Resource allocation techniques for non-orthogonal multiple access systems

Marie Rita Hojeij

UNIVERSITE BRETAGNE LOIRE / MATHSTIC

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

THESE DE DOCTORAT DE

L'ÉCOLE NATIONALE SUPERIEURE MINES-TELECOM ATLANTIQUE
BRETAGNE PAYS DE LA LOIRE - IMT ATLANTIQUE
COMUE UNIVERSITE BRETAGNE LOIRE

ECOLE DOCTORALE N° 601
Mathématiques et Sciences et Technologies de l'Information et de la Communication

Spécialité : *Télécommunications*

Par

Marie-Rita Hojeij

« Resource allocation techniques for non-orthogonal multiple access systems »

Thèse présentée et soutenue à IMT Atlantique, le 30 mai 2018
Unité de recherche : Lab-STICC
Thèse N° : 2018IMTA0085

Rapporteurs avant soutenance :

Didier LE RUYET	Professeur, Conservatoire National des Arts et Métiers - Paris
Georges KADDOUM	Professeur, Ecole de Technologie Supérieure - Canada

Composition du Jury :

Président :	Didier LE RUYET	Professeur, Conservatoire National des Arts et Métiers - Paris
Examinateurs :	Georges KADDOUM	Professeur, Ecole de Technologie Supérieure - Canada
	Mathieu CRUSSIERE	Maître de conférences (HDR), Insa-Rennes
	Charbel ABDEL NOUR	Maître de conférences, IMT Atlantique
Directeur de thèse :	Catherine DOUILLARD	Professeur, IMT Atlantique
Co-directeur de thèse :	Joumana FARAH	Professeur, Université Libanaise

Techniques d'allocation de ressources pour les systèmes à accès multiple non orthogonal

Résumé en français

Avec l'émergence rapide des applications Internet, il est prévu que le trafic mobile mondial augmente de huit fois entre fin 2018 et 2022. En même temps, les futurs systèmes de communication se devront aussi d'améliorer l'efficacité spectrale des transmissions, le temps de latence et l'équité entre utilisateurs. À cette fin, une technique d'accès multiple non orthogonal (NOMA) a été récemment proposée comme un candidat prometteur pour les futurs accès radio. La technique NOMA est basée sur un nouveau domaine de multiplexage, le domaine des puissances. Elle permet la cohabitation de deux ou plusieurs utilisateurs par sous-porteuse ou sous-bande de fréquence.

Cette thèse aborde plusieurs problèmes liés à l'allocation de ressources basée sur NOMA afin d'améliorer les performances du réseau en termes d'efficacité spectrale, de débit et/ou d'équité entre utilisateurs. Dans ce sens, des solutions théoriques et algorithmiques sont proposées et des résultats numériques sont obtenus afin de valider les solutions et de vérifier la capacité des algorithmes proposés à atteindre des performances optimales ou sous-optimales.

Une étude bibliographique des différentes techniques d'allocation de ressources est présentée dans le premier chapitre. L'allocation de ressources pour le cas orthogonal et pour le cas non-orthogonal a été traitée.

Nous avons tout d'abord présenté l'état de l'art des techniques d'allocation de ressources dans les réseaux LTE (*long term evolution*) et LTE-Advanced. Deux méthodes d'accès multiples orthogonaux, OFDMA (*orthogonal frequency-division multiple access*) et SC-FDMA (*single-carrier frequency division multiple access*), ont été examinées car elles exploitent la granularité de fréquence flexible, répondent aux diverses exigences de qualité de service (QoS) des utilisateurs et atteignent une efficacité spectrale élevée. Les techniques OFDMA et SC-FDMA sont utilisées, respectivement, pour la transmission en liaison descendante et en liaison montante. L'OFDMA est robuste à la présence d'évanouissements dus aux trajets multiples et est résistant aux évanouissements sélectifs en fréquence. SC-FDMA a des performances similaires à OFDMA, mais avec un ratio entre la puissance moyenne et la puissance maximale (PAPR) inférieur, et diffère également de OFDMA de par

le fait que, dans SC-FDMA, les symboles d'information sont transmis séquentiellement plutôt qu'en parallèle à l'aide de sous-porteuses orthogonales. Ce mécanisme offre un PAPR inférieur en réduisant les fluctuations d'enveloppe de la forme d'onde du signal transmis, ce qui rend le SC-FDMA mieux adapté à la transmission sur la liaison montante, en particulier pour les équipements à faible coût dont la puissance est limitée.

Dans la suite de ce chapitre, plusieurs techniques d'accès multiple non orthogonales ont été examinées afin de répondre aux exigences strictes de la 5G, telles que la technique *sparse code multiple access* (SCMA) qui est considérée comme une forme à porteuses multiples de NOMA, la technique *multiuser superposition transmission* (MUST) pour LTE qui examine la transmission non orthogonale des utilisateurs et la proposition de récepteurs avancés, ainsi que la méthode NOMA conçue pour le domaine de puissance, qui constitue le point d'intérêt principal de notre mémoire. NOMA est basée sur un codage par superposition (elle superpose plusieurs signaux des équipements utilisateurs ou UE) du côté émetteur et sur une méthode d'annulation d'interférence successive (SIC) qui permet la séparation et le décodage des signaux multi-utilisateurs du côté récepteur. NOMA permet à plusieurs utilisateurs de partager la même sous-bande tout en fournissant à chaque utilisateur une quantité appropriée de puissance. Pour mieux comprendre le principe de NOMA, une description détaillée de son concept, comprenant le multiplexage d'utilisateurs au niveau de l'émetteur de la station de base (BS) et la séparation des signaux au niveau du récepteur a été élaborée dans le premier chapitre. Ensuite, l'allocation de ressources dans les réseaux cellulaires à liaison descendante a été détaillée, car c'est le point d'intéret de notre étude. L'allocation de ressources est considérée comme une solution à l'optimisation des ressources limitées, en fréquence, puissance et temps, afin d'obtenir les performances souhaitées tout en tenant compte de contraintes réalistes. Et comme le problème d'allocation de ressources radio consiste à optimiser une fonction d'utilité tout en prenant en compte un ensemble de contraintes, nous avons présenté dans le chapitre 1 plusieurs fonctions d'utilité classiques, y compris celles utilisées dans la thèse. Les fonctions d'utilité présentées sont la maximisation du débit, la probabilité de succès, la somme des débits pondérés et l'équité, ainsi que le nombre minimal de sous-bandes attribuées. L'optimisation des ressources radio pour un système OFDM et un système NOMA a également été abordée dans le premier chapitre afin de donner un aperçu des techniques d'allocation de puissance existantes dans la littérature afin de proposer de nouvelles techniques adaptatives dans les autres chapitres. L'ordonnanceur à équité proportionnelle, dit *scheduler* PF, a été détaillé puisqu'il est le plus couramment utilisé dans la majorité des travaux qui traitent de NOMA. Le PF est connu pour offrir un bon compromis

entre la maximization du débit et de l'équité. L'allocation de puissance a également été abordée dans le chapitre 1, et nous avons mis l'accent sur le principe de *waterfilling*, étant donné qu'il sera utilisé dans nos propositions des chapitres suivants.

Dans le deuxième chapitre, un nouveau principe de traitement de la bande passante et de l'allocation de puissance dans le cadre d'un scénario d'accès multiple non orthogonal a été introduit. Il vise à minimiser l'utilisation du spectre tout en satisfaisant les utilisateurs quant aux débits demandés. Dans ce sens et pour cette approche proposée, plusieurs problèmes de conception ont fait l'objet d'une analyse approfondie. La sélection de l'utilisateur est le premier problème de conception abordé dans ce chapitre. La sélection est effectuée en accordant une priorité plus élevée aux utilisateurs dont le débit est eloigné de leur débit cible. Après cela, nous avons discuté l'allocation des sous-bandes et le couplage des utilisateurs. Le choix du couplage des utilisateurs est l'un des problèmes de conception les plus critiques que nous avons abordés dans ce chapitre car les performances de NOMA sont très sensibles aux gains de canal des utilisateurs choisis pour être placés ensemble sur la même sous-bande. Deux options de couplage ont été étudiées et celle qui choisit les utilisateurs avec la différence de gain de canal la plus élevée s'est avérée mieux adapté à NOMA. L'allocation de puissance est un autre problème de conception traité dans le deuxième chapitre. Une allocation de puissance statique entre sous-bandes a été d'abord considérée, pour laquelle la puissance est également répartie entre les sous-bandes. Cependant, le débit pouvant ainsi être obtenu est pénalisé du fait que le principe du *waterfilling* n'est pas appliqué. Pour cette raison, une allocation optimale de puissance entre les sous-bandes basée sur le *waterfilling* et prenant en compte les gains de canal des deux utilisateurs appariés au sein de chaque sous-bande a été proposée. De plus, plusieurs stratégies de pondération dans l'optimisation de la somme des débits ont également été développées. Ces propositions optimales effectuent une allocation de puissance conjointe inter et intra-sous-bande qui prend en compte les gains de canal de tous les utilisateurs couplés et se révèle être relativement complexe à mettre en œuvre. Par conséquent, une solution sous-optimale, où la puissance est attribuée à des utilisateurs en deux étapes consécutives, constituées de l'allocation inter-sous-bande puis intra-sous-bande de puissance, a été proposée. Dans la première étape, le gain de canal le plus élevé est pris en compte dans chaque sous-bande afin de déterminer la quantité de puissance à attribuer à chaque sous-bande. Au cours de la deuxième étape, la puissance attribuée à chaque sous-bande est répartie entre les utilisateurs co-localisés en se basant sur les techniques d'allocation de puissance intra-sous-bande. En ce sens, plusieurs techniques d'allocation de

puissance intra-sous-bande ont été étudiées. En plus de l'allocation de puissance, nous avons aussi testé si l'amélioration de l'efficacité spectrale lors de l'application de NOMA était systématique. En effet, nous avons constaté que parfois, sur certaines sous-bandes et pour certains utilisateurs co-localisés, l'utilisation de NOMA n'est pas bénéfique. Ainsi, nous avons proposé une métrique de décision à tester sur chaque sous-bande afin de vérifier s'il est préférable de passer à une signalisation orthogonale pour cette sous-bande. Ensuite, un mécanisme de contrôle *a posteriori* a été appliqué afin d'ajuster la puissance de certaines sous-bandes pour lesquelles le débit des utilisateurs attribués a dépassé le débit demandé. Afin d'évaluer la performance des contributions proposées, nous avons réalisé une analyse comparative considérant un système basé sur une signalisation orthogonale pour lequel chaque sous-bande ne peut être attribuée qu'à un seul utilisateur. Nous avons considéré quatre indicateurs de performance au niveau du système afin de procéder à l'évaluation : la capacité du système en termes de débit, le débit des utilisateurs situés au bord de la cellule, la probabilité de succès (le succès est atteint lorsque tous les utilisateurs ont atteint leur débits demandés) et la quantité de bande passante utilisée. Les résultats de simulation ont montré que le cadre proposé permettait une augmentation significative de l'efficacité spectrale et de la probabilité de succès, en particulier par rapport à un système basé uniquement sur une signalisation orthogonale ou non orthogonale. En outre, l'allocation de puissance conjointe inter et intra-sous-bande, obtenue en résolvant numériquement un problème d'allocation optimisée, procure un gain de performance substantiel par rapport aux solutions sous-optimales. De plus, l'adoption de pondérations appropriées dans la métrique de somme des débits montre des améliorations prometteuses en termes d'efficacité spectrale, de débit utilisateur en périphérie de cellule et du nombre nécessaire de SICs par utilisateur. De telles stratégies de pondération peuvent être utilisées dans plusieurs applications pratiques qui nécessitent de fournir des priorités aux utilisateurs.

Les résultats obtenus dans le deuxième chapitre ont montré le grand intérêt de NOMA par rapport à l'accès multiple orthogonal (OMA) dans un contexte où la bande passante utilisée doit être minimisée et la quantité maximale de puissance délivrée au niveau de la station de base doit être respectée. L'équité a été obtenue en fixant des taux cibles pour les utilisateurs. Toutefois, cela peut s'avérer inapproprié pour certaines applications ou services où les utilisateurs ne sont pas censés se voir garantir des débits de données fixes. Par conséquent, au chapitre 3, nous avons ciblé l'optimisation du débit et de l'équité dans un contexte où la totalité de la bande passante peut être utilisée et où il n'y a pas de contrainte de débits cibles.

Pour ce faire, les performances du *scheduler* PF ont d'abord été évaluées, ce *scheduler* étant le plus utilisé dans la majorité des travaux traitant NOMA.

Le *scheduler* PF a pour objectif de maximiser la somme logarithmique des débits des utilisateurs afin d'assurer un équilibre entre l'équité entre utilisateurs et le débit total. Ce faisant, le *scheduler* PF vise l'équité à long terme car, parmi les éléments d'optimisation d'une métrique PF classique, l'indice d'équité est calculé *a priori* et fixé avant la répartition réelle. Cependant, atteindre l'équité à court terme et une convergence rapide vers les performances requises sont des questions importantes à traiter dans les normes mobiles à venir. En outre, le *scheduler* PF teste tous les utilisateurs candidats pour chaque sous-bande, de sorte qu'un utilisateur peut être sélectionné plusieurs fois et peut se voir attribuer plusieurs sous-bandes, mais il peut également arriver qu'un utilisateur ne se voit attribuer aucune sous-bande lorsque son débit historique est élevé. Dans un tel cas, aucun débit ne sera attribué à cet utilisateur pour plusieurs créneaux d'allocation. Ce comportement peut être très problématique dans certaines applications, notamment celles nécessitant une qualité d'expérience (QoE) quasi constante, telles que les transmissions multimédia. En ce qui concerne les exigences mentionnées, nous avons proposé dans le chapitre 3 plusieurs modifications qui visent à augmenter la capacité de la cellule, à améliorer l'équité à long et à court terme entre les utilisateurs, à améliorer la qualité d'experience et à offrir différents niveaux de qualité de service (QoS). Pour ce faire, la métrique du PF a été modifiée en introduisant des poids qui tiennent compte de l'allocation en cours. Par conséquent, la priorité donnée à chaque utilisateur a été basée non seulement sur son débit historique, mais également sur son débit total actuel atteint (débit atteint pendant l'allocation en cours). L'objectif principal de la version pondérée est d'assurer l'équité entre les utilisateurs dans chaque allocation. Cela permettra d'éviter l'occurrence d'un débit de données de transmission nul pour n'importe quel utilisateur à n'importe quelle échelle de temps. Au chapitre 3, nous avons également envisagé d'appliquer la métrique pondérée proposée à un système OMA au lieu de NOMA, afin d'évaluer la contribution de NOMA dans notre cadre d'étude. Comme certaines applications souhaitent fournir différents niveaux de QoS, ce qui est impossible à obtenir avec la métrique classique PF, des modifications de la métrique PF ont également été développées afin de satisfaire les contraintes de QoS en fournissant plusieurs niveaux de priorité aux différents ensembles d'utilisateurs. Les résultats de simulation ont montré que les métriques pondérées proposées présentent une meilleure performance que le *scheduler* PF classique en termes de débit total et d'équité entre utilisateurs. La QoE a également été évaluée et s'est révélée améliorée grâce aux métriques pondérées proposées.

Dans le *scheduler* PF, l'attribution de chaque sous-bande nécessite l'estimation d'une métrique pour chaque candidat utilisateur possible (cas OMA) ou pour chaque ensemble de candidats (cas NOMA). Ces estimations nécessitent des calculs de débits qui, à leur tour, nécessitent que les niveaux de puissance soient prédits sur la sous-bande considérée, pour chaque candidat. Cela devient encore plus problématique à mesure que le nombre de sous-bandes et/ou d'utilisateurs augmente. Pour toutes ces raisons, dans tous les travaux antérieurs traitant l'allocation de ressource en utilisant le *scheduler* PF, on supposait une répartition de puissance égale entre sous-bandes afin de contourner le problème de l'estimation de puissance, empêchant ainsi le système d'optimiser la puissance entre sous-bandes. En ce sens, des modifications de la métrique PF au niveau de l'allocation de puissance ont également été proposées au chapitre 3 afin d'améliorer encore les performances. Puisque le principe du *waterfilling* dans sa formulation classique ne peut pas être appliqué directement au *scheduler* PF, le chapitre 3 vise à proposer une technique basée sur le *waterfilling* à faible complexité qui permet l'incorporation de la solution sous-optimale basée sur le *waterfilling* dans le PF, afin d'en améliorer les performances. Les résultats de simulation montrent que la technique proposée permet une augmentation du débit total des utilisateurs et de l'équité du système, par rapport à un système basé sur OMA et/ou par rapport à un système NOMA prenant en compte une répartition de puissance égale entre les sous-bandes.

Le débit total de la cellule, le débit des utilisateurs situés au bord de la cellule et l'équité dépendent fortement de la façon dont la puissance et la bande passante sont allouées aux utilisateurs par le *scheduler*, ainsi que le couplage des utilisateurs par le multiplexage NOMA. Il est largement admis que le débit de la cellule est maximal lorsque la différence de gain de canal entre utilisateurs multiplexés est élevée, bien que cette affirmation soit mise en doute dans quelques articles. Par conséquent, dans le chapitre 4, nous cherchons d'abord à vérifier si l'augmentation de la différence de gain de canal est toujours en faveur du débit atteint par NOMA, lorsque différentes techniques d'attribution de puissance intra-sous-bande sont envisagées. Ensuite, guidés par les résultats obtenus, nous avons mis au point plusieurs techniques d'allocation de ressources visant à être pleinement adaptées aux systèmes NOMA. Pour ce faire, l'impact de la différence de gain de canal entre utilisateurs multiplexés sur le débit obtenu pour différentes techniques d'allocation de puissance intra-sous-bande a été analysé. À cette fin, nous nous sommes concentrés sur deux schémas d'allocation de puissance intra-sous-bande: FPA (*fixed power allocation*) et FTPA (*fractional transmit power allocation*). Comprendre le fonctionnement des technques FPA et FTPA nous a aidé à trouver

les meilleurs couples d'utilisateurs à multiplexer par NOMA et à vérifier si l'augmentation de la différence de gain de canal est toujours en faveur de NOMA en termes de débit atteint. Pour mener à bien cette analyse, nous avons considéré le cas où la différence de gain de canal entre utilisateurs multiplexés augmente. Nous avons découvert que lorsque le FPA est utilisé, le débit des utilisateurs couplés est une fonction croissante de la différence de gain de canal. Cependant, dans le cas de FTPA, la relation entre le débit total atteint sur une sous-bande et la différence de gain de canal entre utilisateurs appariés est imprévisible et peut évoluer dans le même sens ou dans le sens opposé. Une vérification supplémentaire a été effectuée pour vérifier si le débit des utilisateurs couplés dépend de la différence de gain de canal lorsque seul l'un des gains de canal des deux utilisateurs varie. Les résultats montrent que le comportement des débits des utilisateurs planifiés est asymétrique avec la différence de gain de canal. Cette observation vient souligner l'importance de la prise en compte du débit et de la dépendance du gain de canal lors du couplage des utilisateurs.

Motivés par cette observation, nous avons également proposé au chapitre 4 une nouvelle technique d'allocation de puissance visant à améliorer le débit total de la cellule, à optimiser l'équité à long terme, à assurer l'équité entre les utilisateurs au cours de chaque allocation et à réduire le temps de convergence vers une performance d'équité requise (cela peut être particulièrement utile lorsqu'un débit utilisateur quasi constant est visé). Cela a été conçu en créant un système compatible avec les contraintes de faible temps de latence, en développant de nouvelles métriques de *scheduling* pouvant être facilement associées à des techniques d'allocation de puissance inégale telles que le *waterfilling*, et en réduisant le nombre de paires d'utilisateurs testés pour chaque attribution de sous-bande. Pour ce faire, la sélection de l'utilisateur et les allocations de sous-bandes ont été effectuées de manière à créer des conditions propices à NOMA. Nous avons ensuite proposé deux réalisations différentes de la métrique de couplage. La première métrique de couplage dénommée FTFMM (*flexible throughput vs fairness maximization metric*) permet de traiter les applications pour lesquelles la maximisation du débit par rapport à l'équité est essentielle. FTFMM vise à trouver un équilibre flexible et contrôlable entre le débit et l'équité. La flexibilité est introduite par le biais de paramètres qui peuvent être choisis pour favoriser le débit au détriment de l'équité et *vice versa*. Ces paramètres sont définis *offline* en fonction des besoins de l'application considérée. La deuxième métrique de couplage, la métrique de maximisation d'équité, désignée par FMM (*fairness maximization metric*), est proposée pour cibler les applications dans lesquelles un niveau élevé d'équité est requis. En ce sens, elle tend à minimiser la différence de débit entre les utilisateurs dans chaque allocation, tout en maximisant le débit

de données moyen des utilisateurs. Par conséquent, FMM vise à fournir des débits de données presque identiques pour tous les utilisateurs. La complexité calculatoire est évaluée pour chaque méthode de couplage et comparée à la métrique PF classique. Les résultats de simulation montrent que les nouvelles mesures proposées surpassent le *scheduler* PF classique en termes de capacité du système, d'équité à long et à court terme et de QoE.

Les infrastructures sans fil sont traditionnellement caractérisées par deux types de modes de transmission distincts: des modes fournissant une connectivité *unicast* point à point (par exemple, les systèmes cellulaires) et d'autres faisant appel à des transmissions *multicast* ou *broadcast* point à multipoint (par exemple, les systèmes de diffusion vidéo numériques ou DVB). De plus en plus, la demande de services mixtes augmente et une transmission hybride *broadband* et *broadcast* commence à faire sens. En effet, les fournisseurs de services sans fil offrent désormais plusieurs services simultanés tels que la diffusion de télévision numérique et l'accès Internet à large bande. Les services *broadcast* fournis sont par exemple les mises à jour logicielles, les messages d'urgence, les avertissements publics, etc. La technique NOMA ayant démontré qu'elle permettait une augmentation significative des débits, un système hybride fournissant des services *broadband* et *broadcast* en utilisant simultanément le schéma NOMA au-dessus de la couche OFDMA a été proposé et évalué au chapitre 5. Autrement dit, les signaux *broadband* et *broadcast* sont autorisés à être transmis simultanément sur les mêmes bandes de fréquence, et pour aller plus loin, nous avons proposé différentes techniques permettant de transmettre les messages *broadband* et les messages *broadcast* sur la même sous-bande en multiplexant leurs signaux dans le domaine de puissance en utilisant NOMA. Le récepteur SIC a pour but d'assurer une réception sans interférence des signaux *broadband*. Dans la plupart des travaux de la littérature, la transmission *broadcast* est traitée dans une étape *a priori*, avec une priorité plus élevée. Ainsi, la transmission *broadband* n'est pas assurée si toutes les ressources (fréquence et puissance) sont consommées par la transmission *broadcast*. Ceci entraîne une réallocation de toutes les sous-bandes à chaque fois qu'un message de diffusion doit être envoyé à partir de la station de base. Par conséquent, dans le chapitre 5, nous avons considéré deux familles de techniques pour le *broadcast*, soit *a priori* soit *a posteriori*. La première famille consiste à effectuer le *broadcast* dans une première étape, jusqu'à ce que le débit *broadcast* cible soit atteint, puis à appliquer la transmission *broadband* aux sous-bandes restantes. La seconde famille de techniques considère que *broadcast* est appliquée une fois que toutes les sous-bandes sont allouées pour la transmission *broadband*. Pour des raisons de simplicité et de

bonnes performances, nous avons supposé que les sous-bandes étaient allouées en se basant sur le *scheduler* PF où la puissance est répartie de manière égale entre les sous-bandes et la technique FTPA est utilisé pour l'allocation de puissance intra-sous-bande. L'action *a posteriori* est effectuée en choisissant des sous-bandes appropriées pour la transmission *broadcast* qui sont extraites de l'ensemble des sous-bandes dédiées aux services *broadband*, en fonction de paramètres de décision spécifiques. Deux métriques de décision différentes ont été proposées, la métrique de minimisation des pertes en broadband (BLMM) et la métrique d'allocation basée sur le canal (CBAM). La première est conçue de manière à minimiser la perte de débit *broadband*, à maximiser le débit de messages *broadcast* et à éviter de récupérer des sous-bandes pour la diffusion à partir d'utilisateurs ayant des débits historiques faibles. La seconde mesure de décision se concentre sur la maximisation du débit de diffusion plutôt que sur la minimisation de la perte de débit sur le haut débit.

L'action *a priori* proposée effectue l'attribution en *broadcast* dans une première étape, jusqu'à ce que le débit *broadcast* cible soit atteint, puis l'attribution *broadband* utilise les sous-bandes restantes. Dans ce cas, l'allocation de puissance est optimisée lors de l'étape d'allocation de *broadcast*. Pour aller plus loin, l'objectif du chapitre 5 étant de permettre une diffusion hybride monodiffusion sur une plate-forme mono-fréquence, nous avons proposé une attribution hybride *broadband* sur des sous-bandes choisies pour *broadcast*. Pour ce faire, nous avons proposé d'ajouter un signal large bande à multiplexer avec le message de *braodcast* commun en utilisant NOMA. Cependant, le choix du signal large bande à multiplexer avec le message de *broadcast*, c'est-à-dire l'utilisateur pour lequel un accès large bande supplémentaire est fourni sur une sous-bande de diffusion, n'est pas trivial. Il doit être choisi de manière à ce que tous les utilisateurs désirant lire le message commun n'aient pas besoin de décoder le signal privé à large bande. Les résultats de simulation ont montré que le système hybride assure de meilleures performances, en termes de débit atteint et de l'équité à long terme, par rapport aux systèmes classiques précédents. Il présente également une légère perte de performances par rapport au *scheduler* PF classique à large bande. L'incorporation d'une technique d'allocation de puissance personnalisée, pour les signaux broadband et broadcast, basée sur le principe du *waterfilling* a donné les meilleures performances.

Le manuscrit se conclue par une récapitulation des résultats obtenus et une proposition de travaux futurs avec un plan à court et à long terme.

Contents

List of Figures

List of Tables

ABSTRACT

Resource allocation techniques for non-orthogonal multiple access systems

Marie-Rita Hojeij

Department Électronique, Telecom - Bretagne

30 May 2018

With the proliferation of Internet applications, total mobile traffic is expected to increase by 8 times between the end of 2018 till 2022. At the same time, future communication systems are required to further enhance system efficiency, latency, and user fairness. To this end, Non-orthogonal multiple access (NOMA) has recently emerged as a promising candidate for future radio access. By exploiting an additional multiplexing domain, the power domain, NOMA allows the cohabitation of two or more users per subcarrier, based on the principle of signal superposition. This dissertation addresses several radio resource allocation problems in mobile communication systems using NOMA, in order to improve network performance in terms of spectral efficiency, throughput, and/or fairness. Theoretical analysis and algorithmic solutions are derived. Numerical results are obtained to validate our theoretical findings and demonstrate the algorithms ability of reaching optimal or sub-optimal performance.

After a tour of state of the art review of power and resource allocation techniques for NOMA provided in the first chapter, the second chapter of this thesis investigates several new strategies for the allocation of radio resources (bandwidth and transmit power) using the NOMA principle, where a minimization of the total amount of used bandwidth is targeted. While taking into account various design issues, we propose and compare several optimum and suboptimum power allocation schemes. These are jointly implemented with multiple user scheduling strategies. Also, to increase the total achieved system throughput, a hybrid orthogonal-non orthogonal scheme is introduced. This hybrid scheme enables a dynamic switching to orthogonal signaling whenever the non-orthogonal cohabitation in the power domain does not improve the achieved data rate per subband. Extensive simulation results show that the proposed strategies for resource allocation can improve both the spectral efficiency and the cell-edge user throughput, especially when compared to schemes employing either orthogonal signaling or NOMA with static inter-subband power allocation. They also prove to be robust in the context of crowded areas.

A context where the total bandwidth is available to be used should also be studied. Therefore, we investigate in the third chapter the performance of the proportional fairness (PF) scheduler, and we propose modifications to the PF, at the level of user scheduling and power allocation. In this sense, several weighted versions of the PF scheduling metric are developed. The proposed schemes introduce weights that adapt the classical PF metric to the NOMA scenario, therefore improving performance indicators and enabling new services. In addition to the weighted metrics, a low-complexity waterfilling-based power allocation (PA) technique, incorporated within the PF, is also proposed. The aim of the proposed joint PA and scheduling scheme is to maximize the achieved average throughput through a quasi-optimal repartition of the transmit power among subbands, while guaranteeing a high level of fairness in resource allocation. The distinguishing value of the proposals resides in their ability to improve long term fairness and total system throughput while achieving a high level of fairness in every scheduling slot.

In the fourth chapter, we first aim to verify if the common belief that the increase in the channel gain difference between scheduled users is always in favor of NOMAs achieved

throughput when using NOMA, when different intra-subband power allocation techniques are considered. Then, guided by the obtained results, two new user pairing metrics are designed, incorporating a non-uniform inter-subband power allocation scheme. The first proposed metric introduces flexibility in throughput/fairness maximization, whereas the second metric aims at providing a high level of fairness among users. The two proposed metrics show enhancements at the level of system capacity, user fairness, and computational complexity.

In the last chapter different techniques that allow a hybrid broadcast/broadband transmission on the same frequency platform are proposed and compared with the state of the art. For this purpose, we consider two families of techniques, *a-priori* and *a-posteriori* multicasting. The proposed techniques aim to minimize the number of used subbands for broadcasting and to maximize the broadband achieved rate, while targeting guaranteeing the reception of the broadcast message by all active users with a 100% success rate. Simulation results confirm the claimed objectives making the proposal an appealing alternative for a hybrid broadcast/broadband system.

DEDICATION

To my father's soul...

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deep and sincere gratitude to my director at the Lebanese university, Prof. Joumana Farah, for providing me excellent guidance and continuous support during theses years. I have learned many valuable lessons from such an outstanding researcher who always selflessly shares her research experience and expertise with me.

I would like also to reveal my deepest appreciation to my director at IMT Atlantique, Prof. Catherine douillard, an exemplary researcher and human being, who offered to me irreplaceable support. Her guidance was a catalyst for the completion of this dissertation. The knowledge and the attitude on research I have learned from her will benefit me a lot in my future career development.

I would like to thank my supervisor at IMT Atlantique, Assoc. Prof. Charbel Abdel Nour for his invaluable contributions to my scientific development. He always encouraged me to move forward, develop myself and take the further step. Without his comments and contributions the work of this thesis could not be achieved.

I would like to take this opportunity to show my gratitude to every person's willingness in helping me become what I am today especially to my university, the Holy Spirit University of Kaslik. I would like also to show my sincere gratitude and appreciation to the person who has always had my back, Father Marwan Azar, former dean of the faculty of engineering, for his invaluable support during the past years.

Last but not least, I would like to thank my family. I am deeply grateful to my mother, who helped me to become what I am today.

"Success is a staircase, not a doorway."

ABBREVIATIONS

3GPP	The Third Generation Partnership Project
5G	The Fifth Generation
BS	Base Station
CSI	Channel State Information
ETU	Extended Typical Urban
FDM	Frequency Division Multiplexing
FFR	Fractional Frequency Reuse
FPA	Fixed Power Allocation
FSPA	Full Search Power Allocation
FTPA	Fractional Transmit Power Allocation
FTPC	Fractional Transmit Power Control
ICI	Inter-Cell Interference
LDM	Layer Division Multiplexing
LTE	Long Term Evolution
LTE-A	LTE-Advanced
M2M	Machine-to-Machine
MA	Multiple Access
Mbps	Mega bits per second
MIMO	Multiple Input Multiple Output
ms	milli seconds
MUST	Multi-User Superposition Transmission
NoA	Number of Additions
NoM	Number of Multiplications
NOMA	Non-Orthogonal Multiple Access
OFDM	Orthogonal Frequency Division Multiplexing
OFDMA	Orthogonal Frequency Division Multiple Access
OMA	Orthogonal Multiple Access
OS	Orthogonal Signaling
PA	Power Allocation
PAPR	Peak-to-Average Power Ratio
PD	Power Domain
PF	Proportional Fairness
QoE	Quality of Experience
QoS	Quality of Service
RA	Radio Access
RB	Resource Block
SC-FDMA	Single Carrier Frequency Division Multiple Access
SCMA	Sparse Code Multiple Access
SFR	Soft Frequency Reuse
SIC	Successive Interference Cancellation
SINR	Signal to Interference and Noise Ratio
TDM	Time Division Multiplexing
TTI	Transmission Time Interval
UE	User Equipment

Introduction

Motivation

Nowadays, human activities become highly dependent on the mobile internet services such as social media services, multimedia streaming and many real-time interactive services. According to the Cisco Visual Networking Index [1], a standard for broadband insights and trends related to spectrum, there will be 11.6 billion mobile connected devices by 2021, including Machine-to-Machine (M2M) modules and exceeding the worlds projected population at that time (7.8 billion). Furthermore, the global mobile data traffic will increase sevenfold between 2016 and 2021, reaching 49.0 exabytes per month by 2021. On the other side, with such a tremendous growth, the spectrum demanded by these devices is likely to outstrip capacity, resulting in spectrum congestion, which will lead to slower data speeds, dropped calls, and increased prices for both consumers and providers. However, the licensed frequency bands is a finite and scarce resource and making use of millimeter bands is a complex challenge. Therefore, in order to meet the increasing number of connected devices and their demands, an efficient use of the limited available spectrum has to be addressed in developing the next generation of mobile systems since the existing system is reaching its performance limits and is not able to address these challenges.

For an efficient use of the available spectrum, several Non-Orthogonal Multiple Access (NOMA) schemes are under evaluation for 5th Generation (5G) of mobile communication systems requirements. 3GPP initiated a study on downlink MultiUser Superposition Transmission (MUST) for LTE [2]. Its main objective is to scrutinize different implementations of superposition coding and the proposal of advanced receivers [3]. Another example of NOMA is the Sparse Code Multiple Access (SCMA) technique, which is regarded as a multicarrier form of NOMA [4]. Because of its superior spectral efficiency, NOMA has also been applied to other types of wireless networks. For example, a version of NOMA termed Layer Division Multiplexing (LDM), has been proposed for the next general digital TV standard ATSC 3.0 [5]. This thesis focuses on the Power Domain (PD) NOMA schemes, which will be usually simply denoted by NOMA throughout this dissertation. The concept is proposed in [6–8]. NOMA exploits an additional

degree of freedom, the power domain, by allowing multiple users to be multiplexed in the power domain such that they share simultaneously the same frequency resource. Signal separation is done by employing the Successive Interference Cancellation (SIC) which allows the separation and decoding of multi-user signals at the receiver side [9]. Its implementation becomes crucial for NOMA due to the technological improvement of end-user receivers. NOMA is known to greatly increase the system capacity which makes NOMA a hot topic for many research activities [10].

Motivated by the importance and the arising challenges of spectrum efficiency in 5G systems, this dissertation addresses several radio resource allocation problems dedicated to communication systems using the PD-NOMA technique, which in some cases can also be applied to Orthogonal Multiple Access (OMA) schemes. In this dissertation, we aim to meet the high number of connected devices, achieve users demands, and optimize several system performance metrics such as system capacity, user data rate, and user fairness. These objectives are met by providing theoretical and algorithmic solutions under multiple realistic constraints.

DISSERTATION OUTLINE AND ORGANIZATION

Chapter 1 gives the background theory and literature review necessary for the following chapters. The principles of multiple access techniques for LTE and beyond are presented in the first section of this chapter, where the basis of OFDMA, SC-FDMA, and NOMA are introduced. In the second section, resource and power allocation in OFDM and NOMA systems are introduced: some classic utility functions, including those used in this dissertation are first defined. Then, various resource allocation schemes considered in the literature for multiuser OFDM systems are presented. Multiple scheduling and power allocation schemes for downlink NOMA systems are also reviewed in this section. The third and last section summarizes the main concepts introduced in chapter 1.

In chapter 2, several allocation techniques for the minimization of the occupied bandwidth in a NOMA system are proposed. The resource allocation problem is first formulated, then, the proposed algorithm is described: it begins with an initialization and priority assignment for each user based on users Channel State Information (CSI); user selection, subband assignment and user pairing are then detailed. Afterwards, several power allocation techniques are proposed: optimum and sub-optimum waterfilling-based power allocation, weighting strategies for the optimized waterfilling-based power allocation, power allocation according to the actual achieved throughput, and a static power allocation. In this chapter, we also derive a decision metric that allows an adaptive switching to orthogonal signaling whenever it is needed; moreover, a data rate estimation for each user is carried out and we come up with a control mechanism to adjust its corresponding power. The proposed scheme is compared to an orthogonal signaling-based system and a performance evaluation is shown in the last part of the chapter.

In chapter 3, two proposals to improve the conventional proportionally fair (PF)

scheduler for a NOMA-based system are presented. The chapter starts by investigating the PF scheduler, its advantages and drawbacks. Improvements are then proposed at the level of user scheduling by introducing adaptive weights to the PF metric, followed by a performance evaluation of the proposed technique. Another proposal is then introduced at the level of power allocation based on the efficient incorporation of the waterfilling principle in the PF scheduler. A performance analysis and an evaluation of the computational complexity are performed afterwards.

In chapter 4, new throughput and fairness allocation metrics are proposed and assessed. The chapter begins by analyzing the impact of the channel gain difference between multiplexed users on the achieved throughput for different intra-subband power allocation techniques. Inspired by this analysis, multiple allocation techniques for a flexible throughput and/or fairness maximization are then proposed. The complexity and the performance of the proposed techniques are given in the last part of the chapter.

Chapter 5 investigates a hybrid broadcast and unicast system based on NOMA. It starts with a formulation of the resource allocation problem. Then, two categories of broadcast allocation techniques are proposed: *a posteriori* and *a priori*. These techniques are compared to some benchmarking schemes from the literature and are evaluated afterwards.

Finally, conclusions are drawn and future directions for research are discussed at the end of the dissertation.

CONTRIBUTIONS OF THE THESIS

Below is a summary of the main findings and contributions of my PhD thesis work:

- Several new strategies for the allocation of radio resources (bandwidth and transmission power) using a NOMA scheme with SIC in a cellular downlink system have been proposed and assessed. While taking into account various design issues, we proposed and compared several optimum and suboptimum power allocation schemes, jointly implemented with multiple user scheduling strategies. The key contributions of the proposed schemes are:
  - employing a joint inter and intra subband power allocation obtained by numerically solving an optimized allocation problem,
  - minimizing the total amount of used bandwidth while achieving a requested target data rate for each user,
  - Enabling a dynamic switching to orthogonal signaling whenever the non-orthogonal cohabitation in the power domain does not improve the achieved data rate per subband.
- A weighted PF scheduling method is proposed in the context of a NOMA system where weights are adaptively introduced to the classical PF metric in order to

make it best suited to a NOMA scenario. The distinguishing value of the proposal resides in its ability to improve long term fairness and total system throughput while achieving a high level of fairness in every scheduling slot.

- A novel low-complexity waterfilling-based power allocation technique, incorporated within the PF scheduler, is proposed and applied to a downlink NOMA system. The objective of the proposed scheme is to maximize the achieved average throughput through a quasi-optimal repartition of the transmit power among subbands, while guaranteeing a high level of fairness in resource allocation.

- Proving that the achieved throughput in a downlink NOMA system does not always increase with the channel gain difference between paired users, as it is commonly assumed in the majority of existing researches dealing with NOMA.

- Novel resource allocation schemes providing a controllable flexibility between throughput and fairness maximization have been proposed. The main appeal of these techniques is that they consider selecting users for assignment based on the previously mentioned finding. In addition to that, these techniques are incorporated with an iterative low-complexity power allocation techniques based on the waterfilling principle.

- A hybrid system that delivers broadband and broadcast services simultaneously using NOMA on top of the OFDM layer has been proposed. The novelty of the proposed techniques stems from two facts: allowing the transmission of broadband and broadcast signals on the same subbands, and considering broadcasting in *a posteriori* step, differenty from the most previous works, where OMA and NOMA-based multicasting schemes treat multicast messages in an *a priori* step.

Chapter 1

Background theory and literature review

A key technology that affects the channel capacity performance is the system's multiple access scheme [11]. In multi-carrier wireless communication systems, orthogonal frequency division multiplexing (OFDM) [12] has been widely adopted as the multiplexing technique [13], for the benefit it provides by transforming a frequency selective fading channel into a number of narrowband flat fading subchannels. A system applying OFDM benefits from many advantages: it provides orthogonality between subchannels, supports users with diverse quality of service (QoS) requirements, exploits the flexible frequency granularity, and achieves high spectral efficiency.
Despite all the benefits OFDM provides to a system, it does not allow frequency reuse within one cell [14], since a subcarrier in an OFDMA cell is allocated only to one user, which significantly limits cell throughput.
Unlike OFDM, non-orthogonal multiple access (NOMA) technique allows multiple users to share a subcarrier at the same time within one cell, which provides higher throughput and fairness due to subcarrier reuse within a cell. Thus, NOMA is a promising technique in future wireless communications [15, 16].
It is in general difficult to improve the system performance using single technology. Performance gains of multiple technologies, working in harmony, need to be combined to realize a required quality. NOMA is an advantageous technology in this aspect as well, since it can work in conjunction with various technologies. In this sense, we tend in this thesis to employ NOMA on top of OFDM in order to benefit from the advantages provided by both technologies and to further improve the system performance. For this sake, we describe in this chapter multiple access techniques in LTE and beyond and we especially detail the NOMA principle. Resource allocation techniques are also reviewed in this chapter in order to give an overview of the existing scheduling and power allocation techniques in the literature, for the sake of proposing new resource and power allocation techniques for NOMA systems throughout this dissertation.

The chapter is organized as follows: In Section 1.1 multiple access techniques in LTE and beyond are presented. Orthogonal multiple access in LTE and NOMA towards 5G are detailed in Section 1.1.1 and 1.1.2 respectively. In Section 1.2 radio resource optimization is described in the context of cellular networks. In Section 1.2.1 some classical performance metrics are summarized, including those used in the dissertation. Resource allocation in OFDM networks is presented in Section 1.2.2 and resource and power allocation in NOMA are detailed in Section 1.2.3. Finally, Section 1.3 concludes the chapter.

1.1 MULTIPLE ACCESS TECHNIQUES IN LTE AND BEYOND

Multiple Access allows multiple users to share a communication channel efficiently. A standardized multiple access (MA) scheme is usually considered as the representative feature for a cellular system in each generation. For example, code division multiple access CDMA [17] and OFDMA/SC-FDMA [18] characterize respectively the 3^{rd} and the 4^{th} generations of mobile communication systems [19, 20]. Non-orthogonal multiple access, abbreviated NOMA, is a newly introduced access scheme, that has been recognized as a promising candidate for the 5^{th} generation of radio access (RA) technologies [21]. We introduce in this chapter the basis of OFDMA, SC-FDMA, and NOMA, and we detail the resource allocation techniques used in power-domain NOMA.

1.1.1 ORTHOGONAL MULTIPLE ACCESS

In LTE systems, which was standardized in the third generation partnership project (3GPP) Release 8 [22], two orthogonal multiple access (OMA) schemes, OFDMA and SC-FDMA, have been adopted as the standard multiple access schemes for downlink and uplink transmission, respectively [18]. Both multiple access schemes are considered as suitable techniques to exploit the flexible frequency granularity, support users´ diverse quality of service (QoS) requirements, and reach high spectral efficiency. In 4G LTE systems, sometimes also referred to as LTE-A which was adopted as a standard in the third generation partnership project (3GPP) Release 10 [23], OFDMA and SC-FDMA are still adopted and the frequency bandwidth lies between 1.25 MHz and 20 MHz [24]. A peak data rate in Gbps [25, 26] can be supported by adopting advanced technology such as multiple-input multiple-output (MIMO).

1.1.1.1 ORTHOGONAL FREQUENCY DIVISION MULTIPLE ACCESS (OFDMA)

LTE uses the popular OFDMA scheme for downlink transmission. In the frequency domain, a given channel is divided into many narrower subcarriers (or subchannels) with bandwidth of 15 kHz each in both LTE downlink and uplink. The spacing is done such

that the subcarriers stay orthogonal and do not interfere with one another. In this sense, guard-bands wont be needed to be placed between adjacent subchannels. In the time domain, transmission time is divided into frames with length 10 milliseconds (ms) each. Each frame is divided into 10 subframes of equal length. Each subframe, corresponding to one transmission time interval (TTI), is divided into two equally sized time slots of length 0.5 ms. Each time slot consists of six or seven orthogonal frequency division multiplexing (OFDM) symbols depending on the choice of cyclic prefix [25]. The smallest resource unit adopted in LTE is a resource block (RB) consisting of 12 subcarriers with a total bandwidth of 180 KHz in the frequency domain and one 0.5 ms slot in the time domain [27]. Active users in a cellular network can transmit or receive data in a given time slot by using such time-frequency RBs.
Another advantage of OFDMA is its robustness to the presence of multipath fading. In data transmission, the high-speed data stream is divided into multiple substreams with lower data rate. These bitstreams are modulated into data symbols and transmitted at the same time over different subcarriers. Each subcarrier only experiences relatively constant channel gain during each TTI, since the bandwidth of each subcarrier is considered much smaller than the coherence bandwidth. Therefore, OFDMA shows efficient resistant to frequency-selective fading. In [18, 28], more detailed discussions of OFDMA can be found.

1.1.1.2 SINGLE CARRIER FREQUENCY DIVISION MULTIPLE ACCESS (SC-FDMA)

In LTE uplink, the high peak-to-average power ratio (PAPR) in transmitted OFDM signals is seen as one of the disadvantages of OFDMA. Consequently, a need for a highly linear power amplifier [29] becomes essential. High PAPR decreases the power efficiency, under a power consumption constraint on UE, and therefore reduces the battery life. Moreover, it can lead to in-band distortion in the transmitted signal and spectral broadening onto adjacent channels. These limitations are not a serious matter for downlink transmission because of high-precision power amplifiers as well as the availability of power supply at Base stations (BS). However, in uplink transmission, a mobile UE is usually limited by its battery capacity and the linearity of its power amplifier, leading to a major concern in the power consumption. To overcome these disadvantages, SC-FDMA, a modified version of OFDMA, has been adopted as the standard MA scheme for LTE uplink transmission. SC-FDMA has similar performance as OFDMA but with lower PAPR [30]. SC-FDMA and OFDMA differ also by the fact that in SC-FDMA, information symbols are transmitted sequentially rather than in parallel using orthogonal subcarriers. This mechanism offers lower PAPR by reducing the envelope fluctuations of the transmitted signal waveform, which makes SC-FDMA more suitable for uplink transmission especially for low-cost equipments with limited power. Performance investigations and comparisons between OFDMA and SC-FDMA can be found in [31, 32].

1.1.2 NON-ORTHOGONAL MULTIPLE ACCESS (NOMA)

The 4G network has seen considerable growth because of the distributions of the profatable LTE networks worldwide. The need to develop next generation communication system such as the 5G neywork stems from the growing demand of connected mobile devices and progressive data traffic [33]. For this reason, the advancement of 5G has gained momentum in research and development. 5G requirements can be identified to sufficiently support wireless communication. Compared to LTE networks, 5G should have the capability to provide 1000-fold gains in system capacity, peak data rate of fiber-like 10 Gbps and 1 Gbps for low mobility and high mobility, respectively, with at least 100 billion devices connections, low energy consumption and latency [34, 35].

In order to comply with these demanding requirements, the 5G network architecture must differentiate itself from LTE, whilst progressing current OMA systems. To this end, several non-orthogonal multiple access schemes are under evaluation for 5G requirements. For example, sparse code multiple access (SCMA) is regarded as a multicarrier form of NOMA [4], which primarily focus on creating a factor graph matrix for mapping users to the limited subcarriers. As stipulated in Release 13, 3GPP initiated a study on downlink multiuser superposition transmission (MUST) for LTE [2]. Its main contention is scrutinize multi-user non-orthogonal transmission and the proposal of advanced receivers [3]. In addition to its application in cellular networks, NOMA has also been applied to other types of wireless networks, because of its superior spectral efficiency. For example, a version of NOMA, termed Layer Division Multiplexing (LDM), has been proposed for the next general digital TV standard ATSC 3.0 [5].

This thesis focuses on the power domain non-orthogonal multiple access scheme and will be denoted simply by NOMA throughout this dissertation. The concept is proposed in [6–8]. This scheme applies superposition coding to superpose multiple UEs signals at the transmitter side. By relying on the technological improvement of end-user receivers and their processing proficiencies, the implementation of Successive Interference Cancellation (SIC) becomes crucial for NOMA which allows the separation and decoding of multi-user signals at the receiver side [9]. In figure 1.1, an illustration for single-cell OMA and NOMA in the power (as well as frequency) domain can be seen. They differentiate from each other in that NOMA permits two or more users to be attributed the same subband whilst providing a proper amount of power to each respective user, whereas each subband in OMA can only be assigned to one user.

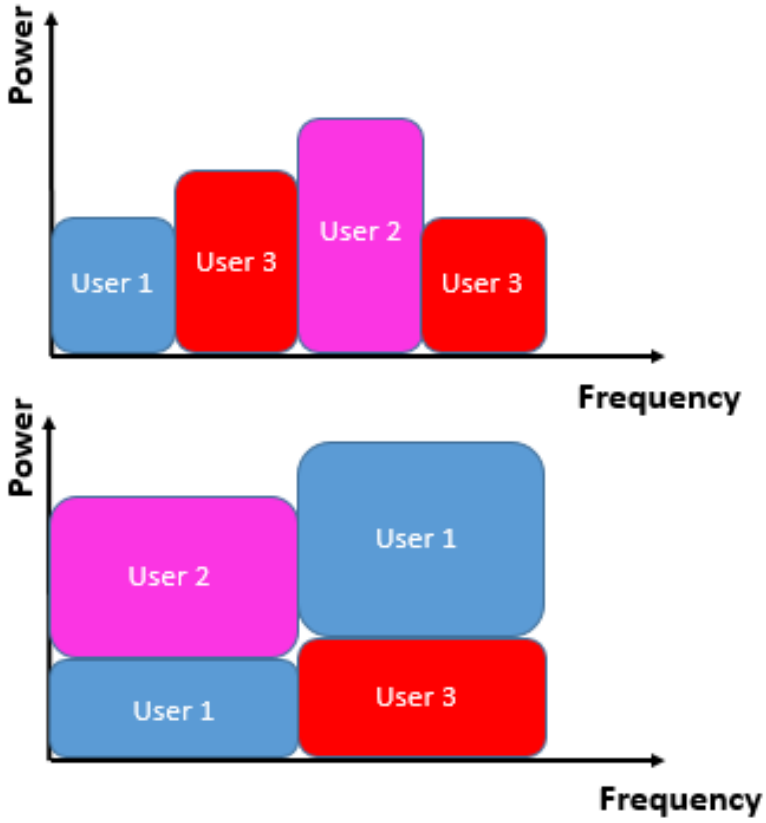


Figure 1.1: An illustration of OMA and NOMA

1.1.2.1 BASIC DESCRIPTION OF NOMA WITH SIC

In this section, we describe the general concept of NOMA including user multiplexing at the BS transmitter and signal separation at the user terminal. We assume a downlink system with a single transmitter and single antenna receiver. We consider K users per cell, and a system bandwidth B divided into S subbands.
We assume that the multiuser scheduler selects, among the K users, a set of users $N_s = \left\{k_1; k_2; ...; k_n; ...; k_{n(s)}\right\}$, to be scheduled at a frequency subband s, $(1 \leq s \leq S))$. Term k_n indicates the n-th $(1 \leq n \leq n(s))$ user scheduled at subband s, and $n(s)$ denotes the number of scheduled users at frequency subband s. At the BS transmitter side, the information sequence of each scheduled user at subband s is independently coded and modulated. The coded and modulated symbol of the n-th scheduled user at subband s is x_{s,k_n} . Thus, the signal x_s, transmitted by the BS at a certain subband s

is a simple power multiplexing of the symbols generated by the $n(s)$ scheduled users:

$$x_s = \sum_{n=1}^{n(s)} x_{s,k_n} \text{ with } E\left[|x_{s,k_n}|^2\right] = P_{s,k_n} \tag{1.1}$$

where P_{s,k_n} is the power allocated to user k_n at subband s. The received signal vector of user k_n at subband s, y_{s,k_n} , is represented by:

$$y_{s,k_n} = h_{s,k_n} x_{s,k_n} + w_{s,k_n} \tag{1.2}$$

where h_{s,k_n} is the channel coefficient between the n-th user, k_n, at subband s and the BS. w_{s,k_n} represents the received Gaussian noise plus inter-cell interference by user k_n on subband s.

Multi-user signal separation is conducted at the receiver side using the SIC process [36]. The optimal order for user decoding is in the increasing order of the users channel gains normalized by the noise and inter-cell interference $h^2_{s,k_n}/n_{s,k_n}$, where h^2_{s,k_n} is the equivalent channel gain and n_{s,k_n} the average power of w_{s,k_n}. Therefore, any user can correctly decode signals of other users whose decoding order comes before that user. In other words, user k_n at subband s can remove the inter-user interference from the j-th user, k_j, at subband s, whose $h^2_{s,k_j}/n_{s,k_j}$ is lower than $h^2_{s,k_n}/n_{s,k_n}$, and treats the received signals from other users whose $h^2_{s,k_j}/n_{s,k_j}$ is higher as noise [37, 38].

Assuming successful decoding and no error propagation, and supposing that inter-cell interference is randomized such that it can be considered as white noise [39, 40], the throughput of user k_n according to Shannon's capacity formula, at subband s, R_{s,k_n}, is given by:

$$R_{s,k_n} = \frac{B}{S}\log_2\left(1 + \frac{h^2_{s,k_n} P_{s,k_n}}{\sum\limits_{j \in N_s,\ \frac{h^2_{s,k_j}}{n_{s,k_j}} > \frac{h^2_{s,k_n}}{n_{s,k_n}}} h^2_{s,k_n} P_{s,k_j} + n_{s,k_n}}\right) \tag{1.3}$$

Fig. 1.2 shows the case of two users scheduled per subband, where k_1 (UE1) is geographically much closer to the BS than k_2 (UE2), thus has a better channel condition than k_2 ($h^2_{s,k_1} > h^2_{s,k_2}$). In this case, user 2 does not perform SIC since it comes first in the decoding order, but rather considers the signal designated to user k_1 as noise. On the contrary, user k_1 first decodes the signal x_{s,k_2} designated to user k_2 and subtracts its component from the received signal x_s. Then, it decodes its own signal without interference from x_{s,k_2}. Based on Eq. 1.3, the throughput of user k_n at subband s, R_{s,k_n} (n=1,2) can be written as:

$$R_{s,k_1} = \frac{B}{S}\log_2\left(1 + \frac{h^2_{s,k_1} P_{s,k_1}}{N_0 \frac{B}{S}}\right) \tag{1.4}$$

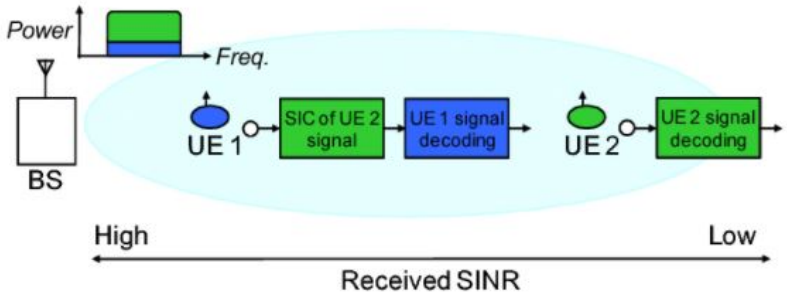


Figure 1.2: Basic NOMA applying SIC for UE receivers in downlink, figure taken from [11].

$$R_{s,k_2} = \frac{B}{S}\log_2\left(1 + \frac{h_{s,k_2}^2 P_{s,k_2}}{h_{s,k_2}^2 P_{s,k_1} + N_0 \frac{B}{S}}\right) \tag{1.5}$$

where N_0 is the power spectral density of the additive white Gaussian noise, including inter-cell interference, and assumed to be constant over all subbands.
Note that, in most papers dealing with resource allocation in downlink NOMA [38, 41–43], the number of users per subband was set to two in order to limit the SIC complexity in the mobile receiver, except for [39, 44] where this number can respectively reach three and four.
It can be seen from Eq. 1.4 and 1.5 that the choice of the multiplexed users over subband s, as well as the amount of power allocated to each user, significantly affects user throughput performance. Therefore, resource allocation, scheduling, and power allocation should be explored carefully in order to reach a gain with NOMA compared to orthogonal-based systems.

1.2 RESOURCE ALLOCATION IN DOWNLINK CELLULAR NETWORKS

The key function of performance improvement of a cellular network fundamentally involves radio resource allocation or scheduling. The main objective of resource allocation is the optimization of the limited frequency/power/time resources assignment in order to achieve desired performance while taking into account realistic constraints. A radio resource optimization problem consists of a utility function as the objective, and a set of constraints to be optimized. The utility can be selected from a range of performance metrics. Various physical limitations in cellular networks or QoS requirements are in practice the reasons behind setting the set of constraints. When combined with with

the optimization variables, a feasible solution region for the optimization problem can be then defined.
At large, we aim to find optimal solutions from the feasible region, or develop near-optimal solutions. In this dissertation, we address several radio resource optimization problems in NOMA.
In this section, we first present some widely used utility metrics.

1.2.1 CLASSIC UTILITY FUNCTIONS

Utility, in general, can be viewed as an abstract concept, e.g., satisfaction and fairness, or a real performance measure, e.g., achieved throughput in bps. A utility function is mainly employed to quantify and provide a performance metric that will be the object of an optimization problem. Summarized below are some classic utility functions, including those used in this dissertation.

- Throughput maximization
  Throughput, also referred to as sum-rate utility, epitomizes the aggregate data rate of the users in a cellular network. The corresponding utility function is expressed by $\sum_k R_k$, where R_k is the data rate achieved by user k. This is usually computed by Shannon channel capacity equation in bits per second. In addition, spectrum efficiency in bits per second per Hz, can also be used to quantify throughput in unit bandwidth. For applications where throughput is the lone performance metric that requires to be high achieved, radio resources will be attributed to users that maximize the sum-rate utility function.

- Probability of success
  In some applications, there is a need to respect specific users QoS requirements by providing each user with a requested service data rate. In this case, the corresponding utility function is expressed by the probability of success which is defined as as $\Pr(R_k = R_{k,requested})$ where $R_{k,requested}$ denotes the data rate requested by user k from the base station.

- Weighted sum-rate and fairness
  In some application scenarios, users priority and fairness in resource allocation need to be taken into account, instead of simply considering maximum throughput. In order to maintain fairness among users, a weight W_k is introduced for each user, corresponding utility function can be expressed as $\sum_k W_k R_k$. For instance, a user with a bad channel condition differentiates itself from a user with a good channel condition and could be allocated with higher weight to avoid excessive imbalance

in resource allocation and provide fairness among users. In resource scheduling over a time duration, weights can be used to achieve fairness. For example, one can update $W_k = 1/{R_k}^{avg}$ for each k in each time slot, where ${R_k}^{avg}$ is the average rate of UE k. There are numerous measurements to quantify fairness [52]. Gini [53] and Jain fairness [54] index are ones of the most common measures to represent fairness.
Gini fairness index measures the degree of fairness that a resource allocation scheme can achieve. It is defined as:

$$G = \frac{1}{2K\overline{R}} \sum_{x=1}^{K} \sum_{y=1}^{K} |R_x - R_y| \tag{1.6}$$

with

$$\overline{R} = \frac{\sum_{k=1}^{K} R_k}{K} \tag{1.7}$$

R_k is the total throughput achieved by user k. It is calculated as follow:

$$R_k = \sum_{s=1}^{S} R_{s,k} \tag{1.8}$$

Gini fairness index takes values between 0 and 1, where $G = 0$ corresponds to the maximum level of fairness among users, while a high value of G indicates that the resource allocation scenario is highly unfair.
Jain fairness index is very similar to the Gini index and represents the fairness among users in terms of the allocated resources. It is defined as follow:

$$J = \frac{\left(\sum_{k=1}^{K} R_k\right)^2}{K \sum_{k=1}^{K} {R_k}^2} \tag{1.9}$$

J is between 0 and 1, and is closer to 1 when the difference in the capacity of each user channel is small and fairness is achieved.

- Minimum number of allocated subbands
  Taking the scarcity of frequency resource into account, additional type of utility metric can be presented by minimizing the number of allocated subbands needed to meet each users data request [55]. There are two main benefits: first, the inter-cell interference (ICI) is mitigated by minimizing the used channels in each cell, second, resources will be available to be allocated as much as possible in order to

allow elastic traffic. while guaranteeing the rates for real-time applications. The benefits consist of two aspects. First, the inter-cell interference (ICI) is mitigated by minimizing the used channels in each cell. Second, it is relevant to make as much resource available as possible for elastic traffic, while guaranteeing the rates for real-time applications. The utility function is expressed as $\sum_k |N_k|$ for OFDM-based systems through exclusive channel access, and $\left|\bigcup_k N_k\right|$ for NOMA systems, where N_k is the set of allocated subbands of user k.

1.2.2 RESOURCE ALLOCATION IN OFDM NETWORKS

Resource allocation schemes for multiuser OFDM systems profit from various forms of diversity such as frequency diversity and multiuser diversity. Frequency diversity refers to different subcarriers within a wireless link having different channel gains due to the frequency selective nature of the channel. On the other hand, multiuser diversity refers to different users facing diverse channel conditions on a certain frequency because of their different locations in the network. These diversities imply that a subcarrier that is in a deep fade for one user may not be in deep fade for the other users. These diversities can be explored further by assigning the subcarriers to users based on the channel conditions experienced by the users on the subcarriers. It was shown in [56] that adaptive resource allocation can improve the OFDM's system performance. Adaptive resource allocation strategies allows an efficient use of the available resources, by taking into consideration the channels conditions. Since, in OFDM-based systems, signals can be transmitted simultaneously, the allocation problem tends to be simplified, and users are orthogonally assigned different subsets of subcarriers. In this section, we will present various adaptive resource allocation schemes considered in the literature.

In the context of multiple user OFDMA networks, subcarriers and power are assigned in such a way to optimize a service utility function. The multiuser resource allocation can be perfomed using waterfilling-based techniques. In [57], an iterative waterfilling algorithm is proposed for the sake of maximizing the sum capacity of a Gaussian multiple access channel. In [58], the data rate maximization is performed otherwise; each subcarrier is attributed to its best user (i.e., the user with the highest channel gain), and then waterfilling is applied on subcarriers assigned to each user individually.
The waterfilling-based technique aims at maximizing the system capacity but does not guarantee fairness among users. In other words, the users that are away from the base station or with bad channel condition will be disadvantaged. In this sense, several works have taken into account the individual rate or the QoS requirements. Some works tend to minimize the transmit power while guaranteeing a service data rate for each user [59,60]. Other works, aims to maximize the throughput under a power constraint with some fairness criteria (e.g., proportional fairness among the users). For example, in [61] a rate

maximization problem is considered with the objective of maximizing the minimum rate among the users, for a given power budget. The proportional fairness (PF) scheduler is also used in a huge number of existing works dealing with resource allocation for OFDM system. The PF scheduler allocates subbands to users in such a way to provide a fair balance between throughput and fairness. In [62] the rate maximization is extended to include proportional fairness.

1.2.3 RESOURCE AND POWER ALLOCATION IN NOMA

Unlike OMA, multiple users in NOMA are allowed to simultaneously share the same subband, and one user may need to be multiplexed on several subbands, with different cohabiting users and in various multiplexing orders within each subband. Even after SIC processing, co-channel interference in NOMA is non-negligible. Therefore, radio resource optimization in the case of NOMA is not straightforward. In this sense, several design aspects should be taken into consideration, such as user pairing and multiuser frequency scheduling, power allocation among subbands, power repartition between scheduled users within a subband, etc., as well as the interaction of these different design issues. Some algorithms and schemes are proposed to optimize the channel and power allocation for NOMA dowlink and uplink [63–65].
Most of the previous approaches address the radio resource optimization problems in NOMA by making assumptions to reduce the complexity of the optimization process, e.g., assuming uniform power allocation, predefining fixed groups of users or/and channels before the optimization process. Some researches also consider splitting the difficult optimization procedure into several problems considered easier to be solved. By doing so, the overall problem becomes tractable, however, the optimality is sacrificed, e.g., splitting the joint channel and power allocation into two separate steps: channel allocation and power allocation.

1.2.3.1 SUBBAND ASSIGNMENT AND USER PAIRING

In a large number of works dealing with NOMA, the proportional fairness (PF) scheduler [66, 67] has been considered for user pairing and subband assignement due to the good tradeoff it provides between system capacity and user fairness. Below is a summarized representation of the PF scheduler technique. Note that a more detailed description is given in Chapter 3.
In an OMA system, user k^* is chosen among possible users K to be scheduled on subband s from the set of available subbands based on the following metric:

$$k^* = \arg\max_{K} \frac{R_{s,k}(t)}{T_k(t)} \tag{1.10}$$

where $R_{s,k}(t)$ is the achievable throughput of user k at time instance t and $T_k(t)$ its historical rate. $T_k(t)$ is calculated using a moving average window of length t_c and it is updated at every new scheduling slot according to Eq. 1.11:

$$T_k(t+1) = \left(1 - \frac{1}{t_c}\right) T_k(t) + \frac{1}{t_c} \sum_{s=1}^{S} R_{s,k}(t) \tag{1.11}$$

In NOMA systems, Eq. 1.10 is replaced by Eq. 1.12 where U_s is the set of scheduled users among all possible candidate sets U chosen to be scheduled on subband s, in a way to maximize the PF metric.

$$U_s = \arg\max_{U} \sum_{k \in U} \frac{R_{s,k}(t)}{T_k(t)} \tag{1.12}$$

1.2.3.2 INTER-SUBBAND POWER ALLOCATION

The total bandwidth granted to serve users is divided over a fixed number of subbands. Every subband is then attributed a fraction of the available transmit power that the base station is allowed to use. An equal repartition is the simplest way to divide the power among subbands since it reduces the complexity of the scheduling process; therefore, a great number of papers dealing with NOMA consider this repartition [68, 69].
In [70], the performance of downlink NOMA with wideband and subband frequency scheduling is evaluated under wide-area cellular system configurations. Power allocation is done such that the amount of power attributed to a subband is common to all subbands, and intra-subband power repartition between each pair of scheduled users is done as a subsequent stage using fractional transmit power allocation (FTPA) [71].
In [40], the throughput improvement using non-orthogonal superposition of users on top of an uplink OFDMA-based system is studied. All mobile users are considered to be transmitting signals at the same power level. Optimized scheduling techniques are proposed to maximize the system throughput. In this sense, a cost function is assigned to each possible pair of users and the Hungarian method is used to solve the problem of user pairing. Cost functions are chosen so as to optimize either the sum-rate or the weighted sum-rate. These two scheduling techniques are shown to provide significant improvements in sum-rates and cell-edge rates compared to orthogonal signaling.
In [43], a non-orthogonal multiuser beamforming system is proposed for improving the system capacity. Degradation in the sum capacity due to inter-cluster interference and inter-user interference is assessed, leading to a clustering and power allocation algorithm that aims at reducing interference and improving capacity. Clustering is done by selecting two users to be paired together, in the same beamforming vector, as having a large gain difference and a high correlation. In addition, power is equally divided between clusters, and then allocated among scheduled users in each cluster, in such a way to maximize the sum capacity, while guaranteeing a minimum capacity for the weakest user.

In [39], the system-level throughput performance of non-orthogonal access with a SIC in downlink is investigated, using a scheduling algorithm based on the proportional fair scheduler. Power is equally partitioned among subbands, and an optimal power allocation strategy is proposed in order to allocate power iteratively among scheduled users within the same subband. This method has shown to achieve a good tradeoff between total and cell-edge user throughputs. In addition, it allows enhancing system-level throughput compared to orthogonal access. However, although the optimal power allocation strategy shows good performance, it is computationally complex. For this purpose, two simplified suboptimal intra-subband power allocation techniques are also proposed to reduce the complexity of the optimal scheme.
However, some other works have proposed non-equal power allocation among subbands for NOMA [14, 72, 73]. In [14], an inter-cell interference-aware transmission power control is proposed and conducted in two steps, followed by user selection based on the PF metric. In the first step, the transmission power of a user per subband is determined by the fractional transmit power control (FTPC) [74] used in LTE. The power is then updated in a second step by taking into consideration the candidate set of scheduled users. Simulation results show that NOMA combined with the proposed power allocation greatly enhances the system-level throughput, compared to orthogonal access. In [72], the system-level performance of downlink NOMA in small cells is investigated, where the full search power allocation scheme in [75] is conducted within the PF scheduler with 10 power sets in order to select the best combination of user pairs and power allocations.

1.2.3.3 INTRA-SUBBAND POWER ALLOCATION

At this level, users are paired together on a subband s that is attributed a certain amount of power P_s. The last step to be accomplished is to divide this power among scheduled users, in a way to respect the NOMA principle: the user with the highest channel gain must be attributed the lowest amount of power, and vice-versa. We can identify several intra-subband power allocation techniques. The less complex one is called Fixed Power Allocation (FPA) [81] where a coefficient β controls the power attributed to every user such as Eq. 1.13 and Eq. 1.14.

$$P_{s,k_1} = \beta P_s \tag{1.13}$$

$$P_{s,k_2} = (1-\beta)\, P_s \tag{1.14}$$

β should be less than or equal to 0.5 so that the power P_{s,k_1} attributed to the user $k_1$with the highest channel gain on subband s is always less than or equal to the power P_{s,k_2} attributed to the user with the lowest channel gain. β should be optimized prior to the power allocation process and it is fixed throughout it.

Another technique called Fractional Transmit Power Allocation (FTPA) [69] can be also used to divide P_s between paired users. This technique is quite more complex than FPA since the amount of power attributed depends on the channel gains of scheduled users as given by Eq. 1.15 and Eq. 1.16.

$$P_{s,k_1} = \frac{\left(\frac{h^2_{s,k_1}}{N_0 B/S}\right)^{-\alpha}}{\left(\frac{h^2_{s,k_1}}{N_0 B/S}\right)^{-\alpha} + \left(\frac{h^2_{s,k_2}}{N_0 B/S}\right)^{-\alpha}} P_s \tag{1.15}$$

$$P_{s,k_2} = \frac{\left(\frac{h^2_{s,k_2}}{N_0 B/S}\right)^{-\alpha}}{\left(\frac{h^2_{s,k_1}}{N_0 B/S}\right)^{-\alpha} + \left(\frac{h^2_{s,k_2}}{N_0 B/S}\right)^{-\alpha}} P_s \tag{1.16}$$

α is a factor ranging between 0 and 1 that controls the amount of power attributed to every user. Its value is fixed throughout the power allocation step. Here, FPA and FTPA expressions are given for the case of two multiplexed users. However they can be extended to any number of paired users such that the summation of all powers attributed to users remains equal to P_s.

Another well-known intra-subband power allocation technique is called Full Search Power Allocation (FSPA) [81]. The complexity of this technique is very high compared to FPA and FTPA since it requires, in each intra-subband power allocation step, a search among all possible power repartitions and chooses the repartition that results in the highest achieved throughput. During this search, all possible values of β are considered. Even though FSPA imposes a high penalty in terms of complexity, it was found in [68] that its resulting throughput does not sufficiently surpass FPAs neither FTPAs achieved throughput.

After a tour of state of the art review of power and resource allocation techniques for NOMA, we have been motivated to go deeper with our research in order to find the best NOMA-based resource and power allocation techniques. Most of the researches related to NOMA have considered improving the system capacity but few of them have targeted a joint improvement of the capacity and the short-term fairness (fairness at every scheduling time instance). When dealing with user pairing, the majority of papers claim that users have to be chosen to allocate a subband such that the channel gain difference between them is maximized. Few are the papers that studied this issue carefully by evaluating how the user rate is affected by the channel gain difference. When it comes to power allocation, a lot of works have been done, but few of them have reached an optimal power repartition. For this sake, we have decided to work on several aspects of resource and power allocation in order to achieve high performance in the NOMA system.

1.3 CONCLUSION

This chapter introduced a background theory and a literature review of a NOMA-based system. Since in the following chapters, all the proposed techniques related to NOMA will be compared to an LTE-based system, multiple access techniques for downlink and uplink scenarios were first presented in this chapter. Then, the principle of NOMA with SIC was well detailed for a downlink system. Radio resource optimization, for an OFDM and a NOMA system, was also considered in this chapter in order to give an overview of the existing scheduling and power allocation techniques in the literature for the sake of proposing new adaptive ones in the following chapters.

Chapter 2

Techniques for the minimization of the occupied bandwidth in a NOMA system

Future 5G and beyond wireless communication systems will require a flexible radio interface with high constraints in terms of spectral and power efficiencies. Therefore, we aim throughout this chapter to propose new techniques for the dynamic assignment of available subbands that seek to achieve two goals. The first one is to reduce the amount of used bandwidth. In fact, the optimization of the amount of occupied bandwidth is of a great interest for the upcoming generations of mobile communications, in order to cope with the increasing scarcity in spectrum and, at the same time, meet the exploding demands of higher data rates. The second goal is to achieve user fairness by providing each user with a requested service data rate.

By combining the potential gain achieved by NOMA through power multiplexing, with a proper optimization of the amount of used bandwidth, we seek to boost the spectral efficiency, while respecting a set of traffic requirements. However, using NOMA within such strategies for resource optimization is not straightforward. This is mainly due to the fact that, in order to respect the resource requirements of a set of users, each one of the latter needs to be multiplexed on several subbands, with different cohabiting users and in various multiplexing orders within each subband. For this purpose, several design aspects will be taken into consideration throughout this chapter, such as user pairing and multiuser frequency scheduling, power allocation among subbands, power repartition between scheduled users within a subband, etc., as well as the interaction of these different design issues.

Our proposed allocation techniques also take into consideration the fact that, in certain situations and on specific subbands, NOMA may not constitute the appropriate solution. Therefore, a dynamic switching between NOMA and a classical OFDMA (i.e. OFDMA without the additional non-orthogonal signaling layer) is allowed under certain conditions.

The chapter is organized as follows: In Section 2.1 the resource allocation problem is formulated. In Section 2.2 a detailed description of the proposed strategy for resource allocation is given. Then, in Section 2.3 an OMA system is presented and compared to NOMA. Afterwards, in Section 2.4 the performance evaluation results of the proposed scheme are presented. Finally, Section 2.5 concludes the chapter.

2.1 FORMULATION OF THE RESOURCE ALLOCATION PROBLEM

In addition to maximizing system throughput, as performed in the majority of existing literature on NOMA, this chapter targets minimizing the amount of used bandwidth. In other words, the proposed allocation technique tends to provide each user with its requested data rate using the minimum number of subbands, under the constraint of a maximum allowed transmit power. Let S_A be the actual number of available subbands ($1 \leq S_A \leq S$), i.e. $S - S_A$ subbands are supposed to be occupied by another system.
$R_{k,requested}$ is the data rate requested by user k from the BS, $1 \leq k \leq K$.
$P_{s,k}$ is the transmit power of user k over subband s ($P_{s,k} \neq 0$ if k is scheduled on s).
$R_{s,k}$ is the achieved data rate by user k over subband s. And S_k denotes the set of all subbands allocated to user k.
The optimization problem can be formulated as follows:

$$\underset{P_{s,k}}{\text{minimize}} \sum_{k=1}^{K} card(S_k) \tag{2.1}$$

Subject to:

$$\sum_{s \in S_k} R_{s,k} = R_{k,\text{requested}},\ \forall\ k,\ 1 \leq k \leq K \tag{2.2}$$

$$\sum_{k=1}^{K} \left(\sum_{s \in S_k} P_{s,k} \right) \leq P_{\max} \tag{2.3}$$

$$P_{s,k} \geq 0,\ \forall s \in S_k,\ 1 \leq k \leq K \tag{2.4}$$

Where $card(S_k)$ represents the cardinality of the set of subbands allocated to user k.
If user k has a channel gain over s that allows him to perform SIC, his data rate $R_{s,k}$ is computed based on Eq. 1.4. Otherwise, it is computed based on Eq. 1.5.
Eq. 2.1 represents the main design function. It aims at minimizing the number of allocated subbands under the rate and power constraints expressed by Eq. 2.2, 2.3, and 2.4. This optimization problem is combinatorial and cannot be resolved in such a way to yield closed-form solutions. Besides, an exhaustive search for the values of $P_{s,k}$ and S_k is impractical, because of the large number of parameters and constraints involved. Moreover, compared to previous orthogonal-based allocation techniques [7,8,9], resource

allocation for the non-orthogonal system must also consider the following additional design constraints: the choice of user pairing scheme, the power distribution between allocated subbands, and the power division between paired users within a subband. In [110], a mixed combinatorial non-convex optimization problem for the maximization of the weighted sum throughput of the system was formulated. The optimization problem was solved optimally using Monotonic optimization, and the resulting optimal power and subcarrier allocation policy has served as a performance benchmark due to its high computational complexity.
Therefore, in the next section, we propose an allocation technique aiming at efficiently taking all these design constraints into consideration.

2.2 DESCRIPTION OF THE PROPOSED ALGORITHM FOR RESOURCE ALLOCATION

In this section, the allocation technique is proposed in Fig. 2.1, where the overall optimization problem is divided into several steps. Those steps are carefully detailed in the sequel.

2.2.1 INITIALIZATION AND PRIORITY ASSIGNMENT

Since each user needs to reach a requested service data rate, users must not be treated equally. For instance, users having the largest rate distance (or gap) towards their requested service data rates should be given higher priority over users that are close to reach their targets. But since at the beginning of the allocation process, transmit powers $P_{s,k}$ and user rates $R_{s,k}$ are all set to zero, priorities are first assigned based on users Channel State Information (CSI), i.e. the channel gains between mobile terminals and the BS. Therefore, it is assumed in this thesis that CSI information is periodically transmitted from mobile users to the BS on dedicated control channels. The channel gains are grouped in a matrix **H**, represented in Fig. 2.2, with dimensions $S_A \times K$, where $h_{s,k}$ is the channel gain experienced by user k on subband s.

In this sense, priority assignment at the beginning of the allocation process is defined as follows:

- For each user k, select the highest channel gain $h_{s_{best},k}$ among the elements in the k^{th} column of matrix **H** (denoted by a circle in Fig. 2.2).

- The user with highest (resp. lowest) priority is the one having the lowest (resp. highest) best channel gain among circled elements.

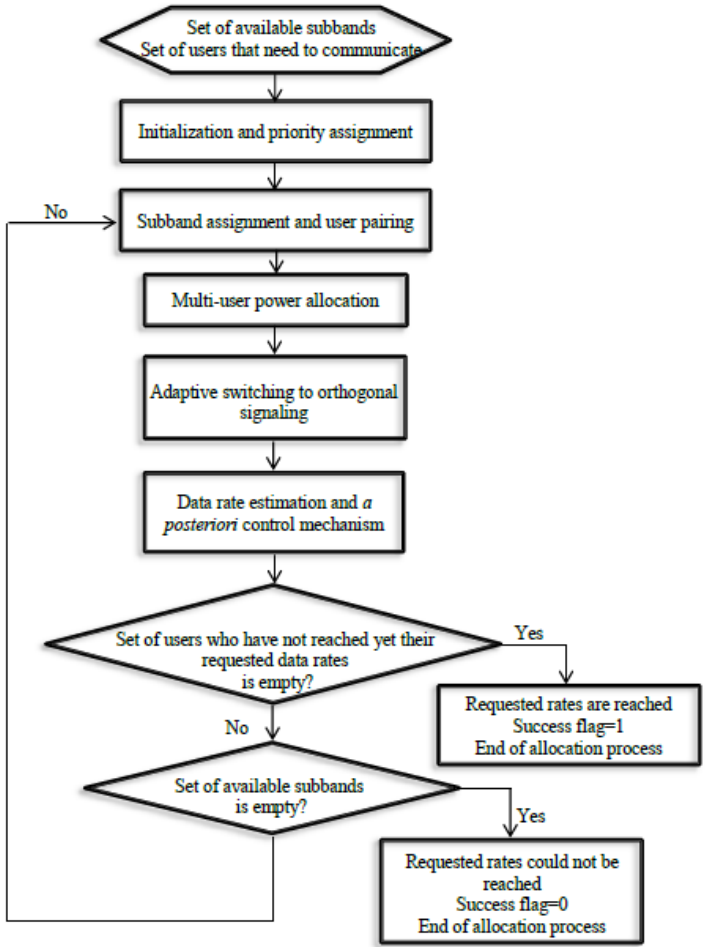


Figure 2.1: Proposed Resource Allocation Technique

2.2.2 SUBBAND ASSIGNMENT AND USER PAIRING

During the iterative process depicted in Fig. 2.1, a candidate set of users denoted by U_{s^*} is identified to be assigned over s^* by applying the following steps, where $card(U_{s^*}) = 2$:

Step 1: User Selection
Select the first user, denoted by k_1, among the set of users that need to communicate, and that have not reached their target rate yet. Selection is based on one of the following criteria:

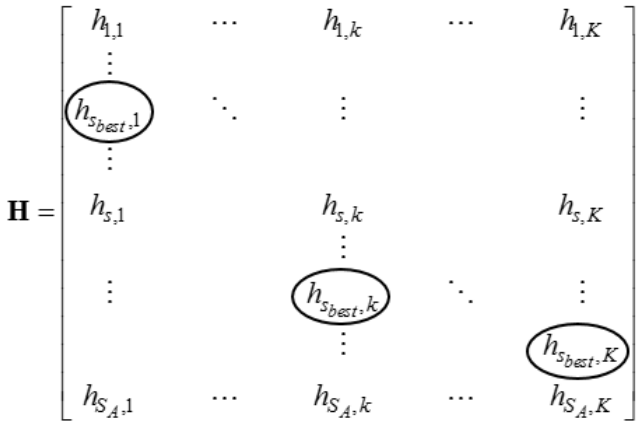


Figure 2.2: Channel gain matrix **H**

- While there exist at least 2 users whose data rates are zero, select user k_1 among them based on their channel gains, as defined in section 2.2.1;

- Once the above criterion is no longer verified, i.e. data rates of all users are non-zero (each user has at least one attributed subband, with a non-zero transmission power on this subband), or it still exists only one user whose data rate is equal to zero, select user k_1 as the one showing the largest rate distance towards its requested service data rate.

Step 2: Subband assignment
Attribute the most favorable subband (the one with the highest channel gain), denoted by s^*, to user k_1. s^* is then removed from the set of available subbands S_A.

Step 3: User pairing
Select a second user k_2 to be multiplexed in the power domain with user k_1 on the current selected subband s^*. User pairing can be done in different ways. In this chapter, we have evaluated two options:

Pairing 1:
User k_2 is chosen as the user having the next lowest channel gain over s^*, when compared to the one of k_1.

Pairing 2:

User k_2 is chosen as the user having the worst channel gain over s^*.

In both pairing options, the channel gain of user k_2 is lower than that of k_1. Therefore, user k_2 does not perform SIC. Instead, his corresponding receiver considers the signal of user k_1 as interfering noise with $P_{s,k_1}h^2_{s,k_2}$ as the interfering term.
In the case where we extend the number of scheduled users per subband to be more than two, steps 1 and 2 are kept the same, and step 3 is modified so that $card(U_{s^*})-1$ users are chosen to be multiplexed with k_1 on the current subband s^*, with $card(U_{s^*})$ the number of scheduled users on s^*.
User scheduling can still be performed in two ways:

- The $card(U_{s^*})-1$ other users are chosen as those having channel gains strictly less than that of k_1;
- Users having channel gains lower than that of k_1 are divided into $card(U_{s^*})-1$ groups, and users to be scheduled on s^* are chosen as those having the worst channel gain in each group.

Step 4: Inverting roles
If, during the allocation process, it happens that user k_1 has the lowest channel gain on its attributed subband s^* (i.e. all other users have higher gains on this subband, compared to k_1), user k_2 is then chosen as the user having the highest gain on this subband, if pairing 2 is adopted in step 3. Otherwise, i.e. if pairing 1 is used, k_2 is chosen as the user having the next highest channel gain over s^*, when compared to the one of k_1.

2.2.3 MULTI-USER POWER ALLOCATION

In order to distribute power among users, several power allocation techniques are proposed in this section: Optimum and sub-optimum waterfilling-based power allocation, weighting strategies for the optimized waterfilling-based power allocation, power allocation according to the actual achieved throughput, and a static power allocation. These listed techniques are detailed in the sequel.

2.2.3.1 OPTIMUM WATERFILLING-BASED POWER ALLOCATION

In [22, 26], static inter-subband power allocation is used for NOMA, where the total transmit power is identically divided between subbands. However, it is stated that the resulting achievable throughput is penalized since the waterfilling principle [29,30] is not applied. For this reason, we propose to apply a waterfilling-based optimal subband power allocation that takes into consideration the channel gains of the two paired users within each subband. It is described by the following optimization problem:

At each stage of the allocation process, maximize the total achieved throughput for users that have not yet reached their requested data rate, under the constraint of the total remaining power:

$$\underset{\{P_{s,k_1},\, P_{s,k_2},\, \forall s \in S_u\}}{\text{maximize}} \quad \sum_{s \in S_u} (R_{s,k_1} + R_{s,k_2}) \tag{2.5}$$

Subject to:

$$\sum_{s \in S_u} (P_{s,k_1} + P_{s,k_2}) = P_{rem} \tag{2.6}$$

S_u is the set of subbands attributed to users whose target data rates have not been reached so far (those users constitute a set U), and P_{rem} denotes the remaining transmit power to be distributed between subbands, at the current stage of the allocation technique.
Solving this optimization problem using Lagrange multipliers leads to the following formulation of the objective function, where λ is the Lagrange multiplier:

$$\begin{aligned} J = & \sum_{s \in S_u} \frac{B}{S} \log_2 \left(1 + \frac{P_{s,k_1} h_{s,k_1}^2}{N_0 \frac{B}{S}} \right) + \sum_{s \in S_u} \frac{B}{S} \log_2 \left(1 + \frac{P_{s,k_2} h_{s,k_2}^2}{P_{s,k_1} h_{s,k_2}^2 + N_0 \frac{B}{S}} \right) \\ & + \lambda \left(P_{rem} - \sum_{s \in S_u} (P_{s,k_1} + P_{s,k_2}) \right) \end{aligned} \tag{2.7}$$

Generally, power multiplexing in NOMA is done such that the highest power is given to the user with the weakest channel gain (user k_2 in our case) [4,12,16]. Therefore, we adjust the power allocation ratio between and by setting a parameter β_s such that:

$$P_{s,k_2} = \frac{1 - \beta_s}{\beta_s} P_{s,k_1}, \text{ with } 0 < \beta_s < \frac{1}{2} \tag{2.8}$$

By substituting Eq. 2.8 in Eq. 2.7, then differentiating J with respect to and λ, and by setting the resulting expressions to zero, we obtain a non-linear system described in Eq. 2.9 by $N_u + 1$ equations with $N_u + 1$ unknowns, P_{s,k_1} and λ, N_u being the current number of elements in S_u.

$$\begin{cases} \dfrac{\frac{h_{s,k_1}^2}{N_0 \frac{B}{S}}}{1 + \frac{h_{s,k_1}^2}{N_0 \frac{B}{S}} P_{s,k_1}} + \dfrac{\frac{1-\beta_s}{\beta_s} \frac{h_{s,k_2}^2}{N_0 \frac{B}{S}}}{\left(1 + \frac{h_{s,k_2}^2}{N_0 \frac{B}{S}} P_{s,k_1} \right) \left(1 + \frac{1}{\beta_s} \frac{h_{s,k_2}^2}{N_0 \frac{B}{S}} P_{s,k_1} \right)} = \lambda \ln 2 \dfrac{1}{\beta_s} \dfrac{S}{B} \\ \sum_{s \in S_u} \dfrac{1}{\beta_s} P_{s,k_1} = P_{rem} \end{cases} \tag{2.9}$$

Note that, in the particular case where $P_{s,k_2} = 0 (\beta_s = 1)$, we get a linear system:

$$\frac{\log_2(e)\frac{B}{S}}{P_{s,k_1} + \frac{N_0 \frac{B}{S}}{h^2_{s,k_1}}} = \lambda, \ s \in S_u \tag{2.10}$$

This linear system corresponds to the solution of the optimization problem using the classical OFDM with orthogonal signaling as a multiple access technique (i.e. with no user cohabitation). In order to solve the above non-linear system of equations described in Eq. 2.9, we used a numerical solver, the trust-region method based on the Dogleg algorithm [31,32]. This algorithm determines the values of the $N_u + 1$ unknowns P_{s,k_1} and λ. However, it does not always guarantee non-negative solutions. To overcome this problem, we propose two alternatives that can be applied when at least one negative power is found at a certain stage of the iterative allocation process:
Alternative 1: switch, at this stage of the iterative process, to a sub-optimum solution for power allocation such as the one subsequently described in section 3.2.
Alternative 2: substitute the negative powers by zeros, and re-distribute the remaining power P_{rem} to the set of subbands where power was found to be positive. This redistribution is done using the same trust-region method, and the process is iterated until only positive solutions are found.
This power allocation strategy, including both alternatives, is depicted in fig. 2.3. To extend the number of scheduled users per subband to more than two, Eq. 2.5 and Eq. 2.6 are to be respectively replaced by Eq.2.11 and 2.12, with the appropriate rate expressions expressed using Eq. 1.3.

$$\underset{\{P_{s,k_n}, \ \forall \ k_n \in N_s, 0 \leq s \leq \ S_u\}}{\text{maximize}} \ \sum_{s \in S_u} \sum_{n=1}^{n(s)} R_{s,k_n} \tag{2.11}$$

Subject to:

$$\sum_{s \in S_u} \sum_{n=1}^{n(s)} P_{s,k_n} = P_{rem} \tag{2.12}$$

2.2.3.2 SUB-OPTIMUM WATERFILLING-BASED POWER ALLOCATION

The optimum solution, described in the preceding section, performs a joint inter and intra-subband power allocation that takes into consideration the channel gains of all paired users. This solution reveals to be rather complex to implement. Therefore, we also propose a sub-optimum solution, where the power is allocated among users in two consecutive stages, inter-subband and intra-subband allocation, as shown in Fig. 2.4

Stage 1: Inter-subband power allocation
In this first step, we propose to consider only the highest channel gain (i.e. for user k_1)

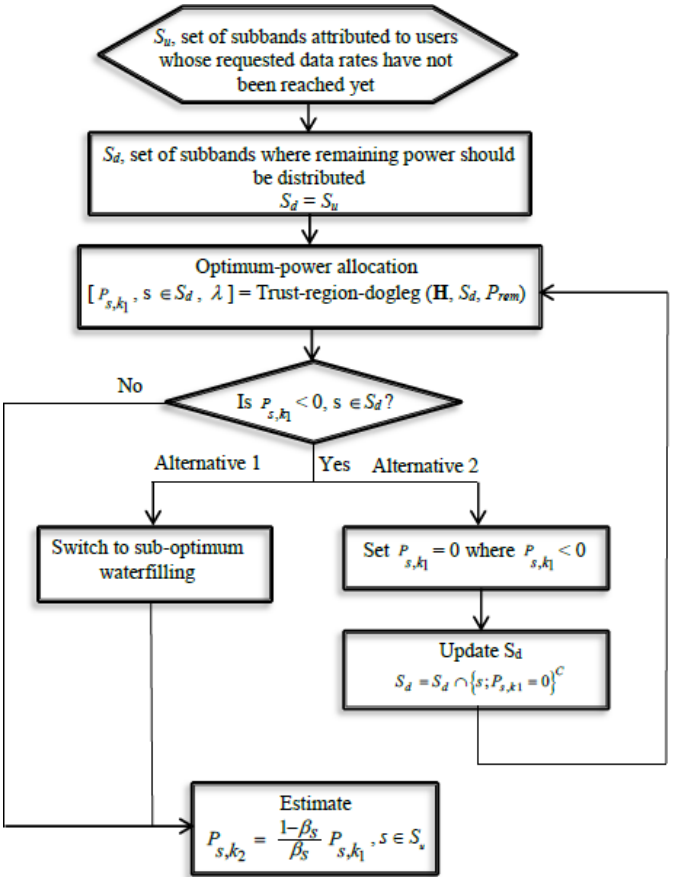


Figure 2.3: Optimum waterfilling-based power allocation

within each subband. In other words, the channel gain of k_1 on subband s determines the total amount of power, P_s, that will be attributed to s, using a waterfilling process, and that will be subsequently partitioned between the two paired users on s. The waterfilling process is performed in an iterative way as in [9]. Even though this allocation technique represents a sub-optimum solution, it is expected to perform better than static power allocation that equally partitions power between subbands.

Stage 2: Intra-subband power allocation

The power allocated in the inter-subband allocation step is now to be partitioned between paired users within each subband. Intra-subband repartition can be done in a static way, according to a fixed threshold, or in a dynamic way, based on paired users channel gains.

Static intra-subband power allocation: Fixed Power Allocation (FPA):

The repartition is done in a static way over all subbands, where the total transmit power

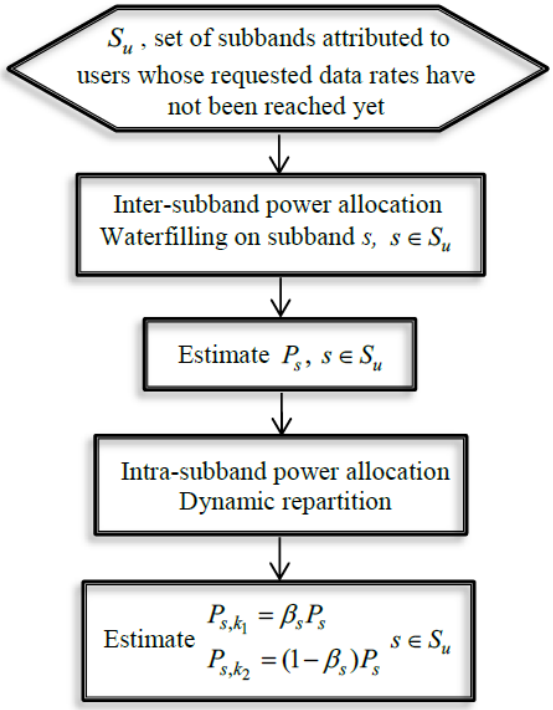


Figure 2.4: Sub-optimum waterfilling-based power allocation

P_s, allocated in stage 1 to subband s, is divided between paired users according to $(\beta.P_s, (1-\beta)P_s)$, with $\beta(0 \leq \beta \leq 0.5)$ being a constant parameter over all subbands. The user with the highest channel gain will be given $\beta.P_s$ and the paired user will be given the rest.

Dynamic intra-subband power allocation: Fractional Transmit Power Allocation (FTPA):

The repartition is done in a dynamic way, similarly to the algorithm in [22] which is based on the channel gains of the two multiplexed users, such that β_s in Eq. 2.8 is

$$\beta_s = \frac{h_{s,k_1}^{-2\alpha}}{h_{s,k_1}^{-2\alpha} + h_{s,k_2}^{-2\alpha}} \tag{2.13}$$

where $\alpha(0 \leq \alpha \leq 1)$ is a decay factor that accounts for the amount of power attributed to user k_2 (this amount is increased with α). α is kept constant over the subbands and is determined *a priori* via computer simulations, such that the achieved spectral efficiency is maximized.

If the number of multiplexed users is to be extended to more than 2, inter-subband power allocation can be performed similarly to the case of two multiplexed users. In other words,

the highest channel gain among the $n(s)(s \in S_u)$ scheduled users is considered in the waterfilling process. As for the intra-subband power allocation, power is allocated to each scheduled user $k_n(n \in N_s)$, within each subband s, based on FTPA, such that:

$$P_{s,k_n} = P_s \frac{h_{s,k_n}^{-2\alpha}}{\sum\limits_{j \in N_s} h_{s,k_j}^{-2\alpha}},\ s \in S_u \tag{2.14}$$

2.2.3.3 OPTIMUM WEIGHTED WATERFILLING-BASED POWER ALLOCATION

Similarly to the optimum solution proposed in section 2.2.3.1, we also propose to apply a weighted waterfilling-based power allocation. The weighted version still takes into consideration the channel gains of the two paired users within each subband to achieve inter and intra-subband allocation, but it also adds a proper weight to each user. It is described by the following optimization problem:

$$\underset{\{P_{s,k_1},\, P_{s,k_2}\}}{\text{maximize}} \sum_{s \in S_u} (a_{s,k_1} R_{s,k_1} + a_{s,k_2} R_{s,k_2}) \tag{2.15}$$

Subject to:

$$\sum_{s \in S_u} (P_{s,k_1} + P_{s,k_2}) = P_{rem} \tag{2.16}$$

Where a_{s,k_1} and a_{s,k_2} are positive weights that can be chosen using differing criteria, and that verify:

$$\sum_{s \in S_u} (a_{s,k_1} + a_{s,k_2}) = 1 \tag{2.17}$$

Solving this optimization problem using Lagrange multipliers leads to the formulation of the corresponding objective function J such as:

$$J' = \sum_{s \in S_u} a_{s,k_1} \frac{B}{S} \log_2 \left(1 + \frac{P_{s,k_1} h_{s,k_1}^2}{N_0 \frac{B}{S}}\right) + \sum_{s \in S_u} a_{s,k_2} \frac{B}{S} \log_2 \left(1 + \frac{P_{s,k_2} h_{s,k_2}^2}{P_{s,k_1} h_{s,k_2}^2 + N_0 \frac{B}{S}}\right) + \lambda \left(P_{rem} - \sum_{s \in S_u} (P_{s,k_1} + P_{s,k_2})\right) \tag{2.18}$$

As in 2.7, after differentiating J with respect to P_{s,k_1} and λ, and setting the results to zero, we obtain a non-linear system of $N_u + 1$ equations with $N_u + 1$ unknowns and λ. We also solve this non-linear system using the trust-region Dogleg algorithm, in the same way as was done in section III.B.3.1. The problem of negative solutions is resolved using the second alternative (substituting the negative powers by zero and re-distributing the remaining power). In fact, as it will be shown by the simulation results, the two

alternatives have been tested and the second alternative shows better results.
As for the choice of the weight attributed to every user, two possible schemes are proposed in the sequel.

Weights based on the actual rates of considered users
In order to give importance to users who are far from reaching their requested data rates, a weight is assigned to each user based on the quadratic distance between its actual achieved throughput and its requested data rate, such as:

$$a_{s,k_1} = \frac{(|R_{k_1,requested} - R_{k_1,tot}|)^{2\eta}}{\sum\limits_{s \in S_u} \left(\left|R_{k_1(s),requested} - R_{k_1(s),tot}\right|\right)^{2\eta}},\ 0 \leq a_{s,k_1} \leq 1 \tag{2.19}$$

$$a_{s,k_2} = \frac{(|R_{k_2,requested} - R_{k_2,tot}|)^{2\eta}}{\sum\limits_{s \in S_u} \left(\left|R_{k_2(s),requested} - R_{k_2(s),tot}\right|\right)^{2\eta}},\ 0 \leq a_{s,k_2} \leq 1 \tag{2.20}$$

where $(\eta \geq 0)$ is a control parameter to be determined *a priori* via simulations such that the achieved spectral efficiency is maximized. $R_{k_1,tot}$ (resp. $R_{k_2,tot}$) is the actual total data rate of user k_1 (resp. k_2), and $R_{k_1,requested}$ (resp. $R_{k_2,requested}$) is the requested data rate by user k_1 (resp. k_2). Results in Section IV provide insight on the sensitivity of the system performance to the value of η. In general, increasing η increases the power difference between users being far and users being near to reach their requested data rates.

Weights based on the position of users within the cell In this option, weights are based on the geographical distance between each mobile user and the base station:

$$a_{s,k_1} = \frac{{r_{s,k_1}}^{2\phi}}{\sum\limits_{s \in S_u} {r_{s,k_1}}^{2\phi}},\ 0 \leq a_{s,k_1} \leq 1 \tag{2.21}$$

$$a_{s,k_2} = \frac{{r_{s,k_2}}^{2\phi}}{\sum\limits_{s \in S_u} {r_{s,k_2}}^{2\phi}},\ 0 \leq a_{s,k_2} \leq 1 \tag{2.22}$$

where r_{s,k_1} (resp. r_{s,k_2}) is the Euclidean distance between user k_1 (resp. k_2) and the base station. $\phi(\phi \geq 0)$ is a parameter to be determined offline via simulations such that the cell-edge user throughput and/or the spectral efficiency is maximized.

2.2.3.4 POWER ALLOCATION ACCORDING TO THE ACTUAL ACHIEVED THROUGHPUT

In this technique, power is allocated to users in two successive steps. First, it is partitioned among subbands in a proportional way to the squared distance between the

actual achieved throughput and the requested data rate, using:

$$P_s = \frac{d_{s,k_1}^2 + d_{s,k_2}^2}{\sum\limits_{\mathsf{s} \in S_u} d_{s,k_1}^2 + \sum\limits_{\mathsf{s} \in S_u} d_{s,k_2}^2} \times P_{rem} \text{ , s } \in \text{ S}_u \tag{2.23}$$

where d_{s,k_1} (resp. d_{s,k_2}) is the distance between the actual throughput of user k_1(resp. k_2) over subband s and its requested data rate. Then, power is divided within each subband based on the FTPA algorithm used in section 2.2.3.2.

2.2.3.5 STATIC POWER ALLOCATION

In the majority of existing works related to NOMA, static power allocation is used, where the total transmit power is identically divided between subbands. In order to test a static power allocation scheme within the proposed resource allocation framework, we propose to equally distribute, at each stage of the iterative process, the remaining power P_{rem} among subbands containing users that have not reached their requested data rates yet. The power allocated to subband s is computed as:

$$P_s = \frac{P_{rem}}{N_u}, \text{ s} \in \text{S}_u \tag{2.24}$$

Then, P_s is distributed between paired users on subband s using FTPA.

2.2.4 ADAPTIVE SWITCHING TO ORTHOGONAL SIGNALING

Improvement in spectral efficiency thanks to NOMA is not systematic. Indeed, sometimes the loss in data rate experienced by user k_1, when sharing its subband with user k_2, is greater than the data rate gain achieved by k_2 on this subband. In this case, NOMA is not the appropriate solution; therefore, we propose to allocate this subband to user k_1 alone. The decision to switch to orthogonal signaling (OS) can be made by

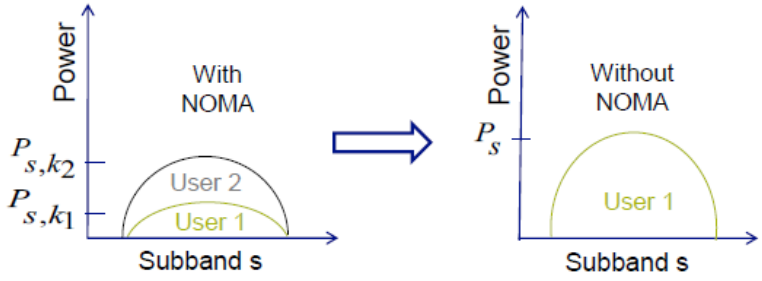


Figure 2.5: Adaptive switching from NOMA to OS

testing the following condition:

$$\gamma\left(R_s - R_{s,k_1}\right) > R_{s,k_2} \tag{2.25}$$

with

$$R_s = \frac{B}{S}\log_2\left(1 + \frac{P_s h_{s,k_1}^2}{N_0 \frac{B}{S}}\right) \tag{2.26}$$

R_s is the data rate achieved on subband s without NOMA. When condition (2.25) is satisfied, the resource allocation technique automatically switches to orthogonal signaling (see Fig. 2.5, for the current subband s). $\gamma(0 < \gamma < 1)$ is a control parameter to be determined *a priori* via simulations such as to maximize the achieved spectral efficiency. In the simulation results, we will study the influence of γ on the system performance. In general, with increasing values of γ, the allocation process tends to switch more often to orthogonal signaling.

2.2.5 DATA RATE ESTIMATION AND CONTROL MECHANISM

At the end of each subband assignment stage, with its power allocation to users k_1 and k_2 (as well as the other users from the set S_u), the algorithm re-estimates the data rates for users in the set S_u (Fig. 1). Then, it verifies if user k_1 (who has been attributed a subband during the latest stage of the algorithm) has reached its requested data rate, that is if the actual total data rate of k_1, $R_{k_1,tot}$, is equal to $R_{k_1,requested}$. In such a case, user k_1 is removed from the set S_u, and the amounts of allocated power on subbands assigned to user k_1 (for k_1 and for the paired users on those subbands) are kept unvaried for the rest of the allocation process. In other words, such subbands will no longer participate in the power distribution step within the allocation process. Such subbands are then removed from the set S_u and their allocated powers are subtracted from the remaining power P_{rem}.

If it happens that the actual data rate is higher than the requested data rate ($R_{k_1,tot} > R_{k_1,requested}$), the total amount of power allocated to user k_1 should be reduced in such a way to reach the exact requested rate. Among the subbands allocated to user k_1 that remain modifiable (i.e. on which k_1 is not paired with a user that has reached its requested rate), we adjust the power on subband s_a having the least channel amplitude. A similar procedure is also applied on user k_2 if it reaches its requested data rate.

When adjusting the transmit power of user k_1 on subband s_a, we encounter two cases.
The first case occurs when user k_1 exhibits the highest channel gain over s_a. The

adjustment is then done as follows:
First, the transmission rate of k_1 over s_a is estimated using:

$$R_{s_a,k_1} = \frac{B}{S}\log_2\left(1 + \frac{P_{s_a,k_1}h^2_{s_a,k_1}}{N_0\frac{B}{S}}\right) \tag{2.27}$$

Then, this rate is subtracted from the actual total rate of user k_1, yielding:

$$R_{rem} = R_{k_1,tot} - R_{s_a,k_1} \tag{2.28}$$

Now, the necessary data rate on s_a is estimated as:

$$R_{k_1,\text{requested}} - R_{rem} \tag{2.29}$$

The power of user k_1 over s_a is modified in such a way to yield the above estimated data rate:

$$P_{s_a,k_1} = \frac{2^{(R_{k_1,\text{requested}} - R_{rem})\frac{S}{B}} - 1}{h^2_{s_a,k_1}} N_0\frac{B}{S} \tag{2.30}$$

Since the power of user k_1 over s_a has been modified, the power of the collocating user should also be reduced according to Eq. 2.8 in order to maintain the same power ratio $(1-\beta_{s_a})/\beta_{s_a}$.

For the second case, when user k_1 exhibits the lowest channel gain over s_a, i.e. not performing SIC over s_a, power adjustment is done by modifying Eq. 2.27 using Eq. 1.5, and keeping Eq. 2.28 and Eq. 2.29 unchanged. Eq. 2.30 is then replaced by Eq. 2.31 using Eq. 2.8 and 1.5.

$$P_{s_a,k_1} = \frac{\left(2^{\left(R_{k_1,\text{requested}} - R_{\text{rem}}\right)\frac{S}{B}} - 1\right) N_0\frac{B}{S}}{\left(1 - \left(\left(2^{\left(R_{k_1,\text{requested}} - R_{\text{rem}}\right)\frac{S}{B}} - 1\right)\frac{\beta_{s_a}}{1-\beta_{s_a}}\right)\right) h^2_{s_a,k_1}} \tag{2.31}$$

Also, the power of the first user on s_a is modified such that to respect the power ratio condition.
Sometimes, when trying to adjust the power of user k_1 over s_a, it can happen that R_{rem} is still greater than $R_{k_1,requested}$. In this case, another subband having a channel gain higher than that of s_a is chosen for power adjustment.
The same kind of power adjustment is performed for user k_2 in case its total actual power is higher than its target data rate ($R_{k_2,tot} > R_{k_2,requested}$).

2.3 COMPARISON WITH OMA

In the majority of existing works dealing with NOMA, the system level-performance is mostly evaluated with respect to OMA [20,22,33], i.e. when a subband is orthogonally

divided in bandwidth and in power between scheduled users. When we assume that OMA signaling is used instead of NOMA within our framework, the bandwidth of subband s is partitioned between users k_1 and k_2 using two factors δ and 1-δ respectively, where $0 < \delta < 1$. In addition, power allocation is performed in two stages: inter-subband power allocation based on the waterfilling principle (sub-optimum technique), followed by intra-subband power allocation based on an equal division of the power between collocating users (i.e. $P_{s,k_1} = P_{s,k_2}$). Since no interference between collocating users occurs in this case, the throughput of user k_i, $(i = 1, 2)$ is computed by:

$$R_{s,k_1} = \frac{B}{S}\delta \log_2\left(1 + \frac{P_{s,k_1} \mathsf{h}^2_{s,k_1}}{\delta N_0 \frac{B}{S}}\right) \tag{2.32}$$

$$R_{s,k_2} = \frac{B}{S}(1\text{-}\delta) \log_2\left(1 + \frac{P_{s,k_2} \mathsf{h}^2_{s,k_2}}{(1-\delta) N_0 \frac{B}{S}}\right) \tag{2.33}$$

In the results of this chapter, OMA signaling is tested when substituting it to NOMA within our allocation technique and comparison towards NOMA is done for the two pairing cases described in Step 3 within Section 2.2.2.

2.4 NUMERICAL RESULTS

In this section, we will evaluate the different design aspects of our proposed algorithm for resource and power allocation.

2.4.1 PERFORMANCE EVALUATION

In this chapter, we mainly consider four system-level performance indicators: the achieved system capacity, the amount of used bandwidth, the probability of success, and the cell-edge user throughput. The first two indicators can be merged into a single metric, the spectral efficiency, calculated as:

$$\text{Spectral\_Efficiency} = \frac{\text{Achieved system capacity}}{\text{Amount of used bandwidth}} \tag{2.34}$$

Several techniques with different combinations of user pairing and multi-user power allocation schemes are evaluated and compared. The following acronyms will be used to refer to the main studied methods:

- NO-O-Opt-Alternative1: Combination of NOMA and OS, with pairing 2 scheme and optimum waterfilling-based power allocation using alternative 1.
- NO-O-Opt-Alternative2: Combination of NOMA and OS, with pairing 2 scheme and optimum waterfilling-based power allocation using alternative 2.

- NO-O-Weighted-Opt-rate-distance: Combination of NOMA and OS, with pairing 2 scheme and optimum weighted waterfilling-based power allocation, where weights are based on distances to target rates of considered users.
- NO-O-Weighted-Opt-position: Combination of NOMA and OS, with pairing 2 scheme and optimum weighted waterfilling-based power allocation, where weights are based on the geographical position of users in the cell.
- NO-O-WF: Combination of NOMA and OS with pairing 2 and sub-optimum waterfilling.
- NO-O-EP: Combination of NOMA and OS with pairing 2 and a static inter-subband power allocation scheme where power is equally divided among subbands (followed by FTPA for intra-subband power distribution).
- NO-O-rate-distance: Combination of NOMA and OS with pairing 2, the power being allocated based on the actual achieved throughput, as was described in Section 2.2.3.4.
- NO-WF: This technique refers to the case when switching to OS (as described in Section 2.2.4) is not allowed (i.e. $\gamma = 0$ in Eq. 2.25). In this case, the allocation process is purely based on NOMA. Pairing 2 and sub-optimum waterfilling are used.
- O-WF: Only OS (classical OFDM) is applied and non-orthogonal cohabitation is not allowed; sub-optimum waterfilling is used.
- O-rate-distance: Classical OFDM is applied, the power being allocated based on the actual achieved throughput in the same way as it is done in NO-O-rate-distance, except that cohabitation is not allowed (k_2 is not present).
- OMA-WF: This technique corresponds to the case when OMA is used instead of NOMA. Bandwidth is equally divided between paired users ($\gamma = 0.5$). Waterfilling is applied as an inter-subband power allocation technique, followed by an equal repartition of power between paired users within each subband.

In this chapter, our simulation setup is chosen such as K users are randomly positioned following a uniform distribution in a 10 km radius cell, with a maximum path loss difference of 20 dB between users. K varies between 5 and 20. The system bandwidth B is 100 MHz and the maximum number of available subbands S is 128. The total transmit power of the BS is 1000 mW, and the requested data rate is set to 5 Mbps for each user. The noise power spectral density is 4.10^{-18} mW/Hz. The transmission medium is modeled by a frequency-selective Rayleigh fading channel with a root mean square delay spread of 500 ns. Perfect knowledge of the channel gains of all users (i.e. perfect CSI) by the BS is assumed in this chapter.

2.4.2 SIMULATION RESULTS

For the performance evaluation process, extensive simulations were conducted and results are given as follows.

2.4.2.1 VALIDATION OF THE DESIGN PARAMETERS

Simulations were first performed to validate the choices of different design parameters within our framework in this chapter, in terms of user pairing, multi-user power allocation, and adaptive switching to OS.

First, we start by identifying the optimal values of the FTPA decay factor α and of the adaptive switching to OS parameter γ. Fig. 2.6 shows the obtained spectral efficiency when NO-O-WF is evaluated for different α and γ values with $K = 10$ and the actual number of available subbands S_A is equal to 128 ($S_A = S$). Spectral efficiency is maximized for $\alpha = 0.5$ and $\gamma = 0.5$. Similar optimal values were observed for different values of K and S_A. Therefore, these values of α and γ are adopted in the sequel. Now, the impact of user pairing and intra-subband power allocation strate-

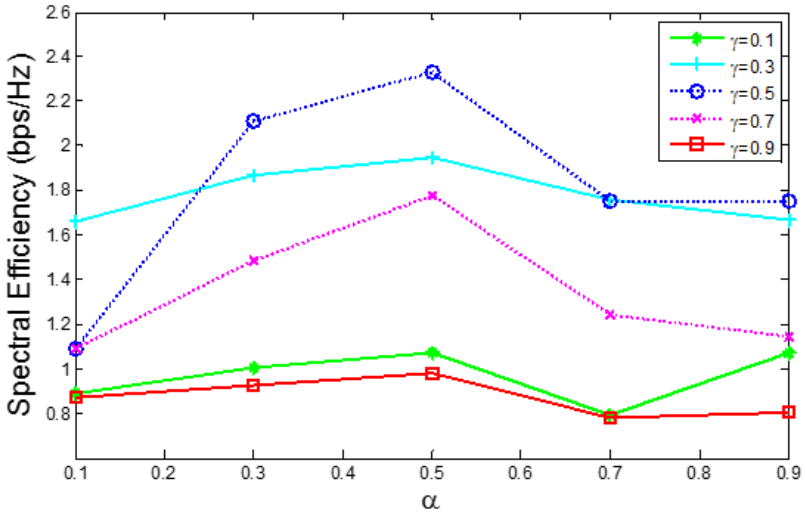


Figure 2.6: Spectral efficiency of NO-O-F for different values of α and γ.

gies on system performance is evaluated, for $S_A = 128$. Fig. 2.7 shows the spectral efficiency of NO-O-WF when FTPA and FPA are used. The effect of the two pairing techniques presented in section III.B.1 is also shown in the same graph. We notice that the combination between FTPA and pairing 2 outperforms FPA for different values of β, with a gain ranging from 16% when the number of users is high, up to 40% when

the number of users per cell is equal to 5. This is due to the fact that FTPA allocates powers dynamically, by taking into account the encountered channel states by all users, whereas static repartition within a subband may be inconvenient for some paired users. In addition, the use of pairing 2, as implemented within the proposed allocation technique, is a reasonable choice since the performance gain of NOMA compared to orthogonal signaling increases with the difference in channel gains between scheduled users [16]. Therefore, using pairing 2, together with FTPA, shows the best performance. In fact, when users having the largest possible gain difference within each subband are

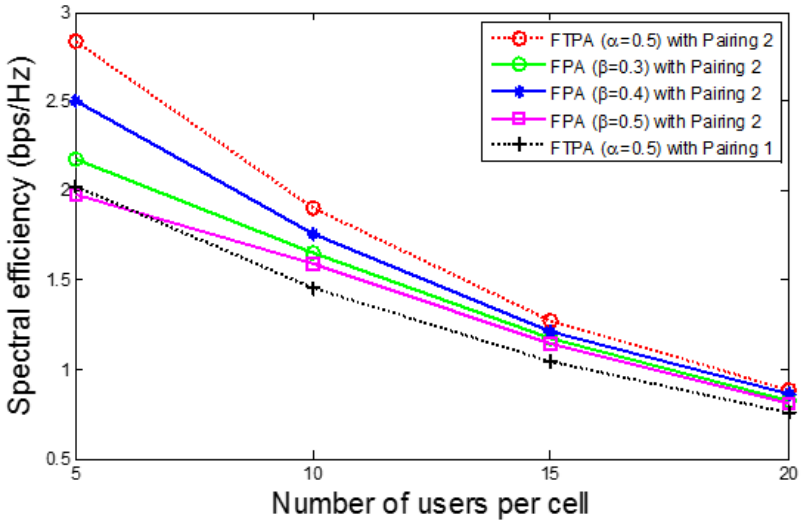


Figure 2.7: Spectral efficiency of NO-O-WF for dynamic and fixed intra-subband power allocation schemes, and for different channel gain differences between paired users.

paired together (pairing 2), the power difference between their received signals should be large too thanks to the application of FTPA (see eq. (19)). Therefore, the amount of inter-user interference experienced by user k_2 (i.e. the $P_{s,k_1}h^2_{s,k_2}$ term in the denominator of R_{s,k_2}) is reduced, not only due to the choice of user k_2 (by reducing h^2_{s,k_2}) but also because the signal power P_{s,k_1} of user k_1 is lowered.

Next, in order to identify the optimal value of η that should be used as a design parameter, we evaluate the spectral efficiency of NO-O-Weighted-Opt-rate-distance for different values of η, when the number of users per cell varies between 5 and 20, and for an actual number of available subbands equal to 128. In fact, increasing η increases the amount of power attributed to the users that are far from reaching their target, at the expense of decreasing the amount of power allocated to the users whose data rates are close to the target. Fig. 2.8 shows that the incidence of η on system performance is especially significant when the number of users per cell is small, and that the spectral

efficiency is maximized when $\eta = 1$; therefore, this value is adopted in the sequel.

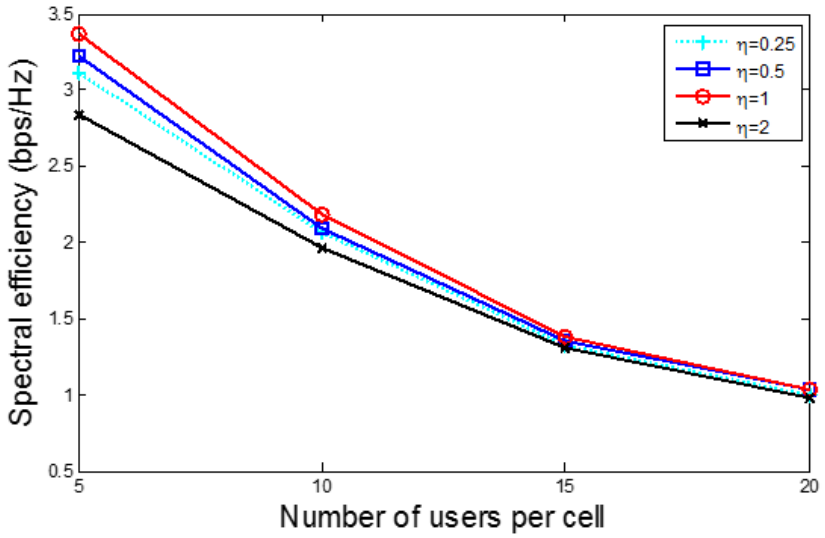


Figure 2.8: Spectral efficiency of NO-O-Weighted-Opt-rate-distance for different values of η

2.4.2.2 PERFORMANCE EVALUATION OF THE DIFFERENT PROPOSED TECHNIQUES

Now that the main system parameters have been chosen, the performance gain of the different proposed techniques is investigated for two different setups:
Case 1: The number of users per cell is equal to 10 and the actual number of available subbands ranges from 16 to 128, with a fixed subband bandwidth at $100/128$ MHz.
Case 2: The actual number of available subbands is 128 ($S_A = S$) and the number of users per cell is varied between 5 and 20.

Fig. 2.9 and Fig. 2.10 compare the spectral efficiency of the simulated methods, for the simulation setups 1 and 2 respectively.
Let us first compare the four proposed allocation schemes NO-O-WF, NO-WF, NO-O-EP, and O-WF. In both setup cases, NO-O-WF outperforms the other 3 simulated methods. This gain in spectral efficiency is due to several reasons:

- The reduction in the amount of used bandwidth due to non-orthogonal cohabitation in the power domain makes NO-WF and NO-O-WF clearly outperform O-WF;

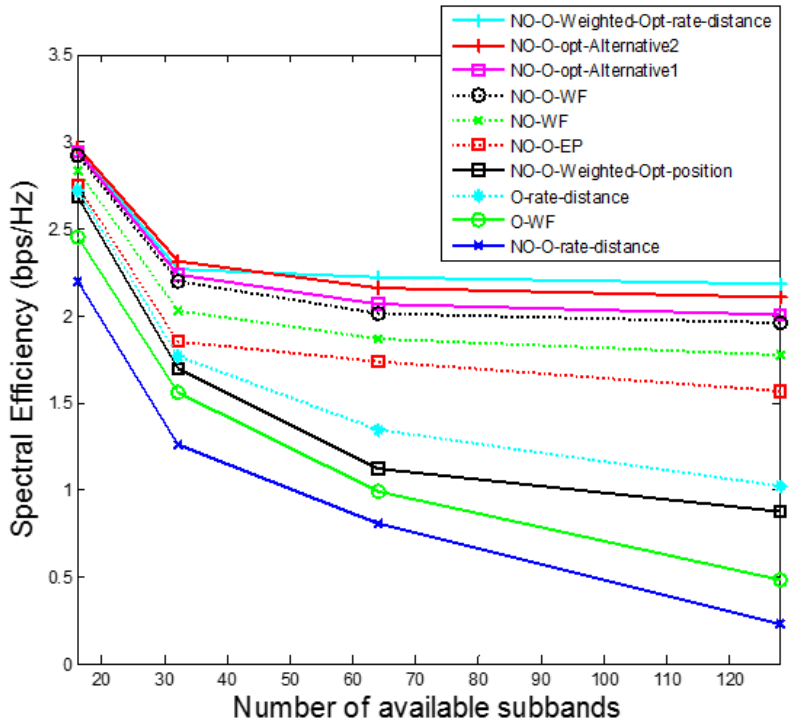


Figure 2.9: Spectral efficiency of the proposed methods in terms of number of available subbands

i.e., a smaller total number of subbands is needed to achieve the data rate targets.

- The improvement in system capacity due to the waterfilling process helps NO-WF and NO-O-WF outperform NO-EP. This capacity is expressed by the total throughput $\sum_{k=1}^{K} \sum_{s \in S_k} R_{s,k}$ achieved at the end of the allocation process, which is also the sum of target data rates reached by users in case of success.

- The use of a dynamic adaptive switching to orthogonal-based system improves NO-O-WF performance with respect to NO-WF.

For instance, with 32 available subbands, and for a number of users per cell equal to 10, NO-O-WF has a spectral efficiency of 2.2 bps/Hz compared to 2.02, 1.85, and 1.5 bps/Hz with NO-WF, NO-EP, and O-WF, respectively. When the number of available

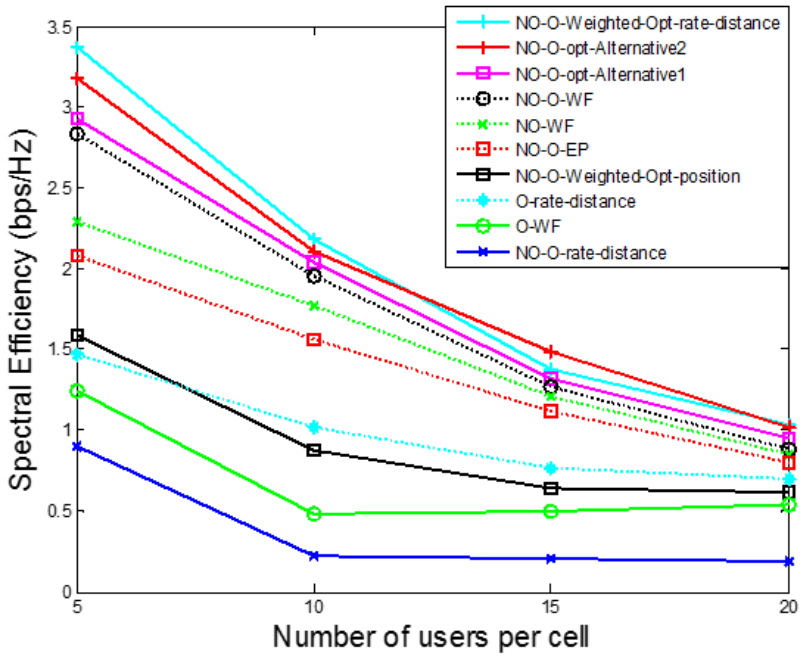


Figure 2.10: Spectral efficiency of the proposed methods in terms of number of users per cell

subbands decays, the advantages of NO-O-WF, in terms of the achieved total data rate and the reduced amount of necessary bandwidth, are maintained. For example, when this number drops to 16, the observed spectral efficiency remains in favor of NO-O-WF and is respectively 2.9, 2.8, 2.7, and 2.4 bps/Hz for NO-O-WF, NO-WF, NO-EP, and O-WF. This shows the efficiency of non-orthogonal signaling in congested areas (i.e. when the amount of bandwidth is low with respect to the number of users).

The two proposed optimal techniques, NO-O-Opt-Alternative1 and NO-O-Opt-Alternative2 show a significant improvement in the system performance with respect to the sub-optimum technique NO-O-WF, especially when the second alternative is applied, i.e., when negative values are replaced by zero followed by a power re-distribution. This result is due to the fact that these two techniques allow a joint inter and intra power distribution by resolving the optimization problem formulated in (2.5), rather than dividing the allocation process into two separate inter and intra stages, as in NO-O-WF and NO-WF. By doing so, the channel gains of all paired users are jointly taken into consideration when distributing power between subbands, instead of considering only the channel gain of the user with the highest gain in each subband.

When weights are introduced, the performance in the spectral efficiency of NO-O-Weighted-Opt-rate-distance is improved with respect to NO-O-Opt-Alternative1 and NO-O-Opt-Alternative2. Taking into account the distance between the actual throughput and the requested data rate for each user, in the power distribution, reveals to be more efficient that the non-weighted solutions, since it allows giving more priority and thus more power to users that are far from achieving their requested data rates. At the same time, users weighted throughputs are being maximized using the waterfilling process. The gain obtained with NO-O-Weighted-Opt-rate-distance with respect to NO-O-Opt-Alternative2 is observed for a low number of users and a high number of available subbands.

As for NO-O-Weighted-Opt-position, this method presents an important loss in the average spectral efficiency, compared to NO-O-WF, due to the use of geographical position as weight. This weighted version tends to give more priority in the power allocation to cell-edge users. In fact, users being far from the BS require a high level of power in order to reach their requested data rates, since their channel gains on their allocated subbands generally tend to be low. In this case, increasing their priority will cause the remaining users to be given low levels of power; therefore, those users will need more subbands in order to reach their requested data rates. At the same time, the amount of power taken from the close users does not significantly increase the data rate of far users (because of their low channel gains), leading to a decrease in the overall spectral efficiency. However, as will be shown later in this section, the advantage of this technique resides in a slightly increased cell-edge user throughput, compared to the other allocation techniques.

On the other hand, when the power is directly distributed according to actual rates of users, without maximizing the total achieved throughput as done in NO-O-rate-distance, we fail to outperform O-WF. In NOMA, allocating power to a user in a proportional way to the squared distance between its actual throughput and its target data rate, without using the waterfilling principle, does not directly guarantee a maximization of the worst user throughput neither the total user throughput. This is mainly due to the collocation principle: Suppose k_1 is the least privileged user (the farthest from his target), at a certain stage of the algorithm. Since its distance d^2_{s,k_1} in eq. (2.23) is high, this user should get a higher amount of power on its allocated subbands, compared to other users. However, since the weight on any of its subbands also takes into consideration the distance of the collocating user d^2_{s,k_2}, the amount of power attributed to a certain subband of k_1 will be lower than necessary if d^2_{s,k_2} is low. This leads to a power allocation far from being optimal. However, when applying this allocation technique to OS (O-rate-distance), i.e. without user cohabitation, this strategy shows to be of a great interest for classical OFDM, since it outperforms the widely used waterfilling technique

(O-WF). This is explained by the fact that our target in this work is to allow users reach their requested data rates, and not only to maximize the average throughput, which is commonly achieved by the waterfilling process. For classical OFDM, this target is reached much more quickly (i.e. with a lower number of subbands) with a power allocation proportional to rate distances, than with waterfilling.

Since our goal throughout this study is to allow users reach their requested data rates,

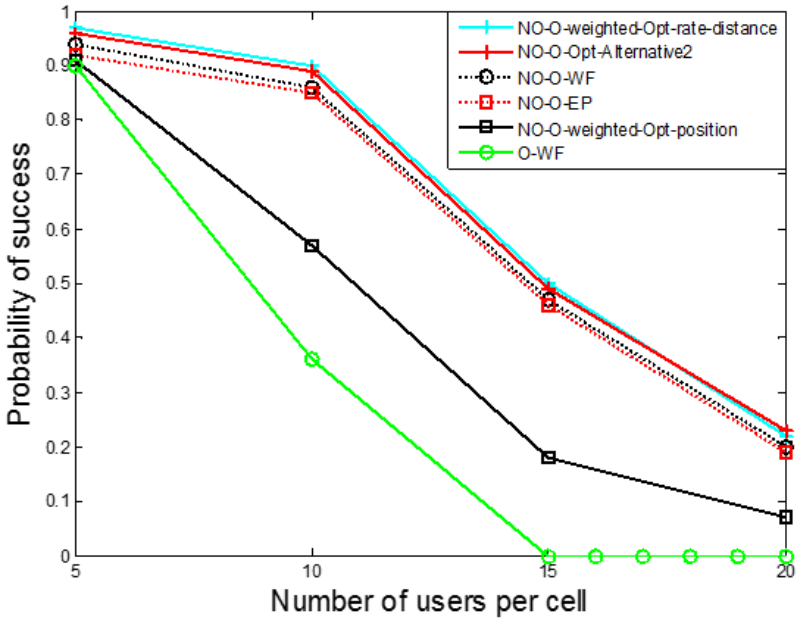


Figure 2.11: Probability of success of the proposed methods in terms of the number of users per cell

while maximizing the total achieved throughput, the proposed techniques should also be compared based on their abilities to reach this goal. When users reach their requested data rates, the success flag of the proposed allocation technique is set to 1, otherwise, it is set to zero (Fig. 2.1). For a high number of conducted simulations, Fig. 2.11 shows the average success probability of the proposed techniques for different numbers of users per cell, and an actual number of available subbands equal to 128. Although the success rate decreases when the number of users per cell increases, the proposed techniques based on NOMA greatly outperform O-WF for a large numbers of users. For instance, starting from a number of users of 15, O-WF fails to provide a solution for the allocation problem, whereas the other proposed techniques based on NOMA still succeed with a probability between 23% (for $K = 20$) and 50% (for $K = 15$).

The cell-edge user throughput is an important fairness evaluator for the allocation

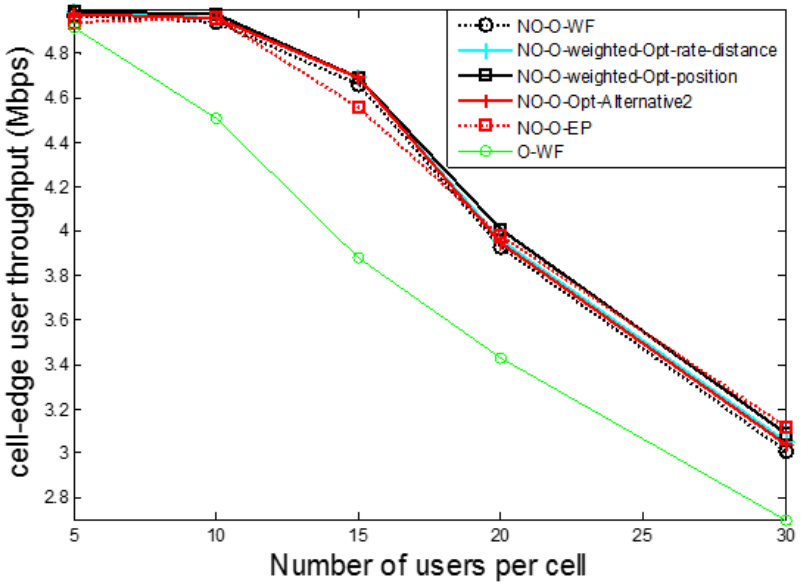


Figure 2.12: Cell-edge user throughput as a function of the number of users per cell.

process. Fig. 2.12 shows this metric as a function of the number of users per cell, where the number of available subbands is fixed to 128.

The cell-edge user throughput when using NOMA is always higher than that of orthogonal signaling. The gain can reach 21% approximately with NO-O-Weighted-Opt-position, with respect to O-WF. The proposed technique for power allocation based on user position NO-O-Weighted-Opt-position shows the best performance. This is due to the fact that the rates in the optimized sum-rate function are weighted proportionally to the distance of users towards the BS. In this case, cell-edge users are given a high level of power, in order to increase their chance to reach the requested data rates.

It should also be noted that, when the number of users per cell is limited, waterfilling-based power allocation (especially NO-O-Weighted-Opt-position) shows a slightly higher cell-edge user throughput compared to equal power allocation. However, when the number of users per cell becomes large, the success rate of all allocation strategies generally decreases (Fig. 2.11). In such conditions, adopting a uniform power allocation or allocating powers while taking into consideration the geographical distribution of users yields similar cell-edge user throughput, which is higher than that of other allocation techniques. This is due to the fact that non-weighted waterfilling-based algorithms generally optimize the average throughput and may not give the best fairness to the cell-edge user, especially for large values of K.

Therefore, we can conclude that when it comes to optimizing the cell-edge user throughput in crowded areas, equal power repartition may perform as good as a weighted

optimized repartition, and better than a repartition that does not consider geographical positions. Nevertheless, the proposed non-weighted approaches still present important gains when compared to orthogonal signaling, not only in terms of spectral efficiency, but also in the degree of fairness.

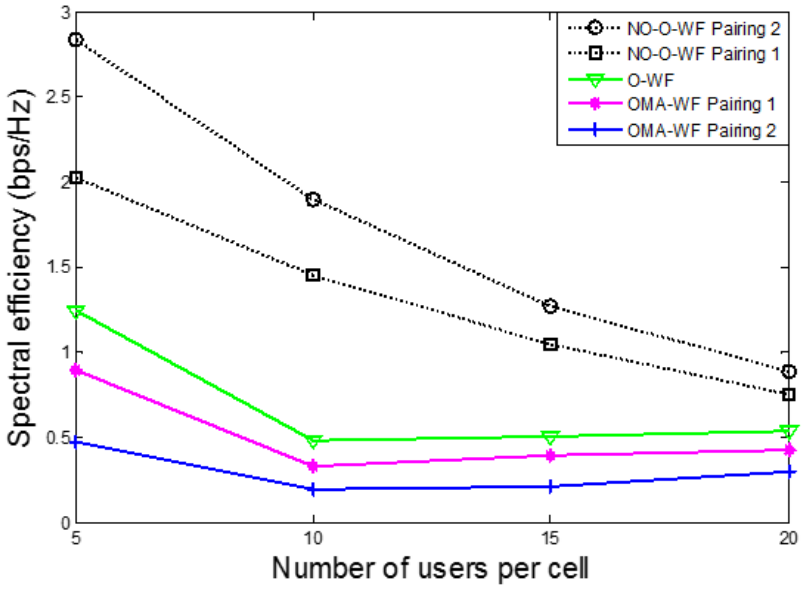


Figure 2.13: Comparison of the spectral efficiency of NOMA and OMA for different channel gain differences between paired users.

2.4.2.3 COMPARISON WITH OMA

In order to compare the proposed allocation techniques based on NOMA to the case where OMA is used within our allocation framework (Section 2.3), Fig. 13 shows the comparative results between NO-O-WF, O-WF, and OMA-WF, with the two pairing options described in Step 3 within Section 2.2.2.
NO-O-WF, with the two pairing options, shows significantly better performance than both O-WF and OMA-WF. On the other hand, as opposed to NOMA, OMA-WF based on pairing 1 gives better performance than pairing 2, because in OMA, there is no interference between users scheduled on the same subband, since the collocation is done by dividing both power and bandwidth. Therefore, the channel gain of the second collocating user does not need to be minimized, with respect to the first user, as is done in NOMA. Furthermore, it is shown that O-WF gives better performance that OMA-WF.

This is due to the fact that the total throughput achieved per subband, when using OMA, is often lower than when assigning the subband to the user with the best channel gain on that subband. For instance, consider the practical example given in [16], where two users were assumed per subband with user 1 a cell-interior user, and user 2 a cell-edge user, such that $h_1^2 = 10h_2^2$. If power and bandwidth are equally allocated among the two users using OMA, the user rates are found in [16] as 3.33 bps and 0.50 bps, respectively for user 1 and user 2. On the other hand, using NOMA, when the power is allocated as $P_1 = P_2/4$, user rates are found to be 4.39 bps and 0.74 bps, for user 1 and user 2 respectively. In this case, NOMA provides higher sum rate than OMA. However, what the authors did not consider in this example is the case where the subband is allocated to only 1 user, for instance the user with the highest channel gain on this subband. In such a case, its achieved rate would be equal to 6.66 bps, which is higher than 5.13 bps, the total rate achieved by the two users when NOMA is used, and also higher than 3.83 bps, the total rate achieved by the two users when OMA is used. In this case, orthogonal signaling without subband division is the most appropriate solution. This explains the superiority of O-WF on OMA-WF in our work, and also the importance of integrating an adaptive switching from NOMA to orthogonal signaling in the allocation technique, in order to account for such situations.

The power delivered by the base station is as critical to system performance as system

Figure 2.14: Comparison of the spectral efficiency of NOMA and OMA for different channel gain differences between paired users.

bandwidth. For this sake, we propose to study the influence of increasing the total transmit power P_{max} and decreasing system bandwidth B. As an example, we take $B = 50$ MHz and $P_{max} = 40$ W (46 dBm).
Fig. 2.14 shows the spectral efficiency for different numbers of users per cell and an actual number of available subbands equal to 128. The proposed optimal algorithms NO-O-Opt-Alternative2 and NO-O-Weighted-Opt-rate-distance still show significantly better performance than sub-optimum techniques.
Besides, a higher gain in performance is noticed when compared to NO-O-EP and O-WF, with respect to the case where $B = 100$ MHz and $P_{max} = 1$ W, especially for a high number of users.

In the aim of assessing the implementation feasibility of the different proposed allocation techniques, we measured the computational load of the main resource allocation methods to be integrated at the Base Station. At the downlink receiver side, we estimated the average number of SIC procedures that are needed for a user to recover its useful information. The second measurement allows us to gain insight into the complexity increase, in mobile receivers, with respect to a resource allocation based on classical orthogonal signaling. Measurements, reported in Table 2.1, were conducted using Mat-

Table 2.1: Complexity analysis of the simulated methods

Simulated Methods	Execution time (ms) at the BS	Average number of SIC per user at the terminal user
NO-O-weighted-opt-rate-distance	3858	0.57
NO-O-weighted-opt-position	3857	5.38
NO-O-opt-alternative1	3570	0.65
NO-O-opt-alternative2	3580	0.62
NO-O-WF	150	0.68
NO-WF	100	10.80
NO-O-EP	50	4.54
O-W	36	0

lab, run under windows 8, on an intel core i3 CPU, for the case where $K = 10$ users and $S_A = 128$ available subbands. The results in the second row correspond to the average execution time of one simulation of the whole allocation process.

We can see that the additional complexity driven by NOMA, when applied with a suboptimum waterfilling process, is affordable in comparison to NO-EP and O-WF. However, when optimum power allocation is used, the computational load increases significantly, due to the use of the numerical solver. Therefore, a compromise has to be made between the complexity and the accuracy of the proposed techniques. Such a choice would depend on the requirements of the application in use. However, it should be noted that the allocation algorithms only need to be applied at the BS side, where sufficient hardware and memory resources are supposed to be available to allow efficient and real-time implementations.

When observing the results in the third row of Table 2.1, together with the comparative results of Fig. 2.9 and Fig. 2.10, we can say that the increase in complexity at the base station allows a better allocation of spectral resources, and therefore a reduction in the average number of subbands allocated to each user. This, in turn, yields a significant decrease in complexity at the user terminal due to the reduced number of executed SIC per user. For instance, the optimum weighted waterfilling-based power allocation requires an average of 0.57 SIC per user, compared to 0.62 or 0.65 SIC with the non-weighted optimum approaches, to 0.68 SIC when suboptimum waterfilling is applied, to 4.54 SIC when an equal power inter-subband allocation is used, and to 10.80 SIC when suboptimum waterfilling is applied without the dynamic switching to orthogonal signaling.

2.5 CONCLUSION

This chapter introduces a new framework for bandwidth and power allocation under a non-orthogonal multiple access scenario. It targets minimizing spectrum usage while satisfying requested data rates by a set of users. Several design issues are thoroughly investigated within the proposed approaches: the choice of user pairing, optimal or sub-

optimum power allocation, fixed and adaptive intra-subband power allocation, dynamic switching from NOMA to orthogonal signaling, weighting strategies for the optimized sum-rate function, etc. Simulation results show that the proposed framework allows a significant increase in spectral efficiency and in the probability of success, especially when compared to a system purely based on either orthogonal or non-orthogonal signaling. Furthermore, a joint inter and intra subband power allocation obtained by numerically solving an optimized allocation problem yields a substantial gain in performance compared to the suboptimum solutions. Moreover, the adoption of appropriate weights in the optimized sum-rate metric shows promising enhancements to either spectral efficiency, cell-edge user throughput, or the necessary number of SIC per user. Such weighting strategies can be used in several practical applications to allow different types of user prioritization. Channel and power allocation should also be studied in the context of a fixed amount of total used bandwidth, where no target rates constraints are set, and where the emphasis is on maximizing the total cell throughput as well as the user fairness. Consequently, attention will be devoted in the following chapter toward such scenarios.

Chapter 3

Proposals to improve the PF scheduler for a NOMA system

Results of chapter 2 have shown the great interest of NOMA compared to orthogonal multiple access (OMA) in a context where the amount of used bandwidth has to be minimized, constrained by a total transmit power budget at the base station (BS). Fairness was achieved by setting user target rates. However, this can be inappropriate for certain applications or services where users are not supposed to be guaranteed fixed data rates. Therefore, in the current chapter, we aim to maximize throughput and fairness in a context where the total bandwidth is available to be used and where there is no target data rates constraint.

To do so, we first investigate the performance of the proportional fairness (PF) scheduler, since it is the mostly used scheduler in the majority of papers dealing with NOMA. It is known to provide a good tradeoff between total user throughput and user fairness. Then, we propose modifications to the PF scheduler, in order to improve its performance in terms of total system throughput and user fairness. Modifications are proposed at the level of user scheduling and power allocation. In this sense, several weighted versions of the PF scheduling metric are developed and an iterative waterfilling process is designed to be incorporated within the PF. The proposed techniques were designed for NOMA schemes but are proven to be also applicable, and of interest, for orthogonal multiple access (OMA) schemes.

This chapter is organized as follows: In section 3.1, the conventional PF scheduler is described. Proposals to improve the PF scheduler at the level of user scheduling are given in Section 3.2. Weighted versions of the PF metric for a NOMA-based system and OMA-based system are proposed and developed in Section 3.2.1 and 3.2.2, respectively. The first scheduling slot is treated differently and a corresponding scheduling metric is derived in Section 3.2.3. Delivering different levels of service is also proposed in Section 3.2.4. The complexity analysis of the proposed metrics is carried out in Section

3.2.5. The performance analysis related to these enhancements takes place in Section 3.3. Afterwards, proposals to improve the PF scheduler at the level of power allocation are investigated in Section 3.4, and a corresponding iterative waterfilling-based power allocation technique is proposed to be incorporated within the PF in Section 3.4.1. Its performance analysis is performed in Section 3.4.2 and its computational complexity is detailed in Section 3.4.3. Finally, Section 3.5 concludes the chapter.

3.1 CONVENTIONAL PROPORTIONAL FAIRNESS (PF) SCHEDULING SCHEME

The proportional fairness scheduler has been used extensively in OMA based systems in order to manage the assignment of radio resources between users. The choice of the PF scheduler is reasonable due to the balance that it provides between fairness and system capacity. Kelly et al. [66] have defined the proportional fair allocation of rates, and used a utility function to represent the degree of users' satisfaction. In [67], the operation of the PF scheduler is detailed: at the beginning of each scheduling slot, each user provides the base station with its channel state or its feasible rate. The scheduling algorithm keeps track of the average throughput $T_k(t)$ of each user in a past window of length t_c. $T_k(t)$ is also known to be the fairness index which is fixed prior to the current allocation In the scheduling slot t, user k^* is selected to be served based on:

$$k^* = \arg\max_k \frac{R_k}{T_k} \tag{3.1}$$

where R_k is the feasible rate of user k in scheduling slot t, and T_k is its moving average.

In [82], an approximated version of the PF scheduler for multiple users transmission is presented. This version has been adopted in the majority of the works dealing with NOMA [6, 8, 83] in order to select users to be non-orthogonally scheduled on available resources.
It is assumed that, for a subband s under consideration, the PF metric is estimated for each possible combination of users U, and the combination that maximizes the PF metric will be denoted by U_s:

$$U_s = \arg\max_U \sum_{k \in U} \frac{R_{s,k}(t)}{T_k(t)} \tag{3.2}$$

U is a possible candidate user set, $R_{s,k}(t)$ is the instantaneous achievable throughput of user k on subband s at time instance t, the time index of a subframe.
Note that the total number of combinations tested for each considered subband s is:

$$N_U = \binom{1}{K} + \binom{2}{K} + ... + \binom{n(s)}{K} \tag{3.3}$$

where K is the total number of users per cell, and $n(s)$ is the maximum number of multiplexed users at subband s.
$R_{s,k}(t)$ is calculated based on 1.3 (defined in Chapter 1), whereas $T_k(t)$ is updated as follows [82] :

$$T_k(t+1) = \left(1 - \frac{1}{t_c}\right) T_k(t) + \frac{1}{t_c} \sum_{s=1}^{S} R_{s,k}(t) \tag{3.4}$$

Parameter t_c defines the throughput averaging time window, i.e. the number of simulated subframes. In other words, it is the time horizon in which we want to achieve fairness. t_c is chosen such that a tradeoff between system performance (in terms of fairness) and system capacity is guaranteed.

3.2 IMPROVING THE PF METRIC AT THE LEVEL OF USER SCHEDULING

Although the PF scheduler represents a good compromise between system throughput and fairness, it has been always a matter of further studies in order to improve its performance.
Aiming at enhancing the gain of the cell-edge user, a research group has proposed in [84], a weighted PF-based multiuser scheduling scheme in a non-orthogonal access downlink system. A frequency block access policy is proposed for cell-interior and cell-edge user groups using fractional frequency reuse (FFR), with significant improvements in the user fairness and system frequency efficiency. In [85], an improved downlink NOMA scheduling scheme based on the PF scheduler is proposed and evaluated. The proposed scheme aims at taking the fairness of the target frame into consideration. It shows an improved performance compared to the conventional PF scheduler. Similarly to the work done in [84], several papers have proposed weighted versions of the PF scheduler, in the aim of improving user fairness. Those papers have considered, in their majority, an OMA-based scheduling. In [86], fair weights have been implemented for opportunistic scheduling of heterogeneous traffic types for OFDMA networks. For designing fair weights, the proposed scheduler takes into account average channel status as well as resource requirements in terms of traffic types. Simulation analysis demonstrates the efficiency of the proposed scheduler in terms of resource utilization, and flexibility to network characteristics change due to user mobility. In [87], the problem of fairness deficiency encountered by the PF scheduler when the mobiles experience unequal path loss is investigated. To mitigate this issue, a modified version of the PF scheduler that introduces distance compensation factors has been proposed. It was shown that this solution achieves both high capacity and high fairness. In [88], a weighted proportional fair algorithm is proposed in order to maximize best-effort service utility. The reason behind introducing weight factors to the PF metric is to exploit the inherent near-far di-

versity given by the pathloss. The proposed algorithm enhances both best-effort service utility and throughput performance while maintaining similar complexity when compared to the PF metric.

Inspired by the aforementioned studies done on the PF scheduler, we aim in this section to stress on its drawbacks in order to improve its performance in a NOMA-based system as well as on an OMA-based system. The objective of the PF is to maximize the logarithmic sum of users throughputs [89] or, equivalently, the long-term averaged user rates, in order to ensure balance between user fairness and cell throughput. By doing so, the PF scheduler targets long-term fairness, since, among the optimization elements in a conventional PF metric, the fairness index is calculated and fixed prior to the current allocation. However, short-term fairness-aware scheduling and fast convergence towards required performance are important issues to be addressed in upcoming mobile standards [90]. For this sake, the fairness index should be considered and updated during scheduling.

Since all possible combinations of candidate users are tested for each subband, a user might be selected more than once and attributed multiple subbands during the same time slot. However, it can also happen that a user will not be allocated any subband whenever its historical rate is high. In such a case, the user will not be assigned any transmission rate for multiple scheduling slots. This behavior can be very problematic in some applications, especially those requiring a quasi-constant Quality of Experience (QoE) such as multimedia transmissions. In order to cope with these challenges, buffering may be needed. However, such a scenario may not be compatible with applications requiring low latency transmission.

In addition, some applications call for schedulers which provide Quality of Service (QoS) constraints such as fairness and minimum throughput guarantees [91]. However, the PF scheduler in its conventional formulation does not provide service differentiation, so the user could not be aware of what he has paid for and the operators cannot assure that they deliver what they promise. Although providing these requirements is a hard nut to crack, in this chapter, we propose modifications to the PF metric in a way to allow service differentiation.

In order to respond favorably to the aforementioned requirements, we propose several weighted proportional fairness scheduling metrics that aim at:

- Improving the user capacity, thus enhancing the total achieved user throughput,
- Improving the convergence time towards required fairness performance, thus QoE,
- Enhancing fairness among users (both long-term and short-term fairness),

- Ensuring steadier user data rates,
- Giving the possibility of delivering different levels of Quality of Service (QoS).

3.2.1 Proposed Weighted NOMA-based Proportional Fairness Scheduler (WNOPF)

The proposed scheduler consists of introducing fair weights to the conventional NOMA-based PF metric. The main goal of the weighted version is to ensure fairness among users in every scheduling slot. This will allow avoiding the occurrence of a zero transmission data rate for any user at any time scale. To do so, we modify the PF metric expression to take into account the status of the current assignment in time slot t. Therefore, the priority given for each user must be based not only on its historical rate but also on its current total achieved rate (throughput achieved during the current scheduling slot t), as it was done in [85].

The scheduling is performed subband by subband and on a time slot basis. For each subband s, the conventional PF metric PF_s^{NOMA} and a weight factor $W(U)$ are both calculated for each candidate user set U. Then, the scheduler selects the set of scheduled users U_s that maximizes the weighted metric $PF_s^{NOMA}(U) \times W(U)$. The corresponding scheduling method is referred to as Weighted NOMA PF scheduler, denoted by WPF^{NOMA}. The resource allocation problem can be formulated as follows:

$$\begin{aligned} WPF_s^{NOMA}(U) &= PF_s^{NOMA}(U) \times W(U) \\ U_s &= \arg\max_U WPF_s^{NOMA}(U) \end{aligned} \tag{3.5}$$

The scheduling metric PF^{NOMA}, defined in Eq. 3.2, can guarantee the proportional fairness criterion by maximizing the sum of users service utility which can be formally written by [92]:

$$PF^{NOMA} = \max_{scheduler} \sum_{k=1}^{K} \log T_k \tag{3.6}$$

Where T_k denotes the average historical rate of user k, for a total observation time t_c.

So, the proposed weighted metric WNOPF achieves higher service utility compared to the conventional PF scheduler, if:

$$\sum_{k=1}^{K} \log T_k \geq \sum_{k=1}^{K} \log T_k^{'} \tag{3.7}$$

where the historical rates T_k and $T_k^{'}$ refer to the schedulers using the WNOPF metric and the conventional PF metric, respectively.

Proposition 1: To make Eq. 3.7 valid, for a NOMA-based system, the following inequality should be verified:

$$\prod_{k=1}^{K} E\left[W\left(U_k\right) / \sum_{U} W\left(U\right)\right] \prod_{k=1}^{K} E\left[R_{s,k}\right] \geq \prod_{k=1}^{K} E\left[R_{s,k}^{'}\right] \tag{3.8}$$

where $E[R_{s,k}]$ and $E[R_{s,k}^{'}]$ are the statistical average of the instantaneous transmittable rate of user k on a subband s, when WNOPF and the conventional PF scheduler are applied respectively, U_k denotes a scheduled user set containing user k, U is a possible candidate user set, and $E\left[W\left(U_k\right) / \sum\limits_{U} W\left(U\right)\right]$ is the statistical average of the normalized weight of the set U_k.

Proof:
Eq. 3.7 can be written as

$$\prod_{k=1}^{K} T_k \geq \prod_{k=1}^{K} T_k^{'} \tag{3.9}$$

If we consider that $T_k = R_{k,tot}/N_{tot}$, where $R_{k,tot}$ is the total amount of information receivable by user k, for a total observation time t_c, and N_{tot} is the total number of time slots within t_c, we obtain:

$$\prod_{k=1}^{K} \frac{R_{k,tot}}{N_{tot}} \geq \prod_{k=1}^{K} \frac{R_{k,tot}^{'}}{N_{tot}} \tag{3.10}$$

If we denote by N_k the number of allocated time slots by user k within t_c, and n_k the statistical average of the number of allocated subbands by user k per time slot, Eq. 3.10 can be re-written as:

$$\prod_{k=1}^{K} \frac{N_k n_k E\left[R_{s,k}\right]}{N_{tot}} \geq \prod_{k=1}^{K} \frac{N_k^{'} n_k^{'} E\left[R_{s,k}^{'}\right]}{N_{tot}} \tag{3.11}$$

By doing a simple rearrangement, we get:

$$\frac{\prod\limits_{k=1}^{K}\left(N_k/N_{tot}\right) S\left(n_k/S\right)}{\prod\limits_{k=1}^{K}\left(N_k^{'}/N_{tot}\right) S\left(n_k^{'}/S\right)} \geq \frac{\prod\limits_{k=1}^{K} E\left[R_{s,k}^{'}\right]}{\prod\limits_{k=1}^{K} E\left[R_{s,k}\right]} \tag{3.12}$$

If Pr_k $(= N_k/N_{tot})$ denotes the probability of user k being scheduled per time slot and pr_k $(= n_k/S)$ the statistical average probability of user k being scheduled per subband,

where S is the total number of subbands per time slot, Eq. 3.12 can be formulated as:

$$\frac{\prod_{k=1}^{K} Pr_k pr_k}{\prod_{k=1}^{K} Pr_k^{'} pr_k^{'}} \geq \frac{\prod_{k=1}^{K} E\left[R_{s,k}^{'}\right]}{\prod_{k=1}^{K} E\left[R_{s,k}\right]} \tag{3.13}$$

pr_k can be regarded as the statistical average probability of a set U_k, $(= E\left[\Pr\left(U_k\right)\right])$, being chosen among all possible candidate sets U to be scheduled per subband. it is calculated as follow:

$$pr_k = E\left[\Pr(U_k)\right] = E\left[\Pr\left(PF^{NOMA}\left(U_k\right) W\left(U_k\right)\right)\right] \tag{3.14}$$

Since the conventional PF metric PF^{NOMA} and the weight calculation are independent, pr_k will be equal to:

$$\begin{aligned} pr_k &= E\left[\Pr\left(PF^{NOMA}\left(U_k\right)\right)\right] E\left[\Pr\left(W\left(U_k\right)\right)\right] \\ &= pr_k^{'} E\left[W\left(U_k\right) / \sum_{U} W\left(U\right)\right] \end{aligned} \tag{3.15}$$

where $E\left[W\left(U_k\right) / \sum_{U} W\left(U\right)\right]$ is the statistical average of the normalized weight of a set U_K.
Thus, we obtain:

$$\frac{\prod_{k=1}^{K} Pr_k pr_k^{'} E\left[W\left(U_k\right) / \sum_{U} W\left(U\right)\right]}{\prod_{k=1}^{K} Pr_k^{'} pr_k^{'}} \geq \frac{\prod_{k=1}^{K} E\left[R_{s,k}^{'}\right]}{\prod_{k=1}^{K} E\left[R_{s,k}\right]} \tag{3.16}$$

Note that under the proposed weighted metric or the conventional PF metric, for a NOMA-based system, the probability of a user being scheduled per time slot remains the same, since for each time slot, users are distributed with uniform and random probability over the entire network. Fig. 3.1 shows the ratio between Pr_k and $Pr_k^{'}$, for different number of users per cell. Results show that this ratio is very close to 1 ($Pr_k \simeq Pr_k^{'}$). Therefore, Eq. 3.16 can also be formulated as Eq. 3.8; Q.E.D.

Weight calculation for each candidate user set U relies on the sum of the multiplexed users' weights:

$$W(U) = \sum_{k \in U} W_k(t) \tag{3.17}$$

with

$$W_k(t) = R_{avg}^{e}(t) - R_k(t),\ k \in U \tag{3.18}$$

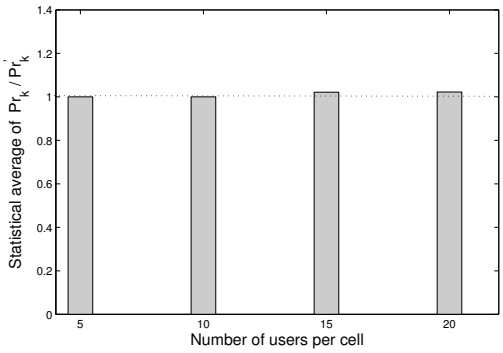


Figure 3.1: The ratio between Pr_k and $Pr_k^{'}$ for a NOMA-based system, in terms of the number of users per cell

$R_{avg}^{e}(t)$ is the expected achievable bound for the average user data rate in the current scheduling slot t. It is calculated as follows:

$$R_{avg}^{e}(t) = c.R_{avg}(t-1) \tag{3.19}$$

Since we tend to enhance the total user achieved rate in every time slot, each user must target a higher rate compared to the rate previously achieved. Therefore, parameter c is chosen to be greater than 1.

The user average data rate, $R_{avg}(t)$, used in Eq. 3.19, is updated at the end of each scheduling slot based on the following:

$$R_{avg}(t) = \frac{1}{K}\sum_{k=1}^{K}\sum_{s=1}^{S} R_{s,k}(t) \tag{3.20}$$

where $R_{s,k}(t)$ is the data rate achieved by user k over subband s.

On the other hand, $R_k(t)$, the actual achieved data rate by user k during the actual scheduling slot t, is calculated such as:

$$R_k(t) = \sum_{s \in S_k} R_{s,k}(t),\ k \in U \tag{3.21}$$

with S_k the set of subbands allocated to user k during time instance t. At the beginning of every scheduling slot, S_k is emptied. Each time user k is being allocated a new subband, S_k and $R_k(t)$ are both updated.

The main idea behind introducing weights is to minimize the rate gap among scheduled users in every scheduling slot, thus maximizing fairness among them. User set U

is provided with the highest priority among candidate user sets if and only if it contains non-orthogonally multiplexed users having good channel quality on subband s with regards to their historical rates or/and having the largest rate distances between their actual achieved rates and their expected average user throughput. The highest level of fairness is achieved, during the actual time slot t, when all users reach the expected user average rate $R^e_{avg}(t)$. By applying the proposed scheduling procedure, we aim to enhance the long-term and the short-term fairness at the same time.

Another configuration of weights can also be introduced, by replacing Eq. 3.5 and 3.17 with the following equation:

$$U_s = \arg\max_{U} \sum_{k \in U} \frac{R_{s,k}(t)}{T_k(t)} W_k(t),\ k \in U \tag{3.22}$$

The conventional NOMA-based PF metric and the weights are jointly calculated for each user k in the set U. By doing so, we assign to each user its weight while ignoring the cross effect $\frac{R_{s,k}(t)}{T_k(t)} W_{k'}(t)$ produced by Eq. 3.5, where k and k' are non-orthogonally multiplexed users in the same candidate user set U. This joint-based incorporation of weights is denoted by J-WNOPF for the following evaluations.

3.2.2 PROPOSED WEIGHTED OMA-BASED PROPORTIONAL FAIRNESS SCHEDULER (WOPF)

In the majority of existing works dealing with fair scheduling, OMA-based system is considered. For this reason, we consider applying the proposed weighted proportional fair scheduling metric introduced in this chapter to an OMA-based system instead of NOMA, in order to evaluate the contribution of NOMA within our framework. In this sense, non-orthogonal cohabitation is not allowed, instead a subband s is allocated to only one user based on the following metric:

$$k^* = \arg\max_{k} \frac{R_{s,k}(t)}{T_k(t)} W_k(t) \tag{3.23}$$

where $W_k(t)$ is the weight assigned to user k, calculated in a similar way to the weights in WNOPF. An ODFMA-based algorithm combined with conventional PF scheduler is denoted by PF^{OMA}, whereas the resulting algorithm from the combination of OMA with weighted PF is denoted by WOPF.

An OMA-based system can be regarded as a special case of a NOMA-based system where only one user is allowed to be scheduled per subband. So in order for the WOPF to achieve a higher user service utility compared to the conventional PF scheduler, PF^{OMA}, Proposition 1 detailed and proven in Section 3.2.1 should also be verified

for an OMA-based system. In this sense, Eq. 3.8 will be modified as follow:

$$\prod_{k=1}^{K} E\left[W_k / \sum_{k} W_k\right] \prod_{k=1}^{K} E\left[R_{s,k}\right] \geq \prod_{k=1}^{K} E\left[R_{s,k}^{'}\right] \tag{3.24}$$

where W_k is the weight assigned to user k.

Note that, we assume that we have equal probability of user k being scheduled per time slot, between the proposed weighted metric, WOPF, and the conventional PF scheduler, PF^{OMA}. For greater certainty, Fig. 3.2 shows the statistical average of the ratio between Pr_k and $Pr_k^{'}$, for different number of users per cell, where the total number of subbands is equal to 16. Results show that this ratio is very close to 1 in practice. Thus, the validity of our assumption is confirmed.

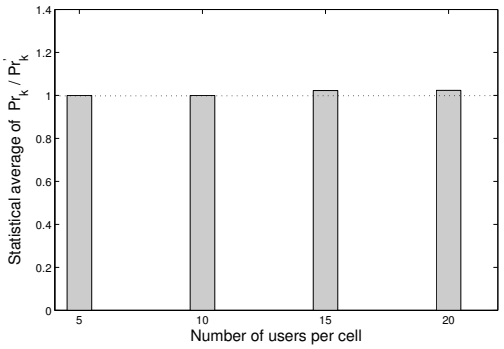


Figure 3.2: The ratio between Pr_k and $Pr_k^{'}$ for an OMA-based system, in terms of the number of users per cell, for $S = 16$

3.2.3 PROPOSED SCHEDULING METRIC FOR THE FIRST SCHEDULING SLOT

Note that, in the first scheduling slot, the historical rates and expected user average data rate are all set to zero. Hence, the selection of users is based only on instantaneous achievable throughputs. Therefore, fairness is not achieved for the first scheduling slot, and the following slots are penalized accordingly. In this sense, we propose to treat the first scheduling slot differently, for all the proposed weighted metrics:
For each subband s, the proposed scheduling process selects U_s among the candidate user sets based on the following criterion

$$U_s = \arg\max_{U} \sum_{k \in U} \frac{R_{s,k}(t=1)}{R_k(t=1)} \tag{3.25}$$

Note that when WOPF is considered, the maximum number of users per set U is limited to 1. $R_k(t=1)$, the actual achieved throughput, is updated each time a subband is allocated to user k during the first scheduling slot. By doing so, we are giving priority to the user experiencing a good channel quality with regard to its actual total achieved data rate, thus enhancing fairness.

3.2.4 SERVICE DIFFERENTIATION

We have already discussed that the desire of some applications to have a QoS guarantee calls for scheduler which provide the ability of delivering different levels of QoS. The PF scheduler, in its classical formulation, fails to meet such requirements. For this sake, we propose in this section some modifications to the proposed weighted metrics in order to satisfy the QoS constraints. In other words, the proposed metrics should have the ability to provide different priorities to different users and to guarantee a certain level of performance to a data flow. Thus, we will detail an example of 3 services, although the proposed modification can be applied for any number of services. To do so, users are classified into 3 groups, namely gold **G**, silver **S**, and bronze **B**. Eq. 3.18 will be modified as follow:

$$W_k(t) = R_{service} - R_k(t),\ k \in U \tag{3.26}$$

where $R_{service}$ is the minimum guaranteed data rate (level of performance) requested by a certain group of users. It is defined as follows:

$$R_{service} = \begin{cases} R_{bronze} & \text{if } k \in \mathbf{B} \\ R_{silver} & \text{if } k \in \mathbf{S} \\ R_{gold} & \text{if } k \in \mathbf{G} \end{cases},\ k \in U \tag{3.27}$$

Note that the proposed weighted metrics have the ability to be easily modified in such a way to meet an application's requirements, by either enhancing the overall achieved fairness (with R^e_{avg}), or providing a requested QoS (with $R_{service}$). Moreover, the aforementioned modification reported in Eq. 3.26 tends also to provide fairness among users that belong to the same group, i.e. asking for the same service.

3.2.5 COMPLEXITY ASSESSMENT

In the aim of assessing the implementation feasibility of the different proposed scheduling techniques, we measured the computational load of the main allocation techniques to be integrated at the BS.

From a complexity point of view, the proposed scheduling metric WNOPF differs from the conventional PF metric in the weight calculation. For a number of users per subband limited to 2 in NOMA, the number of candidates per subband is $C_K^2 + C_K^1$. Hence, our proposed metric WNOPF increases the PF computational load by $\frac{26}{3}KS+S$ ($\simeq O(KS)$) multiplications and $-K^3S+\frac{3}{2}K^2S^2-\frac{4}{6}K^2S-\frac{3}{6}KS$ ($\simeq O(\frac{3}{2}K^2S^2-K^3S)$) additions.
From a complexity point of view, the proposed scheduling metric WNOPF differs from the conventional PF metric in the weight calculation. For a number of users per subband limited to 2 in NOMA, the number of candidates per subband is $C_K^2 + C_K^1$. Hence, our proposed metric WNOPF increases the PF computational load by $\frac{26}{3}KS+S$ ($\simeq O(KS)$) multiplications and $-K^3S+\frac{3}{2}K^2S^2-\frac{4}{6}K^2S-\frac{3}{6}KS$ ($\simeq O(\frac{3}{2}K^2S^2-K^3S)$) additions.

In order to compute the PF metric for a candidate user set containing only 1 user, $4+S$ multiplications and $1+\frac{3}{2}S$ additions are needed. For each candidate user set containing 2 multiplexed users, $13+2S$ multiplications and $6+3S$ additions are required.

By taking account of the calculations of $h^{-2\alpha}$, h^2, and $h^2/(N_0B/S)$ performed at the beginning of the allocation process, the classical NOMA PF requires a total of $3KS+C_K^1S(4+S)+C_K^2S(13+2S)$ multiplications which is equal to $K^2S^2+\frac{1}{2}KS+\frac{13}{2}K^2S$ ($\simeq O(K^2S^2)$) and $C_K^1S(1+3S/2)+C_K^2S(6+3S)$ additions which is equal to $\frac{3}{2}K^2S^2+\frac{1}{2}KS+\frac{13}{2}K^2S$ ($\simeq O(K^2S^2)$). The computational load of the main simulated methods in terms of the number of operations is summarized in Table 3.1.

Table 3.1: Computational load of the simulated methods in terms of the number of additions and multiplications

	Number of multiplications	Number of additions
Classical PF with NOMA	$K^2S^2+1/2KS+13/2K^2S$	$3/2K^2S^2+3K^2S-2KS$
WNOPF	$K^2S^2+29/6KS+13/2K^2S+S$	$-K^3S+3K^2S^2+14/6K^2S-15/6KS$

3.3 PERFORMANCE ANALYSIS OF THE PROPOSED WEIGHTED PF METRICS

In this section, we define the system model adopted to evaluate the performance of the proposed weighted metrics, for both access techniques OMA and NOMA. In particular, the parameters used in our simulations are detailed.

3.3.1 SYSTEM MODEL PARAMETERS

The system model considered in this chapter is different from the one used in the first chapter. The parameters considered are presented in Table 3.2 and are based on existing LTE/LTE-Advanced specifications [93]. Users are deployed randomly in the cell and Extended Typical Urban (ETU) channel model is assumed.

Table 3.2: Simulation Parameters

Parameters	Values
Antenna Pattern	Omnidirectional Antenna
Number of transmitter antennas	1
Number of receiver antennas	1
Carrier Frequency	2 GHz
Sampling Frequency	15.36e$^6 Hz$
System Bandwidth	10 MHz
Cell Radius	500 m
Minimum distance between UE and BS	35 m
Channel Estimation	Ideal
UE speed	50 Km/h
Distance-dependent Path loss	$PL = 128.1 + 37.6log_{10}(d)$, d in Km
BS transmit Power	46 dBm
Number of UEs per cell	K=5, 10, 15, 20
Maximum number of multiplexed UEs	1 (OMA) 2 (NOMA)
Number of subbands	$S = 128$
Noise power spectral density	4.10$^{-18} mW/Hz$
α_{FTPA}	0.7

3.3.2 PERFORMANCE EVALUATION

In this part, we mainly consider four system-level performance indicators: achieved system capacity, long-term fairness, short-term fairness, and cell-edge user throughput. Several techniques, including the proposed modified PF metrics, are evaluated and compared. The following acronyms will be used to refer to the main studied methods:

- PF^{NOMA}: The conventional PF scheduling metric combined with a NOMA-based system.
- $WNOPF$: The proposed weighted PF scheduling metric combined with a NOMA-based system.

- $J-WNOPF$: The proposed Weighted PF scheduling metric with a joint incorporation of weights, combined with a NOMA-based system.
- $PF^{NOMA}_{modified}$: A modified version of the PF scheduling metric proposed in [85], where the actual assignment of each frame is added to the historical rate.
- PF^{OMA}: The conventional PF scheduling metric combined with an OMA-based system.
- $WOPF$: The proposed weighted PF scheduling metric combined with an OMA-based system.

In order to assess the achieved fairness performance of the different algorithms, we used the Gini fairness index [53], defined in Chapter 1 (Section 1.2.1).

In order to evaluate the proposed weighted metrics, we consider equal repartition of power among subbands, in both OMA and NOMA scenarios, as was done in [89, 94]. The bandwidth is divided into 128 subbands. In the case of NOMA, fractional transmit power allocation (FTPA) [21] is used to allocate power among scheduled users within a subband. Without loss of generality, results are shown, in NOMA, for the case where the maximum number of scheduled users per subband is set to 2 ($n(s) = 2$).
As for parameter c in Eq. 3.19, after several testings, the best performance is observed for c equal to 1.5. In fact, the system has a rate saturation bound, since when we further increase c, similar performance is maintained.

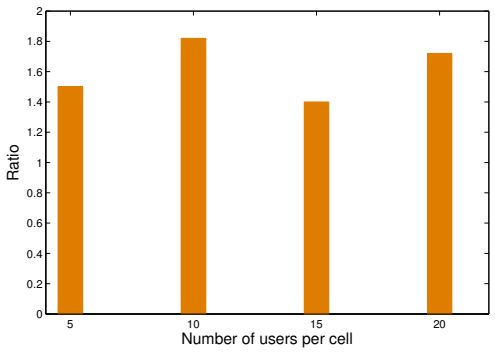


Figure 3.3: Observed ratios in Eq. 3.8, for a NOMA-based system, in terms of the number of users per cell

We start the evaluation process by checking the validity of Proposition 1, detailed in Section 3.2.1 and 3.2.2. Fig. 3.3 shows the ratio between the left hand and the right hand of Eq. 3.8. Results show that this ratio is found to be greater than 1 regardless

of the number of users per cell, which verifies Proposition 1, defined in Eq. 3.8. Similar verification is observed in Fig. 3.4 for an OMA system.

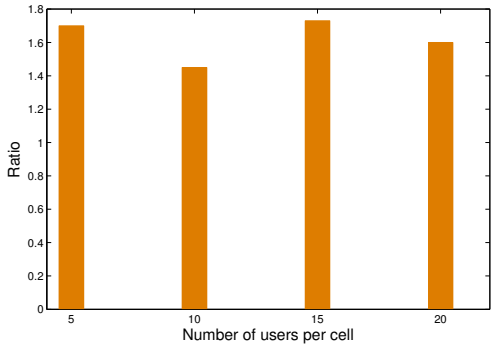


Figure 3.4: Observed ratios in Eq. 3.24, for an OMA-based system, in terms of the number of users per cell

Fig. 3.5 shows the system capacity achieved by each of the simulated methods in terms of the number of users per cell. Curves in solid lines represent the NOMA case, whereas curves with dotted lines refer to OMA.

It is clear that, even though the total number of used subbands is fixed, the throughput of all the simulated methods increases as the number of users per cell is increased. This is due to the fact that the higher the number of users per cell, the better the multi-user diversity is exploited by the scheduling scheme, as also observed in [95].

The gain achieved by WNOPF, when compared to other weighted simulated metric based on NOMA such as J-WNOPF, is mainly due to the fact that J-WNOPF does not take into account the cross effect produced by the non-orthogonally multiplexed users.

The gain in performance obtained by the introduction of weights in the scheduling metric, compared to the conventional PF^{NOMA} metric, stems from the fact that for every channel realization, the weighted metrics try to ensure similar rates to all users, even those experiencing bad channel conditions. With PF^{NOMA}, such users wouldn't be chosen frequently, whereas appropriate weights give them a higher chance to be scheduled more often.

Fig. 3.5 shows also an improved performance compared to the modified PF scheduling metric $PF^{NOMA}_{modified}$ proposed in [85]. Although our proposed weighted metrics and $PF^{NOMA}_{modified}$ consider the current assignment in their metric calculation, they still differ by

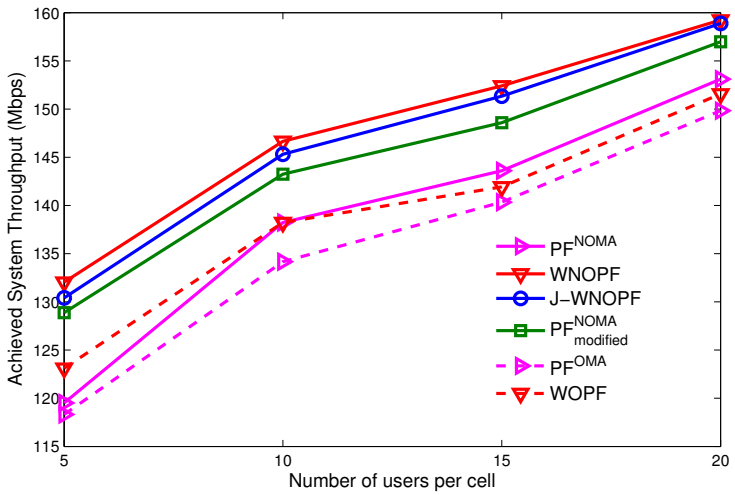


Figure 3.5: Achieved System throughput of the proposed scheduling schemes in terms of the number of users per cell

the fact that the proposed weighted metrics tend in every slot to enhance the achieved user rate, by targeting a higher rate compared to the rate previously achieved.

When the proposed scheduling metrics are applied using OMA as a multiple access technique, it can be clearly noticed that WOPF gives a higher throughput compared to PF^{OMA}, due to the same reason why WNOPF outperforms PF^{NOMA}. Another important result to note in Fig. 3.5 is the performance gain achieved by NOMA over OMA. Weighted scheduling metrics applying NOMA outperform the other simulated metrics based on OMA, including WOPF. This gain is due to the efficient non-orthogonal multiplexing of users. It should also be noted that the gain achieved by WNOPF over PF^{NOMA} is greater than the one achieved by WOPF over PF^{OMA}: combining fair weights with NOMA reveals to give the best performance.

Long-term fairness is an important evaluator for the allocation process performance. Fig. 3.6 shows this metric as a function of the number of users per cell. Long-term fairness is improved when fair weights are introduced, independently of the access technique (OMA or NOMA). This is due to the fact that, when aiming to enhance fairness in every scheduling slot, long-term fairness will be enhanced consequently. It should be noted that WNOPF gives the best performance (Gini index is nearly 0.03, which indicates a high level of fairness). In addition to that, the proposed weighted metrics outperform the modified PF metric [85], $PF^{NOMA}_{modified}$. The reason behind this is that WNOPF and J-WNOPF do not only consider the current rate assignment, but tend also to minimize

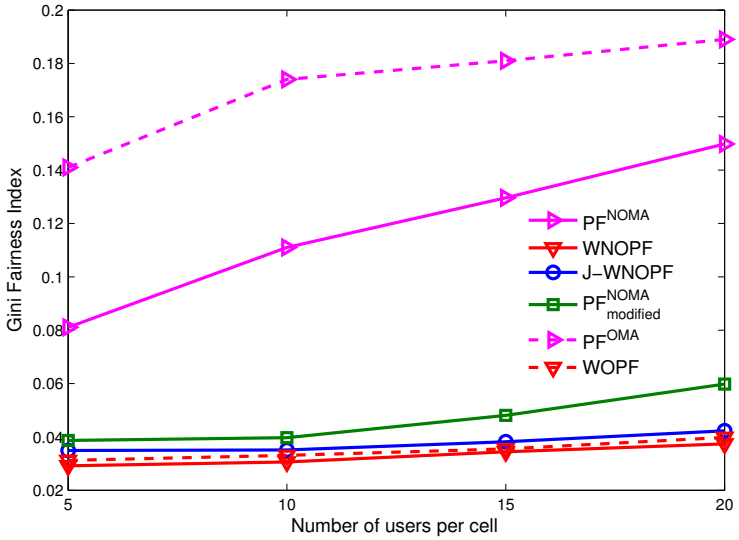


Figure 3.6: Gini Fairness Index in terms of the number of users per cell

the rate gap among scheduled users in every channel realization, thus maximizing fairness among them.

Fig. 3.7 shows the achieved system throughput as a function of S, for 15 users per cell. We can see that the proposed weighted metrics outperform the conventional PF scheduling scheme, for both access techniques OMA and NOMA, even when the number of subbands is limited.

Since one of the goal of this chapter is to achieve fairness in every scheduling slot, the proposed weighted techniques should be compared with the non-weighted PF metrics based on the time they take to reach this goal. Fig. 3.8 shows the Gini fairness index with respect to scheduling time index t. Weighted metrics such as WNOPF, J-WNOPF, and WOPF achieve a high fairness from the beginning of the allocation process, and converge to the highest level of fairness (lowest value of index $G = 0.0013$) in a limited number of allocation steps or time. On the contrary, PF^{NOMA} and PF^{OMA} show unfairness among users for a much longer time. Weighted metrics not only show faster convergence to a high fairness level, but also give a lower Gini indicator at the end of the window length, when compared to non-weighted PF metrics.

In order to assess the quality of experience achieved by the proposed scheduling schemes, we evaluate the time each user takes in order to be served for the first time,

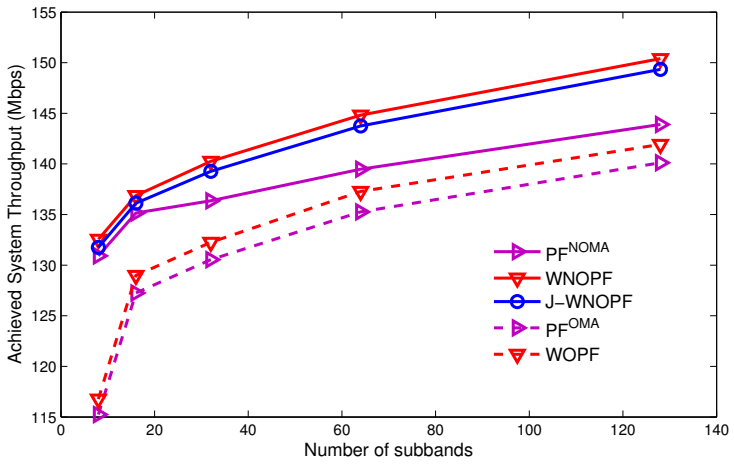


Figure 3.7: Achieved system throughput in terms of S, for $K = 15$.

denoted by the rate latency, and the variation of user achieved rate versus the scheduling index t. For this purpose, Fig. 3.9 shows the achieved rate of the user experiencing the largest rate latency versus time, for different scheduling schemes.

We can observe that, when the conventional PF^{NOMA} is used, no rate is provided for such users, for the first two scheduling slots. In addition, severe rate fluctuations are observed through time. However, when weighted metrics are considered, a rate is assigned for the least privileged users from the first scheduling slot, and remains stable for all the following slots. This behavior results from the fact that at the beginning of the scheduling process (first scheduling slot), historical rates are set to zero, and PF^{NOMA} uses only instantaneous achievable throughputs to choose the best candidate user set. Therefore, users experiencing bad channel conditions will have a low chance to be chosen. Their achieved data rates will then be equal to zero. On the other side, using weighted metrics, the treatment of the first scheduling slot is conducted differently and users are chosen depending on their actual rates (measured during the actual scheduling period). In this case, zero-rates are eliminated. Hence, latency is greatly reduced.

For the next scheduling slots, historical rates will be taken into account. For PF^{NOMA}, users experiencing a large $T_k(t)$ will have less chance to be chosen, and can not be chosen at all. In this case, the use of buffering becomes mandatory. It should be noted that the size of the buffer should be chosen adequately to prevent overflow when peak rates occur, as a result of a high achieved throughput (high $R_{s,k}(t)$). Based on calculation, the average size of the buffer should be around 110 Mbit, for the parameters defined in Table 3.2. However, in the case of the weighted proposed metrics,

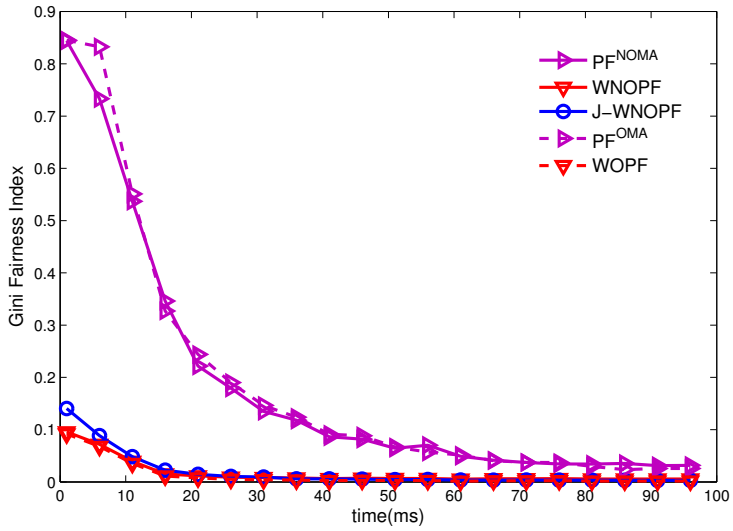


Figure 3.8: Gini fairness index with respect to scheduling time index t

buffering is not needed, since steadier users data rates are almost ensured, and a better QoE is observed. Similar performance is observed for the orthogonal case for the same aforementioned results.

In order to further exploit the achieved fairness, we evaluate the cell-edge user throughput. The motivation for our analysis can be found in the results reported in Fig. 3.10. The proposed weighted metrics outperform the conventional PF scheduling scheme for the case of OMA and NOMA. In addition to that, WNOPF shows the best performance. So, we can say that the incorporation of fair weights with a NOMA-based system stands to be the best combination.

The computational load generated by the simulated methods, in terms of the number of multiplications (NoM) and additions (NoA), is evaluated for 15 users per cell and for multiple numbers of subbands. Results are reported in Table 3.3. It is very clear that the proposed weighted technique incite a slight increase in the complexity compared to the conventional PF scheduler in terms of multiplications and additions.

In order to evaluate the performance of the proposed weighted metrics when premium services are considered, Table 3.6 and 3.7 show the long-term fairness achieved among users that belong to the same group. Results are reported for two different scenarios, where three levels of services are requested: bronze, silver, and gold. Note that, we consider 15 users per cell and 5 users per service.

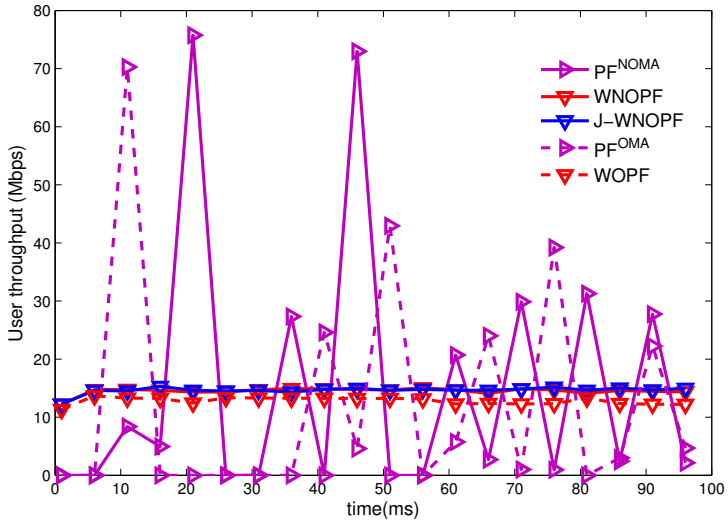


Figure 3.9: User Throughput versus time

Table 3.3: Computational load of the simulated methods, in terms of the number of multiplications and additions

S	8		16		32		64		128	
	NoM	NoA	NoM	NoA	NoM	NoA	NoM	NoA	NoM	NoA
Classical PF with NOMA	**26K**	**26K**	**81K**	**96K**	**277K**	**366K**	**1M**	**1.4M**	**3.8M**	**5.6M**
WNOPF	**26K**	**20K**	**82K**	**126K**	**279K**	**598K**	**1M**	**2.5M**	**3.8M**	**10.6M**

Scenario 1:
The corresponding data rates of the three levels are set at 5 Mbps, 10 Mbps, and 15 Mbps respectively.
Scenario 2:
The corresponding data rates of the three levels are set at 10 Mbps, 20 Mbps, and 30 Mbps respectively.

In scenario 1, all users succeed to reach their requested service data rates, with 153 Mbps as a cell capacity and results of Table 3.4 show a high level of fairness achieved among users requesting the same service. However, when scenario 2 is applied, no success could be obtained. The cell capacity is around 155 Mbps and fairness is still maintained among users.

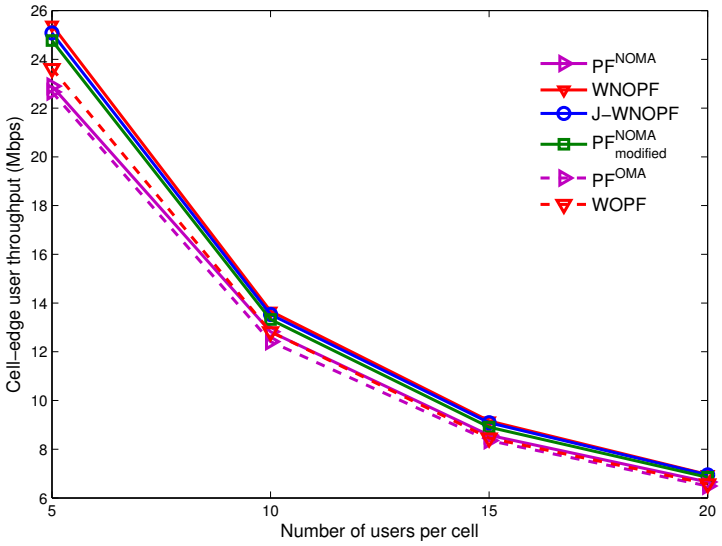


Figure 3.10: Cell-edge user throughput in terms of the number of users per cell

Table 3.4: Gini Fairness index and data rate achieved per group for Scenario 1 (100% success)

Service	Gini Fairness Index	Achieved data rate per group (Mbps)
Bronze	0.0491	25.7
Silver	0.0724	51
Gold	0.0042	76.3

3.4 PROPOSALS TO IMPROVE THE PF SCHEDULER AT THE LEVEL OF POWER ALLOCATION

In the PF scheduler, the allocation of each subband requires the estimation of a scheduling metric for each possible user candidate (in OMA) or candidate set (in NOMA). These estimations necessitate rate calculations which, in turn, require the power levels to be predicted on the considered subband, for each candidate. This becomes even more problematic as the number of subbands and/or users increases. For these reasons, in all previous works dealing with PF scheduling equal power distribution was assumed between subbands [94, 96] to circumvent the power estimation problem, thus preventing the system from the possibility of inter-subband power optimization. In this sense, we propose modifications to the PF scheduler at the level of power allocation in order to improve further its performance.

Indeed, waterfilling algorithm applied in the chapter sec:2, in its classical formulation,

Table 3.5: Gini Fairness index and data rate achieved per group for Scenario 2 (No success)

Service	Gini Fairness Index	Achieved data rate per group
Bronze	0.0522	30.2
Silver	0.0613	49.6
Gold	0.0049	75.2

cannot be directly used within the PF, since this would necessitate performing a separate waterfilling procedure, for each attributed subband and each candidate set, leading to a prohibitive complexity.
On the contrary, in [97], candidates are first sorted based on a certain priority scheme; then, at each step, the user with the highest priority is attributed its most favorable subband. This allowed us avoiding the high number of testings necessary to determine the best candidate for each subband, on the one hand, and allowed the incorporation of more elaborate power allocation schemes, on the other. Power allocation was performed either by means of an optimal scheme, achieving joint inter and intra subband power allocation, or by means of a sub-optimal solution where inter-subband power allocation was first performed using waterfilling, followed by intra-subband power allocation.
In this chapter we aim at introducing a low-complexity iterative waterfilling technique that will allow the incorporation of the waterfilling sub-optimal solution proposed in [98] within the PF, and therefore enhance its performance.

3.4.1 PROPOSED POWER ALLOCATION SCHEME

The allocation of each subband requires the power level to be predicted on the considered subband, for each candidate. Therefore, the power allocation proposed in this section is jointly performed with subband allocation. We first perform the power allocation for each candidate user set U, and then compute the PF metric for the considered candidate. The candidate user set maximizing Eq. 3.2 is selected to be scheduled on the considered subband.
As it was shown in [99], the waterfilling process was the solution obtained to allocate a fixed power P_{max} among the various subchannels of a multichannel transmission system so as to optimize the bit rate of the entired system. For an OMA-based system of S subbands, $S+1$ equation with $S+1$ unknowns are needed to be solved in order to optimally distribute the power among subbands. Specifically, the amount of power that should be allocated on each subband and the waterline level, denoted by W should be

found. The corresponding system can be expressed as follows:

$$\begin{cases} P_1 + \frac{N_0 B/S}{h_{1,k}^2} = W \\ P_2 + \frac{N_0 B/S}{h_{2,k}^2} = W \\ \vdots \\ P_s + \frac{N_0 B/S}{h_{s,k}^2} = W \\ \vdots \\ P_S + \frac{N_0 B/S}{h_{S,k}^2} = W \\ P_1 + P_2 + \cdots + P_S = P_{\max} \end{cases} \tag{3.28}$$

where P_s denotes the power that should be allocated on subband s, and $h_{s,k}^2$ the channel gain of user k on subband s.

Chapter sec:2 details the optimal power allocation when NOMA is applied and describes a sub-optimal solution as well. However, as noted in [97], the optimal solution incurs an increase in the computational load by a factor of more than 20, compared to the suboptimal solution, while yielding very close simulation results (in terms of spectral efficiency and outage probability). Therefore, we aim at introducing a low-complexity iterative waterfilling technique that will allow the incorporation of the waterfilling sub-optimal solution within the PF scheduler. The proposed technique consists in predicting the waterline level recursively from the previous level and from the channel gain of the considered strongest user scheduled on the current subband.

Indeed, maximizing the achieved throughput through an optimal sharing of the total transmit power among subbands can be achieved if [99]:

$$P_s + \frac{N_0 B/S}{h_{s,k^*}^2} = W(S(i)),\ s \in S(i) \tag{3.29}$$

where $S(i)$ is the set of allocated subbands at allocation stage i, $W(S(i))$ the corresponding waterline level at stage i, and h_{s,k^*}^2 the channel gain of user k^* showing the highest channel gain among scheduled users on subband s.
During the allocation process, the total transmit power P_{max} is distributed, at each stage, among allocated subbands based on Eq. 3.28, resulting in:

$$P_{\max} = \sum_{s \in S(i)} \left(W(S(i)) - \frac{N_0 B/S}{h_{s,k^*}^2} \right) \tag{3.30}$$

Since the same amount of total power is redistributed each time the scheduler allocates a new subband denoted by s_{new}, the waterline is updated by $W(S(i+1))$ only if

$\frac{N_0B/S}{h^2_{s_{new},k}} < W(S(i))$, otherwise it keeps its previous value $W(S(i))$. In case the waterline is updated, P_{max} is distributed at stage $i+1$ as follows:

$$\begin{aligned} P_{\max} = & \sum_{s\in S(i)} \left(W\left(S(i+1)\right) - \frac{N_0B/S}{h^2_{s,k^*}} \right) \\ & + \left(W\left(S(i+1)\right) - \frac{N_0B/S}{h^2_{s_{new},k^*}} \right) \end{aligned} \tag{3.31}$$

If we denote by $N(i)$ the number of subbands in the set $S(i)$, Eq. 3.30 can be re-written as:

$$P_{\max} = N(i).W(S(i)) - \sum_{s\in S(i)} \frac{(N_0B/S)}{(h^2_{s,k^*})} \tag{3.32}$$

And Eq. 3.31 as:

$$P_{\max} = N(i)W(S(i+1)) - \sum_{s\in S(i)} \frac{N_0B/S}{h^2_{s,k^*}} + W(S(i+1)) - \frac{N_0B/S}{h^2_{s_{new},k^*}} \tag{3.33}$$

Hence, by comparing Eq. 3.32 and Eq. 3.33, we obtain:

$$\begin{aligned} N(i)W(S(i)) - \sum_{s\in S(i)} \frac{N_0B/S}{h^2_{s,k^*}} = & N(i)W(S(i+1)) \\ - \sum_{s\in S(i)} \frac{N_0B/S}{h^2_{s,k^*}} + & \left(W(S(i+1)) - \frac{N_0B/S}{h^2_{s_{new},k^*}} \right) \end{aligned} \tag{3.34}$$

Therefore, the waterline at stage $i+1$ can be formulated as:

$$W(S(i+1)) = \frac{1}{N(i)+1} \left(N(i).W(S(i)) + \frac{N_0B/S}{h^2_{s_{new},k^*}} \right) \tag{3.35}$$

Fig. 3.11 shows the main steps of the proposed resource allocation technique that incorporates the introduced waterfilling power allocation within the PF.

At each step of the scheduling process, for every candidate user set U, the corresponding waterline level is derived from Eq. 3.35, while taking into account user k^* showing the highest channel gain among scheduled users in the set U, over subband s_{new}. Once the waterline level at the current stage, $i+1$, is determined for each candidate user set U, power is then assigned for each U as $P_{s_{new}|U}$ using:

$$P_{s_{new}|U} = W(S(i+1)) - \frac{N_0B/S}{h^2_{s_{new},k^*|U}} \tag{3.36}$$

Afterwards, $P_{s_{new}|U}$ is divided among scheduled users in the set U based on the chosen intra-subband power allocation technique, e.g. FTPA, the scheduling PF metric is calculated for each U and the best candidate user set U_s is selected based on Eq.3.2. Note that, at each allocation step, the power estimation using Eq. 3.36 is performed only for

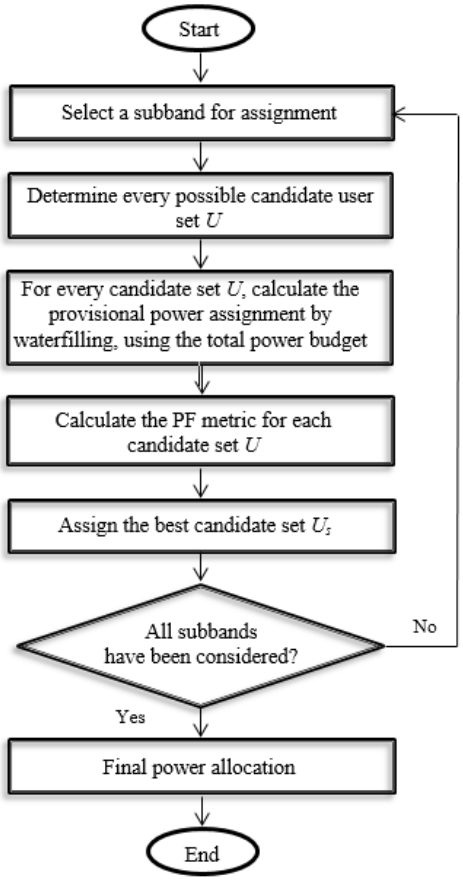
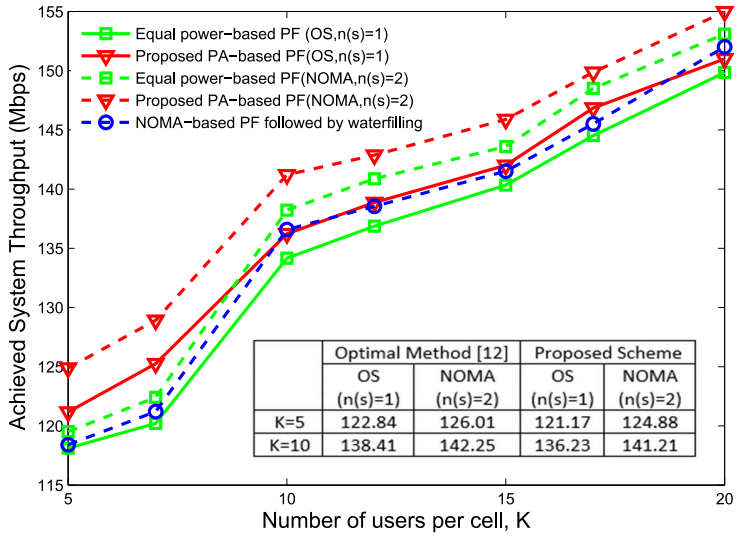


Figure 3.11: Flow chart of the considered waterfilling-based allocation scheme.

subband s_{new} in order to choose the best candidate user set, i.e., there is no need to update, at each iteration, all users powers on previously allocated subbands. Subbands are sequentially tested (as in the PF), until they all have been considered for allocation. At this point, the final users' power levels on all subbands are updated using the final waterline level.

3.4.2 PERFORMANCE ANALYSIS OF THE ITERATIVE WATERFILLING-BASED PF SCHEDULER

The system model used in this section is the same as the system model used in section 3.3.2.
In order to evaluate our proposed power allocation scheme for NOMA, we compared it with an OMA-based system (which is known as orthogonal signaling based system) and with a NOMA system using equal repartition of power among subbands followed by an intra-subband power allocation based on FTPA. OMA system can be regarded as a special case of NOMA where $n(s) = 1$, i.e., only one user can be scheduled over a subband. Note that in the NOMA case, some subbands can also be assigned to single users, leading to a hybrid scheme.



	Optimal Method [12]		Proposed Scheme	
	OS (n(s)=1)	NOMA (n(s)=2)	OS (n(s)=1)	NOMA (n(s)=2)
K=5	122.84	126.01	121.17	124.88
K=10	138.41	142.25	136.23	141.21

Figure 3.12: Achieved system throughput in terms of the number of users per cell, for 128 subbands.

Fig. 3.12 shows the achieved system throughput in terms of number of users per cell, with a number of subbands equal to 128. Dotted lines represent the OMA case,

whereas solid lines refer to NOMA.
The proposed power allocation scheme is also compared with an alternative method where equal inter-subband power allocation is considered within the PF scheduling process (to assign all subbands) and waterfilling is only applied once at the end to determine the final power levels. The alternative method, NOMA-based PF followed by waterfilling, is represented by a blue dotted line.

The throughput increases with the number of users per cell, for all the simulated methods. In fact, when the number of users per cell increases, the scheduling schemes exploit the multi-user diversity more efficiently. This was also observed in [95]. Simulation results show that the proposed NOMA system always outperforms the OMA-based system.

When compared to an equal inter-subband power allocation algorithm, our proposed power allocation scheme shows improved performance regardless of the number of users per cell. The gain in performance is observed in the NOMA case as well as in the OMA case. For the NOMA case, the gain in throughput can reach 5 Mbps for 5 users per cell, i.e. 1 Mbps per user.

A comparison of our proposed scheme with the optimal solution described in [97], and incorporated within the PF, is also presented in the table included in Fig. 3.12. The gap between the total throughput achieved by the two methods is shown to be only 1%.

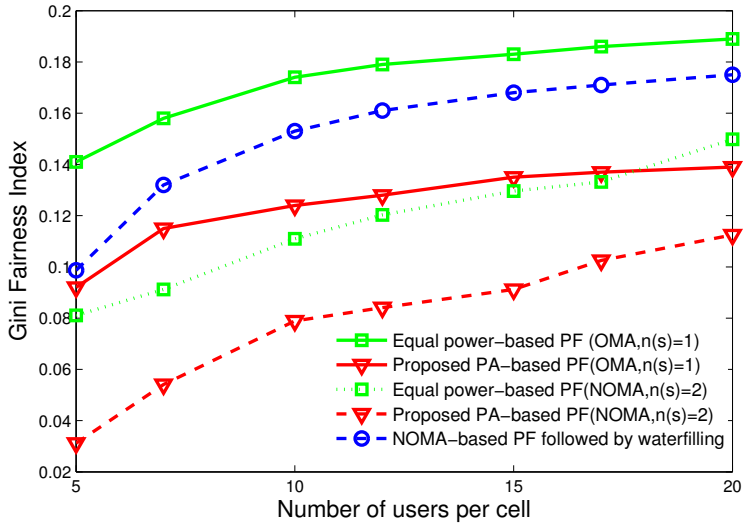


Figure 3.13: Gini fairness index in terms of the number of users per cell, for 128 subbands.

Fig. 3.13 shows the Gini metric as a function of the number of users per cell, K. In addition to system capacity, fairness is also significantly improved when power is dynamically distributed among subbands, independently of the access technique (OMA or NOMA).

However, the fairness level of NOMA is better than that of OMA case. In fact, in NOMA, users located at the edge of the cell (hence having a low channel gain) are given the possibility of being paired (as second users) with other users on certain subbands, and are in this case attributed a power level higher than that of the users close to the center of the cell (hence having a high channel gain). On the contrary, when PF scheduling is used with OMA, only one user is scheduled on each subband, therefore depriving cell-edge users from having access (as second users) to a significant number of subbands that can significantly increase their achieved data rate. From this perspective, we can see that NOMA is fairer to users than OMA, since it compensates for the distance effects on the user channel quality by offering appropriate power levels.

Note that the alternative method NOMA-based PF followed by waterfilling shows degraded performance with respect to the incorporated waterfilling process (see Fig. 3.12 and Fig. 3.13). This is due to the fact that users experiencing bad channel conditions but having low historical rates can be considered by the Equal power-based PF as having high priority on certain subbands. When applying waterfilling at the end of the allocation, such users will be allocated a low level of power (depending on their channel gains), leading to a low spectral efficiency. The incorporation of waterfilling within the PF allows avoiding such cases.

For 10 users per cell, Fig. 3.14 shows the achieved system throughput where the number of subbands varies between 8 and 128. We can see that the proposed joint power allocation and scheduling scheme still outperforms conventional NOMA PF even when the number of subbands is limited. As for the long-term fairness presented in Fig. 3.15, the gain of the proposed power allocation technique compared to the equal power repartition is almost constant for NOMA, regardless of the number of subbands.

3.4.3 EVALUATION OF THE COMPUTATIONAL COMPLEXITY

From a complexity point of view, the proposed scheduling scheme differs from the classical PF by the waterline calculation and the power estimation step for each candidate user set. Assuming that the calculation of N_0/h^2 is carried out once at the beginning of the algorithm using $2SK$ real multiplications, for each candidate user set and each new subband, estimating the new waterline using Eq. 3.30 requires 2 multiplications and 2 additions, and power allocation using Eq. 3.31 takes only 1 addition. For a number of users per subband limited to 2 in NOMA, the number of candidates per subband

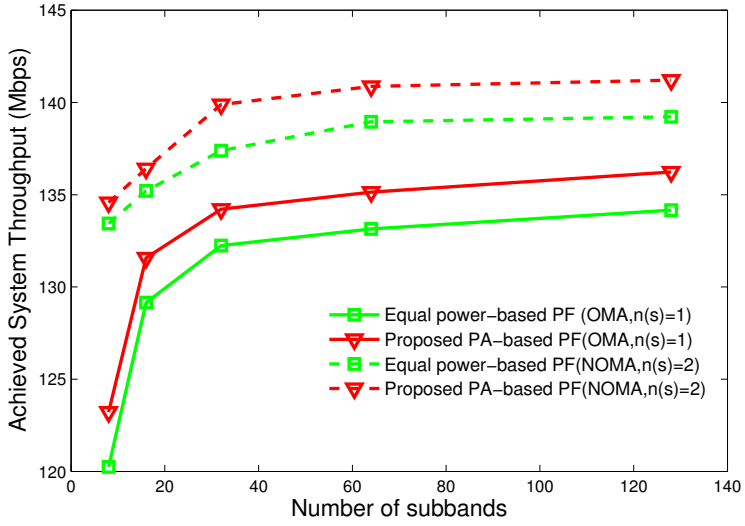


Figure 3.14: Achieved system throughput in terms of the number of subbands, for 10 users per cell.

Table 3.6: The additional computation burden in terms of multiplications and additions of the iterative waterfilling-based PF scheduler compared to the conventional NOMA PF, for 10 users per cell

S	Multiplications	Additions
8	0.9%	0.8%
128	8.93%	11.14%

is $C_K^2 + C_K^1$. Hence, our proposed technique increases the PF computational load by $2SK + 2S(C_K^2 + C_K^1)$ multiplications and $3S(C_K^2 + C_K^1)$ additions.

As for classical NOMA PF, the calculation of the PF metric given by Eq. 3.2 depends on the number of multiplexed users in the candidate user set. The classical NOMA PF requires a total of $3KS + C_K^1 S(4+S) + C_K^2 S(13+2S)$ multiplications and $C_K^1 S(1+3S/2) + C_K^2 S(6+3S)$ additions.

The additional complexity burden of the iterative waterfilling-based PF scheduler in terms of multiplications and additions is reported in Table 3.6. Results are considered for 10 users per cell and two different numbers of subbands, for one allocation step over LTE channel simulation.
It is clear that our scheduling method incurs a minor increase in complexity compared

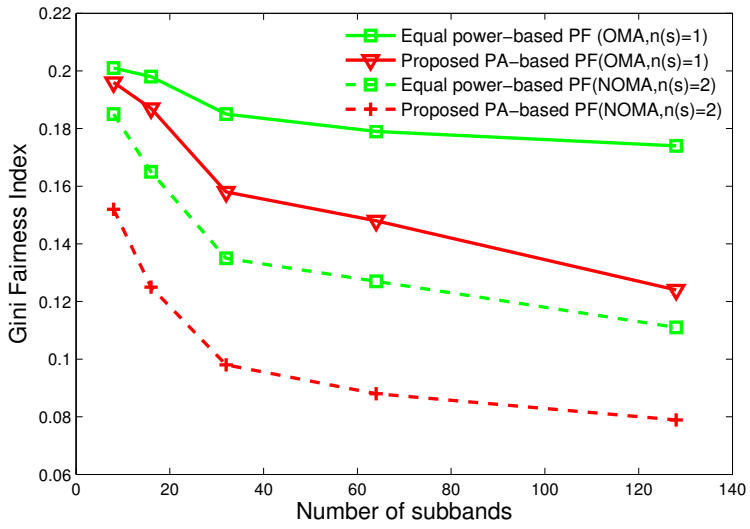


Figure 3.15: Gini fairness index in terms of the number of subbands, for 10 users per cell.

to the classical NOMA PF.

When it comes to the optimal solution [97], since it includes a numerical solver, it is not possible to compare its complexity towards that of the suboptimal method in terms of the number of additions and multiplications. Instead, a measure of the average execution time of one complete allocation cycle (i.e. in one timeslot), for the case of 10 users per cell and 8 subbands, yielded 3 min for the optimal solution and 68 ms for the proposed suboptimal scheme.

3.5 Conclusion

In this chapter, we have investigated the performance of the proportional fairness (PF) scheduler. Several drawbacks were highlighted, especially in a NOMA context, and modifications were proposed, at the levels of subband allocation and power allocation.

In this sense, we have developed several weighted scheduling schemes for both NOMA and OMA multiplexing techniques.They target maximizing fairness among users, while improving the achieved capacity. Several fair weights designs have been investigated. Simulation results show that the proposed schemes allow a significant increase in the total user throughput and the long-term fairness, when compared to OMA and conven-

tional NOMA-based PF scheduler. Combining NOMA with fair weights shows the best performance. Furthermore, the proposed weighted techniques achieve a high level of fairness within each scheduling slot, which improves the QoE of each user.

In addition to that, we have proposed modifications at the power level where we have incorporated a low-complexity iterative waterfilling process within the PF scheme. Simulation results show that the proposed scheme allows an increase in the total user throughput and in the system fairness, when compared to an OMA-based system and to a NOMA system considering an equal power repartition among subbands.

Note that the weighted metrics proposed in this chapter at the level of user scheduling were described in a journal paper accepted in in Wireless Communications and Mobile Computing, Special Issue on Efficient Spectrum Usage for Wireless Communications (October 2018), and have been the subject of the following European patent filing:

M.-R. Hojeij, J. Farah, C. Abdel Nour and C. Douillard, "Method and apparatus for user distribution to sub bands in NOMA systems", Patent application EP 16306722.6-1857, Dec. 19, 2016.

As for the proposed iterative waterfilling process, proposed within the PF scheduler scheme, at the power allocation level, has been published in the IEEE Wireless communication Letters, and has been the subject of the following European patent filing:

M.-R. Hojeij, J. Farah, C. Abdel Nour and C. Douillard, "Method and apparatus for power and user distribution to sub-bands in NOMA systems", Patent application EP 16305929.8-1855, July 19, 2016.

Chapter 4

New throughput and/or fairness maximization metrics

The total cell throughput, the cell-edge user throughput and the fairness are highly dependent on the way both the power and the bandwidth are allocated to users by the scheduler, as well as on the user pairing through NOMA multiplexing [104]. It is widely agreed that the cell throughput is maximized when the channel gain difference between multiplexed users is high [105, 106], although this statement is questioned in a few number of papers. In [107], the performance of two intra-subband power allocation schemes was evaluated: the first one determines the power levels in a static manner based on the average Signal to Interference and Noise Ratio (SINR), while the second one dynamically attributes power to users according to the instantaneous SINR. For the fixed scheme, it is found that the gain of downlink NOMA over OMA increases when the channel gain difference between multiplexed users increases. As for the dynamic scheme, this difference is limited by some constraints. In [108], the authors underlined that the incorrect choice of the power coefficients of paired users or the users target data rates leads to a perceived penalty for allocated users in terms of Quality of Service.

In this chapter, motivated by these contradicting views, we first aim to verify if the common belief that the increase in the channel gain difference is always in favor of NOMA's achieved throughput, when different intra-subband power allocation techniques are considered. Then, guided by the obtained results, we design several resource allocation techniques that aim to be fully adapted to NOMA.

Moreover, it is well known that the proportional fairness scheduler provides a good tradeoff between system capacity and user fairness. However, some applications may need to favor one at the expense of the other. Therefore, the resource allocation techniques introduced in this chapter are designed to introduce flexibility in throughput/fairness levels of a NOMA-based system.

Chapter 4 is organized as follows: In Section 4.1 and 4.2, the dependency of throughput on the channel gain difference and on the channel gain values is observed and analyzed. Flexible throughput/fairness metrics are proposed and detailed in Section 4.3. The complexity analysis of the proposed metrics is carried out in Section 4.4. The performance analysis of the proposed metrics is conducted in Section 4.5. Finally, Section 4.6 concludes the chapter.

4.1 DEPENDENCY OF THROUGHPUT ON CHANNEL GAIN DIFFERENCE IN NOMA

In this section, we analyze the impact of the channel gain difference between multiplexed users on the achieved throughput for different intra-subband power allocation techniques. To this end, we focus on two intra-subband power allocation schemes: Fixed Power Allocation (FPA) and Fractional Transmit Power Allocation (FTPA). For the sake of targeting an implementable scenario, we do not consider the Full Search Power Allocation (FSPA) scheme in our study since its complexity is significantly higher than that of FPA and FTPA. Besides, its resulting throughput is only slightly higher than the ones achieved by FPA and FTPA [109].

The FPA technique is the least complex of all methods. The power is divided between the users on a certain subband according to a fixed ratio β [109]. In the case of two non-orthogonally multiplexed users on a subband s, the total power attributed to this subband, P_s , is divided according to $(\beta P_s, (1-\beta P_s))$, β $(0 \leq \beta \leq 0.5)$ being a constant parameter over all subbands. The user with the highest channel gain is given βP_s and the paired user is given the rest. FPA allocates the power to users regardless of their channel gains.

The FTPA technique is yet another way of dynamically dividing the power between multiplexed users on subband s , based on their channel gains. In the case of two multiplexed users, P_{s,k_1} and P_{s,k_2} are now attributed as follows [109]:

$$P_{s,k_1} = \frac{\left(h_{s,k_1}^2/N_0\right)^{-\alpha}}{\left(h_{s,k_1}^2/N_0\right)^{-\alpha} + \left(h_{s,k_2}^2/N_0\right)^{-\alpha}} P_s \tag{4.1}$$

$$P_{s,k_2} = \frac{\left(h_{s,k_2}^2/N_0\right)^{-\alpha}}{\left(h_{s,k_1}^2/N_0\right)^{-\alpha} + \left(h_{s,k_2}^2/N_0\right)^{-\alpha}} P_s \tag{4.2}$$

α is a factor ranging between 0 and 1 that controls the amount of power attributed to every user; its value is kept constant throughout the power allocation process.

Understanding how FPA and FTPA work helps us find the best couples of users to be multiplexed by NOMA, and verify whether the increase in the channel gain difference is always in favor of NOMA in terms of achieved throughput. In order to carry out this analysis, we consider the case where the channel gain difference between multiplexed users, $d = h^2_{s,k_1} - h^2_{s,k_2}$, is increasing. We assume that, since none of the multiplexed users has been previously chosen, the increase in d results both from an increase in h^2_{s,k_1} and a decrease in h^2_{s,k_2}.
When FPA is used, for a fixed value of P_s, P_{s,k_1} and P_{s,k_2} are constant regardless of the choice of users. In the rate calculation of user k_1 (the user having a higher channel gain compared to user k_2)(Eq.1.4), an increase in h^2_{s,k_1} results in an increase in R_{s,k_1} as well. In the rate calculation of user k_2 (Eq.1.5), the decrease of h^2_{s,k_2} leads to the decrease of both the numerator and the denominator inside the logarithm. However, since $P_{s,k_2} > P_{s,k_1}$, the whole term inside the log function decreases, leading to the decrease of R_{s,k_2}. Nevertheless, since R_{s,k_1} is higher than R_{s,k_2}, the increase in R_{s,k_1} is higher than the decrease in R_{s,k_2}, leading to an increase in the overall achieved rate on subband s, R_s. To sum up, when FPA is used, R_s is an increasing function of d. When FTPA is used, if h^2_{s,k_1} increases and h^2_{s,k_2} decreases, P_{s,k_1} decreases according to Eq. 4.1 since $h^2_{s,k_1} > h^2_{s,k_2}$. Thus, the monotony of the product $h^2_{s,k_1} P_{s,k_1}$ in the rate calculation of user k_1 cannot be predicted: it mainly depends on the value of α. Therefore, the monotony of R_{s,k_1} with d cannot be predicted either. The same applies for R_{s,k_2} given by Eq. 1.5. In short, in the case of FTPA, the relationship between the total achieved throughput on a subband s and the channel gain difference between paired users is unpredictable and could evolve either in the same direction or in the opposite one.

In the will of understanding the relationship between the achieved data rate and the multiplexed users channel gain difference, Fig. 4.1, Fig. 4.2 and Fig. 4.3 represent respectively the behavior of R_{s,k_1} , R_{s,k_2} and R_s with respect to the variation of d for FPA and FTPA. These curves were obtained by simulation, using the following scenario: In order to have a sufficient number of channel gains coefficients, and only in sections 4.1 and 4.2, 50 users are uniformly distributed in the cell and divided into two groups: 25 users with the lowest channel gains are grouped together as UE-2 group, and the remaining 25 users as UE-1 group. Then, the user having the highest available channel gain in UE-1 group is matched with the user having the lowest available channel gain in UE-2 group until we obtain 25 couples of users, each with a unique channel gain difference.

Fig. 4.1 confirms that R_{s,k_1} always increases with d when FPA is used, regardless of the choice of β. Quite the contrary, when FTPA is used, the choice of α is critical for R_{s,k_1} since different directions of variation are shown by the simulations for different values of α.

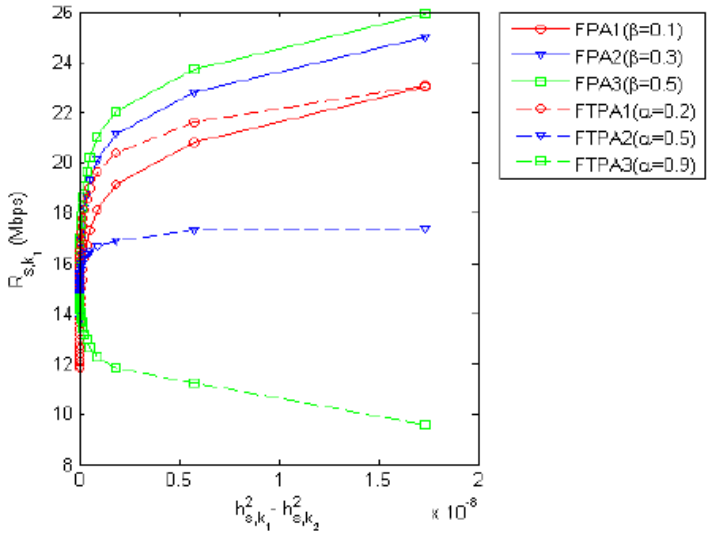


Figure 4.1: R_{s,k_1} achieved for different FPA and FTPA patterns, as a function of the channel gain difference d.

The behavior of R_{s,k_2} shown in Fig. 4.2 complies with the preceding analysis. When FPA is used, R_{s,k_2} is a decreasing function of d. However, when FTPA is deployed, the monotony of R_{s,k_2} depends on the value of d: a fast increase in R_{s,k_2} is obtained at low channel gain differences, but for higher values of d, the increase in R_{s,k_2} slows down and a decrease is even observed for high values of α.

Fig. 4.3 shows the variations of the overall rate R_s. We can observe that FPA can achieve a higher overall throughput than FTPA and that, when FTPA is used, the increase in the channel gain difference between scheduled users, due to an increase in the channel gain of user k_1 and a decrease of channel gain of user k_2, is not always in favor of NOMA in terms of achieved throughput, contrary to what is commonly assumed.

4.2 DEPENDENCY OF THROUGHPUT ON CHANNEL GAIN VALUES FOR A NOMA SYSTEM

In the previous section, couples were chosen such that the channel gain of user k_1 and user k_2, h^2_{s,k_1} and h^2_{s,k_1} respectively, were changing every time, leading to a channel gain difference variation. However, it is also useful to verify whether or not R_{s,k_1}, R_{s,k_2}, and R_s depend on the channel gain difference between user k_1 and user k_2, when only one of the two users' channel gains varies. For this sake, in this section, we will consider two scenarios:

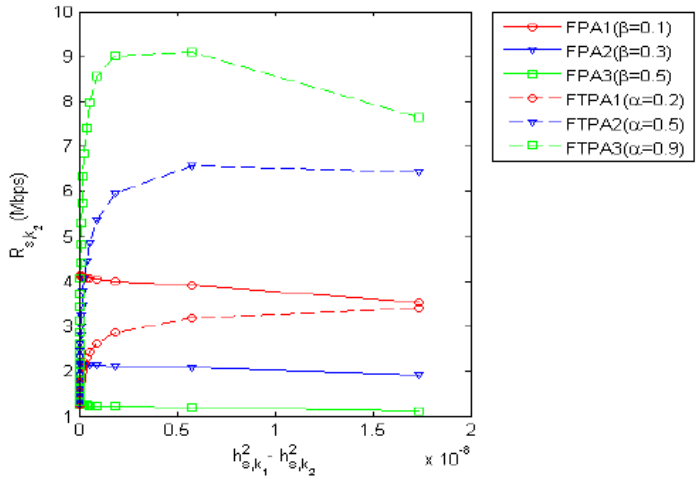


Figure 4.2: R_{s,k_2} achieved for different FPA and FTPA patterns, as a function of the channel gain difference d.

Scenario 1: every user in the user set is considered as k_1, and is matched with all the users whose channel gain is lower than its own channel gain. By doing so, we study the impact of varying the channel gain of user k_2, assuming user k_1 is fixed.

Scenario 2: every user in the user set is considered as k_2, and is matched with all the users whose channel gain is higher than its own channel gain. By doing so, we study the impact of varying the channel gain of user k_1, assuming user k_2 is fixed.

Considering scenario 1, if FPA is adopted, decreasing the channel gain of user k_2 yields a constant value of R_{s,k_1} since P_{s,k_1} and h^2_{s,k_1} are fixed. However, R_{s,k_2} decreases, for the reasons detailed in Section 4.1. Consequently, R_s decreases when h^2_{s,k_2} decreases. If FTPA is considered, when decreasing the channel gain of user k_2, the power allocated to user k_1, P_{s,k_1}, decreases too, for a fixed h^2_{s,k_1}. Thus, the resulting throughput R_{s,k_1} is found to be decreasing when h^2_{s,k_2} decreases (i.e. when d increases). Unfortunately, applying a similar reasoning for R_{s,k_2} is a hard nut to crack. Since P_{s,k_2} increases and h^2_{s,k_2} decreases, the monotony of $P_{s,k_2}h^2_{s,k_2}/\left(P_{s,k_1}h^2_{s,k_2}+N_0B/S\right)$ cannot be predicted. Referring to all the above statements in Section 4.1 and 4.2, we can now conclude that R_{s,k_2} has one or more critical points on its curve. These points can be found by calculating the derivative of R_{s,k_2} with respect to the channel gain difference d, expressed as a function of the power allocated to subband s, P_s, as shown in the following calculation: First, using Eq. 4.4 and 4.2, R_{s,k_2} is re-written as:

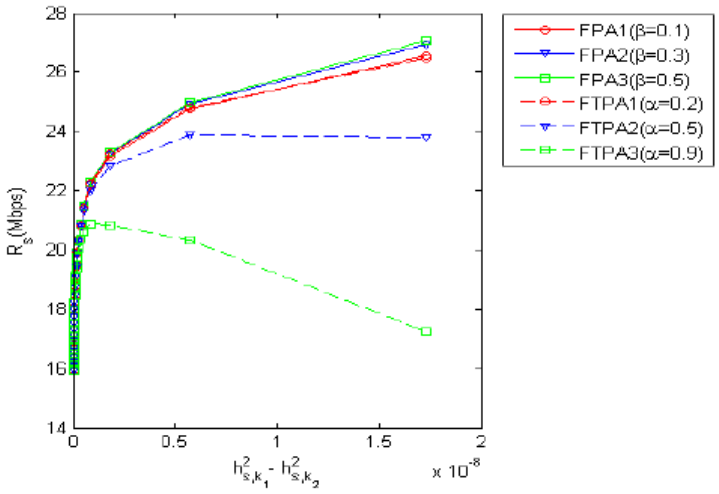


Figure 4.3: R_s achieved for different FPA and FTPA patterns, as a function of the channel gain difference d.

$$R_{s,k_2} = \frac{B}{S}\log_2\left(1 + \frac{h_{s,k_2}^{2-2\alpha}P_s}{h_{s,k_1}^{-2\alpha}h_{s,k_2}^2 P_s + \frac{N_0 B}{S}\left(h_{s,k_1}^{-2\alpha} + h_{s,k_2}^{-2\alpha}\right)}\right) \tag{4.3}$$

The derivative can be written as:

$$\frac{dR_{s,k_2}}{d\left(h_{s,k_1}^2 - h_{s,k_2}^2\right)} = \frac{1}{2h_{s,k_1}}\frac{dR_{s,k_2}}{dh_{s,k_1}} - \frac{1}{2h_{s,k_2}}\frac{dR_{s,k_2}}{dh_{s,k_2}} \tag{4.4}$$

By setting Eq. 4.4 to zero, we can show that the critical points of R_{s,k_2} verify:

$$\alpha P_s h_{s,k_2}^2\left(\frac{h_{s,k_2}}{h_{s,k_1}}\right)^2 + \alpha N_s\left(\frac{h_{s,k_2}}{h_{s,k_1}}\right)^2 + \alpha P_s h_{s,k_2}^2 + (2\alpha - 2)\,N_s = N_s\left(\frac{h_{s,k_2}}{h_{s,k_1}}\right)^{-2\alpha} \tag{4.5}$$

where $N_s = N_0 B/S$.
The detailed derivation of the proof is given in Appendix 4.A.
Eq. 4.5 shows that the critical points are dependent on h_{s,k_1}^2 and h_{s,k_2}^2. This means that the particular values of the channel gains of paired users play an important role in the throughput variation, and not only their difference. Fig. 4.4 shows a set of curves representing R_{s,k_2} as a function of d, each curve corresponding to a fixed user k_1 (and therefore to a fixed value of h_{s,k_1}^2) with all the possibilities of pairing with any other user k_2, having a lower channel gain than k_1. Only pattern FTPA2, corresponding to $\alpha = 0.5$, is considered, but the same behavior was observed for the other values of α. We can observe that, for every specific user k_1, R_{s,k_2} increases with d until reaching a

maximum value and then falls. It is also shown that the critical point from which R_{s,k_2} starts to decrease does not only depend on the channel gain difference but also on the choice of user k_1. The set of critical points, represented by the dashed curve, verify Eq. 4.5 and were proven to be local maxima using the Hessian matrix.

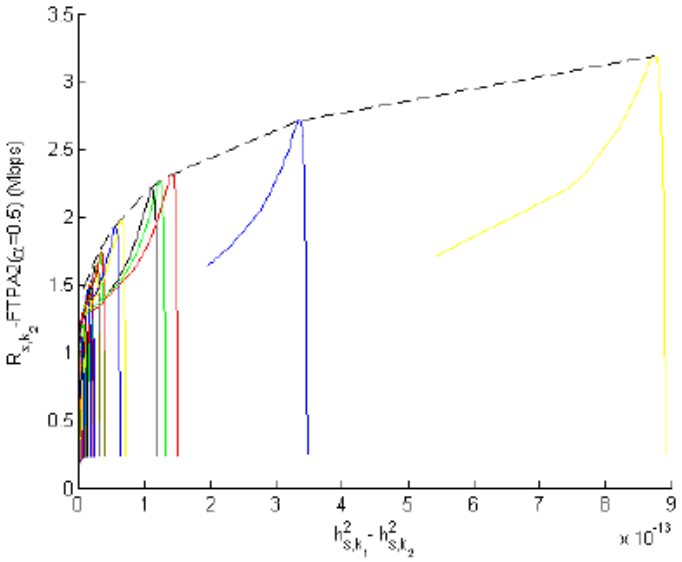


Figure 4.4: R_{s,k_2} achieved when h^2_{s,k_1} is fixed and h^2_{s,k_2} is varied, tested using the FTPA2 pattern $(\alpha = 0.5)$.

As for scenario 2, if FPA is considered, increasing the channel gain of user k_1 yields an increase of R_{s,k_1} since P_{s,k_1} is fixed and h^2_{s,k_1} is increasing. However, R_{s,k_2} remains constant. Thus, R_s will increase consequently.
If FTPA is studied, when increasing the channel gain of user k_1, the power allocated to user k_1, P_{s,k_1}, decreases. This can be verified by taking the derivative of P_{s,k_1} in terms of h_{s,k_1} in Eq. 4.1. This derivative is always negative regardless of h_{s,k_1}. However $P_{s,k_1}h^2_{s,k_1}$ is an increasing term, which can also be verified by calculating the derivative of $P_{s,k_1}h^2_{s,k_1}$ in terms of h_{s,k_1} (this derivative is always positive). Thus the resulting throughput R_{s,k_1} increases as well. Also R_{s,k_2} is found to be increasing since P_{s,k_2} increases and P_{s,k_1} decreases, for a fixed channel gain of user k_2. Consequently, R_s will increase with h^2_{s,k_1}.

It is clear that the behavior of scheduled users data rates in the two aforementioned scenarios is asymmetric. This observation comes to stress the importance of taking into account the throughput and channel gain dependency when pairing users together.

4.3 FLEXIBLE THROUGHPUT VS. FAIRNESS METRICS

At this level of our study, we have investigated in previous chapters multiple allocation and pairing techniques that can improve NOMA performance each on its own. Therefore, in this section, we intend to better exploit them so as to propose a new joint power and subband allocation technique that aims at:

- Improving the total cell throughput,
- Enhancing long-term fairness,
- Achieving fairness among users in any time scale of interest and reducing the convergence time towards a required fairness performance (this can be particularly beneficial when a quasi-constant user rate is required),
- Creating a system compatible with low-latency constraints,
- Developing new scheduling metrics that can be easily associated with unequal power allocation techniques such as waterfilling,
- Reducing the number of tested user pairs for each subband attribution.

To do so, user pairing and subband assignment are conducted in such a way to create prosperous conditions for NOMA. The common design steps of the proposed resource allocation techniques are depicted in Fig. 4.5 and detailed in the following.

4.3.1 STEP 1: INITIALIZATION AND PRIORITY ASSIGNMENT

Since, at the beginning of the process, allocated powers and rates are all set to zero, users are chosen based on the priority assignment technique that we have proposed in chapter 2, section 2.2.1.
Priorities are assigned in the BS based on the channel gains experienced by users on available subbands, prior to user selection. A priority list is used for the selection of the first user on each subband. Several sorting methods can be applied (e. g. random sorting). However, we chose a sorting scheme based on the lowest best subband channel gain, because it is expected to provide good performance in terms of cell-edge user throughput and total cell throughput:

- For each user k, select its highest channel gain, $h_{s_{best},k}$, among the channel gains experienced over all subbands,

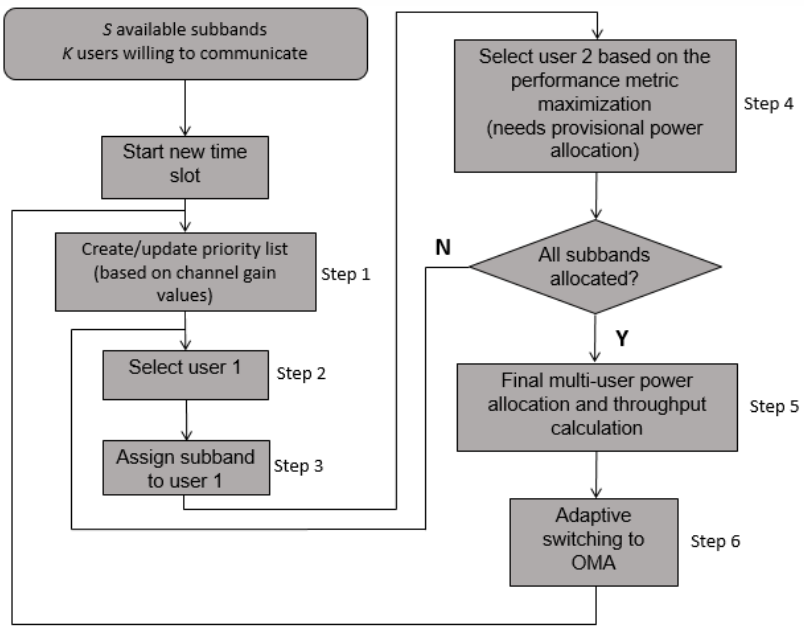


Figure 4.5: Proposed resource allocation algorithm.

- The user with the highest (rep. lowest) priority is the one having the lowest (resp. highest) best channel gain.

The resulting priority list is used while at least one user has not been assigned any subband during the assignment process. At the first time slot, the priority list is created with all K users. Users are removed from this priority list as soon as they are selected in step 2 and step 4. At subsequent time slots, the sorting of users in the priority list is updated using only the users that have not been selected yet (i.e. having zero data rates), until the list is empty.

4.3.2 STEP 2: USER SELECTION

At this step, we need to select the first user, denoted by k^*, to be scheduled next. In this sense, two cases are considered:

Case 1: the priority list is not empty (all the users have not been assigned a subband or, equivalently, any throughput, yet)
The selection of the first user is carried out according to the order given by the priority list.

Case 2: the priority list is empty (all the users have now been assigned at least a subband or, equivalently, a non-zero throughput)

The first user is selected among the set of users that need to communicate as the user experiencing the lowest projected throughput. User selection can be formulated as follows:

$$k^* = \min_k R_{k,projected}(t) \tag{4.6}$$

where $R_{k,projected}(t)$ is computed as the cumulative update of the historical rate (at the beginning of the current time slot with index t), adjusted with the achievable throughputs for the user on the set of subbands to which the user has been attributed in the current time slot. Note that the incorporation of the current rates in the PF scheduler metric was also proposed in [85]. It is calculated as follows:

$$R_{k,projected}(t) = \left(1 - \frac{1}{t_c}\right) T_k(t) + \frac{1}{t_c} R_{k,current}(t) \tag{4.7}$$

$R_{k,current}(t)$ is the current throughput, so far achieved by user k, during the current scheduling slot t, and is derived as:

$$R_{k,current}(t) = \sum_{s=1}^{S} R_{s,k}(t) \tag{4.8}$$

$R_{s,k}(t)$ is the throughput of user k on subband s, at time index t. It is calculated based on Eq. 1.3, and can be equal to zero if user k is not scheduled on subband s. Note that $R_{k,current}(t)$, $1 \leq k \leq K$, is updated each time a user k has been attributed a new subband, within the current scheduling slot t.

User selection is therefore based on the historical rate as well as on the current total achieved rate. Taking into account the historical rate in the user selection tends to achieve long-term fairness, while considering current rates allows avoiding the occurrence of a zero data rate for any user at any time scale. This can be explained by the fact that, without this term, users having high historical data rates will not have the chance to be scheduled on the current slot. But if their current rates are taken into consideration, those users will be chosen accordingly. This will ensure fairness on a short-term standpoint.

Note that in the first scheduling slot, the historical rates are zero. Hence, user selection is only based on current data rates.

4.3.3 STEP 3: SUBBAND ASSIGNMENT

User k^*, identified in the previous step, is attributed his most favorable subband, i.e., the one with the highest channel gain, and denoted by s_f. The latter will then be removed from the set of available subbands and user k^* will be removed from the priority list (unless the user has already been removed from the list).

4.3.4 STEP 4: USER PAIRING

The aim of this step is to determine the best user to be eventually paired with user k^* on subband s_f.
Rather than testing all the possible users that can be multiplexed with user k^*, we propose, in this section, a user pairing approach based on a reduced number of tested candidates. We limit our study to two scheduled users per subband ($n(s) = 2$), since it was shown in [109] that the improvement in cell throughput is 1% when the number of scheduled users per subband is equal to 3, compared to the case where it is equal to 2. However, the proposed pairing techniques described in the following have the ability to be adjusted in such a way to allow considering more than 2 users per subband if needed.

4.3.4.1 TESTED USERS PAIRS

In order to reduce the number of tested users to be scheduled with user k^*, we start by sorting, in the descending order, the channel gains experienced by all users over s_f. H_{max,s_f} (resp. H_{min,s_f}) denotes the highest (resp. lowest) channel gain over the selected subband s_f. For this purpose, let S_1 (resp. S_2) be the subset of users having channel gains higher (resp. lower) than user k^*. Two cases are investigated:

Case 1: If user k^* has a channel gain higher than $(H_{max,s_f} + H_{min,s_f})/2$, it will be considered as k_1 (i.e. as first user on s_f), and user k_2 will be selected from subset S_2. To do so, we sort the users in subset S_2 in the descending order and we proceed as follows:

- Step 1.1: Find the critical point defined in Section 4.2 (Scenario 1), from which the achievable rate on subband s_f starts to decrease. For this purpose, since h^2_{s,k_1} is known, we let $x = h^2_{s,k_2}$ in Eq. 4.5, and resolve the resulting cubic equation in order to determine the value of h^2_{s,k_2} that corresponds to the inflexion point.
- Step 1.2: Remove users from subset S_2 having channel gains on s_f less than the channel gain of the critical point (corresponding to higher distances d than that of the critical point), found in Step 1.1.

In case subset S_2 is empty, we switch to OMA and user k^* solely occupies subband s_f. Otherwise, each user in the subset S_2 constitutes with k_1 a candidate user set U which is considered for potential scheduling on subband s_f. The selection of the best candidate user set is done using one of the performance maximization metrics proposed in the following sections.

Case 2: If user k^* has a channel gain lower than $(H_{max,s_f} + H_{min,s_f})/2$, it will be considered as k_2 on subband s_f, and user k_1 will be selected from subset S_1. In this

case, we don't calculate a critical point, since it was proven in Section 4.2 (Scenario 2) that when the channel gain of user k_2 is fixed and the channel gain of user k_1 increases, the achievable throughput per subband increases accordingly. One may consider in this case choosing the user with the highest channel gain on s_f, in order to maximize the rate on s_f. However, such choice would induce a very low fairness level in the system, since certain users (with favorable channel conditions) will tend to be allocated as first users on most subbands. To avoid this problem, each user in set S_1 together with user k_2 ($= k^*$) constitutes candidate user set U on s_f, that can be scheduled on subband s_f, if selected by one of the performance maximization metrics proposed in the following sections.

4.3.4.2 PROPOSED PROVISIONAL POWER ALLOCATION (PA) FOR USER PAIRING

Having determined the candidate user pairs to be tested for subband s_f, the following step is to choose the best user pair among all candidates, through appropriate metric calculations that target throughput and/or fairness optimization. However, such metric evaluations necessitate the estimation of the achievable user rates on the allocated subbands, which in turn, include power estimates. In this context, one could resort to a joint inter and intra-subband PA similar to the one we previously proposed in section 2.2.3.1. However, applying such optimal joint PA for each candidate would result in a prohibitive complexity. For this purpose, within the pairing phase (step 4 of the algorithm), a suboptimal inter-subband PA is provisionally conducted to determine the amount of power to be attributed to the current subband s_f using a low-complexity waterfilling-based power allocation algorithm, similar to the one that was previously proposed in Section 3.4. However, in the current method, the provisional PA is done such that the amount of total distributed power when allocating the ith subband is equivalent to that of the equal power allocation: $i.P_{max}/S$, since the amount of power allocated by equal PA is P_{max}/S per subband. In other terms, at the first iteration, the total allocated power is P_{max}/S, at the second $2P_{max}/S$, etc. At the last iteration, the total distributed power is P_{max}, which corresponds to the final power allocation in Step 5 of the Algorithm.

This PA algorithm predicts the waterline level recursively from the previous level (corresponding to the most recent subband allocation stage) and from the channel gain of the user k_1 to be scheduled as first user on the current subband s_f. k_1 could be the selected as user k^* (the latter is first user on s_f in this case) or another user tested for pairing with k^* (which is the second user on s_f in this case). This leads to the following waterfilling calculation:
Eq. 3.30 will be modified such that the total distributed power at allocation stage i is

$i.P_{max}/S$:

$$\frac{iP_{\max}}{S} = \sum_{s \in S(i)} \left(W\left(S\left(i\right)\right) - \frac{N_0 B/S}{h^2_{s,k_1}} \right) \tag{4.9}$$

where $W(S(i))$ is the waterline level at the ith subband allocation stage.
At stage $i+1$, a new waterline level $W(S(i+1))$ needs to be determined. The total amount of power at the $i+1th$ stage, $(i+1)P_{max}/S$, can be written as:

$$\frac{(i+1)\,P_{\max}}{S} = \sum_{s \in S(i)} \left(W\left(S\left(i+1\right)\right) - \frac{N_0 B/S}{h^2_{s,k_1}} \right) + W\left(S\left(i+1\right)\right) - \frac{N_0 B/S}{h^2_{s_{new},k_1}} \tag{4.10}$$

s_{new} is the subband allocated at stage $i+1$.
If we denote by $N(i)$ the number of subbands in the set $S(i)$, Eq. 4.9 can be re-written as:

$$\frac{iP_{\max}}{S} = N\left(i\right) W\left(S\left(i\right)\right) - \sum_{s \in S(i)} \frac{N_0 B/S}{h^2_{s,k_1}} \tag{4.11}$$

And Eq. 4.10 as:

$$\frac{(i+1)\,P_{\max}}{S} = N\left(i\right) W\left(S\left(i+1\right)\right) - \sum_{s \in S(i)} \frac{N_0 B/S}{h^2_{s,k_1}} + W\left(S\left(i+1\right)\right) - \frac{N_0 B/S}{h^2_{s_{new},k_1}} \tag{4.12}$$

By comparing Eq. 4.11 and Eq. 4.12, the waterline at stage $i+1$ is found to be formulated as:

$$W\left(S\left(i+1\right)\right) = \frac{1}{N(i)+1} \left(N(i) W(S(i)) + \frac{P_{\max}}{S} + \frac{N_0 B/S}{h^2_{s_{new},k_1}} \right) \tag{4.13}$$

At each step of the scheduling process, for every candidate user set U, the corresponding waterline level is derived from Eq. 4.13, while taking into account user k_1 showing the highest channel gain among scheduled users in the set U, over subband s_f. Once the waterline level at the actual stage, $i+1$, is determined, power is then assigned for each candidate set U as $P_{s_f|U}$ using Eq. 3.36.
Afterwards, $P_{s_f|U}$ is divided among paired users in the set U based on the Fractional Transmit Power Allocation (FTPA) technique [21]. Then, the user rates are computed for each candidate set and the best candidate, hence the best user to be paired with k^*, is selected based on one of the metrics proposed in the following subsections.

4.3.4.3 FLEXIBLE THROUGHPUT VS. FAIRNESS MAXIMIZATION METRIC (FTFMM)

The first pairing metric denoted by FTFMM is proposed in order to cope with applications where throughput versus fairness maximization is essential. FTFMM aims at striking a flexible balance between throughput and fairness, since it is based on the instantaneous achievable rate (on the considered subband s_f), on the current user rates (on the so far attributed subbands within the current allocation slot), and on the historical rates of each possible set of candidates. A user is selected to be paired with user k^* (already chosen during the user selection step, i.e. step 2 of the algorithm), on the considered subband s_f based on the following:

$$\begin{cases} k_2 = \max\limits_{k \in S_2} \frac{R_{s_f,k}+R_{s_f,k^*}}{aR_{k,projected}+bR_{k^*,projected}} \text{ if } k^* > \frac{h^2_{\max,s_f}+h^2_{\min,s_f}}{2} \\ k_1 = k^* \end{cases} \tag{4.14}$$

$$\begin{cases} k_1 = \max\limits_{k \in S_1} \frac{R_{s_f,k}+R_{s_f,k^*}}{aR_{k,projected}+bR_{k^*,projected}} \text{ if } k^* \leq \frac{h^2_{\max,s_f}+h^2_{\min,s_f}}{2} \\ k_2 = k^* \end{cases} \tag{4.15}$$

where $R_{s_f,k}$ is the data rate achievable by user k over subband s_f. It is calculated based on Eq. 1.4 or Eq. 1.5. In order to find $R_{k,projected}$ for user k, we first start by estimating the new waterline level obtained by the allocation of subband s_f to the candidate user set U, using Eq. 4.13. U is constituted by the already chosen user k^* and the candidate user to be paired with him (as first or second user) on s_f. Then, only the powers of the two users in the candidate set U are updated according to the new waterline level, and their achievable rates on s_f are found. Finally, their projected rates are estimated using Eq. 4.7. The same procedure is repeated for each candidate user set, where the possible candidates to be tested are determined according to section 4.3.4.1.

Weights a and b can take values between 0 and 1 and verify $a + b = 1$. They are used to provide flexibility between throughput and fairness maximization.
By considering Eq. 4.14, if a is set to 1 and b to 0, user k_2 will be chosen as follows:

$$k_2 = \max_{k, h_{s_f,k} < h_{s_f,k_1}} \frac{R_{s_f,k} + R_{s_f,k_1}}{R_{k,projected}} \tag{4.16}$$

Since user k_1 has already been chosen during the user selection phase, user k_2 will be selected in a fair manner by taking the instantaneous rates as well as the projected rates of the possible candidates into consideration.
On the contrary, if a is set to 0 and b to 1, the selection of user k_2 is formulated as:

$$k_2 = \max_{k, h_{s_f,k} < h_{s_f,k_1}} \frac{R_{s_f,k} + R_{s_f,k_1}}{R_{k_1,projected}} \tag{4.17}$$

In this case, user k_2 will be chosen to be paired with k_1 according to its instantaneous rate achieved over s_f, and therefore the choice is only based on the rate maximization on the current subband. Neither his historical rate nor his current rate are taken into account. In such a situation, the notion of fairness is completely flawed.
Between these two extreme cases, and depending on the requirement of the application, parameters a and b can be varied in such a way to favor either the fairness or the throughput or to ensure balance between them.

4.3.4.4 FAIRNESS MAXIMIZATION METRIC (FMM)

We propose another pairing metric, Fairness Maximization Metric, denoted by FMM, that targets the applications where a high level of fairness is required. In this sense, we tend to minimize the rate difference among users in every scheduling slot, while maximizing the average user data rate. Therefore, FMM aims at providing almost similar data rates for all users.
To do so, the best candidate set of paired users, denoted by U_{s_f}, is selected to be scheduled over s_f based on the following metric:

$$U_{s_f} = \min_{U_p} \frac{\sum_{k=1}^{K} \left| R_{k,projected|U} - AVG_{|U} \right|}{AVG_{|U}} \tag{4.18}$$

U is a possible candidate user set, $R_{k,projected|U}$ represents the projected data rate of user k, if candidate U was chosen to be scheduled on subband s_f. For each candidate user set, we need first to estimate the waterline level based on Eq. 4.13. However, unlike the FTFMM metric, the powers over all allocated subbands, including s_f, need to be estimated according to Eq. 3.36, in order to calculate the projected values of $R_{k,projected|U}$, $1 \leq k \leq K$. Once the projected data rates of all users are estimated, the user average data rate, denoted by $AVG_{|U}$, is computed as follows:

$$AVG_{|U} = \frac{1}{K} \sum_{k=1}^{K} R_{k,projected|U} \tag{4.19}$$

FMM tends to provide the highest level of fairness among users, while improving their average rates by jointly maximizing the user average data rate and minimizing the rate difference between users when reaching that average.

4.3.5 ADAPTIVE SWITCHING TO OMA

The estimation of power conducted within each stage of the allocation process is considered as temporary until the final stage is reached. Once all subbands have been allocated

(at the final stage), the last waterline level is used to assign the definitive amount of power that should be allocated to each subband using Eq. 4.13 (Step 5 of the Algorithm). Power is then repartitioned on each subband, among paired users, in a second stage using FTPA, and data rates achieved by users are computed accordingly.

However, the improvement in spectral efficiency incurred by the use of NOMA is not systematic towards OMA. Indeed, sometimes the loss in data rate experienced by user k_1 when sharing its subband with user k_2, is significantly greater than the data rate gain achieved by k_2 on this subband. In such cases, NOMA is not the appropriate solution, and it would be better to allocate this subband to user k_1 alone. In our previous work [97], we have proposed an adaptive switching to OMA if the rates achieved by users k_1 and k_2 on subband s (R_{s,k_1} and R_{s,k_2}) verify:

$$\gamma\left(R_s - R_{s,k_1}\right) > R_{s,k_2} \;\rightarrow\; R_s > R_{s,k_1} + \frac{R_{s,k_2}}{\gamma} \tag{4.20}$$

where R_s is the data rate achieved on subband s without NOMA, i.e., by solely attributing the subband to the unique user determined in Step 2 of the algorithm (section 4.3.2). γ is a control parameter to be determined *a priori* via simulations so as to maximize system performance.

Let us consider the PF scheduler: it chooses, for subband s, either a couple of users k_1 and k_2, or one user k so as to get the maximum of the metrics: $\left[\frac{R_{s,k_1}}{T_{k_1}} + \frac{R_{s,k_2}}{T_{k_2}}; \frac{R_s}{T_{k_1}}\right]$. Therefore, in the PF scheduler, the switching from NOMA to OMA is done if:

$$\frac{R_s}{T_{k_1}} > \frac{R_{s,k_1}}{T_{k_1}} + \frac{R_{s,k_2}}{T_{k_2}} \;\rightarrow\; R_s > R_{s,k_1} + \frac{T_{k_1}}{T_{k_2}} R_{s,k_2} \tag{4.21}$$

Inspired from Eq. 4.21 and taking into account the current user rates, we propose to replace γ in Eq. 4.19 by $\frac{T_{k_2}}{T_{k_1}} + \frac{R_{k_2,projected}}{R_{k_1,projected}}$, computed at the end of the allocation process (once all subbands have been allocated). By doing so, on one hand, we avoid conducting extensive simulations in order to determine the optimal value of γ, as was done in [97]. On the other hand, we have an adaptive control parameter that depends on the users' rates (instead of a constant common value). In this case, the switching condition is tested only once at the end of the allocation process, for each time instance.

Another possibility is to include the switching within the scheduling process, by testing condition (4.21) after each subband attribution. Both possibilities have been tested and the results have shown better performance if the switching decision is delayed after the final allocation stage, since the definitive user rates achieved at the final stage are more reliable for decision.

4.4 COMPLEXITY ASSESSMENT

In the aim of assessing the implementation feasibility of the different proposed scheduling techniques, we estimated the computational load of the main allocation techniques to be integrated at the BS.

For FTFMM and FMM, user selection is done in the same way, based on Eq. 4.6: At the beginning of the allocation stage i, $8(i-1)$ multiplications and $10(i-1)$ additions are needed in order to calculate the total achieved throughput of all users using equations 1.4, and 1.5, defined in Chapter 1. Therefore, user selection in the whole S (total number of subbands) stages of the algorithm requires $8\frac{S}{2}(S-1)$ multiplications and $10\frac{S}{2}(S-1)$ additions.

User pairing based on FTFMM is computed either based on Eq. 4.14 or on Eq. 4.15:
If Eq. 4.14 is used, waterline calculation and power estimation are performed once at each allocation stage, using equations 4.13 and 3.36, with 2 multiplications and 4 additions, since the user with the highest channel gain remains the same for each candidate user set. The number of possible users k_2 to be considered for pairing with user k_1 depends on the number of users in subset S_2. In the worst case scenario, FTFMM is supposed to be calculated $K-1$ times (the number of candidate user sets). Thus, $2+11(K-1)$ multiplications and $4+7(K-1)$ additions are needed in this case.
If Eq. 4.15 is used, waterline calculation, power estimation, and FTFMM computation are performed a maximum of $K-1$ times at each allocation stage, using $13(K-1)$ multiplications and $11(K-1)$ additions, since for each candidate user set the user with the highest channel gain will be different.
In this sense, supposing that Eq. 4.14 and Eq. 4.15 are used with equal probability, user pairing based on FTFMM requires $S+12S(K-1)$ multiplications and $2S+9S(K-1)$ additions, for S allocation stages.

User pairing based on FMM is computed based on Eq. 4.18:
At each allocation stage i, a maximum of $(K-1)$ candidates user sets are also searched. At this stage, $1+(1+8i)(K-1)$ multiplications and $1+(10i+K^2+1)(K-1)$ additions are needed in order to choose the best pairing option. Thus, user pairing based on FMM requires $\sum_{i=1}^{S}[1+(1+8i)(K-1)]$ multiplications and $\sum_{i=1}^{S}[1+(10i+K^2+1)(K-1)]$ additions, for S allocation stages.

Once all subbands have been attributed, adaptive switching to OMA is tested on each subband using $11S$ multiplications and $12S$ additions.

Assuming that the calculations of $h^{-2\alpha}$, h^2, and $h^2/(N_0B/S)$ are performed only once at the beginning of the allocation process, using $3KS$ real multiplications, the resource allocation technique based on FTFMM requires a total of $4S^2-4S+15KS$ multiplications and $5S^2+9KS$ additions. Therefore, the complexity of the FTFMM-based allocation technique is $O(S^2+KS)$. On the other hand, resource allocation based

Table 4.1: Computational load of the simulated methods in terms of the number of additions and multiplications

	Number of multiplications	Number of additions
Classical PF with NOMA	$K^2S^2 + 1/2KS + 13/2K^2S$	$3/2K^2S^2 + 3K^2S - 2KS$
FTFMM	$4S^2 + 15KS - 4S$	$5S^2 + 9KS$
FMM	$4KS^2 + 11S$	$5KS^2 + K^3S - K^2S - 4KS + 12S$

on FMM necessitates $4KS^2+11S$ multiplications and $5KS^2+K^3S-K^2S-4KS+12S$ additions, which leads to an overall complexity of $O(KS^2 + K^3S)$.

As for classical NOMA PF, the calculation of the PF metric given by Eq.3.2 depends on the number of multiplexed users in the candidate user set:

In order to compute the PF metric for a candidate user set containing only 1 user, $4 + S$ multiplications and $1 + 3S/2$ additions are needed. For each candidate user set containing 2 multiplexed users, $13+2S$ multiplications and $6+3S$ additions are required.

By adding the calculations of $h^{-2\alpha}$, h^2, and $h^2/(N_0B/S)$ performed at the beginning of the allocation process, the classical NOMA PF requires a total of $3KS + C_K^1S(4 + S) + C_K^2S(13 + 2S)$ multiplications which is equal to $K^2S^2 + 1/2KS + 13/2K^2S$ and $C_K^1S(1 + 3S/2) + C_K^2S(6 + 3S)$ additions which is equal to $3/2K^2S^2 + 3K^2S - 2KS$. The complexity of NOMA PF is therefore $O(K^2S^2)$. The computational load of the main simulated methods in terms of the number of operations is summarized in Table 4.4.

4.5 PERFORMANCE ANALYSIS OF THE THROUGHPUT VS. FAIRNESS MAXIMIZATION METRICS

The proposed FTFMM and FMM metrics differ greatly from the conventional PF scheduler, whether in the way users are paired together or in the way by which the available transmit power is divided between subbands.
Those differences can enhance NOMA's performance each on its own. Therefore, in this section we aim to evaluate the proposed metrics performance, where enhancements are applied all together in a NOMA system. In addition to that, we consider applying the proposed framework for the case where the pairing step is skipped. By doing so, we can evaluate the contribution of NOMA within our framework, compared to an OMA-based system. The simulation conditions used in this section are similar to the ones in the preceding chapter (cell, channel, etc.).

Fig. 4.6 shows the system capacity achieved by each of the simulated methods in

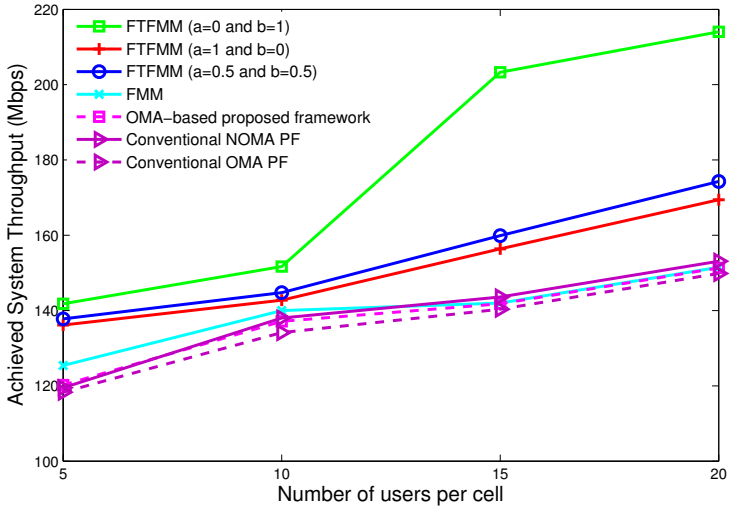


Figure 4.6: Achieved system throughput of the scheduling schemes in terms of the number of users per cell, for $S = 128$.

terms of the number of users per cell. Simulation results show that the proposed NOMA systems always outperform the OS-based system. It is also clear that, the throughput of all the presented methods increases as the number of users per cell is increased. This is due to the fact that, when the number of users per cell increases, the multi-user diversity is better exploited by the scheduling scheme, as was also observed in [95].

The gain in performance obtained by the proposed metric based on FTFMM, when compared to a conventional NOMA PF, has several reasons: First, for every channel realization, the proposed metric tries to provide rates to all users, even those experiencing bad channel conditions, since their current achieved rates (computed in the current scheduling slot) are taken into account during scheduling. On the contrary, with the conventional NOMA PF, such users wouldn't be chosen frequently. Hence, the introduction of current rates in user selection, as well as in user pairing, tends to improve the overall system capacity. Another reason for the significant gain in performance is the fact that our allocation method relies on user prioritization, whereas PF scheduling is based on sequential subband allocation. A third reason is the incorporation of unequal PA within the scheduling process, whereas PF relies on equal power allocation. This gain is observed for several values of the couples of parameters a and b. System capacity is maximized for $a = 0$ and $b = 1$ (upper throughput bound), then decreases progressively when a increases and b decreases, and reaches its lowest value for $a = 1$ and $b = 0$. Similar results were observed for different values of K. This is due to the fact that when

parameter a is set to 0, user pairing will be based on Eq. 4.13, where the maximization of the sum rates of paired users is targeted.

As for FMM, it achieves better performance compared to classical PF when the number of users per cell is limited, and almost similar performance for higher values of K.

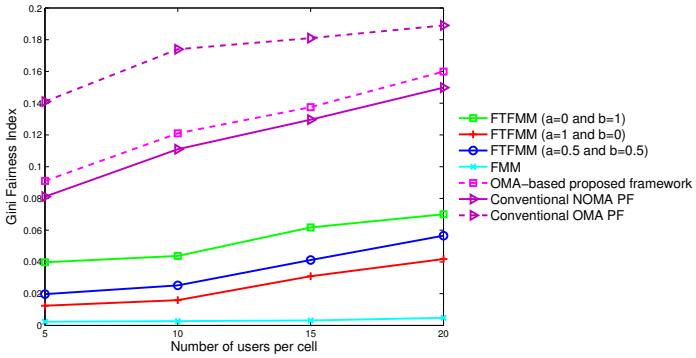


Figure 4.7: Gini Fairness Index of the proposed scheduling schemes in terms of the number of users per cell, for $S = 128$.

Fig. 4.7 shows the long-term fairness of our proposed schemes for different values of the number of users per cell. We can observe that FMM gives the best performance (Gini index is near zero, which indicates the highest level of fairness) compared to all the other simulated methods. This is due to the fact that this pairing metric tends, at each time slot, to minimize the rate difference between users, which leads to a similar achieved throughput for all users. This metric yields a throughput comparable to classical PF with NOMA but yields the highest level of fairness among all simulated methods. Therefore, FMM can be the appropriate solution for a system where fairness represents the main design constraint to be addressed. This solution may be appealing for certain scenarios of the future communication systems where all users are to be served equally, regardless of their position in the cell.

FTFMM always shows better performance when compared to a conventional NOMA PF, since when tending to achieve fairness in every scheduling slot, by taking into account current rates, long-term fairness is enhanced accordingly.
On the other hand, the impact of the parameters a and b on the user fairness is also shown in Fig. 4.7. The highest level of fairness is achieved when $a = 1$ and $b = 0$ (upper fairness bound). The observed result is in compliance with Eq. 4.11, where the projected rates are considered in user pairing, thus providing a high level of fairness.

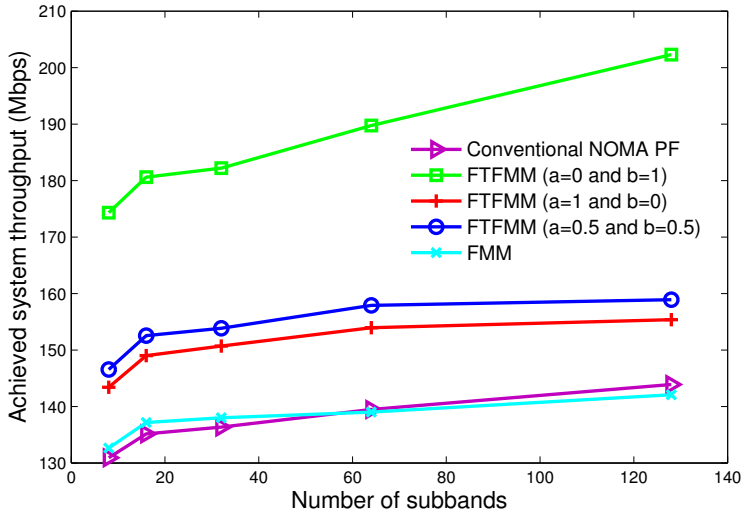


Figure 4.8: Achieved system throughput of the proposed schemes in terms of the number of subbands, for 15 users per cell.

The proposed schemes are also evaluated for multiple numbers of subbands, with a fixed total system bandwidth B of 10 MHz. Fig. 4.8 shows the achieved system throughput of the simulated methods where the number of subbands varies between 8 and 128, for a number of users per cell equal to 15. We can see that the proposed pairing technique based on FTFMM maintains good performance when compared to the conventional NOMA-based PF, even for a limited number of subbands. For instance, FTFMM with $a = 0$ and $b = 1$ shows 34% (resp. 41%) gain when the number of subbands is equal to 8 (resp. 128). On the other side, FMM shows throughput performance close to the conventional NOMA PF, with an improvement around 2.5 Mbps at $S = 16$, and a loss of almost 1.8 Mbps at $S = 128$. This shows that, compared to the conventional PF, FMM can be particularly important when the system is congested (i.e. the ratio of the number of subbands to the number of users is low).

Fig. 4.9 represents the long-term fairness of all the simulated methods in terms of the number of subbands, where the number of users per cell is equal to 15. The proposed allocation methods, based on FTFMM and FMM pairing techniques, still ensure better fairness when compared to conventional NOMA PF, for any value of the number of subbands. Note that FMM provides the highest level of fairness (Gini index lower than 0.02), even when the number of subbands is limited (S=8 and 16).

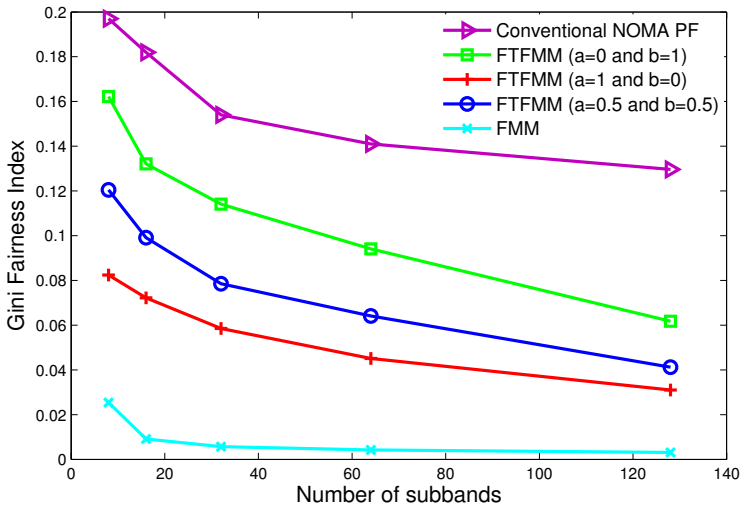


Figure 4.9: Gini Fairness index of the proposed schemes in terms of the number of subbands, for 15 users per cell.

Fig. 4.10 evaluates the short-term fairness achieved by our proposed allocation schemes with respect to the scheduling time index t, for the case where $K = 15$ and $S = 128$ In this sense, the proposed techniques are compared based on the time they take before converging to a high level of fairness. FMM shows high performance from the beginning of the allocation process ($G = 0.05$) and converges to the highest level of fairness (lowest value of index $G = 0.0010$) in a limited number of allocation steps. FTFMM shows performance close to the one achieved by FMM, when $a = 1$ and $b = 0$ and always better fairness performance compared to the conventional PF, for all simulated values of a and b. On the other hand, conventional PF with NOMA or OMA shows unfairness among users for a much longer time. Proposed allocation metrics not only show faster convergence to a high fairness level, but also provide a lower Gini indicator at the end of the scheduling window, when compared to conventional PF.

In order to assess the users' Quality of Experience (QoE) achieved through the proposed scheduling schemes, we evaluate the time it takes each user to be served for the first time, called the *rate latency*, and the variation of the user achieved rate versus the scheduling time index t. For this purpose, Fig. 4.11 shows the achieved rate of the user experiencing the largest rate latency versus time, for the different scheduling schemes, with $K = 15$ and $S = 128$

It is clear that when classical PF is used, no rate is provided for such users, for the first five scheduling slots. In addition, severe rate fluctuations are observed through

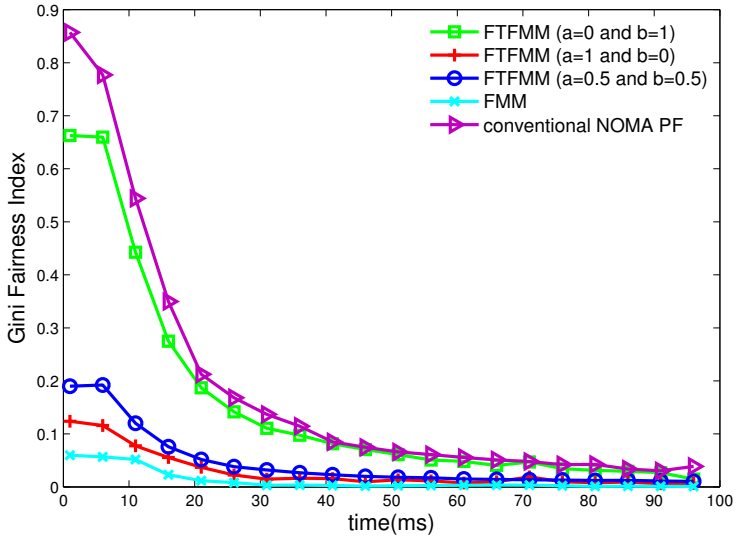


Figure 4.10: Gini Fairness Index versus time for the proposed schemes, for $K = 15$ and $S = 128$.

time. However, when the proposed metrics are considered, a rate is assigned for the least privileged users from the first scheduling slot, and remains stable for all the following slots. This behavior results from the fact that at the beginning of the scheduling process (first scheduling slot), historical rates are set to zero, and the conventional PF scheduler uses only instantaneous achievable throughputs to choose the best candidate user set. Therefore, users experiencing bad channel conditions will have a low chance to be chosen. Their achieved data rates will then be equal to zero. On the other side, using the proposed metrics (FTFMM and FMM), the treatment of the first scheduling slot is conducted differently and users are chosen depending on their current rates (measured during the current scheduling period). In this case, zero rates are avoided. Hence, latency is greatly reduced.

For the next scheduling slots, even if the historical rates are taken into account, in the conventional PF scheduling, users experiencing a large $T_k(t)$ will have less chance to be chosen, or may not to be chosen at all. In this case, in a practical system, the use of buffering becomes mandatory. It should be noted that the size of the buffer should be chosen adequately to prevent overflow when peak rates occur, as a result of a high achieved throughput (high $R_{s,k|U}(t)$). Based on calculation, the average size of the buffer should be around 118 Mbits, for the simulation setup at hand (i.e. $K = 15$ and $S = 128$). However, in our proposed schemes, buffering is not needed, since there is almost no rate fluctuation, and a better QoE is observed.

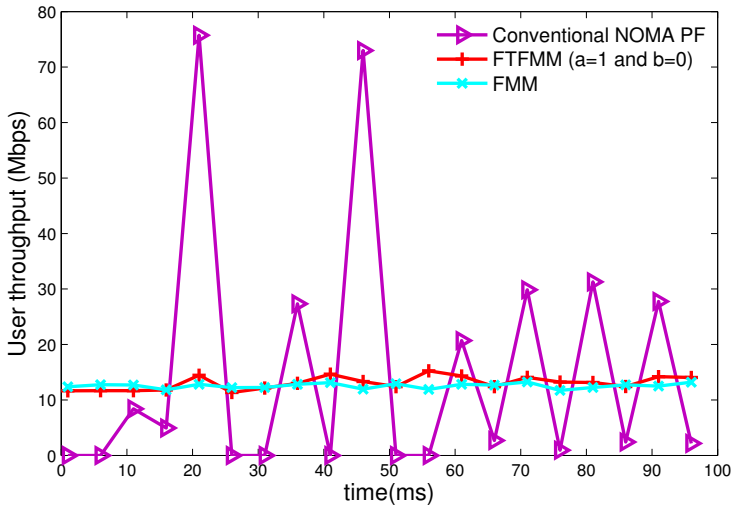


Figure 4.11: User throughput versus time for the user experiencing the largest rate latency, for $K = 15$ and $S = 128$.

Table 4.2: Computational load of the simulated methods, in terms of the number of multiplications and additions, for $K = 15$

S	8		16		32		64		128	
	NoM	NoA	NoM	NoA	NoM	NoA	NoM	NoA	NoM	NoA
Classical PF with NOMA	**31K**	**26K**	**81K**	**96K**	**277K**	**366K**	**1M**	**1.4M**	**3.8M**	**5.6M**
FTFMM (a=0.5 and b=0.5)	**2K**	**1.4K**	**4.5K**	**3.4K**	**11K**	**9.4K**	**30K**	**29K**	**93K**	**99K**
FMM	**3.9K**	**29K**	**15K**	**68K**	**61K**	**176K**	**246K**	**505K**	**984K**	**1.6M**

The computational load generated by the simulated methods, in terms of the number of multiplications (NoM) and additions (NoA), is evaluated for 15 users per cell and for multiple values of the number of subbands. Results are reported in Table 4.2.

In addition, the computational load of one allocation cycle (100 time slots) of the simulated methods, in terms of the execution time (in seconds), is reported in Table 4.3. Measurements were conducted using Matlab, run under Windows 8, on an Intel core i3 CPU, for 15 users per cell. Results reported in Table 4.2 and 4.3 show that the complexity driven by our proposed techniques is significantly reduced when compared to classical PF with NOMA, in terms of the number of multiplications and additions, as well as in terms of execution time. This gain in complexity is particularly due to the proposed user prioritization and subband allocation techniques that avoid the sequential

Table 4.3: Computational load (in seconds) of the simulated methods, for $K = 15$

S	8	16	32	64	128
Classical PF with NOMA	**4.90**	**8.30**	**15.28**	**29.97**	**61.21**
FTFMM (a=0.5 and b=0.5)	**0.65**	**1.21**	**2.45**	**6.12**	**14.99**
FMM	**0.71**	**1.65**	**3.00**	**6.13**	**15.87**

testing of a large number of candidate user sets for each subband, as is done in the conventional PF scheduler.

4.6 CONCLUSION

In this chapter, we first highlighted the main drawbacks of the PF scheduler, particularly in terms of channel gain distances between paired users on subbands. Based on the results of this study, we proposed several modifications to the PF, including user prioritization, and adequate unequal intersubband PA. Also, subband assignment and user pairing were enhanced using two proposed scheduling metrics: the first is particularly suitable for striking a balance between cell throughput and fairness, while the second allows to significantly enhance long-term and short-term fairness, especially when compared to the conventional PF scheduler. The latter was also proven to necessitate a much higher complexity compared to our proposed methods.

Chapter 5
Hybrid Broadcast-Unicast system

Wireless infrastructures are historically characterized by two distinct types of transmission modes: those that provide point-to-point unicast connectivity (e.g., cellular systems) and others that are inherently point-to-multipoint 'broadcast' (e.g., DVB system). However, with the high demands for mixed services, a reliable integration of broadcast and multicast services with the mobile broadband wireless networks has become a must [117–119]. The incorporation of broadcast can offer significant benefits to different services such as software updates, emergency messages, and public warnings, etc. In this sense, more and more wireless service providers are now offering multiple simultaneous services such as digital TV broadcasting and broadband internet access [120].

In order to achieve high transmission rates, orthogonal multiple access (OMA) based on multi-carrier modulation such as OFDM has been widely used in the majority of the broadband communication systems (DVB, WiMAX, 3G-LTE) [121]. Since NOMA has shown to greatly improve the broadband achieved rate, we aim in this chapter to propose a hybrid system that delivers broadband and broadcast services simultaneously using the non-orthogonal multiple access (NOMA) scheme, on top of the OFDMA layer. Recently, the application of NOMA to multicasting transmission has attracted some attention. Recent works in this context [122] and [123] have considered the application case of NOMA multicast. In [122], two data streams of different priorities are sent by the base station to two users via NOMA. The high priority data stream is intended to be decoded by both users while the low priority one is intended for only one user. [122] proposes also a beamforming scheme that aims to minimize the transmit power under QoS constraints of both users. This work is limited to a two-user case of NOMA multicast, whereas in practical networks, a large number of users may have the same interest to receive the same broadcasted data. The work in [123] has considered the application of NOMA to a multi-user network with mixed multicast and unicast traffic. The aim is to improve the unicast performance while maintaining the reception reliability of multicasting. This paper also investigates how the use of NOMA can prevent multicast receivers from intercepting unicast messages, since the latters are broadcasted to all the

users. However, the proposed strategy cannot cope with cases where more than one unicast user is served together with multicast users.

In this chapter, unicast and broadcast signals are allowed to be transmitted simultaneously on the same frequency platform, and to go further, we propose different techniques that allow the transmission of the broadband and broadcast messages on the same subband by multiplexing their signals in the power domain using NOMA. In addition, successive interference cancellation is assumed to provide interference-free reception of unicast signals. In most previous works, OMA and NOMA-based multicasting schemes treat multicast messages in an *a-priori* step, with a higher priority [123], e.g., unicasting will not be incorporated if all resources (frequency and power) are consumed by multicasting. This will lead to a reallocation of all the subbands within a time slot whenever a broadcast message has to be sent from the base station. Thus, in this work, we consider two families of techniques, *a-priori* and *a-posteriori* multicasting. The first one considers performing the broadcast allocation in a first step, until the broadcast rate is met, and then performing broadband on the remaining subbands. The second one considers performing broadcasting once all subbands are allocated for unicast transmission. The *a-posteriori* action is done by choosing appropriate subbands for broadcast transmission, that will be retrieved from the set of subbands dedicated to broadband services, based on specific decision metrics.

This chapter is organized as follows: In Section 5.1, the resource allocation problem is formulated. Section 5.2 describes the proposed *a-posteriori* broadcast allocation techniques. In Section 5.3, *a-priori* broadcast allocation techniques are introduced. In Section 5.4, a performance evaluation is conducted, and Section 5.5 concludes the chapter.

5.1 FORMULATION OF THE RESOURCE ALLOCATION PROBLEM

This study considers a hybrid system, where broadcast and unicast information are transmitted on the same frequency platform. We mean by broadcast the common information delivered from the base station to all users, and by unicast the delivery of private information from the base station to an intended user. By taking advantage of the incorporation of NOMA on top of the OFDMA layer, we intend throughout this chapter to propose a hybrid system supporting both broadband and broadcast transmissions based on NOMA.

The objective therein is to maximize the achievable broadband rate, also known as unicast sum rate, under the constraints of a fixed broadcast rate and a maximum allowed transmit power. The corresponding optimization problem is formulated as follows:

$$\underset{S_{broadband}}{\text{maximize}} \sum_{s \in S_{broadband}} R_s \tag{5.1}$$

Subject to

$$\sum_{s \in S_{broadcast}} R_s = R_M \tag{5.2}$$

And

$$\sum_{s \in S_T} P_s = P_{\max} \tag{5.3}$$

Where R_s is the rate achieved on subband s,
R_M is the data rate needed to transmit a broadcast message,
P_s is the power attributed to subband s, P_{max} is the maximum allowable transmit power,
$S_{unicast}$ is the set of subbands where unicast information is transmitted,
$S_{broadcast}$ is the set of subbands where broadcast information is transmitted,
and S_T is the set of total available subbands.
Note that $S_{unicast}$ and $S_{broadcast}$ are not necessarily two disjoints subsets of S_T.

5.2 A-POSTERIORI BROADCAST ALLOCATION TECHNIQUES

In this section, we assume that the broadcasting message is incorporated after all subbands have been allocated for unicast transmission. For simplicity and the good performance it provides, we assume that in a first step, subbands are allocated for unicasting based on the proportional fairness scheduler, where power is equally repartitioned among subbands, and FTPA is considered for intra-subband power allocation. Although equal power repartition is done in this section, other power allocation techniques can also be considered in the first step. *A-posteriori* action is then performed in order to choose the appropriate subbands for broadcast transmission, that will be retrieved from the broadband set of subbands, based on the two different decision metrics.

5.2.1 BROADBAND LOSS MINIMIZATION METRIC (BLMM)

A decision metric is applied in order to choose the appropriate subband, denoted by s_b, that will be excluded from the broadband set of subbands in order to be used for broadcasting. This metric is designed so as to:

- minimize the loss in broadband rate,
- maximize the broadcast message rate on s_b,

- avoid retrieving subbands for broadcasting from users having low historical rates.

The decision metric is formulated in such a way that each term represents one of the above mentioned design objectives:

$$s_b = \min_{s \in S_{unicast}} Q_s, \text{ with } Q_s = \frac{\left(R_{s|U_s} - R_{s,k\,\min}\right)^{\alpha}}{R_{s,k\,\min}{}^{\beta} \sum_{k \in U_s} R_{k,projected}} \tag{5.4}$$

Where $R_{s|U_s}$ is the data rate achieved by the set of users U_s previously chosen by the PF scheduler to be multiplexed on subband s.
$R_{s,k_{min}}$ is the achievable data rate of the worst user in the cell, i.e. showing the lowest channel gain among all users, on subband s. If subband s is chosen for broadcasting, $R_{s,k_{min}}$ will then be the broadcast rate achieved on subband s. It is assumed, in this case, that all users having a channel gain on s higher than that of user k_{min} can correctly decode the broadcast message part transmitted on subband s. Therefore, all users in the cell will be able to retrieve this message.
$R_{s|U_s} - R_{s,k_{min}}$ is the difference between the broadband rate loss and the brodcast rate gain, since users in U_s will be retrieved from subband s (if chosen for broadcast) and replaced by k_{min}.
$R_{k,projected}$ is the cumulative update of the historical rate (at the beginning of the current time slot with index t), adjusted with the achievable throughputs for the user on its allocated subbands in the current time slot. $R_{k,projected}$ is formulated as follows:

$$R_{k,projected}(t) = \left(1 - \frac{1}{t_c}\right) T_k\,(t) + \frac{1}{t_c} R_{k,current}\,(t) \tag{5.5}$$

where $T_k(t)$ is the historical rate averaged on a t_c window length, calculated by:

$$T_k(t) = \left(1 - \frac{1}{t_c}\right) T_k(t-1) + \frac{1}{t_c} \sum_{s=1}^{S} R_{s,k}(t-1) \tag{5.6}$$

where $R_{s,k}(t)$ represents the throughput of user k on subband s, at time instance t. It is calculated based on Eq. ().
$R_{k,current}$ is the rate achievable by user k on the set of unicast subbands, to which the user has been attributed in the current slot, and is calculated by:

$$R_{k,current}(t) = \sum_{s \in S_{unicast}} R_{s,k}(t) \tag{5.7}$$

where $R_{s,k}(t) = 0$ if k is not scheduled on $s \in S_{unicast}$.
It is worth pointing out that when taking $R_{k,projected}$ into account in the decision metric, we tend not only to minimize the loss in broadband rate when assigning subbands to broadcasting, but also to avoid retrieving subbands from users having low historical rates.

Once a subband s_b is found to minimize the decision metric Q_s, it will be directly added to the set of subbands dedicated for broadcasting:

$$S_{broadcast} = S_{broadcast} \cup \{s_b\} \tag{5.8}$$

And it will be retrieved from the set of subbands for unicast transmission:

$$S_{unicast} = S_{unicast} \cap \{s_b\}^c \tag{5.9}$$

After s_b has been selected, the total broadcast rate is evaluated over the set $S_{broadcast}$. If the broadcast message rate R_M has not been reached yet, another subband will be retrieved from the set of unicast subbands using the previous procedure. Fig.5.1 shows a detailed representation of the proposed BLMM technique.

Also, parameters α and β in Eq. 5.4 must be chosen adequately so as to strike a bal-

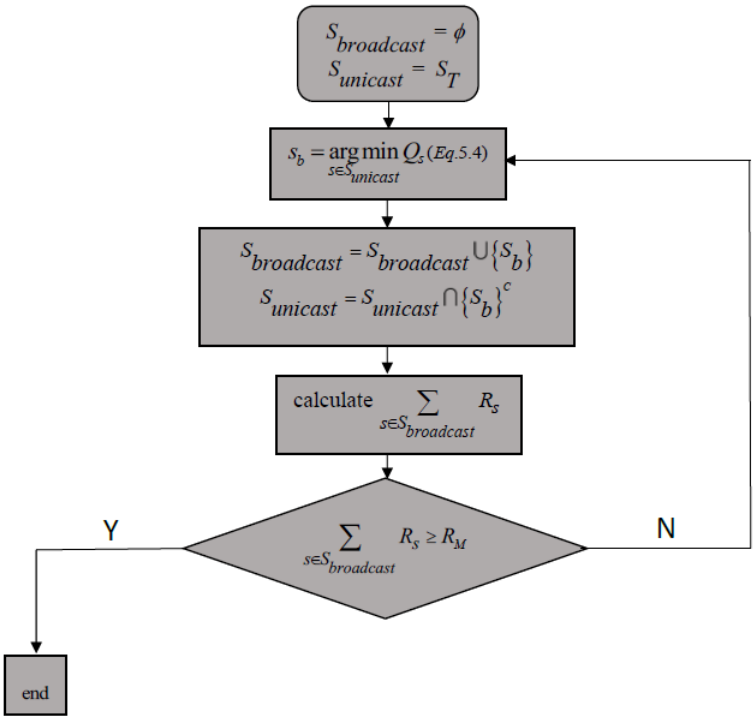


Figure 5.1: Block diagram of the broadcasting technique based on the BLMM metric

ance between maximizing the broadcast rate on s and minimizing the loss in broadband rate.

Eventhough the proposed metric works on minimizing the loss in broadband rate, it doesn't optimize adequately the broadcast achieved rate. To do so, another metric is proposed in the following section.

5.2.2 CHANNEL-BASED ALLOCATION METRIC (CBAM)

This metric focuses on maximizing the broadcast rate rather than minimizing the loss in broadband throughput. Similarly to the first proposed metric, the broadband allocation is conducted in a first step over all subbands using the PF scheduler. Then, appropriate subbands for broadcasting are sequentially retrieved from the set $S_{unicast}$ using a channel-based decision metric until the broadcast message rate is met. The channel-based decision metric is described as follows:

- For each subband s, select the lowest channel gain $H_{s,min}$ among the elements in the s^{th} row of the channel matrix **H**, of size $S \times K$, considered at time instance t.

- The subband that will be chosen for broadcast is the one having the highest worst channel gain among selected elements.

Therefore, the channel-based decision metric can be expressed as:

$$s_b = \underset{s}{argmax} \{H_{1,\min}, H_{2,\min}, ..., H_{s,\min}, ..., H_{S,\min}\} \tag{5.10}$$

Fig. 5.2 shows a detailed description of the proposed CBAM technique.

Since broadcast transmission will be based on the user with the lowest channel gain on the chosen subband, in order for all other users in the cell to be able to decode the common message, the subband having the best-worst channel gain will correspond to the subband achieving the highest broadcast rate. This metric tends to maximize the broadcast rate, thus minimizing the number of subbands needed to reach the required broadcast rate. Therefore, the broadband loss will also be minimized. However, by doing so, it can happen that we retrieve a subband for broadcasting from the broadband set that was essential for its scheduled users, and that its retrieval will significantly decrease their broadband rate. In addition to that, if we re-perform the resource allocation based on the PF scheduler, on the subbands remaining for broadband transmission, the broadband allocation will be different from the original assignment on these subbands. Based on the aforementioned points, we propose another set of allocation techniques that perform broadcast allocation as an *a-priori* step before broadband allocation, as will be described in the sequel.

5.3 A-PRIORI BROADCAST ALLOCATION TECHNIQUES

In this section, we propose to perform the broadcast allocation in a first step, until the broadcast rate is met, and then, broadband allocation takes place on the remaining

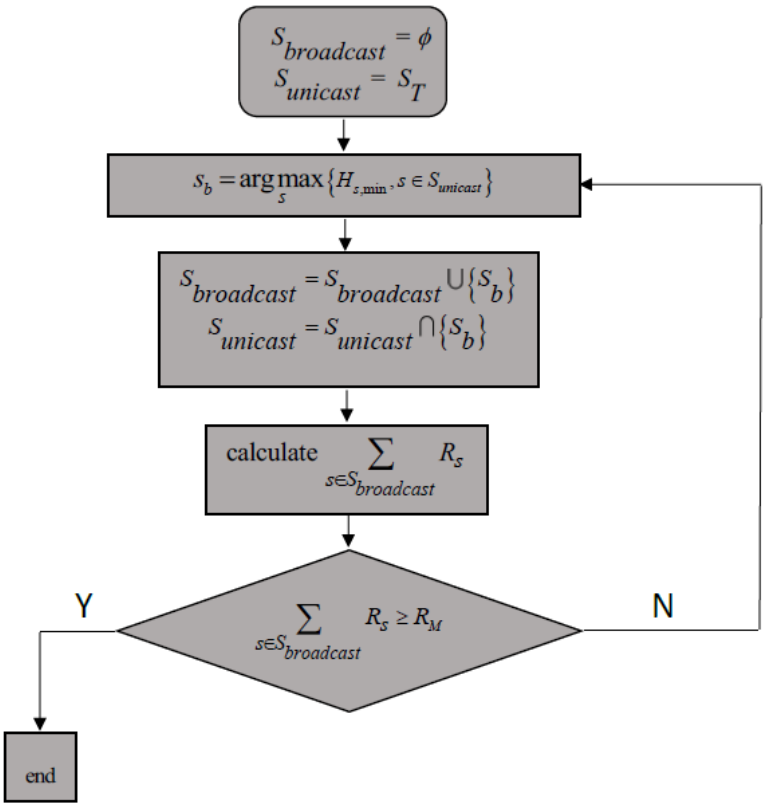


Figure 5.2: Block diagram of the broadcasting technique based on the CBAM metric

subbands. Two power allocation techniques will be evaluated herein.

5.3.1 EQUAL-POWER BASED BROADCAST ALLOCATION TECHNIQUE (EPBAT)

We start with the broadcast allocation step by allocating subbands sequentially until the broadcast message rate is met, based on the channel-based decision metric previously defined by Eq. 5.10, and using equal power repartition among subbands.
In a second step, broadband allocation is performed on the remaining set of subbands ($S_T - S_{broadcast}$) based on the PF scheduler, with equal inter-subband power allocation, and FTPA-based intra-subband power allocation.

Since a fixed amount of transmit power is attributed to each subband, it may happen that in some cases, at the end of the broadcast allocation stage, the achieved broadcast

rate is higher than the requested rate (R_M). In such cases, a power control mechanism is incorporated so as to recover the wasted additional power consumption and that will be exploited in the subsequent broadband allocation phase.

5.3.1.1 POWER CONTROL MECHANISM

This mechanism is performed at the end of the broadcast allocation step. When the broadcast rate is surpassed, power is retrieved from the latest subband allocated for broadcasting, s_{adjust}, and the recovered power is added to the power attributed to the broadband transmission.
Power is adjusted on s_{adjust} according to:

$$P_{s_{adjust}} = \frac{N_0 \frac{B}{S} \left(2^{R_M - R} - 1\right)}{h^2_{s_{adjust},k_{\min}}} \tag{5.11}$$

Where $h^2_{s_{adjust},k_{\min}}$ represents the channel gain of the lowest-channel gain user on subband s_{adjust}.
$P_{s_{adjust}}$ is the exact amount of power that should be allocated over s_{adjust}, in order to reach exactly the broadcast message rate R_M.
Let us denote by R the data rate achieved on the subbands attributed to broadcast transmission, without considering s_{adjust}. R is calculated as:

$$R = \sum_{\substack{s \in S_{broadcast} \\ s \neq s_{adjust}}} R_s \tag{5.12}$$

The amount of power recovered from subband s_{adjust} ($P_{max}/S - P_{s_{adjust}}$) is added to the amount of power considered for broadband allocation, and is equally distributed among subbands in the set $S_{broadband}$.
Let $N_{broadcast}$ be the number of subbands attributed for broadcasting. Thus, the power allocated to each subband in the broadband set is formulated as:

$$P_s = \frac{(S - N_{broadcast}) \frac{P_{\max}}{S} + \frac{P_{\max}}{S} - P_{s_{adjust}}}{S - N_{broadcast}}, \; s \in S_{broadband} \tag{5.13}$$

Fig. 5.3 shows a detailed description of the EPBAT technique.

5.3.2 EQUAL-POWER BASED HYBRID BROADCAST-BROADBAND ALLOCATION TECHNIQUE (EPHBAT)

Since, this chapter considers a hybrid system where broadcast and broadband transmissions share a single frequency platform, in this section we aim to go further and allow a

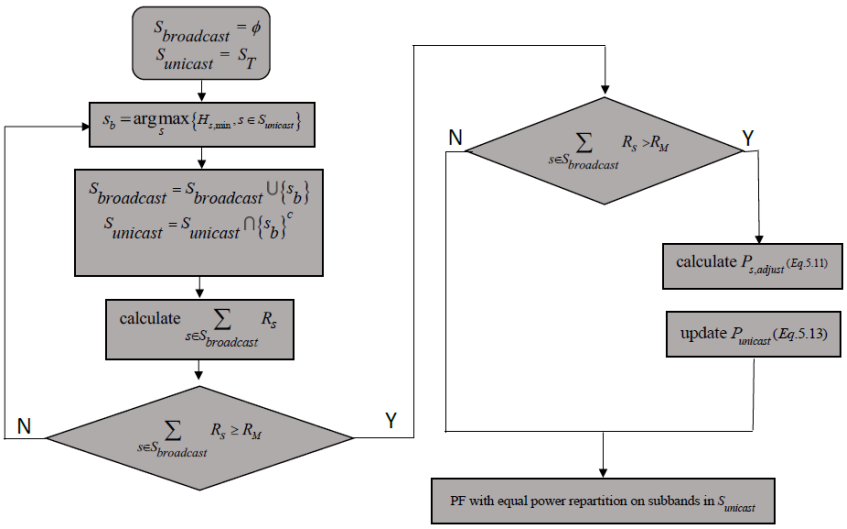


Figure 5.3: Block diagram of the broadcasting technique based on the EPBAT metric

hybrid broadcast-broadband allocation on subbands chosen for broadcast transmission. Similarly to the previous section, broadcast allocation is conducted in an *a-priori* step, where subbands are iteratively chosen for broadcasting based on the priority constraint defined in Eq. 5.10. To incorporate a hybrid broadcast-broadband allocation within the subbands chosen for broadcasting, we propose to add a broadband signal to be power-multiplexed with the common broadcast message using NOMA. The choice of the broadband signal to be multiplexed with the broadcast message, i.e. the user for which an additional broadband access is given on a broadcast subband, is not trivial. It should be chosen in such a way that all users wanting to read the common message, will not need to decode the private broadband signal. Based on the NOMA principle, if the broadband signal is transmitted on subband s, based on the highest channel gain $h^2_{s,k_{max}}$, it won't be decoded by all other users having lower channel gains on s; instead, it will be considered by them as noise. Note that the broadband user on the subband s is chosen prior to the estimation of the achievable broadcast rate on this subband, since this rate will depend on the amount of interference caused by the broadband user.
Broadband allocation then takes place on the remaining subbands, in a second step, where users are multiplexed based on NOMA, and power is equally distributed among subbands and divided within each subband based on FTPA. An illustrated example of the hybrid transmission is given in Fig. 5.4.

It is worth pointing out that the amount of power attributed for a broadcast subband will be repartitioned between the broadband signal and the common signal on this subband using FTPA. Moreover, we must verify if all users can still decode the common

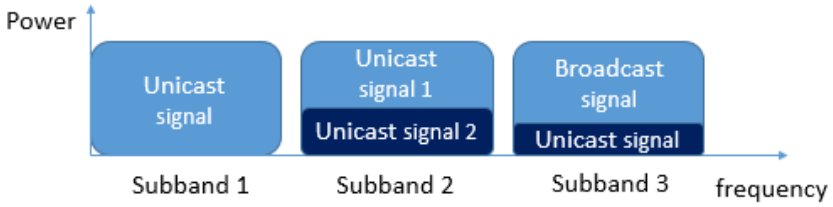


Figure 5.4: An example of EPHBAT with three subbands

message transmitted on a broadcast subband according to the channel gain $h^2_{s,k_{min}}$, using its corresponding attributed power, in spite of the presence of the broadband signal on this subband. In other words, in order to guarantee that all users having channel gains higher than $h^2_{s,k_{min}}$ can decode the broadcast message, they should achieve a data rate greater than that intended for k_{min}. This verification is detailed in Appendix 5.A.

5.3.2.1 POWER CONTROL MECHANISM

As was previously done in section 5.3.1.1, when the broadcast achieved rate surpasses the exact requested rate R_M, power should be retrieved from the latest subband allocated for broadcasting, s_{adjust}, and added to the amount of power attributed to the broadband transmission.
Let $P_{s_{adjust}}$ be the exact power that should be attributed to subband s_{adjust} after the modification; hence, $R_{s_{adjust}}$, the corresponding broadcasting rate, is expressed as follows:

$$R_{s_{adjust}} = \frac{B}{S}\log_2\left(1+\frac{\left(\beta_{s_{adjust},k_{\min}}P_{s_{adjust}}\right)h^2_{s_{adjust},k_{\min}}}{\left(\beta_{s_{adjust},k_{\max}}P_{s_{adjust}}\right)h^2_{s_{adjust},k_{\min}}+N_0\frac{B}{S}}\right) \tag{5.14}$$

with

$$\beta_{s_{adjust},k_{\min}} = \frac{\left(h^2_{s_{adjust},k_{\min}}\right)^{-\alpha}}{\left(h^2_{s_{adjust},k_{\min}}\right)^{-\alpha}+\left(h^2_{s_{adjust},k_{\max}}\right)^{-\alpha}} \tag{5.15}$$

And

$$\beta_{s_{adjust},k_{\max}} = \frac{\left(h^2_{s_{adjust},k_{\max}}\right)^{-\alpha}}{\left(h^2_{s_{adjust},k_{\min}}\right)^{-\alpha}+\left(h^2_{s_{adjust},k_{\max}}\right)^{-\alpha}} \tag{5.16}$$

Where $h^2_{s_{adjust},k_{\min}}$ is the channel gain of the least channel gain user on subband s_{adjust} and $h^2_{s_{adjust},k_{\max}}$ is the channel gain of the highest channel gain user on subband s_{adjust}. In addition, $R_{s,adjust}$ is equal to $R_M - R$, where R is the rate achieved before applying the power control mechanism over all subbands attributed for broadcasting, without considering s_{adjust}. Since broadcast subbands are multiplexed with unicast signals, R is no longer calculated as in Eq. 5.11, but based on the following equation:

$$R = \sum_{s \in S_{broadcast}} R_{s,k_{\min}} \tag{5.17}$$

After some calculations, detailed in Appendix 5.B, $P_{s_{adjust}}$ is estimated using Eq. 5.18.

$$P_{s_{adjust}} = \frac{N_0 \frac{B}{S} \left(1 - 2^{\frac{S}{B}(R_M - R)}\right)}{\left(\beta_{s,k_{\max}} 2^{\frac{S}{B}(R_M - R)} - \beta_{s,k_{\max}} - \beta_{s,k_{\min}}\right) h^2_{s,k_{\min}}} \tag{5.18}$$

Afterwards, the power recovered from subband s_{adjust} is added to the total power allocated for broadband transmission. The new amount of power that will be granted to each subband in the broadband set will be calculated based on Eq. 5.13.
At this stage, since the user with the highest channel gain is always allowed to be multiplexed with the broadcast message, on the set of broadcast subbands, the fairness, and more precisely the short-term fairness, will be penalized. Therefore, in order to mitigate this concern, we propose to take the actual achieved rates (in the current slot t) into consideration in the broadband allocation stage, as was also proposed for the enhancement of the PF scheduler in [85].

By doing so, we maintain the same broadcast allocation stage and we modify the broadband allocation stage in order to incorporate the users actual rates in the PF metric. For this purpose, the broadband allocation step will be based on the weighted NOMA-based PF (WNOPF) scheduler already proposed in Chapter 3, Section 3.3.1. In this sense, we boost the short-term fairness as well as the long-term fairness. Fig. 5.5 shows a detailed representation of the proposed EPHBAT technique.

5.3.3 WATERFILLING-BASED BROADCAST ALLOCATION TECHNIQUE (WBAT)

For all the previous methods proposed in this chapter, we have considered an equal power allocation for the set of broadcast subbands as well as for the set of broadband subbands. The number of subbands needed for broadcasting was determined iteratively, since the power per subband and the requested broadcast rate were known *a-priori*. However, in order to maximize the broadcast and broadband rates, the waterfilling principle should be utilized. In this sense, a balanced proportion between the amount of power allocated

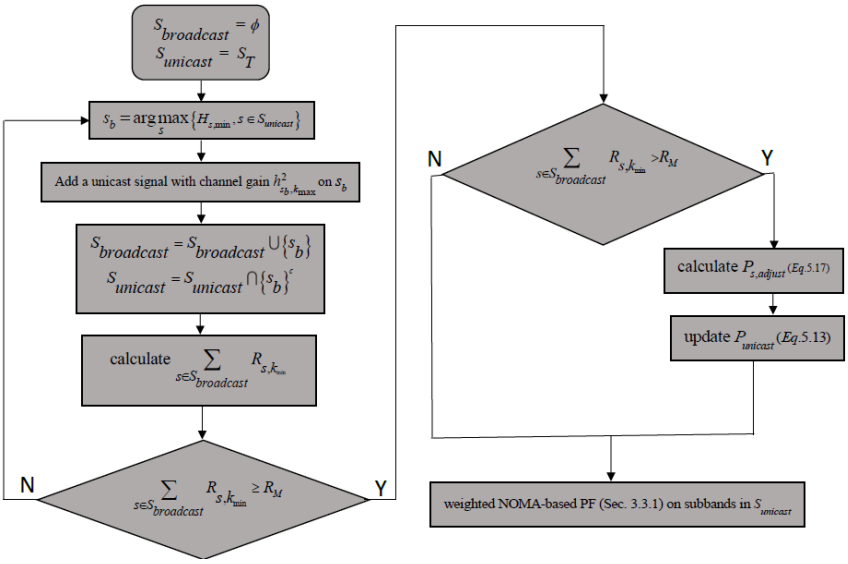


Figure 5.5: Block diagram of the broadcasting technique based on the EPHBAT metric

for broadcasting and the one allocated for broadband should be found. Let $P_{broadcast}$ and $N_{broadcast}$ be respectively the total amount of power and the total number of subbands attributed for broadcasting, in order to reach the requested rate R_M. Starting from the broadcast step, we should minimize the amount of power $P_{broadcast}$ necessary to achieve R_M, so as to maximize the remaining power to be allocated in the broadband step. Also, we should determine the optimal value of $N_{broadcast}$ so as to maximize the broadband achieved rate.
Indeed, when $N_{broadcast}$ increases, the necessary broadcast power to reach R_M generally decreases; therefore, the amount of remaining power allocated by the PF for broadband transmission is increased; at the same time, the number of subbands $N_{broadband}$ attributed by the PF scheduler for broadband transmission decreases. This explains the necessity to search for the optimum value, $N_{broadcast,optimal}$, of parameter $N_{broadcast}$.

For this purpose, we propose a new metric representing the broadband rate loss, denoted by $loss_{BB}$, and calculated as follows:

$$loss_{BB}\left(N_{broadcast}\right) = R_{BB_0} - R_{BB}\left(N_{broadcast}\right) \tag{5.19}$$

Where R_{BB_0} is the broadband achieved rate at the current allocation stage (at time slot t) in the absence of broadcasting, and R_{BB} is the broadband rate achieved in the actual allocation stage in the presence of broadcasting.
For each possible value of $N_{broadcast}$, we determine the corresponding broadband rate loss. Then, we can observe the behavior of the broadband rate loss in relation with

$N_{broadcast}$ or, equivalently, with $P_{broadcast}$. The value of $N_{broadcast}$ (or $P_{broadcast}$) that minimizes the broadband rate loss is considered as the optimal one and is used in the sequel for the hybrid broadcast-broadband resource allocation.
Determining the amount of power $P_{broadcast}$ that corresponds to a tested value of $N_{broadcast}$ is a power minimization problem on the set of subbands in $S_{broadcast}$, constrained by the target rate R_M. Its optimal solution is a waterfilling applied to $S_{broadcast}$. This waterfilling could be performed using the gradual dichotomy-based approach in [126]. However, since a new waterfilling needs to be applied each time $N_{broadcast}$ is increased by an additional subband, we can reduce the computational load by applying the low-complexity iterative waterfilling method proposed in [127]. This method allows the waterline level to be updated and the new total broadcast power to be estimated iteratively, without the need for a complete waterfilling procedure. The diagram shown in Fig. 5.6 details the steps followed in order to find the optimal value $N_{broadcast,optimal}$. The latter is determined in an offline phase, using a large number of simulation tests, with different configuration scenarios (i.e. different values of S, K, etc. and a large number of user deployments). As it will be shown in the simulation results in Section 5.5., the curve of the broadband rate loss always presents a minimum value (that corresponds to the best broadband-broadcast tradeoff) and this optimal value $N_{broadcast,optimal}$ mainly depends on the value of the total number of subbands S. Therefore, the optimal value found in the offline phase, for each value of S, can be applied in the online procedure of the WBAT technique.
Note that, we assume as initialization value for $loss_{BB}(N_{broadcast})$, $loss_{BB}(-1)$, which is equal to ∞.
Also note that the aim of the test $loss_{BB}(N_{broadcast}) > loss_{BB}(N_{broadcast}-1)$ in Fig. 5.6 is to avoid having to test all possible values of $N_{broadcast}$ (from 1 to S), since we are certain of the existence of a minimum value of $loss_{BB}$.
Besides, since the value of R_{BB_0} in (5.19) is the same regardless of the value of $N_{broadcast}$ (and of the current allocation scheme), we do not need to estimate R_{BB_0}. In this case, the minimization of $loss_{BB}$ can be replaced by a maximization of R_{BB}. However, $loss_{BB}$ will be represented in the simulation results in order to analyze its general behavior.
In the online phase, depending on the value of S, $N_{broadcast,optimal}$ subbands are chosen for broadcasting , in a sequential way, using the channel-based decision metric in Eq. 5.10. Then, once the broadcasting subbands have been determined, the necessary broadcasting power on $S_{broadcast}$ is obtained by a waterfilling constrained by the rate R_M. The remaining $N_T - N_{broadcast,optimal}$ subbands will be used for unicasting. For this purpose, the waterfilling-based PF scheduler is used to perform joint subband and unequal power allocation of the unicast signals. Fig. 5.7 details the online phase of the WBAT technique.

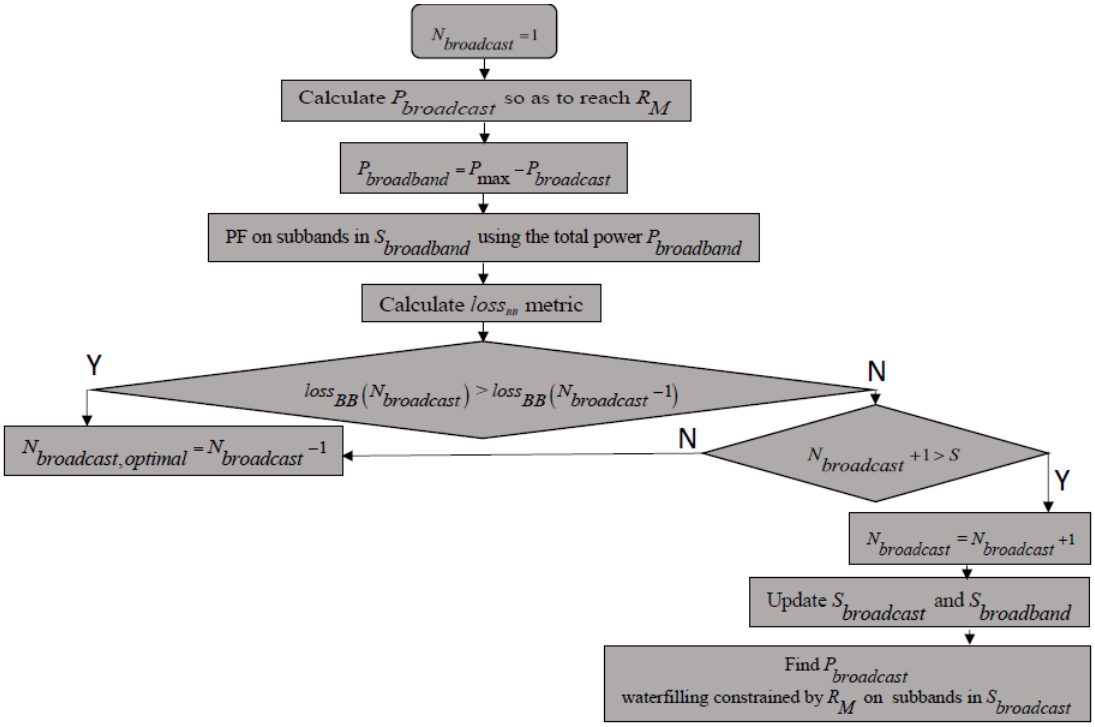


Figure 5.6: Offline procedure for determining the optimal value of $N_{broadcast}$ in WBAT

5.4 BENCHMARKING SCHEMES FOR HYBRID BROADCAST-BROADBAND SYSTEMS

There are two types of schemes that can be used for benchmarking. One is based on the use of predefined orthogonal bandwidth blocks for broadcasting, such as frequency channels with fixed bandwidths as in Frequency-division multiplexing FDM [128], or time slots with fixed durations such as in time-division multiplexing TDM [128]. Hybrid transmission systems based on FDM and TDM are shown in Fig.5.8.
Another benchmarking scheme consists on dynamically adjusting the amount of bandwidth resources allocated for broadcasting and unicasting with a higher priority given for broadcasting, as was done in [123]. For a fair comparison with our proposed schemes, we will use the dynamic FDM scheme as a benchmark, since it also treats broadcasting with a higher priority compared to broadband and tends to reach a broadcast service with a guaranteed success. In other words, both the dynamic and the proposed schemes consider that unicasting does not happen if all the power is consumed by broadcasting. In this sense, we consider two scenarios for comparison: the first one is an OMA-based dynamic FDM scheme where NOMA is not allowed neither on the broadcast nor on the broadband subbands and will be referred to as "OMA dynamic FDM". The second one

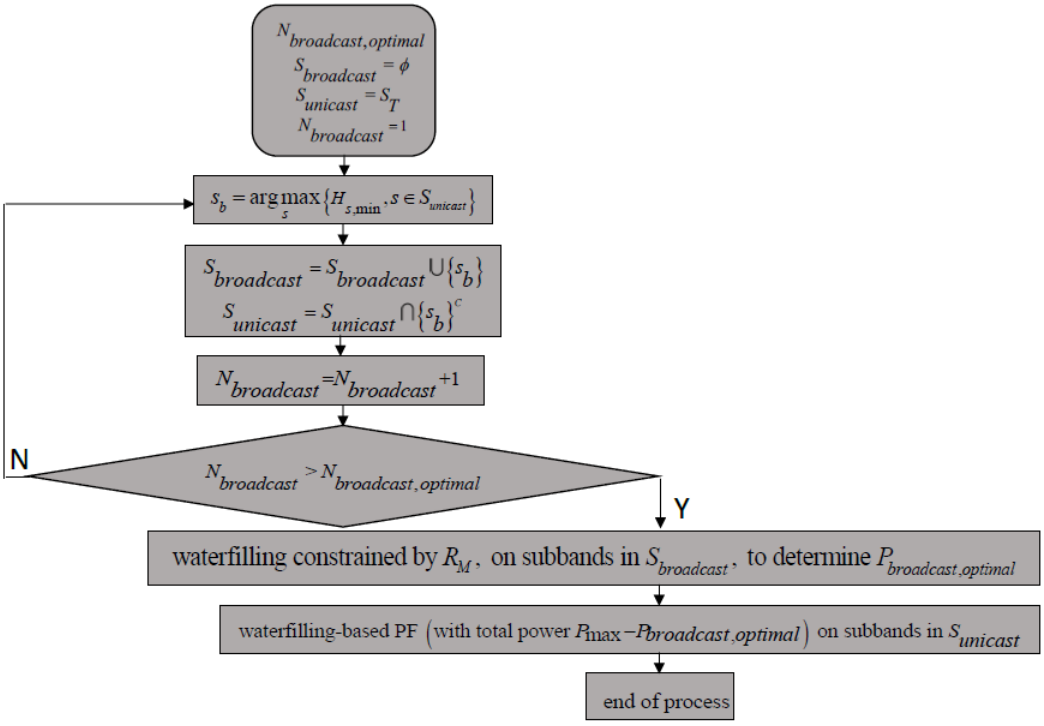


Figure 5.7: Online phase of the WBAT technique

is a NOMA-based dynamic FDM scheme, referred to as "NOMA dynamic FDM", where NOMA is allowed only on the broadband subbands. In both benchmarking scenarios, broadcasting is conducted in a first step, and subbands are sequentially chosen until the broadcast message rate is met. Then, unicasting is performed in a second step, on the remaining subbands. Note that equal power repartition is considered among subbands in both schemes and FTPA is used when NOMA is allowed.
In addition to these two benchmarking schemes, the results of the proposed techniques will also be compared to those obtained by the PF scheduler in the absence of broadcasting (i.e. $R_M = 0$), in order to show the average rate loss due to the incorporation of broadcasting in the allocation schemes. The PF scheduler metric also incorporates the actual achieved rates in the current time slot, as was done in [129].

5.5 PERFORMANCE EVALUATION

In this section, we first aim to evaluate the proposed techniques in terms of the achieved broadband throughput, the number of subbands used for broadcasting, and the achieved fairness, as a function of the number of users, the total number of subbands, and the

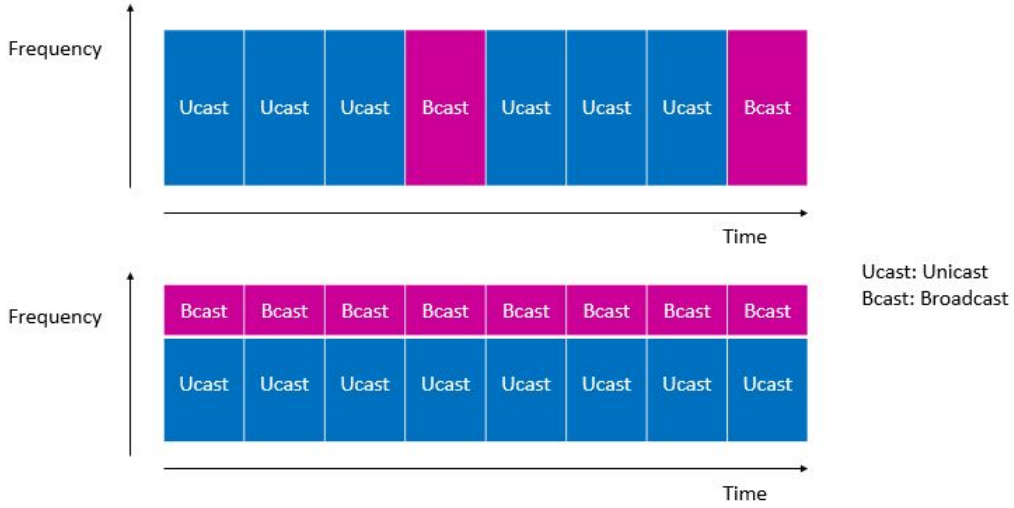


Figure 5.8: Multiplexing of broadcast and unicast: (a)TDM; (b)FDM

broadcast requested rate R_M. Simulation parameters are given in Table 5.1

Before evaluating the different proposed techniques, we start by searching for the optimal value of $N_{broadcast}$ that will be used in the evaluation of the WBAT proposed scheme. This search is performed using the offline procedure described in Section 5.3.3. For this sake, Fig. 5.9 shows the broadband rate loss in terms of $N_{broadcast}$, where the total number of subbands S is equal to 64, for different values of K.

The figure shows that the broadband rate loss is minimized when the number of subbands allocated for broadcasting is equal to 6. This optimal value is found to be almost the same regardless of the value of K. When the number of broadcast subbands is limited ($<$6), the broadcast transmission takes a great portion of power in order to reach R_M, thus the broadband achieved rate decreases, and the loss is important. Also, when the number of broadcast subbands becomes very high, the loss also increases due to the fact that the broadband transmission is deprived from an important part of its bandwidth, even though it is granted a higher amount of power.

Fig. 5.10 shows a similar behaviour, for the case of 128 subbands. In this case, 12 subbands are needed for broadcasting in order to minimize the loss in broadband rate.

We show in Fig. 5.11 the achieved broadband throughput for the allocation techniques for the case where $S = 128$ and the broadcast service rate R_M is 5 Mbps.
We can observe that the *a-priori* techniques outperform the *a-posteriori* techniques, for different number of users per cell. However, *a-posteriori* techniques present the advan-

Table 5.1: Simulation Parameters

Parameters	Values
Antenna Pattern	Omnidirectional Antenna
Number of transmitter antennas	1
Number of receiver antennas	1
Carrier Frequency	2 GHz
Sampling Frequency	15.36e$^6 Hz$
System Bandwidth	10 MHz
Cell Radius	500 m
Minimum distance between UE and BS	35 m
Channel Estimation	Ideal
UE speed	50 Km/h
Distance-dependent Path loss	$PL = 128.1 + 37.6log_{10}(d)$, d in Km
BS transmit Power	46 dBm
Number of UEs per cell	K=5, 10, 15, 20
Maximum number of multiplexed UEs	1 (OMA) 2 (NOMA)
Noise power spectral density	4.10$^{-18} mW/Hz$
α_{FTPA}	0.7

tage of enabling broadcasting by making small changes to the broadband allocation. In other words, there is no need to reallocate all the subbands within a time slot whenever a broadcast message needs to be sent. In this case, the signaling cost is minimized. Application cases with low user mobility can greatly profit from the proposed *a-posteriori* techniques.
EPHBAT shows a higher broadband rate, compared to other equal-power repartition hybrid schemes. This behavior is not surprising since a broadband message was multiplexed with the broadcast message over all broadcast subbands. WBAT yields the highest performance for different values of the number of users per cell, due to the incorporation of the waterfilling principle in broadcast and in unicast subbands, as well as the adoption of the optimal number of broadcast subbands that minimizes the broadband rate loss, with its corresponding amount of power attributed for broadcasting. Also, the loss of WBAT compared to the classical PF does not exceed 2 Mbps. All the proposed techniques show higher performance compared to the OMA and NOMA-based benchmark FDM dynamic schemes.

The proposed techniques are also evaluated, in Fig. 5.12, for different values of the total number of subbands, where the number of users is set equal to 15, and the broadcast message rate R_M is set to 5 Mbits. Again, the *a-priori* proposed techniques show better performance compared to the *a-posteriori* ones. WBAT still shows the highest

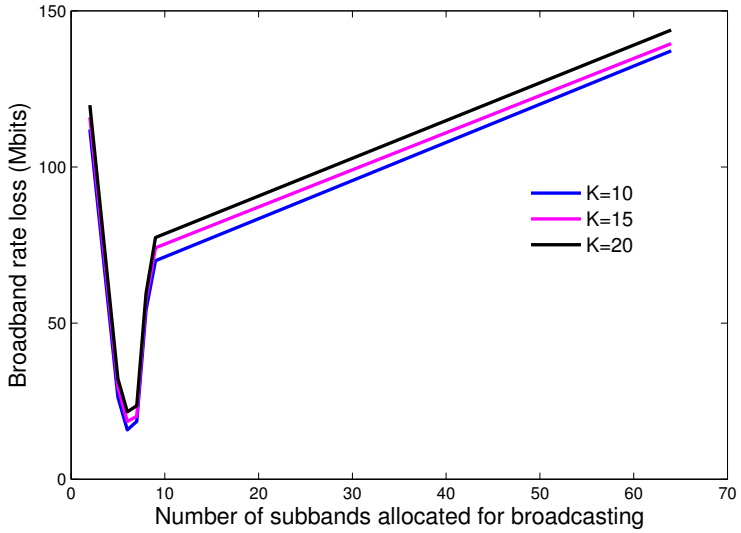


Figure 5.9: Broadband rate loss in terms of the number of broadcast subbands, where $S = 64$ and $R_M = 5$ Mbps.

achieved broadband throughput, which is very close to that of the broadband classical PF. As for the dynamic FDM schemes, they still show the lowest performance.

In order to evaluate the long-term fairness, we have used the Gini fairness index, previously defined in Chapter 1 (Section 1.2.1). Fig. 5.13 shows the Gini fairness index of the proposed techniques in terms of the number of users per cell, where S is equal to 8, and R_M is equal to 5 Mbps.

The performance hierarchy observed in Fig. 5.11 and 5.12 is still valid in Fig. 5.13 for all the proposed techniques except for EPBAT and EPHBAT. In Fig. 5.11 and 5.12, EPHBAT shows a throughput higher than that of EPBAT, since a broadband signal, corresponding to the user having the highest channel gain on a broadcast subband, has been multiplexed with the broadcast message. However, the performance is contrasting at the level of long-term fairness.
The *a-priori* techniques including the NOMA-based dynamic FDM scheme outperform the *a-posteriori* techniques, when fairness is investigated. The OMA-based dynamic FDM shows the worst fairness and the classical NOMA PF scheduler shows the highest fairness.

The number of subbands allocated for broadcasting is an important criterion that should also be evaluated. In this sense, Fig. 5.14 and 5.15 show this criterion for differ-

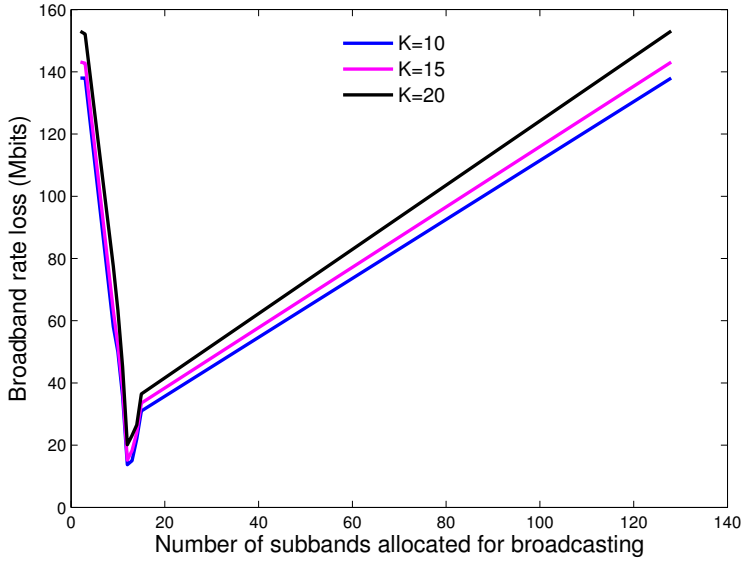


Figure 5.10: Broadband rate loss in terms of the number of broadcast subbands, for $S = 128$ and $R_M = 5$Mbps.

ent values of S and R_M, respectively. The proposed *a-posteriori* schemes allocate more subbands for broadcasting than *a-priori* ones. This result is unsurprising since *a-priori* schemes tend to optimize the broadcast allocation, thus, to minimize the number of allocated subbands. EPBAT outperforms EPHBAT for the fact that EPHBAT allocates unicast signal on the subbands used for broadcasting. Thus, the broadcast rate achieved on those subbands will decrease compared to EPBAT, because of the interference incurred by the unicast signal on the NOMA multiplexed broadcast message. WBAT still shows the best performance since it tends to optimize the broadcast allocation by finding the optimal value of $N_broadcast$ and by incorporating the waterfilling optimization.

Fig. 5.16 shows the achieved broadband throughput in terms of the broadcast message rate (R_M). The performance hierarchy observed in Fig. 5.11 and 5.12 is still valid , even when the broadcast message rate becomes high, and WBAT still shows the best performance. Therefore, the incorporation of the waterfilling principle is very efficient either at the level of broadband achieved rate or at the level of fairness.

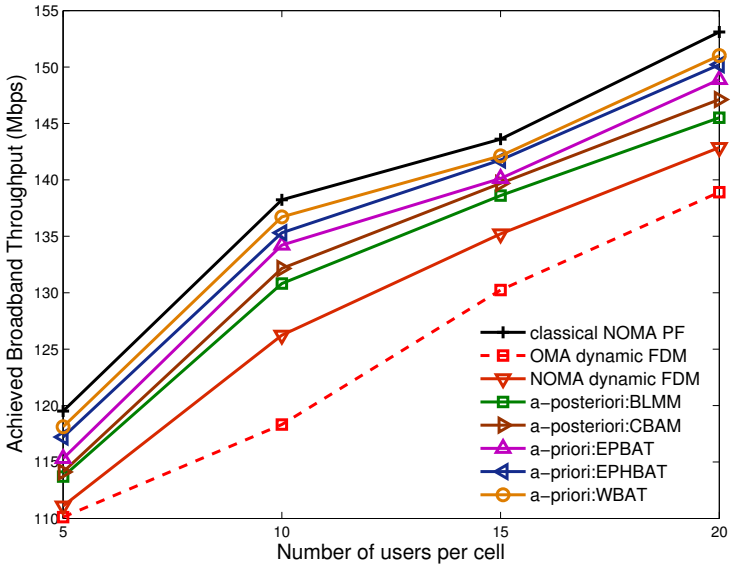


Figure 5.11: Broadband achieved rate in terms of the number of users per cell, for $S = 128$ and $R_M = 5$ Mbps.

5.6 Conclusion

In this chapter, we have proposed several techniques that enable the broadband and the broadcast messages to be transmitted on a single frequency platform. This hybrid system has shown to ensure better performance, at the level of achieved broadband throughput and long-term fairness, when compared to previous broadcasting schemes. It also exhibits a minor loss in performance, with respect to the classical broadband PF scheduler. The incorporation of a customized power allocation technique, for broadband and broadcast signals, based on the waterfilling principle has shown the best performance.

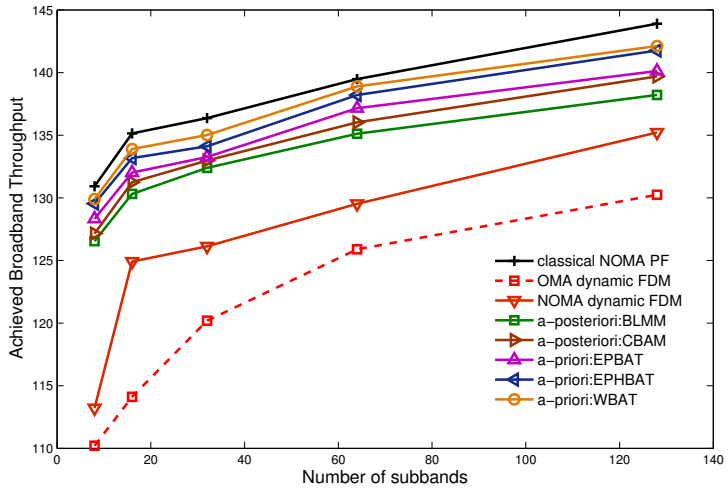


Figure 5.12: Broadband achieved rate in terms of the number of subbands, for $K = 15$ and $R_M = 5$ Mbps.

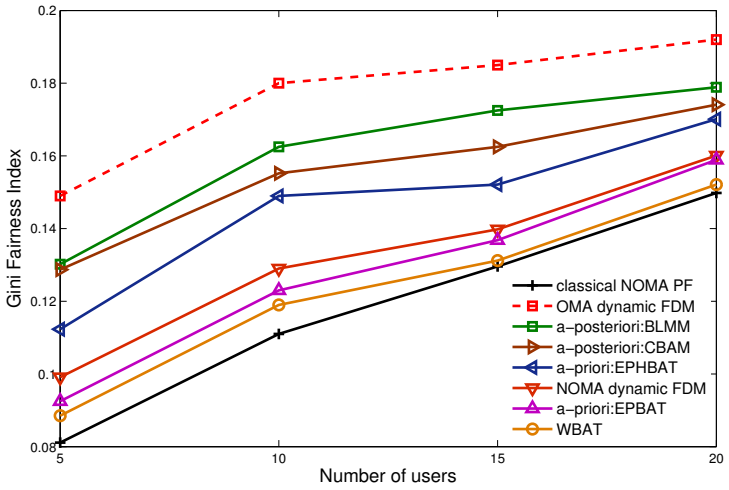


Figure 5.13: Gini fairness index in terms of the number of users per cell, for $S = 128$ and $R_M = 5$ Mbps.

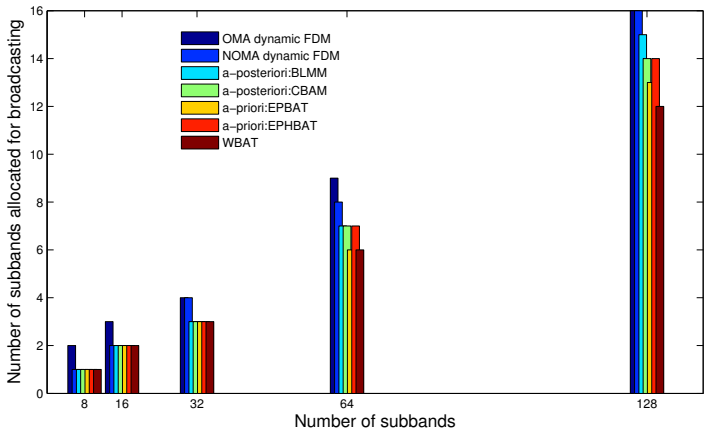


Figure 5.14: Number of subbands allocated for broadcasting, for different values of S, for $K = 10$ and $R_M = 5$ Mbps

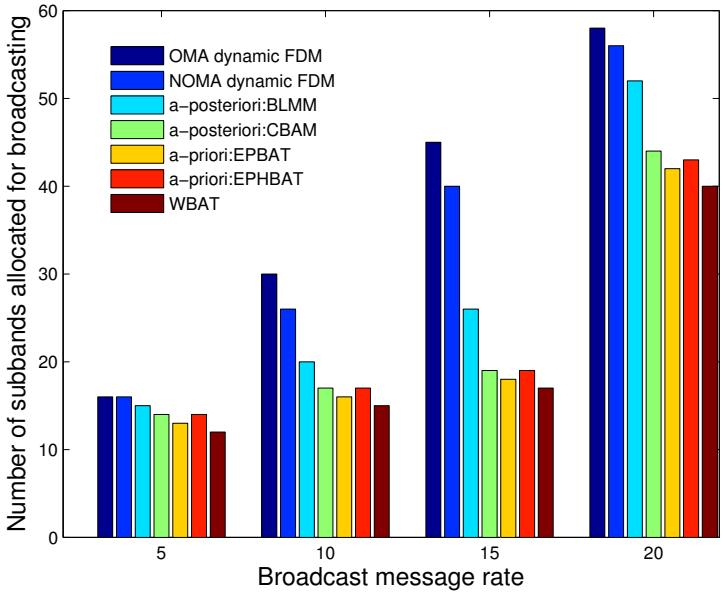


Figure 5.15: Number of subbands allocated for broadcasting, for different values of R_M, for $S = 128$ and $K = 10$

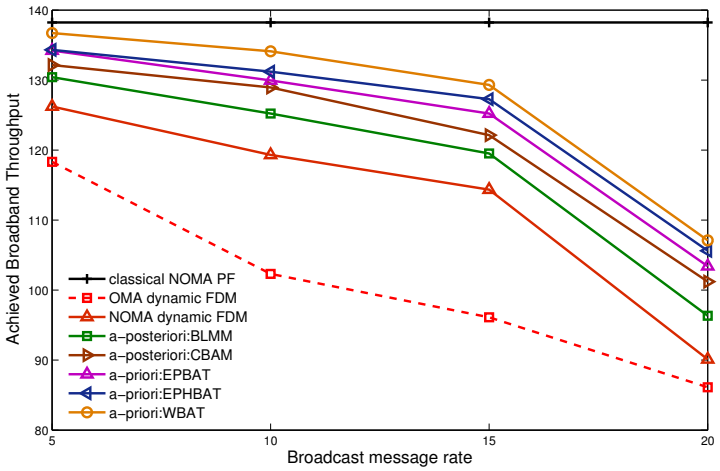


Figure 5.16: Broadband achieved rate in terms of broadcast message rate R_M, for $S = 128$ and $K = 10$.

Conclusion and future works

Conclusion

In this dissertation, we have investigated multiple resource allocation designs for downlink NOMA networks by proposing efficient solutions to corresponding optimization problems that need to be solved.

In Chapter 1, the background theory and literature review necessary for the following chapters were given. The principles of multiple access techniques for LTE and beyond were introduced, as well as the basis of OFDMA, SC-FDMA, and NOMA. Then, we have presented multiple resource and power allocation techniques considered in the literature for OFDM and NOMA systems.

In chapter 2, we considered minimizing the spectrum usage while satisfying requested data rates by a set of users. In order to meet the optimization objectives, we have investigated several design issues: the choice of user pairing, optimal or sub-optimum power allocation, fixed and adaptive intra-subband power allocation, dynamic switching from NOMA to orthogonal signaling, weighting strategies for the optimized sum-rate function, etc. An algorithm taking into account all the mentioned design issues was proposed and evaluated. Performance evaluation of the proposed scheme has shown a significant improvement in the spectral efficiency, and in the probability of success to meet each user requirements, when compared to a classical system purely based on either orthogonal or non-orthogonal signaling. In addition, the proposed optimum power allocation obtained by numerically solving an optimized allocation problem showed a substantial gain in performance when compared to the suboptimum solutions. Moreover, the proposed design for resource allocation allows different types of user prioritization by the adoption of appropriate weights in the considered optimized sum-rate metric.

In chapter 3, attention was devoted to a new scenario where channel and power allocation are studied in the context of a fixed amount of total used bandwidth, where no target rates constraints are set, and where the emphasis is on maximizing the total cell throughput as well as the user fairness. Therefore, many proposals to improve the performance of the PF scheduler for a NOMA-based system were investigated. Modifications were first done by introducing adaptive fair weights to the PF scheduling metric

with the aim of improving the user fairness while maximizing the system capacity. Then, modifications were implemented at the level of power allocation where a low-complexity waterfilling-based power allocation technique, incorporated within the proportional fairness scheduler, is proposed. Simulation results have shown an improved total user throughput, when compared to an orthogonal-based signaling and to the conventional NOMA-based PF scheduler.

In chapter 4, we have first shown that the increase in the channel gain difference between multiplexed users is not always in favor of NOMA´ s achieved throughput, when different intra-subband power allocation techniques are considered. This observation highlighted the importance of taking into account the throughput and channel gain dependency when pairing users together. Then, guided by this observation, we have designed several resource allocation techniques that aim to be fully adapted to NOMA. Those techniques consider users pairing based on controllable metrics that aim to achieve a flexible throughput and/or fairness maximization. In addition to that, the new scheduling metrics are developed such that they can be easily associated with unequal power allocation techniques such as waterfilling. As a result, the designed techniques have shown an improvement at the level of user throughput, long-term and short-term fairness.

In chapter 5, we have used the principle of superposition given by NOMA in order to allow the transmission of the broadband and broadcast messages on the same frequency platform, and to go further, we have proposed different techniques that allow the hybrid transmission to be done on the same subbands. We have considered broadcasting in an *a posteriori* step, differently from the most previous works, where OMA and NOMA-based multicasting schemes treat multicast messages in an *a priori* step. This hybrid system has shown to ensure better performance, at the level of achieved broadband throughput and long-term fairness, when compared to previous broadcasting schemes.

FUTURE WORKS

The work presented throughout this dissertation showed that NOMA is an advantageous technology for future wireless communications. Yet, there are many other interesting directions that the research of this dissertation can be extended to:

Short-term studies:

- Combining the low-complexity waterfilling-based power allocation technique proposed in chapter 3 (Section 3.4) with the weighted PF scheduling metrics proposed also in the same chapter (Section 3.2), since both propositions have provided a significant performance improvement when compared to a classical NOMA-based PF scheduler. Results are expected to be better than when considering each proposition separately.

- Introducing modifications to the fairness maximization metric proposed in chapter 4 in order to allow a certain flexibility between throughput and fairness. This can be done by introducing a controllable fairness level that can be modified for the sake of throughput improvement.

Long-term studies:

- Investigating the resource allocation techniques proposed in this thesis while considering an imperfect channel estimation. Many modifications should be done to the proposed techniques in this case: user selection, user pairing and power allocation may be different from the proposed ones. A degradation in the overall system performance is expected to be found, compared to a perfect channel estimation scenario.

- The study of a NOMA system has so far been conducted on a single-cell scenario. However, downlink NOMA should be evaluated in even more realistic scenarios, such as multi-cell environments. This implies the presence of Inter-Cell Interference (ICI) that was treated as white noise in this study. What we look forward to accomplish in our future works is to propose a new adaptive way to mitigate ICI based on the Soft Frequency Reuse (SFR) pattern. This new method will primarily enable us to mitigate ICI when downlink NOMA is used without losing the advantages of NOMA over OMA in terms of total system throughput and outer users throughput.

- Proposing muti-cell resource allocation techniques that increase the multi-cell system resistance to ICI and that are adaptive to NOMA. In this sense, we can modify our proposed resource allocation techniques, such as, FTFMM, and implement it with a corresponding ICI pattern. Classical PF can also be tested with the proposed ICI pattern for a downlink NOMA-based system.

List of publications

Patents

M-R. Hojeij, C. Abdel Nour, J. Farah, C. Douillard, Method and apparatus for power and user distribution to sub-bands in NOMA systems, European Patent Application EP16305929, 2016 July 19.

M-R. Hojeij, C. Abdel Nour, J. Farah, C. Douillard, Method and apparatus for user distribution to sub-bands in NOMA systems, European Patent Application EP16306722, 2016 December 19.

M-R. Hojeij, C. Abdel Nour, J. Farah, C. Douillard, " Method and apparatus for user distribution to sub-bands in multiple access communications systems ", European Patent Application EP18305352, 2018 March 28.

M-R. Hojeij, C. Abdel Nour, J. Farah, C. Douillard, " User distribution to sub-bands in multiple access communications systems ", European Patent Application EP18305353, 2018 March 28.

M-R. Hojeij, C. Abdel Nour, J. Farah, C. Douillard, Method and apparatus for user distribution to sub-bands in multiple access communications systems", European Patent Application EP18305354, 2018 March 28.

M-R. Hojeij, C. Abdel Nour, J. Farah, C. Douillard, Method and apparatus for power distribution to sub-bands in multiple access communications systems", European Patent Application EP18305355, 2018 March 28.

Journals and letters:

M-R. Hojeij, C. Abdel Nour, J. Farah, and C. Douillard, "Weighted proportional fair scheduling for downlink non-orthogonal multiple access", to appear in Wireless Communications and Mobile Computing, Special Issue on Efficient Spectrum Usage for Wireless

Communications, in October 2018.

M-R. Hojeij, C. Abdel Nour, J. Farah, and C. Douillard, "Waterfilling-based proportional fairness scheduler for downlink non-orthogonal multiple access", IEEE Wireless Communications Letters, February 2017.

M.-R. Hojeij, J. Farah, C. Abdel Nour, C. Douillard, New optimal and suboptimal resource allocation techniques for downlink non-orthogonal multiple access, Wireless Personal Communications, DOI 10.1007/s11277-015-2629-2.

Journal papers in preparation:

M-R. H M-R. Hojeij, C. Abdel Nour, J. Farah, and C. Douillard, "Hybrid Broadcast-Unicast system based on Non-orthogonal Multiple Access", to be submitted to IEEE Transaction on broadcasting.

M-R. Hojeij, C. Abdel Nour, J. Farah, and C. Douillard, "New resource allocation techniques for Downlink Non-Orthogonal Multiple Access".

International Conference papers:

M-R. Hojeij, C. Abdel Nour, J. Farah, and C. Douillard, "Advanced resource allocation technique for a fair downlink non-orthogonal multiple access system", accepted at 25th International Conference on Telecommunication (ICT 2018), St Malo, France, June 2018.

Z. Eddo, M. Hojeij, C. Abdel Nour, J. Farah, and C. Douillard, "Evaluation of intra-subband power allocation for a downlink non-orthogonal multiple access (NOMA) system", IEEE Globecom (GC Workshop), Washington DC, USA, Dec. 2016.

M. Hojeij, J. Farah, C. Abdel Nour, and C. Douillard, "Resource allocation in downlink non-orthogonal multiple access (NOMA) for future radio access", IEEE Vehicular Technology Conference, VTCSpring, Glasgow, Scotland, May 2015 (according to Google Scholar, this paper was cited 34 times by 15 April, 2018).

Appendix 4.A: Calculation of the derivative of user data rate with respect to the channel gain difference

This appendix describes the chain rule for computing the derivative of R_{s,k_2} with respect to $d = h_{s,k_1}^2 - h_{s,k_2}^2$.

Let $B_{sub} = B/S$ and $N_s = N_0 B/S$

$$R_{s,k_2} = B_{sub} log_2 \left(1 + \frac{P_{s,k_2} h_{s,k_2}^2}{P_{s,k_1} h_{s,k_2}^2 + N_s}\right)$$

$$R_{s,k_2} = B_{sub} log_2 \left(1 + \frac{P_s \frac{h_{s,k_2}^{-2\alpha} h_{s,k_2}^2}{h_{s,k_1}^{-2\alpha} + h_{s,k_2}^{-2\alpha}}}{P_s \frac{h_{s,k_1}^{-2\alpha} h_{s,k_2}^2}{h_{s,k_1}^{-2\alpha} + h_{s,k_2}^{-2\alpha}} + N_s}\right)$$

By replacing P_{s,k_1} and P_{s,k_2} in terms of P_s, using Eq. 4.1 and 4.2, we can write:

$$R_{s,k_2} = B_{sub} log_2 \left(1 + \frac{P_s h_{s,k_2}^{2-2\alpha}}{P_s h_{s,k_1}^{-2\alpha} h_{s,k_2}^2 + N_s h_{s,k_1}^{-2\alpha} + N_s h_{s,k_2}^{-2\alpha}}\right)$$

Let $x = h_{s,k_1}$ and $y = h_{s,k_2}$

$$R_{s,k_2} = B_{sub} log_2 \left(1 + \frac{P_s y^{2-2\alpha}}{P_s x^{-2\alpha} y^2 + N_s y^{-2\alpha} + N_s x^{-2\alpha}}\right)$$

$$\frac{dR_{s,k_2}}{d\left(x^2 - y^2\right)} = \frac{dR_{s,k_2}}{dx}\frac{dx}{d\left(x^2 - y^2\right)} + \frac{dR_{s,k_2}}{dy}\frac{dy}{d\left(x^2 - y^2\right)}$$

$$\frac{dR_{s,k_2}}{d\left(x^2 - y^2\right)} = \frac{1}{2x}\frac{dR_{s,k_2}}{dx} - \frac{1}{2y}\frac{dR_{s,k_2}}{dy}$$

$$\frac{dR_{s,k_2}}{dx} = \frac{1}{B_{sub}\ln(2)\left(1 + \frac{P_s y^{2-2\alpha}}{P_s x^{-2\alpha} y^2 + N_s x^{-2\alpha} + N_s y^{-2\alpha}}\right)}\, u$$

where $u = \dfrac{-P_s y^{2-2\alpha}\left(-2\alpha P_s x^{-2\alpha-1} y^2 - 2\alpha N_s x^{-2\alpha-1}\right)}{\left(P_s x^{-2\alpha} y^2 + N_s x^{-2\alpha} + N_s y^{-2\alpha}\right)^2}$

$$u = \frac{2\alpha {P_s}^2 y^{4-2\alpha} x^{-2\alpha-1} + 2\alpha P_s N_s y^{2-2\alpha} x^{-2\alpha-1}}{\left(P_s x^{-2\alpha} y^2 + N_s x^{-2\alpha} + N_s y^{-2\alpha}\right)^2}$$

Therefore:

$$\frac{dR_{s,k_2}}{dx} = \frac{2\alpha {P_s}^2 y^{4-2\alpha} x^{-2\alpha-1} + 2\alpha P_s N_s y^{2-2\alpha} x^{-2\alpha-1}}{B_{sub}\ln(2)\left(P_s x^{-2\alpha} y^2 + N_s x^{-2\alpha} + N_s y^{-2\alpha}\right)\left(P_s x^{-2\alpha} y^2 + N_s x^{-2\alpha} + N_s y^{-2\alpha} + P_s y^{2-2\alpha}\right)}$$

Similarly: $\dfrac{dR_{s,k_2}}{dy} = \dfrac{1}{B_{sub}\ln(2)\left(1 + \frac{P_s y^{2-2\alpha}}{P_s x^{-2\alpha} y^2 + N_s x^{-2\alpha} + N_s y^{-2\alpha}}\right)}\, v$

$$v = \frac{(2-2\alpha) P_s y^{1-2\alpha}\left(P_s x^{-2\alpha} y^2 + N_s y^{-2\alpha} + N_s x^{-2\alpha}\right) - P_s y^{2-2\alpha}\left(2y P_s x^{-2\alpha} - 2\alpha N_s y^{-2\alpha-1}\right)}{\left(P_s x^{-2\alpha} y^2 + N_s y^{-2\alpha} + N_s x^{-2\alpha}\right)^2}$$

$$\frac{dR_{s,k_2}}{dx} = \frac{-2\alpha {P_s}^2 y^{3-2\alpha} x^{-2\alpha} + 2P_s N_s y^{1-4\alpha} + (2-2\alpha) P_s N_s y^{1-2\alpha} x^{-2\alpha}}{B_{sub}\ln(2)\left(P_s x^{-2\alpha} y^2 + N_s y^{-2\alpha}\right)\left(P_s x^{-2\alpha} y^2 + N_s x^{-2\alpha} + N_s y^{-2\alpha} + P_s y^{2-2\alpha}\right)}$$

Finally we can write:

$$\frac{dR_{s,k_2}}{d\left(x^2 - y^2\right)} =$$

$$\frac{\alpha {P_s}^2 y^{4-2\alpha} x^{-2\alpha-2} + \alpha P_s N_s y^{2-2\alpha} x^{-2\alpha-2} + \alpha {P_s}^2 y^{2-2\alpha} x^{-2\alpha} - P_s N_s y^{-4\alpha} - (2-2\alpha) P_s N_s y^{-2\alpha} x^{-2\alpha}}{B_{sub}\ln(2)\left(P_s x^{-2\alpha} y^2 + N_s y^{-2\alpha} + N_s x^{-2\alpha}\right)\left(P_s x^{-2\alpha} y^2 + N_s x^{-2\alpha} + N_s y^{-2\alpha} + P_s y^{2-2\alpha}\right)}$$

To find the critical point of R_{s,k_2}, let $\dfrac{dR_{s,k_2}}{d\left(x^2 - y^2\right)} = 0$

Hence, we obtain:

$$\alpha {P_s}^2 y^{4-2\alpha} x^{-2\alpha-2} + \alpha P_s N_s y^{2-2\alpha} x^{-2\alpha-2} + \alpha {P_s}^2 y^{2-2\alpha} x^{-2\alpha} - P_s N_s y^{-4\alpha} - (2-2\alpha) P_s N_s y^{-2\alpha} x^{-2\alpha} = 0$$

After simplification and rearrangement, we get:

$$\alpha P_s y^2 \left(\frac{y}{x}\right)^2 + \alpha N_s \left(\frac{y}{x}\right)^2 + \alpha P_s y^2 + (2\alpha - 2) N_s = N_s \left(\frac{y}{x}\right)^{-2\alpha}$$

Therefore P_s can be written as:

$$P_s = \frac{N_s \left(\frac{y}{x}\right)^{-2\alpha} - \alpha N_s \left(\frac{y}{x}\right)^2 - (2\alpha - 2) N_s}{\alpha y^2 \left(\frac{y}{x}\right)^2 + \alpha y^2}$$

Appendix 5.A

In this appendix, we want to verify that when we add a broadband message on a subband attributed for broadcasting, all users can still decode the broadcast message.
Suppose that a user k wants to decode the common message on subband s. $h_{s,k}$ is the channel gain of k on subband s. The broadcast message is transmitted on subband s according to $h^2_{s,k_{min}}$ and with power $P_{s,k_{min}}$. The broadband message is added to subband s with a power $P_{s,k_{max}}$ corresponding to the highest user channel gain on s, $h^2_{s,k_{max}}$. $P_{s,k_{min}}$ and $P_{s,k_{max}}$ are calculated as follows, using FTPA:

$$P_{s,k_{\min}} = \frac{\left(h^2_{s,k_{\min}}\right)^{-\alpha}}{\left(h^2_{s,k_{\min}}\right)^{-\alpha} + \left(h^2_{s,k_{\max}}\right)^{-\alpha}} P_s \tag{5.20}$$

$$P_{s,k_{\max}} = \frac{\left(h^2_{s,k_{\max}}\right)^{-\alpha}}{\left(h^2_{s,k_{\min}}\right)^{-\alpha} + \left(h^2_{s,k_{\max}}\right)^{-\alpha}} P_s$$

where P_s is the power attributed to subband s. In order for a user k to be able to decode the broadcast message on s, its achieved data rate on $s$corresponding to $h^2_{s,k}$ and $P_{s,k_{min}}$ should be higher than the data rate achieved by the user k_{min} on s. In other words, we should verify that:

$$\log_2\left(1 + \frac{P_{s,k_{\min}} h^2_{s,k}}{P_{s,k_{\max}} h^2_{s,k} + N_0 B/S}\right) \overset{?}{>} \log_2\left(1 + \frac{P_{s,k_{\min}} h^2_{s,k_{\min}}}{P_{s,k_{\max}} h^2_{s,k_{\min}} + N_0 B/S}\right) \;\forall\; h_{s,k_{\min}} < h_{s,k} < h_{s,k_{\max}} \tag{5.22}$$

Therefore, we only need to prove that:

$$\frac{P_{s,k_{\min}} h^2_{s,k}}{P_{s,k_{\max}} h^2_{s,k} + N_0 B/S} > \frac{P_{s,k_{\min}} h^2_{s,k_{\min}}}{P_{s,k_{\max}} h^2_{s,k_{\min}} + N_0 B/S} \tag{5.23}$$

Let $A = \frac{P_{s,k_{\min}} h_{s,k}^2}{P_{s,k_{\max}} h_{s,k}^2 + N_0 B/S} - \frac{P_{s,k_{\min}} h_{s,k_{\min}}^2}{P_{s,k_{\max}} h_{s,k_{\min}}^2 + N_0 B/S}$, is A > 0?

$$\begin{aligned} A &= \frac{h_{s,k}^2 \left(P_{s,k_{\max}} h_{s,k_{\min}}^2 + N_0 B/S\right) - h_{s,k_{\min}}^2 \left(P_{s,k_{\max}} h_{s,k}^2 + N_0 B/S\right)}{\left(P_{s,k_{\max}} h_{s,k}^2 + N_0 B/S\right)\left(P_{s,k_{\max}} h_{s,k_{\min}}^2 + N_0 B/S\right)} \\ &= \frac{h_{s,k}^2 P_{s,k_{\max}} h_{s,k_{\min}}^2 + (N_0 B/S) h_{s,k}^2 - h_{s,k_{\min}}^2 P_{s,k_{\max}} h_{s,k}^2 - h_{s,k_{\min}}^2 (N_0 B/S)}{\left(P_{s,k_{\max}} h_{s,k}^2 + N_0 B/S\right) + \left(P_{s,k_{\max}} h_{s,k_{\min}}^2 + N_0 B/S\right)} \\ &= \frac{N_0 B/S \left(h_{s,k}^2 - h_{s,k_{\min}}^2\right)}{\left(P_{s,k_{\max}} h_{s,k}^2 + N_0 B/S\right) + \left(P_{s,k_{\max}} h_{s,k_{\min}}^2 + N_0 B/S\right)} \end{aligned} \tag{5.24}$$

Since $h_{s,k}$ is always greater than $h_{s,k_{min}}$, A is always positive. Consequently, any user having a channel gain greater than $h_{s,k_{min}}^2$ can correctly decode the broadcast message. In this sense, the incorporation of the best user's signal over each broadcast subband does not incur a problem in the decoding of the broadcast message.

Appendix 5.B

In this appendix, we present the detailed calculation for determining $P_{s_{adjust}}$ in the EPHBAT method.
$P_{s_{adjust}}$ is the exact power that should be attributed to the subband s_{adjust} after the modification; the corresponding broadcast rate $R_{s_{adjust}}$ is:

$$R_{s_{adjust}} = R_M - R = \frac{B}{S}\log_2\left(1 + \frac{\left(\beta_{s_{adjust},k_{\min}} P_{s_{adjust}}\right) h^2_{s_{adjust},k_{\min}}}{\left(\beta_{s_{adjust},k_{\max}} P_{s_{adjust}}\right) h^2_{s_{adjust},k_{\min}} + N_0 \frac{B}{S}}\right) \quad (5.25)$$

Where R is the broadcast data rate achieved before applying the power control mechanism, over all subbands attributed for broadcasting, without considering s_{adjust}. R is

calculated based on Eq. 4.11. $P_{s_{adjust}}$ is calculated as follows:

$$
\begin{aligned}
2^{\frac{S}{B}(R_M-R)} &= 1 + \frac{P_{s,k_{\min}} h^2_{s,k_{\min}}}{P_{s,k_{\max}} h^2_{s,k_{\min}} + N_0 \frac{B}{S}} \\
&= 1 + \frac{\beta_{s,k_{\min}} P_{s_{adjust}} h^2_{s,k_{\min}}}{\beta_{s,k_{\max}} P_{s_{adjust}} h^2_{s,k_{\min}} + N_0 \frac{B}{S}} \\
2^{\frac{S}{B}(R_M-R)} - 1 &= \frac{\beta_{s,k_{\min}} P_{s_{adjust}} h^2_{s,k_{\min}}}{\beta_{s,k_{\max}} P_{s_{adjust}} h^2_{s,k_{\min}} + N_0 \frac{B}{S}} \\
\left(2^{\frac{S}{B}(R_M-R)} - 1\right) &\left(\beta_{s,k_{\max}} P_{s_{adjust}} h^2_{s,k_{\min}} + N_0 \frac{B}{S}\right) = \beta_{s,k_{\min}} P_{s_{adjust}} h^2_{s,k_{\min}} \\
\beta_{s,k_{\max}} P_{s_{adjust}} h^2_{s,k_{\min}} 2^{\frac{S}{B}(R_M-R)} &- \beta_{s,k_{\max}} P_{s_{adjust}} h^2_{s,k_{\min}} + N_0 \frac{B}{S} 2^{\frac{S}{B}(R_M-R)} - N_0 \frac{B}{S} = \beta_{s,k_{\min}} P_{s_{adjust}} h^2_{s,k_{\min}} \\
\beta_{s,k_{\max}} P_{s_{adjust}} h^2_{s,k_{\min}} 2^{\frac{S}{B}(R_M-R)} &- \beta_{s,k_{\max}} P_{s_{adjust}} h^2_{s,k_{\min}} - \beta_{s,k_{\min}} P_{s_{adjust}} h^2_{s,k_{\min}} = N_0 \frac{B}{S} - N_0 \frac{B}{S} 2^{\frac{S}{B}(R_M-R)} \\
P_{s_{adjust}} &\left(\beta_{s,k_{\max}} h^2_{s,k_{\min}} 2^{\frac{S}{B}(R_M-R)} - \beta_{s,k_{\max}} h^2_{s,k_{\min}} - \beta_{s,k_{\min}} h^2_{s,k_{\min}}\right) = N_0 \frac{B}{S} \left(1 - 2^{\frac{S}{B}(R_M-R)}\right) \\
P_{s_{adjust}} &= \frac{N_0 \frac{B}{S} \left(1 - 2^{\frac{S}{B}(R_M-R)}\right)}{\left(\beta_{s,k_{\max}} 2^{\frac{S}{B}(R_M-R)} - \beta_{s,k_{\max}} - \beta_{s,k_{\min}}\right) h^2_{s,k_{\min}}}
\end{aligned}
\tag{5.26}
$$

Bibliography

[1] Index, Cisco Visual Networking. "Global Mobile Data Traffic Forecast Update, 20162021 White Paper, accessed on May 2, 2017."

[2] A. Benjebbour, A. Li, K. Saito, Y. Saito, Y. Kishiyama, and T. Nakamura, NOMA: From concept to standardization, in IEEE Conference on Standards for Communications and Networking (CSCN), Oct. 2015, pp. 1823.

[3] 3GPP, Study on downlink multiuser superposition transmission (MUST) for LTE, ETSI 3GPP TR 36.859, 2015.

[4] Hosein Nikopour, Eric Yi, Alireza Bayesteh, Kelvin Au, Mark Hawryluck, Hadi Baligh, and Jianglei Ma. SCMA for downlink multiple access of 5G wireless networks. In 2014 IEEE Global Communications Conference, pages 39403945. IEEE, dec 2014. 49

[5] L. Zhang, W. Li, Y. Wu, X. Wang, S. I. Park, H. M. Kim, J. Y. Lee, P. Angueira, and J. Montalban, -Division-Multiplexing: Theory and Practice," IEEE Trans. Broadcast., vol. 62, no. 1, pp. 216232, March 2016.

[6] Benjebbour, A., Saito, Y., Kishiyama, Y., Li, A., Harada, A, Nakamura, T., Concept and practical considerations of non-orthogonal multiple access (NOMA) for future radio access, in *Proc. International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)* (2013).

[7] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi. Non-Orthogonal Multiple Access (NOMA) for Future Radio Access. Vehicular Technology Conference (VTC Spring), pp.1-5, 2013.

[8] Saito, Y., Benjebbour, A., Kishiyama, Y., Nakamura, T., System-level performance evaluation of downlink non-orthogonal multiple access (NOMA), in *Proc. IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*, pp. 611,15 (2013).

[9] G. Wunder, M. Kasparick, S. ten Brink, F. Schaich, T. Wild, Y. Chen, I. Gaspar, N. Michailow, G. Fettweis, D. Ktenas, N. Cassiau, M. Dryjanski, K. Sorokosz, S. Pietrzyk, and B. Eged, System-level interfaces and performance evaluation methodology for 5G physical layer based on non-orthogonal waveforms, in Asilomar Conference on Signals, Systems and Computers, Nov. 2013, pp. 1659163.

[10] Liu, Yuanwei, et al. "Nonorthogonal Multiple Access for 5G and Beyond." Proceedings of the IEEE 105.12 (2017): 2347-2381.

[11] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, Non-Orthogonal Multiple Access (NOMA) for Cellular Future Radio Access, in 2013 IEEE 77th Vehicular Technology Conference (VTC Spring), pp. 15, June 2013.

[12] H. Zhu and J. Wang, Chunk-Based Resource Allocation in OFDMA Systems Part II: Joint Chunk, Power and Bit Allocation, IEEE Transactions on Communications, vol. 60, pp. 499509, February 2012.

[13] N. Abu-Ali, A. Taha, M. Salah, and H. Hassanein, Uplink Scheduling in LTE and LTE-Advanced: Tutorial, Survey and Evaluation Framework, IEEE Communications Surveys and Tutorials, vol. PP, no. 99, pp. 127, 2013.

[14] Y. Endo, Y. Kishiyama, and K. Higuchi, Uplink non-orthogonal access with MMSESIC in the presence of inter-cell interference, in 2012 International Symposium on Wireless Communication Systems (ISWCS), pp. 261265, Aug 2012.

[15] M. Al-Imari, P. Xiao, M. Imran, and R. Tafazolli, Uplink non-orthogonal multiple access for 5G wireless networks, in 2014 11th International Symposium on Wireless Communications Systems (ISWCS), pp. 781785, Aug 2014.

[16] R. Zhang and L. Hanzo, A Unified Treatment of Superposition Coding Aided Communications: Theory and Practice, IEEE Communications Surveys and Tutorials, vol. 13, pp. 503520, Third 2011.

[17] J.J. Caffery and G.L. Stuber, "Overview of radiolocation in CDMA cellular systems", IEEE Communications Magazine, 36(4):3845, apr 1998. 26.

[18] H. Holma and A. Toskala, LTE for UMTS - OFDMA and SCFDMA Based Radio Access. Wiley, 2009.

[19] 3rd Generation Partnership Project (3GPP), "Technical Specification Group Radio Access Network; Physical layer - General description", TS 25.201, release 1999, 1999.

[20] 3rd Generation Partnership Project (3GPP), "Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); LTE Physical Layer - General Description", TS 36.201, release 8, 2007.

[21] A. Benjebbour, A. Li, Y. Saito, Y. Kishiyama, A. Harada, and T. Nakamura "System-level performance of downlink NOMA for future LTE enhancements" in proc. IEEE Globecom, 2013.

[22] LTE, ETSI. "Evolved universal terrestrial radio access (e-utra) and evolved universal terrestrial radio access network (e-utran)(3gpp ts 36.300, version 8.11. 0 release 8), december 2009." ETSI TS 136.300 (2015): V8.

[23] ETSI, Evolved universal terrestrial radio access (E-UTRA) and evolved universal terrestrial radio access network (E-UTRAN); overall description; stage 2 (3GPP TS 36.300 version 10.5.0 Release 10), ETSI TS 136 300 V10.5.0, Nov. 2011.

[24] S. Sesia, I. Toufik, and M. Baker, LTE The UMTS Long Term Evolution: From Theory to Practice. Wiley, 2009.

[25] E. Dahlman, S. Parkvall, and J. Skold, 4G: LTE/LTE-Advanced for Mobile Broadband. Academic Press, 2011.

[26] Q. Li, G. Li, W. Lee, M. il Lee, D. Mazzarese, B. Clerckx, and Z. Li, MIMO techniques in WiMAX and LTE: a feature overview, IEEE Communications Magazine, vol. 48, no. 5, pp. 8692, May 2010.

[27] D. Astely, E. Dahlman, A. Furuskar, Y. Jading, M. Lindstrom, and S. Parkvall, LTE: the evolution of mobile broadband, IEEE Communications Magazine, vol. 47, no. 4, pp. 4451, Apr. 2009.

[28] R. V. Nee and R. Prasad, OFDM for Wireless Multimedia Communications. Artech House, 2000.

[29] T. Jiang and Y. Wu, An overview: Peak-to-average power ratio reduction techniques for OFDM signals, IEEE Transactions on Broadcasting, vol. 54, no. 2, pp. 257268, June 2008.

[30] M. Rana, M. Islam, and A. Kouzani, Peak to average power ratio analysis for LTE systems, in IEEE International Conference on Communication Software and Networks (ICCSN), Feb. 2010, pp. 516520.

[31] G. Berardinelli, L. Ruiz de Temino, S. Frattasi, M. Rahman, and P. Mogensen, OFDMA vs. SC-FDMA: performance comparison in local area IMT-A scenarios, IEEE Wireless Communications, vol. 15, no. 5, pp. 6472, Oct. 2008.

[32] E. Yaacoub and Z. Dawy, A comparison of uplink scheduling in OFDMA and SCFDMA, in IEEE International Conference on Telecommunications (ICT), Apr. 2010, pp. 466470.

[33] W. Nam, D. Bai, J. Lee, and I. Kang, Advanced interference management for 5G cellular networks, IEEE Communications Magazine, vol. 52, no. 5, pp. 5260, May 2014.

[34] J. Andrews, S. Singh, Q. Ye, X. Lin, and H. Dhillon, An overview of load balancing in hetnets: old myths and open problems, IEEE Wireless Communications, vol. 21, no. 2, pp. 1825, Apr. 2014.

[35] G. Wunder, P. Jung, M. Kasparick, T. Wild, F. Schaich, Y. Chen, S. Brink, I. Gaspar, N. Michailow, A. Festag, L. Mendes, N. Cassiau, D. Ktenas, M. Dryjanski, S. Pietrzyk, B. Eged, P. Vago, and F. Wiedmann, 5GNOW: non-orthogonal, asynchronous waveforms for future mobile applications, IEEE Communications Magazine, vol. 52, no. 2, pp. 97105, Feb. 2014.

[36] Tse, D., Viswanath, P. (2005). Fundamentals of wireless communication. Cambridge: Cambridge University Press.

[37] Takeda, T., Higuchi, K. (2011). Enhanced user fairness using non-orthogonal access with SIC in cellular uplink. In IEEE vehicular technology conference.

[38] Tomida, S., Higuchi, K. (2011). Non-orthogonal Access with SIC in Cellular Downlink for User Fairness Enhancement. In International symposium on intelligent signal processing and communication systems (ISPACS), pp. 16.

[39] Otao, N., Kishiyama, Y., Higuchi, K. (2012). Performance of non-orthogonal access with SIC in cellular downlink using proportional fair-based resource allocation. In International symposium on wireless communication systems, pp. 476480.

[40] Schaepperle, J. (2010). Throughput of a wireless cell using superposition based multiple-access with optimized scheduling. In The international symposium on personal, indoor and mobile radio communications, pp. 21221.

[41] Saito, Y., Kishiyama, Y., Benjebbour, A., Nakamura, T., Li, A., Higuchi, K. (2013). Non-orthogonal multiple access (NOMA) for future radio access. In IEEE Vehicular technology conference.

[42] Benjebbour, A., Saito, Y., Kishiyama, Y., Li, A., Harada, A, Nakamura, T. (2013). Concept and practical considerations of non-orthogonal multiple access (NOMA) for future radio access. International symposium on intelligent signal processing and communication systems (ISPACS).

[43] Kim, B., Lim, S., Kim, H., Suh, S., Kwun, J., Choi, S., Lee, C., Lee, S., Hong, D. (2013). Non-orthogonal Multiple Access in a Downlink Multiuser Beamforming System, MILCOM.

[44] Saito, Y., Benjebbour, A., Kishiyama, Y., Nakamura, T. (2013). System-level performance evaluation of downlink non-orthogonal multiple access (NOMA). In The international symposium on personal, indoor and mobile radio communications, pp. 611615.

[45] P. Viswanath, D. Tse, and R. Laroia. Opportunistic Beamforming using Dumb Antennas. IEEE Transactions on Information Theory, vol. 82, n.6, pp. 1277-1294, 2002.

[46] F.P. Kelly, A.K. Maulloo, Tan, and D.K.H., Rate control for communication networks: shadow prices, proportional fairness and stability, Journal of the Operational Research Society, 1998, 49(3).

[47] P. Viswanath, D. Tse, and R. LAroia, Opportunistic Beamforming using Dumb Antennas, IEEE Trans. Inf. Theory, 2002, vol 48(6), pp. 1277- 1294.

[48] M. Kountouris, and D. Gesbert, Memory-based Opportunistic Multiuser Beamforming, in Proc. IEEE Int. Symp. Inf. Theory, 2005.

[49] S. Tomida, and K. Higuchi, Non-orthogonal Access with SIC in Cellular Downlink for User Fairness Enhancement, in Proc. Int. Symp. on Intelligent Signal Processing and Comm. Systems (ISPACS),2011, pp. 1-6.

[50] A. Benjebbour, Y. Saito, Y. Kishiyama, A. Li, A. Harada, and T. Nakamura, Concept and practical considerations of non-orthogonal multiple access (NOMA) for future radio access, in Proc. Int. Symp. on Intelligent Signal Processing and Commun. Systems (ISPACS), 2013.

[51] Y. Saito, A. Benjebbour, Y. Kishiyama, and T. Nakamura, System-level performance evaluation of downlink non-orthogonal multiple access (NOMA), in Proc. IEEE Personal Indoor and Mobile Radio Communications (PIMRC), September 2013.

[52] T. Girici, C. Zhu, J. R. Agre, and A. Ephremides, Proportional fair scheduling algorithm in OFDMA-based wireless systems with QoS constraints, Journal of Communications and Networks, vol. 12, no. 1, pp. 3042, Feb. 2010.

[53] Dianati, M., Shen, X., Naik, S., A new fairness index for radio resource allocation in wireless networks, in Proc. IEEE Wireless Communication and Networking Conference (2005).

[54] A. V. Babu and L. Jacob, Fairness analysis of IEEE 802.11 mesh networks, IEEE trans. Veh. Tech., vol. 56, no. 5, pp. 3073-3088, Sep. 2007

[55] F. Sokmen and T. Girici, Uplink resource allocation algorithms for single-carrier FDMA systems, in IEEE European Wireless Conference (EW), Apr. 2010, pp. 339345.

[56] M. Wahlqvist, H. Olofsson, M. Ericson, C. Ostberg, and R. Larsson. Capacity comparison of an OFDM based multiple access system using different dynamic resource allocation. In Proc. IEEE Veh. Technol. Conf. (VTC), volume 3, pages 16641668,May 1997.

[57] W. Yu,W. Rhee, S. Boyd, and J.M. Cioffi. Iterative water-filling for Gaussian vector multiple-access channels. IEEE Trans. Info. Theory, 50(1):145152, Jan. 2004.

[58] J. Jang and K. B. Lee. Transmit power adaptation for multiuser OFDM systems. IEEE J. Selected Areas Commun., 21(2):171178, Feb. 2003.

[59] D. Kivanc, G. Li, and H. Liu. Computationally efficient bandwidth allocation and power control for OFDMA. IEEE Trans. Wireless Commun., 2(6):11501158, Nov. 2003.

[60] D. Kivanc and H. Liu. Subcarrier allocation and power control for OFDMA. In Proc. Asilomar Conference on Signals, Systems and Computers, volume 1, pages 147151, Oct./Nov. 2000.

[61] W. Rhee and J.M. Cioffi. Increase in capacity ofmultiuser OFDMsystemusing dynamic subchannel allocation. In Proc. IEEE Veh. Technol. Conf. (VTC), volume 2, pages 10851089,May 2000.

[62] Z. Shen, J. G. Andrews, and B. L. Evans. Adaptive resource allocation in multiuser OFDM systems with proportional rate constraints. IEEE Trans. Wireless Commun., 4(6):27262737, Nov. 2005.

[63] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, Non-orthogonal multiple access (NOMA) for cellular future radio access, in IEEE Vehicular Technology Conference (VTC Spring), June 2013, pp. 15.

[64] Y. Saito, A. Benjebbour, Y. Kishiyama, and T. Nakamura, System-level performance evaluation of downlink nonorthogonal multiple access (NOMA), in IEEE International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC), Sept. 2013, pp. 611615.

[65] L. Lei, D. Yuan, C. K. Ho, and S. Sun, Joint optimization of power and channel allocation with non-orthogonal multiple access for 5G cellular systems, in IEEE Global Communications Conference (GLOBECOM), Dec. 2015.

[66] Kelly, F.P., Maulloo, A.K., Tan, D.K.H., Rate control for communication networks: shadow prices, proportional fairness and stability, *Journal of the Operational Research Society*, 49(3) (1998).

[67] Viswanath, P., Tse, D., LAroia, R., Opportunistic Beamforming using Dumb Antennas, *IEEE Trans. Inf. Theory*, vol 48(6), pp. 1277-1294 (2002).

[68] A. Benjebbour, A. Li, Y. Saito, Y. Kishiyama, A. Harada, and T. Nakamura. Concept and Practical Considerations of Non-orthogonal Multiple Access (NOMA) for Future Radio Access. Intelligent Signal Processing and Communications Systems (ISPACS), pp. 770-774, 2013.

[69] N. Otao, Y. Kishiyama, and K. Higushi. Performance of non-orthogonal access with SIC in cellular downlink using proportional fair-based resource allocation. International Symposium on Wireless Communication Systems (ISWSC), pp. 476-480, 2012.

[70] Benjebbour, A., Li, A., Saito, Y., Kishiyama, Y., Harada, A., Nakamura, T. (2013). System-level performance of downlink NOMA for future LTE enhancements. In IEEE Globecom.

[71] Caire, G., Shamai, S. (2003). On the achievable throughput of a multi-antenna Gaussian broadcast channel. IEEE Transactions on Information Theory, 49(7), 16921706.

[72] Y. Saito, A. Benjebbour, Y. Kishiyama, and T. Nakamura System-Level Performance of Downlink Non-orthogonal Multiple Access (NOMA) Under Various Environments in proc. IEEE 81st VTC, 2015

[73] Y. Hayashi, Y. Kishiyama, and K. Higuchi. Investigations on power allocation among beams in non-orthogonal access with random beamforming and intra-beam SIC for cellular MIMO downlink, in proc. 78th IEEE VTC, 2013.

[74] 3GPP, TS 36.213, Evolved Universal Terrestrial Radio Access (EUTRA); Physical layer procedures, May 2009.

[75] A. Benjebbour, A. Li, Y. Kishiyama, H. Jiang, and T. Nakamura, System-Level Performance of Downlink NOMA Combined with SUMIMO for Future LTE Enhancements, in proc. IEEE Globecom, Dec. 2014.

[76] M.-R. Hojeij, J. Farah, C. A. Nour, and C. Douillard, Resource allocation in downlink non-orthogonal multiple access (NOMA) for future radio access, in Proc. 81st IEEE VTC, Glasgow, U.K., 2015, pp. 16.

[77] S. Haykin, (WCS)Communication Systems 4th Edition W/Study Tips Set. New York, NY, USA: Wiley, 2005.

[78] M. Dianati, X. Shen, and S. Naik, A new fairness index for radio resource allocation in wireless networks, in Proc. IEEE Wireless Commun. Netw. Conf., New Orleans, LA, USA, 2005, pp. 712717.

[79] F. P. Kelly, A. K. Maulloo, and D. K. H. Tan, Rate control for communication networks: Shadow prices, proportional fairness and stability, J. Oper. Res. Soc., vol. 49, no. 3, pp. 237252, 1998.

[80] 3GPP, TR25.814, Physical layer aspects for evolved UTRA, 2006.

[81] Y. Saito, A. Benjebbour, Y. Kishiyama, and T. Nakamura. System-level Performance Evaluation of Downlink Non-orthogonal Multiple Access (NOMA). Personal Indoor and Mobile Radio Communications (PIMRC), pp. 611-615, 2013.

[82] Kountouris, M., Gesbert, D., Memory-based Opportunistic Multi-user Beamforming, in *Proc. IEEE International Symp. Inf. Theory (ISIT 2005)*.

[83] Tomida, S., Higuchi, K., Non-orthogonal Access with SIC in Cellular Downlink for User Fairness Enhancement, in *Proc. International Symposium on Intelligent Signal Processing and Comm. Systems (ISPACS)*, pp.1-6(2011).

[84] Umehara, J., Kishiyama, Y., Higuchi, K., Enhancing user fairness in non-orthogonal access with successive interference cancellation for cellular downlink, in Proc. International Conference on Communication Systems (ICCS 2012).

[85] Okamoto, Eiji. "An improved proportional fair scheduling in downlink non-orthogonal multiple access system." Vehicular Technology Conference (VTC Fall), 2015 IEEE 82nd. IEEE, 2015.

[86] Mehrjoo, M., Awad, M.K., Dianati, M., Xuemin, S., Design of Fair Weights for Heterogeneous Traffic Scheduling in Multichannel Wireless Networks, IEEE Trans. Commun., 58(10) (2010).

[87] Gueguen, C., Baey, S., Compensated Proportional Fair Scheduling in Multiuser OFDM Wireless Networks, inProc. IEEE International conference on Wireless Mobile Computing, Networking Communication(2008).

[88] Yang, C., Wang, W., Qian, Y., Zhang X., A Weighted Proportional Fair Scheduling to Maximize Best-Effort Service Utility in Multicell Network, in Proc. IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (2008).

[89] Kelly, F.P., Maulloo, A.K., Tan, D.K.H., Rate control for communication networks: shadow prices, proportional fairness and stability, Journal of the Operational Research Society, 49(3), 1998.

[90] http://www.ericsson.com/res/docs/2015/ericsson-mobility-report-june-2015.pdf

[91] M. Carpin, A. Zanella, K. Mahmood, J. Rasool, O. Grndalen, and O.N. terb, "Service differentiation for improved cell capacity in LTE networks", In World of Wireless,

Mobile and Multimedia Networks (WoWMoM), 2015 IEEE 16th International Symposium on a (pp. 1-8). IEEE, 2015

[92] M. Kountouris, and D. Gesbert, "Memory-based opportunistic multi-user beamforming", Information Theory, 2005. ISIT 2005. Proceedings. International Symposium on IEEE, 2005, pp. 1426-1430.

[93] 3GPP, TR25-814 (V7.1.0), Physical Layer Aspects for Evolved UTRA (2006).

[94] Saito, Y., Benjebbour, A., Kishiyama, Y., Nakamura, T., System-level performance evaluation of downlink non-orthogonal multiple access (NOMA), in Proc. IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, pp. 611615 (2013).

[95] Mehrjoo, M., Awad, M.K., Dianati, M., Xuemin, S., Design of Fair Weights for Heterogeneous Traffic Scheduling in Multichannel Wireless Networks, IEEE Trans. Commun., 58(10) (2010).

[96] N. Otao, Y. Kishiyama, and K. Higuchi, Performance of non-orthogonal access with SIC in cellular downlink using proportional fair-based resource allocation, in proc. Int. Symp. on Wireless Commun. Syst.,2012, 476-480.

[97] M.R. Hojeij, J. Farah, C. Abdel Nour, and C. Douillard, New Optimal and Suboptimal Resource Allocation Techniques for Downlink Nonorthogonal Multiple Access, Wireless Pers. Commun., 2015, 1-31.

[98] M.R. Hojeij, J. Farah, C. AbdelNour, and C. Douillard, Resource Allocation in Downlink Non-Orthogonal Multiple Access (NOMA) for Future Radio Access, in proc. 81st IEEE VTC, 2015.

[99] S. Haykin, Communications Systems, 4th ed. Colorado: Wiley, ch.6, pp. 451-460.

[100] J.M. Holtzman "Asymptotic analysis of proportional fair algorithm", *in Proc. IEEE PIMRC*, October 2001.

[101] D. Avidor, S. Mukherjee, J. Ling, and C. Papadias, "On some properties of the proportional fair scheduling policy", in *Proc. IEEE PIMRC*, September 2004.

[102] K. Khawam, D. Kofman, and E. Altman, "The Weighted Proportional Fair Scheduler", in *Proc. ACM QShine*, August 2006.

[103] C. Jin-Ghoo, and B. Saewoong, "Cell-Throughput Analysis of the Proportional Fair Scheduler in the Single-Cell Environment", IEEE Trans. Veh Technol, vo1.56, no.2, pp.766-778, March 2007.

[104] L. Anxin, L. Yang, C. Xiaohang, and J. Huiling, "Non-orthogonal multiple access (NOMA) for future downlink radio access for 5G", China Communications, vol. 12, pp. 2837, 2015.

[105] Z. Ding, M. Peng, and V. Poor, "Cooperative Non-Orthogonal Multiple Access in 5G Systems" IEEE Commun. Lett., pp. 1462-1465, vol. 19, no. 8, 2015.

[106] A. Benjebbour, A. Li, Y. Saito, Y. Kishiyama, A. Harada, and T. Nakamura, "Concept and Practical Considerations of Non-orthogonal Multiple Access (NOMA) for Future Radio Access", Intelligent Signal Processing and Communications Systems (ISPACS), pp. 770-774, 2013.

[107] Z. Ding, P. Fan, and V. Poor, "Impact of user pairing on 5G nonorthogonal multiple access downlink transmissions", IEEE Transactions on Vehicular Technology, vol. PP, no. 99, pp. 1, 2015.

[108] Z. Ding, Z. Yang, P. Fan, and V. Poor, "On the performance of nonorthogonal multiple access in 5G systems with randomly deployed users", IEEE Signal Processing Letters, vol. 21, no. 12, pp. 1501-1505, 2014.

[109] Y. Saito, A. Benjebbour, Y. Kishiyama, and T. Nakamura, "System-level Performance Evaluation of Downlink Non-orthogonal Multiple Access (NOMA)", in *Proc. Personal Indoor and Mobile Radio Communications (PIMRC)*, pp. 611-615, 2013.

[110] Sun, Yan, et al. "Optimal Joint Power and Subcarrier Allocation for Full-Duplex Multicarrier Non-Orthogonal Multiple Access Systems." arXiv preprint arXiv:1607.02668 (2016).

[111] Y. Zhou, L. Liu, H. Du, L. Tian, X. Wang, and J. Shi. An Overview on Inter-cell Interference Management in Mobile Cellular Networks: from 2G to 5G. IEEE International Conference on Communication Systems (ICCS), pp. 217-221, 2014.

[112] Y. Lan, A. Benjebbour, A. Li, and A. Harada. Efficient and Dynamic Fractional Frequency Reuse for Downlink Non-orthogonal Multiple Access. Vehicular Technology Conference (VTC Spring), pp. 1-5, 2014.

[113] M. Qian, W. Hardjawana, Y. Li, B. Vucetic, J. Shi, and X. Yang. Inter-cell Interference Coordination Through Adaptive Soft Frequency Reuse in LTE Networks. Wireless Communications and Networking Conference (WCNC), pp. 1525-3511, 2012.

[114] W. Ku, T. Kwon, and C. Shin. Joint Scheduling and Dynamic Power Spectrum Optimization for Wireless Multicell Networks. Information Sciences and Systems (CISS), pp. 1-6, 2010.

[115] B. Lee, D. Park, and H. Seo, 2009. Wireless Communications Resource Management, Singapore: John Wiley Sons.

[116] B. Hambebo, M. Carvalho, and F. Ham. Performance Evaluation of Static Frequency Reuse Techniques for OFDMA Cellular Network. IEEE 11th International Conference on Networking, Sensing and Control (ICNSC), pp. 355-360, 2014.

[117] F. Allamandri et al; Service platform for converged interactive broadband broadcast and cellular wireless; IEEE Transaction on Broadcasting, vol.53, no. 1, page(s): 200211, Mar. 2007.

[118] O. Benali et al., Framework for an evolutionary path toward 4G by means of cooperation of networks; IEEE Commun.Mag., page(s): 82 89, May 2004.

[119] R. Keller, T. Lohmar, R. Tonjes, and Thielecke, Convergence of cellular and broadcast networks from a multi-radio perspective; IEEE Personal Commun Mag., vol. 8, no. 2, page(s): 5156, Apr. 2001.

[120] ETSI, TS. "102 796 V1.4.1 (2016-08)." Hybrid Broadcast Broadband TV (201).

[121] Yamasaki, R.; Takada, M.; Hamazumi, H.; Shibuya, K.; High-speed mobile reception of HDTV in ISDB-T with directional diversity reception method using adaptive array antenna; IEEE Radio and Wireless Symposium, 2009.

[122] J. Choi,"Minimum power multicast beamforming with superposition coding for multiresolution broadcast and application to NOMA systems", IEEE trans. on commun., vol.63, No.3, March 2015

[123] Z. Ding, Z. Zhao, M. Peng, and V. Poor, "On the spectral efficiency and security enhancements of NOMA assisted multicast-unicast streaming", IEEE trans. on commun., vol.65, No.7, July 2017

[124] Liu, S., Li, H., Ru, G., Lin, W., Liu, L., Yi, Y. (2011, September). Capacity of Multicarrier Multilayer Broadcast and Unicast Hybrid Cellular System with Independent Channel Coding over Subcarriers. In 2011 IEEE Vehicular Technology Conference (VTC Fall).

[125] Mun, C. Transmit-antenna selection for spatial multiplexing with ordered successive interference cancellation; Communications, IEEE Transactions on, vol.54, issue.3, page(s):423-429, 2006.

[126] J. Farah and F. Marx. Greedy algorithms for spectrum management in ofdm cognitive systems - applications to video streaming and wireless sensor networks. IntechOpen, November 2008.

[127] J. Farah, E. Sfeir, C. Abdel Nour, and C. Douillard, "New resource allocation techniques for base station power reduction in orthogonal and non-orthogonal multiplexing systems." In Communications Workshops (ICC Workshops), 2017 IEEE International Conference on, pp. 618-624. IEEE, 2017.

[128] D. Kim, F. Khan, C. Van Rensburg, Z. Pi and S. Yoon, "Superposition of broadcast and unicast in wireless cellular systems", IEEE Communications Magazine, 46(7), 2008.

[129] http://arxiv.org/abs/1710.10857

UNIVERSITE BRETAGNE LOIRE / MATHSTIC

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

Titre : Techniques d'allocation de ressources pour les systèmes à accès multiple non orthogonal

Mots clés : NOMA, Allocation de ressource, Proportional fairness, Allocation de puissance, Scheduling

Résumé : Avec l'émergence rapide des applications Internet, il est prévu que le trafic mobile mondial augmente de huit fois entre fin 2018 et 2022. En même temps, les futurs systèmes de communication se devront aussi d'améliorer l'efficacité spectrale des transmissions, le temps de latence et l'équité entre utilisateurs. À cette fin, une technique d'accès multiple non orthogonal (NOMA) a été récemment proposée comme un candidat prometteur pour les futurs accès radio. La technique NOMA est basée sur un nouveau domaine de multiplexage, le domaine des puissances. Elle permet la cohabitation de deux ou plusieurs utilisateurs par sous-porteuse ou sous-bande de fréquence. Cette thèse aborde plusieurs problèmes liés à l'allocation de ressources basée sur NOMA afin d'améliorer les performances du réseau en termes d'efficacité spectrale, de débit et/ou d'équité entre utilisateurs. Dans ce sens, des solutions théoriques et algorithmiques sont proposées et des résultats numériques sont obtenus afin de valider les solutions et de vérifier la capacité des algorithmes proposés à atteindre des performances optimales ou sous-optimales. Après une étude bibliographique des différentes techniques d'allocation de ressources présentée dans le premier chapitre, on propose dans le deuxième chapitre plusieurs stratégies d'allocation de ressource où une réduction de la bande utilisée par les utilisateurs est ciblée. Les résultats de simulation montrent que les stratégies proposées améliorent à la fois l'efficacité spectrale et le débit total des utilisateurs par rapport aux systèmes basés uniquement sur des techniques d'accès orthogonales. Quant au troisième chapitre, il étudie la performance du *Proportional Fairness* (PF) *Scheduler* tout en considérant que la bande passante est disponible en totalité. Dans ce sens, plusieurs améliorations basées sur le PF sont proposées, qui offrent au système NOMA des avantages en termes de débit, d'équité entre utilisateurs et de qualité de service. Dans le quatrième chapitre, nous proposons plusieurs techniques d'allocation de ressources qui donnent aux utilisateurs la possibilité de favoriser le débit par rapport à l'équité entre utilisateurs et *vice versa*. Dans le dernier chapitre, différentes techniques permettant une transmission hybride *broadcast/broadband* sur la même bande de fréquence sont proposées et comparées à l'état de l'art.

Title : Resource allocation techniques for non-orthogonal multiple systems

Keywords : NOMA, Ressource allocation, Proportional fairness, Power allocation, Scheduling

Abstract : With the proliferation of Internet applications, between the end of 2016 and 2022, total mobile traffic is expected to increase by 8 times. At the same time, communications networks are required to further enhance system efficiency, latency, and user fairness. To this end, non-orthogonal multiple access (NOMA) has recently emerged as a promising candidate for future radio access. By exploiting an additional multiplexing domain, the power domain, NOMA allows the cohabitation of two or more users per subcarrier, based on the principle of signal superposition. This dissertation addresses several radio resource allocation problems in mobile communication systems, in order to improve network performance in terms of spectral efficiency, throughput, or fairness. Theoretical analysis and algorithmic solutions are derived. Numerical results are obtained to validate our theoretical findings and demonstrate the algorithms ability of attaining optimal or sub-optimal solutions. To this direction, the second chapter of this thesis investigates several new strategies for the allocation of radio resources (bandwidth and transmission power) using NOMA principle, where the minimization of the total amount of used bandwidth is targeted.

Extensive simulation results show that the proposed strategies for resource allocation can improve both the spectral efficiency and the cell-edge user throughput, especially when compared to schemes employing only orthogonal signaling. A context where the total bandwidth is available has also been studied, in the 3rd chapter where we investigate the performance of the proportional fairness (PF) scheduler, and we propose modifications to it, at the level of user scheduling and power allocation that show to improve the system capacity, user fairness and QoS. In the 4*th* chapter, we proposed new pairing metrics that allow to favor the fairness at the expense of the throughput and vice versa. The proposed metrics show enhancements at the level of system capacity, user fairness, and computational complexity. Different techniques that allow a hybrid broadcast/multicast transmission on the same frequency platform are proposed in the last chapter and compared to the state of the art.