Optimisation de l'association des utilisateurs et de l'allocation des ressources dans les réseaux sans fil hétérogènes

Mohamad Zalghout

Thèse

THESE INSA Rennes
sous le sceau de l'Université Bretagne Loire
pour obtenir le titre de
DOCTEUR DE L'INSA RENNES
Spécialité : Télécommunication

présentée par
Mohamad ZALGHOUT
ECOLE DOCTORALE : ***MATHSTIC***
LABORATOIRE : ***IETR***

Optimization of User Association and Resource Allocation in Heterogeneous Networks

Soutenance soutenue le 23 Octobre 2017
devant le jury composé de :

Michel TERRE
Professeur au CNAM de Paris / *Président*
Joumana FARAH
Professeur à l'Université Libanaise, Beyrouth / *Rapporteur*
André-Luc BEYLOT
Professeur à l'ENSEEIHT de Toulouse / *Rapporteur*
Laurent CLAVIER
Professeur à Télécom Lille / *Examinateur*
Samih ABDUL- NABI
Enseignant-Chercheur à l'Univ. Internationale du Liban / *Co-encadrant de thèse*
Ayman KHALIL
Enseignant-Chercheur à l'Univ. Internationale du Liban / *Co-encadrant de thèse*
Matthieu CRUSSIERE
Maître de Conférences à l'INSA Rennes / *Co-encadrant de thèse*
Jean-François HELARD
Professeur à l'INSA Rennes / *Directeur de thèse*

Optimization of User Association and Resource Allocation in Heterogeneous Networks

Mohamad ZALGHOUT

En partenariat avec

Abstract

It is indicated that the expansion of the wireless data traffic requirements exceeds the capacity growth rate of new wireless access technologies. Therefore, next-generation mobile wireless networks are moving toward heterogeneous architectures usually referred to as heterogeneous wireless networks (HWNs). HWNs are usually characterized by the integration of cellular networks and wireless local area networks (WLANs) to meet user requirements and enhance system capacity. In fact, integrating different types of wireless access technologies in HWNs provides flexible choices for users to be associated with the network that best satisfies their needs. In this context, this thesis discusses the user association and downlink resource allocation problem in a heterogeneous wireless system that is based on integrated Wi-Fi access points (APs) and long-term evolution (LTE) base stations (BSs).

The contributions of this thesis could be divided into three main parts. In the first part, a novel user association and resource allocation optimization problem is formulated to maximize the overall user satisfaction in the system. The user satisfaction is based on a weighted profit function that aims at enhancing the relative received signal strength and decreasing the power consumption of mobile terminals (MTs). Since a MT is only allowed to be associated with a single network at a time, the formulated optimization problem is binary with an NP-complete complexity. Then, multiple centralized solutions with polynomial-time complexities are proposed to solve the formulated problem. The proposed centralized solutions are based on heuristic approaches and on the continuous relaxation of the formulated binary optimization problem.

The second part of the thesis aims at providing a distributed solution for the formulated problem. The proposed distributed solution deploys the Lagrangian-relaxation technique in order to convert the global formulated problem into multiple distributed Knapsack problems, each network processes its corresponding Knapsack problem. The sub-gradient method is used in order to find the optimal, or near optimal, Lagrangian multipliers.

Finally, the third part of the thesis studies new perspectives of the formulated optimization problem and its corresponding centralized and distributed solutions. Mainly, a generalized priority-aware user association and resource allocation problem is formulated. The priority-aware problem is then reduced into multiple problems that are solved using the proposed centralized and distributed solutions. Moreover, a novel power efficiency maximization solution is proposed by altering the objectives of the main formulated optimization problem.

Résumé étendu de la thèse en français

Introduction générale

L'expansion des exigences de trafic de données sans fil dépasse le taux de croissance de la capacité des nouvelles technologies d'accès sans fil [HQ14]. Par conséquent, les réseaux sans fil mobiles de la prochaine génération tendent à adopter des architectures hétérogènes généralement appelées réseaux sans fil hétérogènes (HWN) [HQ14]. Dans ces réseaux HWN, les utilisateurs ont le droit de se connecter à différents types de technologies d'accès radio telles que les stations de base LTE (BS) ou les points d'accès Wi-Fi (AP). Une telle architecture augmente la capacité du système en réduisant le nombre d'utilisateurs souhaitant de se connecter aux BS. En outre, les réseaux HWN différent des choix dynamiques pour que l'utilisateur final transfère sa session de données en cours à un réseau plus favorable qui peut satisfaire ses besoins. Par conséquent, les réseaux HWN sont généralement accompagnés du concept de « Always Best Connected » ou ABC [GJ03], qui est le processus d'être connecté au meilleur réseau disponible à tout moment. Cependant, le concept ABC est généralement considéré du point de vue de l'utilisateur pour classer les réseaux candidats et se connecter au meilleur réseau. Habituellement, les algorithmes de sélection de réseau basés sur le concept ABC ne prennent pas en compte les ressources limitées des réseaux et l'effet du handover sur le système. Par conséquent, il est essentiel dans les réseaux HWN d'ouvrir la voie à un système ABC optimisé dans un contexte qui tient compte des besoins du réseau des utilisateurs.

De nos jours, les travaux de recherche se concentrent sur la façon dont les terminaux mobiles (MT) répondent à la variété des technologies d'accès sans fil disponibles en même temps. Dans ce contexte, des études conjointes sur l'association des utilisateurs et l'allocation des ressources ont été introduites ces dernière années. Le problème de l'association des utilisateurs traite le choix du réseau auquel le MT doit être associé parmi plusieurs technologies d'accès disponibles, tandis que l'allocation des ressources traite la quantité de ressources qui doit être attribuée par les réseaux pour chaque MT.

Pour résumer, l'objectif principal de la thèse est de proposer des solutions avec une complexité réduite pour l'association d'utilisateurs et le problème d'allocation de ressources dans des réseaux hétérogènes. Les solutions visent à fournir une vision ABC optimisée qui considère le système dans son ensemble tout en prenant des décisions. Par conséquent, la satisfaction, les besoins et les préférences des

utilisateurs sont pris en considération. En outre, cette thèse explore le problème de la maximisation de l'efficacité énergétique (débit de données par unité de puissance), en plus du problème d'allocation de ressources dans un système avec des utilisateurs de différentes priorités.

Chapitre 1

Dans cette thèse, nous traitons les réseaux HWN intégrant des réseaux Wi-Fi et des réseaux cellulaires. La technologie Wi-Fi évolue pour fournir une connexion au réseau local sans fil, tandis que les réseaux cellulaires sont en cours d'élaboration afin de fournir une connexion sans fil étendue. Les réseaux cellulaires soutiennent fortement la mobilité des utilisateurs tout en fournissant des débits de données plus faibles que les réseaux Wi-Fi et à un coût plus élevé.

Algorithmes de sélection de réseau

Le problème de la sélection du réseau a été largement étudié dans la littérature. Habituellement, les principaux paramètres considérés pour les algorithmes de sélection de réseau sont: la puissance du signal reçu (RSS), le coût, la charge du réseau, la consommation d'énergie, les préférences des utilisateurs et le débit. Plusieurs études classent les réseaux en fonction de paramètres uniques ou multiples. L'idée principale derrière les schémas de sélection de réseau est de sélectionner pour chaque MT le meilleur réseau qui le satisfait de manière égoïste. Par conséquent, les algorithmes de sélection de réseau ne tiennent pas en compte de l'allocation globale des ressources ni des solutions d'association d'utilisateurs à l'échelle du système. Au lieu de cela, ils sont conçus pour satisfaire les besoins de chaque utilisateur individuellement.

Association d'utilisateurs et allocation de ressources à l'échelle du système

Plusieurs études envisagent une association d'utilisateurs et des algorithmes d'allocation de ressources à l'échelle du système. Habituellement, ces études visent à:

- augmenter le nombre d'utilisateurs connectés,
- réduire la consommation totale d'énergie,
- réduire le nombre de réseaux actifs,
- équilibrer la charge entre plusieurs réseaux,
- et maximiser la satisfaction des utilisateurs.

Cependant, un grand nombre de ces études ne tiennent pas compte conjointement des exigences relatives aux débits demandés par les utilisateurs, des préférences des utilisateurs, et des contraintes du réseau.

Conclusion

Les travaux proposés devrait prendre en compte la coordination entre les réseaux Wi-Fi et les réseaux cellulaires pour faire face à la croissance exponentielle du trafic de données demandé. Un aspect important est de suivre les objectifs du concept ABC ; la satisfaction, les demandes et les préférences des utilisateurs devrait être prises en compte afin de maintenir la meilleure connexion en tout temps. Sur la base des paramètres généralement considérés dans le système ABC, nous considérerons principalement la consommation d'énergie du MT et la qualité du

signal. La solution proposée devra être facilement adaptée pour ajouter des paramètres supplémentaires. De plus, la solution proposée devra prendre en compte la quantité de débit demandé par chaque MT, la condition du canal entre les MTs et les réseaux, en plus de la capacité maximale de chaque réseau.

En outre, la solution proposée devra être facilement configurée pour répondre aux objectifs basés sur les réseaux, tels que la maximisation de l'efficacité énergétique.

Chapitre 2

Norme IEEE 802.21.1

L'IEEE a récemment défini une nouvelle norme qui est la norme 802.21.1 pour permettre un handover (HO) entre des réseaux de même type ou de différents types. La norme 802.21.1 définit plusieurs services indépendants des médias (MIS) qui fournissent des informations utiles dans la décision du HO et facilitent l'intégration des HO. L'un des sujets intéressants que la norme IEEE 802.21.1 MIS traite est le transfert indépendant des médias pour les réseaux d'accès radio (SDRAN) définis par logiciel [Jin15]. Le cadre SDRAN associe les capacités des MIS à la mise en réseau définie par logiciel (SDN) afin de fournir les fonctionnalités HO, l'allocation des ressources et la gestion centralisée dans les réseaux HWN et établir une infrastructure fiable pour l'échange de messages entre différentes entités (serveurs, réseaux, MT). En conséquence, nous proposons l'utilisation du paradigme SDRAN pour recueillir des informations liées au problème d'allocation de ressources et d'affectation de réseau et établir un serveur (qui pourrait être un contrôleur SDN) qui maintient une vue globale sur le système, traite l'association d'utilisateurs et de ressources, alloue des ressources aux utilisateurs et déclenche les HO.

Problème d'optimisation

Nous proposons une fonction de profit normalisée $\overline{f}_{mn}$ qui énumère le bénéfice axé sur l'utilisateur contribué en associant MT m au réseau n. La fonction de profit considère la consommation d'énergie le MT et la puissance du signal reçu. Étant donné que les deux paramètres sont d'unités différentes, leurs valeurs normalisées sont considérées. Les préférences de l'utilisateur m sont reflétées par les poids liés à la consommation d'énergie (w_m^{pc}) et à la qualité du signal (w_m^s) tel que $w_m^{pc} + w_m^s = 1$. Les valeurs $\left\lceil \frac{Q_m T_n}{r_{mn}^{tot}} \right\rceil$ et $\left\lceil \frac{Q_m T_n}{B_n^{RB} \log_2(1+\gamma_{mn})} \right\rceil$ indiquent le montant des ressources demandées par le MT m de AP et BS n respectivement en fonction de la quantité de débit demandé par le MT m, les conditions de canal entre le réseau n et le MT m, et le type du réseau n. Le nombre total de ressources dans AP ou BS n est désigné par T_n ou U_n respectivement. L'ensemble de tous les réseaux est désigné par $\mathcal{N}$, et il est divisé en deux ensembles $\mathcal{N}_{\mathcal{AP}}$ et $\mathcal{N}_{\mathcal{BS}}$ pour les APs et les BSs respectivement. L'ensemble de tous les MTs est désigné par $\mathcal{M}$. Chaque MT est autorisé à être associé à un seul réseau à la fois. Par conséquent, nous définissons une variable d'association d'utilisateurs booléens x_{mn} qui indique si le MT m est associé au réseau n ou non. Par conséquent, le problème d'optimisation suivant est proposé pour maximiser le bénéfice global défini par l'utilisateur dans

le système:

$$\textbf{P1: } max \sum_{m\in\mathcal{M}}\sum_{n\in\mathcal{N}} f_{mn}x_{mn} \tag{1a}$$

$$s.\ t. \sum_{m\in\mathcal{M}} \left\lceil \frac{Q_m T_n}{r_{mn}^{tot}} \right\rceil x_{mn} \leqslant T_n \qquad \forall n \in \mathcal{N}_{\mathcal{AP}} \tag{1b}$$

$$\sum_{m\in\mathcal{M}} \left\lceil \frac{Q_m T_n}{B_n^{RB}\log_2(1+\gamma_{mn})} \right\rceil x_{mn} \leqslant U_n \qquad \forall n \in \mathcal{N}_{\mathcal{BS}} \tag{1c}$$

$$\sum_{n\in\mathcal{N}} x_{mn} \leqslant 1 \qquad \forall m \in \mathcal{M} \tag{1d}$$

$$x_{mn} \in \{0,1\} \qquad \forall m \in \mathcal{M}, \forall n \in \mathcal{N} \tag{1e}$$

Les contraintes (1b) et (1c) garantissent que la capacité de chaque réseau n'est pas violée, et les contraintes (1d) et (1e) garantissent que chaque MT n'est associé qu'à un seul réseau à la fois. Le problème **P1** est un problème de programmation linéaire binaire connu pour avoir une complexité NP-hard. Pour estimer la solution optimale du problème **P1**, nous utilisons l'algorithme branch-and-bound. Cependant, comme le nombre de réseaux et de MTs augmente largement, le temps requis pour trouver la solution optimale augmente de façon dramatique. Par conséquent, dans les chapitres 3 et 4, nous proposons respectivement des solutions centralisées et distribuées pour le problème avec des complexités acceptables.

Résultats de simulations

Nous comparons l'effet de la variation de poids sur la performance de la solution triviale profit-fonction (PF) et la solution optimale basée sur la méthode branch-and-bound. Pour chaque solution, trois cas de variation de poids (scénarios) sont simulés. Plus précisément, nous étudions, pour chaque algorithme, l'effet de l'augmentation de nombre de MTs actifs sur les valeurs moyennes de relatif RSS (RRSS) et la consommation d'énergie (Figures 1 et 2). Les résultats de simulations montrent que la fonction de profit s'adapte efficacement aux variations de poids et que la solution optimale est bien meilleure que la solution triviale PF.

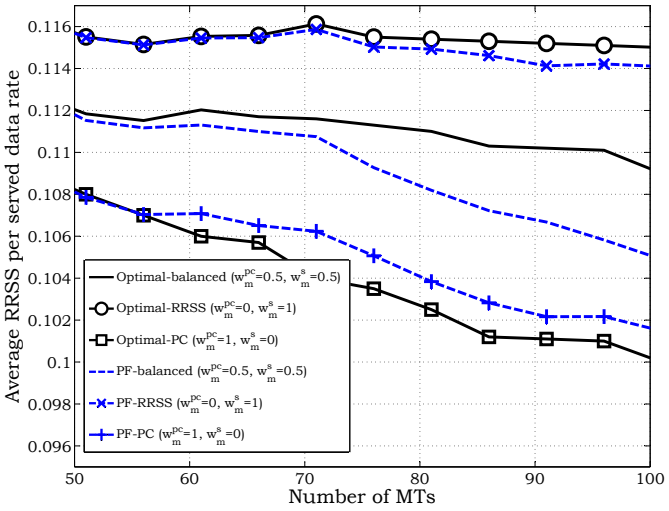


FIGURE 1 – Puissance moyenne du signal.

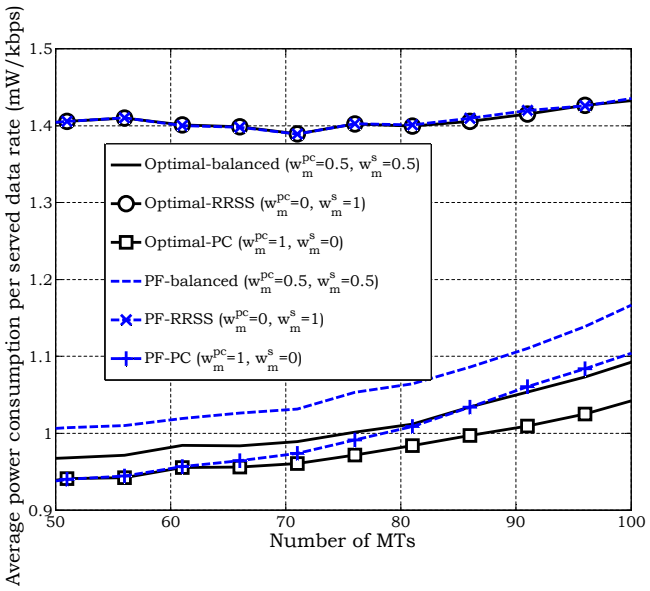


FIGURE 2 – Consommation moyenne d'énergie.

Chapitre 3

Dans ce chapitre, nous proposons quatre nouvelles solutions centralisées pour le problème **P1** présente dans le chapitre 2.

Dépassement continu du problème d'optimisation binaire

La relaxation continue du problème binaire **P1** repose sur le relâchement de la contrainte d'affectation binaire (1e) en une contrainte continue bornée. C'est-à-dire que la contrainte $x_{mn} \in \{0, 1\}$ devient $x_{mn} \in [0, 1]$. En conséquence, le problème continu peut être résolu en utilisant des méthodes destinées à résoudre des programmes linéaires continus tels que la méthode simplex et les méthodes de points intérieurs. Résoudre le problème continu produit trois séries de MTs classées en fonction de leurs résultats d'association:

- Set1: ensemble de MTs non connectés ($\sum_{n\in\mathcal{N}} x_{mn} = 0$).
- Set2: ensemble de MTs connectés à un seul réseau n tel que $x_{mn} = 1$.
- Set3: ensemble de MTs connectés à plusieurs réseaux simultanément ou à un seul réseau n tel que $x_{mn} < 1$, c'est-à-dire que le MT reçoit un débit de données inférieur au taux de données demandé.

Puisque dans notre contexte, un MT doit être connecté à un seul réseau et reçoit toutes les ressources requises, une approche de relaxation continue est utilisée pour associer le MT à des réseaux appropriés et vider Set3. Par conséquent, nous proposons deux algorithmes d'association d'utilisateurs et d'allocation de ressources basés sur la relaxation continue du problème **P1**. La première solution a une complexité indéterminée tandis que la seconde possède une complexité déterminée en termes de temps polynomial. La solution avec une complexité indéterminée peut être considérée comme la solution optimale basée sur l'approche de relaxation continue, et elle est utilisée pour faire du benchmarking uniquement parce qu'elle vise à proposer des solutions d'affectation de ressources et de réseaux avec une complexité déterminée de la durée polynomiale.

Solution optimale de relaxation continue (avec complexité indéterminée)

Principalement, la solution optimale de relaxation continue vise à résoudre le problème d'optimisation linéaire continue pour tous les MT dans Set1∪Set3 jusqu'à $|\text{Set3}| = 0$, i.e. tous les résultats d'association du problème continu sont booléens. Chaque fois que le problème d'optimisation continue est résolu, les valeurs d'association des MTs dans Set2 sont enregistrées, le nombre de ressources gratuites dans chaque réseau est mis à jour en déduisant le nombre de ressources allouées aux MTs dans Set2 et ces MTs ne sont pas déjà considérés dans la fonction d'optimisation continue. Cependant, il est impossible de déterminer de manière analytique le nombre de fois que le programme linéaire continu sera résolu dans cette solution, et donc la complexité n'a pas pu être déterminée.

Solution basée sur la relaxation continue sous-optimale (avec complexité du temps polynomial)

Maintenant, une solution avec une complexité de temps polynomial est proposée pour déterminer les valeurs d'association pour les MTs dans Set1∪Set3. Le problème continu est résolu une seule fois et chaque MT m en Set2 est associé au réseau n où $x_{mn} = 1$.

Tout d'abord, nous commençons à déterminer les valeurs d'association pour tous les MTs dans Set3 uniquement, i.e. les MTs avec des valeurs d'association fractionnaires. La solution itère tous les MTs dans Set3. Dans chaque itération, la solution tente de déterminer la décision d'association pour le MT avec la plus grande valeur d'association parmi tous les MTs dans Set3.

Par la suite, tous les MTs dont la décision d'association n'est toujours pas déterminée sont soumis à des procédures similaires à celles proposées dans le paragraphe précédent. Cependant, l'objectif ici est d'associer le MT au profit le plus élevé à chaque itération au lieu de la valeur d'association fractionnaire la plus élevée.

Approximation-based solution

Après avoir examiné de plus près le problème **P1**, on peut noter qu'il est similaire au problème d'affectation généralisée (GAP) [MT81]. En fait, Martello et Toth, qui ont des contributions importantes dans le domaine des problèmes de GAP, knapsack et bin-packing, ont proposé un algorithme heuristique pour se rapprocher de GAP, basé sur une demande du MT [MT81]. Là, la "désirabilité" d'affecter le MT m au réseau n est mesurée selon le facteur de désirabilité Ω_{mn}. Les facteurs possibles qui pourraient être considérés comme la mesure de désirabilité sont abordés dans la section suivante.

Pour chaque MT, la différence entre la plus haute et la deuxième valeur de Ω_{mn} est calculée, et les MTs sont ensuite attribués dans l'ordre décroissant de cette différence. La solution considère itérativement tous les MTs non associés et détermine le MT m^* ayant la différence maximale entre le plus élevé et le deuxième Ω_{mn}. Le MT m^* est ensuite affecté au réseau pour lequel Ω_{m^*n} est maximum, i.e. réseau n^*. De plus, après avoir pris la décision de chaque association, l'algorithme réévalue, pour chaque MT, la différence maximale entre le plus élevé et le deuxième Ω_{mn}, et associe les MTs en fonction de ces nouveaux résultats. Ainsi, une vision semi-globale sur les réseaux disponibles et leurs bénéfices est maintenue tout en prenant des décisions d'association. De plus, l'algorithme préfère associer les MTs à un seul réseau disponible. Cet aspect de l'algorithme joue un rôle essentiel dans la diminution de la probabilité de blocage.

Nouveau facteur d'efficacité

Le problème **P1** vise à maximiser les bénéfices dans le système. Par conséquent, Martello et Toth ont proposé dans [MT81] d'utiliser le profit (f_{mn}) ou le $\frac{profit}{poids}$ comme facteur de désirabilité pour la solution basée sur l'approximation.

Puisque le problème **P1** traite des MTs ayant différents débits de données, c'est-à-dire des poids différents, le rapport $\frac{profit}{poids}$ est approprié comme facteur de

désirabilité pour ce problème. Cependant, compte tenu du poids des MTs, qui peut être considéré comme le nombre de ressources demandées, n'est pas simple parce que les technologies d'accès ont différents types et quantités de ressources. En fait, la quantité de bande passante qui devrait être fournie par un réseau à un MT est liée aux conditions de canal entre le MT et le réseau, et à la quantité de débit demandé par le MT. En outre, la bande passante (en Hz) est une ressource limitée dans tous les systèmes de communication. Par conséquent, la quantité de bande passante demandée par un MT à partir d'un réseau pourrait être considérée comme un poids. Par conséquent, l'efficacité e_{mn} est introduite pour désigner le bénéfice par poids (bande passante demandée) qui a contribué au système en associant le MT m au réseau n.

Solution greedy simple

La solution greedy est également basée sur la demande du MT selon le facteur de désirabilité, mais elle ne se concentre que sur le facteur de désirabilité le plus élevé au lieu de la plus haute et la deuxième plus élevée. Par conséquent, la complexité est plus simple que la solution basée sur l'approximation parce que les MTs sont ici itérés uniquement selon le facteur de désirabilité le plus élevé.

Résultats de simulations

Comme illustré dans les Figures 3 et 4, la solution basée sur l'approximation maintient la performance la plus proche de la solution binaire optimale. Cela indique que la solution basée sur l'approximation se rapproche efficacement de la solution optimale, et le facteur d'efficacité joue également un rôle essentiel pour stimuler les performances de cette solution. D'autre part, l'écart de performance

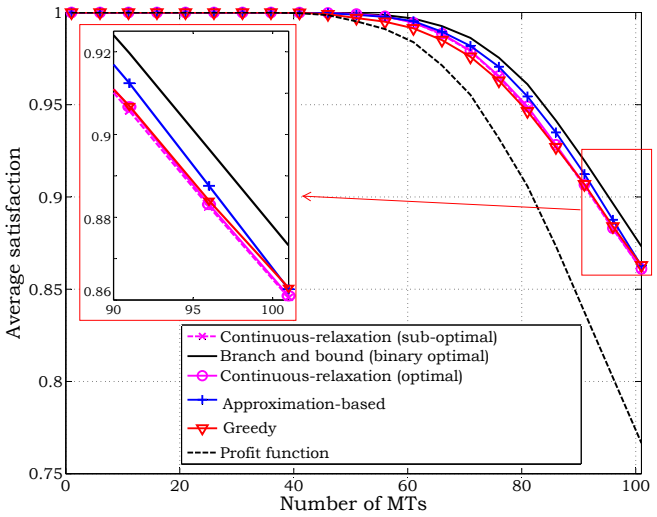


FIGURE 3 – Satisfaction moyenne.

entre la solution binaire optimale et la solution optimale de relaxation continue

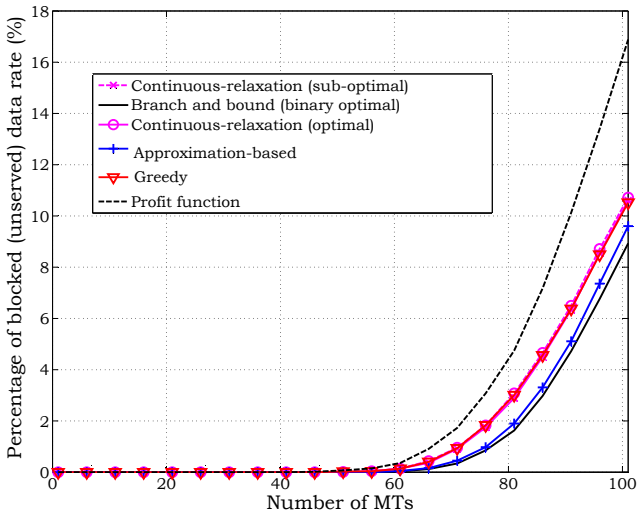


FIGURE 4 – Pourcentage de taux de données bloquées.

indique que la conversion de la contrainte binaire en continu menace l'optimalité de la solution, où la solution greedy pourrait fonctionner de manière similaire à la solution optimale en continu.

Chapitre 4

Dans ce chapitre, la méthode de relaxation lagrangienne est utilisée pour relâcher la contrainte d'affectation du problème **P1**, c'est-à-dire la contrainte (1d). Par la suite, le problème est réparti en plusieurs problèmes de Knapsack, chacun étant résolu de manière indépendante par son réseau correspondant, c'est-à-dire une solution distribuée. Le problème de Knapsack est résolu à l'aide de procédures de programmation dynamique avec une complexité acceptable. Ensuite, en fonction des résultats des problèmes de Knapsack, une solution réalisable est produite en s'assurant que chaque MT est associée à au plus un réseau. Pour trouver les multiplicateurs Lagrangiens optimaux, ou presque optimaux, on utilise la méthode subgradient. Par conséquent, deux solutions distribuées sont proposées en fonction de l'analyse discutée:

- Solution distribuée avec la méthode subgradient: cette solution repose sur la résolution itérative du problème de Knapsack et la recherche de la solution possible jusqu'à ce que des multiplicateurs lagrangiens quasi-optimaux soient détectés.
- Solution distribuée sans la méthode subgradient: cette solution ne vise pas à trouver les multiplicateurs de Lagrange, donc les problèmes de Knapsack sont résolus une fois, et une solution réalisable est basée sur les résultats du problème de Knapsack. Par conséquent, la complexité de cette solution est plus simple que celle de la précédente car la phase de subgradient n'est pas prise en compte.

Résultats de simulations

Les Figures 5 et 6 montrent que la solution distribuée avec la méthode subgradient maintient la performance la plus proche de la solution binaire optimale. Ceci est dû à la méthode de subgradient qui s'approche efficacement de la solution optimale. D'autre part, la solution distribuée sans la méthode subgradient se performe plus pire que la solution basée sur l'approximation, et similaires aux autres solutions.

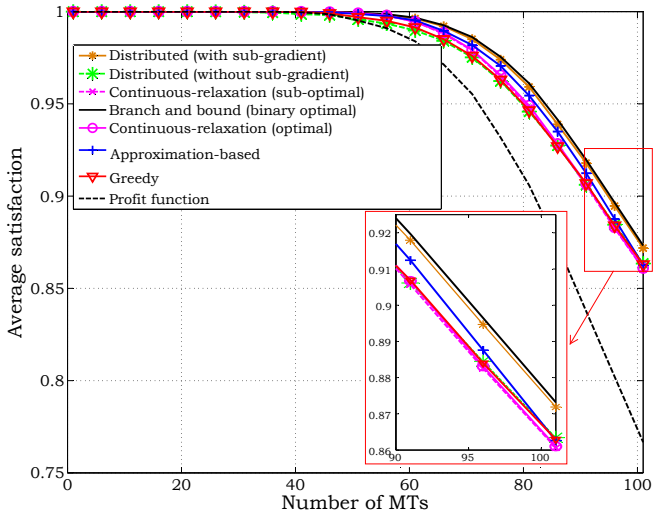


FIGURE 5 – Satisfaction moyenne.

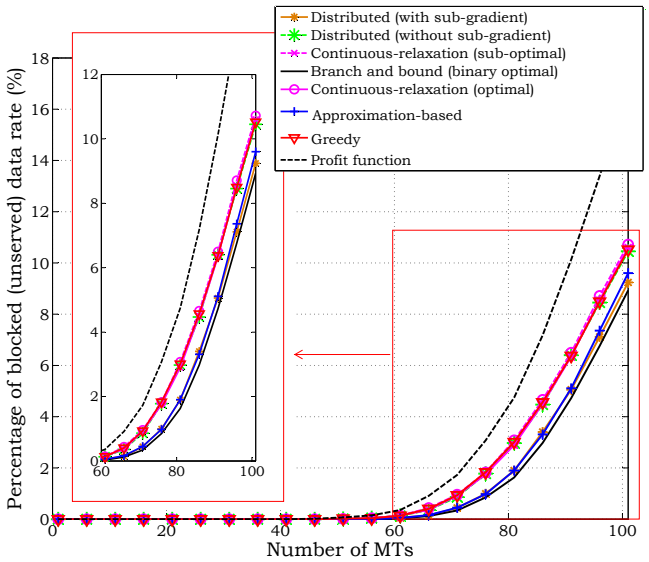


FIGURE 6 – Pourcentage de taux de données bloquées.

Chapitre 5

Dans ce chapitre, nous présentons de nouvelles applications directes pour le problème **P1** proposé dans le chapitre 2 et ses solutions correspondantes proposées dans les chapitres 3 et 4. Tout d'abord, nous proposons une nouvelle solution pour l'association d'utilisateurs prioritaires et l'allocation de ressources dans un système où les utilisateurs ont des priorités différentes. Nous formulons un nouveau problème d'optimisation basé sur les priorités, puis nous le simplifions en un nouveau problème ayant une forme similaire au problème **P1**. En outre, nous discutons de nouvelles perspectives pour l'association d'utilisateurs et l'allocation de ressources pour améliorer l'efficacité énergétique globale (débit par unité de puissance) dans le système.

Gestion axée sur les priorités

Nous considérons que les utilisateurs ont des priorités différentes, de sorte que les utilisateurs ayant la plus haute priorité devraient avoir le meilleur service et le pourcentage de blocage le plus bas. En conséquence, les utilisateurs ayant la plus haute priorité sont autorisés à allouer des ressources utilisées par des utilisateurs ayant des priorités inférieures. Par conséquent, nous formulons un problème d'optimisation pour maximiser le bénéfice pour chaque niveau de priorité, tout en veillant à ce que les MTs avec des niveaux de priorité faibles n'utilisent pas les ressources qui pourraient être utilisées par les MTs avec des priorités plus élevées. En outre, le problème est simplifié en fixant le nombre de ressources qui devraient être allouées aux MTs avec des niveaux de priorité plus élevés. Cela pourrait se faire en répartissant le problème en K problèmes, où K représente le nombre de niveaux de priorité, et résoudre chaque problème séquentiellement dans l'ordre décroissant des niveaux de priorité, c'est-à-dire K, K-1, ... 1. Le nombre de ressources affectées aux MTs avec des niveaux de priorité plus élevés est calculé en fonction des résultats d'association des MTs avec un niveau de priorité supérieur à k, où k est un niveau de priorité spécifique. Par conséquent, le problème simplifié pour chaque niveau de priorité est similaire au problème **P1** et peut donc être résolu en utilisant les solutions proposées dans les chapitres 2, 3 et 4.

Maximisation de l'efficacité énergétique

Nous formulons un problème qui vise à maximiser l'efficacité énergétique, c'est-à-dire le débit de données par unité de puissance. Le problème est similaire au problème **P1**, mais la fonction objective est l'efficacité énergétique plutôt que le profit centré sur l'utilisateur. Ainsi, la fonction de profit centrée sur l'utilisateur ($\overline{f}_{mn}$) dans le problème **P1** est remplacée par le terme $\frac{Q_m}{P_{mn}^{Tx}+pc_{mn}}$ qui est l'efficacité en puissance qui considère le débit de données demandé par le MT (Q_m) divisé par la quantité de puissance qui devrait être transmise par le réseau n pour servir le MT m (P_{mn}^{Tx}) plus puissance consommée par le MT m tout en étant connecté et desservi par le réseau n (pc_{mn}). De même, ce problème peut être résolu en utilisant les solutions proposées dans les chapitres 2, 3 et 4.

Résultats de simulations

La Figure 7 montre que les utilisateurs ayant une priorité plus élevée (SL3) ont une plus grande satisfaction. En outre, la performance de chaque algorithme est semblable à la performance remarquée dans les chapitres précédents. L'allocation optimale des ressources pour les MTs de haute priorité en trouve une utilisation efficace des ressources dans les réseaux. Par conséquent, la possibilité que les MTs avec une faible priorité soient associés à leur réseau préféré diminue. En conséquence, en adoptant la solution optimale, les MTs avec niveau de service (SL) 1, c'est-à-dire le niveau de priorité 1, bénéficient d'un service d'expérience proche de la solution basée sur les bénéfices.

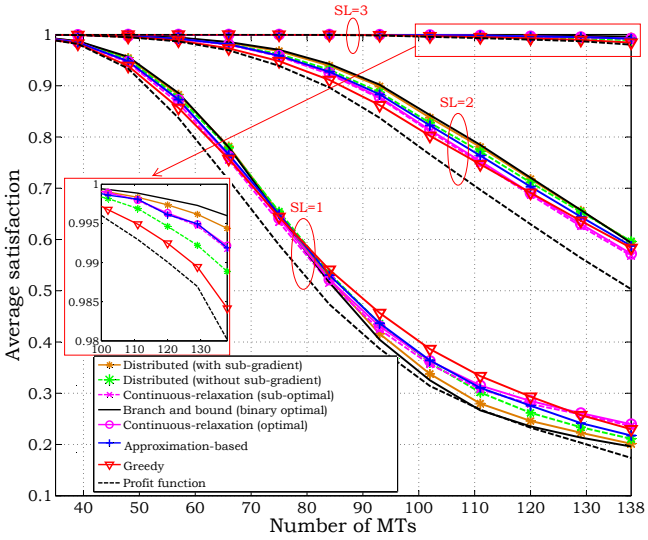


FIGURE 7 – Satisfaction moyenne pour différents niveaux de priorité.

La Figure 8 montre que la solution distribuée proposée avec la méthode subgradient maintient la performance la plus proche de la solution binaire optimale, suivie de la solution optimale de relaxation continue et de la solution basée sur l'approximation qui fonctionne approximativement de la même manière. Ensuite, toujours dans un ordre decroissant de performances, viennent la solution sous-optimale de relaxation continue, la solution greedy et la solution distribuée sans la méthode subgradient. Bien sûr, la solution basée sur les bénéfices est la moins performante.

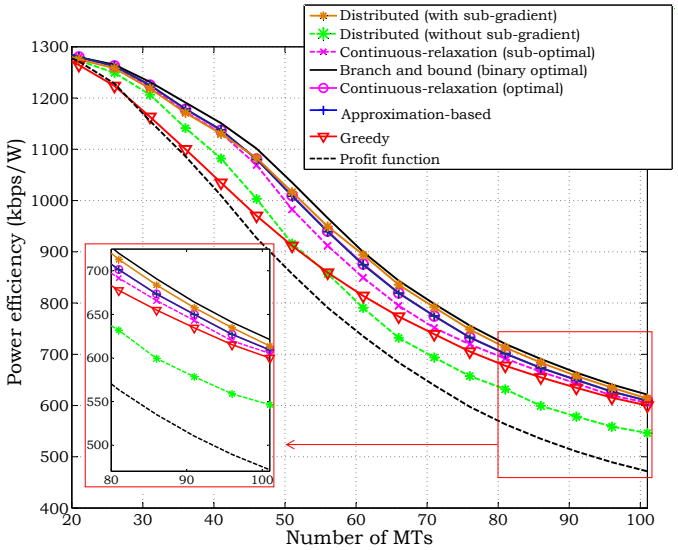


FIGURE 8 – Efficacité énergétique.

Conclusion

Dans cette thèse, nous avons proposé plusieurs solutions avec une complexité tolérable pour l'association d'utilisateurs et le problème d'allocation de ressources dans des réseaux hétérogènes. Les solutions proposées visent à fournir une vision ABC optimisée qui considère le système dans son ensemble tout en prenant en considération la satisfaction, les besoins et les préférences des utilisateurs.

Dans le chapitre 2, nous avons formulé un problème d'optimisation global pour maximiser le bénéfice global du système. Le problème formulé tient compte des besoins de l'utilisateur et des contraintes du réseau. Les résultats de simulations montrent que la solution proposée reflète effectivement les préférences des utilisateurs. En outre, le problème formulé améliore les performances globales du système. Cependant, la solution optimale du problème formulé est obtenue à l'aide de l'algorithme branch-and-bound, qui ne s'améliore pas avec l'augmentation du nombre de MTs dans le système. Par conséquent, les chapitres 3 et 4 proposent respectivement des solutions centralisées et distribuées du problème avec une complexité acceptable. Dans le chapitre 3, nous avons proposé quatre solutions centralisées pour le problème formulé dans le chapitre 2. Les deux premières solutions sont basées sur la relaxation continue du problème binaire formulé dans le chapitre 2. La première solution est la solution optimale de relaxation continue, qui a une complexité indéterminée, alors que la deuxième solution est la solution sous-optimale de relaxation continue qui a une complexité du temps polynomial. La troisième solution est basée sur l'approximation qui a une complexité inférieure aux solutions précédentes, alors que la dernière solution est la solution greedy simple qui a la plus faible complexité. Les résultats de simulations montrent que la solution basée sur l'approximation permet d'obtenir les performances les plus proches de celles de la solution binaire optimale. En outre, il a été démontré que l'approche de relaxation continue menace l'optimalité de la solution.

Dans le chapitre 4, nous avons proposé deux solutions distribuées pour le problème formulé dans le chapitre 2. En conséquence, la méthode de relaxation lagrangienne permet de relâcher le problème binaire et de le diviser en plusieurs problèmes de Knapsack, chaque problème de Knapsack est résolu par son réseau correspondant. La première solution distribuée dépend de la méthode de sous-gradient pour trouver des multiplicateurs lagrangiens presque optimisés. La deuxième solution distribuée résout simplement le problème sans essayer de trouver les multiplicateurs lagrangiens. Les résultats de simulations montrent que la solution distribuée avec la méthode de sous-gradient réalise les meilleures performances parmi toutes les solutions précédentes, alors que la deuxième solution distribuée atteint des performances acceptables avec une complexité inférieure.

Dans le chapitre 5, nous avons utilisé le problème formulé dans le chapitre 2 et ses solutions correspondantes présentées dans les chapitres 3 et 4, avec de nouvelles perspectives. La première perspective est le problème de l'allocation ressources à des utilisateurs ayant des priorités différentes. La deuxième perspective est le problème de maximisation de l'efficacité énergétique.

Table of Contents

Acronyms

2G	Second Generation
3G	Third Generation
4G	Fourth Generation
5G	Fifth Generation
ABC	Always Best Connected
AP	Access Point
AuC	Authentication Center
BC	BS Controller
BLP	Binary Linear Problem
BS	Base Station
CAPEX	Capital Expenditure
CCK	Complementary Code Keying
CDMA	Code Division Multiple Access
CSMA/CA	Carrier Sense Multiple Access with Collision Avoidance
DSSS	Direct-Sequence Spread Spectrum
EDGE	Enhanced Data for GSM Evolution
EIA IS	Electronic Industries Association Interim Standard
EIR	Equipment Identity Register
eNodeB	Evolved Node B
EPA	Equal Power Allocation
ETSI	European Telecommunications Standard Institute
FTP	File Transfer Protocol
GAP	Generalized Assignment Problem
GGSN	Serving GPRS Support Node
GMSC	Gateway Mobile Switching Center
GMSK	Gaussian Minimum Shift Keying
GPRS	General Packet Radio System
GSM	Global System for Mobile communication
HLR	Home Location Register
HO	Handover

HSDPA	High Speed Downlink Packet Access
HWN	Heterogeneous Wireless Network
IP	Internet Protocol
ISDN	Integrated Services Digital Network
ISM	Industrial Scientific and Medical
LAN	Local Area Network
LTE	Long-Term Evolution
M2M	Machine to Machine
MAC	Medium Access Control
MIH	Media-Independent Handover
MIMO	Multiple Input Multiple Output
MIS	Media-Independent Services
MISF	MIS Framework
MME	Mobility Management Entity
MOS	Mean Opinion Score
MSC	Mobile Switching Center
MT	Mobile Terminal
NIC	Network Interface Card
NP	Non-deterministic Polynomial-time
O&M	Operation and Maintenance
OFDM	Orthogonal Frequency Division Multiplexing
OPEX	Operation Expense
PHY	Physical Layer
PoA	Point of Attachment
PoS	Point of Service
PSK	Phase Shift Keying
PSN	Public Safety Network
PSTN	Public Switched Telephone Network
QoE	Quality of Experience
QoS	Quality of Service
RB	Resource Block
RRSS	Relative Received Signal Strength
RSS	Received Signal Strength
RSSI	Received Signal Strength Indicator
SAP	Service Access Point
SB	Scheduling Block
SC	Switching Center
SDN	Software-Defined Networking
SDRAN	Software-Defined Radio Access Networking
SGSN	Serving GPRS Support Node
S-GW	Serving Gateway
SHF	Super High Frequency
SINR	Signal to Interference Noise Ratio
SIP	Session Initiation Protocol
SL	Service Level
SLA	Service Level Agreement

TDMA	Time Division Multiple Access
TIA	Telecommunications Industry Association
TTI	Transmission Time Interval
UHF	Ultra High Frequency
UMTS	Universal Mobile Telecommunication System
VHO	Vertical Handover
VLR	Visitor Location Registry
VoIP	Voice over IP
Wi-Fi	Wireless Fidelity
WLAN	Wireless Local Area Network
WWAN	Wireless Wide Area Network

List of Figures

List of Tables

List of Algorithms

Introduction

This introduction provides the thesis overview, presenting a brief history about wireless access technologies evolution during the last two decades, and the thesis motivation and objectives. Moreover, the novelty of this thesis is highlighted. Finally, the thesis organization is provided.

Brief History

In the last two decades, wireless and cellular networks have experienced magnificent evolution. Network providers and operators have served billions of subscribers by supplying them with different communication services. The development of the telecommunication industry has direct impact on the people's daily life. Users depend on their mobile terminals (MTs) to communicate with their relatives, colleagues, friends, emergency services, companies, *etc.* Therefore, MTs are seen as the "digital interface" of the people, enabling communication between people at an acceptable monetary cost any time and any place.

The new telecommunication story began in the 1980s upon the introduction of the first analogue cellular system, being very limited in mobility, monetary cost, and coverage. The main telecommunication revolution started with the deployment of the second generation (2G) global system for mobile communication (GSM) which was introduced by the European telecommunications standard institute (ETSI). The main aspect for the global GSM success is that it allows users across countries to communicate with each other. Later, upon the deployment of the general packet radio service (GPRS), data services were integrated with GSM. The enhanced data for GSM evolution (EDGE) is considered as a transitional step from 2G to third generation (3G) systems, by introducing optimized experience for data services.

In the last decade, universal mobile telecommunication system (UMTS) was introduced as a major next-level communication system, overcoming the 2G world wide success. The major improvement is the higher data rates that allows better Internet access experience, video calls, file transfers, real-time applications, *etc.* During the same time period, MTs started to be equipped with cameras (photo and video), audio players, and large memories. Therefore, data hungry applications became a necessary requirement for each user. To face this excessive growth in data traffic requirement, 3G systems kept evolving towards the fourth generation (4G) long-term evolution (LTE) communication system. The major enhancements are providing high throughput, low latency, and higher mobility support.

Several other wireless access technologies have been also evolving simulta-

neously with cellular networks to provide users with different services. For example, wireless wide area network (WWAN) is another vastly growing technology providing high-capacity wide area coverage. The IEEE 802.11 wireless local area networks (WLANs) have been evolving rapidly to provide local-area high-speed wireless access. Since 1999, the 802.11 standard has been developing from the original 802.11 protocol to provide high data rate support (up to 866 Mbit/s). Moreover, other 802.11 amendments have been proposed to enhance the interoperability between different WLAN access points (APs) and supply better service for users.

Thesis Motivation and Objectives

Nowadays, in communication research fields, studies are mainly considered at the core network level to provide and deploy an all Internet protocol (IP) packet-based network architecture. New research fields have been focusing on how MTs will respond to the variety of wireless access technologies available at the same time. In this context, joint user association and resource allocation studies have been introduced into research fields. The user association problem deals with the aspect of choosing the network the MT should be associated to upon having multiple available access technologies, while resource allocation deals with the amount of resources that should be allocated by networks for each MT. Therefore, the effect of the user association and resource allocation algorithms should be studied and analyzed on a system-wide scale.

This dissertation is precisely encouraged by the new all IP network view of the next generation wireless communication networks, where the same core network is used by various wireless access technologies, leading to remarkable enhancements in the user experience and the overall performance of the system.

With the recent widespread deployment of the fourth generation (4G) wireless communication systems, the fifth generation (5G) mobile and wireless communication technologies are emerging into research fields. It is indicated that the expansion of the wireless data traffic requirements exceeds the capacity growth rate of new wireless access technologies [HQ14]. That is, modern modulation and coding schemes and access technologies can not provide the huge amount of requested data rates.

Cellular networks and WLANs are being evolved simultaneously to provide users with different services. WLANs are evolving rapidly to provide local-area high-speed wireless access. Cellular networks are another vastly growing technology that provides higher-capacity wide area coverage. Both technologies co-exist, with coverage areas overlapping with each other forming some sort of hybrid wireless access media. The efficiency of wireless links is approaching its theoretical limits, and the amount of requested data rate is severely increasing. Therefore, next-generation mobile wireless networks are moving towards heterogeneous architectures usually referred to as heterogeneous wireless networks (HWNs) [HQ14]. In HWNs, users have the right to connect to different types of radio access technologies like LTE base stations (BSs) or Wi-Fi APs. Such architecture increases the capacity of the system by reducing the number of users competing for resources at BSs. Moreover, HWNs provide flexible choices for the end user to transfer his ongoing data session to a more preferable network that satisfies his needs.

To access the Internet through HWNs, current MTs are equipped with multiple wireless access network interfaces. One type of terminals widely used nowadays is that with multiple data interfaces but can benefit from a single interface at a time, usually referred to as a multi-mode terminal [AM16]. By contrast, multi-homed terminals use multiple interfaces to share the load requested by a single MT. However, a realistic implementation for the multi-homing scenario is still far from deployment and imposes extra complexity on the system. Therefore, the multi-mode terminals are considered. Upon using multi-mode terminals, transferring an ongoing active connection to a new network is probably desired. The transfer in connection could be due to the user mobility, network congestion, user equipment status, *etc.* The process of transferring an active connection between networks is called handover (HO). If the networks that are participating in the HO are of different access technologies, *e.g.* handoff from an LTE BS to Wi-Fi AP, the connection transfer is usually referred to as vertical HO (VHO) [WK13].

In fact, combining and integrating different types of access technologies in HWNs provides flexible choices for the user to associate with his most preferred available network. In general, users prefer to be associated with the network that provides lower power consumption, better signal quality, better quality of service (QoS), security, *etc.* Consequently, HWNs are usually accompanied with the concept of always best connected (ABC) [GJ03], which is the process of being connected to the best available network at any time. However, the ABC concept is usually taken from the user perspective to rank candidate networks and connect to the best one. Usually, ABC-based network selection algorithms do not consider the limited resources of the networks and the effect of the HO algorithm on the system. Therefore, it is essential in HWNs to pave the way for an optimized context-aware ABC scheme that considers both user and network requirements.

Transferring the connection between different networks develops three main issues. The first one is to decide to which network the connection should be transferred, taking into consideration the user preferences and needs. Collecting information needed for the HO decision leads us to the second issue which is establishing a reliable infrastructure that provides a centralized global supervision on different network entities while maintaining cooperation and information exchange between them. While the last issue is to determine the amount of resources that should be allocated for each user based on the access technology of the networks, the channel condition between MTs and networks, the amount of resources requested by all MTs, in addition to the maximum capacity of each network. In this thesis, we discuss solutions for these issues.

To sum it up, the main objective of the thesis is to propose solutions with tolerable complexity for the user association and resource allocation problem in heterogeneous networks. The solutions aim to provide an optimized ABC vision that considers the system as a whole while making decisions. Therefore, users' satisfaction, needs, and preferences are considered.

Novelty

According to the research environment, this thesis claims novelty upon introducing new ideas, solutions, and algorithms. First, we discuss a new standard that

intends to provide a platform for information exchange, resource allocation, and handover management in HWNs. Then, we introduce new architecture that depends on a centralized entity to make decisions. LTE BSs and Wi-Fi APs are considered as candidate access technologies in HWNs.

In order to optimize user experience, it is necessary to provide a mechanism to enumerate user satisfaction, and specify the parameters involved within the satisfaction metric. Based on the literature review, the received signal strength (RSS) and the MT power consumption are considered as the most popular aspects for network selection decision making. Therefore, a profit function is introduced to combine these two parameters. Extra parameters could be easily added to the profit function (such as monetary cost). Other essential parameters such as signal to interference noise ratio (SINR) and the requested data rate by MTs, are also considered while formulating the optimization problem.

To optimize the overall user satisfaction in the system, a novel user association and resource allocation optimization problem is formulated. The formulated problem considers the channel conditions between MTs and networks, the amount of data rate requested by MTs, network capacity constraints, and network resource allocation constraints. However the formulated problem is of high complexity (NP-hard).

In order to provide a solution with acceptable complexity, three new centralized approximation algorithms are proposed. The complexity of each algorithm is studied, and the performance of these algorithms is compared to the optimal NP-hard solution.

Additionally, this dissertation claims another innovative aspect that aims at providing a distributed solution for the formulated optimization problem. The main idea is to process several optimization problems with low complexity instead of processing the main formulated optimization problem.

Another novel requirement is that future networks should be able to deal with users having different priorities. Therefore, a new optimization problem is formulated and solved in order to ensure better service for users with high priorities.

Finally, this thesis also addresses the power efficiency maximization problem in HWNs. Mainly, the optimization problem objective is shifted towards power efficiency. This objective is extremely important for communication system operational cost, having also significant impacts on the environment.

Thesis Organization

This thesis is organized in 5 chapters, excluding this introduction and the conclusion.

In Chapter 1, the main wireless access technologies involved in the studied heterogeneous wireless systems are briefly described, including their evolution, characteristics and architecture. Then the Wi-Fi offloading and the ABC concepts are discussed. A state of the art about existing network selection algorithms, in addition to user association and resource allocation algorithms in HWNs is presented. Finally, we come up with a conclusion regarding the main aspects that should be considered in this thesis.

In Chapter 2, we discuss an existing standardized framework and architecture that defines the communication mechanism between different network entities (servers, MTs, networks). Based on this framework, we will formulate a user association and resource allocation optimization problem. The algorithm runs on a server that maintains global view on the system. Our decision algorithm is based on two attributes: the power consumption at the MT and the signal quality at the MT. We use a reliable power consumption model to estimate the consumed power at the MTs in different wireless networks. The formulated optimization problem considers user preferences, requirements, and network limitations and constraints.

In Chapter 3, we discuss and propose multiple centralized solutions for the user association and resource allocation optimization problem proposed in Chapter 2. First, we propose two new solutions based on the linear-programming-relaxation of the formulated problem. The first proposed solution is with undetermined complexity, and is considered as the optimal solution based on the linear-programming-relaxation methodology. The second solution is with determined polynomial-time complexity, and is considered as a sub-optimal solution based on the linear-programming-relaxation approach. Then, a novel approximation-based solution is proposed to approximate the original optimization problem proposed in Chapter 2. In addition, a new simple greedy heuristic algorithm is also proposed. The performance of the approximation-based solution and greedy solution is boosted through a new proposed efficiency factor that estimates the gain contributed upon associating users to networks. The efficiency factor considers the data rate requirement of users, and the channel conditions between the MT and the BS or AP.

In Chapter 4, we propose two novel distributed user association and resource allocation solutions for HWNs. Mainly, the optimization problem formulated in Chapter 2 is distributed into several easier-to-solve problems, where each problem is independently solved by its corresponding network in an acceptable amount of time. The first distributed solution is based on the Lagrangian-relaxation of the optimization problem, and on the sub-gradient method that finds near-optimal Lagrangian multipliers. On the other hand, the second distributed solution is similar to the first one but without the sub-gradient method. The advantage of the second solution is that it eliminates the iterative sub-gradient phase, which leads to a lower complexity, and permits studying the effect of the sub-gradient method efficiently.

In Chapter 5, we discuss new direct applications for the optimization problem proposed in Chapter 2, and its corresponding solutions proposed in Chapters 3 and 4. First, we propose a new application for the priority-based user association and resource allocation problem in a system where users have different priorities. We formulate a new priority-based optimization problem. Then we simplify it into a new problem with form similar to the problem simplified in Chapter 2, and therefore could be solved with the solutions proposed in this thesis. Moreover, we discuss new perspective for the user association and resource allocation problem to enhance the overall power efficiency (data rate per unit power) in the system.

Finally, the main contributions of the thesis are summarized and future perspectives are proposed.

List of publications

Journals:

- M. Zalghout, A. Khalil, M. Crussière, S. Abdul-Nabi, and J.-F. Hélard, "Context-Aware and Priority-Based User Association and Resource Allocation in Heterogeneous Wireless Networks", *submitted to IEEE Transactions on Mobile Computing, under review.*
- M. Zalghout, A. Khalil, M. Crussière, S. Abdul-Nabi, and J.-F. Hélard, "Novel Distributed Resource and Network Assignment Solution in Heterogeneous Wireless Networks", *under preparation, to be submitted to the IEEE Transactions on Wireless Communications.*

International Conferences:

- M. Zalghout, A. Khalil, M. Crussière, S. Abdul-Nabi, and M. Hélard, "SDRAN-based User Association and Resource Allocation in Heterogeneous Wireless Networks," *in IEEE Wireless Communications and Networking Conference (WCNC), Doha, Qatar, pp. 1–7, April 2016.*
- M. Zalghout, S. Abdul-Nabi, A. Khalil, M. Hélard, and M. Crussière, "Optimizing Context-Aware Resource and Network Assignment in Heterogeneous Wireless Networks," *in IEEE Wireless Communications and Networking Conference (WCNC), San Francisco, United States of America, March 2017.*
- M. Zalghout, J.-F. Hélard, A. Khalil, S. Abdul-Nabi, and M. Crussière, "Load-Aware Power Efficiency Maximization in Heterogeneous Wireless Networks," *in 24th IEEE International Conference on Telecommunications (ICT), Limassol, Cyprus, May 2017.*
- M. Zalghout, M. Crussière, S. Abdul-Nabi, J.-F. Hélard, and A. Khalil, "Context-Aware Network Selection Algorithm for Heterogeneous Wireless Networks," *in IEEE Sensors Networks Smart and Emerging Technologies (SENSET), Beirut, Lebanon, Sep. 2017.*
- M. Zalghout, J.-F. Hélard, M. Crussière, S. Abdul-Nabi, and A. Khalil, "A Greedy Heuristic Algorithm for Context-Aware User Association and Resource Allocation in Heterogeneous Wireless Networks," *in 86th IEEE Vehicular Technology Conference (VTC), Toronto, Canada, Sep. 2017.*

Chapter 1

Heterogeneous Wireless Networks

1.1 Introduction

In this chapter, the main wireless access technologies involved in the studied HWN are briefly described, including their evolution, characteristics and architecture. Then the Wi-Fi offloading and the ABC concepts are discussed. A state of the art about existing network selection algorithms, in addition to user association and resource allocation algorithms in HWNs is presented. Finally, we come up with a conclusion regarding the main aspects that should be considered in this thesis.

1.2 Wireless local area networks

1.2.1 Architecture

A WLAN is a network that connects several equipments using the wireless media within a bounded region such as a company, university, house, hospital, *etc.* WLANs give users the ability to move around inside the local coverage region while being associated to the local network or even the Internet.

Initially, WLANs were proposed to provide a wireless substitute for the wired computer networks that are usually referred to as local area networks (LANs). Therefore, the WLAN main properties are similar to the LAN ones, mainly in terms of protocol layers and supported services. A basic architecture of a WLAN system is shown in Figure 1.1. Each wireless communication device communicates with its corresponding wireless AP. In order to communicate with remote devices on other APs, a network that connects all APs should be established. Normally, routers are needed to forward the packets.

Current WLANs are based on IEEE 802.11 standards and are advertised under the wireless fidelity (Wi-Fi) trademark. The IEEE 802.11 standard defines a set of media access control (MAC) and physical layer (PHY) specifications for a Wi-Fi network. The core specifications of the 802.11 standard was released in 1997, and has been evolving since then.

The IEEE 802.11 standard provides two basic operating modes: the infrastructured and ad hoc.

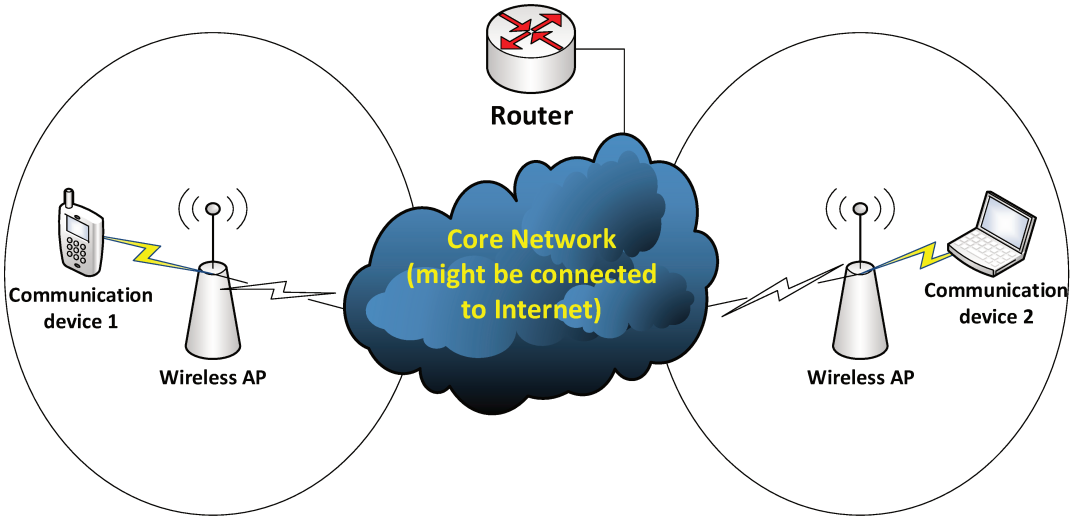


FIGURE 1.1 – Basic architecture of a WLAN system.

– Infra-structured mode: Most Wi-Fi networks are deployed in infrastructure mode where wireless users communicate through an AP as shown in Figure 1.2. Ordinarily, APs are fixed and serve users within their coverage range. To provide a connection to the Internet, APs should be connected to the backbone network through a wired or a wireless link. Wireless users, such as mobile phones, tablets, laptops *etc.* should connect to the AP to join the network.

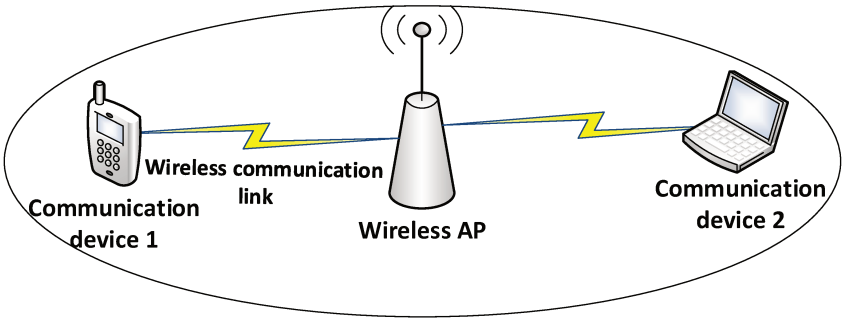


FIGURE 1.2 – Infra-structure mode for WLANs.

– Ad hoc mode: In the ad hoc mode, users can establish direct communication links with each other using the wireless interface without communicating with the AP as shown in Figure 1.3. The ad hoc mode is useful for transferring data between nearly-located users. In this context, coverage extension can be established ; a user connected to both the backbone and ad hoc network could serve as a gateway for other users connected through the ad hoc connection.

The main service provided by Wi-Fi networks is data transfer. However, other services, as voice and video applications, are being supported more often due to the high capacity provided in these networks.

Wi-Fi networks are mainly meant to be deployed in indoor environments. Thus,

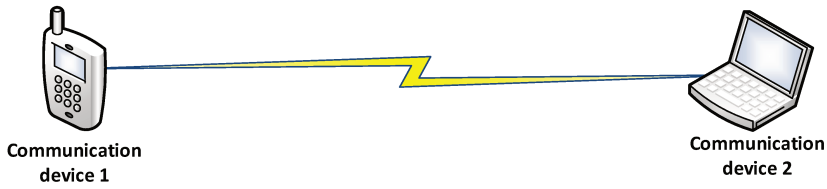


FIGURE 1.3 – Ad hoc mode for WLANs.

the coverage of these networks is limited to typically several 10 meters. Besides, since Wi-Fi networks operate on the limited unlicensed band, the maximum allowed transmission power is quite low (0.2-1 W) to reduce interference with other APs operating on the same band. Thus, the limited transmission power also plays a vital role in reducing the coverage area of these networks. However, this potential drawback limits the number of users being connected to an AP, which in turns supplies connected users with higher data rates. To sum it up, Wi-Fi networks supply users with high data rates when compared to other types of access technologies ; however, this property is accomplished by sacrificing other aspects, like coverage.

To extend the coverage range of Wi-Fi APs while maintaining user mobility, the 802.11 standard supports handover in homogeneous networks, *i.e.* handover between Wi-Fi APs. The coverage range of several interconnected APs could extend from a range as small as multiple rooms to as large as square kilometers. To cover a large area, a group of interconnected APs with overlapping coverage regions is required.

Nowadays, the performance of Wi-Fi networks is remarkable ; the performance of all supported services is excellent. Currently, other types of radio access technologies are taking advantage of this remarkable performance. For example, when a MT being connected to a cellular network detects the presence of a Wi-Fi network, it will probably handover to the Wi-Fi to support the MT with an expected higher QoS. However, this aspect is not always preferable ; mainly, users with high mobility prefers cellular networks. Moreover, it is not always guaranteed that the Wi-Fi AP could provide better performance than other systems. For example, the case where the Wi-Fi AP is highly congested.

One main advantage for Wi-Fi networks is the low cost at the operational and maintenance level. When compared to cellular networks, the infrastructure equipment cost is much lower. Therefore, Wi-Fi networks are considered as a low-cost competitive radio access technology in low-mobility hotspot areas.

An important disadvantage of Wi-Fi networks is that the throughput of a MT decreases when the distance between a MT and an AP increases. Therefore, Wi-Fi networks deployment should be carefully designed, in order to maintain uniform throughput in all the area covered by Wi-Fi APs.

1.2.2 Radio interface

The IEEE 802.11 family consists of a several half-duplex wireless modulation techniques. The 802.11-1997 [IEE97] is the basic Wi-Fi networks standard in the family and the 802.11b standard is the first widely-spread one. There are currently

five main standards in the family: 802.11a [IEE99a], 802.11b [IEE99b], 802.11g [IEE03], 802.11n [IEE09], and 802.11ac [IEE13].

Wi-Fi usually uses the 2.4 GHz ultra high frequency (UHF) and 5 GHz super high frequency (SHF) bands. Mainly, Wi-Fi uses the unlicensed industrial, scientific and medical (ISM) radio bands. Due to this choice of frequency bands, Wi-Fi devices usually suffer interference from devices operating on these frequencies such that microwave ovens and Bluetooth devices. To mitigate the effect of this interference, Wi-Fi devices oftenly use orthogonal frequency-division multiplexing (OFDM) signaling methodology. The 5 GHz band offers more than 23 non-overlapping channels while the 2.4 GHz band offers only three non-overlapping channels.

In order to connect to a Wi-Fi APs, the computing device should be equipped with a wireless network interface controller (NIC). A station is defined as a combination of the computing device and NIC. All stations share the same radio frequency communication channel. To transmit data, all the NICs use a carrier wave, thus, all transmissions are received by all the stations within the coverage range communicating on this channel. Data are organized in packets similar to the communication protocol used on an Ethernet link. All Wi-Fi standards use the carrier sense multiple access with collision avoidance (CSMA/CA) protocol for media sharing.

Initially, the modulation method that was used in 802.11 is phase-shift keying (PSK). The 802.11 b products that have been advertised in the market in 1999 depends on a direct extension of the direct-sequence spread spectrum (DSSS) modulation technique defined in the original standard.Technically, the 802.11b standard uses the complementary code keying (CCK) as its modulation technique because it allows higher data speeds, and is less susceptible to multi-path propagation interference. The more recent Wi-Fi standards use OFDM and multiple-input and multiple-output (MIMO) techniques which allow much higher data rates. Table 1.1 presents the main parameters and specifications of the physical layer in the main 802.11 standards.

Table 1.1 – 802.11 networks physical layer specifications

802.11 protocol	Release date	Frequency (GHz)	Bandwidth (MHz)	Data rate (Mbit/s)	Modulation
a	1999	5	20	54	OFDM
b	1999	2.4	22	11	DSSS
g	2003	2.4	20	54	OFDM
n	2009	2.4/5	20	72	MIMO OFDM
			40	150	
ac	2013	5	20	96	MIMO OFDM
			40	200	
			80	433	
			160	866	

The IEEE 802.11 family also includes amendments that extend the scope of existing standards. Although the 802.11 community defines a lot of amendments,

the main amendments and those that are related to the context of this dissertation are presented in Table 1.2.

TABLE 1.2 – 802.11 main amendments

802.11 Amendment	Short description
e	QoS definition, also supports packet bursting (2005)
f	Inter-Access Point Protocol (2003), interoperability between APs (withdrawn in 2006)
i	Security enhancement (2004)
r	Fast handover between APs (2008)
s	Mesh networking, extended service set (ESS) (2011)
u	Improvements related to HotSpots and authorizing non-IEEE users, e.g., offloading from cellular networks (2011)

1.3 Cellular networks

1.3.1 Architecture

The generalized architecture of cellular networks, shown in Figure 1.4, is characterized mainly by the BS, the controller of the BSs (BC), and the switching center (SC) which is considered as the most important element because it is responsible for processing all the communications and it is equipped with a data base system. This architecture enables the communication between remote MTS, and between MTs and other networks as public switched telephone network (PSTN) and integrated services digital network (ISDN). The operator uses the operation and maintenance (O&M) center to manage the network. This generalized system architecture could be divided in two main ones ; the hierarchical architecture where the BC controls multiple BSs (used in 2G and 3G communication systems), and the architecture where BSs integrate mobility and control functionalities while being directly connected to the SC (4G communication systems).

Currently, the GSM is the most widely-spread cellular system that supports both circuit-switched and packet-switched services. An overview of the connections in GSM and GPRS (2G), UMTS (3G), and LTE (4G) is shown in Figure 1.5 [3GP10b][3GP10a] where it is possible to observe the various interfaces between the data base system, network controllers, and switching systems.

The mobile switching center (MSC), which is connected to the main data bases in the system, is responsible for maintaining the circuit-switched services. The MSC is referred to as gateway mobile switching center (GMSC) when it plays the role of gateway between the current serving system and other fixed or mobile networks.

The most basic service provided by cellular networks is voice calls. However, cellular networks are currently supporting reliable data and multimedia services ranging from a small machine to machine (M2M) synchronization messages to a real-time video conference or virtual reality applications. The main network elements responsible for data services in the 2G/3G core network are the serving

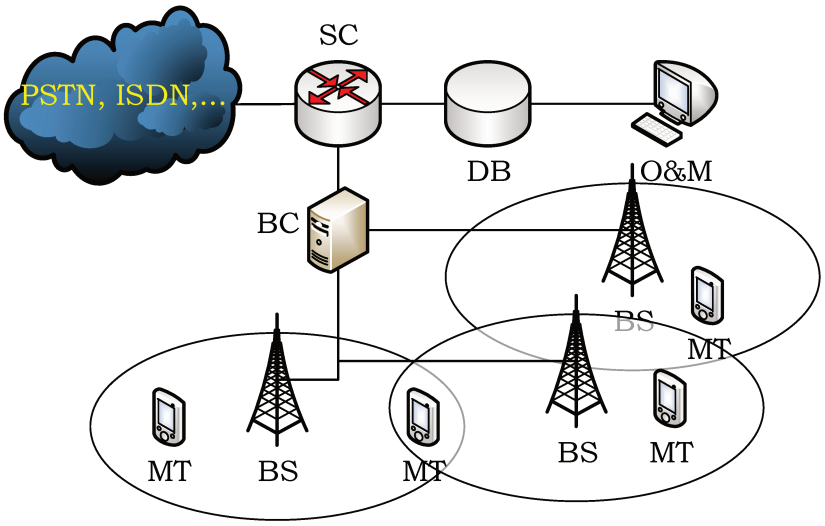


FIGURE 1.4 – Generalized architecture of cellular networks.

GPRS support node (SGSN) and gateway GPRS support node (GGSN). The SGSN is responsible for:

- user authentication,
- delivering packets from/to UEs within coverage range,
- packet routing,
- mobility management (handover and location management),
- logical link layer management,
- monetary charging functions ,

while the GGSN holds the responsibility of inter-networking the external network and the GPRS one. However, the most recent LTE system adopts a much simpler network architecture ; the evolved nodes B (eNodeBs) hold more network coordination responsibilities, and interact directly with each other (using the X2 interface) and with the serving gateway (S-GW). Here, the S-GW executes packet routing and mobility anchoring. The mobility management entity (MME) is responsible for various functionalities such as coordinating user mobility, authentication, and roaming [3GP10a].

All the subscribers that roam into the MSC are saved in the database of the visitor location register (VLR). The home location register (HLR) is a centralized database that queries information about each MT that have the authority to benefit from the core network. The authentication center (AuC) is responsible for authenticating all the MTs that attempt to connect to the core network. When the MT is authenticated, the HLR is allowed to manage the services of this MT. In case of GSM, once authenticated, an encryption key is generated to encrypt all the communications between the MT and the GSM core network (voice calls, messages, *etc.*). The equipment identity register (EIR) queries a list of MTs that are banned from using the network. This is useful in case of stolen MTs. The EIR is usually combined with the HLR.

The main goal of cellular networks is to cover wide area while supporting high user mobility. For this reason, these networks are designed in a way that ensures

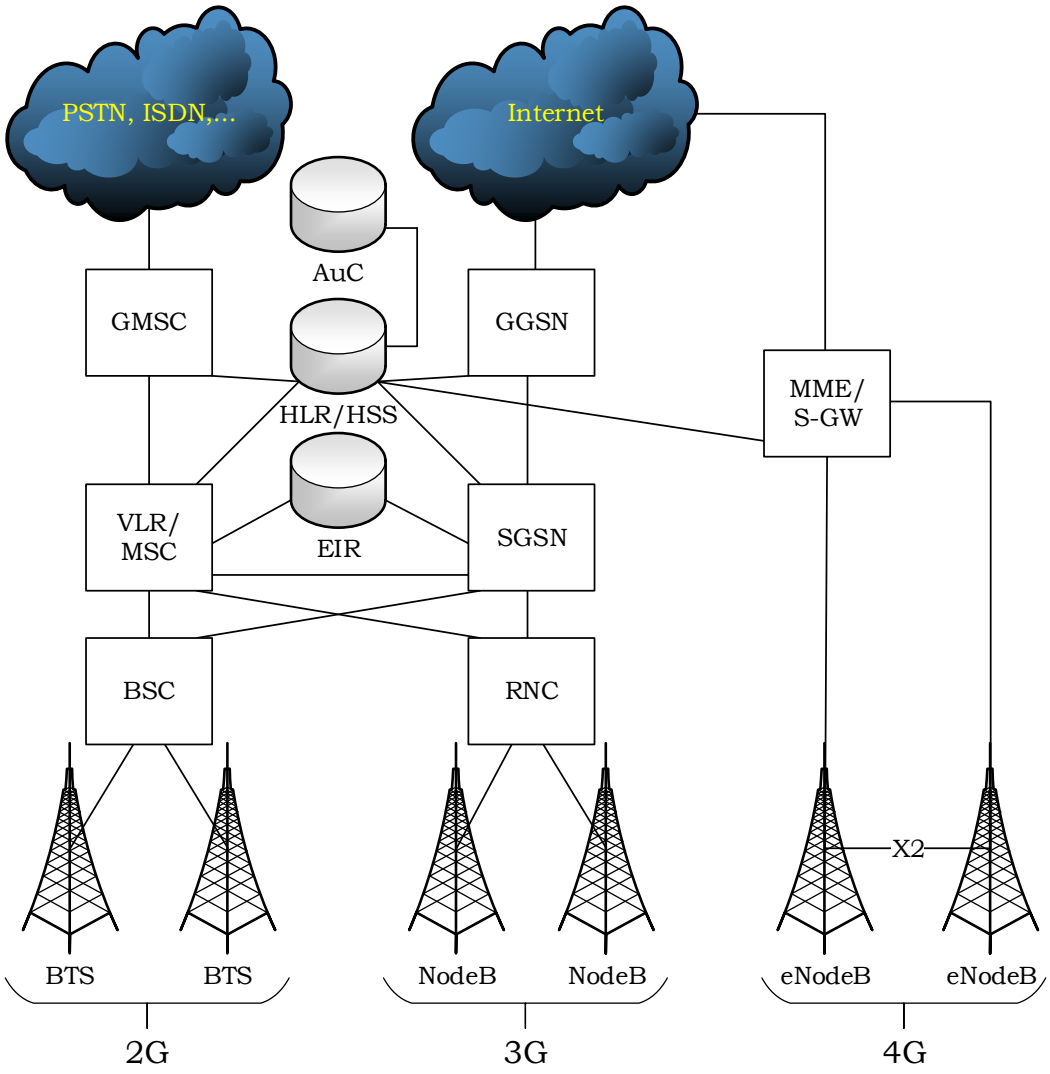


FIGURE 1.5 – Overview of the architecture of 2G, 3G, and 4G communication systems.

"good" service for users experiencing high mobility. The importance of cellular networks is mainly characterized by the large geographical coverage region and the huge number of users served by these networks.

In terms of capacity, cellular networks can be easily adapted according to the required capacity. Cellular networks support multiple hierarchical structures. For example, a single cell could be adjusted to operate in a hot spot area, rural, urban, and suburban areas, by using what so called pico, micro, and macro cells. As the requested capacity increases largely, the amount of radio resources that is offered to users should be increased also. Usually, additional cells are installed to offer new radio resources for users. However, installing new cells increases the inter cell interference. To reduce the interference, the coverage region of each operating cell should be decreased, *i.e.* cell reduction strategy. Upon congestion in cellular networks, the data rate supplied to each user could be extremely low. Therefore, cellular networks could cooperate with Wi-Fi networks to supply better services

for users.

1.3.2 Radio interfaces

In cellular networks, the radio interface is the main component that has been enhancing throughout the years. A cellular generation increment is announced when the radio interface undergoes a noticeable major improvement.

Concerning the 2G systems, GSM has been upgraded to GPRS and then to EDGE. GPRS has introduced data communication to 2G systems, while EDGE increased the throughput that could be supplied to users. EDGE could supply data rates up to 384 kbps due to the usage of the new 8-phase shift keying (8PSK) modulation technique rather than the gaussian minimum shift keying (GMSK) modulation technique. This enhancement requires deploying new transceivers and software updates.

The telecommunications industries association / electronic industries association interim standard – 95 (TIA/EIA IS-95) family [OP98] have defined a second version of 2G communications systems. This version is called CDMAOne and it is based on code division multiple access (CDMA). CDMAOne operates on the 800 and 1900 MHz frequency bands, and each carrier has a bandwidth of 1.25 MHz. CDMAOne provides voice services and data services that could reach a speed of 14.4 kbps.

The 2.5G equivalent of CDMAOne is the IS-95B system which provides data services with data rate up to 64 kbps. CDMA2000 is considered as the 3G equivalent, and it enhances the data rate of its predecessor by increasing it to 307 kbps. CDMA2000 is fully compatible with its predecessor the IS-95B system.

One of the most important characteristics of the CDMA radio management is that the power is considered as a common resource shared between users ; power is allocated to each user while ensuring that the maximum interference threshold is not exceeded. To optimize the capacity of the cell and boost the performance of the network, fast power control in the downlink is considered mainly for indoor and low-speed outdoor environments. The network entity that holds the responsibility of managing resources should track the instantaneous quality of the channel between the BS and the user.

One of the main issues in cellular networks is to enhance the capacity in terms of number of users per cell per bandwidth unit (in Hz). In general, the capacity depends on the following factors:

- total transmission power,
- available bandwidth,
- channel quality,
- user behavior (mainly related to the requested data rate and QoS).

In order to enhance the capacity and the data rates, 3GPP has proposed the high speed downlink packet access (HSDPA) as an upgrade of the UMTS. HSDPA could provide data rates up to 14 Mbps, while its newest version HSDPA+ increases the data rate to 42 Mbps. HSDPA benefits from several communication advancements such as MIMO techniques, adaptive modulation and coding techniques, and optimized resource management schemes.

The LTE is considered as the fourth generation of mobile networks proposed

by 3GPP. LTE uses OFDM technology as the air interface, and its throughput depends on the channel bandwidth, among other things. The main reasons for the increased data rates in the newest 3GPP releases are the usage MIMO technology and new modulation and coding schemes.

The next step of the LTE technology is already defined as LTE-Advanced which is described in the Release 10 standard. LTE-Advanced is mainly characterized by faster protocol stacks, advanced MIMO technology, bandwidth aggregation up to 100 MHz, cooperation between several kinds of BSs (micro, pico, and femto), among other promising features. LTE-Advanced is expected to enhance the offered data rate, reduce inter-cell interference, supply better QoS and lower delays. LTE releases and their main specifications are summarized in Table 1.3.

Table 1.3 – LTE releases and specifications

Release	year	Main specification
8	2008	All-IP network, new OFDMA and MIMO radio interface. Not backwards compatible with previous CDMA interfaces.
9	2009	All-IP network enhancements, WiMAX and LTE/UMTS interoperability.
10	2011	LTE Advanced fulfilling based on 4G specifications. Backward compatibility with release 8.
11	2012	Heterogeneous networks improvements.
12	2015	Enhanced small cells and carrier aggregation.
13	2016	LTE in the unlicensed band. Machine-to-machine communications. LTE Advanced pro.

1.4 General comparison between systems

In this section, a general comparison is presented between Wi-Fi networks and cellular networks according to the main characteristics of each wireless access technology. A summary of these differences is presented in Table 1.4.

By studying the evolution of the wireless access technologies, it becomes obvious that next generation core network is moving towards a completely IP-based architecture, where different types of access technologies coordinate with each other. Therefore, in order to boost the performance of this integrated heterogeneous system, a resource management entity should coordinate the distribution of resources in these networks.

The main goal of the resource management entity is to serve users in the best way according to the users needs. Those needs could be based on the requirements of the services used by users (QoS, data rate, *etc.*) or according to the user preference (monetary cost, power consumption, *etc.*). Thus, the entity that will coordinate the user association and resource allocation should be aware about the needs of the users and the types of the networks and their characteristics.

Observing Table 1.4, the resource management awareness concept is present, where the higher performance levels dealing with mobility, throughput, cost, coverage or service type are found in different kind of systems.

TABLE 1.4 – Wi-Fi and cellular networks comparison

Characteristic	Wi-Fi	Cellular networks
System architecture	Simple	Complex
Service	Data	Voice-Data
Mobility	Low	High
Throughput	High	Low-Medium
Deployment environment	Indoor	Indoor-Outdoor
Spectrum	Licensed	Unlicensed
Capacity	High	High
Coverage	Low	High
Transmission power	Low	High
Deployment cost	Low	High
Operational cost	Low	High
Billing policy	Free-Volume	Time-Volume

1.5 Wi-Fi offloading

Nowadays, the information and communication technology research field is facing an explosion of data traffic, which is characterized by the unprecedented increase in the mobile data traffic due to the proliferation of smart communication devices. Current mobile networks could transmit data traffic at high rates due to the development of the radio access technologies. Recently, Cisco has announced that the overall world-wide mobile data traffic in 2016 has grown 74% [Cis16]. Moreover, it is predicted that the monthly mobile data traffic demand will reach 30.6 exabytes by 2020 where two-thirds of this traffic is due to video and audio services. This terrifying increase in demand is not only due to the growing number of smart communication devices, but also due to the emerging M2M technology.

To supply this wireless data traffic demand, one solution is to replace the currently installed mobile radio access technologies with the next generation mobile wireless networks. The second solution is to increase the number of BSs while reducing the cell size to increase the capacity of the overall wireless system. However, both solutions suffer from the large value of capital expenditure (CAPEX) and operation expense (OPEX). Hence, it is predicted that deploying heterogeneous networks is a must in any future wireless communication system [O. 09].

Therefore, it is essential to take advantage of the current installed wireless technologies to provide the requested data traffic in a heterogeneous network environment. Figure 1.6 provides an example of wireless access technologies cooperating with each other in heterogeneous networks ; a user equipment could access different types of wireless technologies. Research fields are exploring the opportunity of offloading data traffic from cellular networks to Wi-Fi APs, or to femtocells which is usually referred to as mobile data offloading. Knowing that Wi-Fi APs are widely deployed by users, Wi-Fi offloading is considered as a remarkable solution to fully utilize the capabilities of both cellular and Wi-Fi networks. Interestingly, several operators, including the famous ones (e.g., AT%T, VodafoneVerizon, AT&T, Verizon, *etc.*), have been deploying Wi-Fi offloading recently [Jun13].

Currently, facing the exponential data traffic increase is considered as the main

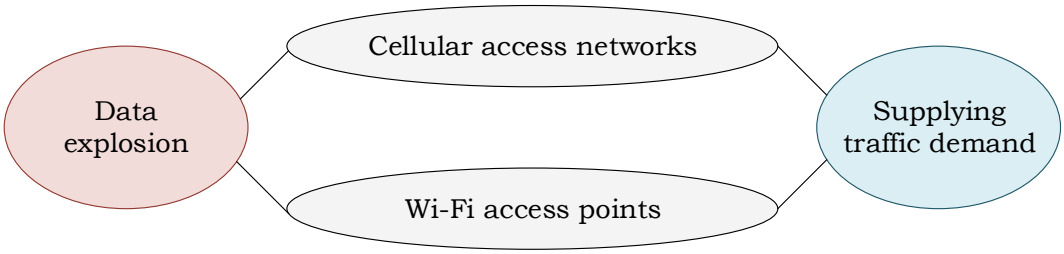


FIGURE 1.6 – Available wireless access technologies that could be used to mitigate the data explosion.

challenge of mobile operators. Wi-Fi offloading stands out as the main low-cost solution to reduce the traffic load on the cellular networks. Wi-Fi offloading is considered as the hybrid paradigm that takes advantage of the existing alternative communication channels.

Data traffic demands are increasing in an exponential way thanks to the increasing popularity of smart mobile devices and the introduction of affordable data plans by mobile network operators. Data hungry services including audio and video streaming, or cloud-based services, are becoming more and more popular among users. The predicted growth in mobile data traffic between 2011 and 2018, is expected to be three times faster than that of the fixed IP networks during the same period [Cis14]. Hence, the traditional cellular networks are trying to cope with this unprecedented traffic explosion. From the economical perspective, upgrading all installed cellular networks is expensive because it requires additional investments to install the latest infrastructure equipment.

One of the aspects that hinders the enhancement of cellular networks is the scarce licensed spectrum. Mobile operators are allowed to use a small portion of the radio spectrum due to frequency regulations. Hence, wireless networks could operate only in limited wireless frequency resources.

The indicated circumstances fostered the interest in alternative methods to alleviate the pressure facing cellular networks. An intuitive approach is to leverage the unused bandwidth across different wireless technologies. Mobile data offloading is characterized by the usage of a complementary wireless technology to transfer data that originally target cellular networks, in order to enhance several main performance indicators.

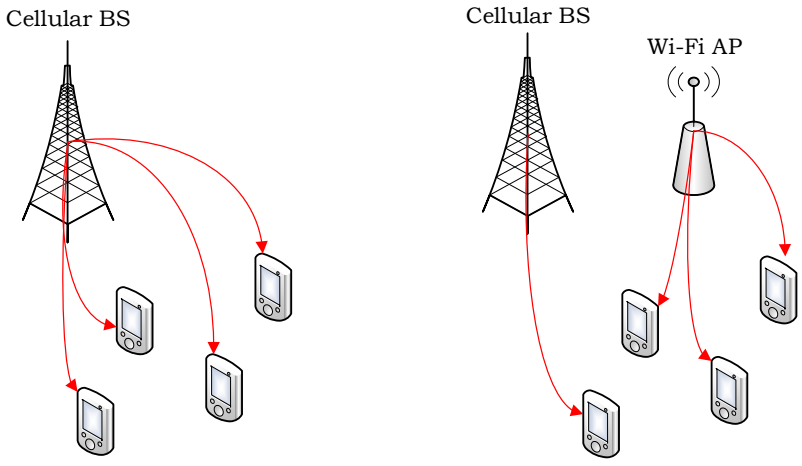
In addition to the mentioned advantages of Wi-Fi offloading, several additional benefits could be contributed due to the cooperation between complementary wireless access technologies. These benefits include:

- enhancing energy efficiency,
- delay reduction,
- increasing the overall throughput,
- coverage extension.

Wi-Fi offloading is usually described as a win-win strategy because these improvements affect both users and network operators [DHHL11]. Unfortunately, it is not easy to deploy such architectures, since several challenges need to be addressed. The main challenges are:

– user mobility,
– enable seamless handover,
– infrastructure coordination.

Figure 1.7 shows two main approaches to serve users in wireless networks. In Figure 1.7a, wireless users are only allowed to be served by one network (cellular BS), while Figure 1.7b shows the situation where Wi-Fi networks offloads some users from the cellular networks to reduce the congestion on the BS. Passing data traffic through Wi-Fi APs is considered as the main solution to reduce the load on cellular networks. MTs located within the coverage area of Wi-Fi APs could use these networks as an alternative to the cellular networks when they have active services. In case the cellular networks are congested, users could experience better service upon offloading to uncongested Wi-Fi networks. However, since the coverage of Wi-Fi networks is limited, the mobility of users is constrained within these cells. Famous cellular providers such as Orange, Verizon, T-Mobile, and AT&T have begun to install more Wi-Fi APs in their cellular networks to support Wi-Fi offloading due to the low monetary cost of installing Wi-Fi APs [Dat13]. Recently, an increased number of applications that support and enhance the Wi-Fi offloading procedures are developed for Android and IOS-based mobile devices, such as BabelTen [Bab] and iPass [iPa]. However, these applications do not solve the mentioned challenges because they prioritize Wi-Fi connections and networks with high RSS without considering the network conditions and user preferences.



(a) (b)

FIGURE 1.7 – (a) Without Wi-Fi offloading. (b) Wi-Fi offloading.

1.6 Always best connected

The widespread deployment of the 2G cellular networks in the 1990s contributed the paradigm of being always connected. The main concern was to supply users with wireless connection for voice services in outdoor, indoor, and mobile

environments. The incredible development of wireless networks in the past two decades provided several types of wireless access technologies that complement each other in different domains as presented in Section 1.5. The heterogeneous wireless environment contributed by the development of these technologies was followed by the development of the ancient concept of being always connected towards being always best connected. That is, not only being always connected, but also being connected to the best available network at any time.

The ABC concept is very broad because identifying the "best" network is actually based on several factors that might be related to the user preference, service preference, or provider preference, *etc.* With the ABC capabilities, MTs could select the access networks that might reduce the monetary cost for the users ; MTs could select the network that reduces the battery consumption ; connecting to network with high traffic overload could be avoided, and thus avoiding congestion and low data rates ; MTs could be associated with network providing the best signal quality, *etc.* Moreover, ABC could benefit the operators. When selecting the best network based on the providers' benefits, the strategy could be based on balancing the load in networks, and thus the congestion could be avoided, and the utilization of the networks is enhanced, hence, the overall data rate increases leading to increase in the revenue of the operators as well. Moreover, the network selection strategy deployed by operators could determine the number of Wi-Fi APs and cellular BSs that should be deployed to supply the data traffic requested by users. Therefore, the ABC concept is seen as a win-win strategy because it could be deployed in a way that enhances the user experience and increases the network operator revenues.

The context-awareness is an essential aspect in ABC. To take the user assignment decisions, contextual user-centric, service-centric, and network-centric informations are needed. The user-centric information is based on:

- the geographical location of users,
- user preferences.

The network-centric services is based on:

- the geographical location of the networks,
- the type of the network,
- amount of available resources,
- transmission power.

Furthermore, the service-centric contextual information is based on the data rate requested by the users.

In the following, a series of scenarios are designed to highlight the importance of the ABC concept in the user's daily life routines. Those scenarios are shown in Figure 1.8 and further illustrated in Table 1.5. In these daily life scenarios, a user with a multi-modal MT works in a company. On a typical working day, the user, who is currently located inside the company, is searching for some information using his MT (scenario (1)). Within this stage, the MT is located inside the transmission range of multiple LTE BSs and Wi-Fi APs. Then, during the morning break, the user moves to the near coffee shop, where the Wi-Fi there is only free for customers buying stuff there (scenario (2)). In the coffee shop, the user calls his friends to organize a trip in the weekend. After finishing his coffee break, the user takes a taxi to another company where he have a meeting (scenario (3)). However,

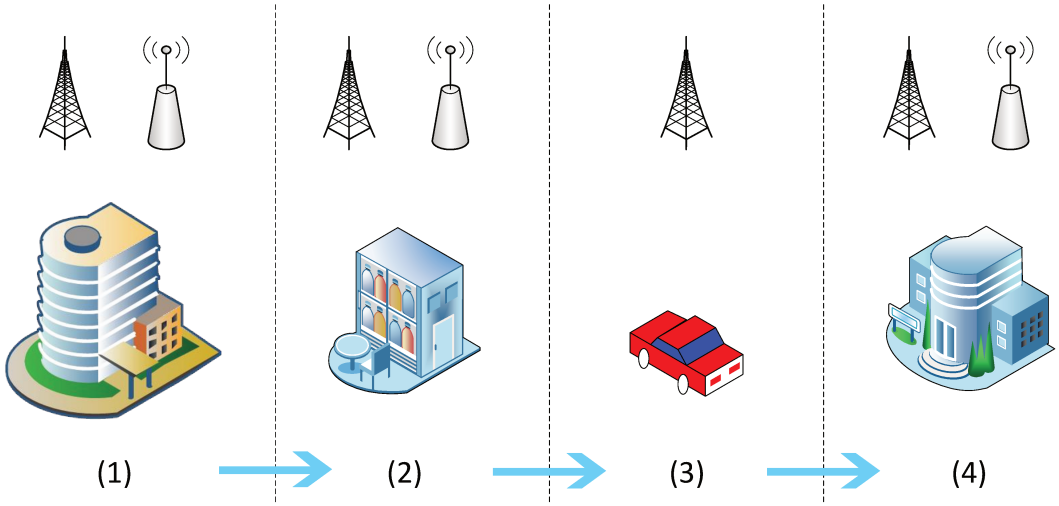


FIGURE 1.8 – Example of a series of scenarios for the ABC concept.

TABLE 1.5 – ABC scenarios information

Scenario	Wi-Fi coverage	LTE coverage	Service	Battery status
(1)	Yes (free)	Yes	WWW	good
(2)	Yes (not always free)	Yes	Voice conversation	good
(3)	No	Yes	Video conference	good
(4)	Yes (free)	Yes	-	Low

due to the traffic jam, the user realizes that he can not arrive on time, so he has to initiate a video conference while he is still inside the taxi. After arriving to the company (scenario (4)), the MT is about to be out of battery, but the user cannot charge the MT during the meeting.

In the example stated in Figure 1.8, as in a lot of other scenarios, there are multiple available networks that might be of the same or different types. In some scenarios, the best available network could be identified easily ; however, for other scenarios, the best network could not be identified easily. That is due to the fact that each network has its own characteristics that will be taken into consideration by the network selection entity. In order to maintain the selection of the best available network, the ABC solution depends on a platform that provides instantaneous information about the networks and users.

An effective ABC solution requires the following functional blocks: network discovery, information gathering, network selection, mobility management, *etc.* as shown in Figure 1.9. Those entities will be described briefly.

Network discovery: The first step towards selecting the best network is actually identifying all available networks. MTs are capable of finding available networks through the broadcast signals transmitted by networks. If the MT is responsible for taking network selection decisions, then the MT can keep these information for itself. However, if the network selection decision is taken by a remote entity, then this entity should have information about the available networks

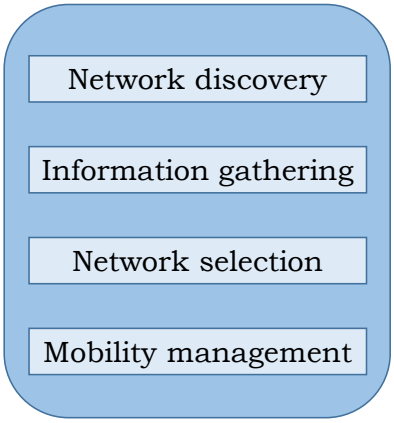


FIGURE 1.9 – Functional blocks of the ABC concept.

of each MT. An important issue here is when to initiate network discovery. This procedure could be initiated periodically or by some trigger events. Those triggers could be related to the MT's conditions, or to external conditions. For example, when the received signal strength drops below a certain threshold, the network discovery procedure is triggered. Or due to congestion in the current serving network, where the MT should find a new network to maintain its required service.

Information gathering: It is one of the main blocks of ABC. In order to maintain the association with the best network at all time, ABC should gather the contextual information discussed earlier. Gathering static, or semi-static information is not a problem. For example, the amount of total resources available at each network, network type, geographical location of networks, user preference, requested data rate ; these information could be supplied upon initiating a connection, and are assumed to be static during the connection interval. However, instantaneous information, such as MTs geographical location, received signal strength, number of free resources at each network, *etc.* Some information could be gathered based on direct feedback from MTs and networks, or could be estimated by networks. The geographical location of MTs could be estimated by networks, as well as the received signal strength.

Network selection: This block uses the gathered information in order to select the best network for each user. In the literature, there is a lot of parameters proposed to be used within the network selection block, in addition to the mechanism or the algorithm used to get a network selection decision based on several parameters. A state of the art about network selection algorithms will be presented later in this chapter (In Section 1.7).

Mobility management: This block handles the transition of an active connection between networks, *i.e.* HO. When the HO is between two networks of the same access technology, usually that technology itself is responsible for the HO. However, if the connection is transferred between networks of different types, *i.e.* from LTE BS to Wi-Fi AP, an external protocol is required to manage the HO. The mobility management block also provides a seamless HO where the continuity of the service is not threatened during the HO.

1.7 Network selection

In this section, we talk about the main parameters used to evaluate available networks. Then we talk about the criteria used in the literature to rank available networks. Finally, a conclusion about the network selection scheme is presented.

Network selection is devoted to decide to what network the MT should connect. It is an essential functionality that improves user experience in heterogeneous networks. If the network selection decision is taken by a MT, those terminals will take autonomous decisions to select the most appropriate networks among available ones. MTs will selfishly aim to maximize their own utility. However, if MTs do not consider the network states, their decisions may lead to a bad experience.

When multiple networks cover the same geographical area, deciding to which network the MT should connect is known as the network selection decision. The network selection procedures could be initiated upon session initialization, weak connection due to user mobility, or upon detecting new available networks.

1.7.1 Network selection parameters

For making network selection decisions, these are the main parameters used in literature.

Recieved signal strength (RSS): RSS, could be also referred to as received signal strength indicator (RSSI), is a classical and inevitable metric used to make handover decisions. RSS is directly related to the power received by the antenna of the MT from the available BSs and APs. The power level decreases as the MT moves away from the BS or AP. Low RSS from the currently serving networks notifies the MT to the importance of searching for another network to handover to. Therefore, RSS plays a vital role in initiating the network selection procedures before the connection is terminated. RSS is currently the main factor used for making horizontal handover decisions. However, its implication in vertical handover is also inevitable. The relative received signal strength (RRSS) is another metric that is less frequently used instead of the RSS for making handover decisions.

Network load: several applications, mainly file transfer protocol (FTP) applications (*i.e.* file transfer), and data hungry services, perform better when the bandwidth supplied by the network is higher. Therefore, networks with low load, *i.e.* having more free bandwidth, are usually preferred.

Monetary cost: users usually prefer to connect to the network that demands lower monetary cost. The monetary cost could differ between operators of the same access technology, or even between different types of access technologies. For example, Wi-Fi APs based on predefined monthly cost, or free, which could be beneficial for the user when compared to other access technology that could use per MB data costs.

Power consumption: This factor is directly related to the type of the access technology, and the distance and interference between the user and the AP or BS. Usually, terminals connected to Wi-Fi APs consume less power than those being connected to LTE systems. Moreover, as the distance and the interference between the BS and the MT increases, the power required to maintain the connection also increases.

User preferences: It is also considered as one of the main criteria used for making network selection decisions. User preferences reflect the importance of the criteria used to make decisions. Those preferences could be defined based on the application requirement, MT status, or the user's specific preferences. For example, it is the user's preference that indicates what is more important, the quality of the current connection or the monetary cost.

Throughput: It is directly related to the SINR and the network load. It refers to the data rate that could e provided to the MTs.

1.7.2 Single-parameter-based network selection

Extensive research has explored the network selection issue in HWNs. Some studies focus only on one parameter to take HO decisions.

1.7.2.1 Received signal strength-based schemes

Several studies consider the RSS as the only factor to determine network selection decisions. Due to the availability of the hardware equipment required for RSS calculations, a large number of studies has participated in this domain of research within the past few years [MZ04], [CC08], [CCHL09], [PKH+00], [YMS08], [YMP05],[ELS08]. Basically, the RSS of the serving network is compared to that of the other available networks, and the MT connects to the network with highest RSS. A generalized scheme for RSS-based network selection decision starts by scanning and checking the availability of wireless networks. Then, the RSS of each network is calculated and compared to a predefined RSS threshold. The handover is initiated if the collected measurements give a satisfactory result for RSS. Otherwise, the MT switches back to the network discovery state. RSS-based network selection schemes also discuss further aspects.

In [PKH+00] and [YMP05], authors proposed an architecture that depends on location and cross-layer information to perform network selection. They compared the performance of two handover algorithms while considering the terminal velocity and its effect on the user experience. Their work is based on a dynamic dwell timer to estimate the time users will maintain connection to networks, and the quality of these connections. Their proposed work is compared to power-based algorithms. The handover is performed when the RSS level of the serving network is always below the threshold and less than the RSS of other available networks. Their proposed solution is shown to be less sensitive to the increase of the handover delay. Moreover, their solution is shown to affect the performance of real time applications, which is considered as a major drawback.

In [ELS08], authors presented a new design using dynamic dwell timer for seamless handover mechanism for MT with high mobility requirements. This scheme is proposed for Wi-Fi networks, and aims at reducing the handover delay. Their proposed scheme consists of also cellular networks that operate on the same frequencies. In order to reduce the overall handover latency, The proposed solution considers all of the basic handover phases. This is done by using a centralized radio control unit (RCU) that estimates the dwell timer based on the coverage area of the networks, the MT speed, and additional handover-related information.

In [MA06], authors proposed a new network selection algorithm that compares the current RSS level to a dynamic RSS threshold value to determine the best available access technology among 3G and Wi-Fi networks. The dynamic RSS threshold plays a vital role in decreasing the number of unnecessary handovers as well as the number of failed handovers. Authors noticed that using a fixed RSS threshold leads to increase the handover failure probability due to the high mobility of the MT or handover signaling delay. Authors took the advantage of the fact that the availability of 3G networks is ubiquitous, so the handover failure probability from Wi-Fi to 3G network is zero. This is why their proposed solution tends to encourage the handover towards Wi-Fi networks whenever Wi-Fi coverage is available. However, this is not always the case, the handover to a Wi-Fi network could be avoided in case this network is congested, or in case the traveling time of the MT in a Wi-Fi coverage area is less than the handover latency.

Similarly, in [ZLS06], the handover is performed when the following two conditions are met: 1) current RSS is less than, or equal to, the new RSS, and 2) the estimated session lifetime is less than, or equal to, the handover latency. The MT keeps measuring the average RSS using the "moving average" method as proposed in equations 3,4 and 5 of [ZLS06].

1.7.2.2 Signal-to-interference noise ratio-based schemes

Another important parameter usually considered for making network selection decisions is the SINR. It is beneficial to highlight here that the RSS is more inclined towards providing connectivity to the MTs, while SINR reflects, or helps in estimating, the throughput that could be achieved in a network. In [AK10], authors used SINR to assess the system performance in terms of throughput by using the equation that relates the maximum achievable data rate and the SINR in Wi-Fi and WCDMA networks. Their proposed algorithm estimates the minimum required SINR in Wi-Fi networks to achieve the same throughput that could be achieved in WCDMA networks. If the SINR of the available Wi-Fi network is higher than the estimated minimum SINR, handover is triggered. Therefore, the proposed scheme provides a network selection decision that is based on the maximum achievable downlink throughput. It is shown that the SINR-based scheme enhances the system throughput when compared to the traditional RSS-based schemes.

1.7.2.3 Quality of service-based schemes

In [CDM04], authors proposed a network selection solution that is based on the user preferences to satisfy users based on their needs in terms of QoS and monetary cost. To select the best network among GPRS and Wi-Fi, Two network selection strategies are presented: (1) the MT will never leave its connection with GPRS network until the connection is forcely terminated, and (2) MTs are always handed over to Wi-Fi APs whenever those APs are available. The first strategy targets those users who are willing to pay money to maintain continuous connection. The second strategy will save the users from paying extra money, (assuming that Wi-Fi is cheaper), but the QoS is threatened due to the high congestion in Wi-Fi networks.

Authors in [LSK+09] propose a network selection solution that is based on the available bandwidth at each network. In the proposed solution, the handover from Wi-Fi to cellular networks is triggered when there is no other available data network. Wi-Fi networks are preferred due to the fact that cellular networks are incapable of handling heavy loads. In case multiple Wi-Fi APs exist, the AP with highest available bandwidth is preferred. This solution plays a vital role in balancing the load among several networks. Also in [EKR15], the HO algorithm selects the network with the highest available bandwidth.

1.7.2.4 Power consumption-based schemes

Many studies focus on the power consumption at the MT level. In [FZZ12], [XPMV14], and [KYII10], each MT decreases its own power consumption and maximizes its battery lifetime through associating to the network that requests the lowest power consumption among available networks. For example, in these studies, the network selection algorithm tends to select Wi-Fi networks because MTs require less power consumption to maintain the connection than the cellular networks.

1.7.3 Multiple parameters-based network selection

In this section, we will discuss studies that consider multiple parameters to rank available networks and select the best one. Usually, each parameter is associated to a weight that indicates its importance among other parameters. The weight of each parameter could be set based on the user preferences.

The first policy-enabled network selection algorithm was proposed by Wang et al. in 1999 [WKG99] where the best network is selected based on multiple parameters combined within a single cost function. Since then, policy-enabled network selection algorithms have been widely used for selecting networks based on multiple criteria. After taking user preferences and applications needs into consideration, network parameters are summed up in order to select the best network. In the following, we present several studies that combines multiple parameters to select the best networks.

In [MZ04], [ZM04], [ZM06], Zhu and McNair proposed multiple cost-function-based network selection algorithms, where the cost function is used to calculate the cost of all available networks. The used cost function considers the services running on all MTs, and the cost for each candidate network is calculated based on the preferences of each service. The total cost of a candidate network is calculated using the sum of the cost of available QoS parameters like bandwidth, battery consumption and network delay. The network selection algorithm will select the network that supplies the best service for the lowest cost. The proposed network selection algorithm is processed at the MT. The main advantage of this scheme is that the MT could select the best network after aquiring the required information without synchronizing with another entity. Thus, the handover for each MT can be initiated independently. However, authors did not specify in these studies the mechanism that is used to assign weights for each parameter. Moreover, knowing that each attribute has its own measuring unit, authors did not discuss also the methodology used to normalize these parameters.

In [HNH05], [HNH06], and [NHH06], the cost function is also used to select the best network in a heterogeneous wireless system. Here, the cost function is based on multiple parameters related to the monetary cost, power consumption at MTs, security, and velocity. However, those parameters have different measuring units. Therefore, the authors specified the mechanism that normalizes these parameters in order to sum them in the cost function. Each parameter is associated to a weight that reflects its importance among other parameters.

Moreover, a policy-enabled network selection solution has been proposed in [SZ06] and [WLSS08] to select the best network based only on the RSS and the available bandwidth parameters.

In [AMC16], authors proposed a network selection strategy based on the monetary cost, power consumption, and QoS parameters. The proposed optimization problem aims mainly at minimizing the power consumption of the MT while considering the QoS requirement of the MT and the monetary cost. The QoS is seen in terms of required data rate. The authors here assume that the MT could be connected to multiple networks at the same time to share the load and receive the requested data rate.

In fact, authors in [TL11b], [TL11a] proposed dynamic context aware network selection solutions that consider both user and service requirements. The context-awareness is based on the information collected about the requirements of the applications running on each MT, the characteristics of the networks, the geographical location of both users and networks.

Weighted cost function is used in [TOM13] to study the trade-off between energy, monetary cost, and QoS of multimedia services in heterogeneous wireless environment. The proposed weighted cost function is used to rank candidate networks.

1.7.4 Main drawback of the network selection scheme

The main idea behind the network selection schemes is to select for each MT the best network that selfishly satisfies the MT. Therefore, all previous studies do not consider a system wide resource allocation and user association solution. Instead, they are designed to satisfy the needs of each user individually. In order to optimize the performance of heterogeneous networks, a system-wide resource allocation and user association scheme should be considered. The system-wide scheme first considers the conditions of all MTs and networks in the system, then the user association algorithm selects the network that each MT should be connected to.

1.8 System-wide resource and network assignment

From a system perspective, several studies with different objectives focus on user association and resource allocation in HWNs.

In [ZDRB13] an optimization problem is formulated to increase the number of connected users, while maintaining a minimum utility for each one. The algorithm was evaluated based on user throughput and number of handovers without studying other specific user-related attributes.

In [GBGF+11], in order to enhance the energy efficiency in heterogeneous networks, authors discussed the problem of BS location and optimal power allocation. Mainly, the formulated problem aims at optimizing the number of BSs and their location to reduce the consumed energy. The problem of reducing the energy consumption has been also studied in [PCS12]. The formulated problem aims at minimizing the number of active networks while considering the data rate requirements. The problem is formulated as an integer programming problem to minimize the overall energy consumption. However, the energy consumption at user level, or other user-related parameters, is not considered in both studies.

In [HC17], authors investigated the profitability of the network operators based on the spectrum-energy efficiency. The authors use techniques such as user association, cell size zooming, and ON/OFF scheme for different BS types in heterogeneous networks to optimize the spectrum-energy efficiency. However, the user-centric benefits is not considered. Instead, the focus is only on the network benefits.

In [LSK+09] authors use joint optimization to maintain load balancing among heterogeneous networks and optimize the MT battery lifetime. However, power consumption rate in each network is assumed to be exponentially distributed and the system model is not realistic but an abstract system model is used. In [rCC14], the objective function of the formulated problem aims at balancing the load as equally as possible among multiple networks. However, user-related attributes and benefits have been not discussed.

In [TRRL15], authors proposed a user association, resource allocation, scheduling, and power allocation algorithm to increase the overall network utility. The formulated problem in [YH15] aims at increasing the total throughput of the system. The idea is to lower the number of active pico-base stations in order to reduce the inter-cell interference with bigger macro-base stations. However, the studies [TRRL15] and [YH15] do not consider the user preferences or the user-centric profit.

In [JSS14], the formulated problem maximizes the total user utility. The user utility is based on the amount of data rate received by the user. However, the amount of data rate requested by each MT is not considered. The optimization problem proposed in [LWH14] considers the data rate requirements of users. The formulated problem aims at minimizing the total resources required to supply the given user traffic demands. However, the formulated problem is based on a system model that do not follow the specific-access-technology resource allocation constraints, instead, time slots or frequency slices are assumed to be infinitely divisible. In practice, taking LTE as an example, resources are discrete, and a single resource unit could not be shared between multiple MTs at the same time.

In [CL16], authors proposed a solution to maximize the user-centric satisfaction while considering the data rate requirements of users. However, their proposed optimization function is also based on an abstract system model that do not follow the constraints of the RATs. In [EWI+16], authors proposed a data rate allocation and MT association solution to maximize the total user utility which depends on the amount of data rate received by each MT. However, the amount of data rate requested by MTs is also not considered.

In [BHW13] and [MAAV14] authors proposed user association and resource

allocation algorithms to minimize the total amount of time required to satisfy user traffic demands. However, both studies do not consider the user preferences. Furthermore, the system model and the formulated problems in [BHW13], and [MAAV14] also do not follow the specific-access-network constraints. Instead, an abstract system model is used.

In [JSS14], the proposed solution increases the overall user-centric utility that is based on the per-user throughput. Increasing the per-user throughput has been also used in [YRC+12] and [ZHY15] as a mean to increase the overall system throughput. However, the studies [JSS14], [YRC+12], and [ZHY15] do not consider the amount of data rate requested by each user. In fact, increasing the per-user throughput does not always contribute to better satisfaction for users. For example, voice over Internet protocol (VoIP) applications usually request a fixed amount of data rate ; increasing the data rate above this amount does not necessarily enhance the performance of the application.

The study [DWY+15] explored the problem of optimizing the user-centric satisfaction while considering user-demand diversity. However, their proposed optimization function and system model do not follow the specific-access-technology constraints. Instead, a generalized problem formulation is adopted.

1.9 Management frameworks

Due to the large number of studies with different objectives that have been introduced to heterogeneous networks, it is essential to differentiate these studies based on their management time. Therefore, in this section, a management framework that classifies these studies based on the time frame is introduced.

Mainly, the studies contributed in heterogeneous networks could be classified into four main time-based categories.

Ultra-slow decisions: In this category, the studies focus on long-term management decisions. Network planning and deployment related research fields in heterogeneous networks are the main field that belongs to this category due to the manual operation required to perform these tasks, and the time required to install BSs and APs. Figure 1.10 summarizes the decision objectives that are classified in each category.

Slow decisions: In this category, the decision is based on the underlying fast decision category to estimate the number of networks, and which networks, that should be functioning in order to serve the active users. Network activation/deactivation decisions should be made after acquiring information about the load and utilization of networks. These informations are gathered based on the underlying fast decision category.

Fast decisions: This category contains the decisions that could be taken based on information gathered from neighboring networks and users. For example, user association and resource allocation decisions could be taken in a specific scheduling interval (for example one second). Moreover, association decisions could be based also on the estimated time that the user will stay within the coverage area of a network.

Ultra-fast decisions: This category contains the decisions that could be taken independently without coordinating or exchanging information with other net-

works. Usually, these decisions are taken within milliseconds time interval. Procedures that could change every transmission time interval (TTI) such as power allocation and modulation and coding schemes for each cell belongs to this category. Moreover, network discovery procedures are also classified in this category because these decisions depend on the instantaneous RSS.

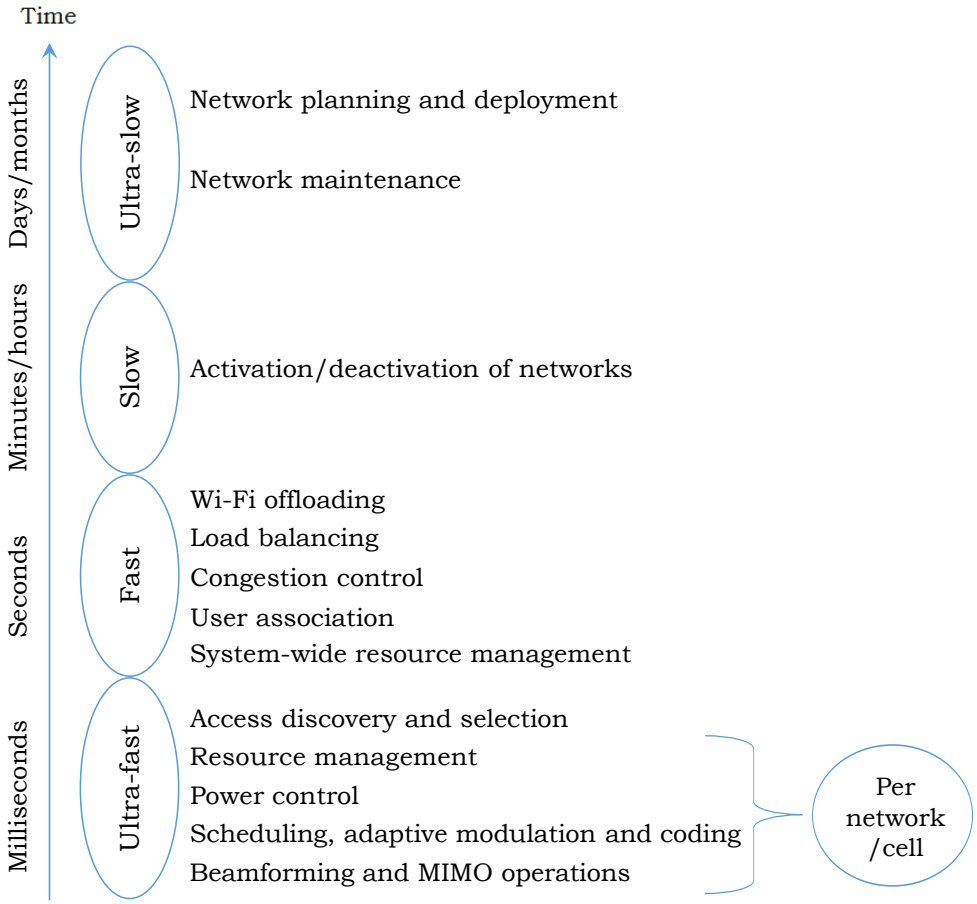


FIGURE 1.10 – Management framework.

1.10 Conclusion

Based on the information provided in this chapter, we can identify the main specifications that should be considered in order to propose new contributions in the HWNs research field.

First, Wi-Fi networks should be seen as an important complementary access technology to the cellular networks due to its wide-spread and the low cost installation and management procedures. Therefore, the proposed work should consider the coordination between Wi-Fi networks and cellular networks to face the exponential growth of the requested data traffic.

The second important aspect is to follow the goals of the ABC scheme ; user satisfaction, demands, and preferences should be considered in order to maintain the best connection at all times. While designing the solution goals, important parameters should be considered based on the proposed work in the literature. For

example, the battery power consumption of MTs should be considered because it is introduced as an essential metric for user satisfaction. Also, the RSS parameter should be considered because it plays a vital role in maintaining the connection to any network. The amount of requested data rate of each MT should be taken into consideration while allocating resources. Therefore, the SINR parameter should be also considered because it is directly related to the amount of data rate received by each MT.

Another important aspect that should be considered is to design the proposed work in a way that any additional user-centric parameter, such as monetary cost, could be added easily to the objective function. Moreover, the proposed work should be able to be configured in a way to address network-utility based objectives, such as power efficiency maximization.

The majority of work that discusses and targets the ABC scheme is based on the network selection paradigm, where we have stated before that the work aims at enhancing the user satisfaction selfishly and without considering the effect of the algorithms on the system as a whole. Therefore, it is essential to introduce and pave the way for an optimized ABC scheme that aims at enhancing the overall user experience and consider the network and MTs conditions. Such solution requires gathering information from the whole system, and coordination between all the networks, therefore it falls in the fast decision category.

In the remaining of this dissertation, those points are taken into consideration.

Chapter 2

A New Resource and Network Assignment Problem Formulation and System Architecture

In this chapter, we discuss the media-independent service (MIS) framework architecture for software-defined radio access networking (SDRAN) and the communication mechanism between different entities. Based on this framework, we will formulate a user association and resource allocation optimization problem. The algorithm runs on the SDN controller that maintains global view on the system. Our decision algorithm is based on two attributes: the power consumption at the MT and the RRSS. We use a reliable power consumption model to estimate the consumed power at MTs in different wireless networks.

2.1 Introduction

To maximize the satisfaction of MTs in HWNs, transferring the connection between different networks is definitely desired, however it develops several issues. The first issue is to decide to which network the connection should be transferred, taking into consideration the limited capacity of networks and the requirements of all MTs. Collecting information needed for the HO decision leads us to the second issue which is establishing a reliable infrastructure that provides a centralized global supervision on different network entities while maintaining cooperation and information exchange between them. The last issue is to provide a common platform that supports mobility management and resource allocation in HWNs. IEEE 802.21 media-independent HO services and software defined networking (SDN) can complement each other to provide a solution for these issues.

The SDN scheme is mainly featured by the separation of data and control planes. Such separation allows the network control to become programmable by a software that might be running on a centralized network entity while the infrastructure is completely abstracted from the software [FRZ14]. This approach enables the management procedures to be implemented in a centralized mode through a SDN controller which maintains global view on network entities. The SDN control-

ler is able to control data packets and manage how such packets are treated by network devices. Usually the SDN functionality is demonstrated using OpenFlow [MAB+08] which is an open standard that enables a separate network entity, the OpenFlow controller, to access the forwarding plane of switches, usually OpenFlow switches, over the network. The forwarding operation still resides within the switch, however it is controlled by software. The OpenFlow communication protocol is already available in several network products and it can be considered as an enabler for SDN.

IEEE has contributed a new standard in 2008 which is the 802.21 media-independent HO (MIH) [47609] to enable seamless HO between networks of same or different types. The 802.21 MIH standard defines several MISs that provide helpful information in the HO decision and facilitate seamless HO. However, the MIS has been recently splitted from the IEEE 802.21-2008 standard into a new stand alone standard called 802.21.1 that mainly deals with MIS. It is necessary to illustrate that at the time of writing this thesis, the IEEE 802.21.1 is still under revision and not published yet. However drafts about some sections that this standard will discuss are already published [Jin15]. Information about the IEEE 802.21.1 project can be found in [IEE].

One of the interesting topics that the IEEE 802.21.1 MIS standard will discuss is the media-independent handover services for SDRANs [Jin15]. The SDRAN framework combines the capabilities of IEEE 802.21 MIH with SDN to provide seamless HO, resource allocation and centralized management in HWNs. The framework maintains independent evolution of both technologies, SDN and MIS, through a clear separation between the SDN control plane and MIS control functions. In the SDRAN paradigm, a controller, different from the SDN-controller, can maintain service continuity during the HO. Depending on the underlying SDN infrastructure, the SDRAN framework also provides a communication protocol between different network entities.

2.2 Background

Nowadays, SDN has been introduced to wireless environment. [YSK+10] and [YKS+10] introduce an open SDN-based wireless architecture that can be shared by different providers, while maintaining a backward-compatibility. In [LMR12] Li *et al.* discuss various challenges that might face the deployment of SDN in mobile networks. SDN-based mobility management and load balancing in WLANs is introduced in [SSZM+12]. In [DKB11], Dely *et al.* introduce a new architecture that combines OpenFlow with wireless mesh networks enabling mesh clients to seamlessly roam between different mesh APs.

Concerning the application of SDN in HWNs, in [GCA+13] the authors propose an architecture that combines the 802.21 MIH with SDN. However, the SDN controller and MIH functionality are not separated as the SDRAN MIS framework suggests. In [KLD], the authors propose a centralized, semi-centralized, and hierarchical approaches to manage handover in SDN-based wireless networks. An application of SDN controller as load balancer among different HWNs was introduced in [TL14] where the load balancing algorithm takes into consideration the requirements of users and the capacity of networks. In [DWA14] the authors

propose a centralized algorithm for Wi-Fi data offloading in HWNs depending on a centralized SDN controller that has global view on the networks and enables centralized coordination and control. However, in [TL14] and [DWA14] the interworking framework is not discussed.

When an active connection is handed off from a network to another one that uses the same wireless access technology, the HO can usually be executed within that access technology itself. For example, a VoIP call over Wi-Fi can be handed over between APs using Wi-Fi standards such as 802.11f and 802.11r. However, if it is required to perform HO between two networks of different access technology, *e.g.* from Wi-Fi AP to LTE BS, then an external protocol is required to manage the HO. In 2008, IEEE has published a new standard which is the 802.21 MIH [47609] to enable seamless HO between networks of same or different types. MIH can communicate with several network protocols to facilitate the HO procedures. Those protocols include the session initiation protocol (SIP) for signaling and mobile IP protocol for mobility management. The standard is intended for HWNs including both 802 and non 802 access technologies.

2.3 Software-defined radio access networking

In this section we discuss the basic entities of the 802.21.1 SDRAN standard, and we propose a novel handover scenario based on the discussed framework.

2.3.1 Media-independent services (MIS)

The IEEE 802.21 standard has been a reliable platform that facilitates mobility management between heterogeneous networks. The entity that supports MIS framework is called MIS entity, *e.g.* networks, MTs, servers, *etc.* The MIS framework defines the following three services that optimize the HO process:

- Media-independent event service (MIES): provides different dynamic physical and link layer events filtering and reporting. For example, the MIES can be used to register for events related to signal quality and specify a certain threshold such that when the signal quality degrades below that threshold the corresponding MIS entity is notified.
- Media-independent command service (MICS): enables some MIS entities to manage and control link layer resources.
- Media-independent information service (MIIS): allows the exchange of static and dynamic information between MIS entities. Informations can be related to cost, QoS, data rate, the geographic location, link layer address, *etc.*

2.3.2 MIS framework (MISF)

To enable HO, all network entities that support MIS standard has a set of MISF functions within their mobility-management protocol stack. The network entity that supports MIS could be also called MISF entity. MISF can be also seen as layer in the protocol stack of the entity. To facilitate handover, the MIS provides a framework and mechanisms by which the MISF entity can discover and

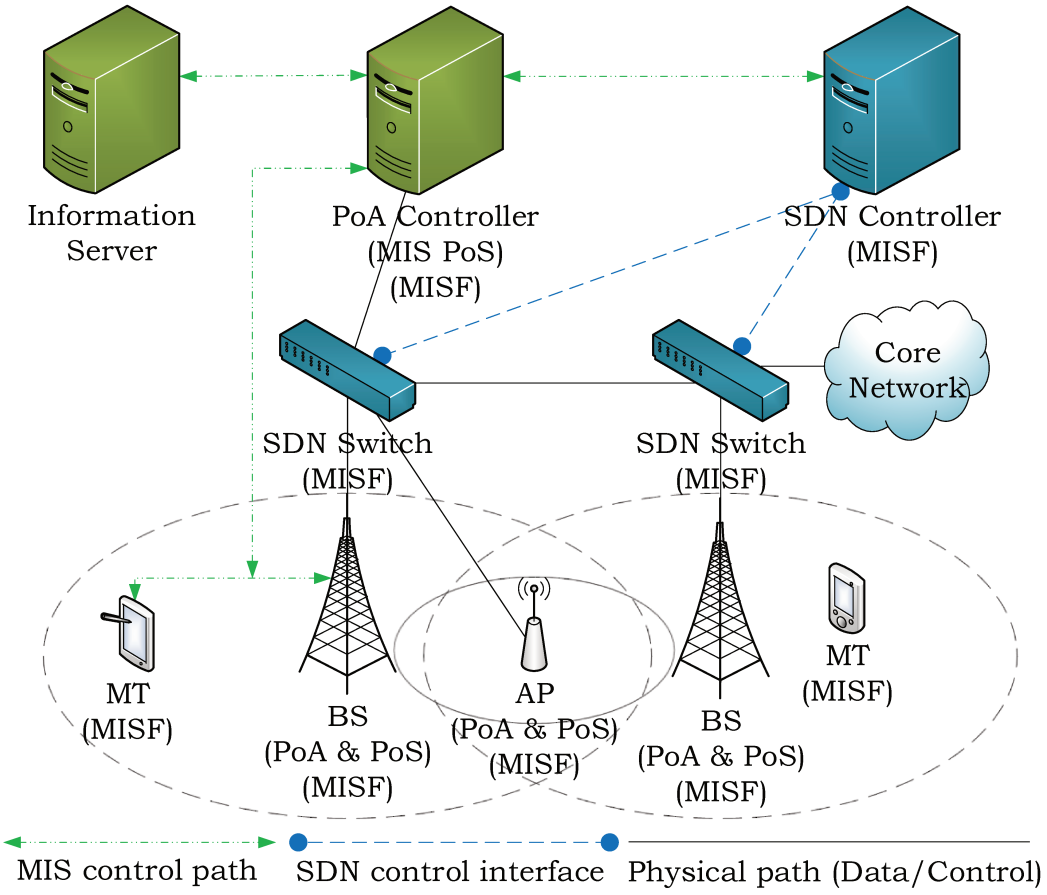


FIGURE 2.1 – The MIS framework architecture for SDRAN.

obtain information related to other entities in the system. The MIS framework for SDRAN is shown in Figure 2.1 and it includes the following entities:

- MT: user mobile device equipped with several radio interfaces such as LTE, Wi-Fi, UMTS, *etc.*
- MIS point of service (MIS PoS): a network entity that can exchange MIS messages with MTs through MISF.
- MIS point of attachment (PoA): the cellular BSs or WLAN AP. It can also act as PoS and directly exchange MIS messages with MTs attached to it.
- PoA controller: it is an MIS PoS that can control and manage HOs and allocate the resources of PoAs. Several PoA controllers might exist, each one manage the resources of several PoAs.
- Information server: a server that collects real time information about MTs and PoAs.
- SDN switch: SDN-based switch that can also operate as a gateway between PoAs and core network.
- SDN controller: a controller that can manage resources of SDN switches and has global view on the system. It can control the forwarding of PoAs' traffic and communicate with all PoA controllers through the SDN switches.

All above entities are equipped with MISF except the SDN controller. MISF entities can communicate with each other using the MIS protocol. The SDN switch

is equipped with MISF interface to allow exchange of MIS messages between remote MISF entities that are not directly connected.

In heterogeneous networks, PoAs have different wireless access technologies, *e.g.* Wi-fi, LTE, 3G, *etc.*, where each access technology has its own mechanism to allocate resources. In SDRAN, each PoA is equipped with MISF to support HO management and resource allocation. The PoA controller can manage PoAs' resources and coordinate HO between PoAs using MIS functions that are sent through the MIS control path. The information server at the core network receives information about PoAs and MTs through their corresponding PoA controller. The PoA controller can use MIS functions to get information about PoAs resources, these information can be helpful for the HO decision. The PoA controller is essential to preserve session continuity during the HO.

The SDN controller is not equipped with MISF in SDRAN framework to separate between the MIS and SDN control planes allowing both technologies to develop independently. We will depend on the SDN controller to make HO decisions since it has global view on the system and can communicate with all PoA controllers through SDN switches. However, it is necessary to illustrate that since the SDN controller cannot communicate with entities using MIS protocol. So, in order to collect information needed for the HO decision and force user association and resource allocation, it will communicate with all PoA controllers. The communication between the SDN controller and PoA controllers can be established through the SDN switches or through a direct interface, called East/West interface, that needs to be standardized in the future.

2.3.3 MISF service access points

Service Access Points (SAPs) is a group of service primitives used by MISF to exchange messages with other entities in the framework and with other layers in the mobility management protocol stack of the entity. Figure 2.2 shows the mobility management protocol stack of MT and PoA entities in SDRAN framework. Moreover, the MISF interface of PoA controller, information server and the SDN switch is shown in the figure, in addition to the OpenFlow interface in both SDN controller and switch.

Mainly we have two types of SAPs: the media-dependent SAP and the media-independent SAPs. MIS_SAP is a media-independent SAP that acts as an interface between the MISF layer and the upper layer in the mobility management protocol stack. The upper layer in the protocol stack of MT and POA, the MIS_User, need to subscribe to MISF in order to receive MISF and link-layer events. This subscription is done using MIES messages sent from the MIS_User to its local MISF using MIS_SAP. The MIS_User can also specify thresholds for these events in the same way. For example, if the MIS_User is interested in knowing when the received signal strength is below a certain threshold, the MIS_User should subscribe to the MISF for this event and specify a certain threshold using the discussed mechanism. Once the lower layer notifies the MISF about this event, the MISF in turns reports it to the MIS_USER using the same mechanism.

The MISF can use services from lower layers in the protocol stack using media-dependent SAPs. Media-dependent SAPs use media-specific SAPs such as the

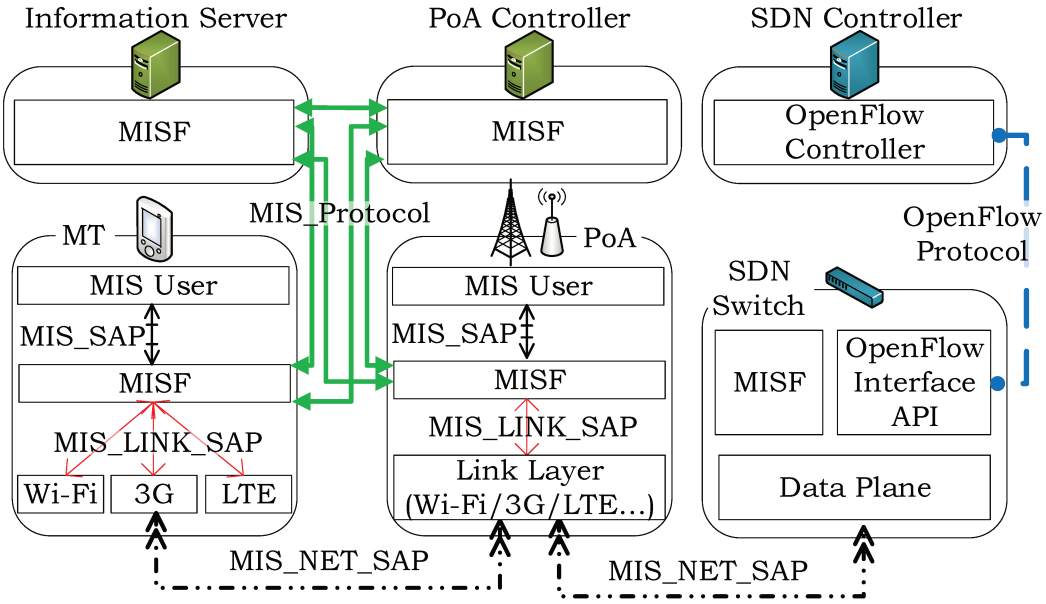


FIGURE 2.2 – The location of MISF in the protocol stack of different SDRAN entities, and its interaction with other MISFs and layers through the MIS SAPs.

MAC_SAPs, PHY_SAPs, and logical link control LLC_SAPs that are specific for each access technology. Two media-dependent SAPs are specified by the SDRAN framework: the MIS_LINK_SAP and the MIS_NET_SAP. Mainly, the MIS_LINK_SAP is an interface between the local MISF and lower link layer, whereas the MIS_NET_SAP enables the exchange of MIS information and messages with remote MISFs over the data plane on the local node. The communication between remote MISFs is done using MIS_Protocol, while the communication between SDN entities uses the OpenFlow protocol. The SDRAN framework enable communication between MISF entities using both layer 2 (link layer) and layer 3 (network layer) communication interfaces. So the MISF entities that does not support MIS_NET_SAP can exchange MIS protocol messages through layer 3 interface.

2.4 Novel handover scenario based on SDRAN

As we have mentioned before, the SDN controller processes the HO and resource allocation algorithm because it's the only entity in the framework that has a global view on the system that might contain several POA controllers. The HO algorithm, which will be discussed in Section 2.7.2, takes into consideration the limited capacity of networks, the different data rates requested by MTs, and the preference of each user. However, in this section we will discuss the mechanism of querying the information needed for our HO algorithm, the mechanism of allocating PoAs' resources, and what might trigger the HO algorithm. Note that this scenario is not discussed in [Jin15], however it is based on the SDRAN framework and HO stages defined in [Jin15].

First, the MIS_User on the MT can be notified when the RSS is below a certain threshold or when a new PoA is available through subscribing for these

events to its local MISF as we have discussed before. Whenever a new PoA is available or the RSS is below the specified threshold, the local MISF is notified by the lower layer through MIS_LINK_SAP interface. Then, the MISF notify the SDN controller about this event. Since the MT do not have an active connection with the SDN controller, the MISF at MT contact the MISF of the SDN switch through the serving PoA's MISF using MIS_NET_SAP as shown in Figure 2.2. The SDN switch then notifies the SDN controller through the OpenFlow protocol. To query the information needed for the HO algorithm, the SDN controller requests these information from a PoA controller that can query MTs' information from the information server through MIIS message sent using MIS_Protocol. The communication between the SDN controller and PoA controllers can be done through the SDN switch or through the East/West interface that needs to be standardized in the future. Once the SDN controller receives the information, the HO algorithm is initiated. After the HO decision is made, the SDN controller notifies PoA controllers about the decision, which in turns command PoAs to allocate resources for MTs (using MICS command) and the MTs are then notified. The communication between PoA controllers, MTs, and PoAs is done through MIS_Protocol. The PoA controller is responsible for maintaining the service continuity during the HO execution phase.

In our algorithm, the required information needed for the HO decision are:

- for networks: capacity, location, coverage range, and access technology.
- for MTs: location, requested data rate, and user preference.

It is necessary to illustrate that the location of MTs can be based on feedback from MTs or can be estimated by networks through modern mechanisms that are based on the angle of arrival, time of arrival, time difference of arrival, and RSS [GG05].

2.5 System model

We focus on the downlink resource allocation in a heterogeneous wireless system. The system consists of several mobile BSs and Wi-Fi APs with overlapping coverage areas. The network sets corresponding to mobile BSs and Wi-Fi APs are denoted by $\mathcal{N}_{\mathcal{BS}} = \{1, 2, ..., G\}$ and $\mathcal{N}_{\mathcal{AP}} = \{G + 1, ..., N\}$ respectively. The total network set is denoted by $\mathcal{N} = \mathcal{N}_{\mathcal{BS}} \cup \mathcal{N}_{\mathcal{AP}} = \{1, 2, ..., N\}$, where $\mathcal{N}_{\mathcal{BS}} \cap \mathcal{N}_{\mathcal{AP}} = \varnothing$. The available mobile BSs and Wi-Fi APs selected by a given MT are those for which this MT is located in their coverage area. Therefore, it is assumed that each network n has a circular coverage area with radius R_n. Hence, we define a set Λ_m which contains a list of all networks reachable by MT m such that $\{n \in \Lambda_m \mid d_{mn} \leq R_n\}$ where d_{mn} denotes the distance between AP or BS n and MT m. The set of MTs located within the system is denoted by $\mathcal{M} = \{1, 2, ..., M\}$. The data rate (kbps) requested by MT m is denoted by Q_m.

2.5.1 Resource allocation in LTE

The downlink of an LTE BS is considered, where the total bandwidth is divided into C_n sub-channels. Each sub-channel is made up of twelve sub-carriers that are grouped into a resource block (RB) whose total bandwidth is B_n^{RB} kHz. Following similar approach as in [BB15], the positive channel power gain between

MT m and BS n is denoted by H_{mn}, where $H_{mn} = G_{mn}F_{mn}$ encompasses the effects of path loss, log-normal shadowing, and antenna gains as large scale fading component (denoted by G_{mn}), and multi-path Rayleigh fading as small scale fading component (denoted by F_{mn}). In [R+96], F_{mn} is modeled as an independent exponentially distributed random variable with a unit variance because the envelope of the signal in Rayleigh fading environment is assumed to follow a Rayleigh distribution. Therefore, it can be assumed that F_{mn} fluctuates fast enough so that a mobile user can average it out in its channel measurements [BB15] [HYM+17]. Thus, the long-term SINR that is measured by MT m from BS n on a RB is [BB15]:

$$\overline{SINR}_{mn} = \frac{P_n^{RB} G_{mn}}{\sum_{i \in \mathcal{N}_{\mathcal{BS}} \setminus n} P_i^{RB} G_{mi} + B_n^{RB} N_0} \tag{2.1}$$

where P_n^{RB} and P_i^{RB} are the transmission power on a RB by BSs n and i respectively, N_0 denotes the thermal noise spectral power, and $\mathcal{N}_{\mathcal{BS}} \setminus n$ is the set of all BSs except BS n. It is assumed that the allocated power for each sub-channel is predefined. For example, equal power allocation (EPA) could be considered [BB15]. Therefore, MT m can measure G_{mn} for all the BSs n. Hence, the long-term spectral efficiency in kbps/Hz between MT m and BS n on a RB is [BB15]:

$$\gamma_{mn} = \log_2(1 + \overline{SINR}_{mn}) \tag{2.2}$$

where the achievable data rate in kbps on a RB could be calculated by multiplying γ_{mn} by the bandwidth of a RB (B_n^{RB}) and the time duration then divided by the scheduling interval. While allocating resources, it is more convenient to consider the long-term achievable data rate instead of the instantaneous one, otherwise, the resource allocation algorithm might run upon any degradation in the instantaneous SINR.

In this thesis, the resource allocation algorithm considers the data rate requested by each MT. The data rate requested by MTs could vary dramatically. For example, a MT running a file download application requests data rate much larger than another MT running a VoIP call. Therefore, it is not convenient to allocate a whole RB to MTs requesting low data rate. In [KK16] the transmission time in LTE is divided into several fractions, where a MT could allocate a specific combination of a time fraction and a RB. Thus, it is assumed that the transmission time, or scheduling interval, is divided into T_n discrete time fractions, where each RB spanning the interval of one time fraction is denoted by scheduling block (SB). Hence, the total number of SBs at a BS n is $U_n = C_n T_n$, and u_{mn} denotes the number of SBs allocated to MT m by BS n. Thus, the long-term achievable data rate (kbps) of MT m in BS n is:

$$r_{mn} = \frac{u_{mn} B_n^{RB} \gamma_{mn}}{T_n} \tag{2.3}$$

2.5.2 Resource allocation in Wi-Fi

In Wi-Fi APs, as in [XGP+12], an enhanced version of distributed coordination function (DCF) [CYCK05] with a reservation-based MAC protocol is considered. MTs can completely avoid collisions by acknowledging their back-off timer value

within the MAC header. Thus, it can be simply seen as if MTs access the AP in a time division multiple access (TDMA) manner. The resource allocation in Wi-Fi APs is also seen as TDMA in [KCL13]. Each MT can occupy the whole bandwidth of AP n, denoted by B_n, in its allocated time slot. The total number of time slots during scheduling time is T_n, and t_{mn} denotes the number of time slots allocated to MT m. Therefore, the achieved data rate (kbps) of MT m in AP n is:

$$r_{mn} = \frac{r_{mn}^{tot} t_{mn}}{T_n} \tag{2.4}$$

where $r_{mn}^{tot} = B_n \gamma_{mn}$ is the total achievable data rate (kbps) in AP n, γ_{mn} is the spectral efficiency (kbps/Hz) between MT m and AP n. The channel model used in [KCL13] for Wi-Fi APs is adopted. The spectral efficiency formula is:

$$\gamma_{mn} = \log_2(1 + \frac{P_n^{AP} g_{mn}(d_{mn})}{\sigma^2}) \tag{2.5}$$

where P_n^{AP} is the transmission power of AP n, $g_{mn}(d_{mn})$ is the channel gain of MT m at distance d_{mn} from the n^{th} AP encompassing the effects of path loss and antenna gains, and σ^2 is the noise power. It is assumed that each AP works on non-overlapping channels so that no interference exists among APs. In addition, since WLANs operate in an unlicensed band, APs do not interfere with the BSs. Full power transmission in each time slot is assumed. Note that the fast fading component follows a Rayleigh distribution and it is averaged out due to the same reason discussed in Section 2.5.1.

2.5.3 User-centric attributes

In this section, we talk about two user-centric attributes that are chosen to calculate the context-aware profit contributed by associating MTs to networks.

2.5.3.1 Signal quality

The signal quality is usually considered as an important attribute for making HO decisions in HWNs. However, it is difficult to compare the quality of the signal among different wireless access technologies because they have various maximum transmission power and receiver power threshold. To overcome this issue, Shen *et al.* have proposed a signal quality formula in [SZ08] that is applicable in different types of wireless technologies. The proposed formula is:

$$s_{mn} = \frac{P_{mn} - P_n^{th}}{P_n^{max} - P_n^{th}} \tag{2.6}$$

where P_n^{th} the receiver power threshold in network n, P_n^{max} the maximum transmitted signal power in network n, and P_{mn} the actual signal power received by MT m from network n. Shen *et al.* have managed to reduce their proposed formula to:

$$s_{mn} = 1 - \frac{\log(d_{mn})}{\log(R_n)} \tag{2.7}$$

Note that network n is unreachable by MT m if $d_{mn} > R_n$.

2.5.3.2 Instantaneous power consumption

MT power consumption is usually considered as an important user-centric attribute. Therefore, it is considered within the profit function. To estimate the instantaneous power consumed by MT m while receiving data from network n, an empirically derived power model proposed by [HQG+12] is used:

$$pc_{mn} = \alpha_n r_{mn} + \beta_n \tag{2.8}$$

where pc_{mn} is the power consumed by MT m while receiving data from network n, r_{mn} the downlink data rate in Kbps, $\alpha_n(mW/kbps)$ and $\beta_n(mW)$ are constants related to the wireless access technology of network n. The highly-cited power model presented in this section has been widely used in research fields and it is shown to achieve a very low estimation error [HQG+12].

2.6 Local (MT-based) network selection solution

In this section, a new MT-based network selection solution that considers network's capacity and resource allocation constraints, user preferences, and user's requested data rate is proposed.

2.6.1 Novel local profit function derivation

In the proposed solution, users prefer to be served by a network with low instantaneous power consumption and high received signal quality. Consequently, a user-centric weighted profit function is defined to combine these two attributes. The weight of each attribute reflects its importance among other attributes in the profit function according to the user preference. The weight of signal quality and instantaneous power consumption for MT m is symbolized by w_m^s and w_m^{pc} respectively. Both weights are subject to the following constraint:

$$w_m^s + w_m^{pc} = 1 \tag{2.9}$$

Note that, both attributes (s_{mn} and pc_{mn}) have different measurement units. Thus, in order to combine them in the weighted profit function, a normalization step is required. While normalizing, it is essential to differentiate between upward and downward attributes; attributes of which their higher value is preferable are called upward attributes; conversely, downward attributes are those we aim at decreasing their value. It is obvious that the signal quality is considered as an upward attribute while the instantaneous power consumption as a downward one. Therefore, based on [HNH06], the normalized forms of the signal quality and instantaneous power consumption, denoted by $\overline{s}_{mn}$ and $\overline{pc}_{mn}$ respectively, are:

$$\overline{s}_{mn} = \frac{s_{mn}}{\max\limits_{n \in \mathcal{N}}(s_{mn})} \tag{2.10}$$

$$\overline{pc}_{mn} = \frac{1/pc_{mn}}{\max\limits_{n \in \mathcal{N}}(1/pc_{mn})} \tag{2.11}$$

Note that increasing the value of $\overline{s}_{mn}$ depends on increasing the value of s_{mn} while $\overline{pc}_{mn}$ can be increased by decreasing pc_{mn}. Note that each attribute is normalized to the local maximum, *i.e.* $\max_{n \in \mathcal{N}}$, because the profit function will be deployed in a local optimization problem that aims at enhancing the experience of a single MT. Moreover, the profit of an unreachable network should be zero. Therefore, the profit function should be also multiplied by a unit step function that indicates whether MT m is within the coverage area of network n. The unit-step function is defined as follows:

$$U(R_n - d_{mn}) = \begin{cases} 1, & if\ R_n \geq d_{mn}, \\ 0, & if\ R_n < d_{mn}. \end{cases} \tag{2.12}$$

Hence, the local profit function that we aim to maximize at MT m is:

$$\overline{f}_{mn} = \left(w_m^s \overline{s}_{mn} + w_m^{pc} \overline{pc}_{mn}\right) .U(R_n - d_{mn}) \tag{2.13}$$

The weight of each parameter is set based on the user preference. For example, if the battery status of MT m is very low, then w_m^{pc} could be set to 1, and w_m^s to 0 to select a network with low power consumption property. If the user is equally concerned about mobility and power consumption, both weights are set to 0.5.

Note that the derived profit function could be extended easily to account for other attributes simply by including the normalized value of the new attribute multiplied by its corresponding weight. However, the sum of all weights should be one, *i.e.* $w_m^s + w_m^{pc} + w_m^{new\ attribute} = 1$.

2.6.2 Local (MT-based) network selection problem formulation

The MT-based network selection algorithm should consider multiple aspects. First, the data rate requirement of the MT should be met. Therefore, after selecting the target network, the algorithm should guarantee that the data rate supplied to the MT is at least equal to the requested data rate. Moreover, the algorithm should associate the MT to the network that best fits its needs based on the user preferences (weights). In addition, the resource allocation constraints of each access technology should be considered.

MTs are permitted to be associated with one network at a time. The network selection algorithm aims at finding the best network for MT m amongst available networks. Therefore, to formulate the optimization problem, we define the boolean association variable x_{mn} such that:

$$x_{mn} = \begin{cases} 1, & if\ MT\ m\ is\ associated\ with\ network\ n, \\ 0, & otherwise. \end{cases} \tag{2.14}$$

The optimization problem finds the optimal values of x_{mn} that maximize the total profit by associating MT m to only one network without exceeding its capacity. Therefore, the following local optimization problem (**PL**) is formulated to be pro-

cessed at MT m in order to select the best available network:

$$\textbf{PL: } max \sum_{n \in \mathcal{N}} \overline{f}_{mn} x_{mn} \tag{2.15a}$$

$$s.\ t. \sum_{n \in \mathcal{N}} r_{mn} x_{mn} \geqslant Q_m x_{mn} \tag{2.15b}$$

$$\sum_{n \in \mathcal{N}} x_{mn} \leqslant 1 \tag{2.15c}$$

$$x_{mn} \in \{0, 1\} \qquad \forall n \in \mathcal{N} \tag{2.15d}$$

$$\sum_{m \in \mathcal{M}} t_{mn} x_{mn} \leqslant T_n \qquad \forall\ n \in \mathcal{N}_{\mathcal{AP}} \tag{2.15e}$$

$$\sum_{m \in \mathcal{M}} u_{mn} x_{mn} \leqslant U_n \qquad \forall\ n \in \mathcal{N}_{\mathcal{BS}} \tag{2.15f}$$

$$t_{mn} \in \mathbb{N} \qquad \forall n \in \mathcal{N}_{\mathcal{AP}} \tag{2.15g}$$

$$u_{mn} \in \mathbb{N} \qquad \forall n \in \mathcal{N}_{\mathcal{BS}} \tag{2.15h}$$

Constraint (2.15b) guarantees that MT m receives data rate at least equal to the requested one. However, we are obliged to multiply both sides of the inequality (2.15b) by x_{mn} because upon congestion, some MTs will not be served. For example, when the number of requested resources exceeds the total number of available resources in the system, some MTs will not be served. Constraints (2.15c) and (2.15d) ensure that a MT is connected to only one network, or not connected at all (upon congestion). Constraints (2.15e) and (2.15f) guarantee that the capacity of APs and BSs respectively is not exceeded. Constraints (2.15g) and (2.15h) assure that a single resource unit is not shared by multiple MTs simultaneously. Hence, the formulated problem considers the resource allocation constraints.

2.6.3 MT-based network selection algorithm

To simplify the problem without violating its goals, the resource allocation variables (t_{mn} and u_{mn}) are fixed while taking into consideration constraints (2.15b), (2.15g), and (2.15h). Therefore, based on (2.3) for LTE BSs and (2.4) for Wi-Fi APs, the variables u_{mn} and t_{mn} are fixed and set to the minimal number of resources requested by MT m from network $n \in \mathcal{N}$ such that:

$$u_{mn} = \left\lceil \frac{Q_m T_n}{B_n^{RB} \log_2(1 + \gamma_{mn})} \right\rceil \tag{2.16}$$

$$t_{mn} = \left\lceil \frac{Q_m T_n}{r_{mn}^{tot}} \right\rceil \tag{2.17}$$

where ($\lceil . \rceil$) is the ceiling of a decimal number, and it is used to preserve the integral constraints (2.15g) and (2.15h). Eqs. (2.16) and (2.17) guarantee that $r_{mn} \geq Q_m$ (constraint (2.15b)) for MT m. The solution of problem **PL** is proposed in Algorithm 1.

The algorithm is based on two main phases. In the initialization phase, association variables are initialized to zero (*line 1.3*), the profit values of all networks

Algorithm 1: Network selection algorithm.

Output: Association and resource allocation decision for MT m

1.1: `// Initialization phase:`
1.2: **foreach** $n \in \mathcal{N}$ **do**
1.3: $x_{mn} = 0$;
1.4: calculate $\overline{f}_{mn}$;
1.5: `//` ψ_n `denotes the total number of free resources in network` $n \in \mathcal{N}$, `and` w_{mn} `the number of resources requested by MT` m `from network` $n \in \mathcal{N}$
1.6: **if** $n \in \mathcal{N}_{\mathcal{BS}}$ **then**
1.7: $\psi_n := U_n - \sum_{m \in \mathcal{M}} u_{mn} x_{mn}$;
1.8: $w_{mn} := c_{mn} = \left\lceil \frac{Q_m T_n}{B_n^{RB} \log_2(1+\gamma_{mn})} \right\rceil$; `//based on Eq.`(2.16)
1.9: **else** `//` $i.e.$ $n \in \mathcal{N}_{\mathcal{AP}}$
1.10: $\psi_n := T_n - \sum_{m \in \mathcal{M}} t_{mn} x_{mn}$;
1.11: $w_{mn} := t_{mn} = \left\lceil \frac{Q_m T_n}{r_{mn}^{tot}} \right\rceil$; `//based to Eq.`(2.17)
1.12: **end**
1.13: **end**
1.14: `// Network selection phase:`
1.15: $\mathcal{K} := \mathcal{J}$;
1.16: **while** $\mathcal{K} \neq \phi$ **do**
1.17: $n\prime = argmax_n(\overline{f}_{mn} \in \mathcal{K})$;
1.18: **if** $\overline{f}_{mn\prime} > 0$ **and** $w_{mn\prime} \leq \psi_{n\prime}$ **then**
1.19: $x_{mn\prime} = 1$;
1.20: **Break**;
1.21: **end**
1.22: $\mathcal{K} = \mathcal{K} - \{n\prime\}$;
1.23: **end**

are calculated (*line 1.4*), as well as the amount of resources requested by the MT (*lines 1.8 and 1.11*) and the amount of free resources at each network (*lines 1.7 and 1.10*). In the network selection phase, in each iteration in the while loop, the algorithm tries to associate MT m to the network $n\prime$ having the highest profit, after selecting the network with highest profit (*line 1.17*). If the network is reachable and the amount of free resources are enough to supply MT m with the requested data rate (*line 1.18*), MT m is associated to network $n\prime$ (*line 1.19*). Otherwise, the next highest-profit network is selected for trial. Since the algorithm aims to associate the MT to a single network, constraints (2.15c) and (2.15d) are preserved. Therefore, the proposed algorithm allocate resources and associates the MT to the top-ranked network while considering all the constraints of problem **PL**.

In the literature, solutions with the same perspectives as Algorithm 1 exists. However, the resource allocation constraints of access technologies are not always considered, or an abstract system model is used without considering the detailed resource allocation constraint. The novelty in Algorithm 1 is the proposition of a new network selection solution that aims at enhancing the power consumption

and RRSS while considering the data rate requirement of each MT and the network capacity constraints. Therefore, Algorithm 1 could be considered as a trivial profit-function-based load-aware network selection solution that will be used to benchmark other algorithms proposed in this thesis.

2.7 Global user association and resource allocation

In this section, the MT-based network selection solution proposed in the previous section is developed into a system-wide global user association and resource allocation solution. The global optimization function is processed at the SDN controller that maintains global view on the system.

2.7.1 Global profit function derivation

Since in this section we aim at optimizing the overall system-wide profit, it is necessary to modify the profit function derived in Section 2.6.1 (Eq. (2.13)) in order to consider system-wide values. Therefore, the global normalized forms of the signal quality and instantaneous power consumption, denoted by $\widehat{s}_{mn}$ and $\widehat{pc}_{mn}$ respectively, should be used,

$$\widehat{s}_{mn} = \frac{s_{mn}}{\max\limits_{m\in\mathcal{M},n\in\mathcal{N}}(s_{mn})} \tag{2.18}$$

$$\widehat{pc}_{mn} = \frac{1/pc_{mn}}{\max\limits_{m\in\mathcal{M},n\in\mathcal{N}}(1/pc_{mn})} \tag{2.19}$$

Note that each attribute is normalized to the global maximum, *i.e.* $\max\limits_{m\in\mathcal{M},n\in\mathcal{N}}$, instead of the local one ($\max\limits_{m\in\mathcal{M}}$) because the profit function will be deployed in a global optimization problem.

The normalized profit function does not differentiate between MTs with unequal data rate requirement. Therefore, the profit function is multiplied by the amount of data rate requested by the MT. Previously, the normalized profit function is not multiplied by the MT's required data rate because it is taken from the user's perspective to rank candidate networks. However, in the global solution, we consider MTs with unequal data rate requirements. Thus, it is important to multiply by the MT's data rate to reflect the real profit of each MT. Hence, the overall profit of MT m in network n is:

$$f_{mn} = \left(w_m^s \widehat{s}_{mn} + w_m^{pc} \widehat{pc}_{mn}\right) .Q_m.U(R_n - d_{mn}) \tag{2.20}$$

Note that $\frac{f_{mn}}{Q_m}$ can be seen as the normalized profit of a MT, or the profit per kbps, which is the real profit without multiplying by the requested data rate.

2.7.2 Novel global optimization problem formulation

It is intended to design an optimization problem that maximizes the total user-centric profit in the system while supplying MTs with their requested data rates

and considering the specific-resource-allocation constraints of access technologies. Hence, the formulated global optimization problem (**PG**) is:

$$\textbf{PG: } max \sum_{m\in\mathcal{M}}\sum_{n\in\mathcal{N}} f_{mn}x_{mn} \tag{2.21a}$$

$$s.\ t. \sum_{n\in\mathcal{N}} r_{mn}x_{mn} \geqslant \sum_{n\in\mathcal{N}} Q_m x_{mn} \qquad \forall m \in \mathcal{M} \tag{2.21b}$$

$$\sum_{n\in\mathcal{N}} x_{mn} \leqslant 1 \qquad \forall m \in \mathcal{M} \tag{2.21c}$$

$$x_{mn} \in \{0,1\} \qquad \forall m \in \mathcal{M}, \forall n \in \mathcal{N} \tag{2.21d}$$

$$\sum_{m\in\mathcal{M}} t_{mn}x_{mn} \leqslant T_n \qquad \forall\ n \in \mathcal{N}_{\mathcal{AP}} \tag{2.21e}$$

$$t_{mn}\ \in \mathbb{N} \qquad \forall m \in \mathcal{M}, \forall n \in \mathcal{N}_{\mathcal{AP}} \tag{2.21f}$$

$$\sum_{m\in\mathcal{M}} u_{mn}x_{mn} \leqslant U_n \qquad \forall\ n \in \mathcal{N}_{\mathcal{BS}} \tag{2.21g}$$

$$u_{mn} \in \mathbb{N} \qquad \forall m \in \mathcal{M}, \forall n \in \mathcal{N}_{\mathcal{BS}} \tag{2.21h}$$

Constraint (2.21b) guarantees that connected MTs receive data rate greater than or equal to their requested one. Constraints (2.21c) and (2.21d) ensure that a MT is connected to only one network, or not connected at all (upon congestion). Constraints (2.21e) and (2.21g) guarantee that the capacity of Wi-Fi APs and LTE eNodeBs is not exceeded, while constraints (2.21f) and (2.21h) assure that a single discrete resource object, *i.e.* time slot or RB, is not shared by multiple MTs at the same time.

2.7.3 Global optimization problem simplification

The formulated problem **PG** contains two types of variables: the association variables x_{mn} and the resource allocation variables u_{mn} and t_{mn}. To simplify the problem without violating its goals, we follow the same approach used in Section 2.6.3. So, the resource allocation variables are fixed while taking into consideration constraints (2.21b), (2.21f) and (2.21h). Therefore, based on Eq. (2.3) for LTE eNodeBs and Eq. (2.4) for Wi-Fi APs, the variables u_{mn} and t_{mn} are fixed and set to the minimal number of resources requested by MT m from network n as seen in Eqs. (2.16) and (2.17), where the ceiling ($\lceil . \rceil$) in both equations is used to preserve the integral constraints (2.21f) and (2.21h). Eqs. (2.16) and (2.17) guarantee that

$r_{mn} \geq Q_m$ at all active MTs. Therefore, problem **PG** simplifies to:

$$\textbf{P1: } max \sum_{m\in\mathcal{M}}\sum_{n\in\mathcal{N}} f_{mn}x_{mn} \tag{2.22a}$$

$$s.\,t. \sum_{m\in\mathcal{M}} \left\lceil \frac{Q_m T_n}{r_{mn}^{tot}} \right\rceil x_{mn} \leqslant T_n \qquad \forall n \in \mathcal{N}_{\mathcal{AP}} \tag{2.22b}$$

$$\sum_{m\in\mathcal{M}} \left\lceil \frac{Q_m T_n}{B_n^{RB}\log_2(1+\gamma_{mn})} \right\rceil x_{mn} \leqslant U_n \qquad \forall n \in \mathcal{N}_{\mathcal{BS}} \tag{2.22c}$$

$$\sum_{n\in\mathcal{N}} x_{mn} \leqslant 1 \qquad \forall m \in \mathcal{M} \tag{2.22d}$$

$$x_{mn} \in \{0,1\} \qquad \forall m \in \mathcal{M}, \forall n \in \mathcal{N} \tag{2.22e}$$

The simplified problem **P1** is a combinatorial optimization problem where the variables are restricted to have discrete binary values. Problem **P1** consists of finding an optimal set of association variables among a finite set of objects. In such problems, exhaustive search is not feasible for even a small-sized system. Therefore, solving the problem is not straightforward. In fact, the problem can be viewed as searching for the best set of association values among a finite, and usually very large, set of objects. Thus, its complexity is $\mathcal{O}(|\mathcal{N}|^{|\mathcal{M}|})$.

P1 operates on the domain of optimization problems where the set of feasible solutions is discrete, and in which the target is to find the best solution. Therefore, standard methods used to solve continuous optimization functions could not be applied for this problem. Since the variables in **P1** are restricted to have binary values and the objective function and constraints are linear, then **P1** is considered as binary linear programming problem (BLP), which could be also referred to as 0-1 integer linear programming problem. In fact, BLP is one of the Karp's 21 NP-complete problems [Kar72]. One class of algorithms used to solve BLPs are variants of the branch and bound method. Moreover, problem **P1** can be also seen as a generalized assignment problem (GAP) [MT90]. It is also indicated that the optimal solution of GAP could be found using the branch and bound algorithm [MT90]. Hence, the GNU linear programming kit (GLPK) is used to estimate the optimal solution based on the branch and bound algorithm. GLPK is intended to solve integer and linear programming optimization problems. However, as the number of variables grows largely, the optimal solution becomes intractable. Therefore, in this thesis, solutions with polynomial-time complexity should be proposed to approximate or solve problem **P1**.

2.8 Performance evaluation

In this section, we compare the effect of weight variation on the performance of the trivial profit-function-based solution (Algorithm 1) and the optimal solution of the global optimization problem **PG** based on the branch and bound method. For each solution, three weight variation cases (scenarios) shown in Table 2.1 are simulated. Specifically, we study, for each algorithm, the effect of increasing the number of active MTs on the average values of RRSS, and power consumption. In order to

focus on the effect of the solutions on the RRSS and MT power consumption, the considered simulation parameters ensure that all MTs will be served without facing any blockage. Other characteristics of the proposed solutions (such as blocking percentage, user satisfaction, processing time, *etc.*), and the effect of congestion on the behavior of the algorithms will be discussed in the next chapters.

TABLE 2.1
Simulated scenario name based on the solution and weights

Weights	Optimal solution (B&B)	Profit-function-based solution
$w_m^{pc} = 0.5,\ w_m^s = 0.5$	Optimal-Balanced	PF-Balanced
$w_m^{pc} = 0,\ w_m^s = 1$	Optimal-RRSS	PF-RRSS
$w_m^{pc} = 1,\ w_m^s = 0$	Optimal-PC	PF-PC

Note that PF denotes the profit-function-based network selection algorithm.

2.8.1 Simulation parameters

The simulation environment consists of two overlapping LTE BSs and two Wi-Fi APs within the service area (SA) (dashed area in Figure 2.3). Focusing on the SA allows us to test all the cases without having to simulate a very large number of networks. That is, we can test the case when MTs are placed near, or far away from, the boundaries of the cell, and when some MTs are in a single or multiple networks coverage area.

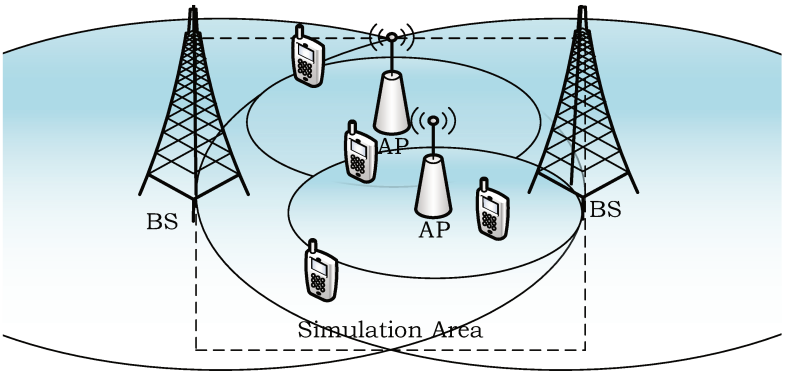


FIGURE 2.3 – Heterogeneous wireless system integrating LTE BSs and Wi-Fi APs. The dashed area is the service area where the simulation is focused.

Each MT is assumed to handle only one session where it randomly selects one of the data rates listed in Table 2.2.

Some of the characteristics and simulation parameters of both access technologies (LTE and Wi-Fi) are listed in Table 2.3. The constants (β_n and α_n) related to the MT power consumption estimation formula (Eq. (2.8)) are based on values collected while deriving the model in [HQG+12]. The coverage ranges of LTE BSs and Wi-Fi APs are rationally selected based on the parameters used in [NVACT13].

The number of available RBs at each BS is $C_n = 80$ and the transmission power per RB is $P_n^{RB} = 26$ dBm. A large number of RBs are considered to guarantee that MTs will not experience any blockage. The bandwidth of one RB is $B_n^{RB} = 180$

TABLE 2.2
Multiple data rates (kbps) for different applications

Voice call	Codec	G.729	G.726	G.711
	Data rate	32	56	87
Video call	Quality	Normal	Good	HD
	Data rate	300	500	1200
File download	Speed	Slow	Medium	Fast
	Data rate	150	700	1000

TABLE 2.3
Simulation parameters

Parameter	Wi-Fi AP	LTE BS
Path loss	$38.2 + 30\log_{10}(d_{mn})$	$34 + 40\log_{10}(d_{mn})$
β_n (W)	0.113286	1.28804
α_n (W/kbps)	0.13701×10^{-3}	0.05197×10^{-3}
Noise (dBm)	-90	-111.45
R_n (m)	100	500

kHz and the noise power at all the receivers in LTE is set to -111.45 dBm, which corresponds to the thermal noise at room temperature and bandwidth of 180 kHz [BB15]. The path loss between the LTE BS and a MT is modeled as $L(d_{mn}) = 34 + 40\log_{10}(d_{mn})$ [BB15].

Concerning Wi-Fi APs, a total bandwidth $B_n = 2000$ kHz is considered for each AP, with a total transmission power of 23 dBm. The path loss model is $38.2 + 30\log_{10}(d_{mn})$ and the noise power at the MT is -90 dBm [XGP+12].

A scheduling interval of 1 second is considered in the simulations [BB15]. In LTE BSs, $T_n = 1000$ so that the duration of one time slot is 1 millisecond which is the duration of one TTI in the LTE standard. For Wi-Fi APs, the scheduling interval is also 1 second and it is divided into 10000 time slots.

MTs are randomly distributed within the SA. The simulation is repeated for 10000 iterations in a Monte Carlo manner. In each iteration, the number of MTs increases from 50 to 100, the location of Wi-Fi APs and MTs changes randomly within the SA, and the initial seed of the random number generator also changes.

2.8.2 Evaluation metrics

The following metrics are used to evaluate the proposed solutions: average signal quality (RRSS) and average instantaneous power consumption. In fact, studying the average value of an attribute is not straightforward in a scenario where MTs request different amounts of data rate. For example, the average power consumption per user, *i.e.* $\frac{\sum_{m\in\mathcal{M}}\sum_{n\in\mathcal{N}} pc_{mn}x_{mn}}{|\mathcal{M}|}$, could be decreased through decreasing the power consumption of MTs with low data rate requirements on the expense of other MTs. Moreover, the average values per served data rate are considered because there is no mean to calculate these values for the blocked data rates. For example, setting a value of zero for the power consumption of unserved

MT will decrease the average consumed power and contribute misleading results. Therefore, the average value per served data rate unit, *i.e.* kbps, is used. Hence, the average MT power consumption per served data rate is studied according to the following formula:

$$\frac{\sum_{m\in\mathcal{M}}\sum_{n\in\mathcal{N}} pc_{mn}x_{mn}}{\sum_{m\in\mathcal{M}}\sum_{n\in\mathcal{N}} Q_m x_{mn}} \tag{2.23}$$

Since the value of the signal quality (s_{mn}) represents a normalized value that is not related to the requested data rate, it is multiplied by Q_m to reflect the actual signal quality per served data rate. Thus, the average relative received signal quality per served data rate is:

$$\frac{\sum_{m\in\mathcal{M}}\sum_{n\in\mathcal{N}} s_{mn}Q_m x_{mn}}{\sum_{m\in\mathcal{M}}\sum_{n\in\mathcal{N}} Q_m x_{mn}} \tag{2.24}$$

Note that as we have mentioned before, the simulation parameters considered in this chapter guarantee that MTs will not experience any blockage. Therefore, there is no difference between using the average value per served data rate and the average value per total requested data rate. In other words, in this chapter, $\sum_{m\in\mathcal{M}}\sum_{n\in\mathcal{N}} Q_m x_{mn} = \sum_{m\in\mathcal{M}} Q_m$ because simulation parameters guarantee that all MTs will be served. However, the formulas in Eqs. (2.23) and (2.24) will be used in the following chapters, where it is necessary to study the blocked data rate percentage, so both formulas consider the case when some MTs are not served.

2.8.3 Simulation results

We simulate all the solution scenarios listed in Table 2.1. First, we would like to discuss some network characteristics. Algorithms that are mainly concerned about power consumption, *i.e.* Optimal-PC and PF-PC, tend to associate MTs with Wi-Fi APs due to their low power consumption property when compared to LTE networks. This could be immediately remarked upon noticing that $\beta_n \backsimeq 1.28$ W for LTE networks, while for Wi-Fi networks $\beta_n \backsimeq 0.11$ W. On the other hand, algorithms concerned about signal quality (RRSS), *i.e.* Optimal-RRSS and PF-RRSS, tend to prioritize LTE BS due to their long transmission range property ($R_n = 500$ m). Moreover, as the number of MTs increases in the system, networks become more congested. Therefore, the probability that each MT is associated to its corresponding preferred network decreases, which in turns reflects an increase in the average MT power consumption and decrease in the average RRSS as shown in Figures 2.4 and 2.5 respectively.

It is shown in Figure 2.4 that Optimal-PC maintains the lowest average MT power consumption all the time, followed by the PF-PC solution. Optimal-PC and PF-PC are mainly concerned about the MT power consumption because the corresponding weight (w_m^{pc}) is set to one. On the other hand, RRSS-concerned solutions (Optimal-RRSS and PF-RRSS) scores the highest average power consumption due to setting its corresponding weight (w_m^{pc}) to zero. Concerning the effect of the global optimization problem **PG**, it is noted in Figure 2.4 that the solutions where $w_m^{pc} \neq 0$ and that are based on the branch and bound algorithm (Optimal-Balanced and Optimal-PC) decreases the average power consumption significantly when compared to PF-Balanced and PF-PC solutions respectively. In

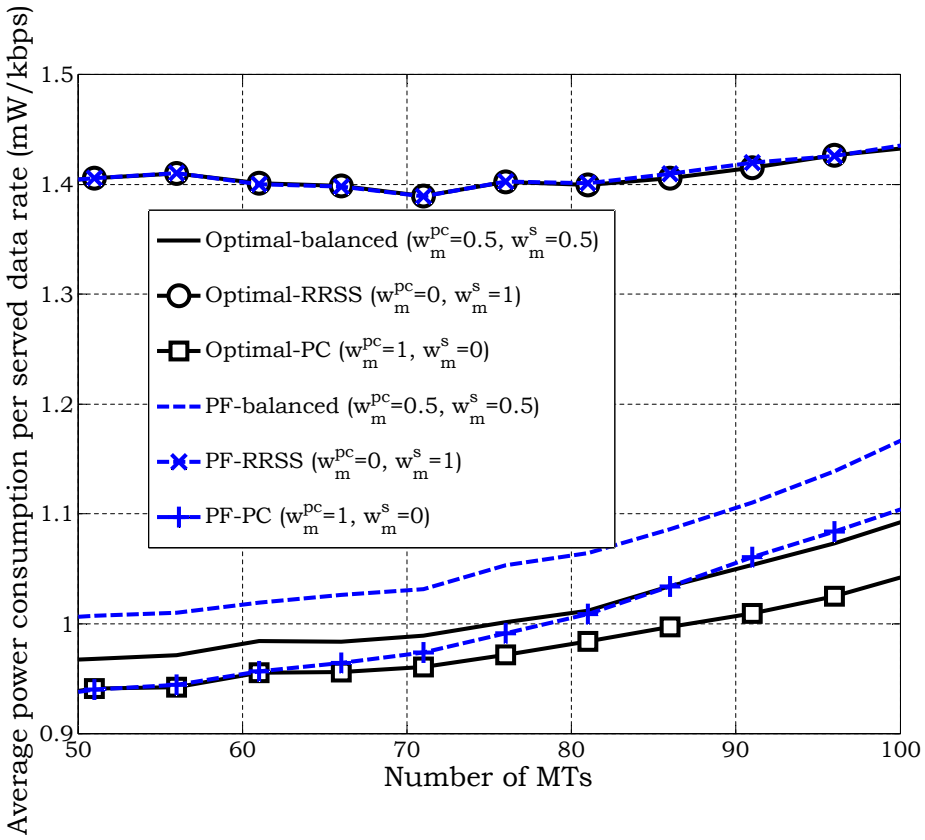


FIGURE 2.4 – Average MT instantaneous power consumption (based on Eq. (2.23)).

fact, increasing the number of MTs also increases the selectivity of the global optimization problem. Thus, the enhancement in performance between the optimized HO algorithms and the profit-function-based ones increases as well. For example, comparing PF-PC to Optimal-PC, the later decreases the power consumption by 7% when the number of MTs reaches 100. A major remark that should be noted is that the Optimal-PC solution, when compared to the Optimal-RRSS solution, reduces the average power consumption by 28% when the number of MTs equal to 100.

Following similar analysis methodology, it is normal to notice in Figure 2.5 that the Optimal-RRSS solution maintains the highest average RRSS value, followed directly by the PF-RRSS solution because w_m^s is set to one in both solutions. Since Optimal-PC, when compared to PF-PC, utilizes Wi-Fi APs more efficiently, then the average RRSS of Optimal-PC is lower than PF-PC due to the low RRSS property of Wi-Fi APs. As the number of MTs increases in the system, the selectivity of the global optimization function increases, which leads to an increase in the performance gap between the optimal solution and the profit-function-based solution. For example, when the number of MTs is between 50 and 60, the difference in the average RRSS between the optimal solution and the profit-function based solution is small. This difference increases remarkably when the number of MTs reaches 100.

It is remarkable that the solutions aiming at jointly enhancing the power

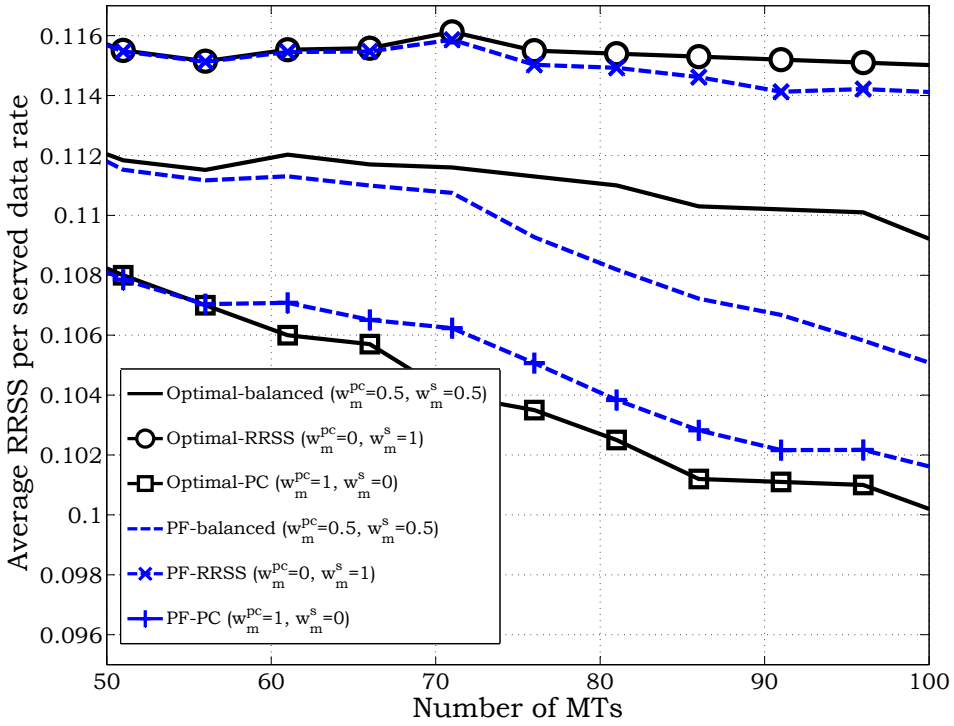


FIGURE 2.5 – Average relative received signal strength (based on Eq. (2.24)).

consumption and signal quality, *i.e.* Optimal-Balanced and PF-Balanced, could effectively decrease the average power consumption and increase the average RRSS simultaneously. Therefore, the profit function responds explicitly to the weight variations which is directly related to the user preferences. Moreover, a major enhancement of the global optimization problem **PG** is that the Optimal-Balanced solution, when compared to the PF-Balanced solution, jointly increases the average RRSS by 6.5% and decreases the average power consumption by approximately 7% when the number of MTs reaches 100. Note that in the following chapters, other important enhancements contributed by the global optimization problem are discussed.

2.9 Conclusion

In this chapter, we have discussed the 802.21.1 MIS SDRAN framework that provides seamless HO, resource allocation, and centralized management in HWNs. Based on this framework we have proposed a novel centralized HO scenario and formulated a new global optimization problem for the user association and downlink resource allocation problem in HWN. The novel formulated optimization problem considers the network's capacity and resource allocation constraints, and aims at maximizing the overall user-centric profit in the system. The user-centric profit is based on a weighted profit function that aims at jointly increasing the RRSS

and decreasing the MT power consumption. The weights of the profit function are set according to the user preferences. Simulation results have demonstrated that the formulated global optimization problem algorithm outperforms significantly the classical profit-function-based solution. Moreover, it is shown that the proposed profit function responds efficiently to the weight variations. Therefore, it can be effectively tuned to meet user preferences. However, in this chapter, the global optimization problem is solved using the branch and bound algorithm where the processing time could increase tragically upon increasing the number of MTs and networks in the system. Therefore, the following two chapters will respectively propose centralized and distributed solutions to the problem with acceptable complexity.

Chapter 3

New Methodologies for Centralized Resource and Network Assignment

In this chapter, we discuss and propose multiple centralized solutions for the user association and resource allocation problem in heterogeneous wireless systems. First, we propose two new solutions based on the continuous-relaxation of problem **P1**. The first proposed solution is with undetermined complexity, and is considered as the optimal solution based on the continuous-relaxation methodology. The second solution is with determined polynomial-time complexity, and is considered as a sub-optimal solution based on the continuous-relaxation approach. Then, a novel approximation-based solution is proposed to approximate the binary problem **P1**. In addition, a new simple greedy heuristic algorithm is also proposed. The performance of the approximation-based solution and greedy solution is boosted through a new proposed efficiency factor that estimates the gain contributed upon associating users to networks. The efficiency factor considers the data rate requirement of users, and the channel conditions between the MT and the BS or AP.

3.1 Brief related works

Usually, for the multi-mode MTs, user association is formulated as a binary matching problem. The user association variable is restricted to have a binary value that indicates if a MT is associated to a network. However, such problems are known to have an NP-complete complexity which makes the solution intractable. A popular approach used to overcome this issue is to relax the binary association variable into a continuous bounded one. Then, the solution of the relaxed problem, which usually has a polynomial-time complexity, is used to get the final association decision. As an example, two of the studies discussed in Chapter 1 in are discussed. In [YRC+12], a simple rounding approach is used to convert the fractional association variables into boolean. In [ZHY15], the MT connects to the network with the highest fractional association value. However, both solutions are not suitable for the case where MTs request a specific data rate ; both approaches could lead to a congested network where the number of available resources is not sufficient

to supply each MT with its requested data rate. Instead, they are designed to increase the total throughput in a system where there is no restriction on each user's data rate. Moreover, relaxing the binary constraint threatens the optimality of the solution. Therefore, it is necessary to provide for problem **P1** (formulated in the previous chapter) solutions that respect the amount of data rate requested by each MT.

3.2 Relaxation of the binary constraint

3.2.1 Continuous-relaxation of the binary constraint

In this section, the discrete (binary) problem **P1** is converted to a continuous optimization problem by relaxing the binary constraint ($x_{mn} \in \{0,1\}$) into a continuous bounded constraint $0 \leq x_{mn} \leq 1$, *i.e.* $x_{mn} \in [0,1]$. Thus, each MT is now allowed to access multiple networks simultaneously, *i.e.* multi-homing. Then, a new methodology is proposed to preserve the single network association constraint by considering only boolean association results. Converting the discrete problem into a continuous permits the usage of typical mathematical methodologies that are intended to solve continuous optimization problems within a polynomial-time complexity. Therefore, the new relaxed problem is:

$$\textbf{P2: } max \sum_{m\in\mathcal{M}}\sum_{n\in\mathcal{N}} f_{mn}x_{mn} \tag{3.1a}$$

$$s.\,t. \sum_{m\in\mathcal{M}} \left\lceil \frac{Q_m T_n}{r_{mn}^{tot}} \right\rceil x_{mn} \leqslant T_n \qquad \forall n \in \mathcal{N}_{\mathcal{AP}} \tag{3.1b}$$

$$\sum_{m\in\mathcal{M}} \left\lceil \frac{Q_m T_n}{B_n^{RB}\log_2(1+\gamma_{mn})} \right\rceil x_{mn} \leqslant U_n \qquad \forall n \in \mathcal{N}_{\mathcal{BS}} \tag{3.1c}$$

$$\sum_{n\in\mathcal{N}} x_{mn} \leqslant 1 \qquad \forall m \in \mathcal{M} \tag{3.1d}$$

$$x_{mn} \in [0,1] \qquad \forall m \in \mathcal{M}, \forall n \in \mathcal{N} \tag{3.1e}$$

where the inequalities (3.1b), (3.1c), and (3.1d) specify a convex polytope over which the profit function is to be optimized. Note that **P2** is similar to **P1**, except for the last constraint (3.1e).

In fact, problem **P2** is a linear optimization problem because the objective function and the constraints are linear, and the problem can be expressed in the canonical form of the standard linear programming problem as shown below:

$$\begin{aligned} max \quad & \boldsymbol{c}^\top \boldsymbol{x} \\ s.\,t. \quad & \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b} \\ & \boldsymbol{x} \geq 0 \end{aligned} \tag{3.2}$$

where:

- $\boldsymbol{c} \in \mathbb{R}^{|\mathcal{M}|.|\mathcal{N}|}$ is a vector that contains all the profit values f_{mn}, and $(\cdot)^\top$ is the matrix transpose.
- $\boldsymbol{x} \in \mathbb{R}^{|\mathcal{M}|.|\mathcal{N}|}$ is a vector that contains all the user association variables x_{mn}.

– $\boldsymbol{A} \in \mathbb{R}^{(|\mathcal{M}|+|\mathcal{N}|)\times(|\mathcal{M}|.|\mathcal{N}|)}$ and $\boldsymbol{b} \in \mathbb{R}^{|\mathcal{M}|+|\mathcal{N}|}$ are respectively a matrix and a vector of coefficients related to constraints (3.1b), (3.1c), and (3.1d); $\boldsymbol{A}$ contains the coefficients at the left side of the inequalities in the constraints, and $\boldsymbol{b}$ the constants on the right side.

It is appropriate to consider constraint (3.1e) as $x_{mn} \geq 0$ alone (without the upper bound) because constraint (3.1d) guarantees that $x_{mn} \leq 1$. For simplicity throughout this thesis, assume that ω_{mn} represents the number of resources requested by MT m from network $n \in \mathcal{N}$, and ζ_n the total number of resources in network $n \in \mathcal{N}$. Note that ω_{mn} can be seen as the weight of MT m in network $n \in \mathcal{N}$, and ζ_n the capacity of network $n \in \mathcal{N}$. Hence, ω_{mn} and ζ_n are respectively defined as:

$$\omega_{mn} = \begin{cases} t_{mn}, & if\ n \in \mathcal{N}_{\mathcal{AP}}, \\ u_{mn}, & if\ n \in \mathcal{N}_{\mathcal{BS}}. \end{cases} \tag{3.3}$$

$$\zeta_n = \begin{cases} T_n, & if\ n \in \mathcal{N}_{\mathcal{AP}}, \\ U_n, & if\ n \in \mathcal{N}_{\mathcal{BS}}. \end{cases} \tag{3.4}$$

Now, for illustrative purposes only, let us consider a heterogeneous system of two networks and three MTs that are placed in the coverage area of both networks, then:

$\boldsymbol{c}^\top = \begin{bmatrix} f_{11} & f_{12} & f_{21} & f_{22} & f_{31} & f_{32} \end{bmatrix}$,

$\boldsymbol{x} = \begin{bmatrix} x_{11} & x_{12} & x_{21} & x_{22} & x_{31} & x_{32} \end{bmatrix}^\top$,

$$\boldsymbol{A} = \overbrace{\begin{bmatrix} \omega_{11} & 0 & \omega_{21} & 0 & \omega_{31} & 0 \\ 0 & \omega_{12} & 0 & \omega_{22} & 0 & \omega_{32} \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}}^{|\mathcal{M}|.|\mathcal{N}|} \begin{matrix} |\mathcal{N}| \\ \\ + \\ \\ |\mathcal{M}| \end{matrix} \begin{matrix} \Big\} (3.1b), (3.1c) \\ \\ \Big\} (3.1d) \end{matrix} , \boldsymbol{b} = \begin{bmatrix} \zeta_1 \\ \zeta_2 \\ 1 \\ 1 \\ 1 \end{bmatrix} \tag{3.5}$$

If a network n is not reachable by some MT m, *i.e.* $f_{mn} = 0$, its corresponding entry is removed from $\boldsymbol{x}$ and $\boldsymbol{c}$ vectors, as well as its corresponding column in matrix $\boldsymbol{A}$. For example, assume that network 2 is not reachable by MT 1, then, x_{12} and f_{12} are removed, in addition to the second column in $\boldsymbol{A}$. Note that expressing problem **P2** in the standard form of a linear program, as shown in (3.2) and based on (3.5), is the first step toward solving the linear program. Moreover, it can serve as a proof of linearity for both problems **P1** and **P2**.

The number of variables in the standard linear program **P2** is $|\mathcal{M}|.|\mathcal{N}|$. In practice, the simplex method performs very well when used to solve this linear program even for a large number of variables. However, its worst-case computational complexity is exponential [PW]. Other methods with polynomial-time complexity have been proposed to solve standard linear programs. The interior-point methods are preferred among them; the theoretical computational complexity is $\mathcal{O}(|\mathcal{M}|^3|\mathcal{N}|^3L)$, where L is the length of the binary coding of the input data [PW]. The fact that the complexity depends on L implies that the time required to solve the problem increases with the required accuracy of the computations.

Actually, solving the relaxed problem **P2** produces three sets of MTs based on their connection state:

– Λ_0: set of unconnected MTs, *i.e.* $\sum_{n\in\mathcal{N}} x_{mn} = 0$.

– Λ_1: set of MTs connected to a single network n such that $x_{mn} = 1$.

– Λ_2: set of MTs connected to several networks simultaneously or to a single network n such that $x_{mn} < 1$, *i.e.* the MT receives data rate less than Q_m.

Since in our context, a MT should be connected to only one network and receives all requested resources, a continuous-relaxation approach is used to associate MTs to appropriate networks and empty the set Λ_2. In this chapter, we propose two user association and resource allocation algorithms based on the continuous-relaxation of problem **P1**. The first solution is with undetermined complexity while the second is with determined polynomial-time complexity. The solution with undetermined complexity can be seen as the optimal solution based on the continuous-relaxation approach, and it is used for benchmarking purposes only because it is intended to propose resource and network assignment solutions with determined polynomial-time complexity.

3.2.2 Novel continuous-relaxation-based solution with undetermined complexity

Algorithm 2 shows the continuous-relaxation-based solution with undetermined complexity.

Algorithm 2: Optimal solution based on the continuous-relaxation approach

Output: Association variables for all MTs
foreach $n \in \mathcal{N}$ **do**
  $\overline{\Gamma_n} := \zeta_n$; // $\overline{\Gamma_n}$ denotes the number of free resources in network n
end
$\overline{\mathcal{M}} := \mathcal{M}$;
while $\overline{\mathcal{M}} \neq \phi$ **do**
  solve problem **P2** $\forall m \in \overline{\mathcal{M}}$ and according to $\overline{\Gamma_n}$;
  $\Lambda_0 := \{m \in \overline{\mathcal{M}} : \sum_{n \in \mathcal{N}} x_{mn} = 0\}$;
  $\Lambda_1 := \{m \in \overline{\mathcal{M}} : \sum_{n \in \mathcal{N}} \lfloor x_{mn} \rfloor = 1\}$; // $\lfloor \cdot \rfloor$ is the floor of a decimal number
  $\Lambda_2 := \overline{\mathcal{M}} - \Lambda_0 - \Lambda_1$;
  if $|\Lambda_2| = \phi$ **then**
    save the association values $\forall m \in \Lambda_1 \cup \Lambda_0$;
    $\overline{\mathcal{M}} = \overline{\mathcal{M}} - \Lambda_0 - \Lambda_1 - \Lambda_2$; // *i.e.* $\overline{\mathcal{M}} = \phi$
  end
  save the association values $\forall m \in \Lambda_1$;
  $\overline{\mathcal{M}} = \overline{\mathcal{M}} - \Lambda_1$;
  foreach $n \in \mathcal{N}$ **do**
    $\overline{\Gamma_n} = \overline{\Gamma_n} - \sum_{m \in \Lambda_1} \omega_{mn} x_{mn}$;
  end
end

Mainly, Algorithm 2 keeps solving the linear optimization problem **P2** for all MTs in $\Lambda_0 \cup \Lambda_2$ until $|\Lambda_2| = 0$, *i.e.* all association results of problem **P2** are boolean.

Every time the optimization problem is solved, the association values of MTs in Λ_1 are saved, the number of free resources in each network is updated by deducting the number of resources allocated to MTs in Λ_1 (*lines 2.16-2.18*), and those MTs are not considered within the optimization function anymore (*line 2.15*). However, it is impossible to determine analytically the number of times the linear program will be solved in Algorithm 2, therefore, the complexity of this algorithm could not be determined.

3.2.3 Novel continuous-relaxation-based solution with polynomial time complexity

In this section, a solution with polynomial-time complexity is proposed to determine the association values for MTs in $\Lambda_0 \cup \Lambda_2$. Problem **P2** is solved only once and each MT m in Λ_1 is associated to the network n where $x_{mn} = 1$. Algorithm 3 determines the association values for all MTs in Λ_2 only, *i.e.* MTs with fractional association values.

Algorithm 3: Convert fractional association variables into boolean

3.1: **foreach** $n \in \mathcal{N}$ **do**
3.2: $\quad \overline{\Gamma_n} := \zeta_n - \sum_{m \in \Lambda_1} \omega_{mn} x_{mn}$;
3.3: **end**
3.4: $X := \{x_{mn} : m \in \Lambda_2 \text{ and } x_{mn} \neq 0\}$;
3.5: **while** $X \neq \phi$ **do**
3.6: $\quad m^* = argmax_m(X)$;
3.7: $\quad n^* = argmax_n(X)$;
3.8: $\quad$ **if** $\sum_{n \in \mathcal{N}} \lfloor x_{m^* n} \rfloor = 0$ **then**
3.9: $\quad\quad$ **if** $\omega_{m^* n^*} \leq \overline{\Gamma_{n^*}}$ **then**
3.10: $\quad\quad\quad x_{m^* n^*} = 1$;
3.11: $\quad\quad\quad \overline{\Gamma_{n^*}} = \overline{\Gamma_{n^*}} - \omega_{m^* n^*}$;
3.12: $\quad\quad$ **else**
3.13: $\quad\quad\quad x_{m^* n^*} = 0$;
3.14: $\quad\quad$ **end**
3.15: $\quad$ **else**
3.16: $\quad\quad x_{m^* n^*} = 0$;
3.17: $\quad$ **end**
3.18: $\quad X = X - \{x_{m^* n^*}\}$;
3.19: **end**

Basically, in each iteration within the while loop, Algorithm 3 attempts to determine the association decision for the MT with the highest association value among all MTs in Λ_2, *i.e.* $x_{m^* n^*}$. If the association decision for MT m^* is not determined yet (*line 3.8*), and the number of resources requested by MT m^* from the target network n^* is less than or equal to the number of free resources in network n^* (*line 3.9*), then MT m^* is associated to network n^* (*line 3.10*), and the number of free resources in network n^* is updated by deducting the number of resources allocated to MT m^* (*line 3.11*).

The procedures in Algorithm 3 could be implemented efficiently by sorting the association values of X in the descending order, then the pointer to the maximum association value becomes immediately available in each iteration within the while loop. Assuming that all MTs are in the coverage areas of all networks, then the total number of variables in X is $|\Lambda_2|.|\mathcal{N}|$. Thus, and based on the quicksort [Ski09] algorithm, the complexity for sorting the items in X is $\mathcal{O}(|\Lambda_2|.|\mathcal{N}|\log(|\Lambda_2|.|\mathcal{N}|))$ and the total complexity of Algorithm 3 is $\mathcal{O}(|\Lambda_2|.|\mathcal{N}|\log(|\Lambda_2|.|\mathcal{N}|) + |\Lambda_2|.|\mathcal{N}|)$ because the complexity of the while loop is $\mathcal{O}(|\Lambda_2|.|\mathcal{N}|)$ and it is executed after the sorting step. If the quicksort algorithm is well-implemented, it can perform 2-3 times faster than its main competitors mergesort and heapsort [Ski09].

Subsequently, all MTs whose association decision is still not determined, *i.e.* each MT m such that $\sum_{n\in\mathcal{N}} x_{mn} = 0$, undergo procedures similar to those proposed in Algorithm 3, however, the goal here, illustrated in Algorithm 4, is to associate the MT with the highest normalized profit ($\frac{f_{mn}}{Q_m}$) at each iteration instead of the highest fractional association value. The normalized profit is used because it reflects a normalized value that is unrelated to the amount of data rate requested by MTs. Similarly, the complexity of Algorithm 4 is $\mathcal{O}(|X|\log|X|+|X|)$ where X is defined in *line 4.4*.

Algorithm 4: Determine association decision for the remaining unassociated MTs

4.1: **foreach** $n \in \mathcal{N}$ **do**
4.2: $\quad \overline{\Gamma_n} := \zeta_n - \sum_{m\in\Lambda_1} \omega_{mn}x_{mn} - \sum_{m\in\Lambda_2} \omega_{mn}x_{mn}$;
4.3: **end**
4.4: $X := \{\frac{f_{mn}}{Q_m} : m \in \mathcal{M}$ *and* $\sum_{n\in\mathcal{N}} x_{mn} = 0\}$;
4.5: *Lines 3.5-3.17*
4.6: $X = X - \{\frac{f_{m^*n^*}}{Q_{m^*}}\}$;
4.7: *Line 3.19*

3.3 Novel approximation-based solution

3.3.1 Generalized assignment problem approach

After taking a closer look at problem **P1**, one may notice that it is similar to the GAP [MT81]. In-fact, Martello and Toth, who have significant contributions in the domain of GAP, knapsack, and bin-packing problems, have proposed a heuristic algorithm to approximate GAP based on an ordering of the MTs [MT81]. There, the "desirability" of assigning MT m to network n is measured according to a desirability factor Ω_{mn}. The possible factors that could be considered as a desirability measure are discussed in the next section. For each MT, the difference between the highest and the second highest value of Ω_{mn} is computed, and MTs are then assigned in the decreasing order of this difference. That is, each MT is assigned to its best network according to the following criteria:

$$\max_{n} \min_{n\neq n\prime} (\Omega_{mn\prime} - \Omega_{mn}) \tag{3.6}$$

or in other words:

$$\begin{aligned} &\min_{n \neq n\prime} \Omega_{mn\prime} - \Omega_{mn} \\ &\text{where} \\ &n\prime = \arg\max_{n} \Omega_{mn} \end{aligned} \tag{3.7}$$

The computational experiments conducted by Martello and Toth have shown that good results are obtained using this algorithm.

3.3.2 Generalized assignment problem adaptation

In fact, the algorithm proposed by Martello and Toth does not exactly suit problem **P1** for two reasons:

- The algorithm is designed to solve GAP while constraint (2.22d), which is $\sum_{n \in \mathcal{N}} x_{mn} \leq 1$, is replaced by $\sum_{n \in \mathcal{N}} x_{mn} = 1$. That is, all MTs should be associated to networks. While upon congestion, some MTs would not be able to associate to any network. Then, the algorithm would fail to approximate problem **P1**.
- The algorithm assumes that all networks are reachable by all MTs. Therefore, it does not differentiate between MTs reachable by a single network and others reachable by multiple networks.

Thus, we modify their proposed algorithm to adapt problem **P1** as shown in Algorithm 5 and explained hereafter.

At first, all association variables for MTs in $\mathcal{M}$ are set to 0. Algorithm 5 iteratively considers all the unassociated MTs, and determines the MT m^* having the maximum difference between the highest and the second highest Ω_{mn} ($n \in F_m$ where F_m is defined in *line 5.9*). MT m^* is then assigned to the network for which Ω_{m^*n} is maximum, *i.e.* network n^*. It is this property of the algorithm which leads to significant results when tested ; the algorithm considers the second maximum Ω_{mn} instead of focusing only on the first maximum. Moreover, after taking each association decision, the algorithm re-evaluates, for each MT, the maximum difference between the highest and the second highest Ω_{mn}, and associates MTs based on these new results. Thus, a semi-global view on the available networks and their profit is maintained while taking association decisions. In addition, the algorithm prefers to first associate MTs with only one available network, *i.e.* $|F_m| = 1$. We add the if block in *lines 5.16-5.20* to associate the MT with highest Ω_{mn} among other MTs with a single available network. Initially, the original algorithm associates any MT with a single available network without taking into consideration the value of Ω_{mn}. This aspect of the algorithm plays a vital role in decreasing the blocking probability.

Algorithm 5 can be implemented efficiently by initially sorting in decreased order, for each MT m, the values Ω_{mn} ($n \in \mathcal{N}$). This requires $\mathcal{O}(|\mathcal{N}| \log |\mathcal{N}|)$ for a single MT. Thus for all MTs $m \in \mathcal{M}$ it requires $\mathcal{O}(|\mathcal{M}||\mathcal{N}| \log |\mathcal{N}|)$. The sorting step makes immediately available, at each iteration in the inner loop, the pointers to the maximum and the second maximum Ω_{mn}. Hence, the main while loop performs the $\mathcal{O}(|\mathcal{M}|)$ associations within a total of $\mathcal{O}(|\mathcal{M}|^2)$ iterations ; whenever a MT is assigned, the decrease in $\overline{\zeta_{n^*}}$ makes it necessary to update the pointers. Since, however, the above maxima can only decrease during execution, a total of

Algorithm 5: Approximation algorithm

Output: Association variables for all MTs in $\mathcal{M}$

5.1: $\mathcal{U} := \mathcal{M}$;
5.2: **foreach** $n \in \mathcal{N}$ **do**
5.3: $\quad \overline{\zeta_n} := \zeta_n$;
5.4: **end**
5.5: **while** $\mathcal{U} \neq \phi$ **do**
5.6: $\quad c^* := -\infty$;
5.7: $\quad d^* := -\infty$;
5.8: $\quad$ **foreach** $m \in \mathcal{U}$ **do**
5.9: $\quad\quad F_m := \{n \in \mathcal{N} : f_{mn} \neq 0 \wedge \omega_{mn} \leq \overline{\zeta_n}\}$;
5.10: $\quad\quad$ **if** $F_m = \phi$ **then**
5.11: $\quad\quad\quad \mathcal{U} = \mathcal{U} - \{m\}$;
5.12: $\quad\quad$ **else**
5.13: $\quad\quad\quad n\prime = argmax_n\{\Omega_{mn} : n \in F_m\}$;
5.14: $\quad\quad\quad$ **if** $/F_m/$=1 **then**
5.15: $\quad\quad\quad\quad d := +\infty$;
5.16: $\quad\quad\quad\quad$ **if** $c^* < \Omega_{mn\prime}$ **then**
5.17: $\quad\quad\quad\quad\quad c^* = \Omega_{mn\prime}$;
5.18: $\quad\quad\quad\quad\quad n^* := n\prime$;
5.19: $\quad\quad\quad\quad\quad m^* := m$;
5.20: $\quad\quad\quad\quad$ **end**
5.21: $\quad\quad\quad$ **else**
5.22: $\quad\quad\quad\quad d := \Omega_{mn\prime} - max_2\{\Omega_{mn} : n \in F_m\}$;
5.23: $\quad\quad\quad\quad$ **if** $d > d^*$ **then**
5.24: $\quad\quad\quad\quad\quad d^* = d$;
5.25: $\quad\quad\quad\quad\quad n^* := n\prime$;
5.26: $\quad\quad\quad\quad\quad m^* := m$;
5.27: $\quad\quad\quad\quad$ **end**
5.28: $\quad\quad\quad$ **end**
5.29: $\quad\quad$ **end**
5.30: $\quad$ **end**
5.31: $\quad$ **if** $d \neq -\infty$ **then**
5.32: $\quad\quad x_{m^*n^*} = 1$;
5.33: $\quad\quad \mathcal{U} = \mathcal{U} - \{m^*\}$;
5.34: $\quad\quad \overline{\zeta_{n^*}} = \overline{\zeta_{n^*}} - \omega_{m^*n^*}$;
5.35: $\quad$ **end**
5.36: **end**

$\mathcal{O}(|\mathcal{M}|^2)$ operations is required for these checks and updates. Thus, we conclude that the overall complexity of Algorithm 5 is $\mathcal{O}(|\mathcal{M}|^2 + |\mathcal{M}||\mathcal{N}| \log |\mathcal{N}|)$. It is clear that the complexity of Algorithm 5 is less than the complexity of solving the relaxed problem **P2**.

3.3.3 New efficiency factor

Problem **P1** aims at maximizing the profit in the system. Therefore, Martello and Toth have proposed in [MT81] to use the profit (f_{mn}) or $\frac{\text{profit}}{\text{weight}}$ as desirability factor in Algorithm 5. Since problem **P1** deals with MTs of different data rate requirements, *i.e.* different weights, the $\frac{\text{profit}}{\text{weight}}$ ratio is suitable as desirability factor for this problem. However, considering the weight of MTs, which can be seen as the number of requested resources, is not straightforward because access technologies have different types and amounts of resources. As a matter of fact, the amount of bandwidth that should be supplied by a network to a MT is related to the channel conditions between the MT and the network, and to the amount of data rate requested by the MT. Moreover, the bandwidth (in Hz) is a limited resource in all communication systems. Therefore, the amount of bandwidth requested by a MT from a network could be considered as a weight. The amount of bandwidth requested by MT m from BS n is $\frac{B_n^{RB} u_{mn}}{T_n}$, and from AP n is $\frac{B_n t_{mn}}{T_n}$. Hence, the efficiency e_{mn} is introduced to denote the profit per weight (requested bandwidth) contributed to the system upon associating MT m to network n such that:

$$e_{mn} = \begin{cases} \frac{f_{mn}}{B_n^{RB}(u_{mn}/T_n)} & \forall n \in \mathcal{N}_{\mathcal{BS}} \\ \\ \frac{f_{mn}}{B_n(t_{mn}/T_n)} & \forall n \in \mathcal{N}_{\mathcal{AP}} \end{cases} \tag{3.8}$$

The normalized profit ($\frac{f_{mn}}{Q_m}$) could be also considered as desirability factor, but it does not consider the number of requested resources. Although the normalized profit reflects the actual profit contributed upon associating a MT, however, the quality of the channel between the MT and the BS or AP is not considered. Therefore, the efficiency is chosen as a main desirability factor. The difference between using the efficiency (e_{mn}) and the normalized profit ($\frac{f_{mn}}{Q_m}$) as desirability factors is discussed in the performance evaluation section.

3.4 Novel greedy solution

Even if the algorithm proposed in the previous section allows to reduce the complexity of the related problem, we may find one greedy approach to further reduce the complexity. In this section, a simple greedy heuristic algorithm is proposed to find a good feasible solution with polynomial-time complexity for problem **P1**.

Ordinarily, the greedy heuristic does not produce an optimal solution, but even so, it may yield to locally optimal solutions that approximate the global one in an acceptable amount of time. To estimate the abstract gain, denoted by Ω_{mn}, contributed upon associating MT m to network n, an abstract criteria is used. The criteria that could be used within the greedy algorithm and the mechanism of calculating Ω_{mn} will be discussed shortly. Algorithm 6 shows the phases of the greedy heuristic.

The proposed greedy algorithm is made up of two phases. In the *preparation phase*, the profit (f_{mn}) and the gain (Ω_{mn}) values for all MTs are calculated, the number of free resources in each network ($\overline{\zeta}_n$) is initialized, the set of gain values

Algorithm 6: Greedy heuristic algorithm

Output: Association variables (x_{mn}) for all MTs

6.1: `// Preparation phase:`
6.2: $\mathcal{E} := \phi$; `//` $\mathcal{E}$ `represents a set that will be filled with all non-zero gain values` (Ω_{mn})
6.3: **foreach** $n \in \mathcal{N}$ **do**
6.4: $\quad \overline{\zeta}_n := \zeta_n$;
6.5: $\quad$ **foreach** $m \in \mathcal{M}$ **do**
6.6: $\quad\quad x_{mn} = 0$;
6.7: $\quad\quad$ **if** $f_{mn} \neq 0$ **then**
6.8: $\quad\quad\quad \mathcal{E} = \mathcal{E} + \{\Omega_{mn}\}$; `//` Ω_{mn} `is the abstract gain based on the criteria used`
6.9: $\quad\quad$ **end**
6.10: $\quad$ **end**
6.11: **end**
6.12: `// Decision phase:`
6.13: **while** $\mathcal{E} \neq \phi$ **do**
6.14: $\quad m\prime = argmax_m(\Omega_{mn} \in \mathcal{E})$;
6.15: $\quad n\prime = argmax_n(\Omega_{mn} \in \mathcal{E})$;
6.16: $\quad$ **if** $\sum_{n \in \mathcal{N}} x_{m\prime n} = 0$ **and** $\omega_{m\prime n\prime} \leq \overline{\zeta}_{n\prime}$ **then**
6.17: $\quad\quad x_{m\prime n\prime} = 1$;
6.18: $\quad\quad \overline{\zeta}_{n\prime} = \overline{\zeta}_{n\prime} - \omega_{m\prime n\prime}$;
6.19: $\quad$ **end**
6.20: $\quad \mathcal{E} = \mathcal{E} - \{\Omega_{m\prime n\prime}\}$;
6.21: **end**

($\mathcal{E}$) is filled, and all association variables (x_{mn}) are set to zero. In the *decision phase*, the algorithm aims at associating the MT with the highest gain in each iteration within the while loop (*lines 6.13-6.21*). After detecting the highest gain, the algorithm, in *line 6.16*, tests if the corresponding MT, *i.e.* $m\prime$, is not associated yet ($\sum_{n \in \mathcal{N}} x_{m\prime n} = 0$), and if the number of free resources in the target network $n\prime$ is sufficient to serve MT $m\prime$ ($\omega_{m\prime n\prime} \leq \overline{\zeta}_{n\prime}$). If both conditions are true, MT $m\prime$ is associated to network $n\prime$ (*line 6.17*), and the number of free resources in network $n\prime$ is updated by subtracting the number of resources allocated to MT $m\prime$ (*line 6.18*).

However, when misused, the greedy algorithm might produce misleading results. For example, considering the profit value as a criteria to estimate the gain in the greedy algorithm, *i.e.* ($\Omega_{mn} = f_{mn}$), is improper because usually MTs with high data rate requirement will have higher profit value, which leads to unfair satisfaction between MTs demanding different amounts of data rate. On the other hand, considering the normalized profit value, *i.e.* $\Omega_{mn} = \frac{f_{mn}}{Q_m}$, is somehow suitable in such case because it reflects a normalized quantity that is not related to the amount of requested data rate. Therefore, the efficiency factor (e_{mn}), introduced in the previous section, is also considered in the greedy algorithm.

Concerning the complexity of the greedy algorithm, *lines 6.2-6.11* could be exe-

cuted within a single step that calculates the gain values, and sorts these values in the descending order while inserting them into the set $\mathcal{E}$. The sorting usually requires an average of $\mathcal{O}(|\mathcal{M}||\mathcal{N}| \log |\mathcal{M}||\mathcal{N}|)$ iterations. Sorting the gain values of $\mathcal{E}$ in the descending order makes the pointer for the maximum gain value immediately available in each iteration in the while loop. Assuming that all the networks are reachable by all the MTs, *i.e.* $|\mathcal{E}| = |\mathcal{M}||\mathcal{N}|$, the average complexity of Algorithm 6 is $\mathcal{O}(|\mathcal{M}||\mathcal{N}| + |\mathcal{M}||\mathcal{N}| \log |\mathcal{M}||\mathcal{N}|)$ because the while loop requires $\mathcal{O}(|\mathcal{E}|)$ iterations and it is executed after the sorting step.

3.5 Implementation feasibility

The user association and resource allocation solutions proposed in this chapter are executed on a centralized SDN controller as shown in Figure 2.1. The centralized controller acquires instantaneous contextual information related to the number of resources available at each network, the data rate requested by MTs, the user preferences, the geographical locations of MTs and networks, and the characteristics of networks (*i.e.* power consumption characteristics, coverage range, *etc.*). After taking association and resource allocation decisions, the centralized controller informs networks and MTs about the results. The solutions proposed in this chapter are designed in a way such that the system-wide resource allocation and user association decisions should be processed on the same entity, *i.e.* the SDN controller. Therefore, these solutions are considered as centralized resource and network assignment solutions.

3.6 Complexity comparison

Some of the discussed complexities are presented in Table 3.1. As we have mentioned before, the complexity of Algorithm 2 could not be determined because it is impossible to estimate the number of times the relaxed problem **P2** is solved. However, the complexity of solving the relaxed continuous problem is $\mathcal{O}(|\mathcal{M}|^3|\mathcal{N}|^3 L)$. Therefore, the complexity of Algorithm 2, which is seen as the optimal solution based on the continuous-relaxation approach, is for sure higher than the complexity of solving **P2**. Moreover, the solution with determined polynomial-time complexity, proposed in Section 3.2.3, which is based on solving the problem **P2** once, in addition to Algorithms 3 and 4, has a complexity also higher than that of solving problem **P2** once. The complexity of the approximation-based solution (Algorithm 5) is lower than the complexity of solving the continuous problem **P2**. Finally, the complexity of the greedy solution (Algorithm 6) is lower than the approximation-based solution and the continuous-relaxation-based solutions. Hence, it is important in the performance evaluation section to discuss the complexity-performance trade-off.

3.7 Performance evaluation

In this section, we compare the performance of the different solutions proposed to solve, or approximate, the binary problem **P1**. Specifically, we study the effect

TABLE 3.1
Solution complexity

	Complexity
The binary problem **P1**	$\mathcal{O}(\lvert\mathcal{N}\rvert^{\lvert\mathcal{M}\rvert})$
Solving the continuous problem **P2**	$\mathcal{O}(\lvert\mathcal{M}\rvert^3\lvert\mathcal{N}\rvert^3 L)$
Algorithm 5 (Approximation)	$\mathcal{O}(\lvert\mathcal{M}\rvert^2 + \lvert\mathcal{M}\rvert\lvert\mathcal{N}\rvert \log \lvert\mathcal{N}\rvert)$
Algorithm 6 (Greedy)	$\mathcal{O}(\lvert\mathcal{M}\rvert\lvert\mathcal{N}\rvert + \lvert\mathcal{M}\rvert\lvert\mathcal{N}\rvert \log \lvert\mathcal{M}\rvert\lvert\mathcal{N}\rvert)$

of increasing the number of active MTs on the average values of profit function, user satisfaction, percentage of the blocked (unserved) data rate, number of HOs, processing time, and the satisfaction fairness index for different data rate classes.

3.7.1 Simulation parameters

The simulation environment consists of two overlapping LTE BSs and two Wi-Fi APs within the SA (dashed area in Figure 2.3).

Each MT is assumed to handle only one session, and randomly selects one of the data rates presented in Table 2.2. Some of the simulation parameters are shown in Table 3.2.

TABLE 3.2
Simulation parameters

Parameter	Wi-Fi AP	LTE BS
Path loss	$38.2 + 30\log_{10}(d_{mn})$	$34 + 40\log_{10}(d_{mn})$
β_n (W)	0.113286	1.28804
α_n (W/kbps)	0.13701×10^{-3}	0.05197×10^{-3}
Noise (dBm)	-90	-111.45
R_n (m)	200	400

The number of available RBs at each BS is $C_n = 10$ and the transmission power per RB is $P_n^{RB} = 26$ dBm. The bandwidth of one RB is $B_n^{RB} = 180$ KHz. Concerning Wi-Fi APs, a total bandwidth $B_n = 4000$ kHz is considered for each AP, with a total transmission power of 23 dBm. A scheduling interval of 1 second is considered in the simulations [BB15]. In LTE BSs, $T_n = 1000$ so that the duration of one time slot is 1 millisecond which is the duration of one transmission time interval (TTI) in the LTE standard. For Wi-Fi APs, the scheduling interval is also 1 second and it is divided into 10000 time slots.

MTs are randomly distributed within the SA. The simulation is repeated for 1000 iterations in a Monte Carlo manner. In each iteration, the number of MTs increases from 1 to 100, the location of Wi-Fi APs and MTs changes randomly within the SA, and the initial seed of the random number generator also changes.

MTs are assumed to be stable and always active in order to focus on the effect of the algorithms on the number of HOs.

3.7.2 Evaluation metrics

The following metrics are used to evaluate the proposed solutions: average profit values, average satisfaction, percentage of the blocked (unserved) data rate, number of HOs, processing time, and the satisfaction fairness index for different data rate classes. The satisfaction of MT m when associated with network n is:

$$\rho_{mn} = \frac{f_{mn}}{f_{mn'}} \tag{3.9}$$

where n' is the index of the network for which MT m achieves the highest profit.

In fact, it is not straightforward to study the average value of an attribute in a scenario where MTs request different amounts of data rate. For example, the average profit per user, *i.e.* $\frac{\sum_{m\in\mathcal{M}}\sum_{n\in\mathcal{N}} f_{mn}x_{mn}}{M}$, could be increased through increasing the profit of MTs with low data rate requirements on the expense of other MTs. Thus, to avert deceptive results, the average profit per requested data rate is studied according to the following formula:

$$\frac{\sum_{m\in\mathcal{M}}\sum_{n\in\Lambda_m} f_{mn}x_{mn}}{\sum_{m\in\mathcal{M}} Q_m} \tag{3.10}$$

Similarly, the average satisfaction per requested data rate is studied according to the following formula:

$$\frac{\sum_{m\in\mathcal{M}}\sum_{n\in\Lambda_m} \rho_{mn}Q_m x_{mn}}{\sum_{m\in\mathcal{M}} Q_m} \tag{3.11}$$

Since ρ_{mn} represents a normalized value, it is multiplied by Q_m in the above formula.

Moreover, we are interested in studying whether the compared algorithms tend to intentionally satisfy UEs with specific data rate requirement at the expense of other UEs. The average satisfaction fairness among different data rates classes is studied based on the Jain's fairness index [JCH84]. Basically, it is intended to study whether MTs of a specific data rate class are satisfied more (or less) than others, or in other words, have an average satisfaction that is far from the average satisfaction values of other data rate classes. Accordingly, the average satisfaction for each data rate class is calculated. Let Ψ denote the set containing all the requested data rates at a certain moment and q_l the set of all MTs whose requested data rate is l, *i.e* $q_l = \{m \in \mathcal{M} : Q_m = l\}$ where $l \in \Psi$. Then, the average satisfaction for all MTs with data rate requirement equal to l is $\mu_l = \frac{\sum_{m\in q_l}\sum_{n\in\mathcal{N}} \rho_{mn}x_{mn}}{|q_l|}$. Hence, the Jain's fairness index is:

$$\frac{(\sum_{l\in\Psi}\mu_l)^2}{|\Psi|\cdot\sum_{l\in\Psi}\mu_l^2} \tag{3.12}$$

The Jain's index is closer to 1 when the values of μ_l are less dispersed.

3.7.3 Simulation results

As we have mentioned before, we will study the performance of the profit-function-based solution (Algorithm 1), the optimal solution based on the branch

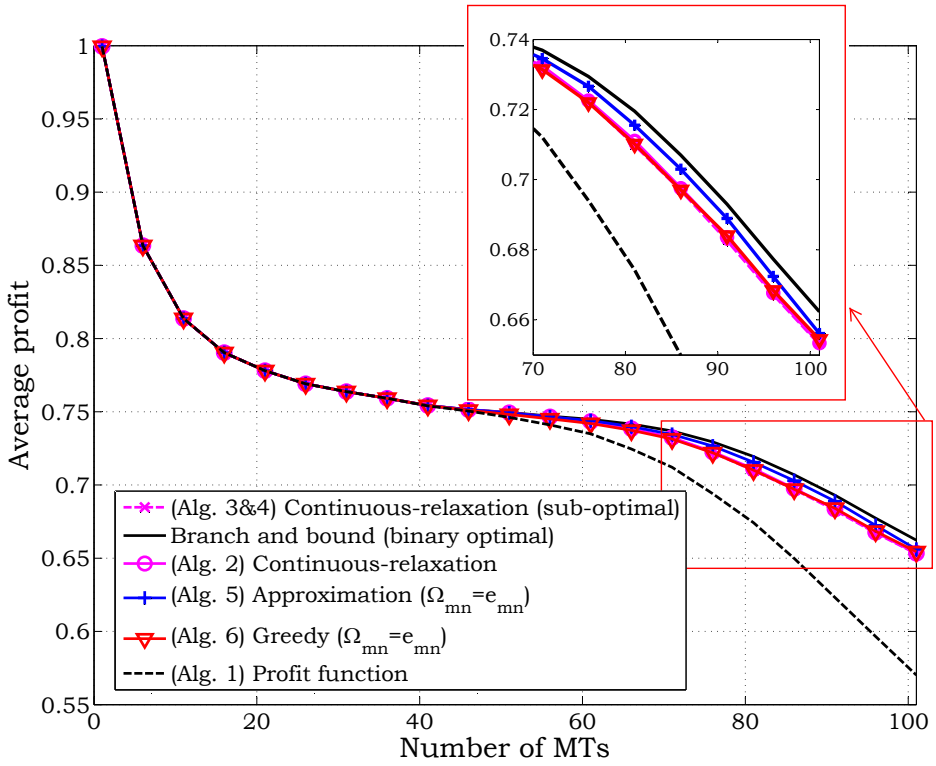


FIGURE 3.1 – Average profit per requested data rate (according to Eq. (3.10)).

and bound algorithm (Section 2.7.3), the optimal continuous-relaxation-based solution (Algorithm 2), the sub-optimal continuous-relaxation-based solution (Algorithms 3 and 4-Section 3.2.3), the approximation-based solution (Algorithm 5), and the greedy solution (Algorithm 6). It is indispensable to note that for the greedy and approximation-based solutions, the efficiency is considered as a desirability factor, *i.e.* $\Omega_{mn} = e_{mn}$, except for Section 3.7.3.4 where the difference between using the efficiency and normalized profit is discussed.

3.7.3.1 General behavior of algorithms

In order to increase the overall profit, the optimization problem **P1** finds the best set of association values for all MTs. Thus, the main performance of the optimization problem and its different solutions could be studied through the profit, and consequently through the satisfaction because it is directly related to the profit. Figure 3.1 shows that the average profit decreases fast as the number of MTs in the system increases from 1 to 10. This decrease is due to the fact that the profit function normalizes attributes by dividing them with the global maximum, *i.e.* $\max_{\forall m,n}$, which will only lead to a profit value of 1 when a single MT exists in the system. However, this behavior does not impact the user satisfaction. In fact, Figure 3.2 shows that the average satisfaction keeps maintaining an optimal value of 1 until the number of MTs reaches 40. An average satisfaction of 1 indicates that MTs are associated to their most preferred RATs. The main performance of

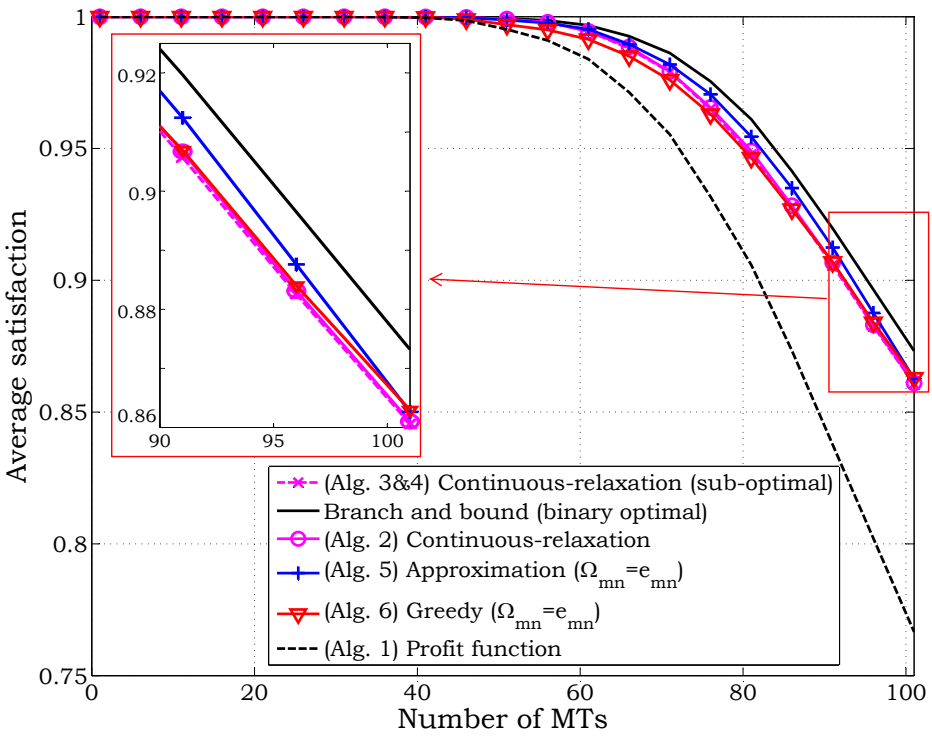


FIGURE 3.2 – Average satisfaction per requested data rate (according to Eq. (3.11)).

the algorithms could be studied when the number of MTs increases beyond 40 causing the networks to become more congested. The average satisfaction starts declining and the difference in both profit and satisfaction between the studied algorithms appears. Well, this aspect is a result of two factors: first, increasing the number of MTs strengthens the competition between MTs to acquire the limited resources of networks. Therefore, MTs have lower chances to be associated with their preferred RATs. The second reason is related to the percentage of blocked data rate. The increase in the percentage of blocked data rate (will be discussed shortly) decreases the average profit and satisfaction values.

It is shown in Figure 3.1 and Figure 3.2 that the proposed approximation-based solution maintains performance near the optimal solution. The greedy solution, optimal continuous-based-relaxation solution, and the sub-optimal continuous-based-relaxation solution performs approximately similarly in terms of average profit and satisfaction. A very close look at the results can show that the optimal continuous-relaxation-based solution performs slightly better than the sub-optimal continuous-relaxation-based solution. Therefore, the proposed approximation-based solution efficiently approximates the optimal solution, and could overwhelm the continuous-relaxation-based solutions (which have higher complexities).

It is very clear that all the simulated solutions perform much better than the trivial profit-function-based solution. For example, as the number of MTs reaches 100, the average satisfaction scored by the profit-function-based solution is 0.77,

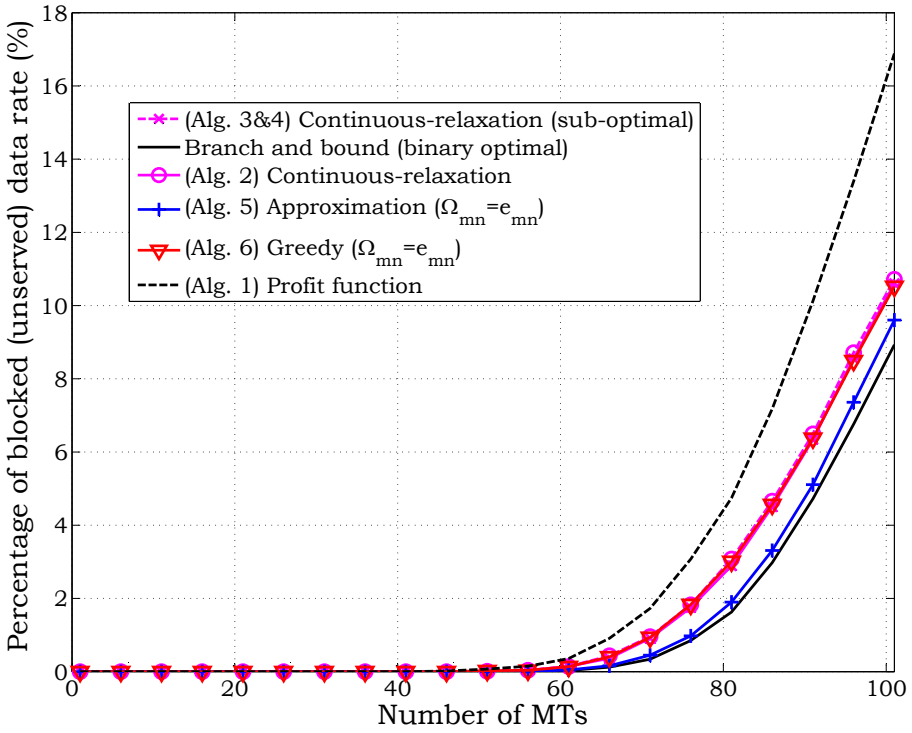


FIGURE 3.3 – Percentage of the blocked data rate.

while other solutions scores higher than 0.86, *i.e.* at least 12% enhancement.

It is important to illustrate that the slight difference in performance, shown in Figure 3.1 and Figure 3.2, between the two continuous-relaxation-based solutions indicates that the continuous-relaxation-based sub-optimal solution proposed in Section 3.2.3, which have polynomial-time complexity, efficiently approximates the optimal continuous-relaxation-based solution proposed in Algorithm 2.

3.7.3.2 Blocking percentage evaluation

In order to fully understand the behavior of the proposed solutions, the percentage of blocked data rate should be studied. It is shown in Figure 3.3 that the proposed approximation-based solution maintains lower blocking percentage after the optimal solution based on the branch and bound algorithm. Followed by the greedy and the optimal continuous-relaxation based solutions which score similar results. The sub-optimal continuous-relaxation based solution performs slightly less than the optimal continuous-relaxation based solution. Of course, the trivial profit-function-based solution scores the worst blocking percentage because it is designed to satisfy individual user needs without considering a system-wide user association and resource allocation scheme. For instance, the profit-function-based solution suffers from 17% blockage when the number of MTs reaches 100. The greedy solution and both continuous-relaxation based solutions lower down this percentage approximately 11, followed by the approximation-based solution

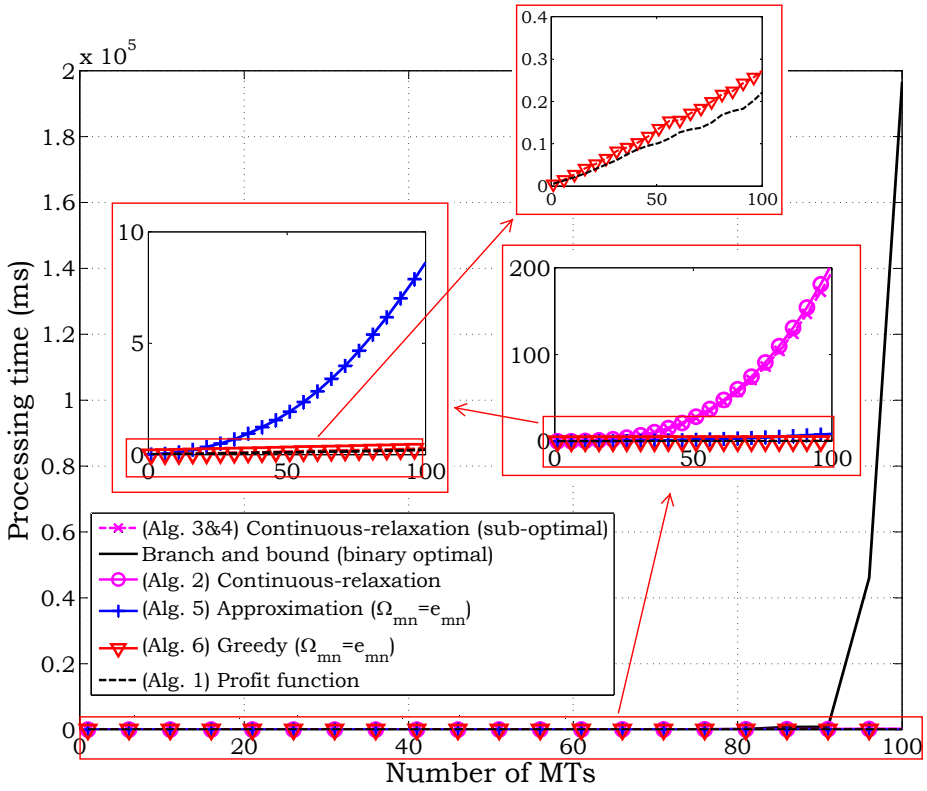


FIGURE 3.4 – Processing time of a system-wide resource and network assignment decision.

which scores 9.5%, while the optimal solution based on the branch and bound algorithm scores about 9%. Therefore, as can be seen in Figure 3.3, the proposed approximation-based solution achieves and maintains the lowest blocking percentage among the tested approaches, except for the optimal one of course.

Considering the efficiency e_{mn} as a desirability factor (abstract gain Ω_{mn}) plays a vital role in decreasing the data rate blocking percentage upon using the greedy and the approximation-based solutions as will be discussed shortly. Moreover, the proposed approximation-based solution achieves low blocking percentage for prioritizing those MTs with a single available network, *i.e.* $|F_m| = 1$ in Algorithm 5, among other MTs. In addition, considering the difference between the highest and the second highest available desirability value (line 5.22 in Algorithm 5) in the approximation-based solution, and maintaining a semi-global view on the system would also contribute in decreasing the blocking percentage in the approximation-based solution.

3.7.3.3 Complexity-performance trade-off

First, concerning the continuous-relaxation-based solutions, simulation results show that relaxing the binary constraint causes degradation in the performance when compared to the optimal solution. This is illustrated in Figures 3.1, 3.2, and 3.3 where it is obvious that there is a huge gap in performance between the

optimal binary solution (based on branch and bound) and the optimal continuous-relaxation-based solution. The approximation-based solution, which has a complexity lower than that of solving the relaxed problem **P2**, could perform better than the continuous-relaxation-based solutions. Even the simple greedy heuristic solution, which has the lowest complexity in this chapter, performs approximately the same as the continuous-relaxation-based solutions. Although the approximation-based solution requires higher complexity than the greedy one, but it has shown more robustness mainly in terms of the data rate blockage and average profit and satisfaction.

To study the effect of the algorithm's complexity on the processing time, the processing time of a system-wide resource and network assignment decision is shown in Figure 3.4. As we have stated in the previous chapter, the processing time of the optimal binary solution based on the branch and bound algorithm increases vastly as the number of MTs increases in the system. For example, when the number of MTs increases form 80 to 100, the decision processing time increases from about 500 millisecond to 198 seconds. Therefore, the branch and bound algorithm could not be used for sure in making resource and network assignment decisions due to its scalability problem. On the other hand, the sub-optimal continuous-relaxation-based solution requires a processing time of 200 milliseconds when the number of MTs reaches 100. This processing time is remarkable reduced to 8 milliseconds and 0.28 milliseconds in case of approximation-based solution and greedy solution respectively.

3.7.3.4 Efficiency factor verses normalized profit

The difference in satisfaction between using the normalized profit, *i.e.* $\frac{f_{mn}}{Q_m}$, and the efficiency in the approximation-based and greedy solutions is shown in Figure 3.5. In fact, associating a MT to the network with the highest efficiency does not guarantee the highest profit. Instead, it guarantees the highest profit per single allocated bandwidth unit, *i.e.* 1 Hz. Therefore, it is normal to notice in Figure 3.5 that using the normalized profit instead of the efficiency in the approximation-based solution contributes to obtain higher average satisfaction when the number of MTs increases from 40 to 70. However, as the number of MTs increases, adopting the efficiency as desirability factor contributes higher profit as shown in Figure 3.5 when the number of MTs increases beyond 70 because it is essential to consider for the requested bandwidth upon congestion. Actually, the effect of using the efficiency in the approximation-based solution is enlarged because it is considered twice in Algorithm 5 where the difference between the highest and the second highest efficiency is used to take a decision, while it is considered only once in the greedy algorithm (when the maximum gain value is chosen). Therefore, it is normal to notice that using the efficiency factor as a gain criteria in the greedy algorithm maintains better performance than the normalized profit at all times.

Using the efficiency value instead of the normalized profit also lowers the blocking percentage for both the greedy and approximation-based solutions as shown in Figure 3.6. In fact, the blocking percentage is remarkably enhanced upon using the efficiency factor in case of the greedy algorithm. For example, as the number of MTs reaches 100, adopting the efficiency factor in the greedy algorithm reduces the blocking percentage form 12 to 10. Hence, the efficiency is chosen as

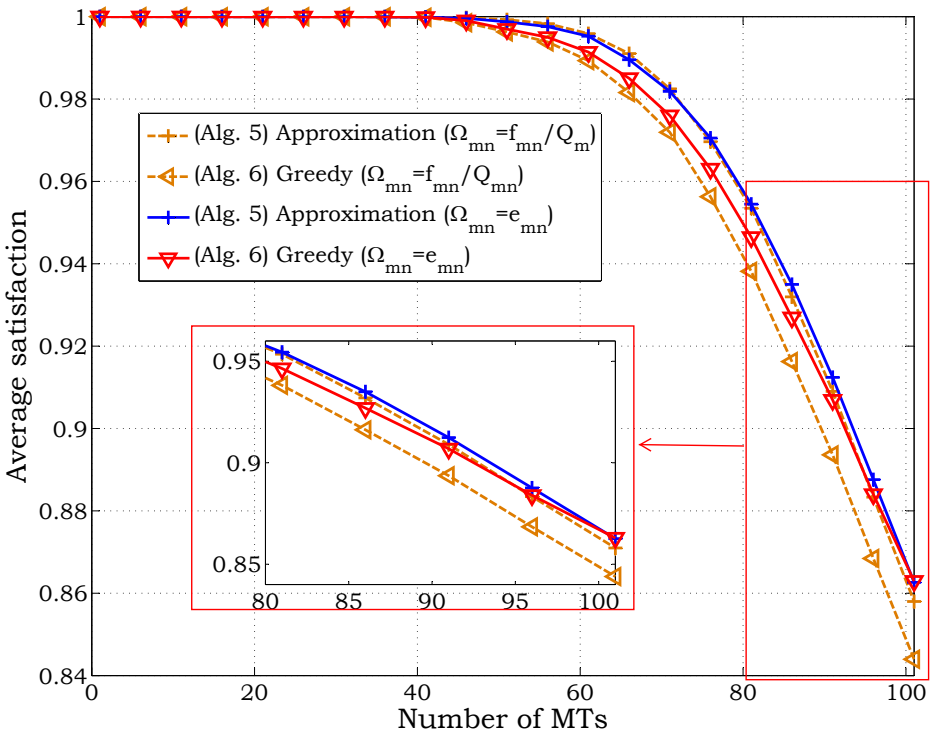


FIGURE 3.5 – Average satisfaction per MT. Comparing the difference between using the efficiency and normalized profit as desirability factor.

the main desirability criteria (abstract gain Ω_{mn}) for both the greedy solution and the approximation-based solution.

3.7.3.5 Effect of the solution on the number of handovers

Concerning the effect of algorithms on the number of HOs, Figure 3.7 shows that the optimal binary solution increases the number of HOs tragically to maintain its optimality. Since it tries to find the global optimal set of binary association values, the optimal solution re-associates a large number of MTs. Conversely, the greedy algorithm, characterized by the search for local optima, requires a much lower number of HOs. The number of HOs at a certain moment depends on the previous and the current association results. Therefore, upon adding new MTs, the number of requested HOs does not vary broadly. Normally, the number of HOs for the profit-function-based solution is zero because MTs are assumed to be stable and always active. So, upon entering the system, the new MT associates with the best available network and does not consider handing over unless a network with higher profit is available, *i.e.* has sufficient resources to serve the MT, which is impossible to happen as the number of MTs is increasing and the networks are becoming more congested.

As we have mentioned before, the optimal binary solution re-associates a large number of MTs to cope with the hard binary association constraint ($x_{mn} \in \{0, 1\}$).

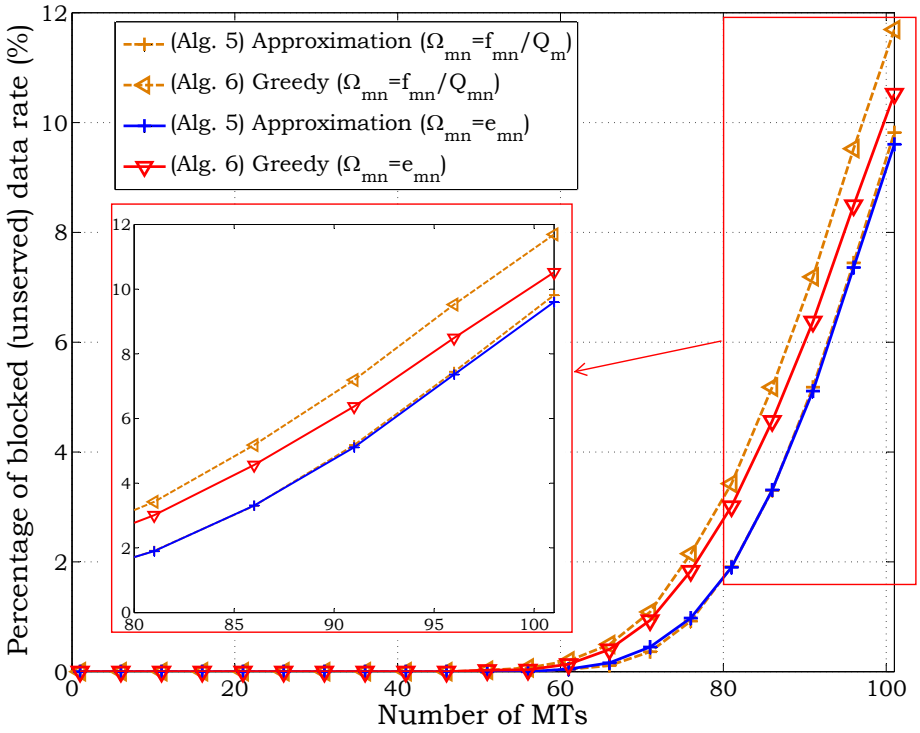


FIGURE 3.6 – Percentage of the blocked data rate. Comparing the difference between using the efficiency and normalized profit as desirability factor.

On the other hand, the relaxed continuous constraint ($x_{mn} \in [0, 1]$) in problem **P2** adds some flexibility on the selected set of association values, which can interpret the lower number of handover in the continuous-relaxation-based solutions. Concerning the proposed sub-optimal continuous-relaxation-based solution, Algorithms 3 and 4 finds the local maxima, *i.e.* the highest fractional association value and the highest normalized profit, while Algorithm 2 searches for the globally optimal set of association values in each iteration. Therefore, the sub-optimal continuous-relaxation-based solution requires slightly lower number of handovers than the optimal continuous-relaxation-based solution.

Regarding the proposed approximation-based solution, since this approximation is based on the highest and second highest gain (Ω_{mn}), and the algorithm behind this approximation aims at maintaining a semi-global view on the system, then the approximation-based solution requires higher number of HOs than the greedy solution. On the other hand, the approximation-based solution requires a slightly higher number of HOs than the continuous-relaxation-based solutions because the latter deals with relaxed continuous constraint, while the former deals with binary hard constraint and aims at maintaining a semi-global view on the system.

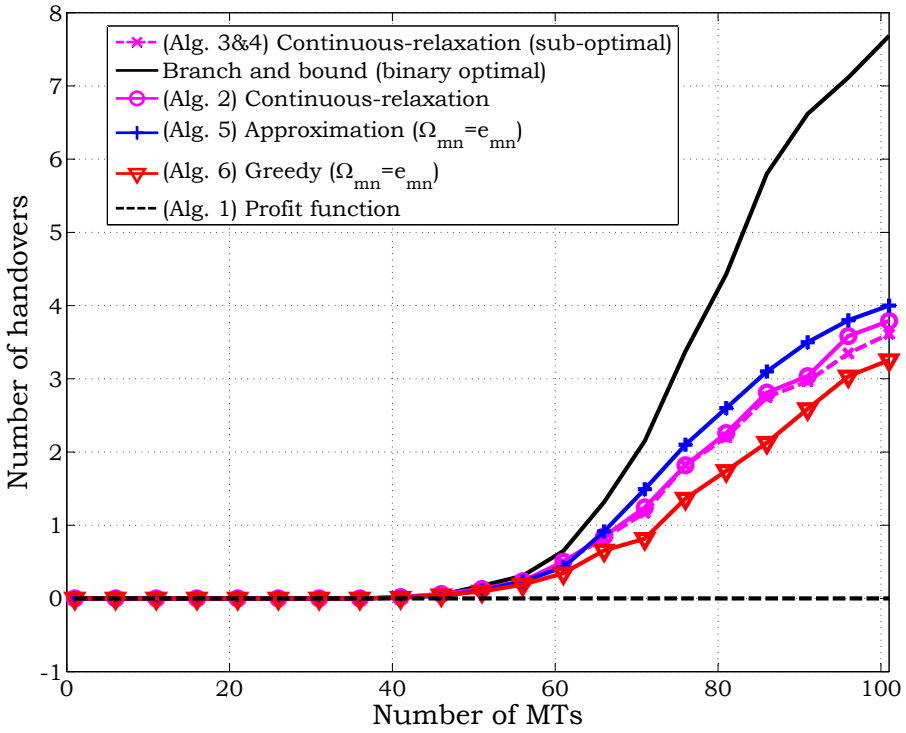


FIGURE 3.7 – Number of handovers.

3.7.3.6 Fairness in satisfaction among different data rate classes

Figure 3.8 studies the fairness in satisfaction among different data rate classes. The goal is to study whether the proposed algorithms tend to intentionally satisfy MTs with specific data rate requirement at the expense of other MTs. Basically, it is intended to study whether MTs of a specific data rate class are satisfied more (or less) than others, or in other words, have an average satisfaction that is far from the average satisfaction values of other data rate classes.

Figure 3.8 shows that using the profit function to evaluate the gain in the greedy algorithm, *i.e.* $\Omega_{mn} = f_{mn}$, leads to an unfair satisfaction between different data rate classes because MTs with high data rate requirements are satisfied more than other MTs as we have discussed earlier in Section 3.3.3.

For the rest of the algorithms, as the number of MTs increases beyond 50, and the networks become more congested, the Jain's fairness index starts deviating slightly away from its ideal value, *i.e.* 1. For instance, the Jain's index of the trivial profit-function-based solution is 0.98 when the number of MTs reaches 100. This deviation (2%) is so small and negligible to be studied for this normalized index. It is normal to observe that the Jain's index starts deviating upon congestion because MTs with high data rate requirements will have lower probability to be connected.

The main conclusion drawn out from Figure 3.8 is that all the proposed solutions (including the efficiency-based and normalized-profit-based greedy and ap-

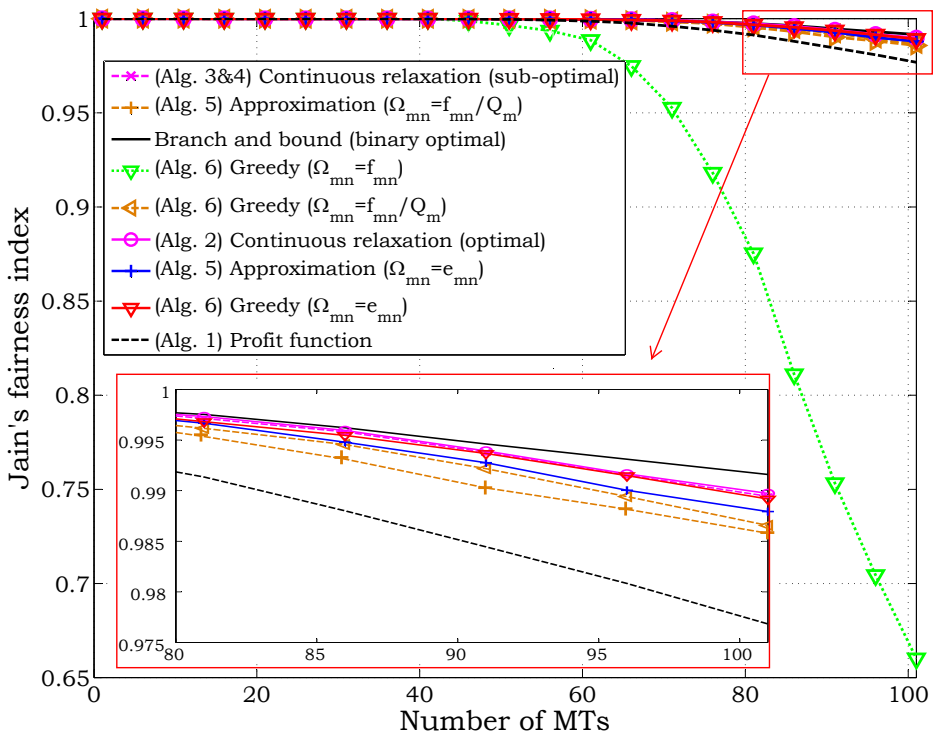


FIGURE 3.8 – Jain's fairness index among different data rate classes (according to Eq. (3.12)).

proximation solutions) do not tend to intentionally satisfy MTs of a certain data rate class to increase the overall satisfaction. Instead, these solutions effectively allocate resources and associate users to maximize the overall user satisfaction in the system. Therefore, the viewed and discussed simulation results are not misleading.

3.8 Conclusion

In this chapter, we have discussed and proposed different centralized solutions for the user association and resource allocation problem in heterogeneous wireless systems. First, we have proposed two new solutions based on the continuous-relaxation of problem **P1**. The first proposed solution is with undetermined complexity, and is considered as the optimal solution based on the continuous-relaxation methodology. The second solution is with determined polynomial-time complexity, and is considered as a sub-optimal solution based on the continuous-relaxation approach. Then, a novel approximation-based solution is proposed to approximate the binary problem **P1**. In addition, a new simple greedy heuristic algorithm is also proposed. The performance of the approximation-based solution and greedy solution is optimized through proposing a new efficiency factor to estimate the gain contributed upon associating users to networks. The efficiency factor considers the data rate requirement of users, and the channel conditions between the MT and

the BS or AP.

Simulation results show that the proposed sub-optimal solution based on the continuous-relaxation of problem **P1** efficiently approximates the optimal solution based on the continuous-relaxation. However, it is shown also that the continuous-relaxation approach threatens the optimality of the solutions. This was obvious in the performance gap between the optimal binary solution and the optimal continuous-relaxation-based solutions. On the other hand, the new approximation-based solution efficiently approximates the optimal binary solution by maintaining the closest performance to the optimal solution. The simpler greedy solution also shows acceptable performance similar to the solutions based on the continuous-relaxation approach, but with much lower complexity. The proposed efficiency factor (e_{mn}) has also contributed to a significant boost in the performance of the greedy and approximation-based solutions. Generally, the approximation-based solution demonstrates a remarkable trade-off between the complexity and performance.

All the solutions proposed in this chapter depends on a centralized entity, the SDN controller, to process the algorithms, allocate resources, and associate users. The following chapter will propose distributed solutions of the binary problem **P1**.

Chapter 4

Novel Distributed Resource and Network Assignment Solution

In this chapter, we propose two novel distributed user association and resource allocation solutions for HWNs. Mainly, the problem **P1** formulated in Chapter 2 will be distributed into $|\mathcal{N}|$ Knapsack problems, where $|\mathcal{N}|$ represents the number of networks and each problem is independently solved by its corresponding network. The Knapsack problem is solved using dynamic programming procedures in an acceptable time. The first distributed solution is based on the Lagrangian-relaxation of problem **P1**, and on the sub-gradient method that finds near-optimal Lagrangian multipliers. On the other hand, the second distributed solution is similar to the first one but without the sub-gradient method. The advantage of the second solution is that it eliminates the iterative sub-gradient phase, which leads to a lower complexity, and permits studying the effect of the sub-gradient method efficiently.

4.1 Simplified problem

We will recall in this section the optimization problem formulated in Chapter 2, and discuss some of its properties. The formulated problem is:

$$\textbf{P1: } max \sum_{m\in\mathcal{M}}\sum_{n\in\mathcal{N}} f_{mn}x_{mn} \tag{4.1a}$$

$$s.\,t. \sum_{m\in\mathcal{M}} \left\lceil \frac{Q_m T_n}{r^{tot}_{mn}} \right\rceil x_{mn} \leqslant T_n \qquad \forall n \in \mathcal{N}_{\mathcal{AP}} \tag{4.1b}$$

$$\sum_{m\in\mathcal{M}} \left\lceil \frac{Q_m T_n}{B^{RB}_n \log_2(1+\gamma_{mn})} \right\rceil x_{mn} \leqslant U_n \qquad \forall n \in \mathcal{N}_{\mathcal{BS}} \tag{4.1c}$$

$$\sum_{n\in\mathcal{N}} x_{mn} \leqslant 1 \qquad \forall m \in \mathcal{M} \tag{4.1d}$$

$$x_{mn} \in \{0,1\} \qquad \forall m \in \mathcal{M}, \forall n \in \mathcal{N} \tag{4.1e}$$

As we have mentioned before, problem **P1** is a combinatorial optimization problem with an NP-complete complexity. In fact, some combinatorial optimization problems are classified as easy problems that could be solved in a time bounded by

polynomial dependent on the number of variables. On the other hand, the majority of combinatorial optimization problems are classified as hard problems for which their solution require exponential time. Even for the hard problems, there exist easier hard problems, like the Knapsack problem, that have pseudo-polynomial algorithms which solve the problem in polynomial time only if specific constraints of the problem are bounded.

One of the most brilliant mathematical methodologies emerged in the 1970s is the observation that several hard optimization problems could be transformed into easier problems that are subject to a relatively small set of side constraints. Dualizing these side constraints leads to a Lagrangian problem whose optimal solution is considered as an upper bound (for maximization optimization problems) to the optimal solution of the original non-relaxed problem. To solve the problem efficiently, the Lagrangian problem should be easier than the original optimization problem.

In the previous chapter, we have used the continuous-relaxation (*i.e.* linear-programming-relaxation) method to relax the binary constraint of problem **P1**. In this chapter, the Lagrangian-relaxation method is used to simplify problem **P1** and distribute it into multiple easier-to-solve problems. As we will notice in this chapter, the Lagrangian-relaxation method offers major advantages over the linear-programming-relaxation method.

4.2 Mathematical discussion

In this section, we will provide a pure mathematical discussion about the methodology that will be used to simplify problem **P1**. In Section 3.2.1 of the previous chapter, we have presented the canonical form of problem **P1**. In this section, we extend the presented canonical problem by assuming that the constraints of the canonical problem are divided into two sets $\boldsymbol{Ax} \leq \boldsymbol{b}$ and $\boldsymbol{Dx} \leq \boldsymbol{e}$ in order to make it easy to solve the Lagrangian problem. Therefore, the canonical form of problem **P1** could be expressed as:

$$\mathbf{P1}: \; Z = \; max \; \boldsymbol{c}^\top \boldsymbol{x} \tag{4.2a}$$
$$s.\,t. \; \boldsymbol{Ax} \leq \boldsymbol{b} \tag{4.2b}$$
$$\boldsymbol{Dx} \leq \boldsymbol{e} \tag{4.2c}$$
$$\boldsymbol{x} \in \{0,1\} \tag{4.2d}$$

By applying the Lagrangian-relaxation method to the set of constraints $\boldsymbol{Ax} \leq \boldsymbol{b}$ the optimization problem becomes:

$$\mathbf{LR}: \; Z_\lambda = \; max \; \boldsymbol{c}^\top \boldsymbol{x} - \boldsymbol{\lambda}(\boldsymbol{Ax} - \boldsymbol{b}) \tag{4.3a}$$
$$s.\,t. \; \boldsymbol{Dx} \leq \boldsymbol{e} \tag{4.3b}$$
$$\boldsymbol{x} \in \{0,1\} \tag{4.3c}$$

where $\boldsymbol{\lambda}$ is a vector of positive Lagrangian multipliers. Of course, problem **LR** should be easier to solve than problem **P1**. For convenience, we assume the following assumptions:

– *Assumption 1*: Problem **P1** is feasible.

– *Assumption 2*: The set $\mathcal{X} = \{x | Dx \leq e, x \in \{0,1\}\}$ containing the solutions for problem **LR** is finite.

Proof of Assumption 1: Notice that constraint (4.1d) considers that upon congestion not all MTs could be connected. Therefore, there exist at least one feasible solution such that $x_{mn} = 0 \ \forall m \in \mathcal{M}, \ \forall n \in \mathcal{N}$.

Proof of Assumption 2: For any given $\boldsymbol{\lambda}$, problem **LR** is a combinatorial optimization problem (because x is binary). Hence, there exist a finite number of solutions for problem **LR**. The overall number of combinations that are considered for all variables x_{mn} is $|\mathcal{N}|^{|\mathcal{M}|}$. Thus, the set $\mathcal{X} = \{x | Dx \leq e, x \in \{0,1\}\}$ is finite.

Therefore, based on these two assumptions, Z_λ is finite for all λ. It is well known that $Z_\lambda \geq Z$. This is easy to show by assuming an optimal solution $\boldsymbol{x}^*$ to problem **P1** and observing that:

$$Z_\lambda \geq \boldsymbol{c}\boldsymbol{x}^* - \boldsymbol{\lambda}(\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{b}) \geq Z \tag{4.4}$$

The inequality in Eq. (4.4) follows from the relations $Z = \boldsymbol{c}\boldsymbol{x}^*$ and $\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b} \leq 0$. Then, we require the Lagrangian multiplier $\lambda \geq 0$ for $Z_\lambda \geq Z$ to hold.

Since the Lagrangian-relaxation aims at finding an upper bound for problem **P1**, then we should find the best values of λ that produces the tightest upper bound to the problem. Hence, it is clear that the best choice for λ would be an optimal solution to the dual problem:

$$\mathbf{DP}: \ Z_{dual} = \min_{\lambda} \ Z_\lambda \tag{4.5a}$$

$$s.\ t.\ \lambda \geq 0 \tag{4.5b}$$

The dual problem **DP**, *i.e.* $\min_{\lambda \geq 0} Z_\lambda$, is shown to be convex and piecewise linear [Fis81]. Thus, problem **DP** could be solved using the sub-gradient method [Fis81]. The sub-gradient method iteratively solves the Lagrangian-relaxed problem and updates the Lagrangian multipliers in an easy systematic way. As BLP is NP-complete and suffers from strictly positive duality gap [JK07], the solutions of the relaxed problem, which are considered as an upper bound to the original problem, are rarely feasible, *i.e.* some MTs will be connected to multiple networks at the same time, or the capacity constraints of some networks are violated. Therefore, the sub-gradient optimization procedure is used as a basis for a heuristic method that maintains the feasibility of the problem in every iteration. Hence, a feasible near-optimal solution to problem **P1** is produced at the end of the sub-gradient iterations.

4.3 Distributed solution

In this section, we apply the procedures discussed in the previous section on the optimization problem **P1**. In order to distribute problem **P1** into $|\mathcal{N}|$ problems, each processed by a network, we choose to relax the assignment constraint (4.1d).

Hence, the relaxed problem is:

$$\mathbf{REL}_\lambda\textbf{:}\ max \quad \sum_{m\in\mathcal{M}}\sum_{n\in\mathcal{N}} f_{mn}x_{mn} - \sum_{m\in\mathcal{M}}\lambda_m(\sum_{n\in\mathcal{N}} x_{mn} - 1) \tag{4.6a}$$

$$s.\ t. \quad x_{mn} \in \{0,1\} \qquad \forall m \in \mathcal{M}, \forall n \in \mathcal{N} \tag{4.6b}$$

$$\sum_{m\in\mathcal{M}} \left\lceil \frac{Q_m T_n}{r_{mn}^{tot}} \right\rceil x_{mn} \leqslant T_n \qquad \forall n \in \mathcal{N}_{\mathcal{AP}} \tag{4.6c}$$

$$\sum_{m\in\mathcal{M}} \left\lceil \frac{Q_m T_n}{B_n^{RB}\log_2(1+\gamma_{mn})} \right\rceil x_{mn} \leqslant U_n \qquad \forall n \in \mathcal{N}_{\mathcal{BS}} \tag{4.6d}$$

After directly manipulating Eq. (4.6a), the problem becomes:

$$\mathbf{REL}_\lambda\textbf{:}\ max \quad \sum_{m\in\mathcal{M}}\sum_{n\in\mathcal{N}} (f_{mn}-\lambda_m)x_{mn} + \sum_{m\in\mathcal{M}}\lambda_m \tag{4.7a}$$

$$s.\ t. \quad x_{mn} \in \{0,1\} \qquad \forall m \in \mathcal{M}, \forall n \in \mathcal{N} \tag{4.7b}$$

$$\sum_{m\in\mathcal{M}} \left\lceil \frac{Q_m T_n}{r_{mn}^{tot}} \right\rceil x_{mn} \leqslant T_n \qquad \forall n \in \mathcal{N}_{\mathcal{AP}} \tag{4.7c}$$

$$\sum_{m\in\mathcal{M}} \left\lceil \frac{Q_m T_n}{B_n^{RB}\log_2(1+\gamma_{mn})} \right\rceil x_{mn} \leqslant U_n \qquad \forall n \in \mathcal{N}_{\mathcal{BS}} \tag{4.7d}$$

Note that for a given set of λ, the expression $\sum_{m\in\mathcal{M}}\lambda_m$ is constant. Therefore, removing $\sum_{m\in\mathcal{M}}\lambda_m$ does not affect the optimality of the solution. Hence, $\mathbf{REL}_\lambda$ separates into $|\mathcal{N}|$ Knapsack problems. For each $n \in \mathcal{N}_{\mathcal{AP}}$, the Knapsack problem is:

$$\mathbf{KNAP}_{\mathcal{AP}}\textbf{:}\ max \quad \sum_{m\in\mathcal{M}} (f_{mn}-\lambda_m)x_{mn} \tag{4.8a}$$

$$s.\ t. \quad x_{mn} \in \{0,1\} \qquad \forall m \in \mathcal{M} \tag{4.8b}$$

$$\sum_{m\in\mathcal{M}} \left\lceil \frac{Q_m T_n}{r_{mn}^{tot}} \right\rceil x_{mn} \leqslant T_n \tag{4.8c}$$

For each $n \in \mathcal{N}_{\mathcal{BS}}$, the Knapsack problem is:

$$\mathbf{KNAP}_{\mathcal{BS}}\textbf{:}\ max \quad \sum_{m\in\mathcal{M}} (f_{mn}-\lambda_m)x_{mn} \tag{4.9a}$$

$$s.\ t. \quad x_{mn} \in \{0,1\} \qquad \forall m \in \mathcal{M} \tag{4.9b}$$

$$\sum_{m\in\mathcal{M}} \left\lceil \frac{Q_m T_n}{B_n^{RB}\log_2(1+\gamma_{mn})} \right\rceil x_{mn} \leqslant U_n \tag{4.9c}$$

Notice that problem $\mathbf{REL}_\lambda$ aims at maximizing the profit of MTs in all networks, while each Knapsack problem $\mathbf{KNAP}_{\mathcal{AP}}$ and $\mathbf{KNAP}_{\mathcal{BS}}$ aims at maximizing the profit in a single network (AP or BS respectively). Each Knapsack problem ($\mathbf{KNAP}_{\mathcal{AP}}$ or $\mathbf{KNAP}_{\mathcal{BS}}$) is processed independently by its corresponding network n.

4.3.1 Multiplier values

To get the best values of the Lagrangian multipliers, the following dual problem is formulated:

$$\begin{aligned} min \quad & \mathbf{REL}_\lambda \\ s.t. \quad & \lambda \in \mathbb{R}_+^{|\mathcal{M}|} \end{aligned} \tag{4.10}$$

The dual problem is solved using the sub-gradient method presented in Algorithm 7.

Algorithm 7: Sub-gradient Method

7.1: **Given** λ;
7.2: **Set** $lb := -\infty$, $ub := +\infty$;
7.3: **while** ***Not*** *stopping_condition* **do**
7.4: | **Solve** relaxation ($\mathbf{REL}_\lambda$), which is distributed to $|N|$ Knapsack problems each solved using Algorithm 8, obtaining x^λ and $v(\mathbf{REL}_\lambda) = \sum_{m\in\mathcal{M}} \sum_{n\in\mathcal{N}} (f_{mn} - \lambda_m) x^\lambda_{mn}$;
7.5: | Obtain a feasible solution x^f and $v_f = \sum_{\forall m\in\mathcal{M}} \sum_{\forall n\in\mathcal{N}} f_{mn} x^f_{mn}$ by applying Algorithm 9 using x^λ;
7.6: | $lb = max[lb, v_f]$;
7.7: | $ub = min[ub, v(\mathbf{REL}_\lambda)]$;
7.8: | **Update** the sub-gradient direction g^λ_m, the step size θ_m, and the multiplier λ;
7.9: | **Make** stopping tests;
7.10: **end**
7.11: $x_{mn} = x^f_{mn}\ \forall m \in \mathcal{M},\ \forall n \in \mathcal{N}$;

The sub-gradient method is an iterative process that aims at finding an optimal, or near-optimal, Lagrangian multipliers. The Lagrangian multipliers are updated in each iteration. The idea is to increase the lower bound (the feasible solution) and decrease the upper bound of the problem while maintaining the feasibility of the solution. In this way, we ensure that the feasible solution is close to the upper bound, and in consequence, close to the optimal solution which is less than or equal to the upper bound (Eq. (4.4)). In order to limit the number of iterations of the sub-gradient method, stopping conditions are used.

First, we start by an initial value for each λ (line 7.1), and initialize the upper and lower bounds (line 7.2). Then, while the stopping conditions are not met, the sub-gradient algorithm iteratively solves ($\mathbf{REL}_\lambda$) finding its association values x^λ, and the overall profit of ($\mathbf{REL}_\lambda$) which is $v(\mathbf{REL}_\lambda) = \sum_{m\in\mathcal{M}} \sum_{n\in\mathcal{N}} (f_{mn} - \lambda_m) x^\lambda_{mn}$ (line 7.4). The solution x^λ of ($\mathbf{REL}_\lambda$) is based on solving the distributed Knapsack problems. Each Knapsack problem is solved independently, and aims at choosing the MTs that will maximize the profit of each network. The solution of the Knapsack problem is detailed in Section 4.3.2. However, the solution of ($\mathbf{REL}_\lambda$) is rarely assignment-feasible because some networks will choose the same MT, which will produce MTs associated to multiple networks at the same time. Therefore, based on the solution of ($\mathbf{REL}_\lambda$), an assignment-feasible solution x^f is developed using a heuristic algorithm described in Section 4.3.3. The feasible solution value $v_f = \sum_{\forall m\in\mathcal{M}} \sum_{\forall n\in\mathcal{N}} f_{mn} x^f_{mn}$ serves as a lower bound for the problem. In each

iteration, the Lagrangian multipliers are updated based on a step size and the sub-gradient direction which will be discussed shortly.

The initial λ used is $\lambda_m = 1\ \forall m \in \mathcal{M}$. Since we have relaxed constraint (4.1d), the sub-gradient direction is:

$$g_m^\lambda = \sum_{n \in \mathcal{N}} x_{mn} - 1 \tag{4.11}$$

The used step size is:

$$\theta_m = \begin{cases} 0, & \text{if } g_m^\lambda = 0, \\ \beta(ub - lb)/||g^\lambda||^2, & \text{otherwise}, \end{cases} \tag{4.12}$$

where β is the Held and Karp control parameter [HK71] defined such that $0 \leq \beta \leq 2$. In each iteration, λ is updated according to the following formula:

$$\lambda_m = \lambda_m + \theta_m g_m^\lambda \tag{4.13}$$

Notice that $\mathbf{REL}_\lambda$ and the Knapsack problems aim at maximizing the value $(f_{mn} - \lambda_m)$. Hence, increasing the value of λ_m makes the association of MT m less appealing, and vice versa. Therefore, if MT m is associated to multiple networks at the same time, the sub-gradient direction g_m^λ takes a positive value, leading to an increase in the value of λ_m which makes the association of MT m less appealing. On the other hand, if MT m is not associated to any network, g_m^λ takes a negative value, and λ_m decreases making the association of MT m more appealing, which is required. If MT m is associated to only one network, λ_m remains unchanged because it is required to associate the MT to only one network.

The initial value of β is two. If after 20 iterations the upper bound ub does not decrease, β is updated to $\beta/2$. In order to limit the number of iterations of the sub-gradient method, the stopping tests used are:

- number of iterations greater than 600,
- $\beta \leq 0.005$,
- $ub - lb < 1$.

The most effective stopping test is $ub - lb < 1$ because it indicates that the difference between the upper and the lower bound is relatively small. In fact, it is the only test that terminates the sub-gradient iterations in the conducted simulations. Finally, the feasible solutions x^f obtained in the last iteration of Algorithm 7 are used for associating MTs, *i.e.* $x_{mn} = x_{mn}^f\ \forall m \in \mathcal{M},\ \forall n \in \mathcal{N}$.

4.3.2 Knapsack problem solution

The Knapsack problem aims at maximizing the total profit subject to the capacity constraint of the network and the boolean association constraint. In fact, the simplified problem **P1**, the relaxed problem $\mathbf{REL}_\lambda$, and the Knapsack problems ($\mathbf{KNAP}_{\mathcal{AP}}$ and $\mathbf{KNAP}_{\mathcal{BS}}$) hold an important property that the number of requested resources and the capacity of the network are strictly positive integers. Hence, we can define an integer weight $i \leq \zeta_n$. We wish to find the maximum profit that could be attained in a network while considering the capacity constraint. Therefore, we also define a two-dimensional array $K[m][i]$ to denote the maximum

profit that can be attained with total number of requested resources less than or equal to i using MTs up to m, *i.e.* MTs from 1 to m. In fact, we can define $K[m][i]$ recursively as follows:

$$\begin{cases} K[0][i] = 0 \text{ and } K[m][0] = 0, & \text{(4.14a)} \\ K[m][i] = K[m-1][i], & \text{if } \omega_{mn} > i, \text{ (4.14b)} \\ K[m][i] = max(K[m-1][i], K[m-1][i-\omega_{mn}] + f_{mn} - \lambda_m), & \text{if } \omega_{mn} \leq i. \text{ (4.14c)} \end{cases}$$

To better understand the recursive function (4.14), $K[m][i]$ could be seen as the overall profit of up to m MTs, *i.e.* MTs 1 to m, while having i available resources. Therefore, $K[0][i] = 0$ represents an initial condition where the contributed profit is zero when no MT is selected. If MT m is not selected, then the profit of MT m is not considered in the overall profit, *i.e.* $K[m][i] = K[m-1][i]$ as in Eq. (4.14b) where MT m is not selected because the number of resources requested by MT m is greater than i. On the other hand, if the number of resources requested by MT m is less than or equal to i, the profit of MT m is not immediately added to the overall profit. Instead, $K[m-1][i-\omega_{mn}] + f_{mn} - \lambda_m$ is compared to $K[m-1][i]$, and the highest value is chosen. Note that $f_{mn} - \lambda_m$ is added to $K[m-1][i-\omega_{mn}]$ because MT m requests ω_{mn} resources. Eq. (4.14c) is seen as the core of the recursive function (4.14). The solution can then be found by calculating $K[m][i]$ $\forall m \in \mathcal{M}$ and $\forall i \leq \zeta_n$ as presented in Algorithm 8 and discussed later in this section. Algorithm 8 shows the dynamic programming procedures used to solve the Knapsack problem for network n.

Algorithm 8 is made up of two main phases. In phase 1, the recursive function (4.14) is implemented and used to fill array K. In *lines 8.6 and 8.7*, $K[m][0] = 0$ because the value $i = 0$ indicates that there are no resources which means the overall profit is zero because no MT could be selected. Phase 2 presents the association solution based on the values stored in array K. In *line 8.20*, the association value is initialized to zero. As we have mentioned before, we can determine whether MT m is selected by comparing the values $K[m][i]$ and $K[m-1][i]$. Both combined phases represent the dynamic programming procedures used to solve the Knapsack problem.

4.3.3 Feasible solution based on the Knapsack problem results

As we have mentioned in Section 4.3.1, the solutions of the Knapsack problems are rarely assignment-feasible because some MTs will be chosen by multiple networks simultaneously. Therefore, in this section, we propose a heuristic algorithm to find assignment-feasible solution based on the results of the Knapsack problems. The feasible solution is presented in Algorithm 9.

Algorithm 9 is mainly divided into two parts, each part is based on two phases (preparation phase and decision phase). In part 1 (*lines 9.1-9.21*), the feasible association results of MTs selected by the dynamic programming procedures, *i.e.* Algorithm 8, are decided. While part 2 (*lines 9.22-9.40*) decides the feasible solutions for the rest of MTs. Both parts have similar structures with minor differences. In preparation phase 1, *i.e.* the preparation phase of part 1, the feasible association values of all MTs are set to zero (*line 9.7*), the set $\mathcal{E}$ is filled with all the profit

Algorithm 8: Dynamic programming procedures

```
8.1: // Phase 1: filling table K
8.2: double K[ ][ ] := double[|M| + 1][ζ_n + 1];
8.3: int i;
8.4: for m = 0 to |M| do
8.5:     for i = 0 to ζ_n do
8.6:         if m = 0 or i = 0 then
8.7:             K[m][i] = 0;
8.8:         else
8.9:             if ω_mn ≤ i then
8.10:                K[m][i] = max(f_mn − λ_m + K[m−1][i−ω_mn], K[m−1][i]);
8.11:            else
8.12:                K[m][i] = K[m − 1][i];
8.13:            end
8.14:        end
8.15:    end
8.16: end
8.17: // Phase 2: association solution based on the values stored
      in array K
8.18: i = ζ_n;
8.19: for m = |M| to 1 do
8.20:    x^λ_mn := 0;
8.21:    if K[m][i] ≠ K[m − 1][i] then
8.22:        x^λ_mn = 1;
8.23:        i = i − ω_mn;
8.24:    end
8.25:    m = m − 1;
8.26: end
```

values f_{mn} whose corresponding $x^{\lambda}_{mn} = 1$, and the number of free resources in each network ($\overline{\zeta}_n$) is initialized. In decision phase 1, the algorithm aims at associating the MT from set $\mathcal{E}$ having the highest profit in each iteration within the while loop (*lines 9.14-9.21*). After detecting the highest profit in set $\mathcal{E}$, the algorithm, in *line 9.16*, tests if the corresponding MT, *i.e.* $m\prime$, is not associated yet ($\sum_{n\in\mathcal{N}} x^{f}_{m\prime n} = 0$), and if the number of free resources in the target network $n\prime$ is sufficient to serve MT $m\prime$ ($\omega_{m\prime n\prime} \leq \overline{\zeta}_{n\prime}$). If both conditions are true, MT $m\prime$ is associated to network $n\prime$ (*line 9.17*), and the number of free resources in network $n\prime$ is updated by subtracting the number of resources allocated to MT $m\prime$ (*line 9.18*).

In part 2, the same procedures of part 1 are repeated. However, this time, the set $\mathcal{E}\prime$ is filled with profit values of MTs whose profit is not equal to zero. Moreover, the feasible association decision for MTs in $\mathcal{E}\prime$ should not taken yet ($\sum_{n\in\mathcal{N}} x^{f}_{mn} = 0$), and their number of requested resources should be less than the number of free resources in the network ($\omega_{mn} \leq \overline{\zeta}_n$).

Algorithm 9: Feasible solution

9.1: // Part 1:
9.2: // Preparation phase 1:
9.3: $\mathcal{E} := \phi$; // $\mathcal{E}$ represents a set that will be filled with all gain values (f_{mn}) whose corresponding $x^{\lambda}_{mn} \neq 0$
9.4: **foreach** $n \in \mathcal{N}$ **do**
9.5: $\quad \overline{\zeta}_n := \zeta_n$;
9.6: $\quad$ **foreach** $m \in \mathcal{M}$ **do**
9.7: $\quad\quad x^{f}_{mn} = 0$;
9.8: $\quad\quad$ **if** $x^{\lambda}_{mn} \neq 0$ **then**
9.9: $\quad\quad\quad \mathcal{E} = \mathcal{E} + \{f_{mn}\}$;
9.10: $\quad\quad$ **end**
9.11: $\quad$ **end**
9.12: **end**
9.13: // Decision phase 1:
9.14: **while** $\mathcal{E} \neq \phi$ **do**
9.15: $\quad m\prime = argmax_m(f_{mn} \in \mathcal{E})$; $n\prime = argmax_n(f_{mn} \in \mathcal{E})$;
9.16: $\quad$ **if** $\sum_{n \in \mathcal{N}} x^{f}_{m\prime n} = 0$ **and** $\omega_{m\prime n\prime} \leq \overline{\zeta}_{n\prime}$ **then**
9.17: $\quad\quad x^{f}_{m\prime n\prime} = 1$;
9.18: $\quad\quad \overline{\zeta}_{n\prime} = \overline{\zeta}_{n\prime} - \omega_{m\prime n\prime}$;
9.19: $\quad$ **end**
9.20: $\quad \mathcal{E} = \mathcal{E} - \{f_{m\prime n\prime}\}$;
9.21: **end**
9.22: // Part 2:
9.23: // Preparation phase 2:
9.24: $\mathcal{E}\prime := \phi$; // $\mathcal{E}$ represents a set that will be filled with all non-zero gain values (f_{mn})
9.25: **foreach** $n \in \mathcal{N}$ **do**
9.26: $\quad$ **foreach** $m \in \mathcal{M}$ **do**
9.27: $\quad\quad$ **if** $f_{mn} \neq 0$ **and** $\sum_{n \in \mathcal{N}} x^{f}_{mn} = 0$ **and** $\omega_{mn} \leq \overline{\zeta}_n$ **then**
9.28: $\quad\quad\quad \mathcal{E}\prime = \mathcal{E}\prime + \{f_{mn}\}$;
9.29: $\quad\quad$ **end**
9.30: $\quad$ **end**
9.31: **end**
9.32: // Decision phase 2:
9.33: **while** $\mathcal{E}\prime \neq \phi$ **do**
9.34: $\quad m\prime = argmax_m(f_{mn} \in \mathcal{E}\prime)$; $n\prime = argmax_n(f_{mn} \in \mathcal{E}\prime)$;
9.35: $\quad$ **if** $\sum_{n \in \mathcal{N}} x^{f}_{m\prime n} = 0$ **and** $\omega_{m\prime n\prime} \leq \overline{\zeta}_{n\prime}$ **then**
9.36: $\quad\quad x^{f}_{m\prime n\prime} = 1$;
9.37: $\quad\quad \overline{\zeta}_{n\prime} = \overline{\zeta}_{n\prime} - \omega_{m\prime n\prime}$;
9.38: $\quad$ **end**
9.39: $\quad \mathcal{E}\prime = \mathcal{E}\prime - \{f_{m\prime n\prime}\}$;
9.40: **end**

4.3.4 Novel distributed solution without the sub-gradient method

In this section, we propose a novel distributed resource and network assignment solution that is not based on the sub-gradient method. The objective is to provide a distributed solution with lower complexity than the solution provided in Algorithm 7, and to effectively study the performance of the sub-gradient method.

Since the solution provided in Algorithm 7 aims at iteratively solving the Knapsack problems to find near-optimal Lagrangian multipliers, in this section we try to directly provide a distributed solution without having to find the Lagrangian multipliers. Therefore, all the Lagrangian multipliers are set to zero, *i.e.* $\lambda_m = 0 \ \forall m \in \mathcal{M}$, where in this case λ is denoted by λ_0. Accordingly, the solution provided in this section is presented in Algorithm 10. As you can see, Algorithm

Algorithm 10: Distributed solution without the sub-gradient method

10.1: **Solve** relaxation ($\mathbf{REL}_{\lambda_0}$), which is distributed to $|N|$ Knapsack problems each solved using Algorithm 8 while setting $\lambda_m = 0 \ \forall m \in \mathcal{M}$, obtaining x^{λ_0};

10.2: Obtain a feasible solution x^f by applying Algorithm 9 using x^{λ_0} instead of x^{λ};

10.3: $x_{mn} = x^f_{mn} \ \forall m \in \mathcal{M}, \ \forall n \in \mathcal{N}$;

10 is also based on Algorithm 8 and Algorithm 9 which allows us to effectively study the impact of the sub-gradient method.

4.4 Complexity analysis

Concerning the complexity of Algorithm 8, it is based on two phases. The complexity of phase 1 is $\mathcal{O}(|\mathcal{M}|.\zeta_n)$. The complexity of phase 2 is $\mathcal{O}(|\mathcal{M}|)$. Therefore, the complexity of Algorithm 8 is $\mathcal{O}(|\mathcal{M}|.\zeta_n + |\mathcal{M}|)$ because the two phases are processed sequentially. Since the dynamic programming procedures are processed by each network n independently and at the same time, then the total complexity of the dynamic programming procedures on all networks is $\mathcal{O}(|\mathcal{M}|.\zeta_{n^*} + |\mathcal{M}|)$, where n^* represents the network with the highest number of resources.

The complexity of Algorithm 9 is based on two parts, each having two phases. The complexity of the preparation phase 1 is $\mathcal{O}(|\mathcal{M}||\mathcal{N}|)$. Concerning the complexity of the decision phase 1, let μ represent the number of variables x^f_{mn} that are not in set $\mathcal{E}$, then the complexity of decision phase 1 is $\mathcal{O}(|\mathcal{M}||\mathcal{N}| - \mu)$. Moreover, the complexity of preparation phase 2 is $\mathcal{O}(|\mathcal{M}||\mathcal{N}|)$. Concerning the complexity of decision phase 2, the maximum number of items in $\mathcal{E}\prime$ is μ. Therefore the maximum complexity of decision phase 2 is $\mathcal{O}(\mu)$. Since the phases of Algorithm 9 runs sequentially, the total complexity of Algorithm 9 is $\mathcal{O}(|\mathcal{M}||\mathcal{N}| + |\mathcal{M}||\mathcal{N}| - \mu + |\mathcal{M}||\mathcal{N}| + \mu)$ which is $\mathcal{O}(3|\mathcal{M}||\mathcal{N}|)$.

Now, for Algorithm 7, let υ denote the number of sub-gradient iterations, where in each iteration Algorithms 8 and 9 are iterated, then the total complexity of Algorithm 7 is: $\mathcal{O}(\upsilon(|\mathcal{M}|.\zeta_{n^*} + |\mathcal{M}| + 3|\mathcal{M}||\mathcal{N}|))$. Since Algorithm 10 is basically

the same as Algorithm 7 without the sub-gradient procedures, the complexity of Algorithm 10 is $\mathcal{O}(|\mathcal{M}|.\zeta_{n^*} + |\mathcal{M}| + 3|\mathcal{M}||\mathcal{N}|)$.

The complexities of all solutions proposed in this thesis are presented in Table 4.1 where L is the length of the binary coding of the input data, υ the number of sub-gradient iterations, and ζ_{n^*} the number of resources in network n^* having the largest number of resources. Since the complexity of algorithms jointly depends on the values of $|\mathcal{M}|$, $|\mathcal{N}|$, L, and ζ_{n^*}, it is hard to exactly compare the complexity of algorithms. However, we can notice that solving the continuous problem **P2** does not scale well upon increasing the number of MTs and networks because its complexity is based on the cube of those values. Although the complexity of Algorithm 5 is based on the square of the number of MTs, we can say that its complexity is lower than that of Algorithm 7 which is basically determined according to the value $\upsilon.|\mathcal{M}|.\zeta_{n^*}$. Mainly, ζ_{n^*} is based on the number of sub-channels and the number of time slots in one scheduling interval. The scheduling interval is usually taken on one second basis, while the duration of time slot is set according to the TTI (one millisecond in the LTE standard), which produces 1000 time slot in one scheduling duration, so the number of resources is the number of sub-channels multiplied by 1000. Moreover, the number of sub-gradient iterations increases upon increasing the number of MTs (as we will see in the simulation results). Thus, we can say that usually the value $\upsilon.\zeta_{n^*} > |\mathcal{M}|$, and $\upsilon.|\mathcal{M}|.\zeta_{n^*} > |\mathcal{M}|^2$. Hence, the complexity of Algorithm 7 is higher than that of Algorithm 5. Similarly, we can directly notice that the complexity of Algorithm 10 is usually higher than Algorithm 5. Hence, the complexities could be sorted in the following order $\mathcal{O}$(solving continuous problem **P2**)>$\mathcal{O}$(Algorithm 7)>$\mathcal{O}$(Algorithm 10)>$\mathcal{O}$(Algorithm 5)>$\mathcal{O}$(Algorithm 6).

TABLE 4.1
Solution complexity for all algorithms

	Complexity
The binary problem **P1**	$\mathcal{O}(\lvert\mathcal{N}\rvert^{\lvert\mathcal{M}\rvert})$
Solving the continuous problem **P2**	$\mathcal{O}(\lvert\mathcal{M}\rvert^3\lvert\mathcal{N}\rvert^3 L)$
Algorithm 5 (Approximation)	$\mathcal{O}(\lvert\mathcal{M}\rvert^2 + \lvert\mathcal{M}\rvert\lvert\mathcal{N}\rvert \log \lvert\mathcal{N}\rvert)$
Algorithm 6 (Greedy)	$\mathcal{O}(\lvert\mathcal{M}\rvert\lvert\mathcal{N}\rvert + \lvert\mathcal{M}\rvert\lvert\mathcal{N}\rvert \log \lvert\mathcal{M}\rvert\lvert\mathcal{N}\rvert)$
Algorithm 7 (Distributed with sub-gradient)	$\mathcal{O}(\upsilon(\lvert\mathcal{M}\rvert.\zeta_{n^*} + \lvert\mathcal{M}\rvert + 3\lvert\mathcal{M}\rvert\lvert\mathcal{N}\rvert))$
Algorithm 10 (Distributed without sub-gradient)	$\mathcal{O}(\lvert\mathcal{M}\rvert.\zeta_{n^*} + \lvert\mathcal{M}\rvert + 3\lvert\mathcal{M}\rvert\lvert\mathcal{N}\rvert)$

4.5 Simulation results

In this chapter, we use the same simulation parameters used in the previous chapter in Section 3.7.1, and we study the same evaluation metrics presented in Section 3.7.2 in addition to the number of sub-gradient iterations. We compare the performance of Algorithms 7 and 10 to the solutions proposed in the previous chapters.

First, we start by studying the Jain's fairness index among different data rate classes. It is shown in Figure 4.1 that as the number of MTs increases to 100, the Jain's fairness index of Algorithms 7 and 10 only declines by 0.7% and 1%

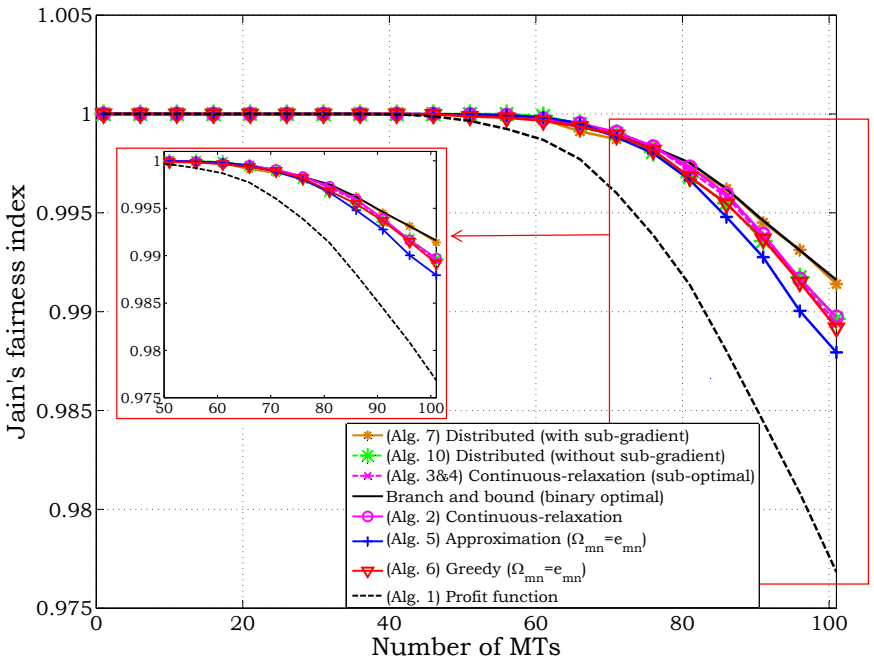


FIGURE 4.1 – Jain's fairness index among different data rate classes (according to Eq. (3.12)).

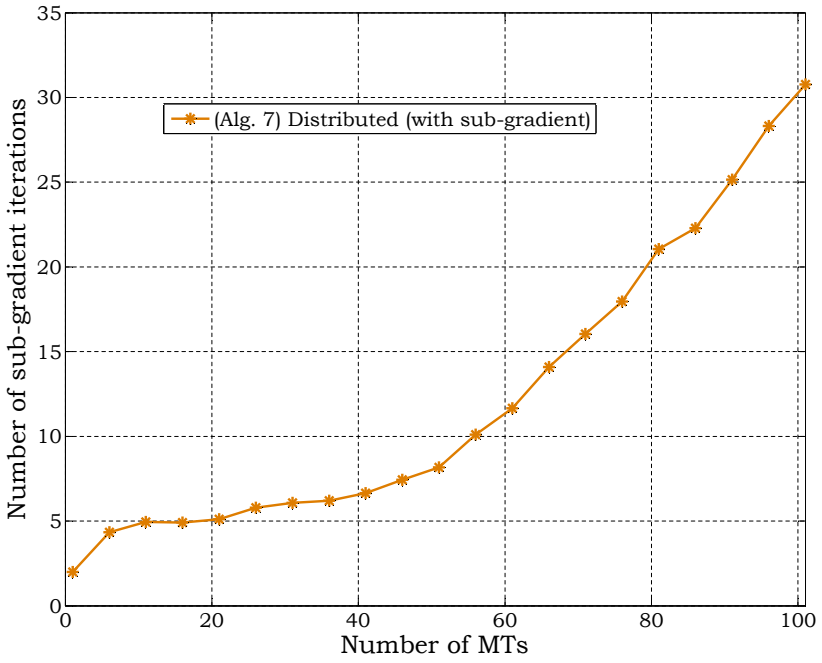


FIGURE 4.2 – Number of sub-gradient iterations.

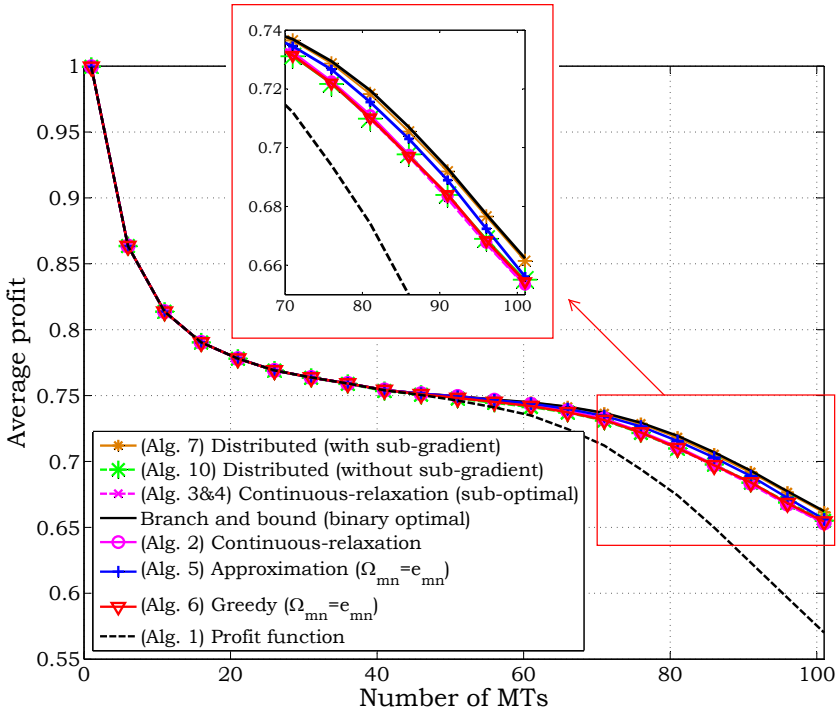


FIGURE 4.3 – Average profit per requested data rate (according to Eq. (3.10)).

respectively. Therefore, Algorithms 7 and 10 do not tend to intentionally satisfy MTs of a certain data rate class to increase the overall satisfaction. Instead, these solutions effectively allocate resources and associate users to maximize the overall user satisfaction in the system. Hence, the viewed and discussed simulation results are not misleading.

Concerning the distributed solution with the sub-gradient method, *i.e.* Algorithm 7, we have proposed in Section 4.3.1 different stopping criteria for the sub-gradient iterations. However, in the conducted simulations, the only test that terminates the sub-gradient iterations is $ub - lb < 1$, *i.e.* the difference between the upper bound (unfeasible solution) and the lower bound (feasible solution) is one, which is relatively small. It is shown in Figure 4.2, that as the number of MTs in the system increases from 1 to 101, the required number of sub-gradient iterations also increases from 2 to 31. The increase in the number of sub-gradient iterations indicates that it becomes harder for Algorithm 7 to approximate the optimal solution when the number of variables increases.

Concerning the behavior of the proposed solutions, we start by studying the average profit per requested data rate. As you can see in Figure 4.3, the proposed distributed solution with the sub-gradient method (Algorithm 7) effectively approximates the optimal binary solution, and performs better than the other solutions proposed in previous chapters. The remarkable performance of Algorithm 7 is due to the Lagrangian-relaxation and the sub-gradient method that reduces the gap between the lower bound (feasible solution) and the upper bound (unfeasible) to approach the optimal solution. We can analyze the effect of the Lagrangian-relaxation and the sub-gradient method by comparing the performance

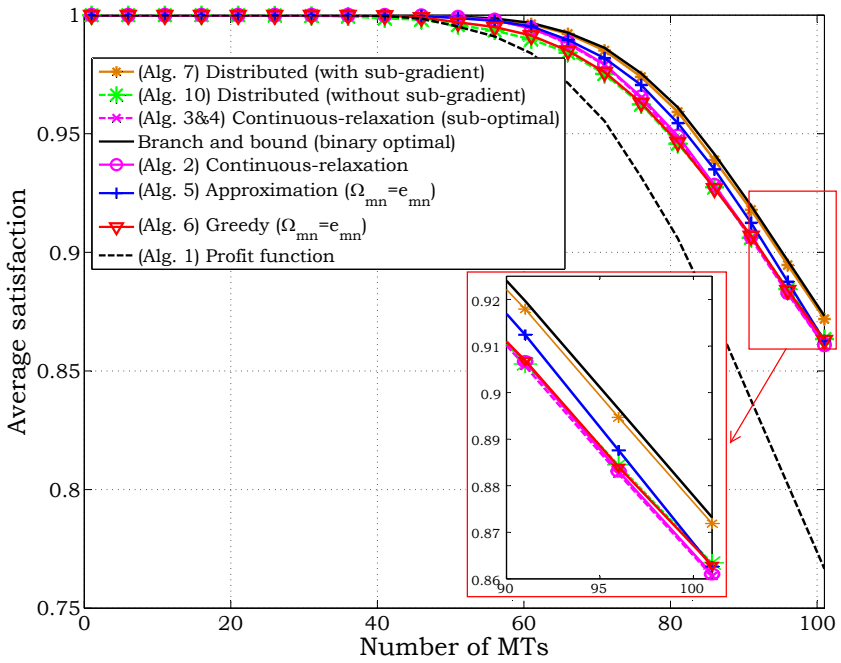


FIGURE 4.4 – Average satisfaction per requested data rate (according to Eq. (3.11)).

of Algorithm 7 and Algorithm 10, *i.e.* distributed solution with and without the sub-gradient method. It is obvious from Figure 4.3 that the average profit achieved by Algorithm 7 is always better than that achieved by Algorithm 10. On the other hand, the distributed solution without the sub-gradient method achieves approximately the same average profit as the centralized efficiency-based greedy solution. Therefore, as an advantage, Algorithm 10 could be used as a distributed solution with low complexity that achieves average profit similar to the centralized efficiency-based greedy solution.

As we have mentioned before in the previous chapter, the average satisfaction is directly related to the average profit. Therefore, the same conclusions drawn out from Figure 4.3 could be seen in Figure 4.4 which shows the average satisfaction results.

Concerning the percentage of blocked (unserved) data rates, it is shown in Figure 4.5 that the optimal solution with sub-gradient iterations achieves slightly higher blocking percentage than the optimal binary solution (based on the branch and bound algorithm) and better than the rest of proposed solutions. This confirms that Algorithm 7 efficiently approximates the optimal solution. Moreover, the distributed solution without the sub-gradient method, achieves blocking percentage approximately similar to the centralized greedy, continuous-relaxation-based, and the sub-optimal continuous-relaxation-based solutions.

On the other hand, the distributed solution is based on decisions taken independently by networks, *i.e.* dynamic programming results of Algorithm 8. Thus, the association results depend highly on the status of MTs in each network, ins-

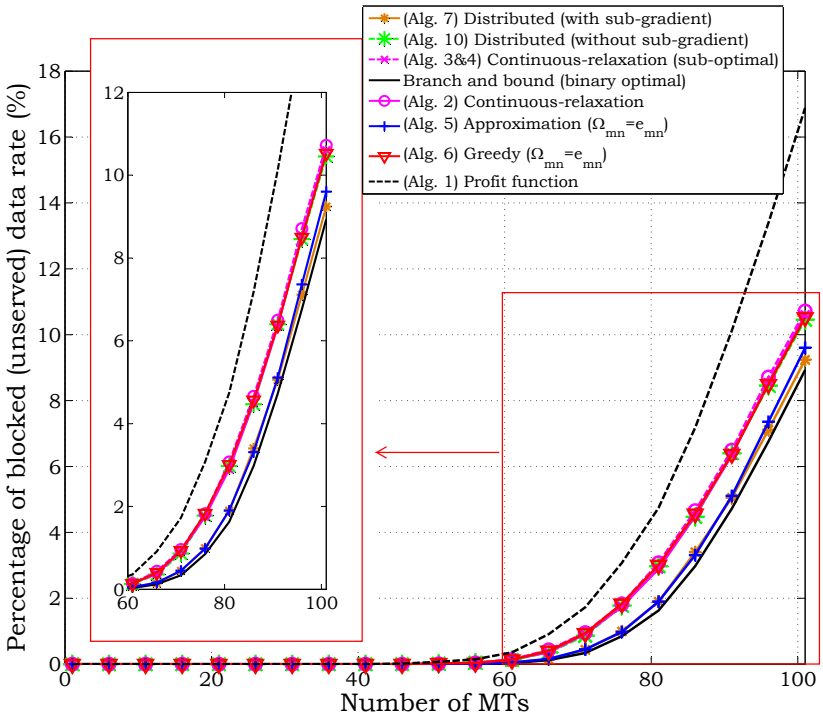


FIGURE 4.5 – Percentage of the blocked data rate.

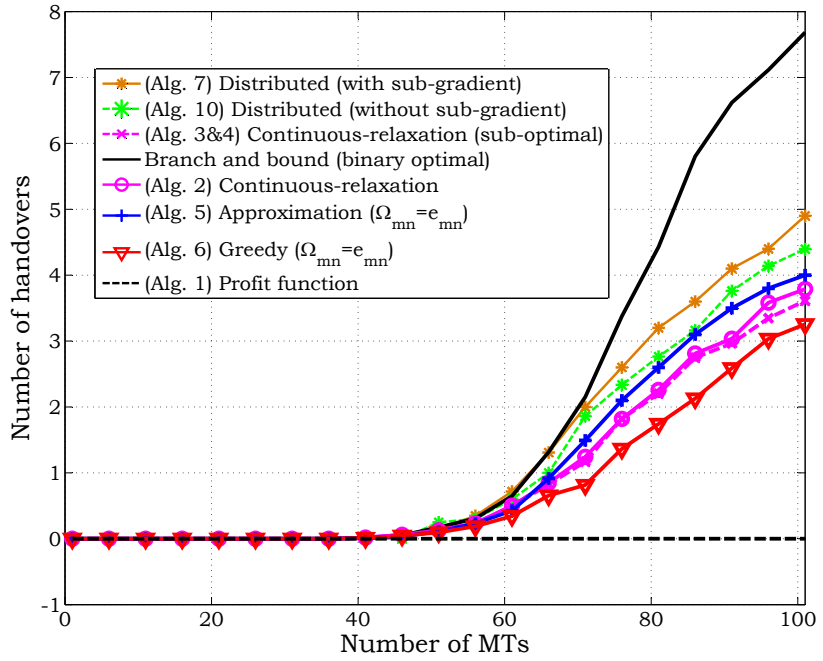


FIGURE 4.6 – Number of handovers.

tead of the status of MTs in the global system. Therefore, we can notice in Figure

4.6 that the distributed solutions require slightly higher number of HOs, than the proposed centralized solution (except for the optimal binary solution based on the branch and bound algorithm). The solution based on Algorithm 7 requires higher number of HOs than the solution in Algorithm 10 because the former iteratively solves the distributed dynamic programming procedures, while the later solves them only once.

4.6 Conclusion

In this chapter, we have proposed two novel distributed user association and resource allocation solutions for HWNs. Mainly, the problem **P1** formulated in Chapter 2 is distributed into N Knapsack problems, where each problem is independently solved by its corresponding network. The Knapsack problem is solved using dynamic programming procedures in an acceptable time. The first distributed solution is based on the Lagrangian-relaxation of problem **P1**, and on the sub-gradient method that finds near-optimal Lagrangian multipliers. On the other hand, the second distributed solution is similar to the first one but without the sub-gradient method. The advantage of the second solution is that it eliminates the iterative sub-gradient phase, which leads to a lower complexity. The complexity of the distributed solution with sub-gradient method is higher than that of the greedy and approximation-based solution, while lower than the continuous-relaxation-based solutions. However, the second proposed distributed solution without the sub-gradient method has a very competitive complexity. Simulation results show that the proposed distributed solution with the sub-gradient method achieves the best performance results in terms of blocking percentage and average profit and satisfaction among all other proposed solutions while requiring a slightly higher number of handovers. The remarkable advantage of the distributed solution without the sub-gradient method is that it achieves performance similar to the centralized efficiency-based greedy algorithm.

Chapter 5

New Applications For Priority-Based Management And Power Efficiency Maximization

In this chapter, we discuss new direct applications for the problem **P1** proposed in Chapter 2, and its corresponding solutions proposed in Chapters 3 and 4. First, we propose a new application for the priority-based user association and resource allocation problem in a system where users have different priorities. We formulate a new priority-based optimization problem, then we simplify it into a new problem with form similar to the problem **P1**, and therefore could be solved with the solutions proposed in this thesis. Moreover, we discuss new perspective for the user association and resource allocation problem to enhance the overall power efficiency (data rate per unit power) in the system. Therefore, the profit function used in problem **P1** is replaced by the power efficiency factor, and the problem is solved using the solutions proposed in this thesis.

5.1 A new priority-based resource and network assignment

It is common that users in communication systems have different priorities. For example, in mobile networks, users experiencing low long-term transmission rate, or users demanding high QoS, are given higher priority [CLL+16]. Moreover, future mobile networks should prioritize the service of emergency applications over the ordinary ones. Note that all previous papers cited in this thesis do not take into consideration different user priorities when making decisions.

In fact, the authors of [TK14] have proposed that upon congestion in public safety networks (PSNs) users handoff to LTE system. To ensure reliable service for those users, they are given higher priority among ordinary commercial users. In [YBC05], authors have introduced the concept of degraded utility to deal with different user priorities ; additional bandwidth is released to high priority users by degrading the low priority traffic. The authors of [CLL+16] have formulated an

optimization problem to associate users with different priorities in heterogeneous networks. However, the studies [CLL+16] and [YBC05] do not consider specific user-related parameters, and the adopted system model is not realistic. Moreover, both studies do not propose a firm mechanism to prevent low-priority users from allocating resources that could be utilized by users with high priority.

In this section, we discuss the user association and downlink resource allocation problem in HWNs. MTs in our context have different service priorities, or service levels (SLs), such that MTs with highest priority should experience the best service. In order to be served, each MT should be supplied with its requested data rate, otherwise, the MT terminates its ongoing session. Typically, users with high priority should encounter the minimal attainable blockage.

In that perspective, we first formulate a novel binary linear programming (BLP) problem that ensures lower blockage and better service for high-priority users. The formulated problem exploits different context information that could be user-centric (power consumption, signal quality, and preferences), service-centric (the amount of requested data rate), and network-centric (number of available resources, geographical location, transmission range, *etc.*). The formulated problem throws firm restrictions to prevent low-priority users from allocating resources that could be utilized by other users with higher priorities. Specifically, the algorithm aims at maximizing, for each SL, the user-centric gain which is based on the received signal quality and instantaneous power consumption at the MT.

In addition, a novel solution management strategy is proposed to minimize the number of times the optimization function is processed without affecting the optimality of the solution.

5.1.1 System model

The system model used for the priority-based assignment is the same as the system model presented in Section 2.5 but with minor changes to adapt users with different priorities. The set of all available SLs is denoted by $\mathcal{K} = \{1, 2, ..., K\}$. Since the priority of a MT at a given moment is determined according to the SL, a MT is assigned a single SL k at a given moment. The SL of MT m is denoted by l_m. For simplicity, higher SL indicates higher priority. We define a set θ_k containing all MTs with SL k such that $\theta_k = \{m \in \mathcal{M} : l_m = k\}$.

5.1.2 Optimization problem

We aim at formulating an optimization problem to maximize the total profit for each SL. The problem should throw firm restrictions to prevent low-priority users from allocating resources that could be utilized by other users with higher priorities. A single network association should be ensured, as well as supplying the connected MT with data rate that is at least equal to its requested data rate

threshold. Thus, the formulated problem is:

$$\textbf{PSL: } max \quad \sum_{k\in\mathcal{K}}\sum_{m\in\theta_k}\sum_{n\in\mathcal{N}} f_{mn}x_{mn} \tag{5.1a}$$

$$s.\ t. \quad \sum_{m\in\theta_k} u_{mn}x_{mn} \leqslant U_n - \sum_{j>k}\sum_{i\in\theta_j} u_{in}x_{in} \qquad \forall\ k\in\mathcal{K},\ n\in\mathcal{N}_{\mathcal{BS}} \tag{5.1b}$$

$$\sum_{m\in\theta_k} t_{mn}x_{mn} \leqslant T_n - \sum_{j>k}\sum_{i\in\theta_j} t_{in}x_{in} \qquad \forall\ k\in\mathcal{K},\ n\in\mathcal{N}_{\mathcal{AP}} \tag{5.1c}$$

$$\sum_{n\in\mathcal{N}} r_{mn}x_{mn} \geqslant \sum_{n\in\mathcal{N}} Q_m x_{mn} \qquad \forall m\in\mathcal{M} \tag{5.1d}$$

$$\sum_{n\in\mathcal{N}} x_{mn} \leqslant 1 \qquad \forall m\in\mathcal{M} \tag{5.1e}$$

$$x_{mn}\in\{0,1\} \qquad \forall m\in\mathcal{M},\forall n\in\mathcal{N} \tag{5.1f}$$

$$u_{mn}\in\mathbb{N}^+ \qquad \forall m\in\mathcal{M},\forall n\in\mathcal{N}_{\mathcal{BS}} \tag{5.1g}$$

$$t_{mn}\ \in\mathbb{N}^+ \qquad \forall m\in\mathcal{M},\forall n\in\mathcal{N}_{\mathcal{AP}} \tag{5.1h}$$

Constraint (5.1b) ensures that the capacity of LTE BSs is not exceeded and the resources allocated to high-priority MTs are not violated. Similarly, constraint (5.1c) guarantees the same aspects in Wi-Fi APs. Constraints (5.1e) and (5.1f) assure that a MT will be associated with a single network, or not connected at all (upon congestion). Constraint (5.1d) guarantees that the data rate received by a MT is at least equal to its requested data rate threshold. However, we are obliged to multiply both sides of the inequality by x_{mn} because upon congestion, some MTs will not be served. Constraint (5.1g) ensures that a single SB (LTE) is not assigned to multiple MTs simultaneously. Similarly, constraint (5.1h) guarantees that a single time slot in an AP is not allocated for multiple MTs at the same time. Note that MTs are distributed in different SL sets θ_k, and constraints (5.1b) and (5.1c) ensure that the resources allocated for MTs with SLs higher than k, *i.e.* MTs $\in\theta_j$ such that $j>k$, are not given to MTs with SL k, *i.e.* MTs $\in\theta_k$. Therefore, it is preferable to show the maximization form in terms of all SLs and all MTs in SL sets instead of directly maximizing for all MTs, *i.e.* "$max\ \sum_{k\in\mathcal{K}}\sum_{m\in\theta_k}\sum_{n\in\mathcal{N}}$" instead of "$max\ \sum_{m\in\mathcal{M}}\sum_{n\in\mathcal{N}}$". This plays a role in clarifying the characteristics of the formulated problem.

5.1.3 Problem simplification and solution

The formulated problem (**PSL**) aims at finding three sets of variables:
- The boolean association variables (x_{mn}).
- The number of SBs allocated for each MT m connected to BS n (u_{mn}).
- The number of time slots allocated for each MT m connected to AP n (t_{mn}).

In the following, the number of resources that should be allocated by each network in order to supply the MT with its requested data rate (if the MT is associated to the network) is calculated. Hence, u_{mn} or t_{mn} can be seen as the weight of MT m in network n. Thus, the optimization problem now aims at finding only the boolean association variables x_{mn}. Therefore, based on constraint (5.1d), we calculate the

number of resources that should be allocated by each network to supply MTs with their minimum requested data rate, *i.e.* $r_{mn} = Q_m \ \forall m \in \mathcal{M}$. Hence, based on Eq. (2.3), and according to the approach adopted by [BB15] and [SA14], the minimum number of resources that should be allocated to MT m if it is connected to BS n is:

$$u_{mn} = \frac{Q_m T_n}{B_n^{RB} \gamma_{mn}} \qquad \forall n \in \mathcal{N}_{\mathcal{BS}} \quad (5.2)$$

Similarly, and based on Eq. (2.4) for Wi-Fi APs:

$$t_{mn} = \frac{Q_m T_n}{r_{mn}^{tot}} \qquad \forall n \in \mathcal{N}_{\mathcal{AP}} \quad (5.3)$$

In fact, both (5.2) and (5.3) can be seen as the weights of MTs in networks. Thus, $\omega_{mn} \in \mathbb{N}^+$ is introduced to indicate the weight of MT m in network $n \in \mathcal{N}$ such that:

$$\omega_{mn} = \begin{cases} \left\lceil \frac{Q_m T_n}{B_n^{RB} \gamma_{mn}} \right\rceil & \forall n \in \mathcal{N}_{\mathcal{BS}} \\ \\ \left\lceil \frac{Q_m T_n}{r_{mn}^{tot}} \right\rceil & \forall n \in \mathcal{N}_{\mathcal{AP}} \end{cases} \quad (5.4)$$

The ceiling ($\lceil . \rceil$) of values in (5.2) and (5.3) is taken to preserve the integral constraints (5.1g) and (5.1h). Similarly, ζ_n denotes the capacity of network $n \in \mathcal{N}$ such that:

$$\zeta_n = \begin{cases} U_n & \forall n \in \mathcal{N}_{\mathcal{BS}} \\ T_n & \forall n \in \mathcal{N}_{\mathcal{AP}} \end{cases} \quad (5.5)$$

Therefore, based on (5.4) and (5.5), **PSL** could be reformulated as:

$$\textbf{PSL}_2\textbf{:} \ max \quad \sum_{k \in \mathcal{K}} \sum_{m \in \theta_k} \sum_{n \in \mathcal{N}} f_{mn} x_{mn} \quad (5.6a)$$

$$s.\ t. \quad \sum_{m \in \theta_k} \omega_{mn} x_{mn} \leqslant \zeta_n - \sum_{j>k} \sum_{i \in \theta_j} \omega_{in} x_{in} \qquad \forall\ k \in \mathcal{K},\ n \in \mathcal{N} \quad (5.6b)$$

$$\sum_{n \in \mathcal{N}} x_{mn} \leqslant 1 \qquad \forall m \in \mathcal{M} \quad (5.6c)$$

$$x_{mn} \in \{0, 1\} \qquad \forall m \in \mathcal{M}, \forall n \in \mathcal{N} \quad (5.6d)$$

Actually $\textbf{PSL}_2$ could be further simplified by fixing the value $\sum_{j>k} \sum_{i \in \theta_j} \omega_{in} x_{in}$ in constraint (5.6b). To do so, a new variable $\varpi_n^{k^+}$ is introduced to express the number of resources that are allocated to MTs with SL $> k$ in network n, *i.e.* the total weight of MTs with SL $> k$. Thus:

$$\varpi_n^{k^+} = \begin{cases} 0 & if\ k = K \\ \sum_{j>k} \sum_{i \in \theta_j} \omega_{in} x_{in} & if\ k < K \end{cases} \quad (5.7)$$

Note that $\varpi_n^{k^+}$ depends on the association results of MTs with SL $> k$. Hence, if the association decision for MTs with SL $> k$ is found, $\varpi_n^{k^+}$ can be considered as a constant value for MTs with SL k. Therefore, $\textbf{PSL}_2$ is distributed to K problems

which will be solved sequentially according to the decreased order of priority, *i.e.* $K,\ K-1,\ldots,1$. Thus, the user association and resource allocation problem for MTs with SL k is:

$$\mathbf{PSL_3}\text{: } max \sum_{m\in\theta_k}\sum_{n\in\mathcal{N}} f_{mn}x_{mn} \tag{5.8a}$$

$$s.\ t. \sum_{m\in\theta_k} \omega_{mn}x_{mn} \leqslant \zeta_n - \varpi_n^{k^+} \qquad \forall n\in\mathcal{N} \tag{5.8b}$$

$$\sum_{n\in\mathcal{N}} x_{mn} \leqslant 1 \qquad \forall m\in\theta_k \tag{5.8c}$$

$$x_{mn}\in\{0,1\} \qquad \forall m\in\theta_k, \forall n\in\mathcal{N} \tag{5.8d}$$

You can notice that problem $\mathbf{PSL_3}$ has the same structure as problem **P1**. The difference between problem $\mathbf{PSL_3}$ and Problem **P1** is that the former aims at maximizing the profit of all MTs in θ_k instead of $\mathcal{M}$. Moreover, the capacity constraint in Problem **P1** is "ζ_n" while in problem $\mathbf{PSL_3}$ "$\zeta_n - \varpi_n^{k^+}$" is considered. Therefore, the same solutions proposed in Chapters 2, 3, and 4 could be used to solve Problem $\mathbf{PSL_3}$. Hence, to solve Problem $\mathbf{PSL_3}$, we propose Algorithms 2′, 3′, 4′, 5′, 6′, 8′, and 9′, which have the same structure as Algorithms 2, 3, 4, 5, 6, 8, and 9 respectively except for $\mathcal{M}$ replaced by θ_k, and ζ_n replaced by $\zeta_n - \varpi_n^{k^+}$. Moreover, Algorithms 7′ and 10′ have the same structure as Algorithms 7 and 10 respectively except for $\mathcal{M}$ replaced by θ_k, and Algorithms 8 and 9 replaced by Algorithms 8′ and 9′ respectively.

5.1.4 Solution strategy

In this section, we discuss the proposed solution management strategy where a MT with low priority is not allowed to utilize resources allocated for MTs with higher priorities. The solution management strategy tries to minimize the number of times the optimization function is processed without affecting the optimality of the algorithm. Moreover, the aspects that trigger the resource allocation and user association algorithm are discussed. Mainly, the solution management strategy tries to decrease the number of MTs that are involved within the optimization problem ($\mathbf{PSL_3}$). The resource allocation and user association algorithm is triggered when one of the following scenarios occurs:

- The current serving network is not able to supply a certain MT with its requested data rate.
- A new connection is initiated.
- A MT with an active connection is about to leave the boundaries of its serving network.

However, it is not always required to run the optimization function. For instance, if a new connection is initiated, the MT could evaluate its candidate networks, and try to connect to the best one. If the target network has sufficient resources to serve the newly admitted connection, the MT will connect without having to run the optimization function. On the other hand, the target network might not be able to serve the MT unless it dissociates some MTs with lower SL. In this case, lets assume that the newly admitted MT has a SL of 3, it might be enough to dissociate

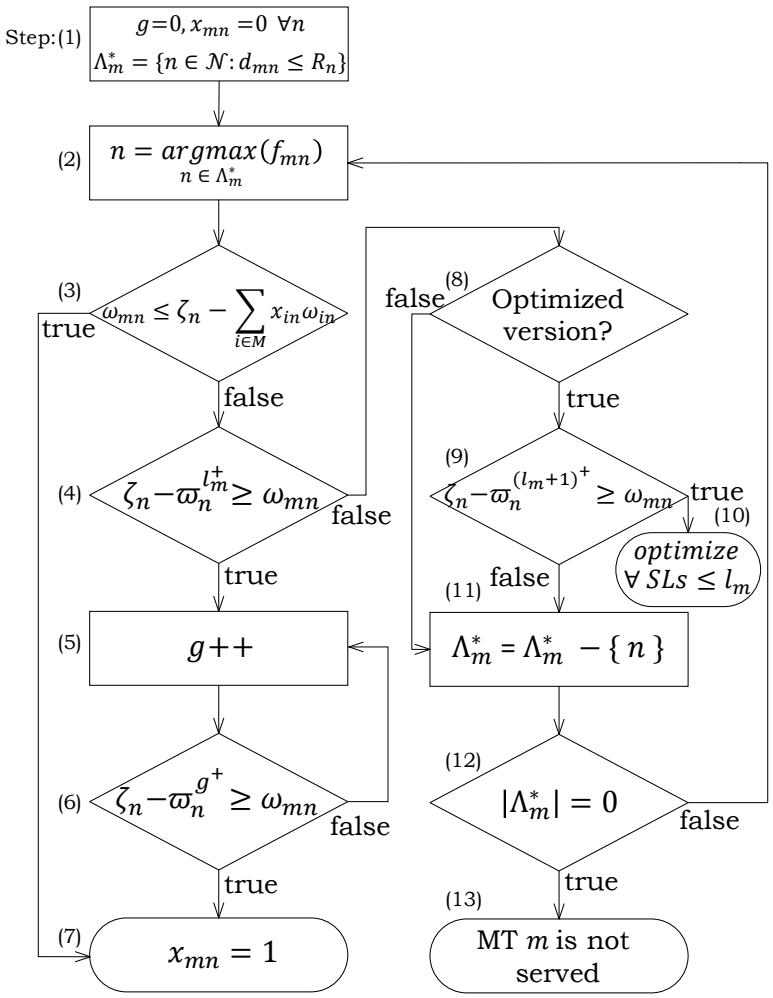


FIGURE 5.1 – The proposed algorithm that determines the association values of MT m when it experience one of the scenarios that trigger the resource allocation and MT association algorithm.

some MTs of SL 1, *i.e.* run the optimization function for MTs with SL 1, without having to encompass MTs with other SLs within the optimization problem. The fact that we have distributed problem $\mathbf{PSL}_2$ into K problems $\mathbf{PSL}_3$, each for a specific SL, enables applying such strategy without violating the optimality of the solution.

MT m undergoes the procedures shown in Figure 5.1 upon experiencing any of the scenarios that trigger the resource allocation and user association algorithm. The flow chart outputs the association variables x_{mn} for MT m, and an integer value g where all MTs with SLs $< g$ undergo the same procedures (shown in Figure 5.1), as well as some or all MTs with SL g. The optimized version of the solution (Figure 5.1-*step 8*) indicates using one of the methods proposed to solve or approximate problem $\mathbf{PSL}_3$. When the optimized version of the solution is deployed, the algorithm in Figure 5.1 finds the minimal number of SLs that will undergo the optimization problem $\mathbf{PSL}_3$. On the contrary, if the optimized version is not deployed, the algorithm describes the profit-function-based solution for the problem.

The algorithm in Figure 5.1 is detailed as follows: *Step 1* is an initialization step where integer g and association variables x_{mn} ($n \in \mathcal{N}$) are set to zero. Λ_m^* denotes the set of all networks for which MT m is within their coverage range. In *step 2*, the algorithm chooses the network n having the highest profit. The algorithm tests in *step 3* if the unallocated bandwidth resources of the selected network are sufficient to serve MT m. If that is the case, x_{mn} is immediately set to 1 (*step 7*), which indicates the association of MT m to network n. Conversely, if the number of requested resources is more than the unallocated ones, the algorithm proceeds to the next step. The association value x_{mn} could be immediately determined in *step 4*. If the number of resources that are not allocated to MTs with SLs $\geq l_m$ is sufficient to serve MT m, then this MT will be surely associated to network n. However, the algorithm enters an iterative process in *step 5* and *step 6* to determine the SL(s) of MTs that might be detached from the selected network. Of course, it is preferable to detach MTs of the lowest SL first. Hence, g increases by each iteration. On the other hand, in *step 4*, if the number of resources that are not allocated to MTs with SLs $\geq l_m$ is less than the number of resources requested by MT m, and if the optimized version of the solution is not deployed (*step 8*), then the algorithm tries to associate MT m to the next top-ranked network (*i.e.* the network with second highest profit in Λ_m^*). To do so, the selected network is removed from the list of available networks in *step 11*. *Step 12* tests if the cardinality of the available networks set is equal to 0, which indicates that the algorithm has already tried to associate MT m to all its reachable networks. If so, MT m will not be served as indicated in *step 13*. Otherwise, the algorithm tries to associate MT m to its next top-ranked network.

On the other hand, upon congestion, the optimized version of the solution allows MT m to use resources allocated for MTs with SLs $\leq l_m$. In the profit-function-based solution, MT m is not allowed to allocate resources utilized by MTs with SL $= l_m$. The idea here is to maximize the profit of MTs with SLs $\leq l_m$ by efficiently utilizing the resources. *Step 9* mainly tests if the number of resources that are not allocated to MTs with SLs $> l_m$ is sufficient to serve MT m. In this case, the optimization problem $\mathbf{PSL}_3$ is sequentially processed, in the decreasing order of SL, for each SL $\leq l_m$ (*step 10*).

5.1.5 User priority assignment

Till now, we have discussed the user association and resource allocation problem in HWNs with users having different priorities. However, the aspects that should be considered upon assigning user priorities are not discussed. Therefore, two general scenarios are presented.

The first scenario is based on a service level agreement (SLA) that could be signed between the user and the system operator. The system operator provides several SLs, each having a different pricing plan. Of course, it is expected that the best SL will have the most expensive pricing plan. Users are assigned to SLs according to their selected pricing scheme and the amount of money they are willing to pay in order to experience better service. The lowest SL is assigned for users who are not willing to pay extra money in order to experience better service.

The second scenario is related to the communication strategy in emergency

situations. Usually, in emergency or disastrous situations, a small number of BSs or APs remain active, and the PSNs suffer from extreme congestion. Therefore, the heterogeneous wireless system that is based on the remaining active BSs and APs becomes an essential alternative to PSNs. Hence, in order to prioritize the data traffic of medical, security, and emergency users, those users should be assigned to different SLs according to their priority. Normally, ordinary commercial users are assigned to the lowest SL in this case.

Several other dynamic scenarios could be also considered in order to assign user priorities. For example, users could be categorized according to their MT's battery status. In order to ensure that MTs in the most critical battery status category are associated to the access technology that requests the lowest power consumption, those MTs are assigned the highest priority, and their corresponding power consumption weight (w_m^{pc}) is set to one.

Note that it is beyond the scope of this thesis to discuss the advantages/disadvantages or the performance of each scenario. Instead, the problem formulated and solved could be applied in any scenario having different user priorities. Moreover, the formulated problem could be flexibly reconfigured to meet operator's objectives. For example, if it is requested to ensure that MTs with lowest SL are not always blocked upon extreme congestion, a specific number of resources in each network could be reserved for MTs with lowest SL. This can be configured in problem $\mathbf{PSL}_3$, by deducting in constraint (5.8b) the number of resources that should be reserved for MTs with lowest priority in network n. In other words, assuming that the number of resources that should be reserved in network n for MTs with lowest priority is denoted by D_n, then, "$\zeta_n - \varpi_n^{k^+}$" in constraint (5.8b) is replaced by "$\zeta_n - \varpi_n^{k^+} - D_n$" if $k \neq 1$ ($k = 1$ indicates the lowest SL).

5.2 A new power efficiency maximization solution

Due to the exponential growth in traffic demands, the power consumption of wireless networks has been subject to extreme expansion. In fact, the information and communication technology industry is estimated to account for 6% of the global CO2 emission in 2020 [LF16]. Thus, the power efficiency, *i.e.* data rate per power unit, has been imposed as an essential metric in the wireless communication research fields. Interestingly, the variety of access technologies in heterogeneous networks reveals different power consumption characteristics. Therefore, heterogeneous networks could efficiently participate in decreasing the amount of power consumed at both networks and MTs.

The power minimization problem has been extensively studied in heterogeneous networks. Many studies focus on the power consumption at the MT level. In [FZZ12] and [KYII10], each MT decreases its own power consumption through associating to the network that requests the lowest power consumption among available networks. These studies are categorized as user-centric network selection strategies ; each MT is satisfied individually without considering a system-wide user association and resource allocation scheme.

On the other hand, several studies propose a system-wide power minimization solution. In [VAN+15], authors have formulated an optimization problem to lower the transmission power of BSs in the downlink of heterogeneous cellular networks.

The optimization problem formulated in [CFC14] aims at minimizing the number of active BSs in the system. In [ZHY15], the formulated optimization problem aims at maximizing the system throughput and minimizing the transmission power of networks. However, the studies [VAN+15, ZHY15, CFC14] do not consider the amount of data rate requested by each MT.

The studies [AF16] and [AEK16] propose solutions to maximize the downlink throughput and energy efficiency. The formulated optimization problems set constraints on the minimum amount of data rate that should be received by each MT. However, the main objective is to enhance the system throughput. Such methodology could lead to an inefficient power utilization. For example, a VoIP application requests a specific amount of data rate, increasing the data rate beyond this amount does not contribute any benefit to the system. Instead, power is wasted on the additional supplied data rate. Hence, in both studies, the overall power consumption in the system is not evaluated. Moreover, all the studies [VAN+15, ZHY15, CFC14, AF16, AEK16] do not consider the power consumption at MTs. Instead, they only consider the power consumption at BSs and APs, *i.e.* network-centric. So, there is a need to complement the network-centric power efficiency maximization scheme by the user-centric approach.

Inspired by the user-centric power minimization methodologies, and encouraged by the network-centric power efficiency maximization solutions, this section proposes a user association and downlink resource allocation algorithm to maximize the power efficiency of the system. The load-aware proposed solution considers the data rate requested by MTs that are permitted to be associated with a single network at a time. Moreover, the power efficiency considers the transmission power of networks and the power consumption of MTs. The formulated optimization problem is similar in form to problem **P1**. Thus, the solutions proposed in this thesis are used to solve the problem.

5.2.1 Problem formulation and solution

In this section, we aim at formulating an optimization problem to maximize the overall power efficiency that considers the power consumption at MTs and networks. Moreover, the constraints of the formulated problem should guarantee supplying each MT with its requested data rate. The transmission power, in W, allocated by network n to MT m is calculated according to the following formula:

$$P_{mn}^{Tx} = \begin{cases} \frac{P_n^{RB} u_{mn}}{T_n} & \forall n \in \mathcal{N}_{\mathcal{BS}} \\ \\ \frac{P_n^{AP} t_{mn}}{T_n} & \forall n \in \mathcal{N}_{\mathcal{AP}} \end{cases} \tag{5.9}$$

Thus, the formulated optimization problem for power efficiency which is similar in form to problem **P1** is:

$$\textbf{PE: } max \sum_{m\in\mathcal{M}}\sum_{n\in\mathcal{N}} U(R_n - d_{mn}) \cdot \frac{Q_m}{P^{Tx}_{mn} + pc_{mn}} \cdot x_{mn} \tag{5.10a}$$

$$s.\ t. \sum_{m\in\mathcal{M}} \left\lceil \frac{Q_m T_n}{B^{RB}_n \log_2(1+\gamma_{mn})} \right\rceil x_{mn} \leqslant U_n \qquad \forall n \in \mathcal{N}_{\mathcal{BS}} \tag{5.10b}$$

$$\sum_{m\in\mathcal{M}} \left\lceil \frac{Q_m T_n}{r^{tot}_{mn}} \right\rceil x_{mn} \leqslant T_n \qquad \forall n \in \mathcal{N}_{\mathcal{AP}} \tag{5.10c}$$

$$\sum_{n\in\mathcal{N}} x_{mn} \leq 1 \qquad \forall m \in \mathcal{M} \tag{5.10d}$$

$$x_{mn} \in \{0,1\} \qquad \forall m \in \mathcal{M}, \forall n \in \mathcal{N} \tag{5.10e}$$

To avoid supplying MTs with excessive data rate, the power efficiency in (5.10a) is based on the requested data rate, *i.e.* Q_m, not on the received data rate as in [ZHY15, CFC14, AF16, AEK16]. Moreover, the power efficiency considers both the allocated power at networks and the power consumed at MTs.

Similar to what we did in Section 5.1.3, we propose Algorithm 1* having the same structure as Algorithm 1 except for $\overline{f}_{mn}$ replaced by $\frac{Q_m}{P^{Tx}_{mn}+pc_{mn}}$. Algorithms 2*, 3*, 4*, 5*, 6*, 8*, and 9*, which have the same structure as Algorithms 2, 3, 4, 5, 6, 8, and 9 respectively except for f_{mn} replaced by $\frac{Q_m}{P^{Tx}_{mn}+pc_{mn}}$. Hence, the efficiency factor e_{mn} in Algorithms 5 and 6 is replaced by e^*_{mn} as follows:

$$e^*_{mn} = \begin{cases} \frac{\frac{Q_m}{P^{Tx}_{mn}+pc_{mn}}}{B^{RB}_n(u_{mn}/T_n)} & \forall n \in \mathcal{N}_{\mathcal{BS}} \\ \\ \frac{\frac{Q_m}{P^{Tx}_{mn}+pc_{mn}}}{B_n(t_{mn}/T_n)} & \forall n \in \mathcal{N}_{\mathcal{AP}} \end{cases} \tag{5.11}$$

Moreover, Algorithms 7* and 10* have the same structure as Algorithms 7 and 10 respectively except for f_{mn} replaced by $\frac{Q_m}{P^{Tx}_{mn}+pc_{mn}}$, and Algorithms 8 and 9 replaced by Algorithms 8* and 9* respectively.

5.3 Performance evaluation (priority-based case)

5.3.1 Simulation parameters

In the simulation parameters, since we have different SLs now, we have increased the maximum number of MTs to 138 in order to better study the characteristics of each SL alone. The number of users in each SL is the same. Therefore, the simulation parameters are the same as those presented in Section 3.7.1, except that we have increased the number of APs in the SA to three, and the number of available RBs in each LTE BS to fifty.

5.3.2 Evaluation metrics

Similar to the metrics used in Section 3.7.2, the following metrics are used to evaluate the proposed solutions: average profit, average satisfaction, average signal

quality, average instantaneous power consumption, and blocking percentage. The satisfaction of MT m when associated with network n is:

$$\rho_{mn} = \frac{f_{mn}}{f_{mn'}} \tag{5.12}$$

where n' is the index of the network for which MT m achieves the highest profit.

In fact, studying the average value of an attribute is not straightforward in a scenario where MTs request different amounts of data rate. For example, the average profit per user, *i.e.* $\frac{\sum_{m\in\theta_k}\sum_{n\in\mathcal{N}} f_{mn}x_{mn}}{|\theta_k|}$, could be increased through increasing the profit of MTs with low data rate requirements on the expense of other MTs. Thus, to avert deceptive results, the average profit per requested data rate is studied according to the following formula:

$$\frac{\sum_{m\in\theta_k}\sum_{n\in\mathcal{N}} f_{mn}x_{mn}}{\sum_{m\in\theta_k} Q_m} \tag{5.13}$$

Similarly, the average satisfaction per requested data rate is studied according to the following formula:

$$\frac{\sum_{m\in\theta_k}\sum_{n\in\mathcal{N}} \rho_{mn}Q_m x_{mn}}{\sum_{m\in\theta_k} Q_m} \tag{5.14}$$

Since ρ_{mn} represents a normalized value, it is multiplied by Q_m in the above formula.

For the signal quality and power consumption, the average values per served data rate are considered because there is no mean to calculate these values for the blocked data rates. For example, setting 0 for the power consumption of blocked data rate will decrease the average consumed power and contribute misleading results. Therefore, for power consumption, the average value per served kbps, in mW/kbps, is:

$$\frac{\sum_{m\in\theta_k}\sum_{n\in\mathcal{N}} pc_{mn}x_{mn}}{\sum_{m\in\theta_k}\sum_{n\in\mathcal{N}} Q_m x_{mn}} \tag{5.15}$$

Since the value of the signal quality is not related to the requested data rate, it is multiplied by Q_m to reflect the actual signal quality per served data rate. Thus the average relative received signal quality per served data rate is:

$$\frac{\sum_{m\in\theta_k}\sum_{n\in\mathcal{N}} s_{mn}Q_m x_{mn}}{\sum_{m\in\theta_k}\sum_{n\in\mathcal{N}} Q_m x_{mn}} \tag{5.16}$$

5.3.3 Simulation results

As we have mentioned before, we will study the performance of the algorithms discussed in Section 5.1.3.

5.3.3.1 Multiple service levels

First, concerning the effect of providing different SLs, it is obvious from Figure 5.2 and Figure 5.3 that the proposed scheme maintains better profit and satisfaction for high-priority users. It is important to note that the average profit is 0.52

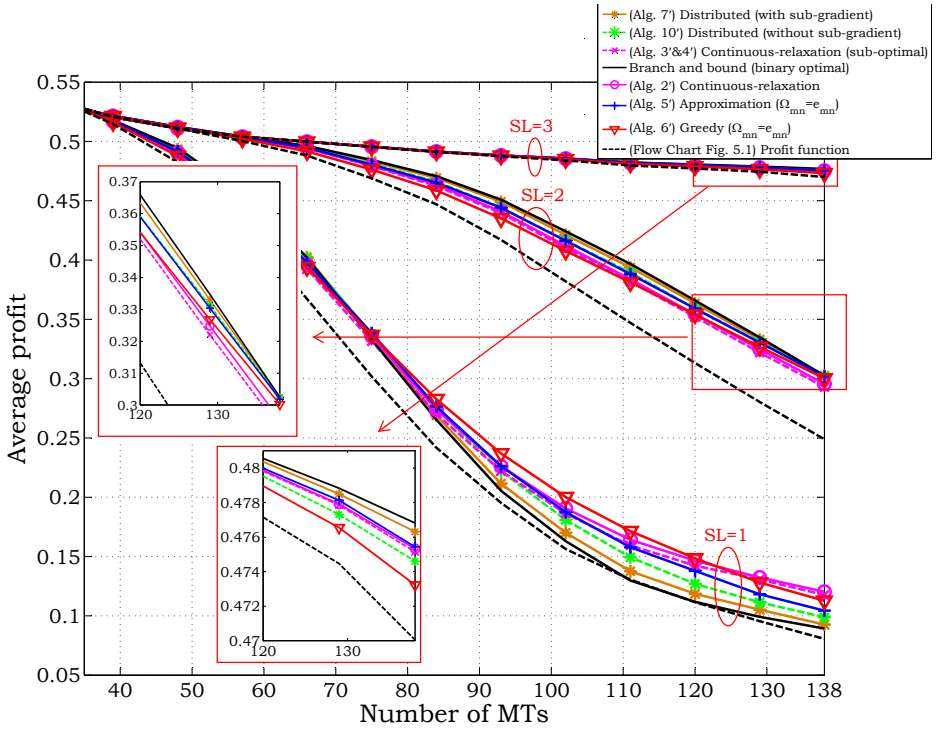


FIGURE 5.2 – Average profit (according to Eq. (5.13)).

when the number of MTs is 35 because the profit function normalizes attributes through dividing them by the global maximum, *i.e.* $\max_{\forall m,n}$, which will only lead to a profit value of 1 when a single MT exists in the system. However, this behavior does not impact the user satisfaction. As the number of MTs reaches 138, the average satisfaction is approximately 0.99, 0.6, and 0.22 for MTs with SLs 3, 2, and 1 respectively (Figure 5.3). Therefore, a remarkable increase in satisfaction is maintained upon subscribing to higher SL. The same aspect is observed for the signal quality (Figure 5.4) and instantaneous power consumption (Figure 5.5). Moreover, MTs with SLs 3 and 2 do not experience any blockage. Therefore, Figure 5.6 shows the percentage of blocked data rate for MTs with SL 1 only.

5.3.3.2 General behavior of algorithms

In general, increasing the number of MTs in the system strengthens the competition to acquire the limited resources of networks. Therefore, the opportunity that MTs connect to their preferred network decreases. Consequently, MTs experience degraded service illustrated by the decrease in profit and satisfaction as shown in Figure 5.2 and Figure 5.3 respectively. In order to increase the overall profit, the optimization problem $\mathbf{PSL}_3$ finds the best set of association values for all MTs in θ_k. Thus, the main performance of the optimization problem and its different solutions could be studied through the profit, and consequently through the satisfaction because it is directly related to the profit. It is shown in Figure 5.2 and Figure 5.3 that the proposed distributed solution with sub-gradient algorithm maintains the nearest performance to the optimal binary solution based on the branch and

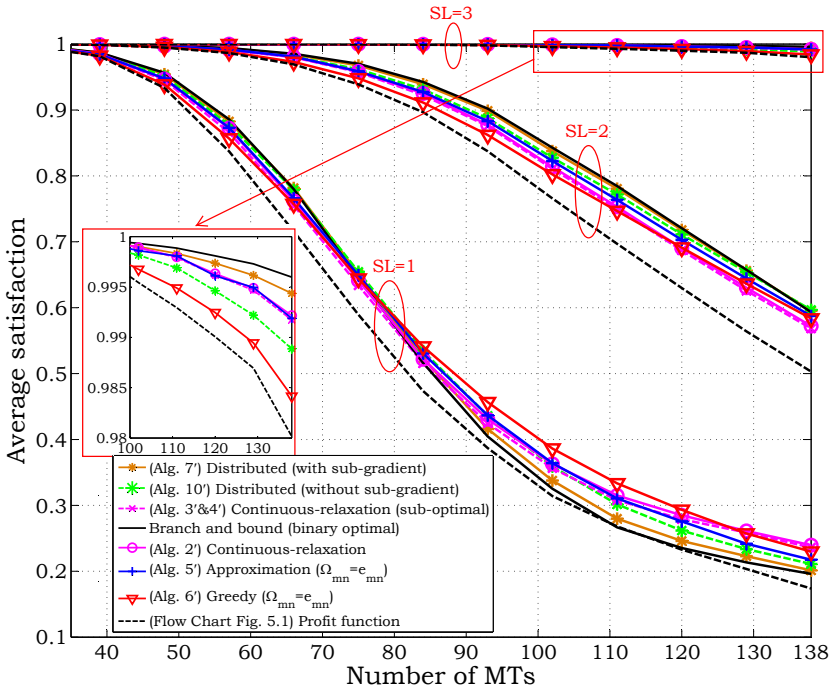


FIGURE 5.3 – Average satisfaction (according to Eq. (5.14)).

bound algorithm for MTs with SL 3. The performance of the proposed distributed solution with sub-gradient algorithm for MTs with SL 3 is respectively followed by the approximation-based solution, the relaxation-based solutions, and the distributed solution without the sub-gradient method. The greedy solution, although it tends to approach the optimal solution, performs near the profit-function-based solution which has the worst performance. Therefore, the proposed distributed solution with the sub-gradient method efficiently approximates the optimal solution, and overwhelms the rest of solutions.

Concerning MTs with SL 2, as the number of MTs increases, the distributed, relaxation-based, approximation-based, and greedy solutions perform near the optimal solution, and far away from the profit-function-based solution. It is remarkable that the proposed distributed solution with the sub-gradient method maintains the nearest performance to the optimal one (Figure 5.2 and Figure 5.3).

Optimal resource allocation for high-priority MTs causes efficient resource utilization in networks. Hence, the chance that MTs with low priority associate to their preferred network decreases. For example, Wi-Fi APs are usually preferred for their low power consumption feature ; efficient resource utilization in these APs lowers the number of unallocated resources, which in turns lowers the chance that MTs with low priority associate to these APs. Consequently, upon adopting the optimal solution, MTs with SL 1 experience service near the profit-function-based solution. This is recognized in Figure 5.2 and Figure 5.3 where the optimal solution starts approaching the profit-function-based solution as the number of MTs increases beyond 80.

As a matter of fact, the profit function depends on the location of the MT,

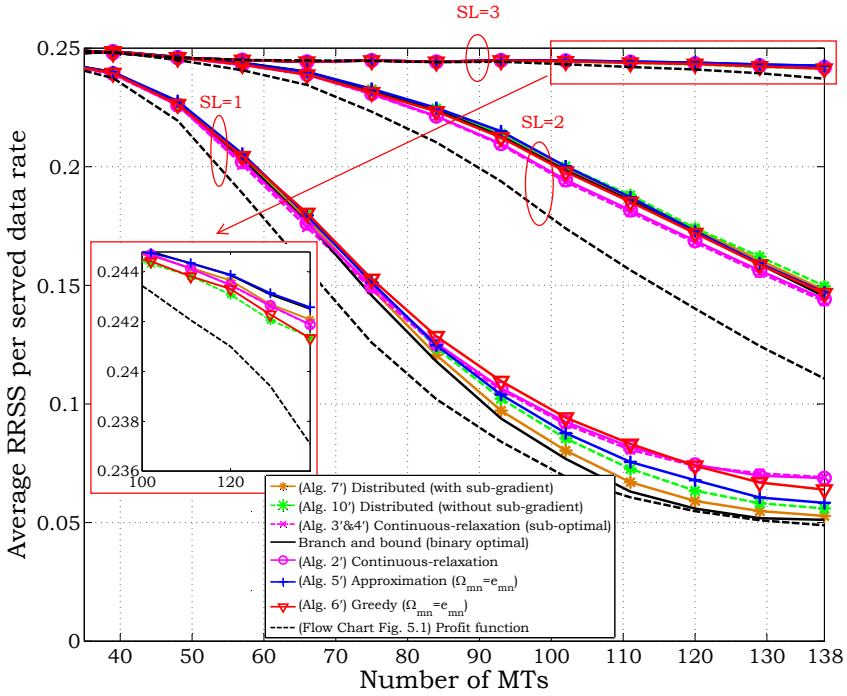


FIGURE 5.4 – Average RRSS (according to Eq. (5.16)).

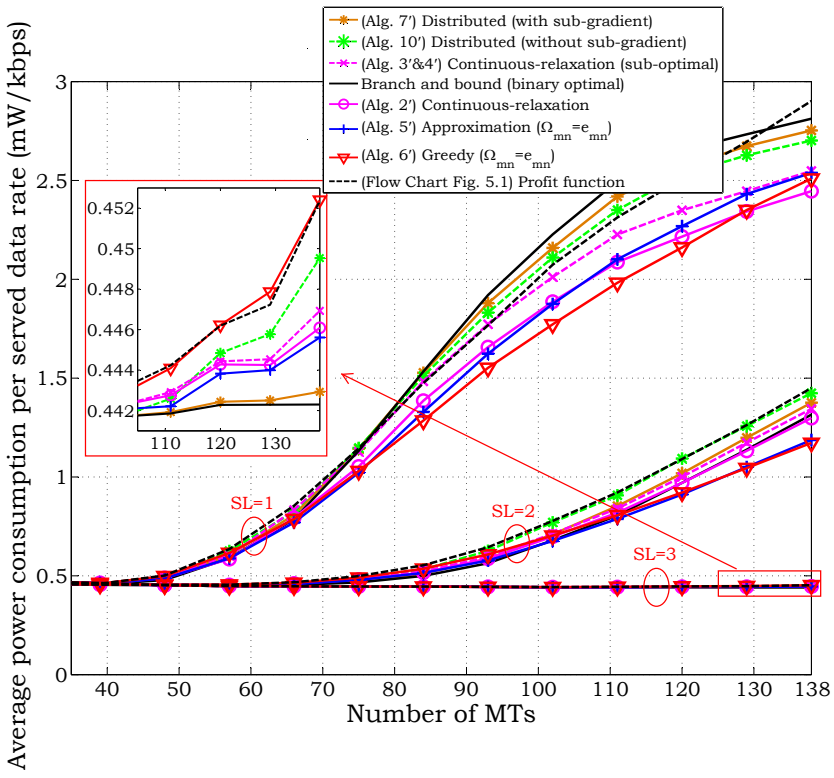


FIGURE 5.5 – Average MT power consumption (according to Eq. (5.15)).

the requested data rate, the normalized values of the signal quality and power

consumption, and the weights w_m^s and w_m^{pc} which are different between MTs. Therefore, it is normal not to notice the same behavior of the satisfaction curve (Figure 5.3) reflected in the curves of the signal quality and power consumption (Figure 5.4 and Figure 5.5 respectively). However, the satisfaction could reflect a general behavior of the compared solutions in terms of signal quality and power consumption. For example, Figure 5.4 and Figure 5.5 show that the proposed distributed solution with sub-gradient method maintains near-optimal performance for MTs with SL 3. This is illustrated through high signal quality and low instantaneous power consumption. Moreover, the degraded service of the optimal solution for MTs with SL 1 in Figure 5.4 and Figure 5.5 is a result for the same reason discussed before for the profit and satisfaction of those MTs.

5.3.3.3 Blocking percentage evaluation

In order to fully understand the behavior of the proposed solutions, the percentage of blocked data rate should be studied. According to the proposed solution, the HOE has the privilege to reassign resources used by low-priority MTs to MTs with higher priorities. Therefore, MTs with SLs 3 and 2 do not suffer from any blockage throughout the simulation. Concerning MTs with SL 1, Figure 5.6 illustrates that the optimal solution achieves the lowest data rate blockage. The distributed solutions maintains the nearest blocking percentage to the optimal binary solution, followed by the approximation-based, relaxation-based, greedy, and profit-function-based solutions respectively. For instance, the profit-function-based solution suffers from 19.7% blockage when the number of MTs reaches 138. The greedy solution and the sub-optimal continuous-relaxation-based solution lower down this percentage to 11, followed by the optimal continuous-relaxation-based and approximation-based solutions that score 10% and 9.8% respectively. The distributed solution without the sub-gradient method achieves 9.4% blocking percentage, while the distributed solution with the sub-gradient method and the optimal solution scores about 8.9%. You can notice that the distributed solution with the sub-gradient method achieves slightly higher blocking percentage than the optimal binary solution. Therefore, as can be seen in Figure 5.6, the proposed distributed solution with the sub-gradient method achieves and maintains the lowest blocking percentage among the tested approaches, except for the optimal one of course. Such result is considered as a major improvement since users subscribing to the lowest SL would be mainly concerned about having a service, without paying much attention to the performance.

5.4 Performance evaluation (power efficiency case)

5.4.1 Simulation parameters

Same as those mentioned in Section 3.7.1, except that we want to focus on the power characteristics when we have zero blockage, so we have increased the number of RBs in an LTE BS to forty and the transmission power of an AP to one W.

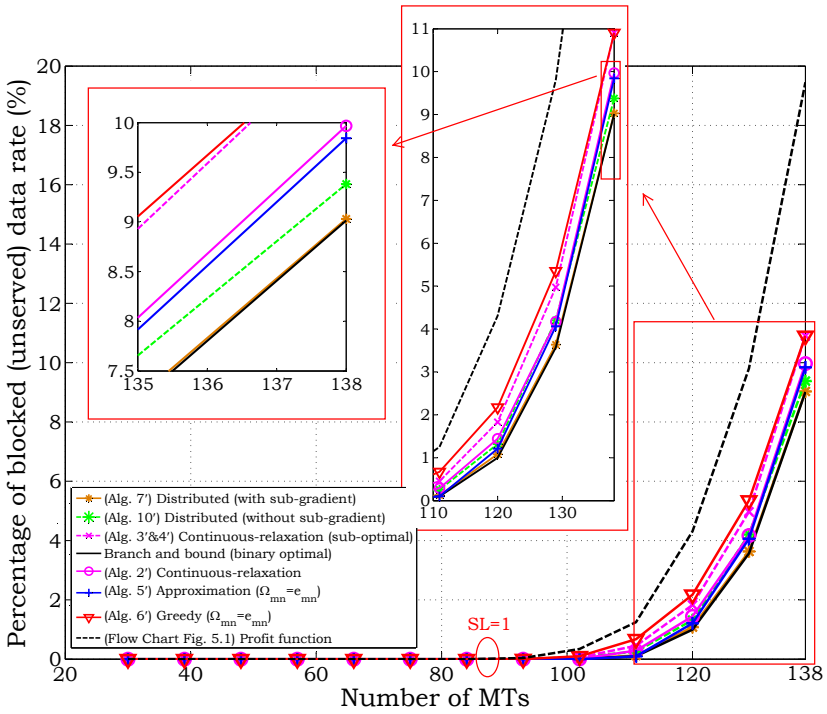


FIGURE 5.6 – Percentage of blocked (unserved) data rate.

5.4.2 Evaluation metrics

The power efficiency is studied according to the following formula:

$$\sum_{m\in\mathcal{M}}\sum_{n\in\mathcal{N}}\frac{Q_m}{P_{mn}^{Tx}+pc_{mn}}.x_{mn} \tag{5.17}$$

The overall network (BSs and APs) allocated transmission power is studied according to the following formula:

$$\sum_{\forall m\in\mathcal{M}}\sum_{\forall n\in\mathcal{N}}P_{mn}^{Tx}x_{mn} \tag{5.18}$$

Since the context of this paper deals with MTs requesting different amounts of data rate, it is not convenient to study the average power consumption per MT. Therefore, the average MT power consumption per requested data rate unit (*i.e.* kbps) is studied according to the following formula:

$$\frac{\sum_{\forall m\in\mathcal{M}}\sum_{\forall n\in\mathcal{N}}pc_{mn}x_{mn}}{\sum_{\forall m\in\mathcal{M}}\sum_{\forall n\in\mathcal{N}}Q_m x_{mn}} \tag{5.19}$$

Note that the average power consumption per data rate (in terms of W/bps) has been also studied in [AEK16].

5.4.3 Simulation results

In this section, we compare the performance of the algorithms discussed in Section 5.2.1.

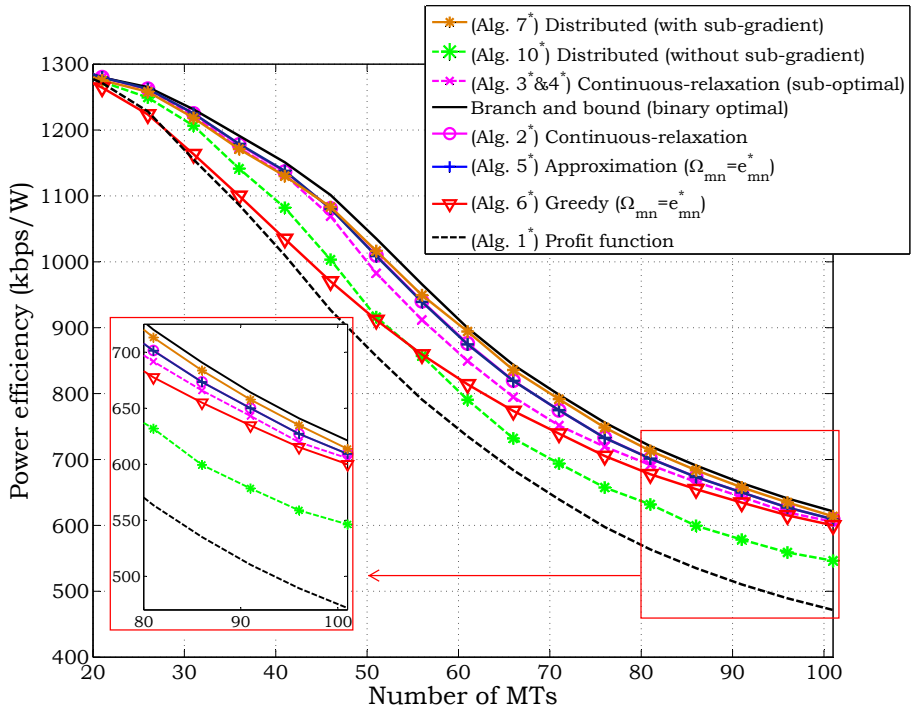


FIGURE 5.7 – Power efficiency (according to Eq. 5.17).

The objective function of the optimization problem **PE** aims at increasing the power efficiency. Therefore, the main behavior of the optimization problem and the solutions could be studied based on the power efficiency. As the number of MTs increases in the system, networks become more congested. Hence, the chance that each MT is assigned to its most power-efficient network decreases, which in turns causes degradation in the power efficiency as shown in Figure 5.7. Moreover, Figure 5.7 shows that the proposed distributed solution with the sub-gradient method maintains the nearest performance to the optimal binary solution, where the optimal binary solution enhances the increases efficiency by 42%, and the distributed solution by 40% when compared to the basic profit-function-based solution and when the number of MTs reaches 101. Followed by the optimal continuous-relaxation-based solution and the approximation-based solution which performs approximately the same and enhances the power efficiency by 37.7% when the number of MTs reaches 101. Then, the performance is followed by the sub-optimal continuous-relaxation-based solution, the greedy solution, and the distributed solution without the sub-gradient method which enhances the power efficiency by 36, 35.5, and 23 % respectively. Of course, the profit-function-based solution performs the worst.

Concerning the transmission power of wireless networks, as the amount of requested data rate increases, networks allocate more downlink resources. Therefore, the total network transmission power increases upon increasing the number of MTs in the system as shown in Figure 5.8. It is shown that when the number of MTs increases beyond 70, the distributed solution with the sub-gradient method achieves the lowest network transmission power, followed by the optimal

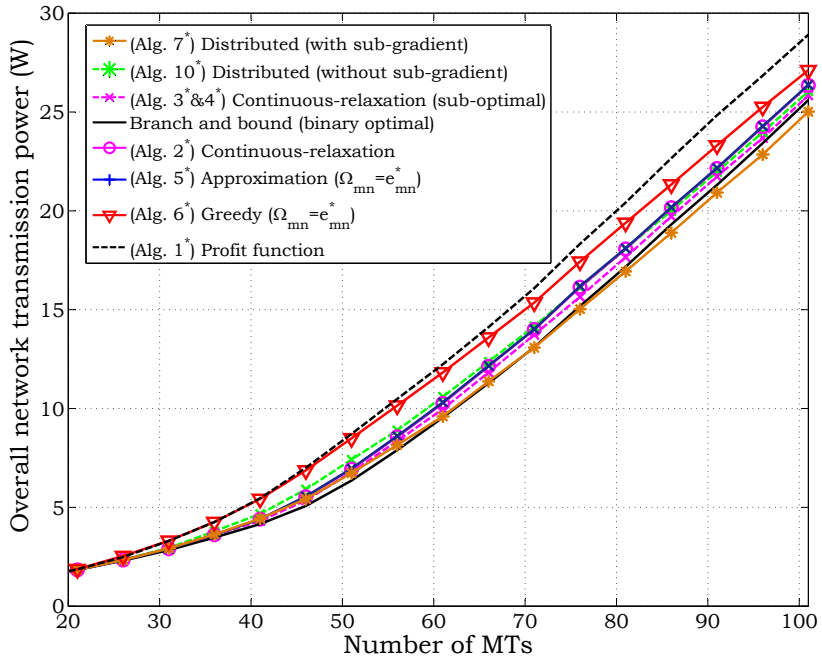


FIGURE 5.8 – Overall network transmission power (according to Eq. (5.18)).

binary solution, the sub-optimal continuous-relaxation-based solution, the distributed solution without the sub-gradient method, the approximation-based and the optimal continuous-relaxation-based solutions, the greedy solution, and finally the profit-function-based solution.

The total amount of power consumed in the system highly depends on the MT power consumption. Since $\omega_n \simeq 1.2$ W for LTE BSs and $\simeq 0.1$ W for Wi-Fi APs, then, approximately 1.1 W could be reduced just by associating a MT to an AP instead of BS. Interestingly, it is shown in Figure 5.9 that the average MT power consumption per data rate unit, *i.e.* kbps, could be enhanced up to 30% upon adopting the optimal binary solution, and when compared to the profit-function-based solution when the number of MTs is 101. Then, the optimal continuous-relaxation-based solution and the approximation-based solution enhances this percentage by 29.3, followed by the greedy solution (28.5 % enhancement), the distributed solution with the sub-gradient method (28 % enhancement), the sub-optimal continuous-relaxation-based solution, and the distributed solution without the sub-gradient method which achieves 28.2% and 15% enhancement respectively.

5.5 Conclusion

In this chapter, we have shown how the solutions proposed in this thesis could be used to inspect new perspectives. Mainly, we have discussed new direct applications for the problem **P1** proposed in Chapter 2, and its corresponding solutions proposed in Chapters 3 and 4. First, we proposed a new application for the

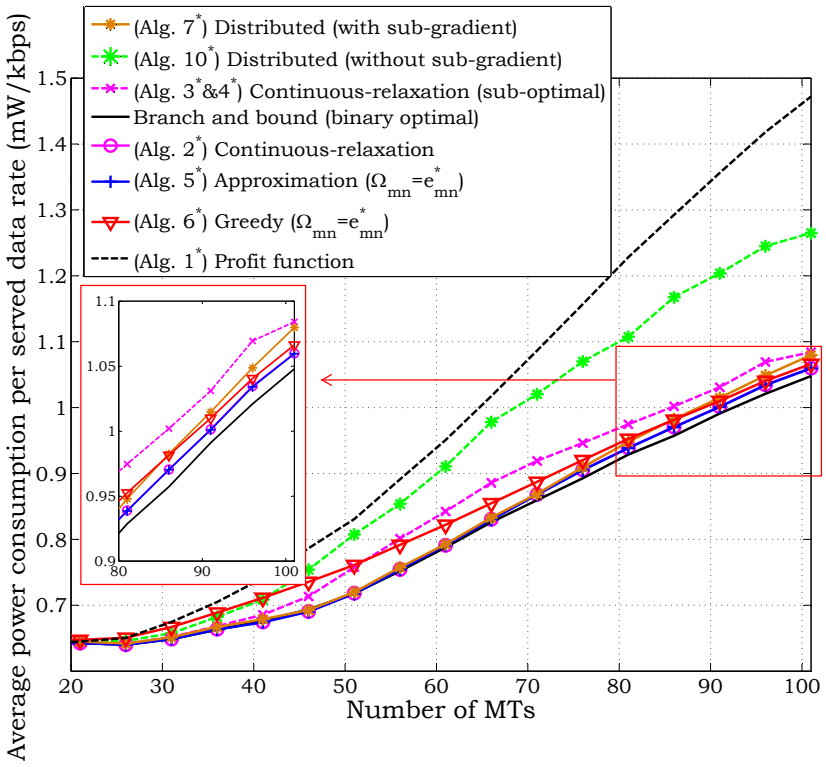


FIGURE 5.9 – Average MT power consumption (according to Eq. (5.19)).

priority-based user association and resource allocation problem in a system where users have different priorities. We formulated a new priority-based optimization problem. Then we simplified it into a new problem with form similar to the problem **P1**, and therefore could be solved with the solutions proposed in this thesis. Moreover, we discussed new perspective for the user association and resource allocation problem to enhance the overall power efficiency (data rate per unit power) in the system. Simulation results confirm that the proposed distributed solution with the sub-gradient method maintains the nearest performance to the optimal solution. Moreover, the proposed centralized approximation-based solution also shows remarkable results. The rest of solutions proposed in this thesis have shown also performance much better than the trivial profit-function-based solution. In particular, the simulation results related to the priority-based management case encourage users to subscribe to the highest priority where they experience lower blockage and better service. In addition, the simulation results of the power efficiency application shows that the proposed solutions plays a vital role in decreasing the overall network transmission power, and the MT power consumption.

Conclusion and Perspectives

Conclusion

In this thesis, we have proposed several solutions with tolerable complexity for the user association and resource allocation problem in heterogeneous networks. The proposed solutions aim at providing an optimized ABC vision that considers the system as a whole while making decisions. Therefore, users' satisfaction, needs, and preferences are considered.

First, in Chapter 2, we have discussed the 802.21.1 MIS SDRAN framework that provides seamless HO, resource allocation, and centralized management in HWNs. Based on this framework we have proposed a novel centralized HO scenario and formulated a new global optimization problem for the user association and downlink resource allocation problem in HWN. The novel formulated optimization problem considers the network's capacity and resource allocation constraints, and aims at maximizing the overall user-centric profit in the system. The user-centric profit is based on a weighted profit function that aims at jointly increasing the RRSS and decreasing the MT power consumption. The weights of the profit function are set according to the user preferences. It is shown through simulations that the formulated global optimization problem algorithm outperforms significantly the classical profit-function-based solution. Moreover, it is shown that the proposed profit function responds efficiently to the weight variations. Therefore, it can be effectively tuned to meet user preferences.

In fact, the global optimization problem proposed in Chapter 2 is solved using the branch and bound algorithm where the processing time could increase tragically upon increasing the number of MTs and networks in the system. Therefore, in Chapter 3, we have proposed different centralized solutions with lower complexities for the formulated user association and resource allocation problem. The first two proposed solutions are based on the continuous-relaxation of problem, *i.e.* relaxing the binary association constraint into a bounded continuous. The first continuous-relaxation-based solution is with undetermined complexity, and is considered as the optimal solution based on the continuous-relaxation methodology. The second continuous-relaxation-based solution is with determined polynomial-time complexity, and is considered as a sub-optimal solution based on the continuous-relaxation approach. Although the proposed solutions based on the continuous-relaxation approach have lower complexity than the branch and bound algorithm, their complexity is still based on the cube of the number of networks multiplied by the cube of the number of MTs in the system. Therefore, both solutions do not scale well upon increasing the number of MTs and networks in the system. Moreover, relaxing the binary constraint into continuous changes

the characteristics of the problem, which in turns threatens the optimality of the solution. Therefore, a novel approximation-based solution with lower complexity have been proposed to approximate the formulated optimization problem without relaxing the binary constraint. In addition, a new simple greedy heuristic algorithm have been also proposed. The approximation-based solution and the greedy solution have been further optimized through proposing a new efficiency factor to estimate the gain contributed upon associating users to networks. The efficiency factor considers the data rate requirement of users, and the channel conditions between the MT and the BS or AP.

Simulation results in Chapter 3 show that the proposed sub-optimal continuous-relaxation-based solution effectively approximates the optimal solution based on the continuous-relaxation approach. Moreover, simulation results confirm that the continuous-relaxation approach threatens the optimality of the solution. On the other hand, the proposed approximation-based solution has shown better performance than the continuous-relaxation-based solutions. The simpler greedy solution have also shown acceptable performance similar to the solutions based on the continuous-relaxation approach, but with much lower complexity. The proposed efficiency factor (e_{mn}) has also contributed a significant boost in the performance of the greedy and approximation-based solutions. Generally, the approximation-based solution demonstrated a remarkable trade-off between the complexity and performance.

All the solutions proposed in Chapter 3 rely on a centralized entity that is responsible for processing the proposed algorithms. Therefore, in Chapter 4, we have proposed two novel distributed algorithms where the processing now is shared between all the networks, *i.e* each network is responsible for processing a part of the algorithm, instead of one centralized entity carrying all the processing. Mainly, in Chapter 4, the optimization problem formulated in Chapter 2 has been distributed into several Knapsack problems, where each problem is independently solved by its corresponding network. The Knapsack problem is solved using dynamic programming procedures in an acceptable time. The first distributed solution is based on the Lagrangian-relaxation of the optimization problem, and on the sub-gradient method that finds near-optimal Lagrangian multipliers. On the other hand, the second distributed solution is similar to the first one but without the sub-gradient method. The advantage of the second solution is that it eliminates the iterative sub-gradient phase, which leads to a lower complexity. The complexity of the distributed solution with sub-gradient method is higher than that of the greedy and approximation-based solution, but lower than the continuous-relaxation-based solutions. The second proposed distributed solution without the sub-gradient method has a very competitive complexity. Simulation results have shown that the proposed distributed solution with the sub-gradient method achieves the best performance results among all other proposed solutions while requiring a slightly higher number of handovers. The remarkable advantage of the distributed solution without the sub-gradient method is that it achieves performance similar to the centralized efficiency-based greedy algorithm.

The optimization problem formulated in Chapter 2, and solved in Chapters 3 and 4, aims at maximizing the overall profit of the users in the system. Therefore, in Chapter 5, we have shown how the solutions proposed in the previous chap-

ters could be used to inspect new perspectives. Mainly, we have discussed new direct applications for the formulated problem and its corresponding solutions. First, we have proposed a new application for the priority-based user association and resource allocation problem in a system where users have different priorities. Moreover, we have discussed new perspective for the user association and resource allocation problem to enhance the overall power efficiency (data rate per unit power) in the system. Simulation results confirms that the proposed distributed solution with the sub-gradient method maintains the nearest performance to the optimal solution. Moreover, the proposed centralized approximation-based solution also shows remarkable results. The rest of solutions proposed in this thesis have shown also performance much better than the trivial profit-function-based solution. In particular, the simulation results related to the priority-based management case encourage users to subscribe to the highest priority where they experience lower blockage and better service. In addition, the simulation results of the power efficiency application shows that the proposed solutions plays a vital role in decreasing the overall network transmission power, and the MT power consumption.

Perspectives

Many interesting areas and prospects can be further investigated for future works based on this thesis:

First, we can directly study new perspectives for the formulated problem and its corresponding solutions, similar to what we did in Chapter 5. For example, we can study the case where the objective is to maximize the overall data rate in the system, and compare the effect of this objective on the user experience, *i.e.* MT power consumption and RRSS. Accordingly, we can propose a new solution that jointly aims at enhancing the network profit (overall data rate) and the user-centric profit.

Second, the optimization problem formulated in Chapter 2 could be easily extended to account for the uplink resource management, instead of just focusing on the downlink resources. This could be achieved by adding two linear constraints to problem **P1**, similar to constraints (2.22b) and (2.22c) but for the uplink resources. Accordingly, the continuous-relaxation approach proposed in Chapter 3 could be also used to solve the problem, but the heuristic algorithms proposed in Chapter 3 should also count for the uplink resources in the same methodology provided for the downlink resources. Moreover, the Lagrangian-relaxation proposed in Chapter 4 could be also used to the newly formulated problem. However, the Knapsack problem now should account for the uplink and downlink resources. The Knapsack problem formulated in Chapter 4 is one-dimensional, *i.e.* account only for the downlink resources. To extend the Knapsack solution to uplink and downlink resources, a multi-dimensional Knapsack problem, in fact two-dimensional, should be formulated. The multi-dimensional Knapsack problem could be also solved using dynamic programming procedures.

Third, in this thesis, we studied the case where users request a specific data rate for different applications (voice, video, and FTP) listed in Table 2.2. In fact, the formulated problem could be altered easily to maximize the overall quality

of experience (QoE). For example, in Table 2.2, we have chosen three different codecs for the voice application. Each codec reflects a specific QoE value, usually the mean opinion score (MOS) is used to quantify the QoE. Therefore, a voice user could now be given one of multiple defined data rates, each for a given codec. Accordingly, a new linear constraint could be added to problem **P1** in order to limit the user for having a single requested data rate that could be chosen by the algorithm in order to maximize the overall QoE. Video application has also its specific QoE criteria similar to voice. For FTP users, a linear equation could be proposed to relate the provided data rate to the QoE. Similar to what we have discussed in the previous paragraph, the solutions proposed in this thesis could be altered to enhance the QoE. Moreover, a multi-dimensional Knapsack problem could be used, where each codec could be seen as a new dimension.

Finally, the optimization problem formulated in this thesis could be extended to account for users with elastic data rate requirements, *i.e.* variable data rate requirement instead of fixed one. However, the solution here is not straight forward as the three previously provided perspectives. The number of resources requested by users with elastic traffic could not be fixed, which leads to non-linear problem for these users. Accordingly, the optimization problem is divided into two major problems, one linear problem for users with inelastic data rate requirements, and a second non-linear problem for users with elastic data rate requirements. The former problem could be solved by the solutions proposed in this thesis, while the later problem is solved by methods specified for non-linear optimization problems. Note that non-linear problems for users with elastic traffic have been widely discussed in HWNs.

Bibliography

[3GP10a] 3GPP. Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN), Overall description Stage 2, Report TS 136.300, V8.12. http://www.3gpp.org, Apr. 2010.

[3GP10b] 3GPP. Network architecture, Technical Specification Group Services and Systems Aspects, Report No 23.002 v3.4.0, France. http://www.3gpp.org, Dec. 2010.

[47609] IEEE Standard for Local and metropolitan area networks - Media Independent Handover Services. *IEEE Std 802.21-2008*, January 2009.

[AEK16] H. Ameur, M. Esseghir, and L. Khoukhi. Fully distributed approach for energy saving in heterogeneous networks. In *2016 8th IFIP International Conference on New Technologies, Mobility and Security (NTMS)*, pages 1–6, Nov 2016.

[AF16] M. Adedoyin and O. Falowo. An energy-efficient radio resource allocation algorithm for heterogeneous wireless networks. In *2016 IEEE 27th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, pages 1–6, Sept 2016.

[AK10] K Ayyappan and R Kumar. Qos based vertical handoff scheme for heterogeneous wireless networks. *International Journal of Research and Reviews in Computer Science (IJRRCS)*, 1(1):1–6, 2010.

[AM16] R. Amin and J. Martin. Assessing performance gains through global resource control of heterogeneous wireless networks. *IEEE Transactions on Mobile Computing*, 15(2):292–305, Feb 2016.

[AMC16] A. Awad, A. Mohamed, and C. F. Chiasserini. User-centric network selection in multi-rat systems. In *IEEE Wireless Communications and Networking Conference*, pages 1–6, April 2016.

[Bab] Guglielmo , BabelTen Application, Reggio Emilia, Italy. http://www.guglielmo.biz/Servizi.aspx ?lan=eng.

[BB15] H. Boostanimehr and V. K. Bhargava. Unified and distributed qos-driven cell association algorithms in heterogeneous networks. *IEEE Transactions on Wireless Communications*, 14(3):1650–1662, March 2015.

[BHW13] S. Borst, S. Hanly, and P. Whiting. Optimal resource allocation in hetnets. In *IEEE International Conference on Communications (ICC)*, pages 5437–5441, June 2013.

[CC08] B. J. Chang and J. F. Chen. Cross-layer-based adaptive vertical handoff with predictive rss in heterogeneous wireless networks. *IEEE Transactions on Vehicular Technology*, 57(6):3679–3692, Nov 2008.

[CCHL09] B. J. Chang, J. F. Chen, C. H. Hsieh, and Y. H. Liang. Markov decision process-based adaptive vertical handoff with rss prediction in heterogeneous wireless networks. In *IEEE Wireless Communications and Networking Conference*, pages 1–6, April 2009.

[CDM04] Andrea Calvagna and Giuseppe Di Modica. A user-centric analysis of vertical handovers. In *Proceedings of the 2nd ACM international workshop on Wireless mobile applications and services on WLAN hotspots*, pages 137–146. ACM, 2004.

[CFC14] J. H. Chu, K. T. Feng, and T. S. Chang. Energy-efficient cell selection and resource allocation in lte-a heterogeneous networks. In *2014 IEEE 25th Annual International Symposium on Personal, Indoor, and Mobile Radio Communication (PIMRC)*, pages 976–980, Sept 2014.

[Cis14] Cisco. Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update (2013–2018), San Jose, CA, USA, 2014.

[Cis16] Cisco. Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update 2015–2020 White Paper. http://www.cisco.com/c/en/us/solutions/service-provider/visualnetworking- index-vni/white-paper-C11-520862.html, Feb. 2016.

[CL16] L. Chen and H. Li. An mdp-based vertical handoff decision algorithm for heterogeneous wireless networks. In *IEEE Wireless Communications and Networking Conference*, pages 1–6, April 2016.

[CLL+16] Y. Chen, J. Li, Z. Lin, G. Mao, and B. Vucetic. User association with unequal user priorities in heterogeneous cellular networks. *IEEE Transactions on Vehicular Technology*, 65(9):7374–7388, September 2016.

[CYCK05] Jaehyuk Choi, Joon Yoo, Sunghyun Choi, and Chongkwon Kim. Eba: an enhancement of the ieee 802.11 dcf via distributed reservation. *IEEE Transactions on Mobile Computing*, 4(4):378–390, July 2005.

[Dat13] Data Offload—Connecting Intelligently, Juniper Research, Hampshire,U.K., White paper. 2013.

[DHHL11] S. Dimatteo, P. Hui, B. Han, and V. O. K. Li. Cellular traffic offloading through wifi networks. In *IEEE Eighth International Conference on Mobile Ad-Hoc and Sensor Systems*, pages 192–201, Octobre 2011.

[DKB11] Peter Dely, Andreas Kassler, and Nico Bayer. Openflow for wireless mesh networks. In *IEEE Proceedings of 20th (ICCCN)*, pages 1–6, 2011.

[DWA14] Xiaoyu Duan, Xianbin Wang, and Auon Muhammad Akhtar. Partial mobile data offloading with load balancing in heterogeneous cellular networks using software-defined networking. In *IEEE 25th Annual International Symposium on PIMRC 2014*, pages 1348–1353, 2014.

[DWY+15] Z. Du, Q. Wu, P. Yang, Y. Xu, J. Wang, and Y. D. Yao. Exploiting user demand diversity in heterogeneous wireless networks. *IEEE Transactions on Wireless Communications*, 14(8):4142–4155, Aug 2015.

[EKR15] J. B. Ernst, S. C. Kremer, and J. J. P. C. Rodrigues. Heterogeneous wireless network rat selection with multiple operators and service contracts. In *IEEE International Conference on Communications (ICC)*, pages 6011–6017, June 2015.

[ELS08] Marc Emmelmann, Tim Langgäertner, and Marcus Sonnemann. System design and implementation of seamless handover support enabling real-time telemetryhighly mobile users. In *Proceedings of the 6th ACM international symposium on Mobility management and wireless access*, pages 1–8. ACM, 2008.

[EWI+16] A. R. Ekti, X. Wang, M. Ismail, E. Serpedin, and K. A. Qaraqe. Joint user association and data-rate allocation in heterogeneous wireless networks. *IEEE Transactions on Vehicular Technology*, 65(9):7403–7414, Sept 2016.

[Fis81] Marshall L Fisher. The lagrangian relaxation method for solving integer programming problems. *Management science*, 27(1):1–18, 1981.

[FRZ14] Nick Feamster, Jennifer Rexford, and Ellen Zegura. The road to sdn: an intellectual history of programmable networks. *ACM SIGCOMM Computer Communication Review*, 44(2):87–98, 2014.

[FZZ12] J. Fan, S. Zhang, and W. Zhou. Energy-friendly network selection in heterogeneous wireless networks. In *IEEE 75th Vehicular Technology Conference (VTC Spring)*, pages 1–5, May 2012.

[GBGF+11] P. Gonzalez-Brevis, J. Gondzio, Y. Fan, H. V. Poor, J. Thompson, I. Krikidis, and P. J. Chung. Base station location optimization for minimal energy consumption in wireless networks. In *IEEE Vehicular Technology Conference (VTC Spring)*, pages 1–5, May 2011.

[GCA+13] Cayley Guimaraes, Daniel Corujo, Rui L Aguiar, Francisco Silva, and Pedro Frosi. Empowering software defined wireless networks through media independent handover management. In *Global Communications Conference, 2013 IEEE*, pages 2204–2209, 2013.

[GG05] Fredrik Gustafsson and Fredrik Gunnarsson. Mobile positioning using wireless networks: possibilities and fundamental limitations based on available wireless network measurements. *IEEE Signal processing magazine*, 22(4):41–53, 2005.

[GJ03] Eva Gustafsson and Annika Jonsson. Always best connected. *IEEE Wireless communications*, 10(1):49–55, 2003.

[HC17] C. C. Hsu and J. M. Chang. Spectrum-energy efficiency optimization for downlink lte-a for heterogeneous networks. *IEEE Transactions on Mobile Computing*, 16(5):1449–1461, May 2017.

[HK71] Michael Held and Richard M Karp. The traveling-salesman problem and minimum spanning trees: Part ii. *Mathematical programming*, 1(1):6–25, 1971.

[HNH05] A. Hasswa, N. Nasser, and H. Hassanein. Generic vertical handoff decision function for heterogeneous wireless. In *Second IFIP International Conference on Wireless and Optical Communications Networks, 2005. WOCN 2005.*, pages 239–243, March 2005.

[HNH06] A. Hasswa, N. Nasser, and H. Hassanein. Tramcar: A context-aware cross-layer architecture for next generation heterogeneous wireless networks. In *IEEE International Conference on Communications*, volume 1, pages 240–245, June 2006.

[HQ14] R. Q. Hu and Y. Qian. An energy efficient and spectrum efficient wireless heterogeneous network framework for 5g systems. *IEEE Communications Magazine*, 52(5):94–101, May 2014.

[HQG+12] Junxian Huang, Feng Qian, Alexandre Gerber, Z Morley Mao, Subhabrata Sen, and Oliver Spatscheck. A close examination of performance and power characteristics of 4g lte networks. In *Proceedings of the 10th international conference on Mobile systems, applications, and services*, pages 225–238. ACM, 2012.

[HYM+17] Q. Han, B. Yang, G. Miao, C. Chen, X. Wang, and X. Guan. Backhaul-aware user association and resource allocation for energy-constrained hetnets. *IEEE Transactions on Vehicular Technology*, 66(1):580–593, Jan 2017.

[IEE] IEEE 802.21.1 Media Independent Services . IEEE 802.21.1 Task Group 21.1 (TG SAUC).

[IEE97] IEEE. Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications, 802.11. http://standards.ieee.org, 1997.

[IEE99a] IEEE. Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications High-speed Physical Layer in the 5 GHz Band, 802.11a. http://standards.ieee.org, 1999.

[IEE99b] IEEE. Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications: Higher-Speed Physical Layer Extension in the 2.4 GHz Band, 802.11b. http://standards.ieee.org, Sep. 1999.

[IEE03] IEEE. Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications: Further Higher Data Rate Extension in the 2.4 GHz Band, 802.11g. http://standards.ieee.org, 2003.

[IEE09] IEEE. Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications Amendment 5: Enhancements for Higher Throughput, 802.11n. http://standards.ieee.org, 2009.

[IEE13] IEEE. Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications–Amendment 4: Enhancements for Very High Throughput for Operation in Bands below 6 GHz, 802.11ac. http://standards.ieee.org, 2013.

[iPa] iPass, iPass Application, Redwood Shores, CA, USA. http://www.ipass.com/.

[JCH84] Raj Jain, Dah-Ming Chiu, and William R Hawe. *A quantitative measure of fairness and discrimination for resource allocation in shared*

computer system. Eastern Research Laboratory, Digital Equipment Corporation Hudson, MA, 1984.

[Jin15] Jin Seek Choi, Hyeong-Ho Lee, Ajung Kim, Kwangho Cho. Revised Draft of Media Independent Handover Service for Software-defined radio access network (SDRAN) Section for IEEE 802.21.1 Media Independent Services Draft Standard. August 2015.

[JK07] Vishv Jeet and Erhan Kutanoglu. Lagrangian relaxation guided problem space search heuristics for generalized assignment problems. *European Journal of Operational Research*, 182(3):1039–1056, 2007.

[JSS14] B. H. Jung, N. O. Song, and D. K. Sung. A network-assisted user-centric wifi-offloading model for maximizing per-user throughput in a heterogeneous network. *IEEE Transactions on Vehicular Technology*, 63(4):1940–1945, May 2014.

[Jun13] Juniper. *Data Offload–connecting intelligently*. White Paper, Juniper Research, Hampshire, U.K., 2013.

[Kar72] Richard M Karp. Reducibility among combinatorial problems. In *Complexity of computer computations*, pages 85–103. Springer, 1972.

[KCL13] S. Kim, S. Choi, and B. G. Lee. A joint algorithm for base station operation and user association in heterogeneous networks. *IEEE Communications Letters*, 17(8):1552–1555, August 2013.

[KK16] P. Y. Kong and G. K. Karagiannidis. Backhaul-aware joint traffic offloading and time fraction allocation for 5g hetnets. *IEEE Transactions on Vehicular Technology*, 65(11):9224–9235, Nov 2016.

[KLD] Slawomir Kuklinski, Yuhong Li, and Khoa Truong Dinh. Handover management in sdn-based mobile networks. In *Globecom Workshops,2014*, pages 194–200. IEEE.

[KYII10] I. Kanno, K. Yamazaki, Y. Ikeda, and H. Ishikawa. Adaptive energy centric radio access selection for vertical handover in heterogeneous networks. In *2010 IEEE Wireless Communication and Networking Conference*, pages 1–6, April 2010.

[LF16] C. H. Liu and K. L. Fong. Fundamentals of the downlink green coverage and energy efficiency in heterogeneous networks. *IEEE Journal on Selected Areas in Communications*, 34(12):3271–3287, Dec 2016.

[LMR12] Li Erran Li, Z Morley Mao, and Jennifer Rexford. Toward software-defined cellular networks. In *European Workshop on EWSDN*, pages 7–12. IEEE, 2012.

[LSK+09] S. Lee, K. Sriram, K. Kim, Y. H. Kim, and N. Golmie. Vertical handoff decision algorithms for providing optimized performance in heterogeneous wireless networks. *IEEE Transactions on Vehicular Technology*, 58(2):865–881, Feb 2009.

[LWH14] C. Liu, P. Whiting, and S. V. Hanly. Joint resource allocation and user association in downlink three-tier heterogeneous networks. In *2014 IEEE Global Communications Conference*, pages 4232–4238, Dec 2014.

[MA06] S. Mohanty and I. F. Akyildiz. A cross-layer (layer 2 + 3) handoff management protocol for next-generation wireless systems. *IEEE Transactions on Mobile Computing*, 5(10):1347–1360, Oct 2006.

[MAAV14] A. Mesodiakaki, F. Adelantado, L. Alonso, and C. Verikoukis. Energy-efficient context-aware user association for outdoor small cell heterogeneous networks. In *IEEE International Conference on Communications (ICC)*, pages 1614–1619, June 2014.

[MAB+08] Nick McKeown, Tom Anderson, Hari Balakrishnan, Guru Parulkar, Larry Peterson, Jennifer Rexford, Scott Shenker, and Jonathan Turner. Openflow: enabling innovation in campus networks. *ACM SIGCOMM Computer Communication Review*, 38(2):69–74, 2008.

[MT81] Silvano Martello and Paolo Toth. An algorithm for the generalized assignment problem. *Operational research*, 81:589–603, 1981.

[MT90] Silvano Martello and Paolo Toth. *Knapsack problems: algorithms and computer implementations*. John Wiley & Sons, Inc., 1990.

[MZ04] J. McNair and Fang Zhu. Vertical handoffs in fourth-generation multinetwork environments. *IEEE Wireless Communications*, 11(3):8–15, June 2004.

[NHH06] N. Nasser, A. Hasswa, and H. Hassanein. Handoffs in fourth generation heterogeneous networks. *IEEE Communications Magazine*, 44(10):96–103, Oct 2006.

[NVACT13] Quoc-Thinh Nguyen-Vuong, Nazim Agoulmine, El Hadi Cherkaoui, and Laura Toni. Multicriteria optimization of access selection to improve the quality of experience in heterogeneous wireless access networks. *IEEE Transactions on Vehicular Technology*, 62(4):1785–1800, 2013.

[O. 09] O. Cabral and F. J. Velez and J. Rodriguez and V. Monteiro and A. Gameiro and N. R. Prasad. Optimal load suitability based RAT selection for HSDPA and IEEE 802.11e. In *2009 1st International Conference on Wireless Communication, Vehicular Technology, Information Theory and Aerospace Electronic Systems Technology*, pages 722–726, May 2009.

[OP98] Tero Ojanpera and Ramjee Prasad. *Wideband CDMA For Third Generation Mobile Communications: Universal Personal Communications*. Artech House, Inc., 1998.

[PCS12] E. Pollakis, R. L. G. Cavalcante, and S. Stańczak. Base station selection for energy efficient network operation with the majorization-minimization algorithm. In *2012 IEEE 13th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, pages 219–223, June 2012.

[PKH+00] K. Pahlavan, P. Krishnamurthy, A. Hatami, M. Ylianttila, J. P. Makela, R. Pichna, and J. Vallstron. Handoff in hybrid mobile data networks. *IEEE Personal Communications*, 7(2):34–47, Apr 2000.

[PW] Florian A Potra and Stephen J Wright. Interior-point methods. *Journal of Computational and Applied Mathematics*, pages 281–302, 2000.

[R+96] Theodore S Rappaport et al. *Wireless communications: principles and practice*, volume 2. Prentice Hall PTR New Jersey, 1996.

[rCC14] S. r. Cho and W. Choi. Coverage and load balancing in heterogeneous cellular networks with minimum cell separation. *IEEE Transactions on Mobile Computing*, 13(9):1955–1966, Sept 2014.

[SA14] S. Sadr and R. S. Adve. Partially-distributed resource allocation in small-cell networks. *IEEE Transactions on Wireless Communications*, 13(12):6851–6862, December 2014.

[Ski09] Steven S Skiena. *The Algorithm Design Manual*. Springer Science & Business Media, p 129, 2009.

[SSZM+12] Lalith Suresh, Julius Schulz-Zander, Ruben Merz, Anja Feldmann, and Teresa Vazao. Towards programmable enterprise wlans with odin. In *Proceedings (HOTSDN*, pages 115–120. ACM, 2012.

[SZ06] Wei Shen and Qing-An Zeng. A novel decision strategy of vertical handoff in overlay wireless networks. In *Fifth IEEE International Symposium on Network Computing and Applications (NCA'06)*, pages 227–230, July 2006.

[SZ08] Wei Shen and Qing-An Zeng. Cost-function-based network selection strategy in integrated wireless and mobile networks. *IEEE Transactions on Vehicular Technology*, 57(6):3778–3788, 2008.

[TK14] Chafika Tata and Michel Kadoch. Efficient priority access to the shared commercial radio with offloading for public safety in lte heterogeneous networks. *Journal of Computer Networks and Communications*, October 2014.

[TL11a] P. TalebiFard and V. C. M. Leung. A dynamic context-aware access network selection for handover in heterogeneous network environments. In *2011 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pages 385–390, April 2011.

[TL11b] Peyman TalebiFard and Victor CM Leung. Context-aware mobility management in heterogeneous network environments. *JoWUA*, 2(2):19–32, 2011.

[TL14] Wenqiang Tang and Qing Liao. An sdn-based approach for load balance in heterogeneous radio access networks. In *In Symposium IEEE SCAC, 2014*, pages 105–108. IEEE, 2014.

[TOM13] R. Trestian, O. Ormond, and G. M. Muntean. Energy-quality-cost tradeoff in a multimedia-based heterogeneous wireless network environment. *IEEE Transactions on Broadcasting*, 59(2):340–357, June 2013.

[TRRL15] S. E. Tajbakhsh, T. Ray, M. C. Reed, and Y. Liu. Joint power control and resource scheduling in wireless heterogeneous networks. In *2015 22nd International Conference on Telecommunications (ICT)*, pages 180–184, April 2015.

[VAN+15] Q. T. Vien, T. Akinbote, H. X. Nguyen, R. Trestian, and O. Gemikonakli. On the coverage and power allocation for downlink in

heterogeneous wireless cellular networks. In *2015 IEEE International Conference on Communications (ICC)*, pages 4641–4646, June 2015.

[WK13] L. Wang and G. S. G. S. Kuo. Mathematical modeling for network selection in heterogeneous wireless networks ; a tutorial. *IEEE Communications Surveys Tutorials*, 15(1):271–292, First 2013.

[WKG99] H. J. Wang, R. H. Katz, and J. Giese. Policy-enabled handoffs across heterogeneous wireless networks. In *Mobile Computing Systems and Applications, 1999. Proceedings. WMCSA '99. Second IEEE Workshop on*, pages 51–60, Feb 1999.

[WLSS08] Y. Wei, X. Li, M. Song, and J. Song. Cooperation radio resource management and adaptive vertical handover in heterogeneous wireless networks. In *2008 Fourth International Conference on Natural Computation*, volume 5, pages 197–201, Oct 2008.

[XGP+12] P. Xue, P. Gong, J. H. Park, D. Park, and D. K. Kim. Radio resource management with proportional rate constraint in the heterogeneous networks. *IEEE Transactions on Wireless Communications*, 11(3):1066–1075, March 2012.

[XPMV14] D. Xenakis, N. Passas, L. Merakos, and C. Verikoukis. Archon: An andsf-assisted energy-efficient vertical handover decision algorithm for the heterogeneous ieee 802.11/lte-advanced network. In *2014 IEEE International Conference on Communications (ICC)*, pages 3166–3171, June 2014.

[YBC05] Xu Yang, J. Bigham, and L. Cuthbert. Resource management for service providers in heterogeneous wireless networks. In *IEEE Wireless Communications and Networking Conference, 2005*, volume 3, pages 1305–1310 Vol. 3, March 2005.

[YH15] Y. Yamazaki and K. Higuchi. Online probabilistic activation control of picocells for system throughput maximization in heterogeneous networks. In *2015 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*, pages 503–507, Nov 2015.

[YKS+10] Kok-Kiong Yap, Masayoshi Kobayashi, Rob Sherwood, Te-Yuan Huang, Michael Chan, Nikhil Handigol, and Nick McKeown. Openroads: Empowering research in mobile networks. *ACM SIGCOMM (CCR)*, 40:125–126, 2010.

[YMP05] Mika Ylianttila, J Mäkelä, and Kaveh Pahlavan. Analysis of handoff in a location-aware vertical multi-access network. *Computer Networks*, 47(2):185–201, 2005.

[YMS08] X. Yan, N. Mani, and Y. A. Sekercioglu. A traveling distance prediction based method to minimize unnecessary handovers from cellular networks to wlans. *IEEE Communications Letters*, 12(1):14–16, January 2008.

[YRC+12] Qiaoyang Ye, Beiyu Rong, Yudong Chen, C. Caramanis, and J. G. Andrews. Towards an optimal user association in heterogeneous cellular networks. In *2012 IEEE Global Communications Conference (GLOBECOM)*, pages 4143–4147, Dec 2012.

[YSK+10] Kok-Kiong Yap, Rob Sherwood, Masayoshi Kobayashi, Te-Yuan Huang, Michael Chan, Nikhil Handigol, Nick McKeown, and Guru Parulkar. Blueprint for introducing innovation into wireless mobile networks. In *Proc. of the 2nd (VIZA), ACM SIGCOMM*, pages 25–32, 2010.

[ZDRB13] A. Zakrzewska, F. D'Andreagiovanni, S. Ruepp, and M. S. Berger. Biobjective optimization of radio access technology selection and resource allocation in heterogeneous wireless networks. In *2013 11th International Symposium and Workshops on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt)*, pages 652–658, May 2013.

[ZHY15] T. Zhou, Y. Huang, and L. Yang. User association with jointly maximising downlink sum rate and minimising uplink sum power for heterogeneous cellular networks. *IET Communications*, 9(2):300–308, 2015.

[ZLS06] Ahmed H Zahran, Ben Liang, and Aladdin Saleh. Signal threshold adaptation for vertical handoff in heterogeneous wireless networks. *Mobile Networks and Applications*, 11(4):625–640, 2006.

[ZM04] Fang Zhu and J. McNair. Optimizations for vertical handoff decision algorithms. In *2004 IEEE Wireless Communications and Networking Conference (IEEE Cat. No.04TH8733)*, volume 2, pages 867–872 Vol.2, March 2004.

[ZM06] Fang Zhu and Janise McNair. Multiservice vertical handoff decision algorithms. *EURASIP Journal on wireless communications and networking*, 2006(2):52–52, 2006.

INSA de RENNES
Service des Formations

AVIS DU JURY SUR LA REPRODUCTION DE LA THESE SOUTENUE

Titre de la thèse:
Optimisation of user association and ressource allocation in hétéregeneous networks

Nom Prénom de l'auteur : ZALGHOUT MOHAMAD

Membres du jury :
- Monsieur CLAVIER Laurent
- Monsieur HELARD Jean-François
- Monsieur CRUSSIERE Matthieu
- Monsieur BEYLOT André-Luc
- Monsieur TERRE Michel
- Monsieur KHALIL Ayman
- Monsieur ABDUL-NABI Samih
- Madame FARAH Joumana

Président du jury : Michel TERRE

Date de la soutenance : 23 Octobre 2017

Reproduction de la these soutenue

☒ Thèse pouvant être reproduite en l'état
☐ Thèse pouvant être reproduite après corrections suggérées

Fait à Rennes, le 23 Octobre 2017

Signature du président de jury

Le Directeur,

M'hamed DRISSI

Résumé

Aujourd'hui, l'extension des exigences du trafic de données sans fil dépasse le taux de croissance de la capacité des nouvelles technologies d'accès sans fil. Par conséquent, les réseaux sans fil mobiles de la future génération proposent des architectures hétérogènes, généralement appelées réseaux sans fil hétérogènes (HWN). HWN se caractérisent par l'intégration des réseaux cellulaires et des réseaux locaux sans fil (WLAN) pour répondre aux besoins des utilisateurs et améliorer la capacité du système. En fait, l'intégration de différents types de technologies d'accès sans fil dans HWN offre des choix flexibles pour que les utilisateurs soient associés au réseau qui répond le mieux à leurs besoins. Dans ce contexte, cette thèse traite le problème d'association d'utilisateurs et le problème d'allocation de ressources dans un système sans fil hétérogène basé sur des points d'accès Wi-Fi intégrés et des stations de base LTE.

Les contributions de cette thèse pourraient être divisées en trois parties principales. Dans la première partie, un nouveau problème d'association d'utilisateurs et d'optimisation de l'allocation des ressources est formulé pour maximiser la satisfaction globale des utilisateurs dans le système. La satisfaction de l'utilisateur est basée sur une fonction de profit pondérée qui vise à améliorer la puissance relative du signal reçu et la diminution de la consommation d'énergie des terminaux mobiles (MT). Étant donné qu'un MT n'est autorisé à être associé qu'à un seul réseau à la fois, le problème d'optimisation formulé est binaire avec une complexité NP complète. Ensuite, plusieurs solutions centralisées avec une complexité à temps polynomial sont proposées pour résoudre le problème formulé. Les solutions proposées sont basées sur des approches heuristiques et sur la relaxation continue du problème d'optimisation binaire formulé.

La deuxième partie de la thèse vise à fournir une solution distribuée pour le problème formulé. La solution distribuée proposée déploie la technique de détente lagrangienne pour convertir le problème global formulé en plusieurs problèmes de Knapsack distribués, chaque réseau traite son problème Knapsack correspondant. La méthode de sous gradient est utilisée pour trouver les multiplicateurs lagrangiens optimaux ou sous optimaux.

Enfin, la troisième partie de la thèse étudie de nouvelles perspectives de la formulation du problème d'optimisation et ses solutions centralisées et distribuées correspondantes. Un problème d'association d'utilisateurs et d'allocation de ressources basé sur la priorité est formulé. Le problème est ensuite réduit en plusieurs problèmes résolus à l'aide des solutions proposées réparties et centralisées. En outre, une nouvelle solution de maximisation de l'efficacité énergétique est proposée en modifiant les objectifs du problème d'optimisation originalement formulé.

Abstract

It is indicated that the expansion of the wireless data traffic requirements exceeds the capacity growth rate of new wireless access technologies. Therefore, next-generation mobile wireless networks are moving toward heterogeneous architectures usually referred to as heterogeneous wireless networks (HWNs). HWNs are usually characterized by the integration of cellular networks and wireless local area networks (WLANs) to meet user requirements and enhance system capacity. In fact, integrating different types of wireless access technologies in HWNs provides flexible choices for users to be associated with the network that best satisfies their needs. In this context, this thesis discusses the user association and downlink resource allocation problem in a heterogeneous wireless system that is based on integrated Wi-Fi access points (APs) and long-term evolution (LTE) base stations (BSs).

The contributions of this thesis could be divided into three main parts. In the first part, a novel user association and resource allocation optimization problem is formulated to maximize the overall user satisfaction in the system. The user satisfaction is based on a weighted profit function that aims at enhancing the relative received signal strength and decreasing the power consumption of mobile terminals (MTs). Since a MT is only allowed to be associated with a single network at a time, the formulated optimization problem is binary with an NP-complete complexity. Then, multiple centralized solutions with polynomial-time complexities are proposed to solve the formulated problem. The proposed centralized solutions are based on heuristic approaches and on the continuous relaxation of the formulated binary optimization problem.

The second part of the thesis aims at providing a distributed solution for the formulated problem. The proposed distributed solution deploys the Lagrangian relaxation technique in order to convert the global formulated problem into multiple distributed Knapsack problems, each network processes its corresponding Knapsack problem. The sub-gradient method is used in order to find the optimal, or near optimal, Lagrangian multipliers.

Finally, the third part of the thesis studies new perspectives of the formulated optimization problem and its corresponding centralized and distributed solutions. Mainly, a generalized priority-aware user association and resource allocation problem is formulated. The priority-aware problem is then reduced into multiple problems that are solved using the proposed centralized and distributed solutions. Moreover, a novel power efficiency maximization solution is proposed by altering the objectives of the main formulated optimization problem.

INSA RENNES

N° d'ordre : 17ISAR 28 / D17 - 28

Institut National des Sciences Appliquées de Rennes
20, Avenue des Buttes de Coësmes - CS 14315 - F-35043 Rennes Cedex
Tél : 02 23 23 82 00 – Fax : 02 23 23 83 96