UNIVERSITÉ DU QUÉBEC À MONTRÉAL

# DÉCHARGEMENT DES TÂCHES DANS L'INFORMATIQUE EN PÉRIPHÉRIE MOBILE ASSISTÉ PAR LES COMMUNICATIONS D2D ET MMWAVE

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN INFORMATIQUE

PAR

CHEIKH IBRAHIMA MBAYE

JANVIER 2024

# REMERCIEMENTS

À la fin de ce projet, je tiens à exprimer ma profonde gratitude et mes sincères remerciements à :

- Mon directeur de recherche, le Professeur Elmahdi Driouch, pour le temps précieux qu'il a consacré à me guider, sa qualité de supervision, ses conseils éclairés, et son soutien financier tout au long de ma formation.

- Les membres du jury pour leurs remarques constructives et leurs suggestions qui ont grandement contribué à l'amélioration de mon travail.

- Mes collègues de laboratoire, avec qui j'ai partagé des moments enrichissants d'échange et de collaboration, en particulier Ababacar Gaye et Ismael Bagayoko.

- M. Mouhamad Dieye, dont les conseils ont été d'une grande importance pour la réussite de ce projet.

- L'ensemble du personnel enseignant et administratif de l'UQAM pour leur appui continu.

- Ma famille, au sens le plus large qu'évoque le terme africain, en particulier Baye Diadia Mbaye, Cheikh Balla, Khadim, Serigne Cheikh, Moustapha, à ma femme Sokhna Diarra Lo et particulièrement à ma maman Soda djitté que son âme repose en paix, pour leur soutien indéfectible et leurs encouragements tout au long de ce parcours.

*Cheikh Ibrahima MBAYE*

TABLE DES MATIÈRES

TABLE DES FIGURES

# LISTE DES TABLEAUX

# LISTE DES ACRONYMES

**ABF** Formation de faisceaux analogique (en anglais analog beamforming)

**AWGN** Bruit blanc, gaussien et additif (en anglais additive white Gaussian noise)

**BS** Station de base (en anglais base station)

**CPU** Processeur central (en anglais central processing unit)

**CSI** Information d'état de canal (en anglais channel state information)

**D2D** Appareil à appareil (en anglais device to device)

**DQN** Apprentissage par renforcement profond (en anglais deep q-network)

**EC** Informatique en périphérie (en anglais edge computing)

**HBF** Formation de faisceaux hybride (en anglais hybrid beamforming)

**HSR** Trains à grande vitesse (en anglais high-speed railway)

**IoT** Internet des objets (en anglais internet of things)

**MEC** Informatique en périphérie mobile (en anglais mobile edge computing)

**PDD** Décomposition duale avec pénalités (en anglais penalty dual decomposition)

**PPTS** Préservation de la confidentialité et fiabilité (en anglais privacy-preserving and trustworthy)

**PSO** Optimisation par essaim particulaire (en anglais particle swarm optimization)

**QdS** Qualité de service (en anglais quality of service)

**RIS** Surface intelligente reconfigurable (en anglais reconfigurable intelligent surface)

**SIC** Annulation successive des interférences (en anglais successive interference cancellation)

**SNR** Rapport signal sur bruit (en anglais signal-to-noise-ratio)

**UE** Équipement utilisateur (en anglais user equipment)

**WP-MEC** Informatique en périphérie mobile avec récolte d'énergie sans fil (en anglais wireless powered MEC)

# RÉSUMÉ

Les appareils mobiles, caractérisés par une capacité de calcul et une autonomie de batterie limitées, peuvent bénéficier grandement du déchargement des tâches. Ce mécanisme leur permet de gérer des applications aux besoins informatiques exigeants tout en réduisant le délai d'exécution et en préservant l'énergie. Dans ce contexte, l'informatique en périphérie mobile (en anglais mobile edge computing MEC), les communications via ondes millimétriques (en anglais millimiter wave mmWave) sont des technologies prometteuses qui se démarquent par leur potentiel à réduire la latence dans les réseaux et à accroître leur capacité.

Dans le cadre de cette étude, nous explorons un système MEC renforcé par des communications d'appareil à appareil (D2D). Non seulement les appareils ont la capacité d'exécuter leurs propres tâches de calcul ou de les décharger vers les serveurs MEC, mais ils peuvent également transférer des tâches vers d'autres appareils à proximité grâce aux connexions D2D.

L'objectif de notre travail est de minimiser la consommation d'énergie des tâches exécutées au sein du système, tout en garantissant le respect des délais. Dans cette optique, nous avons développé un algorithme de déchargement s'appuyant sur l'optimisation par essaim particulaire. Après une série de simulations, l'efficacité de l'algorithme proposé a été démontrée. Afin d'étayer davantage nos résultats, nous confronterons notre solution aux approches gloutonne et aléatoire, mettant ainsi en lumière la pertinence de notre choix méthodologique.

**Mots clés :** déchargement des tâches, informatique en périphérie mobile (MEC), communications d'appareil à appareil (D2D), optimisation par essaim particulaire, ondes millimétriques (mmWave), efficacité énergétique.

# CHAPITRE I

# INTRODUCTION

## 1.1 Mise en contexte

Le déploiement de réseaux 5G à grande échelle, a inauguré une nouvelle gamme d'applications et de services réseaux exigeants en ressources. Ces applications englobent la réalité augmentée, la réalité virtuelle, la conduite autonome, la diffusion vidéo en haute définition, les jeux en ligne en temps réel, ainsi que la vidéosurveillance. Selon un rapport de Cisco (Cisco, 2021), d'ici 2026, près de 82% du trafic de données mondial sera lié aux contenus multimédia, une hausse par rapport aux 73% observés en 2021.

Par ailleurs, le nombre d'équipements utilisateurs (en anglais user equipment UE) est sur une tendance ascendante. D'après les projections, le nombre global d'appareils et de connexions mobiles connaîtra une hausse de 32% en deux ans, passant de 5,1 milliards (soit 66% de la population en 2018) à 5,7 milliards (représentant ainsi 71% de la population en 2023) (Cisco, 2021). Cette croissance conduit à une expansion exponentielle du volume des données échangées. On prévoit que le trafic de données mondial triplera d'ici 2026, pour atteindre un total de 4,5 zettaoctets (Zo) annuellement, par rapport aux 1,5 Zo prévus pour 2022.

Dans cette période marquée par une expansion rapide du trafic de données, les

fournisseurs de réseau font face à des défis grandissants pour assurer une qualité et une expérience utilisateur optimales face à des applications de plus en plus gourmandes. Les attentes vis-à-vis des réseaux 5G sont élevées. Ils devraient offrir des vitesses accrues, s'approchant des gigabits par seconde (Gb/s), tout en minimisant la latence à quelques millisecondes (ms) et en adoptant une architecture souple capable de gérer une densité extrêmement élevée d'UE, soit plusieurs centaines d'appareils au mètre carré.

Pour relever ces défis, il est essentiel de prendre en compte deux problématiques majeures : la rareté du spectre radio disponible et l'augmentation des interférences due à l'accroissement du nombre d'appareils. L'une des ambitions centrales des réseaux de nouvelle génération est d'assurer une expérience utilisateur uniforme. Cependant, maintenir cette continuité pour les utilisateurs situés en lisière de cellule demeure compliqué, même avec l'emploi de systèmes de diffusion sophistiqués (Alliance, 2015a).

### 1.1.1 L'informatique de périphérique mobile

L'informatique mobile de périphérie (en anglais mobile edge computing ou MEC) est un paradigme émergent prometteur qui vise à soutenir les réseaux 5G pour répondre à la demande croissante d'applications sensibles au temps et à forte intensité de calcul. L'utilisation du MEC est nécessaire pour faire face aux défis posés par les limites physiques des appareils, tels que la capacité du processeur, la mémoire et l'autonomie de la batterie.

Le MEC cherche à exploiter les ressources de calcul inutilisées dans les infrastructures périphériques des réseaux sans fil, comme les stations de base et les points d'accès, afin d'exécuter les tâches requises par les utilisateurs du réseau (au lieu d'utiliser des centres de données distants situés dans le nuage informatique). En

utilisant les ressources de calcul à la périphérie du réseau, les UE peuvent transférer leurs tâches à forte intensité de calcul vers les serveurs MEC, économisant ainsi de l'énergie. De plus, grâce à la proximité des serveurs MEC, le temps d'exécution des tâches, incluant le délai de communication, est souvent nettement réduit.

Cependant, il est impératif de gérer la consommation d'énergie et la latence liées au calcul des tâches avec précaution, en tenant compte des contraintes d'autonomie des batteries des équipements utilisateurs qui peuvent s'épuiser, entraînant une dégradation du service. Pour résoudre ce problème, les solutions couramment envisagées sont le rechargement périodique des batteries ou l'utilisation de batteries à plus grande capacité. Néanmoins, recharger les batteries peut se révéler complexe dans certaines situations, en raison de contraintes logistiques ou économiques, et utiliser des batteries plus puissantes n'est pas toujours faisable en raison de contraintes financières ou des dimensions des dispositifs (comme les téléphones intelligents).

Une solution envisagée pour répondre à l'épuisement rapide des batteries consiste à capter de l'énergie ambiante à partir de sources renouvelables (solaire, éolienne, etc.) disposées stratégiquement au sein de la cellule. Toutefois, il est important de souligner que la récolte d'énergie n'est pas toujours constante (par exemple, la production d'énergie solaire dépend de la quantité d'irradiation solaire) et peut connaître des pics de demande à cause du nombre croissant d'appareils dans une cellule. Par conséquent, les appareils mobiles peuvent transférer leurs tâches à l'infrastructure MEC, créant un goulot d'étranglement et augmentant le temps d'exécution des tâches.

De plus, la distance entre les équipements utilisateurs et l'infrastructure MEC n'est pas toujours négligeable et dépend de la distribution des équipements dans la cellule. Cela s'ajoute à la qualité variable du canal de communication entre un

équipement utilisateur et l'infrastructure MEC, en raison de phénomènes tels que l'évanouissement ou le bruit. Ceci peut influencer la latence subie par les requêtes, qui peut dépasser les exigences des applications, malgré le transfert de tâches vers le MEC.

Pour pallier les limitations susmentionnées, deux approches sont actuellement privilégiées :

- le déchargement de tâches en D2D,
- l'ultra-densification des réseaux sans fil grâce aux communications en ondes millimétriques.

Le traitement coopératif centré sur l'utilisateur est une méthode qui encourage les utilisateurs à collaborer pour partager les ressources de calcul et de stockage. Ceci peut être réalisé grâce à la communication directe entre appareils (en anglais device-to-device, D2D) qui leur permet de communiquer entre eux sans passer par une infrastructure centralisée. Cette technique réduit la demande en bande passante sur les réseaux sans fil en minimisant le trafic vers les structures centralisées.

L'ultra-densification des réseaux sans fil via l'utilisation d'ondes millimétriques propose d'accroître drastiquement la densité des cellules. Les ondes millimétriques présentent des avantages notables car elles offrent un débit plus élevé et des latences réduites grâce à des bandes de fréquence plus larges. Elles peuvent aussi être employées pour établir des réseaux ad-hoc assurant la connectivité entre les équipements.

En conclusion, ces stratégies présentent un potentiel significatif pour surmonter les défis liés à la consommation d'énergie et à la latence. Elles permettent aux utilisateurs de collaborer ou de densifier les réseaux afin d'accroître leur capacité.

### 1.1.2 Le traitement coopératif centré sur l'utilisateur

Le traitement coopératif centré sur l'utilisateur est une méthodologie visant à améliorer la répartition de la charge en réduisant la pression exercée sur le réseau d'infrastructure (en anglais core network ou CN) tout en diminuant la latence perçue par les applications. Une illustration concrète de cette approche est la communication appareil à appareil (D2D), qui consiste en une communication directe, sans intervention de la station de base ou d'un module central, entre deux utilisateurs mobiles.

La communication à travers une station de base est généralement adaptée aux services mobiles traditionnels ayant des besoins en termes de débit de données modestes, tels que les appels vocaux ou la messagerie texte, dans lesquels les utilisateurs sont rarement assez proches pour communiquer directement. À l'inverse, les applications et services multimédias récents requérant un débit élevé peut bénéficier d'une communication directe. Dans ce contexte, le D2D permet d'augmenter de manière significative l'efficacité spectrale du réseau, tout en améliorant la vitesse de transmission, la réactivité et l'efficacité énergétique. La communication D2D est particulièrement bénéfique pour les utilisateurs se trouvant en périphérie de la cellule, pour qui la distance à la station de base est considérable ou sont confrontés à une qualité de canal médiocre. Il en résulte une expérience utilisateur homogène et constante pour les utilisateurs de la cellule.

Il est vrai que la communication D2D présente également des inconvénients. En raison de sa nature décentralisée, il n'y a pas de garantie de coopération de la part des dispositifs. Ainsi, un appareil particulier peut refuser de traiter ou de relayer une tâche reçue en fonction de critères tels que le niveau de charge de sa batterie. De plus, une utilisation intensive de la communication D2D peut entraîner des risques d'interférences. Par conséquent, l'absence d'une méthode de coordination

et d'allocation efficace, la communication D2D peut s'avérer instable et peu fiable dans certaines situations.

### 1.1.3 Les réseaux de communications à ondes millimétriques

La nécessité impérieuse de répondre aux exigences de communications à haut débit, en conjonction avec les défis liés à la pénurie de spectre et à la gestion de l'interférence résultant de la prolifération des appareils, a conduit à l'émergence des réseaux de communications à ondes millimétriques (en anglais millimeter wave mmWave). Ces réseaux utilisent la gamme de longueurs d'onde électromagnétiques comprises entre 10 millimètres (30 gigahertz (GHz)) et 1 millimètre (300 GHz), offrant une largeur de bande utilisable d'environ 250 GHz. Les propriétés uniques de ces ondes, notamment une latence très faible et une capacité élevée, en font une solution privilégiée pour les communications sans fil ultra-rapides.

L'utilisation des ondes mmWave, qui se situent dans la bande de fréquences extrêmement élevées, présente de nombreux avantages. Tout d'abord, les antennes doivent être directionnelles pour compenser l'atténuation dans les hautes fréquences, ce qui permet d'obtenir un gain de puissance plus important. En outre, les signaux d'interférence sont fortement atténués lorsqu'ils s'échappent par les lobes latéraux des antennes, rendant les réseaux mmWave particulièrement adaptés pour un déploiement dans des zones denses en termes de population ou d'infrastructures.

Effectivement, l'utilisation de réseaux d'antennes directionnelles sur une bande de fréquences autorisant la transmission de données à haut débit sans interférence notable induite permet d'augmenter rapidement et efficacement la capacité du réseau à travers une approche nommée la densification du réseau. Elle consiste essentiellement à intégrer des cellules supplémentaires dans des zones où la capacité est actuellement limitée et où le trafic existant a le plus besoin d'être déchargé. Il

s'agirait, par exemple, de zones urbaines denses et d'autres zones à forte densité de population.

Pour les réseaux 5G, la densification du réseau à travers le déploiement de petites cellules à faible coût et à faible puissance devrait améliorer les performances globales du réseau en termes d'efficacité énergétique, conséquence de la distance physique réduite entre l'émetteur et le récepteur. En outre, la densification du réseau permet d'améliorer Alliance, 2015b :

- la connectivité du réseau : une zone ayant plus de cellules aura moins de zones mortes ou à mauvaise connectivité ;
- la flexibilité du réseau : l'opérateur peut offrir une couverture lors d'un pic temporaire de trafic ;
- la fiabilité du réseau : les réseaux surchargés peuvent réduire les débits de données ou provoquer des coupures d'appels, alors qu'un réseau dense est plus stable.

Les signaux sans fil peuvent transmettre simultanément des informations et de l'énergie, offrant ainsi aux émetteurs la capacité de transmettre des données à d'autres dispositifs. Cet aspect est particulièrement avantageux pour les appareils sans fil à faible consommation d'énergie.

Cependant, il y a aussi quelques inconvénients pratiques à l'utilisation d'ondes millimétriques. En effet, la portée réduite des communications est un problème majeur, car les signaux à ondes millimétriques sont très sensibles au phénomène d'ombrage (shadowing en anglais) et présentent une forte atténuation liée à la fréquence, ce qui les rend plus susceptibles d'être bloqués. De plus, ils subissent un affaiblissement notable à cause de l'atténuation atmosphérique, parmi d'autres fac-

teurs. Il est donc nécessaire de prendre en compte ces limites lors de la planification et de la mise en place de réseaux mmWave.

### 1.1.4 conclusion

En conclusion, ce chapitre introductif a posé le contexte de la problématique en soulignant la croissance exponentielle du trafic de données, motivée par l'émergence de services exigeants en ressources dans le cadre des réseaux 5G. Les défis liés à la gestion de la rareté du spectre radio, l'augmentation des interférences, et les attentes élevées en termes de débit, latence, et densité d'utilisateurs ont été mis en évidence. Deux approches clés, l'informatique mobile de périphérie (MEC) et les réseaux à ondes millimétriques (mmWave), ont été présentées comme des solutions potentielles, chacune avec ses avantages et ses défis.

## 1.2 Problématique et objectifs

Comme mentionné précédemment, l'augmentation exponentielle du nombre d'utilisateurs, d'applications et de services a entraîné une croissance significative du trafic réseau. Cette situation nécessite des mesures efficaces pour éviter une congestion importante du réseau, qui pourrait avoir des répercussions financières catastrophiques pour les opérateurs. Avec les applications et les services de nouvelle génération qui deviennent de plus en plus exigeants en termes de ressources, l'implémentation de solutions adaptées est primordiale. D'où les attentes élevées en matière financière, opérationnelle et autre envers les réseaux 5G, qui intègrent plusieurs technologies telles que le mmWave, le MEC et les communications D2D. Même si ces approches offrent chacune des avantages distincts, elles sont également associées à des risques et des défis qu'il convient d'évaluer et de surmonter.

Le MEC présente des défis majeurs, notamment la possibilité de goulets d'étran-

glement sur les serveurs de traitement partagés et la gestion de tâches hétérogènes en termes de qualité de service (QoS). L'introduction de D2D dans le MEC offre encore plus d'avantages en réduisant la distance de transmission, mais introduit des contraintes liées à la batterie des appareils. L'équilibrage entre les performances de calcul et la consommation d'énergie devient crucial, tout en tenant compte des limitations matérielles et de la recharge. De plus, une coordination des appareils D2D et des stratégies de gestion de ressources sont essentielles pour éviter les conflits.

Le déploiement de réseaux à ondes millimétriques (mmWave) peut être une solution pour améliorer l'expérience utilisateur et augmenter les débits, mais il pose des défis liés à la portée, aux interférences et à la gestion des ressources radio. Cependant, la complexité de la gestion du réseau augmente avec la densification, nécessitant des approches innovantes pour une allocation des ressources. Il est donc crucial d'analyser minutieusement les différents compromis dans un tel environnement pour garantir une exécution optimale des tâches des utilisateurs, répondant ainsi aux exigences de la nouvelle génération.

L'objectif de ce travail est de présenter une approche basée sur l'optimisation par essaim de particules (en anglais particle swarm optimization) pour orchestrer et minimiser la consommation d'énergie, tout en priorisant les délais de traitement. À cet effet, nous envisageons un environnement réseau comprenant divers éléments, à savoir :

- l'utilisation de la technologie MEC pour offrir des ressources de calcul dédiées au déchargement des tâches des utilisateurs du réseau ;
- le recours à la technologie D2D pour faciliter le déchargement de tâches entre les appareils utilisateurs proches ;

- l'adoption de réseaux à ondes millimétriques pour la densification du réseau.

Dans cette perspective, notre approche vise à aborder simultanément les enjeux relatifs au déchargement de tâches afin de respecter les exigences de QoS(en particulier le délai), tout en minimisant la consommation énergétique.

Notre modèle, qui sera détaillé plus loin, considère un scénario où des appareils utilisateurs peuvent traiter leurs tâches eux-mêmes ou au besoin déléguer l'exécution de leurs tâches en utilisant un appareil D2D ou à travers l'infrastructure MEC suivant deux types de relais : la station de base macro ou une station de base mmWave. En dernier recours, l'exécution d'une tâche peut être rejetée.

## 1.3 Méthodologie

Afin de définir le contexte de notre travail, nous décrivons l'environnement d'étude ainsi que les composants de notre modèle. Ensuite, nous formulons analytiquement les problématiques précédemment décrites sous la forme d'un problème de programmation linéaire en nombres entiers (*Integer Linear Programming - ILP*). Puis, nous montrons à travers une analyse de la complexité algorithmique que le problème résultant est NP-difficile, justifiant la conception d'un heuristique basé sur l'optimisation par essaim particulaire. La performance de notre solution est évaluée à travers des simulations à l'issue desquelles nous fournissons une interprétation des résultats obtenus.

## 1.4 Contribution et organisation du mémoire

En somme, la contribution de ce mémoire consiste à :

1. décrire l'état de l'art pour le déchargement de tâches dans les réseaux mm-

Wave, le MEC et le D2D ;

2. formuler une modélisation mathématique basée sur ILP, tenant compte des contraintes énergétiques, d'interférence et de ressources, pour la maximisation du déchargement des tâches d'utilisateurs ;

3. proposer un nouveau mécanisme de déchargement de tâches s'appuyant sur l'optimisation par essaim particulaire.

Le présent document sera organisé comme suit :

Nous commençons par discuter, au chapitre II, de la littérature couvrant les sujets liés à notre problème.

Ayant posé les bases de notre analyse, nous présentons au chapitre III un modèle du système considéré en préalable à une formulation mathématique soulignant les contraintes et paramètres d'optimisation de notre problème. Par la suite, une analyse de la complexité est faite et dans laquelle nous établissons que notre problème est NP-difficile, justifiant de ce fait, une heuristique que nous détaillons au chapitre IV. Celui-ci est consacré à une description détaillée de l'approche proposée.

Au chapitre V, nous évaluons les performances de notre solution en la comparant avec une approche aléatoire et une approche gloutonne. Dans ce même chapitre, nous analysons les résultats de cette évaluation. Nous fournissons enfin nos conclusions au chapitre VI.

# CHAPITRE II

# REVUE DE LITTÉRATURE

Cette revue de littérature vise à explorer les travaux antérieurs dans ce domaine en se concentrant sur l'utilisation conjointe du D2D, du mmWave et du déchargement vers les serveurs MEC, dans le but précis de réduire la consommation d'énergie. Nous discuterons des approches existantes, mettant en lumière leurs contributions, leurs limitations et les défis non résolus. De plus, nous mettrons en évidence la pertinence et l'originalité de notre propre approche qui se penche spécifiquement sur ce problème tout en explorant la synergie entre le D2D, le mmWave et le déchargement vers les serveurs MEC.

## 2.1 Déchargement des tâches vers le MEC

Nous discutons dans un premier temps, des travaux se focalisant sur le déchargement de tâches, principalement dans les réseaux MEC.

Les auteurs mentionnent dans leur travail (Huang et al., 2020) un scénario qui se concentre sur un système relevant de l'Internet des Objets (IoT) en environnement extérieur, lequel est connecté une infrastructure de calcul à la périphérie du réseau. Au cœur de ce système se trouve un point d'accès qui assume la double responsabilité d'effectuer la transmission d'énergie via des signaux radiofréquences (RF) ainsi que la réception des charges de calcul déléguées par les dispositifs du réseau. De manière

spécifique, les dispositifs suivent une stratégie de déchargement des tâches binaires, comme discuté dans (Mao et al., 2017). Cette stratégie dicte que l'exécution d'une tâche donnée peut être réalisée soit localement sur le dispositif lui-même, soit déchargée vers le serveur MEC en fonction de divers facteurs.

L'article (K. Wang et al., 2020) réalise une analyse approfondie d'une situation de déchargement au sein de laquelle diverses options d'exécution de tâches sont possible. Dans ce cadre, chaque tâche peut être traitée selon une des trois manières suivantes : elle peut être exécutée entièrement de manière locale sur le dispositif d'origine, ou bien elle peut être déchargée de façon totale vers les serveurs MEC, ou encore elle peut faire l'objet d'un déchargement partiel. Ce dernier scénario se matérialise par une division de la tâche initiale en deux sous-tâches distinctes. La première est accomplie localement, exploitant les ressources du dispositif, tandis que la seconde est déléguée au serveurs MEC.

Cependant, il est important de souligner que cette opportunité de déchargement des tâches n'est pas disponible pour l'ensemble des appareils du réseau. La raison en est l'existence de contraintes inhérentes à l'infrastructure MEC, telles que les limitations de bande passante et d'autres ressources essentielles. Cette limitation concerne donc le nombre d'appareils qui peuvent effectivement recourir au déchargement de leurs tâches vers les serveurs MEC. La complexité de l'équilibre entre la demande croissante de déchargement et les contraintes de l'infrastructure MEC constitue un enjeu central de cette étude et de la plupart des efforts de recherches dans ce domaine.

De manière semblable, dans (Wei et al., 2017), les auteurs présentent une approche de collaboration entre les appareils mobiles et l'infrastructure MEC pour mieux gérer le déchargement de calcul. Contrairement à une simple délocalisation complète du traitement, cette approche procède d'abord à la découpe des programmes en

modules distincts. Certains de ces modules sont destinés à être exécutés localement sur le dispositif mobile, particulièrement ceux impliqués dans les interactions avec les périphériques. En revanche, pour les autres modules, une évaluation est entreprise afin de déterminer si leur déchargement vers l'infrastructure MEC serait bénéfique. Cette approche aboutit à une utilisation plus efficace et ciblée des ressources disponibles pour le traitement des tâches.

Dans (N. Zhang et al., 2019), les chercheurs explorent une stratégie avancée de déchargement de calcul, ainsi qu'à la mise en cache efficace des données requises de manière fréquente. L'objectif central poursuivi est double : réduire les délais de latence associés à l'exécution de tâches, tout en optimisant la consommation d'énergie. Dans l'ensemble, l'article présente une analyse détaillée et approfondie de toutes les facettes du déchargement de calcul, en considérant le rôle central des serveurs MEC ainsi que la gestion astucieuse de la mise en cache. Cette approche vise à un déchargement optimal des tâches en offrant des délais de latence réduits et une utilisation efficace de l'énergie.

Dans (H. Wang et al., 2017), les auteurs explorent en détail un élément central du domaine du déchargement de calcul, soit la détermination du moment propice pour cette opération, en capitalisant pleinement sur les avantages inhérents à l'infrastructure MEC. Cette démarche vise, dans l'ensemble, à atteindre un double objectif : réduire les délais d'exécution des tâches et minimiser la consommation énergétique.
L'étude dévoile que l'introduction du MEC en combinaison avec le cloud permet de sensiblement atténuer les délais de latence inhérents au processus de déchargement. Cette approche hybride permet d'exploiter les avantages distinctifs de chaque modèle, ce qui contribue à l'amélioration globale des performances du déchargement de calcul.
Cette étude formule le problème de Processus de Décision de Markov (PDM). Ce

modèle permet de représenter et d'analyser les scénarios de décision séquentiels, notamment en ce qui concerne la planification du déchargement de calcul.

L'article (Aliyu et al., 2017) offre une perspective intéressante en utilisant la théorie des jeux pour aborder la question de l'allocation de ressources au sein des réseaux. Les auteurs présentent une stratégie d'allocation adaptable des ressources de calcul dans le serveur MEC. Cette approche vise à trouver un équilibre entre les exigences variées en matière de qualité de service imposées par les applications de l'IdO, tout en gérant les défis associés à la latence et au coût de traitement.
Cet algorithme minimise le coût de traitement des tâches et la latence. Cette double optimisation reflète la nécessité de jongler entre la réduction des coûts et la satisfaction des contraintes de temps dans les environnements de l'IdO, où des performances optimales doivent être maintenues tout en économisant les ressources.

Une autre facette importante de la problématique de déchargement réside dans la planification en présence de plusieurs serveurs MEC. Elle est relativement moins explorée en raison de sa complexité. Cependant, les auteurs de (Zhu et al., 2020) sont penchés sur cette question en analysant la planification du déchargement de calculs avec plusieurs serveurs de périphérie dans les réseaux avec récolte d'énergie sans fil (en anglais wireless powered MEC ou WP-MEC). Le problème d'optimisation qui se pose ici est comme NP-difficile. Les auteurs ont donc proposé un algorithme d'approximation centralisé pour améliorer les performances globales du système. Cet algorithme offre une approche plus gérable pour traiter la complexité inhérente à la planification avec plusieurs serveurs MEC.
Cette catégorie de recherche revêt une importance particulière car elle reflète la réalité des réseaux WP-MEC, où plusieurs serveurs de périphérie sont déployés pour optimiser le traitement des tâches.

## 2.2 Le déchargement de tâches en D2D

Dans un réseau D2D, les utilisateurs qui se trouvent à proximité les uns des autres communiquent directement sans avoir besoin d'un passage par la BS, et ce dans la même bande spectrale ce qui permet d'améliorer la capacité du réseau. Dans le contexte de l'informatique en périphérie, ce type de communication s'avère d'une grande utilité puisqu'il permet un déchargement plus rapide et des opportunités d'exécution des tâches encore plus variés. Pour illustration, les travaux (Li et al., 2014 ; Hu et Cao, 2017) proposent d'utiliser des stratégies de déchargement de tâches employant un réseau D2D en alternative aux serveurs MEC afin de réduire la latence pour les tâches ayant des exigences strictes en QoS . Dans la même veine, les auteurs de (Y. He et al., 2019) proposent une nouvelle technique D2D-MEC visant à maximiser le nombre d'appareils pris en charge par les réseaux cellulaires via le déchargement coopératif.

Dans (Baek et al., 2020), les auteurs ont étudié la problématique du déchargement de tâches dans un réseau D2D. Lorsque les tâches sont fréquemment déchargées vers un périphérique mobile spécifique, cela peut soulever des préoccupations en matière de confidentialité pour le propriétaire de la tâche, car cela peut potentiellement révéler des informations sensibles. De plus, la fiabilité des résultats de déchargement peut être compromise en raison de la possible malhonnêteté des périphériques mobiles participants. Pour résoudre ces problèmes, l'article propose une solution de déchargement D2D qui respecte la vie privé ( en anglais Privacy-Preserving and Trustworthy D2D Offloading Scheme PPTS), qui se décompose en deux étapes essentielles.

Dans la première étape, elle vise à préserver la vie privée du propriétaire de la tâche tout en obtenant des résultats de déchargement fiables avec un minimum de délai d'exécution. Pour ce faire, le propriétaire de la tâche sélectionne plusieurs

destinataires de déchargement et décharge les tâches de manière redondante vers ces destinataires. De plus, pour renforcer la confiance dans le processus, le propriétaire de la tâche désigne un autre périphérique mobile comme vérificateur.

La deuxième étape repose sur l'utilisation de la technologie blockchain pour garantir que les résultats de déchargement sont correctement traités. La blockchain offre un mécanisme transparent et inviolable pour vérifier l'intégrité des opérations de déchargement.

Dans (Sun et al., 2019), les auteurs se penchent sur un mécanisme de déchargement optimal pour les réseaux de calcul en périphérie multi-accès avec l'aide de la technologie de communication D2D. En particulier, en exploitant la présence de périphériques inactifs dans les réseaux, les tâches de communication peuvent être déchargées vers ces périphériques pour exploiter pleinement leurs ressources de calcul. Afin d'encourager les périphériques inactifs à participer au déchargement collaboratif, un mécanisme d'incitation combinant une stratégie de tarification et une stratégie d'attribution de ressources de calcul est proposé. Une méthode basée sur le jeu de Stackelberg est ensuite utilisée pour atteindre l'équilibre entre les prix et les ressources. Ensuite, pour maximiser les profits de calcul des éditeurs de tâches, le problème d'attribution de tâches de collaboration D2D est formulé, et une méthode d'attribution de tâches optimale basée sur la correspondance bipartite est présentée.

## 2.3 Le déchargement de tâches en mmWave

Les communications à ondes millimétriques (mmWave) suscitent un vif intérêt en tant que candidat potentiel pour les nouvelles fréquences dans les réseaux de prochaine génération, grâce à leurs débits élevés.

Dans cette optique les auteurs (Zhao et al., 2020), ont développé un algorithme

conjoint de formation de faisceaux hybrides et d'allocation de ressources pour un réseau MEC qui utilise des communications mmWave. Plus précisément, ils optimisent conjointement les vecteurs de formation de faisceaux analogiques chez les utilisateurs, les matrices de formation de faisceaux analogiques et numériques à la station de base (BS), les taux de déchargement des tâches de calcul, et l'allocation de ressources au serveur MEC afin de minimiser le délai maximal, tout en respectant le temps de communication et les ressources disponible pour le calcul. Ils ont mis au point un algorithme pour résoudre ce problème complexe et non convexe avec des contraintes interdépendantes, en utilisant la technique de décomposition duale avec pénalités (en anglais penalty dual decomposition PDD).

Dans (Liu et al., 2022), les auteurs explorent les avancées en matière de communications sans fil, notamment grâce aux technologies MEC, mmWave. L'idée clé est de décomposer une tâche en deux parties : la première est exécutée localement par l'utilisateur A, tandis que la seconde est traitée par le serveur MEC via BS. Les résultats sont ensuite partagés avec l'utilisateur B via un lien D2D et un lien BS-B. Pour optimiser cette répartition de calculs, des antennes multiples sont utilisées, avec une formation de faisceau hybride. L'étude propose un nouvel algorithme qui minimise la latence système tout en réduisant la signalisation nécessaire. Il se base sur des matrices de formation de faisceaux analogiques mises à jour en fonction de l'information d'état de canal (en anglais channel state information CSI) qui varie au fil du temps.

Les auteurs de (Ghosh et al., 2014) ont plaidé pour l'utilisation des ondes millimétriques, spécifiquement dans les bandes de fréquences recommandées pour l'accès amélioré en zone locale de la 5G. L'application de ces fréquences plus élevées offre une largeur de bande considérable, permettant au système d'atteindre des débits de données de pointe élevés et des débits de données périphériques significatifs. Cette approche souligne le potentiel des bandes mmWave pour renforcer les performances

et la connectivité dans le contexte de la 5G, en ouvrant la voie à des expériences utilisateur plus rapides et plus fluides.

Dans (Linqian et al., 2022) les auteurs conçoivent un système de communication basé sur des communications mmWave pour les trains à grande vitesse (en anglais high-speed railway HSR). Dans ce scénario, les tâches des utilisateurs peuvent être partiellement déchargées soit vers la BS à côté des rails, soit vers les relais mobiles déployés sur le toit du train. Les RM fonctionnent en mode full-duplex (FD) pour maximiser l'utilisation du spectre. L'objectif central est de minimiser la latence moyenne de traitement des tâches de tous les utilisateurs, tout en respectant les contraintes de consommation d'énergie imposées aux appareils locaux et aux RM.

Pour résoudre ce défi, une solution conjointe d'allocation de ressources et de déchargement de calcul est proposée. Elle comprend deux composantes clés : un algorithme d'allocation de ressources et de déchargement de calcul ainsi qu'un algorithme de contrainte d'énergie (RM). L'algorithme proposé repose sur la théorie des jeux et se décline en deux sous-problèmes : la segmentation des données, l'association des utilisateurs et l'allocation des sous-canaux.

Dans (Rajasekaran et al., 2020) les auteurs proposent un système combinant les technologies mmWave et NOMA (en anglais non-orthogonal multiple access) prenant en compte les capacités de traitement du signal des utilisateurs finaux. L'implémentation de NOMA en liaison descendante nécessite une annulation successive des interférences ( en anglais successive interference cancellation SIC) aux terminaux des utilisateurs, ce qui augmente la complexité. Les auteurs envisagent un système où les utilisateurs rapportent leur capacité de décodage SIC à la BS. Les auteurs examinent le problème de maximisation du débit, en le décomposant en un problème d'ordonnancement des utilisateurs et d'allocation de puissance.

Dans (Yu et Jiali, 2022), les auteurs présentent des schémas d'allocation de

ressources en calcul pour un système de calcul en périphérie mobile à ondes millimétriques (mmWave-MEC) avec l'aide d'une surface intelligente reconfigurable ( en anglais reconfigurable intelligent surface RIS) qui assiste la communication montante des utilisateurs vers la station de base (BS). À travers une analyse théorique, ils dérivent le taux réalisable et l'efficacité de calcul. Ensuite, ils formulent le problème d'optimisation pour maximiser l'efficacité de calcul tout en respectant les contraintes de consommation d'énergie maximale et de fréquence CPU locale. Cela inclut la conception conjointe de faisceaux hybrides à la BS, du façonnage passif de faisceaux à la RIS, ainsi que de l'allocation locale de ressources pour chaque utilisateur. Les auteurs proposent un algorithme itératif basé sur la méthode de descente de coordonnées en bloc inexacte pour obtenir le schéma d'allocation de ressources optimal.

Dans (X. Yu et al., 2023), les auteurs ont étudié les communications à ondes millimétriques et le NOMA dans un réseau MEC pour améliorer les performances du déchargement des tâches. L'objectif est d'améliorer l'efficacité de calcul en assurant l'équité entre les utilisateurs. Ils ont examiné l'optimisation de l'efficacité de calcul , en prenant en compte à la fois les architectures de formation de faisceaux analogique (en anglais analog beamforming ABF) et hybride (en anglais hybrid beamforming HBF) dans le cadre du mode de déchargement partiel.

## 2.4 Discussion

Dans le domaine de la communication D2D et du MEC, plusieurs travaux de recherche ont déjà exploré des approches visant à optimiser différents aspects tels que la consommation d'énergie, le délai ou la gestion des ressources. Cependant, malgré les avancées significatives réalisées, ces travaux présentent encore certains problèmes.

Plusieurs travaux sont concentrés principalement sur la minimisation de la consommation d'énergie, en négligeant potentiellement d'autres paramètres cruciaux tels que la maximisation du nombre de tâches traitées dans un délai donné. D'autres travaux privilégient la réduction du délai, mais au détriment de la consommation d'énergie, sans parvenir à un équilibre optimal entre ces deux aspects essentiels.

De plus, l'exploitation conjointe du D2D et de la technologie millimétrique pour la transmission de données dans le contexte du MEC représente une avenue de recherche prometteuse, mais qui n'a pas encore été explorée de manière approfondie. Les défis techniques liés à la propagation des signaux mmWave, à leur portée limitée et à leurs caractéristiques spécifiques exigent une attention particulière.

C'est dans ce contexte que notre approche de déchargement de tâches dans le serveur MEC en utilisant D2D et mmWave prend tout son sens. Notre objectif est de résoudre de manière équilibrée la minimisation de la consommation d'énergie et la maximisation du nombre de tâches traitées. Nous considérons cette approche comme une contribution importante à la recherche actuelle.

CHAPITRE III

# ENVIRONNEMENT D'ÉTUDE ET FORMULATION ANALYTIQUE DU PROBLÈME

Dans ce chapitre, nous proposons d'abord un aperçu général de l'environnement. À la suite, nous présentons une formulation mathématique du problème avant de proposer une analyse de sa complexité.

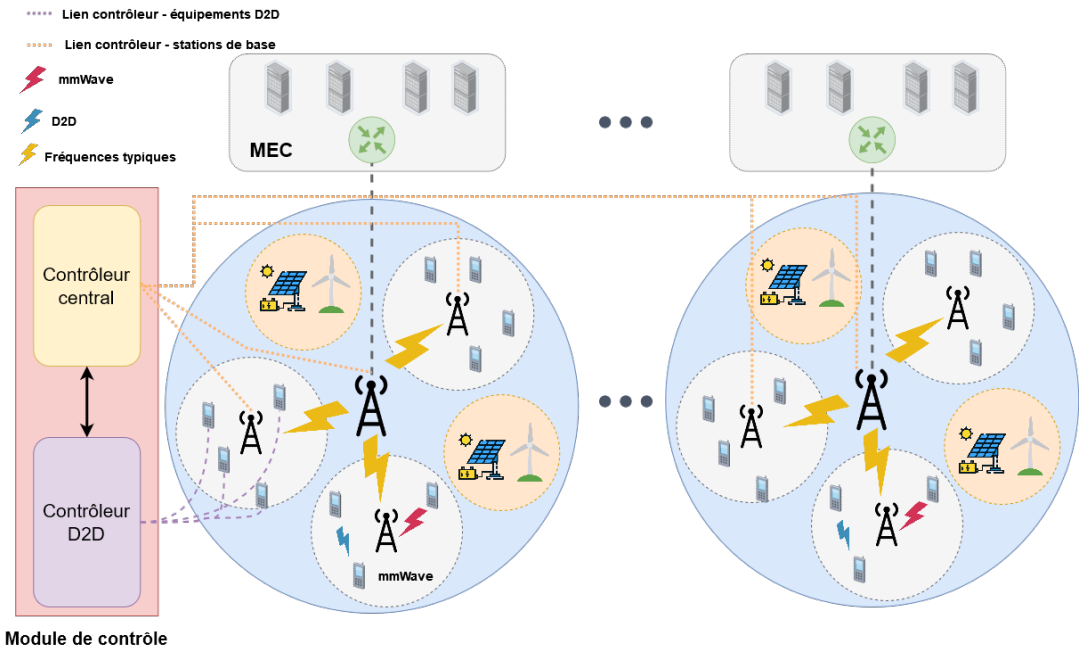


FIGURE 3.1 Modèle du système

## 3.1 Principe de fonctionnement et architecture

Nous considérons le système illustré à la Fig 3.1, partiellement inspiré de (Deng et al., 2017). Le modèle proposé est une architecture hiérarchique visant à permettre une gestion évolutive du réseau, le contrôle de la multi-connectivité, la sélection des relais et des faisceaux, l'allocation des ressources et la gestion des interférences.

Nous considérons le scénario dans lequel des microcellules mmWave sont intégrées au réseau cellulaire macro afin d'améliorer les performances, étendre la couverture, atténuer la surcharge éventuelle de trafic et permettre une expérience utilisateur cohérente pour les utilisateurs ; particulièrement ceux situés aux bords des cellules.

Plus spécifiquement, nous considérons le cas où des utilisateurs désirent exécuter des tâches, tout en respectant les délais de traitement préalablement définis. Par défaut, un appareil utilisateur exécute ses propres tâches lorsqu'il est capable de le faire tout en respectant les exigences de service préalablement définies pour chaque tâche. Si cela s'avère impossible, l'appareil utilisateur demande au module de contrôle (détaillé plus loin) de déterminer un emplacement alternatif permettant de satisfaire les exigences de service et qui tient compte des différentes contraintes d'allocation de ressources.

En résumé, quatre modes d'exécution de tâches sont considérés :

- exécution directe : l'appareil utilisateur est capable d'exécuter ses tâches adéquatement sans assistance externe ;
- exécution sur un appareil utilisateur pair D2D situé à proximité de l'appareil utilisateur propriétaire de la tâche ;
- exécution sur l'infrastructure MEC en utilisant le réseau cellulaire macro ;

- exécution sur l'infrastructure MEC en utilisant les réseaux mmWave comme relais.

La gestion du réseau avec relais mmWave et infrastructure MEC nécessite la prise en compte simultanée de l'association des cellules, la sélection des relais et des faisceaux, l'allocation des ressources et la coordination des interférences. En particulier pour les appareils utilisateurs en bordure de cellule. L'association de cellules dans un tel réseau de multi-connectivité avec relais mmWave devient plus complexe que dans un réseau cellulaire traditionnel.

Effectivement, la sélection des relais et des faisceaux doit tenir compte non seulement des performances de bout en bout souhaitées, mais aussi des problèmes d'interférence. En raison de la variabilité des ressources et des liaisons à planifier, l'allocation des ressources pour des utilisateurs multiples est particulièrement difficile. Pour surmonter les obstacles susmentionnés, notre modèle considère une architecture hiérarchique évolutive qui divise le contrôle du réseau en niveaux distincts en fonction des exigences de latence variables.

L'information sur l'état du réseau, notamment la charge de trafic, la qualité des liaisons et les niveaux d'interférence, est utilisée pour gérer le réseau. En raison des délais et des coûts liés au transport de l'information sur l'état du réseau, il est difficile pour un contrôleur centralisé d'obtenir cette information et de faire des choix de contrôle opportuns dans un réseau à grande échelle.

L'architecture comprend un module de contrôle avec deux sous-modules : un sous-module contrôleur D2D et un sous-module contrôleur central.

Le contrôleur central est l'élément chargé de la prise de décision liée au contrôle de la multi-connectivité (macro et mmWave), la sélection des relais et des faisceaux, l'allocation des ressources et la gestion des interférences. Pour cela, il collecte pério-

diquement en plus des requêtes émises par les appareils utilisateurs, l'information sur l'état du réseau et des différents appareils utilisateurs (p. ex., niveau de batterie restant, localisation, etc.) afin de déterminer les allocations optimales permettant la maximisation du nombre de tâches exécutées tout en respectant les exigences de services prédéfinies.

Plus spécifiquement, le contrôleur central décide dans un premier temps des tâches qui peuvent être traitées de manière adéquate par des appareils D2D situés à proximité de l'appareil utilisateur émetteur de la requête. Dans le cas échéant, le contrôleur central délègue la gestion de l'exécution de la tâche au contrôleur D2D.Cette décision est principalement basée sur les exigences en termes d'énergie et de ressources nécessaires pour une exécution appropriée de la tâche considérée..

Par exemple, l'exécution d'une tâche qui implique une consommation excessive de ressources de calcul peut mener potentiellement un épuisement de l'énergie de l'appareil D2D. Cela pourrait impacter négativement l'exécution future des tâches de l'appareil D2D lui-même. Dans ce contexte, il serait plus approprié de faire exécuter la tâche au niveau du MEC en utilisant soit un réseau mmWave ou le réseau macro.

L'information sur l'état global du réseau est communiquée au contrôleur central par les stations de base. Cette information comprend entre autres :

- le nombre et l'emplacement d'appareils en état actif ;
- le nombre d'appareils actifs en bordure de cellule avec la charge de trafic correspondante ;
- la capacité énergétique des appareils en état actif accompagné de leur charge de trafic ;

- l'emplacement et le nombre de stations de base mmWave disponible.

Le contrôleur central a la charge de déterminer les paramètres de sélection des relais mmWave tels que les règles de sélection des relais et le nombre maximum de liaisons de relais mmWave actives ; les paramètres de coordination des interférences (nombre de ressources, seuils de coordination des interférences) ; et les paramètres d'allocation des ressources.

À noter que les ressources radio du réseau macro sont utilisées pour la signalisation, le contrôle du réseau et la communication de données pour les appareils dont la qualité du signal radio est médiocre, et pour gérer les liaisons mmWave. Les appareils avec un statut actif et bénéficiant d'une qualité de signal acceptable utilisent les liaisons mmWave.

Étant donné les différences en termes d'échelles de temps pour une décision de contrôle ou de types de communication (p. ex., la communication et les mécanismes de gestion en D2D sont décentralisés) selon les types de réseaux et pour limiter la surcharge de contrôle liée à la densification du réseau et des appareils, les responsabilités de gestion sont hiérarchiquement distribuées en plusieurs niveaux granulaires.

À titre d'illustration, l'impact de la mobilité sur la gestion des ressources est différent pour les réseaux mmWave, D2D et macro. Ainsi, à grande échelle, l'effet le plus notable des réseaux mmWave est la variation de l'indice conditions "LOS/NLOS/outage". Si la distance de corrélation LOS est de 10 mètres et que l'appareil se déplace à une vitesse de 30 km/h, la condition LOS peut changer une fois par seconde. En conséquence, suite à un changement d'état (p. ex., topologiques, panne, etc.), il est primordial que les décisions de gestion et de contrôle soient prises en quelques dizaines de millisecondes. En résumé, le paradigme de réseaux avec relais mmWave constitue une méthode séduisante pour fournir des services à

haut débit avec une expérience utilisateur cohérente. Il faut noter que le contrôle du réseau devient plus compliqué lorsque le nombre de sauts augmente. Il faut trouver un équilibre entre les performances et les surcharges liées au contrôle.

## 3.2 Formulation du problème

En partant de l'environnement d'étude précédemment décrit, nous formulons mathématiquement le problème de manière à pouvoir analyser les contraintes mais également évaluer la complexité du problème avant de pouvoir proposer une solution proche de l'optimale.

Soient :

- $\mathcal{N} = \{1, 2, \cdots, |N|\}$, l'ensemble des appareils utilisateurs, chacun équipé d'une antenne ;
- $\mathcal{M} = \{1, 2, \cdots, |M|\}$, l'ensemble des serveurs MEC ;
- $BS$, la station de base couvrant une cellule du réseau macro ;
- $\mathcal{Q} = \{1, 2, \cdots, |Q|$, l'ensemble des stations de base mmWave intégrées à la cellule couverte par $BS$ ;

Sans perte de généralité, nous considérons un système de temps discret représenté par l'ensemble $\mathcal{T} = \{1, 2, \cdots, |T|\}$. Supposons $t \in \mathcal{T}$ comme étant l'indice d'une tranche de temps et $\tau$, la durée d'une tranche de temps.

Soit $\mathcal{A}(t)$, l'ensemble des tâches générées à l'instant $t$ dans le système. Pour une tâche $a \in \mathcal{A}(t)$, nous dénotons par :

- $l(a)$, sa taille ;

- $\tau(a)$, son délai limite d'exécution avec $\tau(a) < \tau$ ;
- $\Omega(a, n, t) = \{0, 1\}$, égale à 1 lorsque la tâche $a$ est générée par l'appareil $n$ à l'instant $t$.

| | |
|---|---|
| $\mathcal{N}$ | ensemble des appareils utilisateurs du système. |
| $\mathcal{M}$ | ensemble des serveurs MEC du système. |
| $\mathcal{Q}$ | ensemble des véhicules également des tâches. |
| $\mathcal{A}(t)$ | l'ensemble des tâches générées à l'instant $t$ |
| $\mathcal{I}$ | variable de décision pour le déchargement. |
| $t$ | indice d'une tranche de temps . |
| $T$ | capacité requis pour exécuter la taches $a$ |
| $l(a)$ | taille de la tâche $a$ |
| $x$ | unité de cycles CPU nécessaire pour traiter un bit |
| $X$ | nombre total de cycles CPUs requis pour exécuter une tâche a |
| $\mathcal{X}$ | fréquences CPU correspondant au nombre de cycles CPUs |
| $w$ | quantité de ressources de traitement disponible sur l'appareil $n$ |
| $W$ | quantité de ressources de traitement disponible sur le MEC $m$ |
| $J$ | fonction retournant la localisation d'un élément du réseau |
| $\kappa(\tilde{n})$ | capacité effective de commutation de l'appareil $n$ |
| $\kappa(m)$ | capacité effective de commutation du serveur MEC m |
| $a$ | indice d'une tâche |
| $m$ | indice d'un MEC |
| $E(a,k,t,trans)$ | énergie consommée par la transmission d'une tâche a sur un appareil D2D k |
| $E(a,q,t,trans)$ | énergie consommée par la transmission d'une tâche à travers le mmWave q |
| $D(a,n,q,t,trans)$ | délai de transmission d'une tâche travers la station mmWave q |
| $D(a,k,t,trans)$ | délai de transmission d'une tâche a sur un appareil D2D k |
| $H(c,y,q,t)$ | gain du canal pour un appareil $y$ couvert par $q$ |
| $\Gamma(\theta(c,k,\Delta,t))$ | vecteur de réponse de l'appareil récepteur de la paire D2D $k$ |
| $\Gamma(\varphi(c,k,\Delta,t))$ | vecteur d'orientation de l'appareil émetteur D2D $k$ |
| $P(q)$ | puissance de transmission totale de la station $q$ |
| $D(a,\tilde{n},t,exec)$ | délai d'exécution d'une tâche $a$ sur un appareil $\tilde{n}$ |
| $D(a,m,t,exec)$ | délai d'exécution d'une tâche $a$ sur un serveur $m$ |
| $E(a,\tilde{n},t,exec)$ | énergie d'exécution d'une tâche a sur un appareil $\tilde{n}$ |
| $E(a,m,t,exec)$ | énergie d'exécution d'une tâche a sur un MEC m |

TABLE 3.1 Tableau de notations

| | |
|---|---|
| $g(n, BS, t)$ | le gain de puissance entre $n$ et $BS$ |
| $g(n, BS, trans, t)$ | gain d'antenne à la transmission entre $n$ et $BS$ à l'instant $t$ |
| $g(n, BS, trans, t)$ | gain d'antenne à la réception entre $n$ et $BS$ à l'instant $t$ |
| $SINR(n, BS, t)$ | rapport signal/interférence plus bruit entre n et BS à l'instant t |
| $P(n, BS, t)$ | puissance d'émission de n vers BS |
| $R(n, BS, t)$ | débit de transmission entre n et BS |
| $D(a, BS, t, trans)$ | délai de transmission d'une tâche $a$ la BS |
| $E(a, BS, t, trans)$ | énergie consommée par la transmission d'une tâche $a$ à la BS |
| $\sigma$ | variance du bruit blanc additif gaussien |
| $B$ | nombre de bande de fréquence. |
| $\eta$ | exposant d'affaiblissement de propagation |
| $\gamma_q$ | nombre d'antennes de la station mmWave q |
| $\phi_q$ | nombre de chaînes RF de la station mmWave q |
| $\|C_q\|$ | nombre de faisceaux de la station mmWave q |
| $Y(q, t)$ | portion d'appareils utilisateurs couverts par la station mmWave $q$ |
| $O(c, q, t)$ | signal superposé via le faisceau (cluster) $c$ . |
| $\varrho(c, y, q, t)$ | signal émis par un appareil $y$ couvert par $q$ |
| $P(c, q)$ | puissance de transmission totale du faisceau $c$. |
| $H(c, y, q, t)$ | vecteur de canal pour un appareil $y$ couvert par $q$ |
| $t_p$ | pénalité sur le temps. |
| $\|\Delta(c, y, q)\|$ | nombre de trajets entre la station mmWave $q$ et l'appareil $y$ |
| $\zeta(c, y, q, t)$ | affaiblissement moyen sur le trajet entre $q$ et l'appareil $y$. |
| $\rho(c, y, q, \Delta, t)$ | gain complexe du trajet $\Delta$. |
| $K(c, t)$ | portion de l'ensemble des paires D2D associé à la grappe $c$. |
| $H^t_{c,k}$ | canal entre une paire D2D avec le faisceau $c$ |
| $\gamma(K)$ | nombre d'émetteurs D2D dans $K(c, t)$ |
| $\|\Delta(c, k)\|$ | nombre de trajets entre l'émetteur et le récepteur D2D $k$ |

TABLE 3.2 Tableau de notations suite

Les tableaux 3.1 et 3.2 donne un résumé de toutes les variables que nous utilisons dans ce document.

Soient $n, \tilde{n} \in \mathcal{N}$. deux appareils et $a$ une tâche, tel que (a,n,t)= 1. Les modes d'exécution de tâches sont représentés par les variables binaires suivantes :

- $I(a, n, \tilde{n}, t)$ ={0,1}, dénotant l'exécution de la tâche $a$ sur l'appareil $\tilde{n}$ à l'instant $t$. À noter que $\tilde{n}$ et $n$ peuvent désigner des appareils D2D distincts ou le même appareil.
- $I(a, BS, m, t)$ ={0,1}, dénotant l'exécution de la tâche $a$ sur un serveur MEC $m \in M$ en utilisant la station de base $BS$.
- $I(a, q, m, t)$ ={0,1}, dénotant l'exécution de la tâche $a$ sur un serveur MEC $m \in M$ en utilisant la station de base mmWave $q \in Q$.
- $I(a, \emptyset, t)$ ={0,1}, dénotant que la tâche $a$ ne sera pas exécutée.

Pour chaque tâche, un seul mode d'exécution est autorisé à l'instant $t$. De ce fait, nous imposons la contrainte suivante :

$$\sum_{\substack{n,\tilde{n}\in N \\ m\in M \\ q\in Q}} I(a,n,\tilde{n},t)+I(a,BS,m,t)+I(a,q,m,t)+I(a,\emptyset,t) = 1, \quad \forall\, a \in \mathcal{A}(t),\ t \in T \tag{3.1}$$

Dénotons par $x(\tilde{n}), \tilde{n} \in N$, le nombre de cycles CPU nécessaire pour traiter un bit. Le nombre total de cycles CPUs requis $X(a, \tilde{n}, t)$ [alternativement $X(a, m, t)$] pour exécuter une tâche $a$ sur un appareil $\tilde{n} \in N$ [alternativement $m \in M$] est alors défini par :

$$X(a, \tilde{n}, t) = x(\tilde{n}) \cdot \Omega(a, n, t) \cdot l(a), \quad \forall\, a \in \mathcal{A}(t),\ t \in T \tag{3.2}$$

$$X(a,m,t) = x(m) \cdot \Omega(a,n,t) \cdot l(a), \quad \forall\ a \in \mathcal{A}(t),\ t \in T \tag{3.3}$$

Nous notons les fréquences CPU correspondant au nombre de cycles CPUs $X(a,\tilde{n},t)$ [alternativement $X(a,m,t)$] par la variable $\mathcal{X}(a,\tilde{n},t)$ [alternativement $\mathcal{X}(a,m,t)$]. À noter que les fréquences CPU sont guidés par le réglage dynamique de la tension et de la fréquence(en anglais dynamic voltage and frequency scaling DVFS) (W. Zhang et al., 2013).

Nous supposons qu'il existe une fréquence CPU maximale pour chaque appareil $\tilde{n}$ (de façon correspondante $m$) de telle sorte que :

$$\mathcal{X}(a,\tilde{n},t) \leq \mathcal{X}(\tilde{n},max) \quad \forall\ a \in \mathcal{A}(t),\ t \in T \tag{3.4}$$

$$\mathcal{X}(a,m,t) \leq \mathcal{X}(m,max) \quad \forall\ a \in \mathcal{A}(t),\ t \in T \tag{3.5}$$

Supposons $\tilde{n}, n \in N$, $m \in M$ et $t \in T$. Soient :

- $w(\tilde{n},t)$, la quantité de ressources de traitement (CPU) disponible à l'instant $t$ sur l'appareil $\tilde{n}$.
- $E(\tilde{n},t)$, la quantité d'énergie disponible à l'instant $t$ au niveau de la batterie de l'appareil $\tilde{n}$.
- $w(m,t)$, la quantité de ressources de traitement (CPU) disponible à l'instant $t$ sur le serveur MEC $m$.
- $W(M,t) = \{w(m,t) \quad \forall m \in M\}$ l'ensemble des ressources de traitement disponible sur l'ensemble des serveurs MEC
- $j(\cdot,t)$, une fonction retournant la localisation d'un élément du réseau à l'instant $t$, avec $\cdot$ pouvant être un appareil $n \in N$ ou une station de base macro ou mmWave $q \in Q$.

- $w(a, \tilde{n}, t)$, la quantité de ressources CPU allouée à la tâche $A_n^t(l_n^t, \tau_n^t)$ pour s'exécuter sur l'appareil $\tilde{n}$.
- $w(a, m, t)$, la quantité de ressources CPU allouée à la tâche $a$ pour s'exécuter sur le serveur MEC $m$.
- $\kappa(\tilde{n}), \kappa(m) > 0$, la capacité effective de commutation de l'appareil $\tilde{n}$/serveur MEC $m$ , dépendant de l'architecture du processeur(Burd et Brodersen, 1996).

Un appareil ou un serveur MEC peut traiter plusieurs tâches distinctes simultanément. Nous devons tenir compte de la capacité totale disponible sur l'appareil ou le serveur MEC :

$$\sum_{a \in \mathcal{A}^t} I(a, n, \tilde{n}, t) \cdot w(a, \tilde{n}, t) \leq w(\tilde{n}, t) \quad \forall\ \tilde{n} \in N,\ t \in T \tag{3.6}$$

$$\sum_{a \in \mathcal{A}^t} I(a, n, \tilde{n}, t) \cdot w(a, m, t) \leq w(m, t) \quad \forall\ m \in M,\ t \in T \tag{3.7}$$

Le délai de traitement $D(a, \tilde{n}, t, exec)$ lié à l'exécution d'une tâche $a$ sur un appareil $\tilde{n}$ qui s'écrit comme suit :

$$D(a, \tilde{n}, t, exec) = I(a, n, \tilde{n}, t) \cdot \frac{X(a, \tilde{n}, t)}{w(a, \tilde{n}, t)} \quad \forall\ a\mathcal{A}(t),\ t \in T \tag{3.8}$$

Le délai de traitement $D(a, m, t, exec)$ lié à l'exécution d'une tâche $a$ sur un serveur MEC $m$ est donné par :

$$D(a, m, t, exec) = I(a, BS, m, t) \cdot \frac{X(a, m, t)}{w(a, m, t)} \quad \forall\ a \in \mathcal{A}(t),\ t \in T \tag{3.9}$$

La consommation énergétique $E(a, \tilde{n}, t, exec)$ induite par le traitement lié à l'exécution d'une tâche $a$ sur un appareil $\tilde{n}$ est :

$$E(a, \tilde{n}, t, exec) = \kappa(\tilde{n}) \cdot w(a, \tilde{n}, t) \cdot X(a, \tilde{n}, t) \quad \forall\ a \in \mathcal{A}(t),\ t \in T \tag{3.10}$$

La consommation énergétique $E(a, m, t, exec)$ induite par le traitement lié à l'exécution d'une tâche $a$ sur un serveur MEC $m$ est :

$$E(a, m, t, exec) = \kappa_m \cdot w(a, m, t) \cdot X(a, m, t) \quad \forall\ a \in \mathcal{A}(t),\ t \in T \tag{3.11}$$

À l'exception du traitement de tâches sur l'appareil émetteur de la tâche lui-même (c.-à.d, $\tilde{n} = n$), tous les autres modes d'exécution induisent des délais liés au temps de transfert ainsi qu'une consommation énergétique correspondante que nous détaillons ci-dessous. .

### 3.2.1 Modèle de communication

Tel qu'illustré à la Fig 3.1, nous considérons un réseau cellulaire macro intégrant à la fois des réseaux mmWave et des communications D2D. une station de base macro est placée au centre de chaque cellule et est connectée aux stations de bases mmWave. Ce dernier est connecté à l'infrastructure MEC. Il convient de noter, qu'afin de simplifier la formulation sans perte de généralité, nous supposons que le délai de transmission entre la $BS$ et les serveurs MEC $m \in M$ est négligeable.

Nous considérions, en outre, trois modes d'exécution de tâches impliquant une transmission réseau :

1. Exécution sur l'infrastructure MEC en utilisant le réseau cellulaire macro :
   (a) $I(a, BS, m, t) = 1$.
   (b) L'appareil $n$ transmet dans un premier temps la tâche $a$ à la $BS$, qui redirige la tâche aux serveurs MEC.
2. Exécution sur un appareil utilisateur pair D2D situé à proximité de l'appareil utilisateur propriétaire de la tâche ;

(a) $I(a, \tilde{n}, t) = 1$ avec $(n, \tilde{n})$, deux appareils formant une paire D2D.

(b) L'appareil $n$ transmet la tâche $a$ à l'appareil $\tilde{n}$ qui exécute la tâche.

3. Exécution sur l'infrastructure MEC en utilisant les réseaux mmWave comme relais.

(a) $I(a, q, m, t) = 1$

(b) L'appareil $n$ transmet la tâche $a$ à la station de base mmWave $q \in Q$, qui redirige la tâche vers $BS$ pour une exécution sur les serveurs MEC.

Ceci implique les délais de transmission suivants : (1) entre $n$ et $BS$ ; (2) entre un pair D2D $n, \tilde{n}$ ; (3) entre $n$ et $q$, puis entre $q$ et $BS$.

#### 3.2.1.1 Modèle de communication du réseau macro

Soit $g(n, BS, t)$, le gain de puissance du canal entre $n$ et la $BS$, modélisé dans ce travail selon l'équation de transmission de Friis (Mudumbai et al., 2009). Il est défini comme suit :

$$g(n, BS, t) = \frac{g(n, BS, trans, t) \cdot g(n, BS, rcpt, t) \cdot \alpha(n, BS)^2}{16\pi^2 (\frac{\delta(n,BS)}{\delta_0})^\eta} \quad \forall\ n \in N,\ t \in T \tag{3.12}$$

où $g(n, BS, trans, t)$ et $g(n, BS, rcpt, t)$ représente respectivement le gain d'antenne à la transmission et le gain d'antenne à la réception pour une communication entre $n$ et la $BS$ à l'instant $t$. La variable $\alpha(n, BS)$ correspond à la longueur d'onde utilisée. Les variables $\delta(n, BS)$ et $\delta_0$ correspondent respectivement à la distance entre $n$ et la $BS$ et à la distance de référence du champ lointain (en anglais, *far field reference distance*). La variable $\eta \in [2, 6]$ représente l'exposant d'affaiblissement de propagation.

Définissons l'ensemble $Z = N \cup Q$. Le rapport signal/interférence plus bruit $SINR(n, BS, t)$ entre $n$ et $BS$ à l'instant $t$ est donc formulé comme suit :

$$SINR(n, BS, t) = \frac{P(n, BS, t) \cdot g(n, BS, rcpt, t)}{\sum_{\substack{z \in Z \\ z \neq n}} P(z, BS, t) \cdot g(z, BS, rcpt, t) + \sigma^2} \quad \forall\ t \in T \tag{3.13}$$

avec $P(n, BS, t)$, la puissance d'émission de $n$ vers $BS$ et $\sigma$, la variance du bruit blanc additif gaussien (AWGN).

La formule de Shannon sur la capacité d'un canal de transmission, nous obtenons alors le débit $R(n, BS, t)$ entre $n$ et la $BS$ comme suit :

$$R(n, BS, t) = \Big(\sum_{\substack{\forall\ m \in M, \\ \forall\ a \in \mathcal{A}^t}} I(a, BS, m, t)\Big)^{-1} \cdot B \log_2(1 + SINR(n, BS, t)) \quad \forall\ t \in T \tag{3.14}$$

défini si $\sum_{\substack{\forall\ m \in M, \\ \forall\ a \in \mathcal{A}^t}} I(a, BS, m, t) \neq 0$ et avec $B$ représente la bande passante et $\sum_{\substack{\forall\ m \in M, \\ \forall\ a \in \mathcal{A}^t}} I(a, BS, m, t)$, dénotant l'ensemble des appareils associés à la $BS$, à l'instant $t$, pour l'exécution de leurs tâches au niveau des serveurs MEC. Ainsi, chaque appareil $n$ associé à $BS$ aura droit à $\big(\sum_{\substack{\forall\ m \in M, \\ \forall\ a \in \mathcal{A}^t}} I(a, BS, m, t)\big)^{-1}$ de la bande de fréquence totale disponible.

Le délai de transmission $D(a, BS, t, trans)$ lié à la transmission d'une requête d'exécution de tâche $a$ à travers la $BS$ est donné par :

$$D(a, BS, t, trans) = I(a, BS, m, t) \cdot \frac{l(a)}{R(n, BS, t)} \quad \forall\ a \in \mathcal{A}(t),\ t \in T \tag{3.15}$$

La consommation énergétique $E(a, BS, t, trans)$ induite par la transmission d'une tâche $a$ à travers la $BS$ est :

$$E(a, BS, t, trans) = P(n, BS, t) \cdot D(a, BS, t, trans) \quad \forall\ a \in \mathcal{A}(t),\ t \in T \tag{3.16}$$

#### 3.2.1.2 Modèle de communication mmWave et D2D

Nous soulignons dans les chapitres précédents qu'en dépit de la largeur de bande substantielle offerte par les ondes mmWave, le nombre de chaînes RF utilisables en pratique demeurait limité, eu égard aux exigences matérielles et énergétiques élevées requises. Effectivement, il existe une limite sur le nombre de chaines RF disponible en fonction du nombre d'appareils utilisateurs pouvant être servis sur un bloc de ressources (Solaiman et al., 2021 ; S. He et al., 2022).

Une approche prônée pour contourner la restriction précédente et permettre l'évolutivité continue du réseau en cas d'ultra-densification consiste à intégrer les réseaux mmWave et D2D dans un réseau cellulaire. À cet effet, nous proposons dans cette sous-section un modèle de communication qui correspond cette combinaison.

Dans chaque microcellule mmWave, gérée par une station de base mmWave $q \in Q$, nous supposons que les appareils sont regroupés en clusters. Chaque cluster est servi par un faisceau avec une seule chaîne RF. Les clusters gérés par une même station de base mmWave $q \in Q$ utilisent des ressources fréquences orthogonales (Solaiman et al., 2021).

Nous notons que notre modèle système suppose la communication D2D sous-jacente (en anglais, *underlay D2D*), une sous-catégorie des communications D2D intrabande(en anglais, *in-band D2D*). Dans une communication D2D sous-jacente, les communications D2D et cellulaire se déroulent sur le même spectre de fréquences sous licence. Cette technique augmente l'efficacité spectrale en utilisant la diversité spatiale, mais nécessite un mécanisme efficace d'allocation de ressources entre les utilisateurs D2D et cellulaires.

Dans ce contexte, afin d'utiliser le spectre disponible, les paires D2D sont appariées aux clusters d'appareils utilisateurs communiquant en mmWave. Par conséquent,

les paires D2D et les appareils utilisateurs d'un même cluster partagent les mêmes ressources spectrales.

La réutilisation extensive des fréquences induit d'importantes interférences. Plus précisément, on observe quatre types d'interférences (SOLAIMAN et al., 2021) :

- l'intrafaisceau (en anglais, *intra-beam*), qui se produit lorsque le signal d'un faisceau d'une station mmWave vers un appareil utilisateur interfère avec d'autres appareils desservis par le même faisceau.
- l'interfaisceau (en anglais, *inter-beam*), qui se produit lorsque le signal d'un faisceau d'une station mmWave vers un appareil utilisateur interfère avec d'autres faisceaux dirigant vers d'autres appareils.
- l'ntracluster (en anglais, *intra-cluster*), qui se produit lorsqu'un émetteur D2D interfère avec d'autres appareils et récepteurs D2D situés dans le même cluster.
- l'intercluster (en anglais, *inter-cluster*), qui se produit lorsqu'un émetteur D2D interfère avec d'autres appareils et récepteurs D2D situés dans d'autres clusters.

Pour la suite, nous considérons que $N = \{U \cup V\}$ avec $U = \{1, 2, \cdots, |U|\}$ désignant l'ensemble des appareils utilisateurs utilisant mmWave et $V = \{1, 2, \cdots, |V|\}$, l'ensemble des paires D2D. Un récepteur D2D comporte une seule antenne de réception et reçoit les signaux d'un réseau d'antennes directionnelles d'un autre émetteur via une chaîne RF indépendante.

Une station de base mmWave $q$ est équipée de $\gamma_q$ d'antennes et de $\phi_q$ chaînes RF qui peuvent générer $|C_q|$ faisceaux hautement directionnels pour desservir un ensemble de clusters $C_q = \{1, 2, \cdots, |C_q|\}$ simultanément, avec $\phi_q < |N| < \gamma_q$. Plus

précisément, le nombre de faisceaux est égal au nombre de chaînes RF ($|C|_q = \phi_q$). Ainsi, l'ensemble $N$ sera partitionné en $|C_q|$ clusters et chaque cluster utilise une chaîne RF distincte.

Soit l'ensemble $Y(q,t) = \{y \in U\}$, les appareils utilisateurs couverts par la station de base $q$. Il est possible d'effectuer l'ordonnancement de $Y(q,t)$ appareils utilisateurs sur le même bloc de ressources temps-fréquence mmWave pour un faisceau $c \in C_q$ donné (SOLAIMAN et al., 2021).

La station de base $q$ envoie un signal superposé $O(c,q,t)$ via le faisceau (cluster) $c$ pour tous les appareils dans $Y^t$ couverts par $q$ :

$$O(c,q,t) = \sum_{y=1}^{Y(q,t)} \sqrt{\beta(c,y,q,t)P(c,q)} \cdot \varrho(c,y,q,t) \quad \forall\ t \in T \tag{3.17}$$

avec $\varrho(c,y,q,t)$ dénote le signal émis par un appareil $y$ couvert par $q$ à travers le faisceau $c$ à l'instant $t$ et $\beta(c,y,q,t)$ représente le coefficient de puissance assigné à l'appareil $y$ desservi par $c$, de sorte que $\sum_{y=1}^{Y(q,t)} \sqrt{\beta(c,y,q,t)P(c,q)} = 1$.

La variable $P(c,q)$ représente la puissance de transmission totale du faisceau $c$. Pour simplicité, nous considérons que la puissance de transmission est équitablement répartie sur les $C$ faisceaux. Ainsi, $P(c,q) = \frac{P(1)}{\gamma(q)}$ avec $P(q)$, la puissance de transmission totale de la station de base $q$.

Pour mmWave, le modèle de canal directionnel typiquement adopté suppose $|\Delta_q|$ diffuseurs/trajets (en anglais, *scatters*) et un réseau linéaire uniforme avec un espacement d'antenne d'une demi-longueur d'onde.

Le vecteur des coefficients du canal entre l'appareil $y$ et la station mmWave correspondante $q$, desservi par un faisceau $c$ est exprimé comme suit (SOLAIMAN et al., 2021) :

$$H(c,y,q,t) = \sqrt{\frac{\gamma(q)}{\zeta(c,y,q,t)}} \sum_{\Delta=1}^{|\Delta(c,y,q)|} \rho(c,y,q,\Delta,t) \cdot \Gamma(\varphi(c,y,q,\Delta,t)) \cdot \Gamma(\theta(c,y,q,\Delta,t)) \quad \forall\, t \in T \tag{3.18}$$

avec $|\Delta(c,y,q)|$ dénotant le nombre de trajets entre la station de base mmWave $q$ et l'appareil $y$ desservi par le faisceau $c$. L'affaiblissement moyen sur le trajet entre $q$ et l'appareil $y$ desservi par le faisceau $c$ est représenté par $\zeta(c,y,q,t)$. La variable $\rho(c,y,q,\Delta,t)$ est le gain complexe du trajet $\Delta$. Les variables $\varphi(c,y,q,\Delta,t)$, $\theta(c,y,q,\Delta,t) \in [0,2\pi]$ constituent respectivement l'angle d'arrivée (en anglais, *angle-of-arrival* ; abrégé *AoA*) et l'angle de départ (en anglais, *angle-of-departure* ; abrégé *AoD*) du trajet $\Delta$. La variable $\Gamma(\varphi(c,y,q,\Delta,t))$ dénote le vecteur d'orientation de l'antenne de la station de base $q$. La variable $\Gamma(\theta(c,y,q,\Delta,t))$ représente le vecteur de réponse de l'appareil $y$ desservi par le faisceau $c$.

Soit l'ensemble $K(c,t)$ une portion de l'ensemble des paires D2D associé au cluster $c$. Similairement à l'équation 3.18, le canal entre une paire D2D avec le faisceau $c$ est exprimé par $H(c,k,t)$ tel que :

$$H(c,k,t) = \sqrt{\frac{\gamma(K)}{\zeta(c,k,t)}} \sum_{\Delta=1}^{|\Delta(c,k)|} \rho(c,k,\Delta,t) \cdot \Gamma(\varphi(c,k,\Delta,t)) \cdot \Gamma(\theta(c,k,\Delta,t)) \quad \forall\, t \in T \tag{3.19}$$

avec $\gamma(K)$, le nombre d'émetteurs D2D dans $K(c,t)$ ; $|\Delta(c,k)|$, le nombre de trajets entre l'émetteur et le récepteur D2D de la paire D2D $k$. L'affaiblissement moyen sur le trajet entre l'émetteur et le récepteur D2D de la paire D2D $k$, desservi par le faisceau $c$ est représenté par $\zeta(c,k,t)$. La variable $\rho(c,k,\Delta,t)$ est le gain complexe du trajet $\Delta$. Les variables $\varphi(c,k,\Delta,t)$, $\theta(c,k,\Delta,t) \in [0,2\pi]$ constituent respectivement l'angle d'arrivée (en anglais, *angle-of-arrival* ; abrégé *AoA*) et l'angle de départ (en anglais, *angle-of-departure* ; abrégé *AoD*) du trajet $\Delta$. La variable $\Gamma(\varphi(c,k,\Delta,t))$ dénote le vecteur d'orientation de l'appareil émetteur de la paire

D2D $k$. La variable $\Gamma(\theta(c,k,\Delta,t))$ représente le vecteur de réponse de l'appareil récepteur de la paire D2D $k$, desservi par le faisceau $c$.

Les appareils utilisateurs reçoivent un signal superposé $\tilde{O}(c,q,t)$ de la station de base mmWave ainsi que d'autres signaux d'interférence incluant les interférences intrafaisceau, interfaisceau, intracluster et intercluster respectivement dénotées $\mathcal{E}(c,\bar{y},q,t), \mathcal{E}(\bar{c},\bar{y},q,t), \mathcal{E}(c,k,t)$ et $\mathcal{E}(\bar{c},k,t)$.

Le signal reçu par l'appareil $y$ couvert par $q$, desservi par le faisceau $c$ est alors défini par :

$$\begin{aligned}\mathcal{H}(c,y,q) = {} & H(c,y,q,t)\nabla\varkappa(c)\sqrt{\beta(c,y,q,t)P(c,q)}\varrho(c,y,q,t) \\ & + \mathcal{E}(c,\bar{y},q,t) + \mathcal{E}(\bar{c},\bar{y},q,t) + \mathcal{E}(c,k,t) + \mathcal{E}(\bar{c},k,t) + \sigma(c,y,q)\end{aligned} \tag{3.20}$$

avec :

$$\mathcal{E}(c,\bar{y},q,t) = H(c,y,q,t)\nabla\varkappa(c)\sum_{\bar{y}=1,\bar{y}\neq y}^{Y(q,t)}\sqrt{\beta(c,\bar{y},q,t)P(c,q)}\varrho(c,\bar{y},q,t) \quad \forall\ t\in T \tag{3.21}$$

$$\mathcal{E}(\bar{c},\bar{y},q,t) = H(c,y,q,t)\nabla\sum_{\bar{c}=1,\bar{g}\neq c}^{C_q}\varkappa_{\bar{c}}\sum_{\bar{y}=1}^{Y_q^t}\sqrt{\beta(\bar{c},\bar{y},q,t)P(\bar{c},q)}\varrho(\bar{c},\bar{y},q,t) \quad \forall\ t\in T \tag{3.22}$$

$$\mathcal{E}(c,k,t) = \sum_{k=1}^{K(c,t)}\sqrt{P(c,k)}(H^t_{(c,k)\to(c,y,q)})\varrho(c,k,t) \quad \forall\ t\in T \tag{3.23}$$

$$\mathcal{E}(\bar{c},k,t) = \sum_{\bar{c}=1,\bar{c}\neq c}^{C_q}\sum_{k=1}^{K(c,t)}\sqrt{P(\bar{c},k)}(H^t_{(\bar{c},k)\to(c,y,q)})\varrho(\bar{c},k,t) \quad \forall\ t\in T \tag{3.24}$$

et considérant $H(c,y,q,t)$ comme étant le gain du canal pour un appareil $y$ couvert par $q$, desservi par un faisceau $c$ ; la variable $\nabla$, de taille $\gamma(q)\times\phi(q)$, est la matrice

de précodage analogique ; la variable $\varkappa(c)$, de taille $\phi(q) \times 1$, est le vecteur de précodage numérique pour le faisceau $c$, la variable $\mathcal{E}(c, \bar{y}, q, t)$ est le brouillage intrafaisceau causé par la station de base ; la variable $\mathcal{E}(\bar{c}, \bar{y}, q, t)$ est le brouillage inter-faisceaux causé par la station de base ; la variable $\mathcal{E}(c, k, t)$ est le brouillage intragrappe causé par l'appareil D2D émetteur utilisant le faisceau $c$ dans la paire D2D $k$, la variable $\mathcal{E}(\bar{c}, k, t)$ est l'interférence intergrappe causée par l'appareil D2D émetteur utilisant le faisceau $c$ dans la paire D2D $k$ et $\sigma(c, y, q)$ est le bruit blanc gaussien additif.

Le signal reçu par l'appareil D2D émetteur utilisant le faisceau $c$ dans la paire D2D $k$ peut être formulée comme suit :

$$\mathcal{H}(c,k) = (H^t_{(c,k)\to(c,k)})\nabla\varkappa(c,k)\sqrt{P(c,k)}\varrho(c,k,t) + \mathcal{E}(c,\bar{k},t) + \mathcal{E}(\bar{c},\bar{k},t) + \sigma(c,k) \tag{3.25}$$

avec

$$\mathcal{E}^t_{c,\bar{k}} = \sum_{\bar{k}=1,\bar{k}\neq k}^{K(c,t)} (H^t_{(c,\bar{k})\to(c,k)})\sqrt{P(c,\bar{k})}\varrho(c,\bar{k},t) \quad \forall\ t \in T \tag{3.26}$$

$$\mathcal{E}^t_{\bar{c},\bar{k}} = \sum_{\bar{c}=1,\bar{c}\neq c}^{C(q)} \sum_{\bar{k}=1,\bar{k}\neq k}^{K(c,t)} (H^t_{(\bar{c},\bar{k})\to(c,k)})\sqrt{P(\bar{c},\bar{k})}\varrho(\bar{c},\bar{k},t) \quad \forall\ t \in T \tag{3.27}$$

et considérant $H^t_{(c,k)\to(c,k)}$ comme étant le gain du canal entre les appareils D2D de la paire $k$ utilisant le faisceau $c$. La variable $\nabla$ est le pré-codage analogique D2D ; la variable $\varkappa(c,k)$ est le pré-codage numérique D2D, La variable $\mathcal{E}(c,\bar{k},t)$ est l'interférence intra-grappe causée à l'appareil D2D émetteur utilisant le faisceau $c$ dans la paire D2D $k$ par les autres appareils D2D utilisant le faisceau $c$, la variable $\sqrt{P(c,k)}$ représente la puissance de transmission et $\sigma_{c,k}$ constitue le bruit blanc gaussien additif.

On formule le SINR respectif de l'appareil $y$ couvert par $q$ et de la paire D2D $k$, desservis par le faisceau $c$ de la manière suivante :

$$SINR(c,y,q,t) = \frac{\sqrt{\beta(c,y,q,t)P(c,q)}\ ||\tilde{H}(c,y,q,t)\varkappa(c)||^2}{\mathcal{E}(c,\bar{y},q,t) + \mathcal{E}(\bar{c},\bar{y},q,t) + \mathcal{E}(c,k,t) + \mathcal{E}(\bar{c},k,t) + \sigma(c,y,q)} \tag{3.28}$$

$$SINR_{c,k} = \frac{\sqrt{P(c,k)}\ ||(\tilde{H}^t_{(c,k)\to(c,k)})\varkappa(c,k)||^2}{\mathcal{E}(c,\bar{k},t) + \mathcal{E}(\bar{c},\bar{k},t) + \sigma(c,k)} \tag{3.29}$$

avec $\tilde{H}$ représentant une approximation du vecteur de gain de canal. Des formulations précédentes, nous dérivons les formules de débit suivantes :

$$R(c,y,q,t) = \log_2(1 + SINR(c,y,q,t)) \tag{3.30}$$

$$R(c,k,t) = \log_2(1 + SINR(c,k,t)) \tag{3.31}$$

### 3.2.2 Délais de transmission pour les communications mmWave et D2D

Le délai de transmission $D(A,n,q,t,trans)$ lié à la transmission d'une requête d'exécution de tâche $a$ à travers la station de base mmWave $q$ est la suivante :

$$D(a,n,q,t,trans) = I(a,q,m,t) \cdot \frac{l(a)}{R(c,y,q,t)} \quad \forall\ a \in \mathcal{A}(t),\ t \in T \tag{3.32}$$

Le délai de transmission $D(a,k,t,trans)$ lié à la transmission d'une requête d'exécution de tâche $a$ sur un appareil D2D $k$ est la suivante :

$$D(a,k,t,trans) = I(a,k,t) \cdot \frac{l(a)}{R(c,k,t)} \quad \forall\ a \in \mathcal{A}(t),\ t \in T \tag{3.33}$$

La consommation énergétique $E(a,q,t,trans)$ induite par la transmission d'une tâche $a$ à travers la station de base mmWave $q$ est la suivante :

$$E(a,q,t,trans) = P(c,q)D(a,q,t,trans) \quad \forall\ a \in \mathcal{A}(t),\ t \in T \tag{3.34}$$

La consommation énergétique $E(a, k, t, trans)$ induite par la transmission d'une tâche $a$ sur un appareil D2D $k$ est la suivante :

$$E(a, k, t, trans) = P(c, k)D(a, k, t, trans) \quad \forall\ a \in \mathcal{A}(t),\ t \in T \tag{3.35}$$

Nous résumons les délais et les consommations énergétiques totaux pour les quatre modes d'exécution de tâches considérés :

- Exécution directe sur l'appareil utilisateur créateur de la tâche :

$$D(a, n, t) = D(a, n, t, exec) \quad \forall\ a \in \mathcal{A}(t),\ t \in T \tag{3.36}$$

$$E(a, n, t) = E(a, n, t, exec) \quad \forall\ a \in \mathcal{A}(t),\ t \in T \tag{3.37}$$

- Exécution directe sur un appareil utilisateur pair D2D situé à proximité de l'appareil utilisateur propriétaire de la tâche :

$$D(a, k, t) = D(a, k, t, exec) + D(a, k, t, trans) \quad \forall\ a \in \mathcal{A}(t),\ t \in T \tag{3.38}$$

$$E(a, k, t) = E(a, k, t, exec) + E(k, t, trans) \quad \forall\ a \in \mathcal{A}(t),\ t \in T \tag{3.39}$$

- Exécution sur l'infrastructure MEC en utilisant le réseau cellulaire macro :

$$D(a, BS, t) = D(a, BS, t, trans) + D(a, m, t, exec) \quad \forall\ a \in \mathcal{A}(t),\ t \in T \tag{3.40}$$

- Exécution sur l'infrastructure MEC en utilisant les réseaux mmWave comme relais :

$$D(a, q, t) = D(a, q, t, trans) + D(a, m, t, exec) \quad \forall\ a \in \mathcal{A}(t),\ t \in T \tag{3.41}$$

### 3.2.3 Contraintes

La problématique détaillée ci-haut comprend plusieurs contraintes. Pour chaque station de base, il existe, à un instant $t$, une puissance maximale à ne pas dépasser :

$$P(t) \leq P(max) \quad t \in T \tag{3.42}$$

Nous devons également nous assurer que le délai maximal prédéfini pour chaque tâche est respecté :

$$\mathcal{D}(a,t) - \tau(n,t) \geq 0 \quad \forall\ a \in \mathcal{A}(t),\ t \in T \tag{3.43}$$

où

$$\begin{aligned} \mathcal{D}(a,t) = I(a,BS,m,t)D(a,BS,t) + I(a,n,t)D(a,n,t) \\ + I(a,k,t)D(a,k,t) + I(a,q,m,t)D(a,q,t) \end{aligned} \tag{3.44}$$

Pour les communications mmWave, le nombre de transmissions concurrentes à travers chaque station de base mmWave et pour les paires D2D est limité par le nombre de faisceaux disponibles :

$$\sum_{a \in \mathcal{A}^t} I(a,k,t) + I(a,q,m,t) \leq |C_q| \quad \forall\ q \in Q,\ t \in T \tag{3.45}$$

L'objectif de notre problème d'optimisation est de :

- maximiser le nombre de tâches qui respecte la contrainte de délai
- minimiser la quantité globale d'énergie utilisée pour l'exécution de ces tâches.

Par conséquent, nous définissons la fonction suivante :

$$\mu(a,t) = \begin{cases} \epsilon, & \text{si } \mathcal{D}(a,t) - \tau(n,t) \geq 0 \\ E(a,t), & \text{sinon,} \end{cases} \quad \forall\ a \in \mathcal{A}(t)t,\ t \in T \tag{3.46}$$

où $\epsilon$, est une pénalité ayant une valeur fixe.

et E(a,t) est l'énergie totale nécessaire pour l'exécution de la tâche $a$ qui est donnée par :

$E(a,t) = I(a,n,t)E(a,n,t) + I(a,k,t)E(a,k,t) + I(a,q,m,t)E(a,q,t) +$
$I(a,BS,m,t)E(a,BS,t)$

D'après sa définition, Il est clair que la fonction assure un équilibre entre le nombre de tâches exécutées et l'énergie nécessaire pour cette exécution.

Ici, $\epsilon$ représente une pénalité fixe. Si l'on choisit un très petit $\epsilon$ , cela revient à minimiser uniquement l'énergie, car les tâches qui dépassent leur délai seront toujours exécutées. Dans ce cas, l'objectif principal est de minimiser la consommation d'énergie. En revanche, en choisissant un plus grand $\epsilon$, l'accent est davantage mis sur le respect des délais, indépendamment de la quantité d'énergie consommée.

La fonction objectif de notre problème est alors définie comme suit :

$$\left( \sum_{a \in \mathcal{A}(t)} \mu(a,t) \right) \tag{3.47}$$

Par conséquent le problème d'optimisation des décisions de déchargement peut s'écrire comme suit

$$\begin{gathered} \min_{\substack{I(a,BS,m,t), I(a,n,t), \\ I(a,k,t), I(a,q,m,t)}} \left( \sum_{a \in \mathcal{A}(t)} \mu(a,t) \right) \\ \text{s. c. (3.44)-(3.48)} \end{gathered} \tag{3.48}$$

## 3.3 Analyse de la complexité

Nous utilisons l'approche de réduction (Kleinberg et Tardos, 2006) afin de montrer la NP-difficulté de notre problème. Ainsi, nous simplifions et reformulons notre problème comme étant un système ayant $n$ processus et $m$ ressources.

À tout moment, chaque processus identifie un ensemble de ressources qu'il souhaite utiliser. Chaque ressource peut être demandée par plusieurs processus simultanément, mais elle ne peut être utilisée que par un seul à la fois. Nous devons alors affecter les ressources aux processus qui en ont besoin (exécution réussie d'une tâche). Dans le cas contraire, le processus est bloqué (non-exécution d'une tâche).

L'objectif du problème d'allocation est de prendre en charge le plus grand nombre possible de processus actifs (maximisation des tâches exécutées). Par conséquent, le problème de réservation des ressources s'énonce comme suit :Est-il possible d'attribuer des ressources aux processus de manière à garantir qu'au moins $k$ processus soient actifs ?, étant donné un ensemble de processus et de ressources, l'ensemble des ressources demandées pour chaque processus, et un nombre $k$ ?

Le problème ci-dessus est NP-complet. Nous pourrions reformuler la complexité du problème de la manière suivante : nous avons une collection de $n$ processus. Chacun des $n$ processus nécessite un sous-ensemble des $m$ ressources pour s'exécuter. L'objectif est alors de déterminer s'il existe un ensemble de $k$ processus tels que leurs temps d'exécution coïncident. Les ressources demandées sont disjointes.

Nous démontrons maintenant que le problème est NP-difficile en décrivant une réduction polynomiale à partir de IndependentSet, un problème d'optimisation prouvé fortement NP-difficile (Tarjan et Trojanowski, 1977). Il reste à démontrer que le problème que nous avons formulé est bien égal au problème original IndependentSet. En supposant que nous avons une instance IndependentSet

sur un graphe $G$ et un entier $k$, nous voulons déterminer si $G$ contient un ensemble indépendant d'au moins $k$ sommets. Les processus correspondent aux nœuds du graphe $G$, tandis que les ressources correspondent à ses arêtes. Un processus (nœud) a besoin d'une ressource (arête) si et seulement si elle est adjacente à l'arête dans $G$. S'il existe $k$ processus dont les ressources nécessaires sont distinctes, alors les $k$ nœuds appartenant à ces processus ne partageront aucun bord, formant ainsi un ensemble indépendant. S'il existe un ensemble indépendant de taille $k$, alors aucun des $k$ nœuds ne partage d'arêtes. Par conséquent, les processus correspondant aux nœuds n'ont pas de ressources partagées, et les $k$ processus disposent de ressources distinctes selon les besoins.

Étant donné la NP-difficulté de notre problème, nous proposons dans le chapitre suivant une solution pour résoudre le problème précédemment formulé.

# CHAPITRE IV

# PRÉSENTATION DE LA SOLUTION PROPOSÉE : PSO

Ce chapitre décrit l'algorithme d'optimisation par essaim de particules **(PSO)** développé pour résoudre le problème formulé dans le chapitre précédent. Avant de présenter l'algorithme proposé, nous offrons un bref aperçu du fonctionnement de cette métaheuristique.

## 4.1 Fonctionnement général du PSO

Tout comme les algorithmes de colonies de fourmis et les algorithmes génétiques, l'optimisation par essaims de particules (en anglais Particle Swarm Optimization, PSO) est un algorithme bio-inspiré. Il repose sur les principes d'auto-organisation permettant à un groupe d'organismes vivants d'agir ensemble de manière complexe à partir de règles simples. Le PSO s'inspire du modèle développé par Craig Reynolds pour simuler le déplacement grégaire de certains animaux (troupeaux de bovins, volées d'oiseaux, etc.). Dans ce modèle, chaque oiseau artificiel, ou essaim, se déplace aléatoirement en suivant trois règles simples :

- La cohésion : les essaims sont attirés vers la position moyenne du groupe ;
- L'alignement : les essaims suivent le même chemin que leurs voisins ;
- La séparation : pour éviter les collisions, les essaims gardent une certaine

distance entre eux.

Concrètement, le plus souvent, les positions et les vitesses des essaims sont représentées comme des vecteurs de nombres en D-dimensions. Les positions et les vitesses initiales sont souvent définies aléatoirement. L'exploration est répétée en mettant à jour la position de chaque essaim puis son vecteur vitesse jusqu'à atteindre une solution satisfaisante. Le niveau de qualité associé à la position de chaque essaim est évalué grâce à une fonction de fitness. On détermine ainsi la meilleure position que chaque essaim a pu rencontrer. Ensuite, le vecteur vitesse de chaque essaim $X$ est mis à jour.

## 4.2 Fonctionnement de l'algorithme proposé : COPSO (computing offloading particle swarm optimization)

### 4.2.1 Définition des caractéristiques de COPSO

La position $\pi^l$ de la particule $l$ est une matrice binaire bidimensionnelle $A \times N$, dont les éléments sont définis par :

$$\pi^l(a,n) = \begin{cases} 1, & \text{si la tâche } a \text{ est traitée par l'appareil } n, \\ 0, & \text{sinon.} \end{cases}$$

**Fonction de fitness** : Pour mesurer la qualité de l'association, $f_t(\pi^l)$ évalue le nombre de tâches associées à la particule $l$.

**Sortie** : *gbest* est une matrice binaire $A \times N$ correspondant à la particule qui maximise la fonction de fitness.

**Critères d'arrêt** : L'algorithme se termine lorsqu'un nombre maximum d'itérations est atteint ou lorsque toutes les tâches ont été exécutées.

**Vitesse des particules** : À chaque itération $k$, chaque particule $l$ se déplace avec une vitesse $v_l^k$ dans l'espace de recherche. Cette vitesse est donnée par une matrice bidimensionnelle de taille $A \times N$, où chaque élément de la matrice est restreint à l'intervalle $[v_{\min}, v_{\max}]$. Après chaque déplacement, la vitesse est mise à jour comme suit :

$$\begin{aligned} v_{k+1}^l(a,n) =& \omega_k v_k^l(a,n) \\ &+ w_1 r_1 \cdot (pbest_k^l(a,n) - \pi_k^l(a,n)) \\ &+ w_2 r_2 \cdot (gbest_k(a,n) - \pi_k^l(a,n)) \\ &\forall\ a \in \mathcal{A}(t),\ n \in N \end{aligned} \tag{4.1}$$

**Composante cognitive (pbest)** : Le terme $w_1 r_1 \cdot (pbest_{lk}(a,n) - \pi_{lk}(a,n))$ représente la composante cognitive de la mise à jour. $w_1$ est le poids de la composante cognitive, $r_1$ est un nombre aléatoire entre 0 et 1, $pbest_{lk}(a,n)$ est la meilleure position individuelle de la particule $l$ à l'itération $k$ pour les indices $a$ et $n$, et $\pi_{lk}(a,n)$ est la position actuelle de la particule pour les mêmes indices.

**Composante sociale (gbest)** : Le terme $w_2 r_2 \cdot (gbest_k(a,n) - \pi_{lk}(a,n))$ représente la composante sociale de la mise à jour. $w_2$ est le poids de la composante sociale, $r_2$ est un autre nombre aléatoire entre 0 et 1, $gbest_k(a,n)$ est la position de la meilleure particule à l'itération $k$ aux indices $a$ et $n$, et $\pi_{lk}(a,n)$ est la position actuelle de la particule pour les mêmes indices.

La figure 4.1 ci-dessous illustre un exemple d'une stratégie de planification des tâches en fonction de plusieurs chemins d'exécution disponibles. Nous définissons par chemin un plan d'exécution des tâches. Chaque tâche $A_1$ jusqu'à $An$ doit être associée à une unique chemin d'exécution, qui peut être soit un appareil spécifique ($D1$ à $Dn$), soit une combinaison de stations mmWave et du MEC par le biais d'un chemin ($V1$ à $Vn$). Alternativement, elle peut être transmise par un chemin reliant une station de base macro vers MEC ($G1$ à $Gn$). Dans cette matrice, une valeur

de 1 signale l'attribution d'une tâche à un chemin spécifique, tandis qu'une valeur de 0 indique que la tâche n'est pas traitée. La figure 4.1 présente un exemple de la représentation.

| chemins | | $D1$ | $D2$ | $D3$ | $D4$ | $V1$ | $V2$ | $V3$ | $G1$ | $G2$ | $\cdot$ | $Gn$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Tâches | $A1$ | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | $A2$ | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | $A3$ | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | $A4$ | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| | $A5$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| | $An$ | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |

FIGURE 4.1 Représentation de la particule

### 4.2.2 Étapes de COPSO

Dans cette section, nous allons aborder l'algorithme COPSO à travers une série d'étapes distinctes, comme décrit ci-dessous :

1. **Initialisation des particules** : Les positions initiales des particules de l'essaim sont définies aléatoirement dans l'espace de recherche selon une distribution uniforme. Cela permet de créer une diversité initiale pour explorer différentes régions de l'espace. Il est aussi possible d'initialiser en utilisant l'algorithme heuristique présent dans la section suivante.

2. **Initialisation des vitesses** : Les vitesses initiales des particules sont également définies de manière aléatoire entre $v_{min}$ et $v_{max}$ . Rappelons que les

vitesses influencent la manière dont les particules se déplacent à travers l'espace de recherche au fil des itérations.

3. **Initialisation de** *pbest* : La meilleure position individuelle (*pbest*) est initialisée pour chaque particule en fonction de sa position initiale. Cela servira de référence pour les mouvements futurs.

4. **Évaluation des Particules** : Chaque particule est évaluée en utilisant la fonction d'objectif $f_t(\pi_0)$, où $\pi_0$ représente la position actuelle de la particule. Cette évaluation mesure la qualité de chaque solution candidate.

5. **Initialisation de** *gbest* : La meilleure position globale (*gbest*) est initialisée par la meilleure position (*pbest*) de toutes les particules. Cela identifie la solution globale la plus prometteuse à ce stade.

6. **Boucle d'Optimisation** : L'algorithme itère tant qu'un critère d'arrêt prédéfini n'est pas atteint. À chaque itération, les particules mettent à jour leur vitesse en fonction des composantes cognitives et sociales, explorent de nouvelles positions, respectent les contraintes, mesurent leur valeur de fitness, et mettent à jour leurs meilleures positions individuelles et globales.

    - **Mise à Jour de la vitesse et de la position** : Les particules mettent à jour leur vitesse en fonction de l'inertie, des influences cognitives (*pbest*) et sociales (*gbest*). Ces vitesses mises à jour guident les particules vers des solutions potentiellement meilleures.

7. **Respect des Contraintes** : Avant la mise à jour des positions, les contraintes du problème sont vérifiées pour s'assurer que les nouvelles positions respectent les contraintes spécifiées.

    **Réparation des Positions** : Si nécessaire, les positions des particules sont réparées pour se conformer aux contraintes du problème. Voici les problèmes

que l'on peut rencontrer et comment les corriger.

- Si une position est inférieure à 0 après la mise à jour avec la vitesse, il convient de la corriger en la fixant à 0.
- Si une position est supérieure à 1 après la mise à jour avec la vitesse, elle doit être ajustée à 1.
- Si la somme des valeurs sur une ligne (correspondant à l'ordonnancement d'une tâche) est strictement supérieure à 1, cela indique qu'une tâche s'exécute sur plusieurs chemins.
  Correction : on ramène la somme à 1 en mettant à 0 les valeurs excédentaires de (somme - 1) chemins.

8. **Mesure de la Valeur de Fitness** : La valeur de fitness est mesurée pour chaque particule en fonction de sa nouvelle position. Cela évalue à quel point la solution est prometteuse.

9. **Mise à Jour de** *pbest* **et** *gbest* : Les meilleures positions individuelles (*pbest*) et globales (*gbest*) sont mises à jour en fonction des nouvelles évaluations de fitness. Cela permet aux particules de retenir les meilleures solutions découvertes.

10. **Itération Suivante** : Le numéro d'itération est incrémenté, et le processus se répète jusqu'à ce que le critère d'arrêt soit atteint.

La figure 4.2 présente un diagramme qui résume les différentes étapes de COPSO.

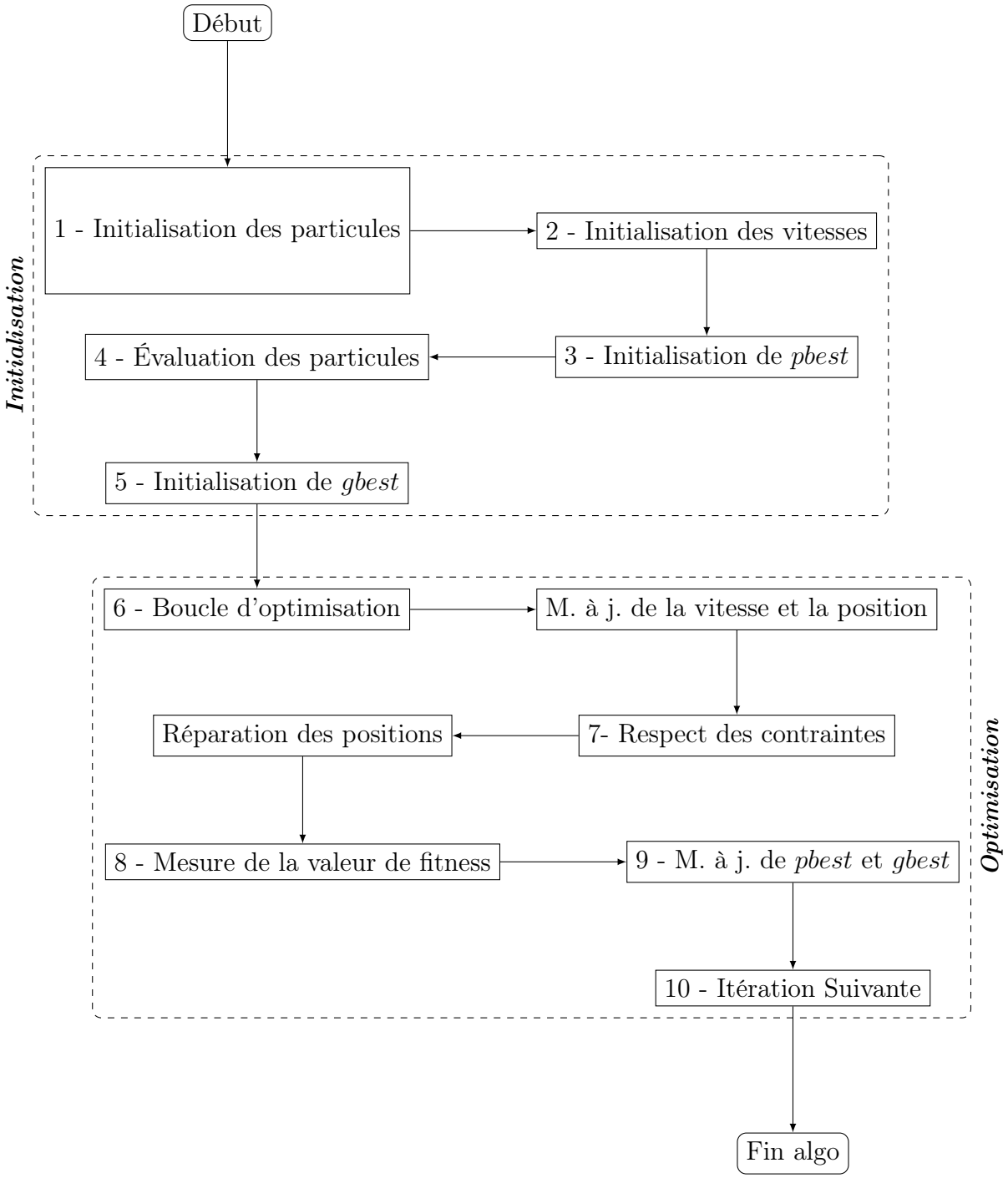


FIGURE 4.2 Diagramme des étapes de COPSO

## 4.3 Une solution basée sur une stratégie gloutonne : **COGA** (computing offloading greedy algorithm)

Les algorithmes gloutons sont des méthodes heuristiques conçues pour résoudre des problèmes d'optimisation en prenant des décisions localement optimales à chaque étape du processus. Bien que ces algorithmes puissent rapidement arriver à une solution acceptable, cette solution n'est pas nécessairement la meilleure sur le plan global. Le choix à chaque étape est fait en se basant sur l'information immédiatement disponible, sans prendre en compte les implications à long terme. Cette approche est particulièrement utile lorsque le temps d'exécution est une contrainte importante ou lorsque le problème est si complexe qu'une solution exacte est difficile à obtenir. Cependant, il faut être prudent lors de l'application de méthodes gloutonnes, car elles peuvent conduire à des solutions sous-optimales pour certains problèmes (Cormen et al., 2009).

Nous présentons dans ce qui suit les différentes étapes de l'algorithme heuristique COGA

1. **Entrées :** Prendre une liste de tâches en entrée, chaque tâche ayant des contraintes spécifiques.
2. **Génération des décisions de déchargement possibles :** Créer une liste des décisions de déchargement possibles en fonction des ressources disponibles.
3. **Boucle principale :** Cette boucle itère sur la totalité de l'ensemble des tâches :
4. **Recherche de la décision de déchargement compatible :** Trouver le premier chemin qui peut exécuter la tâche tout en respectant les contraintes.
5. **Attribution du chemin :** Si un chemin est trouvé, attribuer le chemin à la tâche.
6. **Mise à jour de l'état du système :** Mettre à jour l'état global en tenant compte de la nouvelle attribution.

7. **Passage à la tâche suivante :** Si aucun chemin n'est trouvé, passer à la tâche suivante.

8. **Terminaison :** L'algorithme se termine lorsque toutes les tâches ont été considérées. Algorithme 1 présente le pseudocode de COGA.

**Algorithme 1 :** Algorithme glouton pour le déchargement des tâches

**Données :** Liste des tâches et information sur l'environnement

**Résultat :** Décisions de déchargement

```
Génération de toutes les décisions de déchargement possibles (appreils, mmws, stations de base, MECs)
pour chaque tâche dans la liste des tâches faire
    Recherche du premier chemin capable d'exécuter la tâche tout en respectant les contraintes
    si un chemin est trouvé alors
        Attribuer le chemin à la tâche
        Mettre à jour l'état global du système
    sinon
        Passer à la tâche suivante
```

## 4.4 Une solution basée sur une stratégie aléatoire : **CORA** (computing offloading random algorithm)

### 4.4.1 Stratégie aléatoire

Un algorithme aléatoire, également connu sous le nom d'algorithme probabiliste, est un type d'algorithme qui utilise des éléments de hasard ou de probabilité dans son processus de prise de décision. Contrairement aux algorithmes déterministes qui produisent toujours la même sortie pour une entrée donnée, les algorithmes aléatoires introduisent une composante aléatoire qui peut conduire à des résultats

différents à chaque exécution, même pour les mêmes entrées.

**Algorithme 2 :** Algorithme Aléatoire pour le déchargement des tâches

**Données :** Liste des tâches et information sur l'environnement

**Résultat :** Décisions de déchargement

```
Génération de toutesles décisions de déchargement possibles (appreils, mmws,
  stations de base, MECs)
Générer tous les chemins possibles (appareils, mmws, stations de base, MECs)
pour chaque tâche dans la liste des tâches faire
    Choisir aléatoirement un chemin capable d'exécuter la tâche
    si le chemin est valide alors
        Attribuer le chemin à la tâche
        Mettre à jour l'état du système
    sinon
        Passer à la tâche suivante
```

### 4.4.2 Conclusion

En conclusion, ce chapitre a introduit l'algorithme d'optimisation par essaim de particules (PSO) comme solution pour résoudre le problème formulé précédemment. Une vue d'ensemble du fonctionnement général du PSO a été présentée, soulignant son inspiration à partir du comportement des essaims d'oiseaux. L'algorithme PSO est conçu pour guider les essaims vers des solutions optimales en utilisant des règles simples telles que la cohésion, l'alignement, et la séparation. Les sections suivantes du chapitre exploreront également d'autres approches, notamment une solution basée sur une stratégie gloutonne (COGA) et une solution basée sur une stratégie aléatoire (CORA).

# CHAPITRE V

# ÉVALUATION DE PERFORMANCES

Dans la présente section, nous évaluons les performances de notre solution basée sur PSO, et la comparons aux méthodes de déchargement à la fois gloutonne et aléatoire.

## 5.1 Environnement de simulation (Koo et Lim, 2021 ; Azizi et al., 2022)

Nous avons réalisé nos simulations à l'aide de solutions conçues en Python, exécutées sur un ordinateur doté d'un processeur Intel Core(TM) i7-5500U CPU à 2.4 GHz, de 8 Go de RAM et d'un système d'exploitation windows 10 64 bits.

Dans le cadre de notre recherche, nous explorons un environnement réseau articulé autour d'une cellule, au cœur de laquelle se trouve une station de base macro avec une capacité de couverture s'étendant sur un rayon de 500 m. En complément, cette station est assistée par quatre autres stations de base mmWave, chacune étant capable de couvrir une étendue de 200m. Au sein de cette structure, nous disposons de 50 appareils mobiles, uniformément distribués.

Chacun des appareils mobiles est caractérisé par une file d'attente pouvant héberger dix tâches. Initialement, la capacité énergétique de ces appareils est fixée à 100%, mais celle-ci décroît proportionnellement à l'exécution des tâches. Le traitement

s'effectue à une cadence de 30Mb/s pour chaque appareil.

Parallèlement, la station de base macro est interfacée avec quatre serveurs MEC au moyen de liaisons fibrées. Pour les besoins de notre étude, nous avons choisi de ne pas prendre en compte la latence de cette connexion. Chaque serveur MEC dispose d'une vitesse de traitement de 2Gb/s et d'une file d'attente de 40 tâches.

Nos simulations considèrent un ensemble de tâches dont la taille est comprise entre 200 et 800 Kb. Chaque session de simulation s'étale sur un minimum de 40 intervalles de temps, durant lequel chaque appareil génère une tâche avec une probabilité de 0,05. Chaque tâche a une échéance de 6 secondes après sa génération. la valeur de $\epsilon$ est égal à 5.
Pour le PSO un nombre de 100 particules est utilisé avec une limite de 50 itérations, un poids d'inertie de 0.7, un poids cognitif de 1.4, et un poids social de 1.2 est étudié.

Quant aux spécifications techniques :

- La station de base macro opère sur une bande de fréquence de 100 MHz, avec une puissance de transmission de 20W. Le niveau de bruit est fixé à -101 dBm.
- Les stations de base mmWave travaillent sur une bande de 800 MHz, avec une puissance de transmission de 10W.
- Les appareils en communication D2D opèrent dans une plage de 20 MHz, avec une puissance de transmission de 5 W.

Il convient de souligner que, dans notre évaluation, les violations de délai sont considérées comme étant plus préjudiciables et sont donc pénalisées plus lourdement que la consommation énergétique.

## 5.2 Résultats

Dans un premier temps, nous examinons, pour chaque algorithme, le pourcentage de tâches traitées en fonction du nombre d'appareils présents dans le système à la Figure 5.1. Cette métrique offre un aperçu significatif. En effet, un pourcentage élevé de tâches déchargées reflète une volonté d'optimiser et d'étendre l'autonomie de la batterie, laquelle se dégrade progressivement lors de la réalisation des tâches. Par ailleurs, compte tenu de la saturation potentielle de la file d'attente des appareils, le déchargement des tâches vers les serveurs se présente comme une stratégie pertinente pour minimiser l'énergie.

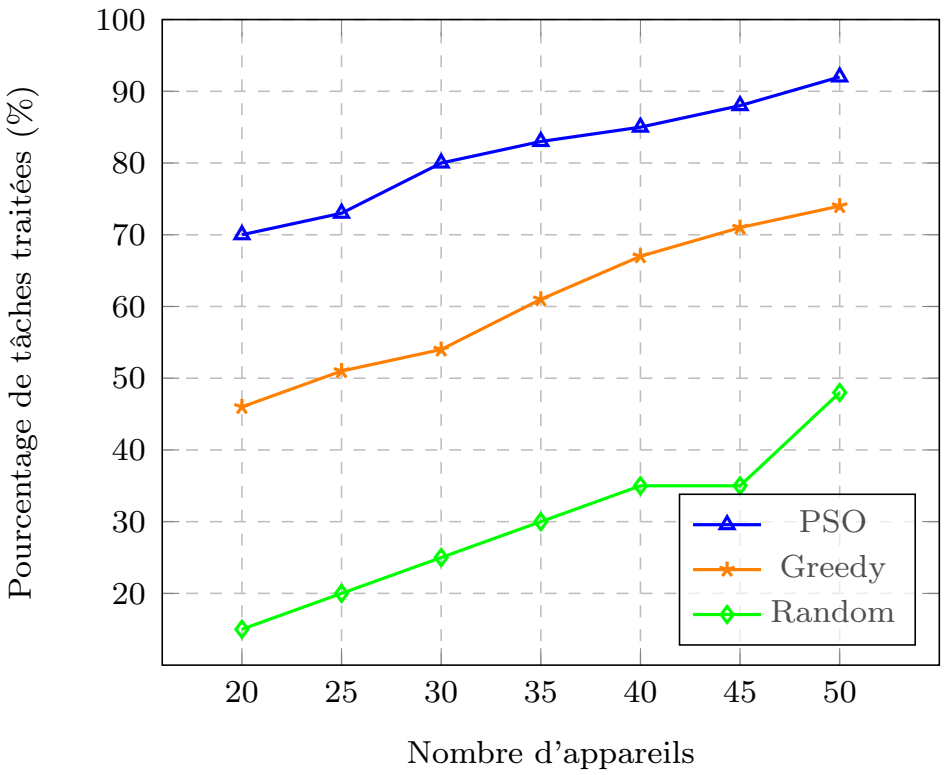


FIGURE 5.1 Pourcentage de tâches traitées en fonction du nombre d'appareils

Il est remarqué que l'algorithme basé sur le COPSO réalise le déchargement le plus

significatif, atteignant un maximum de 92% des tâches traitées. En comparaison, la stratégie gloutonne culmine à 74% des tâches, avec une progression plus modérée que le COPSO. En raison de notre critère de pénalisation plus élevé pour les violations de délai, la stratégie gloutonne montre une propension plus forte à traiter les tâches en local par rapport à la méthode COPSO, entraînant ainsi une consommation énergétique supérieure. Il est essentiel de souligner que cette consommation est intrinsèquement liée au temps nécessaire pour traiter la tâche. En tenant compte de la capacité de traitement nettement supérieure d'un serveur MEC par rapport à un appareil standard, la disparité en termes de consommation d'énergie devient évidente. une tendance prédominante émerge : à mesure que le nombre de d'appareils dans le réseau s'accroît, le pourcentage de tâches déchargées suit une courbe ascendante. Cette dynamique s'observe uniformément à travers les trois algorithmes étudiés.

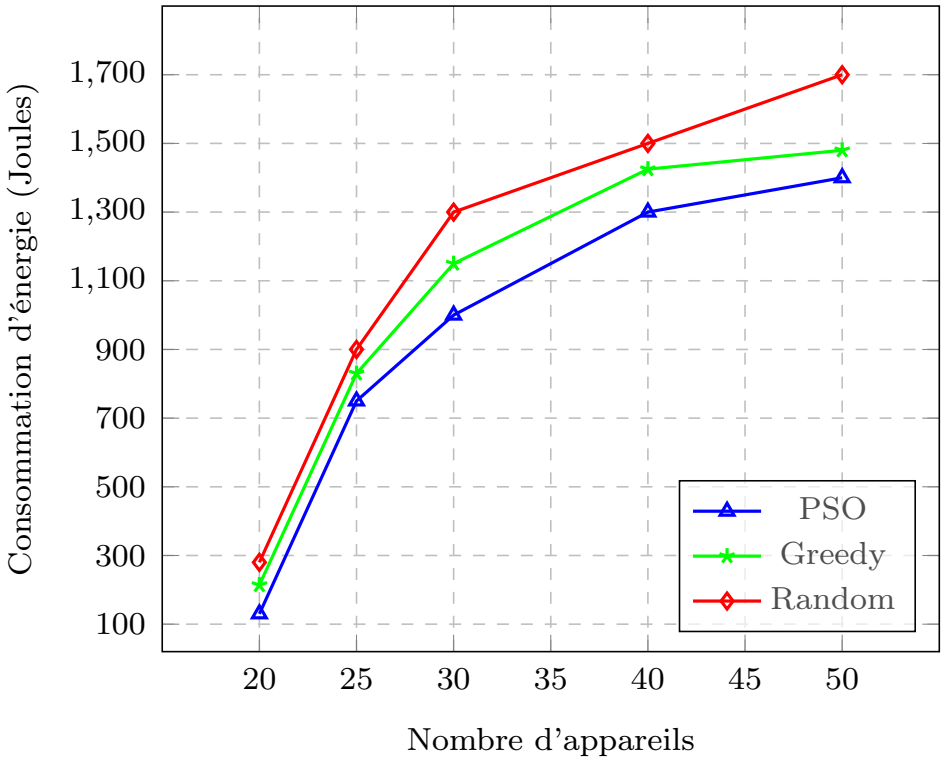


FIGURE 5.2 Consommation d'énergie en fonction du nombre d'appareils.

Dans la Figure 5.2, les paramètres demeurent constants pour l'analyse de la consommation énergétique associée à chaque stratégie. Il est crucial de noter que la consommation est directement corrélée au pourcentage de tâches traitées, augmentant à mesure que le déchargement croît. Ceci est dû aux coûts de transmission liés au déchargement des tâches.

Dans cette figure, on observe que les performances du PSO en termes de consommation d'énergie sont significativement inférieures à celles des autres algorithmes, atteignant une valeur de 1400J, tandis que le glouton se situe autour de 1500J et l'aléatoire atteint 1700J, intrinsèquement liée à l'augmentation du nombre d'appareils.

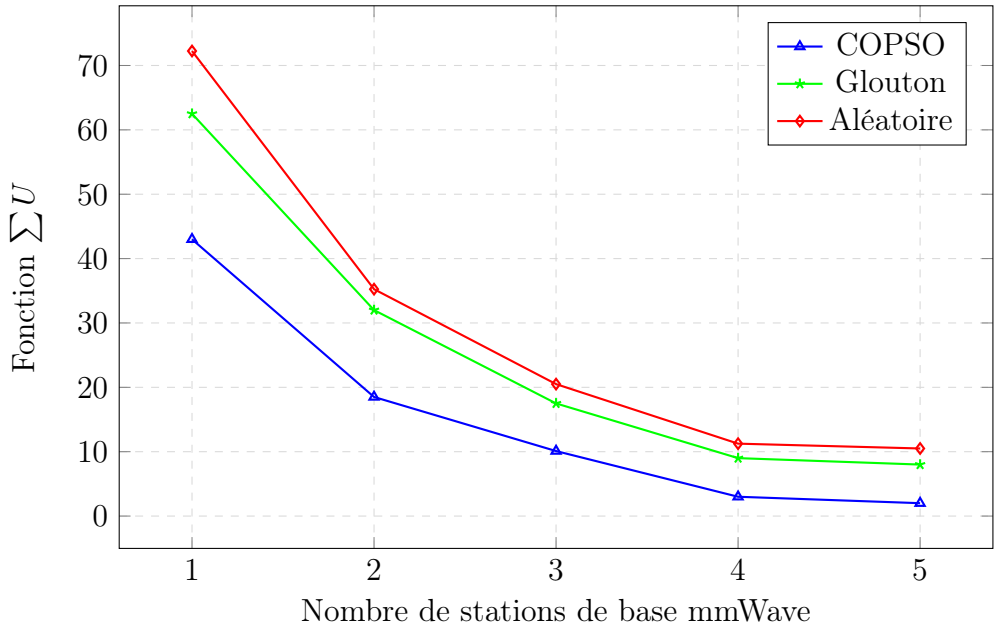


FIGURE 5.3 Impact du nombre de stations de base mmWave sur la fonction objectif

Dans la Figure 5.3, nous examinons l'impact de la variation du nombre de stations de base mmWave sur le coût total déterminé par notre fonction objectif. Nous maintenons la considération d'une pénalité pour les violations de délai, où chaque infraction au délai assigné à une tâche entraîne une pénalité.

La simulation se déroule avec 50 appareils traitant un ensemble de tâches, et la variable d'intérêt est le nombre de stations de base mmWave. L'augmentation de leur nombre améliore la couverture réseau, conduisant à une meilleure transmission radio, bénéfique pour le respect des exigences de délai et la consommation énergétique.

Cependant, il est essentiel de noter que l'effet positif de l'ajout de ces stations atteint un plateau après l'intégration de 4 stations de base mmWave. Au-delà de ce point, l'amélioration de la performance du réseau montre des signes de saturation, indiquant que l'ajout ultérieur de stations ne se traduit pas nécessairement par

des gains proportionnels en termes de couverture ou d'efficacité énergétique.

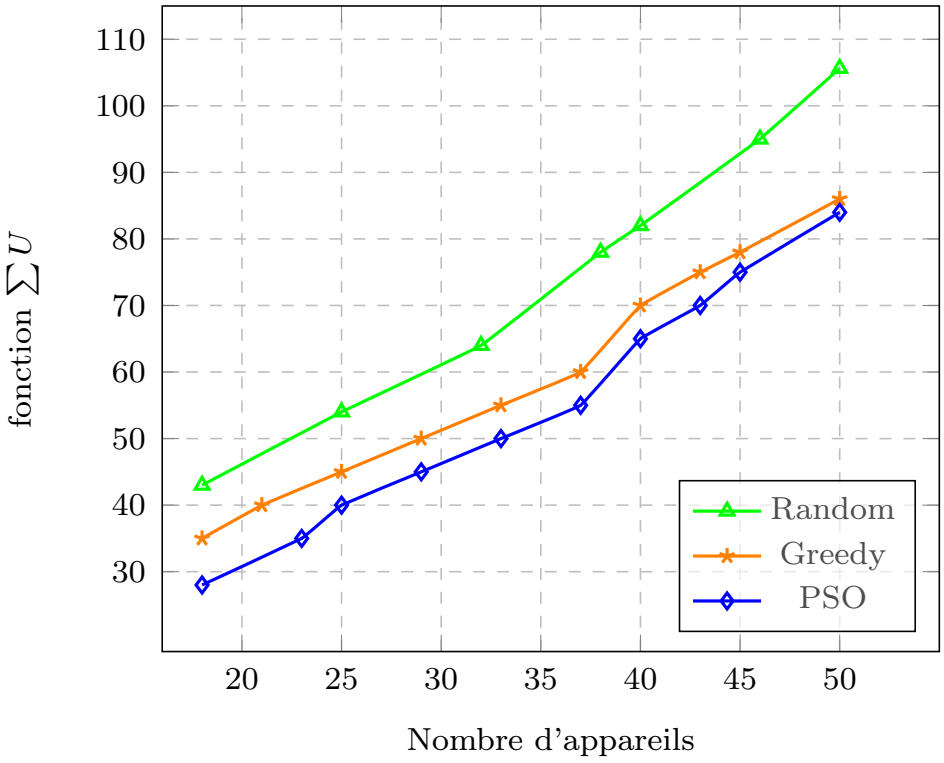


FIGURE 5.4 Impact du nombre d'appareils sur la fonction objectif

Dans la Figure 5.4, nous explorons l'impact de la variation du nombre d'appareils sur notre fonction objectif. Comme prévu, la densité d'appareils dans le système influe directement sur la fonction objectif.

À mesure que le nombre d'appareils augmente, la demande en ressources, que ce soit en termes de capacité de traitement ou de bande passante, augmente également. Cela peut entraîner une saturation plus rapide des ressources disponibles, surtout dans les scénarios où le nombre de stations de base mmWave ou de serveurs MEC est limité. De plus, une densité accrue d'appareils peut accroître les risques d'interférences, potentiellement nuisibles à la qualité de service.

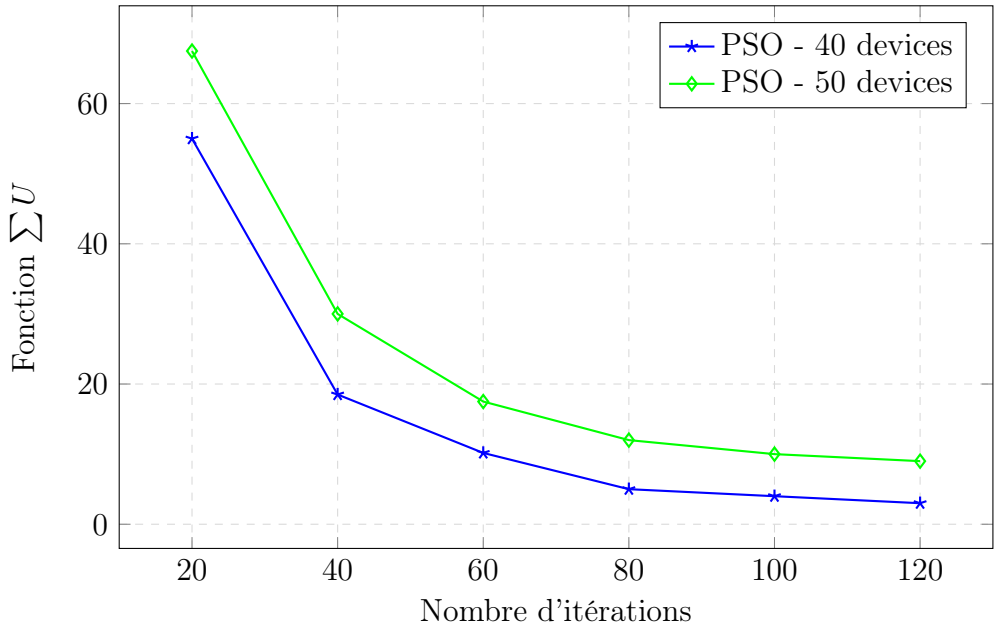


FIGURE 5.5 Courbe de convergence du nombre de taches traitées sur la fonction objectif

L'analyse de la convergence de l'algorithme PSO, illustrée dans la Figure 5.5, met en lumière des gains significatifs de performance initiale du PSO entre la première et la quarantième itération pour divers nombres de tâches. Toutefois, l'amélioration devient moins remarquée au-delà d'un certain nombre d'itérations, suggérant une possible convergence de l'algorithme. Le choix judicieux du nombre d'itérations est crucial, offrant une adaptabilité en fonction de l'application et de l'environnement. On note qu'une différence des deux courbes de convergence du PSO en variant le nombre d'appareils.

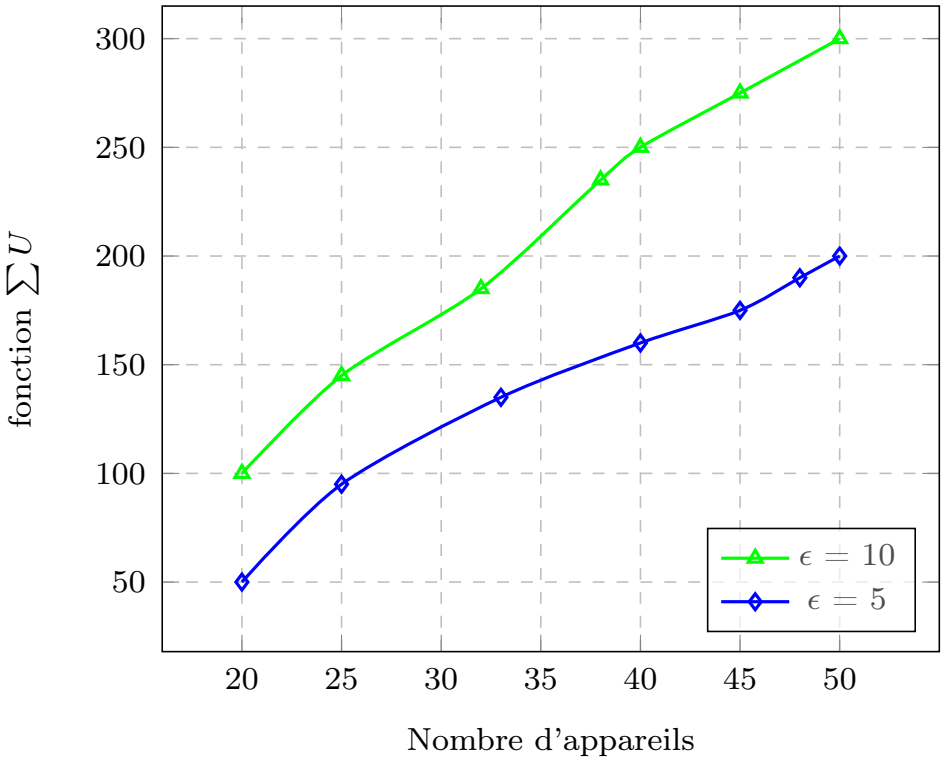


FIGURE 5.6 Impact de $\epsilon$ sur la fonction objectif

Dans cette section, nous examinons l'impact de $\epsilon$ sur la fonction objectif. Le paramètre $\epsilon$ représente une pénalité fixe. Nous avons sélectionné deux valeurs différentes pour $\epsilon$ afin d'évaluer son influence sur la fonction objectif.

Nous observons que le choix de $\epsilon$ a une incidence significative sur la fonction objectif. L'analyse de cette courbe révèle deux tendances distinctes. Lorsqu'un petit $\epsilon$ est choisi, cela revient à minimiser principalement l'énergie, car les tâches dépassant leur délai sont toujours exécutées. En revanche, en optant pour un $\epsilon$ plus élevé, l'accent est davantage mis sur le respect des délais, indépendamment de la quantité d'énergie consommée. Cette observation souligne l'importance du paramètre $\epsilon$ dans l'équilibrage entre la consommation énergétique et la conformité aux contraintes temporelles.

### 5.2.1 conclusion

En conclusion, ce chapitre a été dédié à l'évaluation des performances de notre solution basée sur PSO, en la comparant à des méthodes de déchargement gloutonne et aléatoire. Nous avons détaillé l'environnement de simulation utilisé, en mettant en lumière les caractéristiques de la cellule réseau, des appareils mobiles, des serveurs MEC, et des spécifications techniques des stations de base macro et mmWave.

Les résultats obtenus ont été présentés, mettant en évidence les performances de notre solution par rapport aux méthodes de déchargement alternatives. Nous avons pris en compte des critères tels que la consommation énergétique, le délai d'exécution des tâches. Ces indicateurs ont permis de quantifier l'efficacité de notre approche dans le déchargement des tâches, en soulignant son avantage concurrentiel, notamment en termes de réduction de la consommation d'énergie.

# CHAPITRE VI

# CONCLUSIONS ET TRAVAUX FUTURS

L'évolution rapide des appareils intelligents et des technologies de communication a suscité de nombreuses applications nouvelles exigeant des capacités de calcul élevées et une faible latence, telles que la réalité virtuelle et la réalité augmentée. Pour relever ces défis, nous avons élaboré une approche novatrice : une plateforme centralisée dans un environnement millimétrique.

Cette plateforme délègue certaines tâches à la périphérie du réseau, que ce soit au serveur directement, au serveur MEC en passant par les stations de base mmWave, ou en utilisant la communication directe entre appareils (D2D). Notre stratégie vise à maximiser le nombre de tâches respectant la contrainte de délai tout en minimisant la quantité globale d'énergie utilisée pour leur exécution. L'adoption d'un algorithme d'optimisation par essaim de particules nous a permis d'obtenir un plan de déchargement quasi-optimal, réduisant à la fois le délai de calcul des tâches du système et la consommation totale d'énergie.

Nous exprimons de manière analytique les problématiques mentionnées précédemment en les modélisant comme un problème de programmation linéaire en nombres entiers (ILP - Integer Linear Programming). Par la suite, une analyse approfondie de la complexité algorithmique est menée pour démontrer que le problème résultant est NP-difficile. Cette complexité souligne la nécessité de développer

une heuristique reposant sur l'optimisation par essaim particulaire pour aborder efficacement ces défis computationnels.

Les évaluations ont clairement démontré que notre approche basée sur COPSO surpasse les autres algorithmes, qu'ils soient gloutons ou aléatoires, tant en termes de consommation d'énergie que du pourcentage de tâches traitées avec succès.

À l'avenir, notre recherche s'orientera vers l'intégration d'algorithmes d'apprentissage automatique pour optimiser la prise de décision de déchargement des tâches, avec un accent particulier sur la gestion énergétique des ressources. Cette approche tiendra compte des charges de travail variables et imprévisibles des appareils, visant à développer des modèles prédictifs robustes pour une allocation plus efficiente des ressources informatiques. En résumé, l'utilisation de capacités prédictives de l'apprentissage automatique représente une réponse innovante aux défis posés par les applications émergentes et les besoins évolutifs des utilisateurs.

Parallèlement, notre extension future portera sur l'intégration d'aspects liés à la protection de la vie privée dans le contexte du déchargement au sein des réseaux MEC assistés par D2D.

# BIBLIOGRAPHIE


Aliyu, Suleiman Onimisi et al. (2017). "A Game-theoretic based QoS-Aware capacity management for real-time edgeiot applications". In : *2017 IEEE International Conference on Software Quality, Reliability and Security (QRS)*. IEEE, p. 386-397.

Alliance, NGMN (2015a). "5G white paper". In : *Next generation mobile networks, white paper* 1.2015.

– (2015b). "5G White Paper. Next Generation Mobile Networks, White Paper". In : *URL https ://www. ngmn. org/5g-white-paper*.

Azizi, Sadoon et al. (2022). "Deadline-aware and energy-efficient IoT task scheduling in fog computing systems : A semi-greedy approach". In : *Journal of network and computer applications* 201, p. 103333.

Baek, Hosung, Haneul Ko et Sangheon Pack (2020). "Privacy-Preserving and Trustworthy Device-to-Device (D2D) Offloading Scheme". In : *IEEE Access* 8, p. 191551-191560. doi : 10.1109/ACCESS.2020.3032735.

Burd, Thomas D et Robert W Brodersen (1996). "Processor design for portable systems". In : *Journal of VLSI signal processing systems for signal, image and video technology* 13.2, p. 203-221.

Cisco (2021). *Migration vers la 5G*. URL : https://www.silicon.fr/reseau-cisco-trafic-internet-2022-226519.html# (visité le 11/06/2022).

Cormen, Thomas H et al. (2009). *Introduction to algorithms*. MIT press.

Deng, Junquan et al. (2017). "Resource allocation and interference management for opportunistic relaying in integrated mmWave/sub-6 GHz 5G networks". In : *IEEE Communications Magazine* 55.6, p. 94-101.

Ghosh, Amitava et al. (2014). "Millimeter-wave enhanced local area systems : A high-data-rate approach for future wireless networks". In : *IEEE Journal on Selected Areas in Communications* 32.6, p. 1152-1163.

He, Shiwen et al. (2022). "GBLinks : GNN-based beam selection and link activation for ultra-dense D2D mmWave networks". In : *IEEE Transactions on Communications*.

He, Yinghui et al. (2019). "D2D communications meet mobile edge computing for enhanced computation capacity in cellular networks". In : *IEEE Transactions on Wireless Communications* 18.3, p. 1750-1763.

Hu, Wenjie et Guohong Cao (2017). "Quality-aware traffic offloading in wireless networks". In : *IEEE Transactions on Mobile Computing* 16.11, p. 3182-3195.

Huang, Liang, Suzhi Bi et Ying-Jun Angela Zhang (2020). "Deep Reinforcement Learning for Online Computation Offloading in Wireless Powered Mobile-Edge Computing Networks". In : *IEEE Transactions on Mobile Computing* 19.11, p. 2581-2593. doi : 10.1109/TMC.2019.2928811.

KLEINBERG, Jon et Eva TARDOS (2006). *Algorithm design*. Pearson Education India.

KOO, Seolwon et Yujin LIM (2021). "Optimal Task Offloading Decision in IIoT Enviornments Using Reinforcement Learning". In : *2021 IEEE 3rd Eurasia Conference on IOT, Communication and Engineering (ECICE)*. IEEE, p. 86-89.

LI, Yujin, Lei SUN et Wenye WANG (2014). "Exploring device-to-device communication for mobile cloud computing". In : *2014 IEEE international conference on communications (ICC)*. IEEE, p. 2239-2244.

LINQIAN et al. (2022). "Resource Allocation and Computation Offloading in a Millimeter-Wave Train-Ground Network". In : *IEEE Transactions on Vehicular Technology* 71.10, p. 10615-10630. DOI : 10.1109/TVT.2022.3185331.

LIU, Yanzhen et al. (2022). "Latency Minimization for mmWave D2D Mobile Edge Computing Systems : Joint Task Allocation and Hybrid Beamforming Design". In : *IEEE Transactions on Vehicular Technology* 71.11, p. 12206-12221. DOI : 10.1109/TVT.2022.3192345.

MAO, Yuyi et al. (2017). "A survey on mobile edge computing : The communication perspective". In : *IEEE communications surveys & tutorials* 19.4, p. 2322-2358.

MUDUMBAI, Raghuraman, SK SINGH et Upamanyu MADHOW (2009). "Medium access control for 60 GHz outdoor mesh networks with highly directional links". In : *IEEE INFOCOM 2009*. IEEE, p. 2871-2875.

RAJASEKARAN, Aditya S. et al. (2020). "User Clustering in mmWave-NOMA Systems With User Decoding Capability Constraints for B5G Networks". In : *IEEE Access* 8, p. 209949-209963. DOI : 10.1109/ACCESS.2020.3039276.

SOLAIMAN, Suhare, Laila NASSEF et Etimad FADEL (2021). "User clustering and optimized power allocation for D2D communications at mmWave underlaying MIMO-NOMA cellular networks". In : *IEEE Access* 9, p. 57726-57742.

SUN, Weijie et al. (2019). "Profit Maximization Task Offloading Mechanism with D2D Collaboration in MEC Networks". In : *2019 11th International Conference on Wireless Communications and Signal Processing (WCSP)*, p. 1-6. DOI : 10.1109/WCSP.2019.8928117.

TARJAN, Robert Endre et Anthony E TROJANOWSKI (1977). "Finding a maximum independent set". In : *SIAM Journal on Computing* 6.3, p. 537-546.

WANG, Haixia et al. (2017). "Joint computation offloading and data caching with delay optimization in mobile-edge computing systems". In : *2017 9th International Conference on Wireless Communications and Signal Processing (WCSP)*. IEEE, p. 1-6.

WANG, Kehao et al. (2020). "Joint Offloading and Charge Cost Minimization in Mobile Edge Computing". In : *IEEE Open Journal of the Communications Society* 1, p. 205-216.

WEI, Xiaojuan et al. (2017). "MVR : An architecture for computation offloading in mobile edge computing". In : *2017 IEEE International Conference on Edge Computing (EDGE)*. IEEE, p. 232-235.

YU et Huang JIALI (2022). "Computation Efficiency Optimization for RIS-Assisted Millimeter-Wave Mobile Edge Computing Systems". In : *IEEE Transactions on Communications* 70.8, p. 5528-5542. DOI : 10.1109/TCOMM.2022.3181673.

YU, Xiangbin et al. (2023). "Computation Efficiency Optimization for Millimeter-Wave Mobile Edge Computing Networks With NOMA". In : *IEEE Transactions on Mobile Computing* 22.8, p. 4578-4593. DOI : 10.1109/TMC.2022.3164974.

ZHANG, Ni et al. (2019). "Joint Task Offloading and Data Caching in Mobile Edge Computing". In : *2019 15th International Conference on Mobile Ad-Hoc and Sensor Networks (MSN)*. IEEE, p. 234-239.

ZHANG, Weiwen et al. (2013). "Energy-optimal mobile cloud computing under stochastic wireless channel". In : *IEEE Transactions on Wireless Communications* 12.9, p. 4569-4581.

ZHAO, Cunzhuo et al. (2020). "Mobile Edge Computing Meets mmWave Communications : Joint Beamforming and Resource Allocation for System Delay Minimization". In : *IEEE Transactions on Wireless Communications* 19.4, p. 2382-2396. DOI : 10.1109/TWC.2020.2964543.

ZHU, Tongxin et al. (2020). "Computation scheduling for wireless powered mobile edge computing networks". In : *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*. IEEE, p. 596-605.