k,j}$ are concave and only defined for discrete values. Consequently, the lagrangian of previous equation is convex and the Lagrange dual theory states that minimizing the Lagrange dual problem is equivalent to solve the initial primal problem (1). Therefore, the optimal allocation is obtained when the lagrangian reaches the following optimal value:

$$\mathcal{L}(\mathbf{P}^*, \lambda^*, \mu^*) = \min_{\mathbf{P}} \max_{\lambda,\mu} \mathcal{L}(\mathbf{P}, \lambda, \mu) \tag{3}$$

To obtain the global minimum of a continuous resource allocation problem, usage of centralized solutions is widespread in literature [5], [6] and [7]. Yet, centralized solutions are not suited to heterogeneous systems as they require to gather information at a unique place, then compute an allocation and finally transmit it to every user. On the contrary, it is of highest interest to use the decomposition theory (for more details see [5]), so as to decompose the previous lagrangian into smaller ones. This aims to split the initial complex optimization problem into many simple subproblems.

When the decomposition theory is used, partial lagrangians are obtained. They can be solved independently by the users and the base stations. Consequently, using successive optimizations at the users and the base stations enables to obtain the global minimum by finding the optimal vector of network prices (μ^*) and deduce the associated vectors of data rate to grant to each user k ($\mathbf{R_k^{(n)}}$).

In the following, we formalize mathematically how optimizations, performed alternatively by the users and the base stations, enable to solve the Lagrange dual problem. First of all, the lagrangian in (2) is rewritten. This allows to separate the global optimization into smaller subproblems which can be solved independently by the users:

$$\mathcal{L}(\mathbf{P}, \lambda, \mu) = \sum_{k=1}^{N_u} \mathcal{L}_\mu(\mathbf{P_k}, \lambda_k) - \sum_{j=1}^{N_s} \mu_j P_{max,j}$$

with:

$$\mathcal{L}_\mu(\mathbf{P_k}, \lambda_k) = \sum_{j=1}^{N_s} (1 + \mu_j) P_{k,j} + \lambda_k (R_{min,k} - \sum_{j=1}^{N_s} f_{k,j}(P_{k,j})) \tag{4}$$

Finding the optimal solution to the global problem (2) is equivalent to minimize for each user k the partial lagrangian $\mathcal{L}_{\mu^*}(\mathbf{P_k}, \lambda_k)$ knowing the vector μ^* (weight of requirements on the

7

different base stations). Therefore, the decomposition with respect to the users is explicit and minimizing $\mathcal{L}_\mu(\mathbf{P_k}, \lambda_k)$ is equivalent to solve the following primal problem:

$$\begin{aligned} \mathbf{P_k} = \quad & \arg \min_{\{P_{k,j}\}_{j=1..N_s}} \quad \sum_{j=1}^{N_s} (1+\mu_j) P_{k,j} \\ & \text{subject to :} \qquad \sum_{j=1}^{N_s} f_{k,j}(P_{k,j}) \geq R_{min,k} \end{aligned} \tag{5}$$

This problem is a *discrete* resource allocation problem (functions $f_{k,j}$ are both discrete and concave). The previous problem can be solved at each user k using Non Linear Integer Programming [20]. However, this method reveals to have a high computational cost. Consequently, we propose to solve this problem using a method inspired from Shoham-Gersho Algorithm [19]. This algorithm enables to solve a discrete optimization problem. It was proposed in the context of information source coding and enables to reduce computational complexity by restraining the search to a limited number of configurations. Furthermore it enables a fast convergence taking advantage of the convex nature of our problem. For more details on resolution of problem in (5), see Appendix A.

In previous paragraphs we described the method to optimize resources independently for each user k considering the value of parameters μ_j as known. To achieve the global optimum of the RRM problem, we now describe a manner to obtain the optimal vector of resource weights μ^*. To fulfill this objective, a method based on subgradients is employed. This method consists in updating iteratively each network weight μ_j according to data rate requirements, channel quality of links and available power. This method is shown to converge to the optimal solution.

To obtain the optimal solution the approach proposed consists in, first, obtaining the user k requirements by solving equation (5) considering $\mu_j^{(n)}$ as a constant and then computing for each base station j a subgradient δ_j which reflects the gap between requirements and available power (the larger the subgradient magnitude, the higher the gap). This subgradient is then used to find the new parameter $\mu_j^{(n+1)}$. This process is repeated until convergence. It is known that using a subgradient method enables to converge to the optimal solution [21], [22].

A "Subgradient" algorithm which can be applied to find the optimal allocation, is presented thereafter. A comprehensive proof of convergence, which also provides bounds on the subgradient stepsize, is performed in Appendix B. Before expressing the subgradient, the vector $\tilde{\mu_{\mathbf{jo}}}$ is

8

introduced. This vector is such that every parameter μ_j is optimal excepted μ_{j_0}, that is to say:

$$\begin{aligned} \forall j \neq j_0, \qquad & \tilde{\mu_{\mathbf{j_0}}}(j) = \mu_j^* \\ \text{for } j = j_0, \qquad & \tilde{\mu_{\mathbf{j_0}}}(j_0) = \mu_{j_0} \end{aligned} \tag{6}$$

To obtain the subgradient of network j_0, it is required to study how the lagrangian evolves when only the price of resources changes on this network j_0. Consequently, the difference of previous vector between two successive iterations is expressed:

$$\Delta\mathcal{L}_{j_0} = \mathcal{L}(\mathbf{P}^*, \lambda^*, \tilde{\mu_{\mathbf{j_0}}}^{(\mathbf{1})}) - \mathcal{L}(\mathbf{P}^*, \lambda^*, \tilde{\mu_{\mathbf{j_0}}}^{(\mathbf{0})}) \tag{7}$$

The result obtained cannot be directly employed. However, after some basic substitution and factorization, previous equation leads to the following expression:

$$\Delta\mathcal{L}_{j_0} \leq (\mu_{j_0}^{(1)} - \mu_{j_0}^{(0)})(\sum_{k=1}^{N_u} P_{k,j_0} - P_{max,j_0}) \tag{8}$$

Using the definition of subgradient, one can notice that $(\sum_{k=1}^{N_u} P_{k,j_0} - P_{max,j_0})$ is a subgradient for the base station j_0 and that subgradients are of the following form:

$$\delta_{j_0} = \epsilon_{j_0}(\sum_{k=1}^{N_u} P_{k,j_0} - P_{max,j_0}) \tag{9}$$

where ϵ_{j_0} is the subgradient stepsize and defines the convergence and its speed. Denoting as $[x]^+ = \max(0, x)$, it is known from [23], that using the following update method for each base station j enables to ensure the convergence, provided that the stepsize ϵ_j is sufficiently small:

$$\mu_j^{(n+1)} = [\mu_j^{(n+1)} + \epsilon_j(\sum_{k=1}^{N_u} P_{k,j} - P_{max,j})]^+ \tag{10}$$

The choice of an efficient stepsize is not trivial. Indeed, Annex B sets conditions on ϵ_j to ensure the algorithm convergence. If the stepsize is very small, then the algorithm will converge but at a very low speed. On the contrary, if the stepsize is relatively large, the algorithm will converge quickly for some scenarios but will not converge for other scenarios. More details on the choice of this parameter will be given in Section V.

One can remark that subgradients directly depend on the distance between aggregated power requirements and power available. This can be easily understood as the highest power required to a given base station is, the highest its price of resources is.

9

Simulation results provided in Section V show that this solution requires many negotiations and is not suitable for practical implementation. Thus, a fast convergence approach is the following Section.

C. Fast Convergence Approach

The solution proposed in this Section aims to find an allocation within few negotiations and therefore reduce signaling overhead. This approach reveals more adapted to real devices and practical implementation.

The Fast Convergence approach leans on the same principles than the Optimal approach with differences in the base station part (second part) of the optimization. Thus, as previously, the primal problem (1) is converted into the dual problem (2), that is to say:

$$\begin{aligned}\mathcal{L}(\mathbf{P},\lambda,\mu) &= \sum_{k=1}^{N_u}\sum_{j=1}^{N_s} P_{k,j} \\ &+ \sum_{k=1}^{N_u} \lambda_k (R_{min,k} - \sum_{j=1}^{N_s} f_{k,j}(P_{k,j})) \\ &+ \sum_{j=1}^{N_s} \mu_j (\sum_{k=1}^{N_u} P_{k,j} - P_{max,j})\end{aligned}$$

The first part of the Fast Convergence approach still consists in decomposing this problem into smaller optimizations performed independently by the users:

$$\mathcal{L}(\mathbf{P},\lambda,\mu) = \sum_{k=1}^{N_u} \mathcal{L}_\mu(\mathbf{P_k},\lambda_k) - \sum_{j=1}^{N_s} \mu_j P_{max,j} \tag{11}$$

with:

$$\mathcal{L}_\mu(\mathbf{P_k},\lambda_k) = \sum_{j=1}^{N_s}(1+\mu_j)P_{k,j} + \lambda_k(R_{min,k} - \sum_{j=1}^{N_s} f_{k,j}(P_{k,j})) \tag{12}$$

As previously, each user minimizes its partial dual problem ($\max_{\lambda_k} \min_{\mathbf{P_k}} \mathcal{L}_\mu(\mathbf{P_k},\lambda_k)$) using a discrete convex optimization algorithm inspired from [19]. These partial minimizations lead to obtain vector of requirements for each user. The vector $\widehat{\mathbf{R}}_\mathbf{k} = [\widehat{R}_{k,1} \dots \widehat{R}_{k,N_s}]$ refers to requirements for the user k. Note that these subproblems do not enclose any constraint on power available at the base station. This imply that it is required to check the feasibility of the solution composed of solutions to subproblems treated by the users. Suppose that this solution

is feasible, that is to say that it fulfills power constraints at the base stations. Then the optimal allocation is obtained at the first iteration.

Now suppose that the obtained solution is not feasible by a given base station. In this case, an optimization should be performed by the base stations.

In order to analyze in details the optimization performed in this paper and to provide a feasible solution, let us define the following set:

$$S_R^j = \{R_{k,j} \quad / \quad \text{for} \quad k = 1 \dots N_u, \quad R_{k,j} \leq \widehat{R}_{k,j}\}$$

The solution in (12) is contained on the boundary of S_R^j.

Let us reconsider the original problem (2) and decompose this problem with respect to each base station i.e. we will find the optimal solution viewed by the base stations. Consequently, equation (2) is rewritten as:

$$\mathcal{L}(\mathbf{P}, \lambda, \mu) = \sum_{j=1}^{N_s} \mathcal{L}_\lambda(\mathbf{P_j}, \mu_j) + \sum_{k=1}^{N_u} \lambda_k R_{min,k} \tag{13}$$

where:

$$\mathcal{L}_\lambda(\mathbf{P_j}, \mu_j) = \sum_{k=1}^{N_u} (P_{k,j} - \lambda_k R_{k,j}) + \mu_j (\sum_{k=1}^{N_u} P_{k,j} - P_{max,j}) \tag{14}$$

Each base station should minimize its partial dual problem ($\max_{\mu_j} \min_{\mathbf{P_j}} \mathcal{L}_\lambda(\mathbf{P_j}, \mu_j)$) whose equation is given by (14). Define $\check{\mathbf{P}}_\mathbf{j} = [\check{P}_{1,j} \dots \check{P}_{N_u,j}]$ as a vector of power minimizing (14). One can define the associated vector of rates $\check{\mathbf{R}}_\mathbf{j} = [f_{1,j}(\check{P}_{1,j}) \dots f_{N_u,j}(\check{P}_{N_u,j})]$. From the previous definitions, we define for each base station j the following set:

$$S_P^j = \{R_{k,j} \quad / \quad \forall k, R_{k,j} \leq \check{R}_{k,j} \quad \text{and} \quad \sum_k P_{k,j} \leq P_{max,j}\}$$

The solution in (14) is contained on the boundary of S_P^j.

Therefore, the best possible solution to the optimization problem (2) *is the intersection between* S_R^j *and* S_P^j*. This corresponds to the intersection between the best allocations viewed by the users and the best allocations viewed by the base stations. This intersection can be interpreted as follows:*

First, one can notice that the intersection is not empty ($S_R^j \cap S_P^j \neq \emptyset$). Indeed, the null allocation (such that $\forall k, \forall j, R_{k,j} = 0$) belongs to S_R^j and S_P^j. Then remark that three situations can occur:

11

- exact requirements ($S_R^j \subset S_P^j$)
- excessive requirements $S_P^j \subset S_R^j$
- possible requirements ($S_R^j \cap S_P^j = S^j$)

In the first case where the set S_R^j is included in S_P^j, the requirements performed to the base station can be handled and the given base station only grants them. In the second case, the set S_P^j is included in S_R^j, this means that requirements can not be handled. In this case, the best available behavior consists in finding the best allocation under the bound of S_P^j. Indeed, this bound consists in allocation of maximal available power, that is to say, the best available allocations. Finally, the third case consists in situations where sets have a common intersection. In this case, two situations are possible. If requirements are on the boundary of S_R^j and inside S_P^j, they can be handled exactly as explained for exact requirements. On the contrary, if the requirements are on the boundary of S_R^j and outside S_P^j, they cannot be handled as too much power was required. In this latter case, we propose that the base station selects a feasible allocation on the boundary of S_P^j and within S_R^j.

From the above analysis, one should consider $S_R^j \cap S_P^j$ in order to find the best feasible allocation. In other words, the best feasible allocation is the solution of (14) with respect to the constraint set S_R^j. The optimization is performed for a value of λ obtained from the subproblem previously solved by the mobiles. Consequently, the resource allocation is obtained by solving the following primal problem:

$$\mathbf{P_j} = \arg \min_{\{P_{k,j}\}_{k \in E_j}} \sum_{k \in E_j} [P_{k,j} - \lambda_k R_{k,j}] \quad \text{subject to :} \quad \begin{cases} \sum_{k \in E_j} P_{k,j} \leq P_{max,j} \\ \forall k, \quad R_{k,j} \leq \widehat{R}_{k,j} \end{cases} \tag{15}$$

For each overloaded base station, full power is used by solving equation (15) (i.e. the solution is on the boundary of S_P^j and within S_R^j). This implies that $\sum_{k \in E_j} P_{k,j}(\widehat{i}_{k,j}^{(n)}) = P_{max,j}$. Consequently, for each base station in overload, equation (15) can be simplified as follows:

12

$$\mathbf{P_j} = \arg \max_{\{P_{k,j}\}_{k \in E_j}} \sum_{k \in E_j} \lambda_k R_{k,j}$$
$$\text{subject to : } \begin{cases} \forall k, \quad R_{k,j} \leq \widehat{R}_{k,j} \\ \sum_{k \in E_j} P_{k,j} = P_{max,j} \end{cases} \tag{16}$$

This problem can be solved via a DCO algorithm inspired from [19]. A method to solve equation (5) and therefore obtain the requirements was proposed in Appendix A. The problem (16) can also be solved by this method. Indeed, the first constraint in (16) consists in suppressing modes of too high data rate (i.e. such that $R_{k,j} > \widehat{R}_{k,j}$. Moreover, the power in (16) plays the role of the data rate in (5) and reciprocally.

Once the overloaded base station performs the optimization in (16), its power usage is $P_{max,j}$ and it is then removed from the allocation process. The users that have not yet satisfied their data rate constraint will require data rate to the other base stations. A new iteration will start, where the users will again minimize the partial lagrangians in (12) for the remainder rate constraint, by obtaining resources from the remainder base stations. After they have transmitted their requirements to the base stations, the overloaded base stations will proceed to solve (14), and so on and so forth until convergence (i.e. until all users data rates are satisfied and the power used in each base station is inferior to $P_{max,j}$). It is obvious that this approach will converge in at worse N_s iterations.

IV. Algorithm Implementation

In this Section, the algorithm implementation is discussed in details. First, a frame structure is proposed. Then, algorithm functioning is detailed. Finally, signaling overhead is studied.

A. Frame Structure

The convergence and efficiency of proposed algorithms directly depends on the duration of the configuration and allocation phases. Indeed, the configuration phase should be short compared to the allocation phase to limit signaling overhead. Moreover, the allocation phase should be long compared to the configuration phase to limit signaling overhead and ensure that requirements

can be fulfilled. However, the allocation phase should be sufficiently small to ensure that channel estimations remain relevant.

Figure 2 shows the frame structure adopted for Frequency Duplex Division (FDD) systems at each timeslot. Each FDD frame is composed of a short configuration phase and a long allocation phase. As indicated on the figure, these phases last respectively T_{conf} and T_{alloc}. During the configuration phase each user has access to a predefined resource unit (negotiated before) and addresses its requirements via this dedicated resource unit. Recall that in this paper we only deal with downlink resource allocation.

White bars correspond to downlink communications and gray bars to uplink communications. Moreover, bare bars are dedicated to configuration messages while hashed bars are kept for data transmissions (allocations) and crossed bars refer to unused resources. Thus, a FDD communication on the downlink (what is treated in this paper), is firstly composed of a configuration phase which alternates short configuration messages on the uplink and then on the downlink (while transmitting power on the remainder of resources). Then an allocation phase is performed on the downlink. Note that the FDD communication on the uplink is symmetric to the downlink.

Figure 3 shows the frame structure adopted for Time Duplex Division (TDD) systems. A TDD frame is composed of successive uplink and downlink transmissions. The TDD frame structure is very similar to the FDD one. Indeed, when the users send requirements to the base stations to obtain a downlink configuration, the users can also reply to the base stations to obtain an uplink configuration (and reciprocally). Therefore, using TDD also enables to compute uplink and downlink allocations simultaneously. However, for simplicity purposes, the remainder of the paper focuses on downlink resource allocation on FDD systems.

In the following, each negotiation phase is called “iteration” and aims to minimize the Lagrange dual problem (2). An iteration is composed of one user requirement computation and one feedback computation at the base stations. Therefore, an iteration is composed of two parts in every technology. From what was stated before, one can deduce that the objective pursued is to reduce the number of iterations (to reduce signaling overhead), while finding an efficient configuration.

Figure 4 and Figure 5 show the configuration phase structure for the two algorithms proposed in this paper. One can notice that the “Base Station Removal” algorithm requires more information exchange at each iteration but it has a fix low number of iterations. On the opposite,

the "Subgradient" algorithm requires an unknown number of iterations with low information exchange. Moreover, the "Subgradient" algorithm requires more iterations than the "Base Station Removal" algorithm to obtain an allocation(see Section V). Therefore, it reveals very challenging to design the frame structure for the "Subgradient" algorithm. Indeed, one can wonder whether the configuration phase should be limited and to which value. Moreover, if the network designer restrains the configuration phase duration, one can wonder about the interest of designing an optimal algorithm and do not fulfilling its constraints.

Moreover, suppose that timeslots have a different value on different networks. Then, coordination has to be recovered at the base stations and at the users for "Subgradient" algorithm (and potentially for an huge number of iterations), whereas coordination has to be recovered only at the base stations for "Base Station Removal" algorithm (and only for a small number of iterations). Therefore, "Base Station Removal" algorithm is more adapted to practical implementation than "Subgradient" algorithm.

B. Algorithm Functionning

1) Algorithm 1 ("Subgradient"): At each iteration, the user requirements are computed according to equation (5) and feedback vector $\tilde{\mathbf{i}}_{\mathbf{j}}^{(\mathbf{n})}$ are computed via subgradients according to equation (5). We describe here in details the "Subgradient" algorithm.

Algorithm Implementation

- Initialize algorithm
  - Compute Path Loss
  - Compute Shadowing
  - Initialize $\forall j, \mu_j^{(0)} = 0$
  - n=1
- While $n <= N_{iter}$
  - Each user k, computes $\widehat{\mathbf{P}}_{\mathbf{k}}^{(\mathbf{n})}(\mu^{(\mathbf{n-1})})$ using (5) with the DCO algorithm
  - Each station j, computes $\Delta_j = (\sum_{k=1}^{N_u} \widehat{P}_{k,j}^{(n)} - P_{max,j})$
  - If $\Delta_j <= 0, \mu_j^{(n)} = \mu_j^{(n-1)}$
  - If $\Delta_j > 0, \mu_j^{(n)} = \mu_j^{(n-1)} - \epsilon_j \Delta_j$

– Update: $n = n + 1$

2) Algorithm 2 ("Base Station Removal"): Before describing the "Base Station Removal" algorithm, we first explain its global functioning. The algorithm checks if the sum of powers required to a given base stations exceeds its maximum available power. If this condition is true, the given base station does not grant to the users what they required. Instead, the given base station evaluates the rates that it can allocate to the users according to equation (16). The allocation obtained is sent back to the users via a feedback vector $\tilde{\mathbf{i}}_{\mathbf{j}}^{(\mathbf{n})}$. Consequently, as expected, the algorithm converges at worst in N_s iterations. Simulation results presented in Section V also evaluate the algorithm efficiency in terms of the user data rate satisfaction.

Algorithm Implementation

- Initialize algorithm
  - Compute Path Loss
  - Compute Shadowing
  - n=1
- While it exists some stations
  - Each user k, computes $\widehat{\mathbf{P}}_{\mathbf{k}}^{(\mathbf{n})}(\mu^{(\mathbf{n-1})})$ and $\lambda_k^{(n)}$ of (5) the DCO algorithm
  - Each station j, computes $\Delta_j = (\sum_{k=1}^{N_u} \widehat{P}_{k,j} - P_{max,j})$
  - If $\Delta_j <= 0$, station j allocates $\tilde{\mathbf{P}}_{\mathbf{j}}^{(n)} = [\widehat{\mathbf{P}}_{\mathbf{k}}^{(n)}(j)]$
  - If $\Delta_j > 0$, station j computes $\tilde{\mathbf{P}}_{\mathbf{j}}^{(n)}(\widehat{\mathbf{P}}_{\mathbf{k}}^{(n)}(j), \lambda^{(\mathbf{n})})$ in (16) using a DCO algorithm inspired from [19]. Then the base station j is removed from the configuration process (the base station j does not belong to $V^{(n+1)}$)..
  - Update: $n = n + 1$

C. Signaling Overhead

The "Subgradient" algorithm consists for each base station to compute its new parameter $\mu_j^{(n+1)}$ from the previous one and from the user requirements on the base station j ($\mathbf{P_j} = [\mathbf{P}_1(j) \dots \mathbf{P}_{N_u}(j)]$). Therefore, the exchange of information consists for each user to transmit

its requirements to each base station (maximum $N_u \times N_s$ parameters) and for each base station to transmit its new parameter $\mu_j^{(n+1)}$ (maximum N_s parameters). This algorithm requires the knowledge/transmission of $(N_u + 1) \times N_s$ parameters per iteration.

The "Base Station Removal" algorithm functions as follows. At each iteration n, the algorithm requires to transmit to the base stations the data rate required $\widehat{\mathbf{R}}_{\mathbf{k}}^{(\mathbf{n})}$ and a parameter $\lambda_k^{(n)}$ by the user k. Note $N_s^{(n)}$ the number of base stations concerned by iteration n. Moreover, the base stations which have computed their allocation transmit them to the users (that is to say the vector $\tilde{\mathbf{R}}_{\mathbf{k}}^{(\mathbf{n})}$ for the base stations not removed from the process). Therefore, the algorithm requires the transmission of $2 \times (N_u \times N_s^{(n)}) + N_s^{(n)}$ parameters at iteration n. There is no need to transmit vector $V^{(n+1)}$ as the users know when the base stations use their full capacity.

Table I displays the differences between "Subgradient" and "Base Station Removal" algorithms (Vector $\mathbf{V}$ contains the indices of the base Stations available for next iteration). One can notice that the proposed algorithms present very different characteristics in terms of signaling overhead, time of convergence, computational complexity, optimality and performance (see simulation results for this last characteristic).

V. Simulations

A. Simulation Parameters and Radio Conditions

In the considered scenarios, we study the resource allocation of three base stations to ten users. The considered base stations are located at the same place but function at different bandwidths. The users are in a short or medium range from the base stations. As stated previously, the users are able to receive simultaneously data from the base stations. Detailed parameters (such as the number of Monte Carlo Simulations, the cell range, etc.) are displayed in Table II.

Simulations performed also assume that radio conditions between the users and the base stations are obtained from pedestrian A model of [24] with parameters displayed in Table III.

The Signal to Interference plus Noise ratio is given by the following expression:

$$SINR_{k,j} = \frac{\|g_{k,j}\|^2 P_{k,j}}{I_{inter} + I_{intra} + \sigma_n^2} \tag{17}$$

Figure 8 displays the relationships between target SINR and data rates for the considered base stations [25]. One can notice that SINR and data rates are related by discrete and convex relationships.

17

Furthermore, as underlined in the problem statement, data rates may be expressed from powers using concave functions. In our simulations, we used the following relationship inspired from the Shannon capacity:

$$R_{k,j} = B\log(1 + SINR_{k,j}) = B\log(1 + \alpha_{k,j}P_{k,j}) \quad (18)$$

B. Simulation Results

In Annex B, we computed the conditions that parameter ϵ_{j_0} should verify to ensure "Subgradient" algorithm convergence. These conditions require that the stepsize belongs to a close range which evolves according to iterations. However, in practice the users should have a fixed stepsize for subgradient update. Figure 9 displays the number of iterations required to ensure convergence for three different Monte Carlo simulations of Scenario 1.

One can easily notice the difficulty encountered to find the optimal value for the stepsize. Indeed, the convergence time depends upon simulations. For Simulation 1, the optimal value for the stepsize (i.e. the value which leads to the fastest convergence) is 0.1 whereas this value is equal to 0.2 for Simulation 2. Note that there are multiple efficient stepsizes for Simulation 3. If the network designer wants to set the value of epsilon, then it will result in a variable convergence time. In other words, the number of iterations (signaling exchange between the base stations and the users) will be variable. This is not suitable for practical implementation where the signaling overhead should be fixed and specified regardless the channel variations or any other change of scenario. Moreover, one can see in Figure 9 that, in general, the "Base Station Removal Algorithm" (less than three iterations) converges faster than the "Subgradient" algorithm.

Table IV displays the number of Monte Carlo simulations which converged, depending on the algorithm. We stop the "Subgradient" algorithm after 50 iteration as 50 iterations is not implementable and is equivalent to divergence (even 50 is very high). One can notice that a small number of simulations only converge with "Subgradient" algorithm. One can also remark that very few simulations only converge for "Base Station Removal" algorithm. In these particular cases, the "Subgradient" algorithm does not converge because of the 50 iteration limit.

Figure 11 displays the Cumulative Density Function (CDF) of power when both algorithms converge. One can notice that, when both algorithms converge, granted power is approximatively the same for each algorithm.

Figure 10 displays the Cumulative Density Function (CDF) of power when only the “Subgradient” algorithm converges. One can remark that, when only the “Subgradient” algorithm converges, the base stations of the “Base Station Removal” algorithm function at their maximal power (as the base stations grant every available resource in the latter case) while the base stations of the “Subgradient” algorithm use approximatively all their resources. This can be explained by the fact that the “Base Station Removal” algorithm does not converge to the optimum but to a close solution.

Indeed a proof of this quasi-optimality is obtained by evaluating the difference between the lagrangians obtained by the two algorithms (at last iterations). Thus, Figure 12 proves that when both algorithms convergence the difference is very small. One can effectively notice that only 10% of the simulations have a difference superior or equal to 15%. One can also remark that in the worst case the difference between the two lagrangians do not exceed 30%.

From previous observations one can deduce that “Base Station Removal” algorithm is quasi-optimal (convergence for most of the simulations and same power granted than optimal algorithm when both algorithm converge) and converges quickly. This reveals very useful for practical implementations.

VI. Conlusion

In this paper we provided a distributed discrete resource allocation algorithm minimizing the power usage in heterogeneous systems, while respecting the user minimum data rate constraints and maximum available power at the base stations.

To solve the RRM problem, a distributed framework is proposed in this paper. This framework consists in successive negotiations between the users and the base stations. Two algorithms were proposed: the first one is called “Subgradient” algorithm, is based on successive resource price updates and converges to the optimal solution; the second one is entitled “Base Station Removal” algorithm and consists in searching for each base station the allocation, which is closest to requirements and fulfills the base station power constraints. Therefore, this consists in a projection of the requirements on the space of feasible allocations. In this paper, we also provided a frame structure, which ensures practical implementation of both approaches.

Simulation results show that the “Subgradient” algorithm converges to the optimal solution if the iteration stepsize is sufficiently small. Results also display that choosing an appropriate

stepsize for subgradient updates is not easy and that there is no optimal choice. Furthermore, the "Subgradient" algorithm can reveal to converge very slowly compared to the "Base Station Removal" algorithm. This reveals to be a problem, as any practical implementation requires to use a fixed low number of message exchange, and justifies the interest of the proposed method.

APPENDIX A
OBTAINING REQUESTS WITH DISCRETE CONVEX OPTIMIZATION (DCO)

In this annex, the problem in (5) is solved using a DCO algorithm inspired from [19]. Note that this enables to obtain the best resource requirements on the complex envelope. Mode indices can be omitted as power are directly related to data rates on link (k, j) by formula (17). Furthermore, modes such that $\widehat{P}_{k,j}^{(n)} > P_{max,j} - \sum_{k' \neq k} \tilde{P}_{k',j}^{(n-1)}$ can be removed as they can not be chosen in the optimization process. Finally, parameters μ_j can be set to zero without loss of generality. Consequently, calculation of the user k requirements given by equation (5), can be transformed into the following problem:

$$\begin{array}{ll} \min\limits_{\{\widehat{R}_{k,j}^{(n)}\}_{j=1..N_c}} & \sum_{j=1}^{N_c} \widehat{P}_{k,j}^{(n)} \\ \text{subject to :} & \sum_{j=1}^{N_c}(-\widehat{R}_{k,j}^{(n)}) \leq -R_{min,k}^{(n)} \end{array} \tag{19}$$

This problem can be solved using a DCO algorithm inspired from [19], that is to say by minimizing the following Lagrangian untill constraints are satisfied:

$$L = \sum_{j=1}^{N_c} \widehat{P}_{k,j}^{(n)} + \lambda \sum_{j=1}^{N_c}(-\widehat{R}_{k,j}^{(n)}) \tag{20}$$

Algorithm Implementation

- Step 1: Initialization of λ
  - Although there are better initialization the problem can be initialized by setting $\lambda = 0$.
- Step 2: Unconstrained Problem Resolution
  - Search the vector of allocations $\{-\widehat{R}_{k,j}^{(n)}\}_{j=1..N_c}$ which minimizes equation (20).
- Step 3: Checking Constraints and Updating λ

- For each allocation, compute the aggregated data rate $-R(\lambda) = \sum_{j=1}^{N_c}(-\widehat{R}_{k,j}^{(n)})$.
- a) λ is not singular (single solution)
  - i) $-R(\lambda) = -R_{min,k}^{(n)}$:
    The optimal allocation was found. The algorithm is stopped and the solution returned.
  - ii) $-R(\lambda) > -R_{min,k}^{(n)}$:
    λ is updated so as to obtain the nearest singular value which comes closer to the constraint (smallest increase).
  - iii) $-R(\lambda) < -R_{min,k}^{(n)}$:
    λ is updated so as to obtain the nearest singular value which comes closer to the constraint (smallest decrease).
- b) λ is singular (multiple solutions)
  The lowest constraint obtained is defined as $-R_l(\lambda)$ and the highest constraint as $-R_h(\lambda)$.
  - i) $-R_l(\lambda) \leq -R_{min,k}^{(n)} \leq -R_h(\lambda)$:
    The best allocation on the complex envelope is the highest constraint such that $-R_a(\lambda) \leq -R_{min,k}^{(n)}$. It always exists since $-R_l(\lambda)$ fulfills the constraint.
  - ii) $-R_l(\lambda) > -R_{min,k}^{(n)}$:
    λ is updated so as to obtain the nearest singular value which comes closer to the constraint (smallest increase).
  - iii) $-R_h(\lambda) < -R_{min,k}^{(n)}$:
    λ is updated so as to obtain the nearest singular value which comes closer to the constraint (smallest decrease).

Here is a figure of the optimization scheme:

APPENDIX B
"SUBGRADIENT" ALGORITHM CONVERGENCE PROOF

The subgradient Algorithm is based on the successive update of vector of parameter μ_j at each base station j. This parameter increases if too much requirements were made but does not decrease. To ensure "Subgradient" algorithm convergence it is required that the distance to the optimal vector of parameter decreases at each iteration, this means:

21

$$||\mu^{n+1} - \mu^*|| < ||\mu^n - \mu^*|| \tag{21}$$

The proof which follows aims to find a condition on the iteration stepsize ϵ_j^n to ensure the equivalent inequality:

$$||\mu^{n+1} - \mu^*|| - ||\mu^n - \mu^*|| < 0 \tag{22}$$

Assume that g_j^n is the subgradient of variable j at iteration n and that ϵ_j^n is the stepsize used for the following iteration. The left part of the previous equation can be written as:

$$\begin{aligned} ||\mu^{n+1} - \mu^*|| &= \sum_j (\mu_j^n + \epsilon_j^n g_j^n - \mu_j^*)^2 \\ &= \sum_j (\mu_j^n - \mu_j^*)^2 + 2\sum_j \epsilon_j^n g_j^n (\mu_j^n - \mu_j^*) \\ &+ \sum_j (\epsilon_j^n)^2 ||g_j^n||^2 \\ &= ||\mu^n - \mu^*|| + 2\sum_j \epsilon_j^n g_j^n (\mu_j^n - \mu_j^*) \\ &+ \sum_j (\epsilon_j^n)^2 ||g_j^n||^2 \end{aligned} \tag{23}$$

This can be re-written as:

$$\begin{aligned} ||\mu^{n+1} - \mu^*|| - ||\mu^n - \mu^*|| &= 2\sum_j \epsilon_j^n g_j^n (\mu_j^n - \mu_j^*) \\ &+ \sum_j (\epsilon_j^n)^2 ||g_j^n||^2 \end{aligned} \tag{24}$$

As we know that $g_j^n > g_j^*$, we have $-g_j^n \mu_j^* < -g_j^* \mu_j^*$ Moreover $-g_j^* \mu_j^* = q_j(\mu_j^*)$, therefore we have a sufficient condition to ensure the previous inequality:

$$\forall j, \quad 2\epsilon_j^n (g_j^n \mu_j^n - g_j^* \mu_j^*) + (\epsilon_j^n)^2 ||g_j^n||^2 < 0 \tag{25}$$

After factorization by the iteration stepsize ϵ_j^n which is positive, and some basic manipulation it comes the following condition:

$$\forall j, \quad \epsilon_j^n < \frac{q_j(\mu_j^*) - q_j(\mu_j^n)}{||g_j^n||^2} \tag{26}$$

REFERENCES

[1] I. F. Akyildiz, W.-Y. Lee, M. C. Vuran, and S. Mohanty, "NeXt generation/dynamic spectrum access/cognitive radio wireless networks: a survey," *Elsevier, Computer Networks*, vol. 50, no. 13, pp. 2127–2159, Sept. 2006.

[2] S. Mangold, Z. Zhong, K. Challapali, and C.-T. Chou, "Spectrum agile radio: radio resource measurements for opportunistic spectrum usage," *Proceedings of the IEEE Global Telecommunications Conference*, vol. 6, no. 29, pp. 3467–3471, Nov. 2004.

[3] G. Ganesan and Y. G. Li, "Cooperative spectrum sensing in cognitive radio: Part I: two user networks," *IEEE Trans. on Wireless Communications*, vol. 6, pp. 2204–2213, June 2007.

[4] G. Ganesan and Y. G. Li, "Cooperative spectrum sensing in cognitive radio: Part II: multiuser networks," *IEEE Trans. on Wireless Communications*, vol. 6, pp. 2214–2222, June 2007.

[5] S. Boyd and L. Vandenberghe, "Convex Optimization," *New York: Cambridge University Press*, 2004.

[6] W. Yu, W. Rhee, S. Boyd, and J. Cioffi, "Iterative Water-filling for Gaussian Vector Multiple-access Channels," *IEEE Trans. on Information Theory*, vol. 50, no. 1, pp. 145–152, Jan. 2004.

[7] W. Rhee and J. Cioffi, "Increase in Capacity of Multiuser OFDM System using Dynamic Subchannel Allocation," *Proceedings of the IEEE Vehicular Technology Conference*, pp. 1085–1089, May 2000.

[8] M. Chiang and A. Sutivong, "Efficient Optimization of Constrained Nonlinear Resource Allocation," *Proceedings of the IEEE Global Telecommunications Conference*, vol. 7, pp. 3782–3786, Dec. 2003.

[9] C. Saraydar, N. Mandayam, and D. Goodman, "Efficient Power Control via Pricing in Wireless Data Networks," *IEEE Trans. on Communications*, vol. 50, no. 2, pp. 291–303, 2002.

[10] Z. Han and K. Liu, "Noncooperative Power-control Game and Throughput Game over Wireless Networks," *IEEE Trans. on Communications*, vol. 53, no. 10, pp. 1625–1629, 2005.

[11] J. Nash, "Two-person Cooperative Games," *Econometrica*, vol. 21, no. 1, pp. 128–140, 1953.

[12] H. J. M. Peters, "Axiomatic Bargaining Game Theory," *Kluwer Academic Publishers*, 1992.

[13] L. Zhou, "The Nash Bargaining Theory with Non-convex Problems," *Econometrica*, vol. 65, pp. 681–685, May 1997.

[14] Z. Han, Z. Ji, and K. Liu, "Fair Multiuser Channel Allocation for OFDMA Networks using Nash Bargaining Solutions and Coalitions," *IEEE Trans. on Communications*, vol. 53, no. 8, pp. 1366–1376, 2005.

[15] J. Jang and K. B. Lee, "Transmit Power Adaptation for Multiuser OFDM Systems," *IEEE Journal on Selected Areas in Communications*, vol. 21, no. 2, pp. 171–178, Feb. 2003.

[16] W. Rhee and J. M. Cioffi, "Increasing in Capacity of Multiuser OFDM System Using Dynamic Subchannel Allocation," *Proceedings of the IEEE Vehicular Technology Conference*, vol. 2, pp. 1085–1089, May 2000.

[17] C. Y. Wong, R. S. Cheng, K. B. Letaief, and R. D. Murch, "Multiuser OFDM with adaptive subcarrier, bit, and power allocation," *IEEE Journal on Selected Areas in Communications*, vol. 17, pp. 17471758, Oct. 1999.

[18] D. Kivanc, L. Guoqing, and L. Hui, "Computationally efficient bandwidth allocation and power control for OFDMA," *IEEE Trans. on Wireless Communications*, vol. 2, no. 6, pp. 1150–1158, Nov. 2003.

[19] Y. Shoham and A. Gersho, "Efficient Bit Allocation for an Arbitrary Set of Quantizers," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 36, pp. 1445–1453, Sept. 1988.

[20] D. Li and X. Su, "Nonlinear Integer Programming," *Springer, International Series in Operations Research and Management Science*, 2006.

[21] M. Chiang, "Balancing transport and physical Layers in wireless multihop networks: jointly optimal congestion control and power control," *IEEE Journal on Selected Areas in Communications*, vol. 23, no. 1, pp. 104–116, Jan. 2005.

[22] M. Chiang, S. H. Low, A. R. Calderbank, and J. C. Doyle, "Layering as Optimization Decomposition: A Mathematical Theory of Network Architectures," *Proceedings of the IEEE*, vol. 95, no. 1, pp. 255–312, Jan. 2007.

[23] D. P. Bertsekas, "Nonlinear Programming," *Athena Scientific*, 1999.

[24] H. Holma and A. Toskala, "WCDMA for UMTS: Radio Access for Third Generation Mobile Communications, 3rd Edition," *Wiley*, Sept. 2004.

[25] M. Folke and S. Landstrom, "An NS module for simulation of HSDPA," *Technical Report / Lulea University of Technology*, 2006.

24

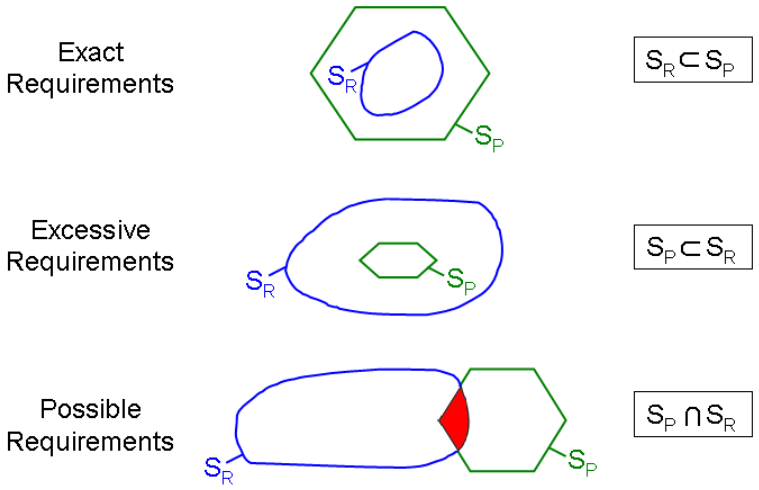

Fig. 1. Optimization at the Base Stations.

	Subgradient	**BS Removal**
To BS	$\widehat{\mathbf{P}}_\mathbf{k}$	$\lambda_k, \widehat{\mathbf{P}}_\mathbf{k}$
To Users	μ_j	$\tilde{\mathbf{P}}_\mathbf{j}, \mathbf{V}$
Nb of Parameters Transmitted	$(N_u + 1) \times N_s$	$3 \times (N_u \times N_s^{(n)}) + (N_s - N_s^{(n)})$
Nb of Iterations	depends on ϵ_j	$\leq N_s$
Notes	optimality, convergence time can be large	quasi-optimal, simple, quick convergence

TABLE I
ALGORITHM COMPARISON

Indeed a proof of this quasi-optimality is obtained by evaluating the difference between the lagrangians obtained by the two algorithms (at last iterations). Thus, Figure 12 proves that when both algorithms convergence the difference is very small. One can effectively notice that only 10% of the simulations have a difference superior or equal to 15%. One can also remark that

25

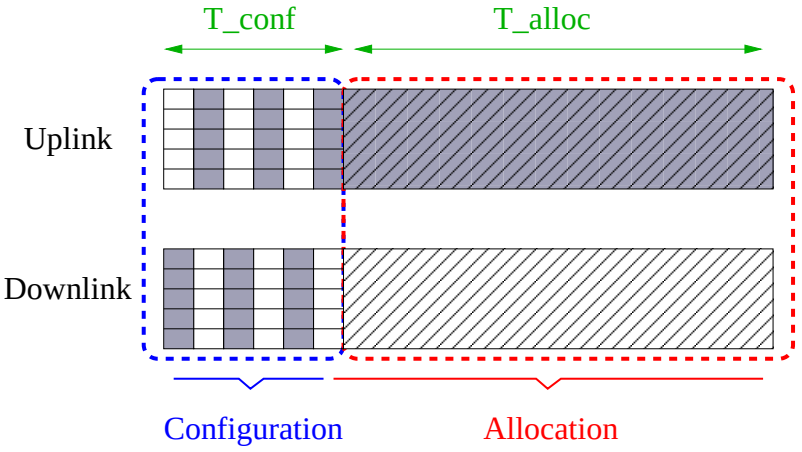

Fig. 2. FDD Frame Structure on each Network.

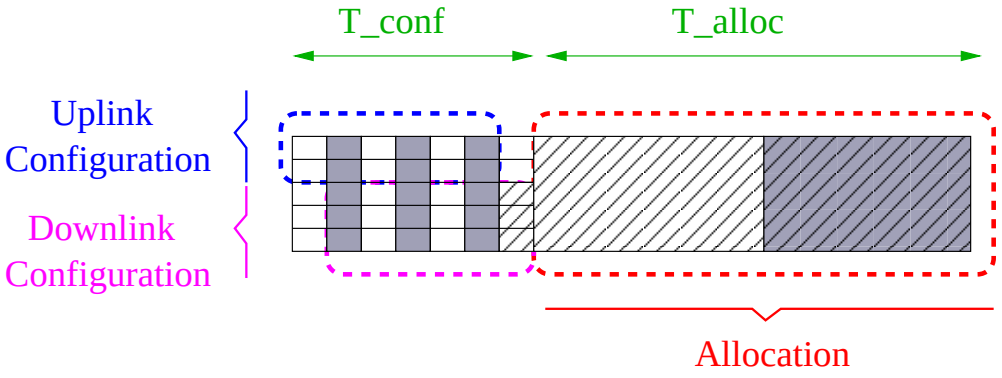

Fig. 3. TDD Frame Structure on each Network.

in the worst case the difference between the two lagrangians do not exceed 30%.

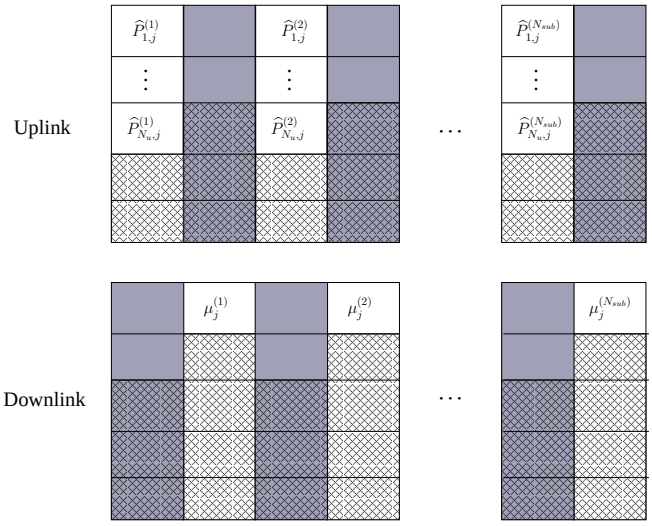

Fig. 4. Configuration Phase of “Subgradient” Algorithm Algorithm for a FDD Frame Structure.

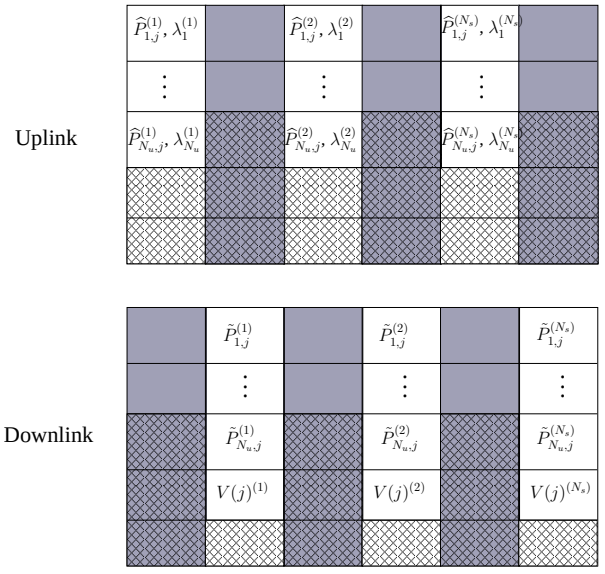

Fig. 5. Configuration Phase of “Base Station Removal” Algorithm for a FDD Frame Structure.

27

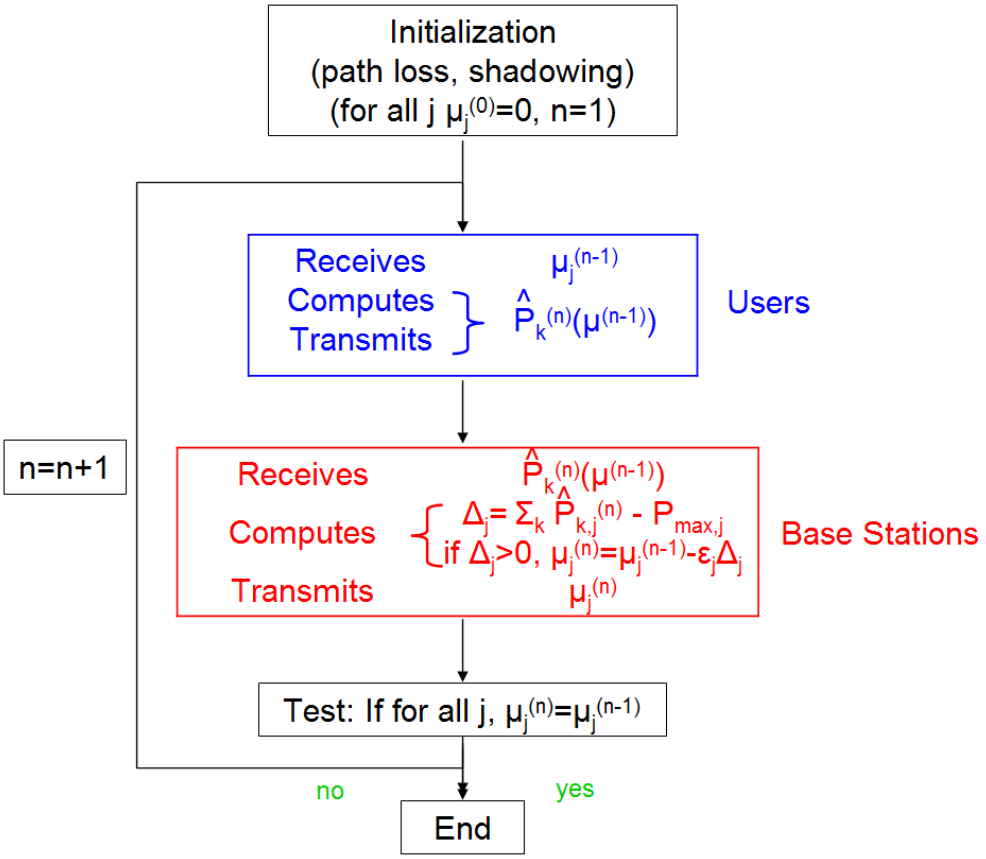

Fig. 6. Flowchart for “Subgradient” algorithm.

28

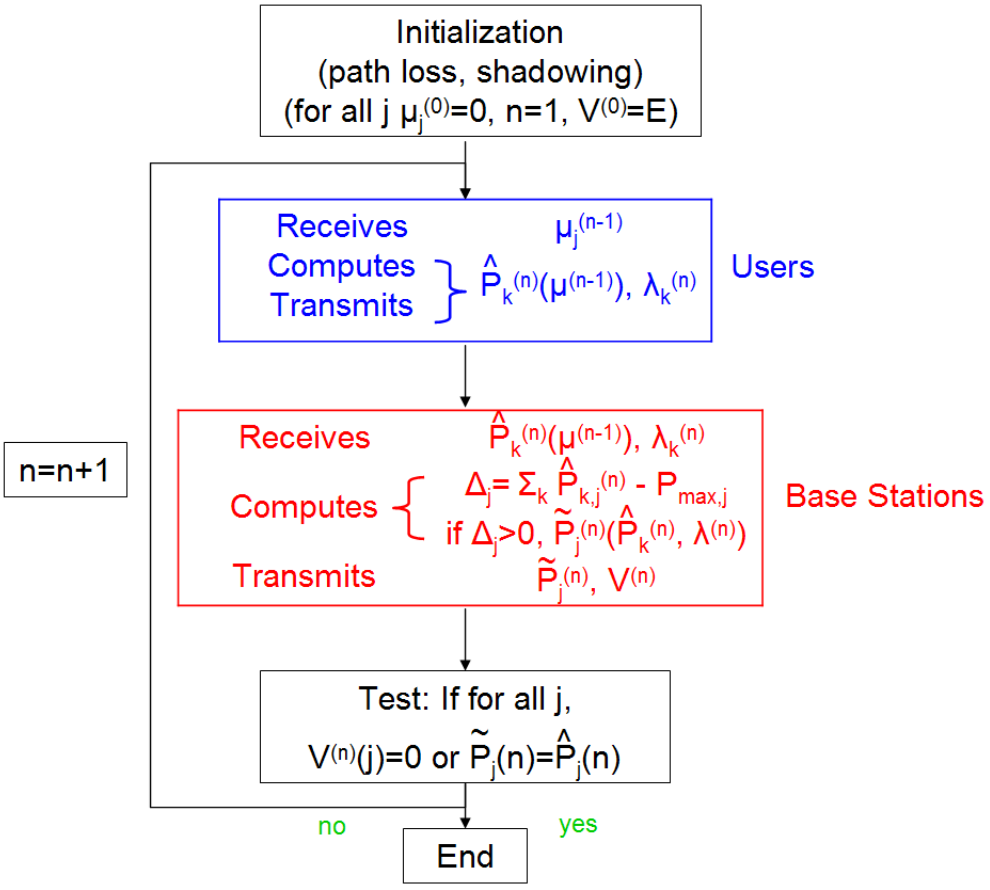

Fig. 7. Flowchart for “Base Station Removal” algorithm.

29

Cell	hexagonal radius = 1000 m
Station 1 [x,y]	[0,0]
Station 2 [x,y]	[0,0]
Station 3 [x,y]	[0,0]
User Distribution	hexagonal around Stations radius = 750 m
User Nb.	10
User Req.	1.024 Mb/s
Monte Carlo Sim	100

TABLE II
SIMULATION PARAMETERS

Carrier frequency	1950 MHz
Bandwidth	4.65 MHz
Thermal Noise	-170.3 dBm
Path loss exponent	3.52
Path loss deviation	4 dB
Transmission power	15 W
Antenna Gain	17 dBi
Intracell Interference	30 dBm
Intercell Interference	-70 dBm

TABLE III
ATTENUATION MODEL PARAMETERS

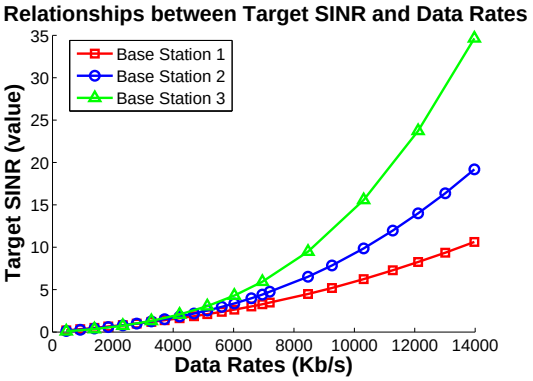

Fig. 8. Relationships between target SINR and data rates.

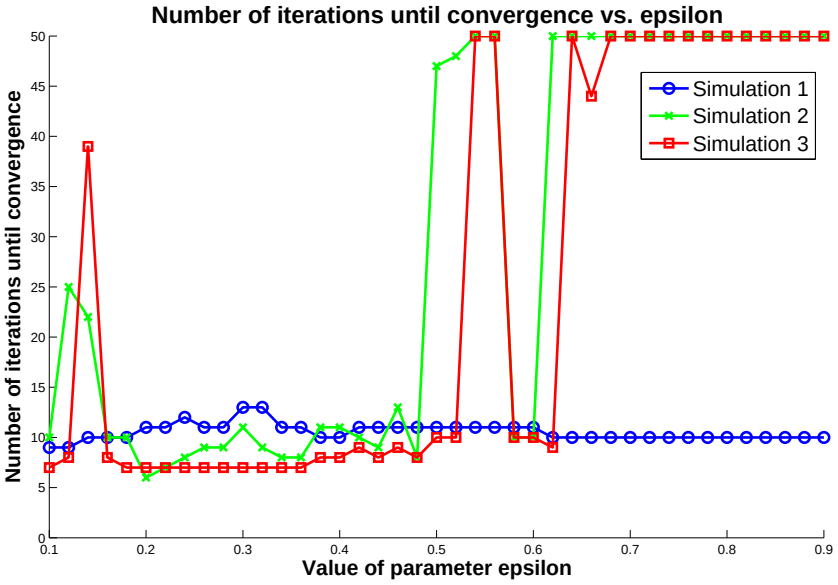

Fig. 9. Convergence time of “Subgradient”

Convergence	Proportion of Simulations
Both	89 %
Algorithm 1 Alone	10 %
Algorithm 2 Alone	1 %

TABLE IV
CONVERGENCE RESULTS.

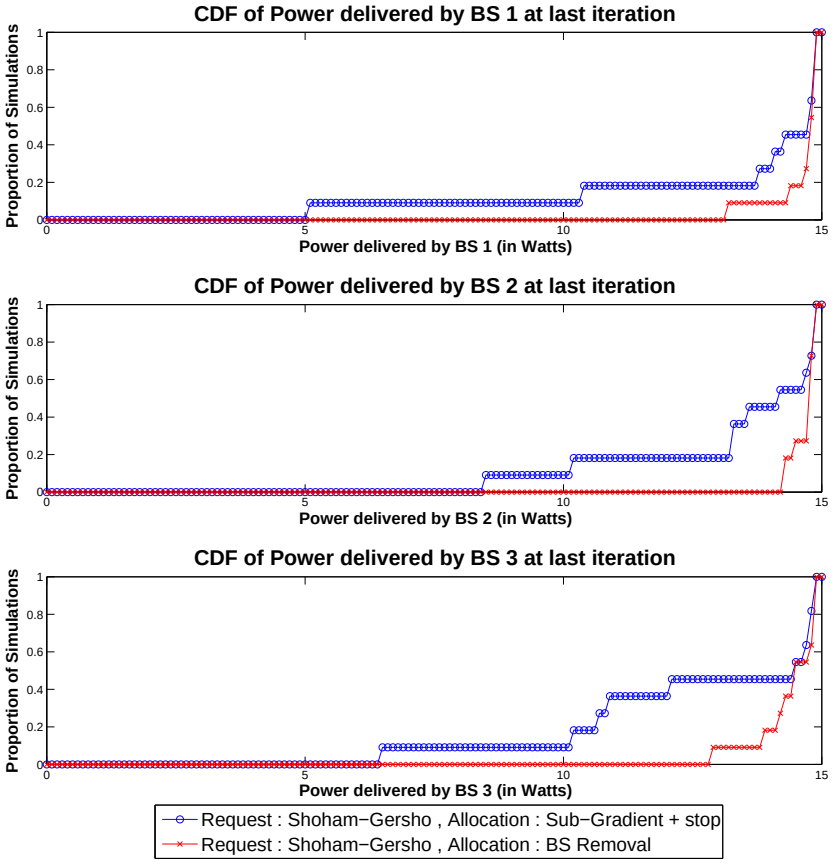

Fig. 10. Cumulative Density Function of Power when only “Subgradient” algorithm converges.

7.3 Annexe 2

Distributed Optimization in Heterogeneous Networks with Proportional Fairness

Christophe Gaie, *Student Member, IEEE,* Mohamad Assaad, *Member, IEEE,*
and Pierre Duhamel, *Fellow, IEEE*

***Abstract*—In this paper we develop a discrete and distributed optimization framework that maximizes the overall spectral efficiency of heterogeneous networks while ensuring a minimum fairness among users. We split the problem into two subproblems: the first one is handled by the users and consists in a transmission by each user of a minimum rate requirement to the base stations and the second is the resource allocation performed by the base stations with respect to the rate constraints generated by the users. The first subproblem is solved using a process where each user tries to adjust its rate requirements to the network capacity. This process is inspired from the TCP Vegas algorithm [1], where the network usage is improved by reducing congestion. Based on the users' rate requirements, the base stations perform a discrete resource optimization in order to allocate power and rate to each user. The rates are selected among a finite set of discrete rates since Adaptive Modulation and Coding (AMC) technique is assumed to be used in each network. The proposed algorithm is compared to an extension of the Proportional Fair algorithm, developed in this paper to allow joint rate and power allocation. Simulation results show that the proposed algorithm provides better performance than the simple PF while ensuring an acceptable fairness among the users.**

***Index Terms*—Discrete Optimization, Resource Allocation, Heterogeneous Systems.**

I. Introduction

During the last decades, many wireless technologies were developed including GSM, GPRS, EDGE, WCDMA, HSPA, WiMAX or LTE. In the meanwhile, many radio frequencies reveal under-utilized as underlined in [2] and [3]. As the needs for new wireless applications continue to grow, the research toward heterogeneous systems has become of particular interest.

In this paper, we deal with the downlink transmission of data from elastic services (such as a downloading of a large file). A relevant problem in this context, is to propose an efficient method to allocate the resources to the users while ensuring a certain fairness among them.

In the state of the art, there are different approaches to allocate resources. The most famous approaches are: 1- Rate Adaptive, which consists in maximizing the total capacity (see [4], [5]), 2- Margin Adaptive, which aims to minimize the overall power usage under minimum user QoS (see [6], [7]). Note that Proportional Fair Scheduling (PFS) is an example of rate adaptive strategy when no power control is used (only rate and timeslot allocation). For more details on Proportional Fair Scheduling see the work of Jalali et al. [8].

Furthermore, many publications in the state of the art deal with centralized approaches based on *continuous* convex optimization which can be solved via Lagrange Theory and water-filling algorithms [9], [10]. The authors assume the existence of a network entity which receives all information of the systems. However, this network structure requires a huge signaling and computational overhead and leads to the denial of this centralized solution.

The problem we study here differs from previous literature in many aspects since it deals with *heterogeneous* networks and *discrete* resource allocation using a *distributed* algorithm. In fact, most of existing wireless systems use Adaptive Modulation and Coding technique where a finite set of Modulation and Coding schemes is available in the system. This AMC technique is tightly coupled with fast power allocation in order to handle fast link adaptation and hence cope with fast radio channel variations (fast fading). This means that the base stations jointly allocate power to users and select a modulation and coding scheme among a discrete set of values (i.e. the bit rates to attribute to users are then selected among a finite set of available discrete rates). This allocation should optimize the overall resources in the whole heterogeneous systems but should be performed *independently* within each system or at least with small cooperation between systems (low computational cost and signaling overhead).

In this paper, the tradeoff between fairness and network throughput in the context of heterogeneous networks with discrete resources is studied. The main contribution of this paper consists in splitting the whole discrete resource optimization problem into two subproblems: the first one performed by the users and the second one by the base stations. A new algorithm is proposed to solve the first subproblem which aims to adjust the rate requirements of the users to the network capacity of the base stations. This process is inspired from the TCP Vegas mechanism algorithm [1], where the authors aim to improve the network usage by reducing congestion. In our proposed algorithm, the sole information exchange required is the transmission of the rate requirements from the users to the base stations. In the second subproblem, the base stations will consider the users' rate requirements as constraints and make the allocation by solving a Discrete Convex Optimization (DCO). These optimizations performed by the base stations use an algorithm inspired from [11] and adapted in this paper to the context of resource allocation with multiple constraints. This algorithm reveals to converge very quickly to the optimal solution on the convex envelop which makes the proposed solution computationally tractable. This novel approach is compared to an extension of the mono cell standard Proportional Fair Scheduler PFS. This extension is developed in this paper in order to enable joint power and

2

discrete rate allocation (not only rate allocation as in the standard PFS).

The remainder of this paper is organized as follows. In Section II, context of the study and problem statement are described. Section III provides description of the suggested solution while Section IV discusses the implementation of the subgradient algorithm. Simulation parameters and results are provided in Section V. Section VI concludes the paper.

II. Problem Statement

We consider a heterogeneous system composed of non interfering base stations (each base station functions on a different bandwidth) with indices $j = [1 \dots N_s]$ and maximal available power $P_{max,j}$. We also assume that the users have indices $k = [1 \dots N_u]$ and seek to obtain a minimum data rate $R_{min,k}$ on the downlink. The users can also use and maintain multiple connections to the different base stations. The connection between the user k and the network j is denoted by (k, j).

Moreover, the base stations use power control and Adaptive Modulation and Coding. Therefore, the user k can only select discrete values of data rate to communicate with the base station j. These discrete values are called "modes" and are of indices $i_{k,j} = [1 \dots I_{k,j}]$. Modes characterize a physical configuration consisting in a suitable modulation type (e.g., BPSK, QPSK, etc.) and a suitable FEC (Forward Error Correction) encoding scheme (e.g. a convolutional code with a given code rate). We also suppose that relationships between powers and data rates are convex.

Our objective is to provide a Mean Fair Allocation adapted to heterogeneous systems using a distributed algorithm. On link (k, j), the instantaneous data rate is denoted as $R^t_{k,j} = R_{k,j}(i_{k,j})$, the channel quality as $g_{k,j}$ and the power as $P^t_{k,j} = P_{k,j}(g^t_{k,j}, i^t_{k,j})$. Note that instantaneous powers and data rates are related by the following relationships: $R^t_{k,j} = f^t_{k,j}(P^t_{k,j})$ (with functions $f_{k,j}$ concave and only defined for discrete modes). Define $T^t_{k,j}$, the average data rate transmitted in the previous timeslots and evaluated through an exponentially weighted low pass filter such that:

$$\forall t, \quad T^t_{k,j} = (1-\beta)T^{t-1}_{k,j} + \beta R^t_{k,j} \tag{1}$$

Note that the mean data rate at the end of slot t is $T^t_{k,j}$. It depends upon the mean data rate at the end of slot $t-1$ (i.e. $T^{t-1}_{k,j}$), the instantaneous data rate (i.e. $R^t_{k,j}$) and the window size $\frac{1}{\beta}$. From all these definitions, the objective to fulfill may be expressed as:

$$\mathbf{I^{t,*}} = \underset{\{i_{k,j}\}}{\arg\max} \sum_{\substack{k=1..N_u \\ j=1..N_s \\ i_{k,j}\in[1\dots I_{k,j}]}}^{N_u} \log\Big(\sum_{j=1}^{N_s} T^t_{k,j}\Big)$$

$$\text{subject to:} \quad \forall j, \quad \sum_{k=1}^{N_u} P_{k,j}(g^t_{k,j}, i^t_{k,j}) \le P_{max,j} \tag{2}$$

III. Proposal

It is quite known that the mono cell Proportional Fair algorithm provides an acceptable trade off between fairness and capacity by maximizing a logarithmic function of the user long term throughput. If one applies the mono cell PFS algorithm in our context, this would result in maximizing the following utility function

$$\sum_{k=1}^{N_u}\sum_{j=1}^{N_s} \log\big(T^t_{k,j}\big)$$

The use of this algorithm will provide then a lower data rate efficiency compared to the optimal solution since

$$\sum_{k=1}^{N_u}\sum_{j=1}^{N_s} \log\big(T^t_{k,j}\big) \le \sum_{k=1}^{N_u} \log\Big(\sum_{j=1}^{N_s} T^t_{k,j}\Big) \tag{3}$$

In other words, solving (2) will improve the performance of the system compared to the mono cell Proportional Fair algorithm. Therefore, in the remainder of this paper, we use this mono cell PF as a basis of comparison. Moreover, to get closer to the optimal solution we propose a new algorithm, inspired from TCP Vegas mechanism [1], that we call "Vegas-DCO Approach".

A. Extended Proportional Fair Resource Allocation

In this section, we propose extension to the PF algorithm in order to take into account the context considered in this paper. In fact, the extended PF algorithm should allow joint power and rate allocation for each user in each system. Each base station can also maintain parallel connections with multiple users (e.g. OFDMA or CDMA system).

Each base station has then to solve the following problem:

$$\mathbf{\tilde{i}^t_j} = \underset{\{i^t_{k,j}\}}{\arg\max} \sum_{\substack{k=1..N_u \\ j=1..N_s \\ i^t_{k,j}\in[1\dots I_{k,j}]}}^{N_u} \frac{R_{k,j}(i^t_{k,j})}{T^{t-1}_{k,j}}$$

$$\text{subject to:} \quad \forall j, \quad \sum_{k=1}^{N_u} P_{k,j}(g^t_{k,j}, i^t_{k,j}) \le P_{max,j} \tag{4}$$

This problem is a discrete resource allocation problem (functions $f_{k,j}$ are both discrete and convex). We propose to solve this problem using a method inspired from Shoham-Gersho Algorithm [11]. This algorithm is a Discrete Convex Optimization (DCO) algorithm, which was proposed in the context of information source coding and enables to reduce computational complexity by restraining the search to a limited number of configurations. We adapt then this algorithm to the context of resource allocation with one or multiple constraints.

B. Vegas-DCO Approach

Solving the problem of fair resource allocation stated in equation (2) is rather difficult, as it aims to grant average average data rates while fulfilling instantaneous power constraints. We split this problem into two subproblems: the first

one handled by the users and the second one by the base stations.

The user subproblem consists in transforming the complicated optimization problem into a simpler one which only depends upon data rates. The idea consists in the fact that the maximum power constraint can be replaced by an ergodic capacity (i.e. max rate) constraint since the max power will determine in somehow the max capacity of the system. Since the channel is variable (fast fading), the instantaneous capacity is variable and the max instantaneous power results in variable max instantaneous capacity constraint. However, the user can estimate/measure the previous system throughput values and determine the average system capacity. The previous system capacity/throughput values are resulted from the resource allocation performed by the base stations (i.e. the second subproblem). Since the fairness constraint is fulfilled in long term, we can use the average capacity in the optimization problem. In other words, the objective function and the constraints of (2) are decomposed into two parts. These two parts are connected using the average system capacity. 1- For the user side: We should maximize the objective function of (2) with respect to the average system capacity which depends on the previous system throughputs in the previous time slots. 2- For the base station side: We should maximize the system capacity with respect to max instantaneous power constraint and minimum rate constraint generated by the users' optimizations.

This decomposition of the problem is explained throughout this section.

1) User optimization subproblem (Vegas): Consider $\widehat{R}^t_{k,j}$, the data rate required by user k to base station j at time t. Consider $\tilde{R}^t_{k,j}$, the data rate granted by station j to user k at time t. Define $\bar{C}^t_j$ as the average capacity of base station j at time t. The user subproblem may be written as:

$$\mathbf{I}^* = \underset{\{i_{k,j}\}}{\arg\max} \sum_{\substack{k=1..N_u \\ j=1..N_s \\ i_{k,j}\in[1...I_{k,j}]}}^{N_u} \log\big(\sum_{j=1}^{N_s} T^t_{k,j}\big)$$

$$\text{subject to:} \quad \forall j, \quad \sum_{k=1}^{N_u} T^t_{k,j} \leq \bar{C}^t_j \tag{5}$$

One can notice that this problem is in somehow similar to TCP mechanism [1], where the authors aim to improve the network usage by reducing congestion. While the Vegas algorithm aims to adjust the arrival of new packets to the size of backlogs to avoid congestion, the solution we propose for problem of equation (5) aims to adapt the user rate requirements to the radio capacity of the base stations. Therefore, in the following, we define the "congestion" at a given base station as exceeding user requirements compared to the base station radio capacity. Note that although there are some similarity between our problem and TCP congestion control problem, our optimization problem is different from TCP (e.g. our utility function $log(\sum T)$ is different from TCP where the utility function is a simple log(T)). This requires us to develop a new algorithm/framework.

Let us explain theoretically how the problem of equation (5) may be solved. First, the lagrangian of equation (5) is expressed, to obtain the operation performed by the users:

$$\begin{aligned}\mathcal{L}(\mathbf{R},\lambda) &= -\sum_{k=1}^{N_u}\log\big(\sum_{j=1}^{N_s} T^t_{k,j}\big) \\ &+ \sum_{j=1}^{N_s}\lambda_j\big(\sum_{k=1}^{N_u} T^t_{k,j} - \bar{C}^t_j\big)\end{aligned} \tag{6}$$

The optimal value of previous lagrangian is $\mathcal{L}(\mathbf{T}^*,\lambda^*) = \min_{\mathbf{T}}\max_{\lambda}\mathcal{L}(\mathbf{T},\lambda)$.

This problem may be solved using a subgradient algorithm based on successive updates of the network parameter λ_j using a sufficiently small stepsize ϵ. Differentiating the lagrangian of equation (6) leads to the following relationship:

$$\lambda^{t+1}_j = [\lambda^{t+1}_j + \epsilon\underbrace{\big(\sum_{k=1}^{N_u} T^t_{k,j} - \bar{C}^t_j\big)}_{\delta_j}]^+ \tag{7}$$

The instantaneous capacity of base station j may be expressed as $C^t_j = \sum_{k=1}^{N_u}\tilde{R}^t_{k,j}$. Then, the average capacity of base station j expresses as $\bar{C}^t_j = (1-\beta)\bar{C}^{t-1}_j + \beta C^t_j$.

At the equilibrium, we have:

$$\frac{\partial\mathcal{L}}{\partial T_{k,j}}(\mathbf{T},\lambda) = \frac{1}{T_{k,j}} - \lambda_j = 0 \tag{8}$$

Replacing into equation (8) λ_j by its value at the equilibrium, it comes after some trivial transformations:

$$T^{t+1}_{k,j} = \frac{T^t_{k,j}}{1 + \epsilon T^t_{k,j}\big(\sum_{k=1}^{N_u} T^t_{k,j} - \bar{C}^t_j\big)} \tag{9}$$

As ϵ is very small, a Taylor development leads to:

$$T^{t+1}_{k,j} = T^t_{k,j}\big(1 - \epsilon T^t_{k,j}\big(\sum_{k=1}^{N_u} T^t_{k,j} - \bar{C}^t_j\big)\big) \tag{10}$$

Express the capacity as $\bar{C}^t_j = \sum_k \bar{C}^t_{k,j}$, whose estimation is $\tilde{T}^t_j = sum_k\tilde{T}^t_{k,j}$ and is based on the actual granted data rates $\tilde{R}^t_{k,j}$. Define the congestion measure for user k on network j as $c^t_{k,j} = \widehat{T}^t_{k,j} - \tilde{T}^t_{k,j}$. Let $c^t_k = \sum_{j=1}^{N_s} c^t_{k,j}$ denote the total congestion measured by user k. Set the stepsize as $\epsilon = \frac{1}{T^t_{k,j}\|\sum_j c^t_{k,j}\|}$, it comes:

$$T^{t+1}_{k,j} = T^t_{k,j}\big(1 - \frac{\big(\sum_{k=1}^{N_u}(T^t_{k,j} - \bar{C}^t_{k,j})\big)}{\|\sum_j c^t_{k,j}\|}\big) \tag{11}$$

As user k has no idea about other user data rate, its decision is simplified as:

$$T^{t+1}_{k,j} = T^t_{k,j}\big(1 - \frac{c^t_{k,j}}{c^t_k}\big) \tag{12}$$

4

We can now define our algorithm which is able to simulate the transmission of the network parameter λ_j by monitoring the perceived congestion. Note that this congestion is positive if the requirements exceed the allocation (there are too many requirements on the link), null if the requirements are equal to the allocation (perfect allocation for the given link), negative if the requirements are lower than the allocation (there are too few requirements on the link). Introducing a parameter α which reflects a tolerance between the current and the target data rate, we propose to implement for each user k the following algorithm:

$$\forall k, \forall j, \quad \widehat{T}_{k,j}^{t+1} = \begin{cases} \widehat{T}_{k,j}^{t}(1+\frac{c_{k,j}^{t}}{c_k^t}\epsilon) & \text{if} \quad \frac{c_k^t}{\widetilde{T}_{k,j}^{t}} < 0 \\ \widehat{T}_{k,j}^{t}(1-\frac{c_{k,j}^{t}}{c_k^t}\epsilon) & \text{if} \quad \frac{c_k^t}{\widetilde{T}_{k,j}^{t}} > \alpha \\ \widehat{T}_{k,j}^{t} & \text{otherwise} \end{cases} \tag{13}$$

A very useful property of the algorithm is that there is no need to transmit congestion information between the users and the base stations. Indeed, if a scheduler ensuring minimum requirements is chosen to schedule the resource allocations at the base stations, then each user is able to determine whether there is congestion at each base station. This ability comes from the fact that each user can detect if a base station can handle with rate requirements (then there is no congestion) or not (then there is congestion). Moreover, the users can determine the amount of congestion by comparison between the requirements and the allocations.

2) base station optimization algorithm: Let now describe the base station part of the Wireless Vegas Resource Allocation algorithm. The algorithm implemented at each base station aims to ensure user minimum requirements while maximizing data rate granted for the remainder power. However, the given base station may suffer from too high requirements. Therefore, mechanisms have to be implemented to grant resources fairly and close to user requirements.

With the proposed algorithm, each base station begins by trying to solve the following problem using an algorithm inspired from [11]:

$$\tilde{\mathbf{i}}_{\mathbf{j}}^{\mathbf{t}} = \arg\max_{\{i_{k,j}^t\}} \sum_k R_{k,j}(i_{k,j}^t)$$

subject to:

$$\begin{cases} \forall k, \quad R_{k,j}(i_{k,j}^t) \geq \widehat{R}_{k,j}^t \\ \sum_{k=1}^{N_u} P_{k,j}(i_{k,j}^t) \leq P_{max,j} \end{cases} \tag{14}$$

If a solution is found, then power is granted according to this solution $\tilde{R}_{k,j}^t = R_{k,j}(\tilde{i}_{k,j}^t)$. Otherwise, the base station decreases every user of one mode and tries to solve problem of equation (14) with a new data rate constraint. This operation is repeated until a solution is found.

IV. Implementation

In this section, the implementation of the two proposed algorithms is discussed in details.

For the Extended mono cell Proportional Fair Resource Allocation algorithm, there is no need to transmit any information between the users and the base stations. This is a very rare and interesting property in the case of distributed optimization. Indeed, each signaling overhead reduces the actual data rate. Moreover, the Extended Proportional Fair Resource Allocation is computationally tractable as it reduces to a simple and fast optimization performed at the base stations (the algorithm inspired from Shoham Gersho algorithm is very efficient).

The Extended mono cell Proportional Fair Resource Allocation functions as follows:

- Initialize algorithm (t=0, $T_{k,j}^0 = 1$)
  - Compute Path Loss
  - Compute Shadowing
  - t=1
- For $j = 1 : N_s$ (for each base station)
  - Compute for all modes of each user k the ratio $r_{k,j}^t(i_{k,j}) = \frac{R_{k,j}(i_{k,j}^t)}{T_{k,j}^{t-1}}$
  - Each base station j, computes $\tilde{\mathbf{i}}_{\mathbf{j}}^{(t)}(r_{k,j}^t(i_{k,j}))$ of equation (4) using a DCO algorithm inspired from [11].
- Update: $t = t + 1$

At each timeslot, the Vegas-DCO approach, requires that each user computes and transmits a single parameter of rate requirements. This signaling overhead is tractable and is counterbalanced by increased rate performances. As for previous algorithm, the Wireless Vegas Resource Allocation consists in a simple comparison at the users, which is completed at the base station side by a fast optimization.

The Vegas-DCO approach can be described as follows:

- Initialize algorithm (t=0)
  - Compute Path Loss
  - Compute Shadowing
  - t=1
- For $k = 1 : N_u$ (for each user)
  - If $t == 1$ then requirements of user k are obtained using a DCO algorithm inspired from [11]
  - Otherwise
    - For each base station j, $c_{k,j}^t = \widehat{R}_{k,j}^t - \tilde{R}_{k,j}^t$
    - Compute $c_k^t = \sum_{j=1}^{N_s} c_{k,j}^t$
    - Compute the ratio $\Delta_{k,j}^t = \frac{c_k^t}{\tilde{R}_{k,j}^t}$
    - If $\Delta_{k,j}^t < 0$ then $\widehat{R}_{k,j}^t = \widehat{R}_{k,j}^{t-1}(1 + \frac{c_{k,j}^{t-1}}{c_k^{t-1}}\epsilon)$
    - If $\Delta_{k,j}^t > \alpha$ then $\widehat{R}_{k,j}^t = \widehat{R}_{k,j}^{t-1}(1 - \frac{c_{k,j}^{t-1}}{c_k^{t-1}}\epsilon)$
    - Otherwise $\widehat{R}_{k,j}^t = \widehat{R}_{k,j}^{t-1}$
- For $j = 1 : N_s$ (for each base station)
- While no solution exists
  - Each base station j, computes $\tilde{\mathbf{i}}_{\mathbf{j}}^{(t)}(\widehat{R}_{k,j}^t)$ of equation (14) using a DCO algorithm inspired from [11].
  - If a solution is found then data rate is allocated
  - Otherwise decrease all users of one mode
- Update: $t = t + 1$

V. Simulations

A. Simulation Parameters and Radio Conditions

We consider three base stations each one uses a given bandwidth. The users are in a short or medium range to the base stations, which are located at the same place. Detailed parameters are displayed in Table I.

Cell	hexagonal radius = 1000 m
Station 1, 2, 3 [x,y]	[0,0]
User Distribution	hexagonal around Stations radius = 750 m
User Nb.	20
Monte Carlo Sim	100

TABLE I
Simulation Parameters

Simulations performed assume that radio conditions between the users and the base stations are obtained from pedestrian A model of [12] with parameters of Table II.

Carrier frequency	1950 MHz
Bandwidth	4.65 MHz
Thermal Noise	-170.3 dBm
Path loss exponent	3.52
Path loss deviation	4 dB
Transmission power	20 W
Antenna Gain	17 dBi
Intracell Interference	30 dBm
Intercell Interference	-70 dBm

TABLE II
Attenuation Model Parameters

The Signal to Interference plus Noise ratio is given by the following expression:

$$SINR_{k,j} = \frac{\|g_{k,j}\|^2 P_{k,j}}{I_{inter} + I_{intra} + \sigma_n^2} \quad (15)$$

Figure 1 displays the relationships between target SINR and data rates for the considered base stations [13].

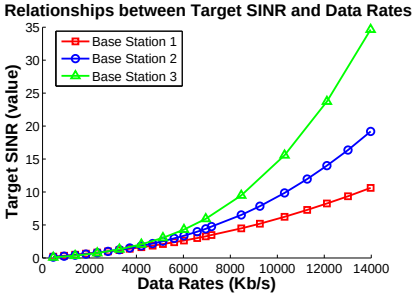

Fig. 1. Relationships between target SINR and data rates.

B. Simulation Results

To assess the algorithm efficiency, the simulations performed have to compare both the data rate obtained by the users and the fairness between the users.

The two following figures show that the Vegas-DCO approach provides more data rate than the Extended Proportional Fair Resource Allocation algorithm. Indeed, Figure 2 shows that 30% of the users get more than 4 Mb/s with the Wireless Vegas Resource Allocation algorithm while only 10% get the same data rate with the Extended Proportional Fair Resource Allocation algorithm. Moreover, one can see in Figure 3 that the Vegas-DCO algorithm increases the data rate of the three users in the best conditions of respectively (30%, 58% and 81%).

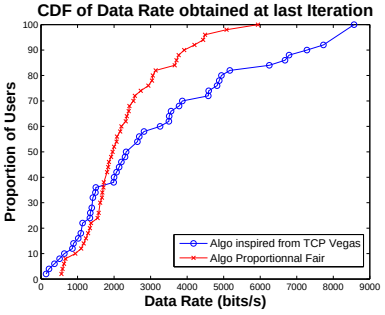

Fig. 2. CDF of Data Rate obtained by the Users.

Concerning the users with the worst channel conditions (i.e. users 1 and 2), the Vegas-DCO algorithm does not impact the their data rates. Indeed, in Figure 3 one can remark that the normalized data rate of user 2 is the same for both approaches. The bit rate of user 1 changes slightly from 0.55 for the Extended Proportional Fair Resource Allocation algorithm to 0.45 for the Vegas-DCO algorithm. This explains why the two CDF nearly overlap at low data rates in Figure 2.

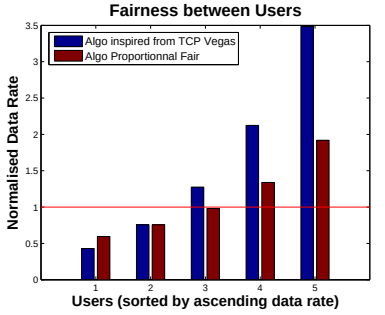

Fig. 3. Fairness between the Users.

VI. Conlusion

In this paper, a distributed discrete resource optimization method adapted to heterogeneous systems was proposed to

6

solve the problem of long term fair resource allocation. we split the optimization problem into two subproblems. The first subproblem consists in adjusting the rate requirements of the users to the network capacity of the base stations. The second subproblem consists in performing a Discrete Resource Optimization (DCO) by the base stations in order to make joint power and rate allocation with respect to the rate requirements generated by the first subproblem. Our framework is compared to an extension of the proportional fair algorithm developed also in the paper in order to allow joint rate and power allocation. Simulation results show that the Vegas-DCO approach has better performance than the extended PF algorithm and ensures both resource efficiency and fairness between the users.

REFERENCES

[1] S. Low, L. Peterson, and L. Wang, "Understanding TCP vegas: a duality model," *SIGMETRICS Perform. Eval. Rev.*, vol. 29, no. 1, pp. 226–235, 2001.

[2] I. F. Akyildiz, W.-Y. Lee, M. C. Vuran, and S. Mohanty, "NeXt generation/dynamic spectrum access/cognitive radio wireless networks: a survey," *Elsevier, Computer Networks*, vol. 50, no. 13, pp. 2127–2159, Sept. 2006.

[3] G. Ganesan and Y. G. Li, "Cooperative spectrum sensing in cognitive radio: Part I: two user networks," *IEEE Trans. on Wireless Communications*, vol. 6, pp. 2204–2213, June 2007.

[4] J. Jang and K. B. Lee, "Transmit Power Adaptation for Multiuser OFDM Systems," *IEEE Journal on Selected Areas in Communications*, vol. 21, no. 2, pp. 171–178, Feb. 2003.

[5] W. Rhee and J. M. Cioffi, "Increasing in Capacity of Multiuser OFDM System Using Dynamic Subchannel Allocation," *Proceedings of the IEEE Vehicular Technology Conference*, vol. 2, pp. 1085–1089, May 2000.

[6] C. Y. Wong, R. S. Cheng, K. B. Letaief, and R. D. Murch, "Multiuser OFDM with adaptive subcarrier, bit, and power allocation," *IEEE Journal on Selected Areas in Communications*, vol. 17, pp. 17471758, Oct. 1999.

[7] D. Kivanc, L. Guoqing, and L. Hui, "Computationally efficient bandwidth allocation and power control for OFDMA," *IEEE Trans. on Wireless Communications*, vol. 2, no. 6, pp. 1150–1158, Nov. 2003.

[8] A. Jalali, R. Padovani, and R. Pankaj, "Data throughput of CDMA-HDR a high efficiency-high data rate personal communication wireless system," *European Transactions on Telecommunications*, vol. 3, pp. 1854–1858, May 2000.

[9] S. Boyd and L. Vandenberghe, "Convex Optimization," *New York: Cambridge University Press*, 2004.

[10] W. Yu, W. Rhee, S. Boyd, and J. Cioffi, "Iterative Water-filling for Gaussian Vector Multiple-access Channels," *IEEE Trans. on Information Theory*, vol. 50, no. 1, pp. 145–152, Jan. 2004.

[11] Y. Shoham and A. Gersho, "Efficient Bit Allocation for an Arbitrary Set of Quantizers," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 36, pp. 1445–1453, Sept. 1988.

[12] H. Holma and A. Toskala, "WCDMA for UMTS: Radio Access for Third Generation Mobile Communications, 3rd Edition," *Wiley*, Sept. 2004.

[13] M. Folke and S. Landstrom, "An NS module for simulation of HSDPA," *Technical Report / Lulea University of Technology*, 2006.

7.4 Annexe 3

DISTRIBUTED DISCRETE RESOURCE OPTIMIZATION IN HETEROGENEOUS NETWORKS

*Christophe Gaie**†, *Mohamad Assaad*†, *Markus Muck**, *Pierre Duhamel*†

*Motorola Labs, Parc Les Algorithmes, Saint Aubin 91193 Gif sur Yvette, France
†SUPELEC, 3 rue Joliot Curie, 91192 Gif sur Yvette, France

ABSTRACT

Nowadays, the emergence of many radio technolgies has increased the research interest towards Radio Resources Management (RRM) in heterogeneous systems. In this context, we seek a distributed scheme for discrete resource allocation. The algorithm proposed should reduce computation and signalling overhead, compared to centralized optimization or distributed solutions based on Game Theory.

Therefore, the solution proposed here consists in splitting the resource allocation problem into two parts. First, Users and Base Stations negotiate a mean allocation using a Multi-system Minimum Mean Rate Scheduling (M^3RS) algorithm presented here. Then, Base Stations allocate power independantly and instantaneously in order to cope with changing radio conditions. This problem is widely known in literature and is not developped here.

1. INTRODUCTION

The objective of this paper is to provide solutions to perform Radio Resource Management (RRM) of multiple Users requiring a minimum data rate in a heterogeneous system composed of multiple networks. To fulfill this objective, two approaches are usually investigated: 1- maximizing total capacity, 2- minimizing overall power usage. Both of these approaches are implemented with respect to user Quality of Service (QoS) constraints. Minimizing power consumption in heterogeneous systems is more attractive for operators since it enables to minimize network usage cost, reduce interference in other cells and therefore increase the number of satisfied Users. Thus, this paper deals with minimizing overall power usage under user minimal data rate and maximum power at Base Stations. To do so, we consider here that Users can use and maintain simultaneously multiple connections with different Base Stations of non-interfering systems (i.e. multi-radio capability) and can re-assemble data from different sources (i.e. multihoming capability).

This work was performed partially in the project E3, which has received research funding from the European Community's Seventh Framework programme. This paper reflects only the authors' views and the Community is not liable for any use that may be made of the information contained therein. The contributions of colleagues from E3 consortia are hereby acknowledged.

In literature, the problem of RRM in single and multi-radio context has attracted much research interest. Many studies propose a centralized approach based on continuous convex optimization which can be solved via Lagrange Theory and water-filling algorithm [1], [2] and [3]. In order to apply water-filling, they assume that there is a network entity which has a complete knowledge of all the systems. These assumptions require a huge signalling and computational cost which makes these solutions not suitable to implement in practice.

To overcome this problem, non-cooperative game theory has been used to analyze wireless networks [4], [5] and [6]. However, distributed solutions based on non-cooperative game models result in a suboptimal solution of the resource optimization problem. Therefore, cooperative game models, where players coordinate to achieve a mutually desirable solution, have been studied in [7], [8] and [9] for analysis of networks and spectrum sharing. However, cooperative game theory requires much signalling overhead due to the coordination between the players which is contradictory with the initial objective.

In this paper we propose a distributed solution to solve the resource optimization problem in heterogeneous multi-radio systems. The proposed solution is based on discrete convex optimization and requires reduced signalling overhead. Simulation results presented in Section 4 show that the solution obtained converges rapidly and efficiently. The idea consists in splitting the RRM problem into two sub-problems. First, Users seek how to obtain their minimum mean data rate among available Base Stations. This process, which is presented here, is referred to as Multiradio Minimum Mean Rate Scheduling (M^3RS). By determining a minimum average user rate in each system, the overall optimization can then be decomposed using the optimization decomposition theory [10] into N_s sub-optimization problems, where N_s is the number of various systems. Therefore, each Base Station can optimize independently its own resources while achieving user minimum data rate determined by our proposed M^3RS. Note that each of the N_s sub-optimization problems is merely a mono-system mono-cell resource optimization problem. This optimization is widely studied in literature and the solution is well known and specific to each system (i.e. CDMA, OFDMA, TDMA, etc.). Consequently, this paper focuses on the presen-

tation of the proposed M^3RS algorithm which enables the decomposition of the overall complicated multisystem optimization into N_s mono-system optimization problems.

The remainder of this paper is organized as follows. In Section 2, context of the study and problem statement are described. Section 3 provides description of the solution proposed in this paper. Simulation parameters and results are provided in Section 4. Section 5 concludes the paper.

2. PROBLEM STATEMENT

In this paper, we consider an independent heterogeneous system composed of non interfering Base Stations (different spectrum used) with indices $j = [1 \dots N_s]$ and maximal available power $P_{max,j}$. We also assume that Users have indices $k = [1 \dots N_u]$ and seek to obtain a minimum data rate $R_{min,k}$ on the downlink. Users can also use and maintain multiple connections to different Base Stations. The connection between User k and network j is denoted by (k, j).

Moreover, Base Stations use power control and Adaptive Modulation and Coding. Therefore, User k can only select discrete values of data rate to communicate with Base Station j. These discrete values are called "modes" and are of indices $i_{k,j} = [1 \dots I_{k,j}]$. Modes characterize a physical configuration consisting in a suitable modulation type (e.g., BPSK, QPSK, etc.) and a suitable FEC encoding scheme (e.g. a convolutional code with a given code rate). We also suppose that relationships between powers and data rates are convex.

In this context, our objective is to minimize overall mean power usage, while ensuring a minimum mean data rate to every User and respecting maximum available power at Base Stations. If on link (k, j), $R_{k,j}(i_{k,j})$ denotes the data rate, $g_{k,j}$ the channel quality and $P_{k,j}(g_{k,j}, i_{k,j})$ the power, the problem consists in obtaining $\mathbf{I}^* = \{i^*_{k,j}\}$, the matrix of modes to use, such that:

$$\mathbf{I}^* = \underset{\{i_{k,j}\}}{\arg\min} \sum_{\substack{k=1..N_u \\ j=1..N_s \\ i_{k,j}\in[1\dots I_{k,j}]}}^{N_u} \sum_{j=1}^{N_s} P_{k,j}(g_{k,j}, i_{k,j})$$

subject to:

$$\begin{cases} \forall k, & \sum_{j=1}^{N_s} R_{k,j}(i_{k,j}) \geq R_{min,k} \\ \forall j, & \sum_{k=1}^{N_u} P_{k,j}(g_{k,j}, i_{k,j}) \leq P_{max,j} \end{cases} \quad (1)$$

For convenience, the subscript $g_{k,j}$ will be omitted when expressing $P_{k,j}$.

3. PROPOSAL

As mentioned in the introduction, centralized solutions for the problem in (1) require high computational complexity and signalling overhead. In this paper, we propose a distributed approach to solve this problem at low cost (low signalling overhead). Any distributed approach called to solve the resource optimization problem of (1) should provide an answer to the following problem: how Base Stations of different systems will optimize resources of all systems if they do not have knowledge of Users' channel conditions in other systems?

The main idea of our proposal consists in using the fact that Users know their channel conditions in different systems. Consequently, Users and Base Stations can proceed to successive negotiations in order to determine the *average* minimum bit rate that each Base Station should allocate to each User. The algorithm proposed in this paper to achieve this task, is referred to as Multisystem Minimum Mean Rate Scheduling (M^3RS). Results afforded by this scheduler are provided in Section 4 and show that the successive negotiation phase converges quickly to the optimal solution (few ms). It is important to mention that M^3RS is used by Users and Base Stations every few seconds in order to reconfigure the overall network (i.e. to determine how Users will connect to systems in the following few seconds). Once the minimum average bit rate of each User in each Base Station is obtained, the optimization problem of (1) can be simplified using the decomposition theory [10] into the following N_s subproblems:

$$\mathbf{i_j} = \underset{\{i_{k,j}\}}{\arg\min} \sum_{k=1}^{N_u} P_{k,j}(g_{k,j}, i_{k,j})$$

subject to:

$$\begin{cases} \forall k, & R_{k,j}(i_{k,j}) \geq R^+_{k,j} \\ & \sum_{k=1}^{N_u} P_{k,j}(g_{k,j}, i_{k,j}) \leq P_{max,j} \end{cases} \quad (2)$$

Each of these optimization problems can be handled independently by each Base Station. In other words, the problem now is to determine how each Base Station will allocate instantaneously the resources (in order to cope with the channel fast fading) to Users in such a way that minimum average bit rate is achieved for each User and that the total power does not exceed a maximum value. In literature, this problem has been widely addressed and the solutions are known and specific to each technology. In general, a gradient based algorithm can be used to solve the subproblem of equation (2). Due to space limit, details on this algorithm are not given in this paper. Readers can refer to [11] for more details. In the remainder of this paper, we focus on the main contribution of this paper i.e. how to obtain the average minimum bit rate of Users in each system using the proposed M^3RS algorithm.

3.1. M^3RS Algorithm

M^3RS is based on successive negotiations between Users and Base Stations. Each negotiation phase is called "iteration". It is clear that the convergence of the algorithm depends upon the number of iterations or negotiation phases. Each iteration n is composed of the two following steps.

The first step of an iteration is denoted here as User Requirements. The idea is to minimize the overall power of all the systems. Each User will try to minimize "independently" the total power required to all systems. Then, Users transmit to Base Stations a vector of average minimum rate requirements $\widehat{\mathbf{R}}_{\mathbf{k}}^{(\mathbf{n})}$ (equivalently $\widehat{\mathbf{i}}_{\mathbf{k}}^{(\mathbf{n})}$).

The second step of an iteration is denoted here as Network Maximal Power Computation and Feedback. In this step, each Base Station j gets requirements from Users. If the total power needed to satisfy the rate requirements of Users does not exceed the maximum available power of Base Station j, then Base Station j can deal with these requirements. Otherwise, Base Station j computes an efficient configuration (under a process described in Section 3.3) and sends to Users the maximum data rate they can use via a feedback vector $\tilde{\mathbf{R}}_{\mathbf{j}}^{(\mathbf{n})}$ (equivalently $\tilde{\mathbf{i}}_{\mathbf{j}}^{(\mathbf{n})}$).

This algorithm converges to a stable solution (configuration) within N_s iterations where N_s is the number of different systems. A configuration is stable if the total power needed to satisfy the rate requirements of all Users does not exceed the total available power of this Base Station. Since the number of systems N_s is small (no more than 3 or 4) the convergence of our algorithm is relatively fast (few ms). Once the User and Base Station configuration is finished (convergence to a stable solution) each Base Station can start its own fast resource allocation (e.g. fast Scheduling).

3.2. User Configuration

The first step of an iteration consists in researching $\widehat{\mathbf{i}}_{\mathbf{k}}^{(\mathbf{n})}$ the vector of requirements of User k on Base Stations $j = [1 \dots N_s]$ at iteration n, which minimizes power required, while ensuring minimum data rate of Users and not exceeding maximal available power at Base Stations. With continuous values, as relationships between powers and data rates are supposed convex, Lagrange Theory would enable to obtain the optimal solution. In our case, we seek to obtain discrete modes, by solving:

$$\widehat{\mathbf{i}}_{\mathbf{k}}^{(\mathbf{n})} = \underset{\substack{\{i_{k,j}^{(n)}\} \quad j=1..N_s \\ i_{k,j}^{(n)} \in [1 \dots I_{k,j}]}}{\arg\min} \sum_{j=1}^{N_s} P_{k,j}(i_{k,j}^{(n)})$$

subject to :

$$\begin{cases} \sum_{j=1}^{N_s} R_{k,j}(i_{k,j}^{(n)}) \geq R_{min,k} \\ \forall j \text{ not in overload}, \quad P_{k,j}(i_{k,j}^{(n)}) \leq P_{max,j} \\ \forall j \text{ in overload}, \quad P_{k,j}(i_{k,j}^{(n)}) \leq P_{k,j}(i_{k,j}^{(n-1)}) \end{cases} \tag{3}$$

with Base Stations initially granting no resource. To solve this problem, it is possible either to proceed via an exhaustive search (which has a high computational cost since modes can be numerous) or to solve it via Discrete Convex Optimization (DCO). Therefore, we transform the problem of equation (3) so as to use Shoham-Gersho algorithm [12], which gives us the best allocation on the convex envelope and is much more efficient than exhaustive search when available modes and Base Stations are numerous.

3.3. Network Maximal Power Computation and Feedback

Once Base Stations have gathered the rate requirements of Users, they allow the allocation if:

$$\sum_{k \in E_j} P_{k,j}(\widehat{i}_{k,j}^{(n)}) \leq P_{max,j} \tag{4}$$

where E_j is the space of Users having sent requirements to Base Station j. If the demanded power exceeds the maximum available power, the given Base Station evaluates the rates that it can allocate to Users, based on available resources. In this section, two algorithms are presented, each proposing a different way of computing the feedback vector $\tilde{\mathbf{i}}_{\mathbf{j}}^{(\mathbf{n})}$ based on data rate requirements, channel quality of links and available power.

Recall that, the algorithms proposed here do not aim to allocate resources instantaneously but to compute a feedback vector to Users which will then be able to improve their requirements (at the following iteration).

3.3.1. Algorithm 1 : Full Max-Rate

At Base Station j, when too much power is required, without knowledge of User requirements on other Base Stations, the best behaviour available consists in finding the feedback vector which will maximize the cumulated data rate:

$$\tilde{\mathbf{i}}_{\mathbf{j}}^{(\mathbf{n})} = \underset{\{i_{k,j}^{(n)} \leq \widehat{i}_{k,j}^{(n)}\}_{k \in E_j}}{\arg\max} \sum_{k \in E_j} R_{k,j}(i_{k,j}^{(n)})$$

subject to:

$$\forall j, \quad \sum_{k \in E_j} P_{k,j}(i_{k,j}^{(n)}) \leq P_{max,j} \tag{5}$$

As for DCO requirements, this problem can be solved using Shoham-Gersho algorithm. Even though this algorithm maximizes the sum rate in system j, a minimum of fairness between Users is achieved in the system. Indeed, Users requiring data rate on Base Station j have good radio conditions on it. Since the rate is a convex function of the power, this algorithm will not result in a huge gap of fairness between connected Users. Note also that Users whose requirements are not satisfied by Base Station j will require more data rate to other Base Stations at the following iteration.

3.3.2. Algorithm 2 : Highest Marginal Power to Data Rate Decrease

The second proposed algorithm consists in decreasing recursively of one mode the User of worst power efficiency. Mathematically speaking, this consists in decreasing of one mode

the User of maximum marginal power to data rate while power required is greater than power available. Thus, the algorithm is the following:

* Initialize for each User k, $\tilde{i}_{k,j}^{(n)} = \hat{i}_{k,j}^{(n)}$
* While $\sum_{k \in E_j} P_{k,j}(\tilde{i}_{k,j}^{(n)}) > P_{max,j}$
* Find $k* = \arg\max_{k \in E_j} \frac{(P_{k,j}(\tilde{i}_{k,j}^{(n)}) - P_{k,j}(\tilde{i}_{k,j}^{(n)} - 1))}{(R_{k,j}(\tilde{i}_{k,j}^{(n)}) - R_{k,j}(\tilde{i}_{k,j}^{(n)} - 1))}$
* Update $\tilde{i}_{k*,j}^{(n)} = \tilde{i}_{k*,j}^{(n)} - 1$

4. SIMULATIONS

In this section performances of the M^3RS proposed are compared to those of a well known algorithm in the case of mono-link called "Best SNR", which consists in associating Users to the Base Station with the best radio conditions. Note that this strategy implies no improvement with iterations.

4.1. Simulation Parameters and Radio Conditions

We consider 3 Base Stations each one uses a given bandwidth. Two simulations scenarios are considered. In the first scenario, we consider that many Users are in a short or medium range to Base Stations, which are located at the same place. In the second scenario, we consider that many Users are in a short or medium range to a Base Station 1 and are relatively far from other Base Stations. Detailed parameters are displayed in the following table:

	Simulation 1	**Simulation 2**
Cell	hexagonal radius = 1000 m	hexagonal radius = 1000 m
Station 1 [x,y]	[0,0]	[-300,0]
Station 2 [x,y]	[0,0]	[0,0]
Station 3 [x,y]	[0,0]	[100,0]
User Distribution	hexagonal around Station 1 radius = 250 m	hexagonal around Station 1 radius = 250 m
User Nb.	10	10
User Req.	2.560 Mb/s	2.560 Mb/s
Monte Carlo Sim	100	100

Table 1. Simulation Parameters

Simulations performed assume that radio conditions between Users and Base Stations are obtained from pedestrian A model of [13] with parameters displayed in Table 2.

The Signal to Interference plus Noise ratio is given by the following expression:

$$SINR_{k,j} = \frac{\|g_{k,j}\|^2 P_{k,j}}{I_{inter} + I_{intra} + \sigma_n^2} \quad (6)$$

Carrier frequency	1950 MHz
Bandwidth	4.65 MHz
Thermal Noise	-170.3 dBm
Path loss exponent	3.52
Path loss deviation	4 dB
Transmission power	15 W
Antenna Gain	17 dBi
Intracell Interference	30 dBm
Intercell Interference	-70 dBm

Table 2. Attenuation Model Parameters

Figure 1 displays the relationship between target SINR and data rates for HSDPA (curve with symbols "+") and for our simulations (curve beneath). In our simulations, we took regularly spaced data rates (multiple of 64 kb/s).

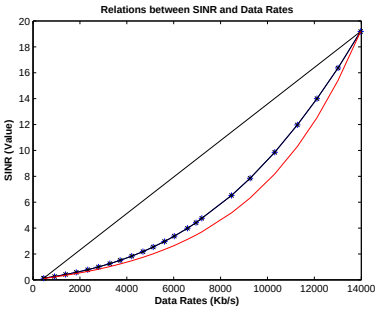

Fig. 1. Relationship between target SINR and data rates.

The relationship between powers and data rates is also convex and is obtained by introducing the following equation in the previous one for the available discrete data rates:

$$R_{k,j} = B\log(1 + SINR_{k,j}) = B\log(1 + \alpha_{k,j}P_{k,j}) \quad (7)$$

4.2. Simulation Results

Figure 2 displays the results obtained for the first simulation. It can be seen that every User is satisfied when DCO requirements are used and nearly every User is satisfied with "Best SNR" requirements. Moreover and as expected, the two first algorithms, which use DCO requirements, use less power than the two other ones. We can also remark that in the case where every User is satisfied at first iteration, no improvement is obtained by iterations.

Figure 3 displays the results obtained for the second simulation. It can be seen that at the first iteration, some Users are unsatisfied for all the algorithms. Algorithms using DCO requirements converge to the optimal solution within 3 iterations (number of iterations equal to the number of various systems) however Algorithms using "Best SNR" never improve

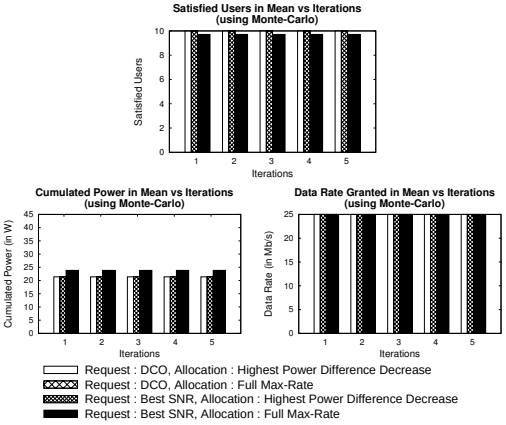

Fig. 2. Results for Simulation 1.

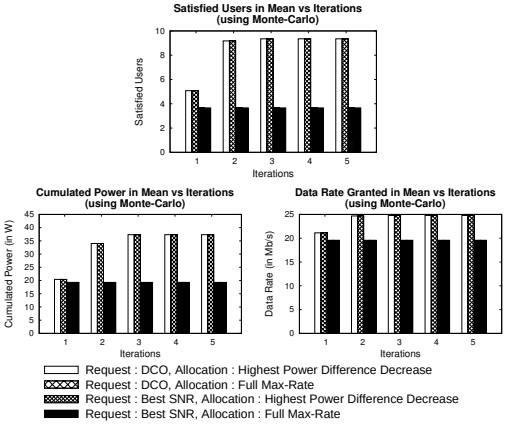

Fig. 3. Results for Simulation 2.

over time. The proposed M^3RS converges then rapidly and allows decreasing the total consumed power by Base Stations.

5. CONLUSION

In this paper we provided a distributed discrete resource allocation algorithm minimizing power usage in heterogenous systems, while respecting minimum data rate constraints of Users and maximum available power at Base Stations. The algorithm proposed lies on the division of the initial problem into two independent subproblems (1- obtaining an allocation satisfying User and Base Stations constraints 2- Allocating power instantaneously to cope with varying radio conditions).

This algorithm converges to a stable solution (configuration) within N_s iterations where N_s is the number of different systems. A configuration is stable if the total power of each system (Base Station) needed to satisfy the rate requirement of all connected Users does not exceed the total available power of this Base Station. Since the number of systems N_s is small (no more than 3 or 4) the convergence of our algorithm is relatively fast (few ms). Once the User and Base Station configuration is finished (convergence to a stable solution) each Base Station can start its own fast resource allocation (e.g. fast Scheduling).

6. REFERENCES

[1] S. Boyd and L. Vandenberghe, "Convex Optimization," *New York: Cambridge University Press*, 2004.

[2] W. Yu, W. Rhee, S. Boyd, and J. Cioffi, "Iterative Water-filling for Gaussian Vector Multiple-access Channels," *IEEE Trans. on Information Theory*, vol. 50, no. 1, pp. 145–152, Jan. 2004.

[3] W. Rhee and J. Cioffi, "Increase in Capacity of Multiuser OFDM System using Dynamic Subchannel Allocation," *Proceedings of the IEEE Vehicular Technology Conference*, pp. 1085–1089, May 2000.

[4] C. Saraydar, N. Mandayam, and D. Goodman, "Efficient Power Control via Pricing in Wireless Data Networks," *IEEE Trans. on Communications*, vol. 50, no. 2, pp. 291–303, 2002.

[5] Z. Han and K. Liu, "Noncooperative Power-control Game and Throughput Game over Wireless Networks," *IEEE Trans. on Communications*, vol. 53, no. 10, pp. 1625–1629, 2005.

[6] J. Nash, "Two-person Cooperative Games," *Econometrica*, vol. 21, no. 1, pp. 128–140, 1953.

[7] H. J. M. Peters, "Axiomatic Bargaining Game Theory," *Kluwer Academic Publishers*, 1992.

[8] L. Zhou, "The Nash Bargaining Theory with Non-convex Problems," *Econometrica*, vol. 65, pp. 681–685, May 1997.

[9] Z. Han, Z. Ji, and K. Liu, "Fair Multiuser Channel Allocation for OFDMA Networks using Nash Bargaining Solutions and Coalitions," *IEEE Trans. on Communications*, vol. 53, no. 8, pp. 1366–1376, 2005.

[10] M. Chiang, S. H. Low, A. R. Calderbank, and J. C. Doyle, "Layering as Optimization Decomposition: A Mathematical Theory of Network Architectures," *Proceedings of the IEEE*, vol. 95, no. 1, pp. 255–312, Jan. 2007.

[11] J. Huang, V. Subramanian, R. Berry, and R. Agrawal, "Downlink Scheduling and Resource Allocation for OFDM Systems," *40th Annual Conference on Information Sciences and Systems*, pp. 1272–1279, 2006.

[12] Y. Shoham and A. Gersho, "Efficient Bit Allocation for an Arbitrary Set of Quantizers," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 36, pp. 1445–1453, Sept. 1988.

[13] H. Holma and A. Toskala, "WCDMA for UMTS: Radio Access for Third Generation Mobile Communications, 3rd Edition," *Wiley*, Sept. 2004.

Bibliography

[Badi 05a] L. Badia, D. Guerra, and M. Zorzi. "Micro-economic strategies for the radio resource management of heterogeneous access techniques". Sep. 2005.

[Badi 05b] L. Badia, C. Taddia, G. Mazzini, and M. Zorzi. "Multi-radio resource allocation strategies for heterogeneous wireless networks". Sep. 2005.

[Bert 99] D. P. Bertsekas. "Nonlinear Programming". *Athena Scientific*, 1999.

[Blau 07] I. Blau, G. Wunder, I. Karla, and R. Siegle. "Cost based Heterogeneous Access Management in Multi-Service, Multi-System Scenarios". pp. 1–5, 2007.

[Boyd 04] S. Boyd and L. Vandenberghe. "Convex Optimization". *New York: Cambridge University Press*, 2004.

[Budd 05] M. Buddhikot, P. Kolodzy, S. Miller, K. Ryan, and J. Evans. "DIMSUMNet: New Directions in Wireless Networking Using Coordinated Dynamic Spectrum Access". June 2005.

[Carn 05] G. Carneiro, C. García, P. Neves, Z. Chen, M. Wetterwald, M. Ricardo, P. Serrano, S. Sargento, and A. Banchs. "The DAIDALOS Architecture for QoS over Heterogeneous Wireless Networks". June 2005.

[Chen 06] B. B. Chen and M. C. Chan. "Resource Management in Heterogenous Wireless Networks with Overlapping Coverage". Jan. 2006.

[Chen 08] J. Chen, G. Murtaza, S. Adimulam, G. Singh, and K. Toyama. "Deadline Constrained Minimum Cost Data Transfer over Heterogeneous Connections". Dec. 2008.

[Chia 03] M. Chiang and A. Sutivong. "Efficient Optimization of Constrained Nonlinear Resource Allocation". *Proceedings of the IEEE Global Telecommunications Conference*, Vol. 7, pp. 3782–3786, Dec. 2003.

[Chia 05] M. Chiang. "Balancing transport and physical Layers in wireless multihop networks: jointly optimal congestion control and power control". *IEEE Journal on Selected Areas in Communications*, Vol. 23, No. 1, pp. 104–116, Jan. 2005.

[Chia 07] M. Chiang, S. H. Low, A. R. Calderbank, and J. C. Doyle. "Layering as Optimization Decomposition: A Mathematical Theory of Network Architectures". *Proceedings of the IEEE*, Vol. 95, No. 1, pp. 255–312, Jan. 2007.

[Coup 08] M. Coupechoux, J.-M. Kelif, and P. Godlewski. "Network Controlled Joint Radio Resource Management for Heterogeneous Networks". pp. 1771–1775, May 2008.

[Cris 03] G. Cristache, K. David, M. Hildebrand, J. Diaz, and R. Sigle. "Aspect for the integration of ad hoc and cellular networks". 2003.

[Ferr 05] R. Ferrus, A. Gelonch, O. Sallent, J. Perez-Romero, N. Nafisi, and M. Dohler. "Vertical Handover Support in Coordinated Heterogeneous Radio Access Networks Conference". June 2005.

[Gela 05a] X. Gelabert, J. Pérez-Romero, O. Sallent, and R. Agustí. "On the Suitability of Load Balancing Principles in Heterogeneous Wireless Access Networks". pp. 18–22, Sep. 2005.

[Gela 05b] X. Gelabert, J. Pérez-Romero, O. Sallent, R. Agustí, and F. Casadevall. "Radio Resource Management in Heterogeneous Networks". July 2005.

[Gozl 08] J. Gozálvez, M. C. Lucas-Estañ, and J. Sánchez-Soriano. "Joint Radio Resource Management in Beyond 3G Heterogeneous Wireless Systems". Sep. 2008.

[Han 05a] Z. Han, Z. Ji, and K. Liu. "Fair Multiuser Channel Allocation for OFDMA Networks using Nash Bargaining Solutions and Coalitions". *IEEE Trans. on Communications*, Vol. 53, No. 8, pp. 1366–1376, 2005.

[Han 05b] Z. Han and K. Liu. "Noncooperative Power-control Game and Throughput Game over Wireless Networks". *IEEE Trans. on Communications*, Vol. 53, No. 10, pp. 1625–1629, 2005.

[Havi 01] P. J. M. Havinga, G. J. M. Smit, L. Vognild, and G.Wu. "The SMART project: Exploiting the heterogeneous mobile world". pp. 346–352, 2001.

[Jang 03] J. Jang and K. B. Lee. "Transmit Power Adaptation for Multiuser OFDM Systems". *IEEE Journal on Selected Areas in Communications*, Vol. 21, No. 2, pp. 171–178, Feb. 2003.

[Joha 05] K. Johansson, J. Zander, and A. Furuskar. "Modelling the cost of heterogeneous wireless access networks". *International Journal of Mobile Network Design and Innovation*, Vol. 2, pp. 58–66, May 2005.

[Kiva 03] D. Kivanc, L. Guoqing, and L. Hui. "Computationally efficient bandwidth allocation and power control for OFDMA". *IEEE Trans. on Wireless Communications*, Vol. 2, No. 6, pp. 1150–1158, Nov. 2003.

[Kuhn 55] H. W. Kuhn. "The Hungarian Method for the assignment problem". *Naval Research Logistics Quarterly*, Vol. 2, pp. 83–97, 1955.

[Li 06] D. Li and X. Su. "Nonlinear Integer Programming". *Springer, International Series in Operations Research and Management Science*, 2006.

[Low 01] S. Low, L. Peterson, and L. Wang. "Understanding TCP vegas: a duality model". *SIGMETRICS Perform. Eval. Rev.*, Vol. 29, No. 1, pp. 226–235, 2001.

[Luca 08] M. Lucas-Estan, J. Gozalvez, and J. Sanchez-Soriano. "Common radio resource management policy for multimedia traffic in beyond 3G heterogeneous wireless systems". pp. 1–5, Sep. 2008.

[Luo 03] J. Luo, R. Mukerjee, M. Dillinger, E. Mohyeldin, and E. Schulz. "Investigation of Radio Resource Scheduling in WLANs Coupled with 3G Cellular Network". *IEEE Communications Magazine*, June 2003.

[Magn 05] P. Magnusson, F. Berggren, I. Karla, R. Litjens, F. Meago, H. Tang, and R. Veronesi. "Multi-radio resource management for communication networks beyond 3G". pp. 1653–1657, 2005.

[Mott 02] N. Motte, R. Rümmler, D. Grandblaise, L. Elicegui, D. Bourse, and E. Seidel. "Joint radio resource management and QoS implications of software downloading for SDR terminals". June 2002.

[Murr 03] K. Murray and D. Pesch. "Adaptive Policy Based Access Management in Heterogeneous Wireless". pp. 325–329, 2003.

[Murr 04] K. Murray and D. Pesch. "Policy based access management and handover control in heterogeneous wireless networks". pp. 3319–3323, Sep. 2004.

[Murr 07] K. Murray and D. Pesch. "Call Admission and Handover Management for Access Network Roaming in Heterogeneous Wireless Networks". 2007.

[Nash 53] J. Nash. "Two-person Cooperative Games". *Econometrica*, Vol. 21, No. 1, pp. 128–140, 1953.

[Papa 06] K. Papadaki and V. Friderikos. "Optimal Vertical Handover Control Policies for Cooperative Wireless Networks". *Journal Of Communication and Networks*, Vol. 8, pp. 442–450, 2006.

[Pere 03] J. Perez-Romero, O. Sallent, R. Agusti, P. Karlsson, A. Barbaresi, L. Wang, F. Casadevall, M. Dohler, H. Gonzalez, and F. Cabral-Pinto. "An architecture for integrating UMTS and 802.11 WLAN". pp. 716–723, 2003.

[Pere 05] J. Perez-Romero, O. Sallent, R. Agusti, P. Karlsson, A. Barbaresi, L. Wang, F. Casadevall, M. Dohler, H. Gonzalez, and F. Cabral-Pinto. "Common radio resource management: functional models and implementation requirements". pp. 2067–2071, Sep. 2005.

[Pere 07] J. Perez-Romero, O. Salient, and R. Agusti. "On the Optimum Traffic Allocation in Heterogeneous CDMA/TDMA Networks". *IEEE Trans. on Wireless Communications*, Vol. 6, pp. 3170–3174, Sep. 2007.

[Pete 92] H. J. M. Peters. "Axiomatic Bargaining Game Theory". *Kluwer Academic Publishers*, 1992.

[Rayc 03] D. Raychaudhuri and X. Jing. "A Spectrum Etiquette Protocol for Efficient Coordination of Radio Devices in Unlicensed Bands". Sep. 2003.

[Rhee 00a] W. Rhee and J. Cioffi. "Increase in Capacity of Multiuser OFDM System using Dynamic Subchannel Allocation". *Proceedings of the IEEE Vehicular Technology Conference*, pp. 1085–1089, May 2000.

[Rhee 00b] W. Rhee and J. M. Cioffi. "Increasing in Capacity of Multiuser OFDM System Using Dynamic Subchannel Allocation". *Proceedings of the IEEE Vehicular Technology Conference, Spring*, Vol. 2, pp. 1085–1089, May 2000.

[Sall 08] O. Sallent, R. Agustí, J. Pérez-Romero, and L. Giupponi. "Decentralized spectrum and radio resource management enabled by an on-demand Cognitive Pilot Channel". *Annales des Télécommunications*, Vol. 63, pp. 281–294, 2008.

[Sara 02] C. Saraydar, N. Mandayam, and D. Goodman. "Efficient Power Control via Pricing in Wireless Data Networks". *IEEE Trans. on Communications*, Vol. 50, No. 2, pp. 291–303, 2002.

[Shen 07] W. Shen and Q.-A. Zeng. "Resource Allocation Schemes in Integrated Heterogeneous Wireless and Mobile Networks". *Journal of Networks*, Vol. 2, pp. 78–86, 2007.

[Shoh 88] Y. Shoham and A. Gersho. "Efficient Bit Allocation for an Arbitrary Set of Quantizers". *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol. 36, pp. 1445–1453, Sep. 1988.

[Suli 04] I. Suliman, C. Pomalaza-Ráez, I. Oppermann, and J. Lehtomäki. "Radio resource allocation in heterogeneous wireless networks using cooperative games". pp. 16–18, Aug. 2004.

[Vuce 08] N. Vucevic, F. Bernardo, A. Umbert, and M. López-Benítez. "End-to-edge QoS across heterogeneous wireless and wired domains ". July 2008.

[Vuli 05] N. Vulic, S. H. de Groot, and I. Niemegeers. "Common radio resource management for WLAN-UMTS integration Radio Access Level". June 2005.

[Wang 07] L. Wang, H. Aghvami, N. Nafisi, J.Pérez-Romero, O. Sallent, and R. Agusti. "Coverage-based Common Radio Resource Management in Heterogeneous CDMA/TDMA Cellular Systems". 2007.

[Wong 99] C. Y. Wong, R. S. Cheng, K. B. Letaief, and R. D. Murch. "Multiuser OFDM with adaptive subcarrier, bit, and power allocation". *IEEE Journal on Selected Areas in Communications*, Vol. 17, pp. 1747–1758, Oct. 1999.

[Yu 04] W. Yu, W. Rhee, S. Boyd, and J. Cioffi. "Iterative Water-filling for Gaussian Vector Multiple-access Channels". *IEEE Trans. on Information Theory*, Vol. 50, No. 1, pp. 145–152, Jan. 2004.

[Zhou 97] L. Zhou. "The Nash Bargaining Theory with Non-convex Problems". *Econometrica*, Vol. 65, pp. 681–685, May 1997.