Université du Québec
Institut National de la Recherche Scientifique
Centre Énergie, Matériaux et Télécommunications

Tactile Internet over Fiber-Wireless Enhanced HetNets using Edge Intelligence

Par

Amin Ebrahimzadeh

Thèse présentée pour l'obtention du grade de
Doctorat en Télécommunications, Ph.D.

Jury d'évaluation

Examinateur externe	Roch H. Glitho (Concordia University)
Examinateur externe	Vincent Duchaine (École de Technologie Supérieure),
Examinateur interne	Tiago H. Falk (INRS ÉMT)
Directeur de recherche	Martin Maier (INRS ÉMT)

To my parents,
who instilled in me the virtues of perseverance and commitment
and relentlessly encouraged me to strive for excellence.

To my family and friends,
especially from my hometown, Tabriz, Iran,
for their endless love, support, and encouragement.

Acknowledgments

"*Let us be grateful to the people who make us happy; they are the charming gardeners who make our souls blossom.*"

Marcel Proust

The completion of this doctoral thesis was indeed a meandering road, which would have never been possible without the support of a number of amazing people, whom I was lucky enough to cross paths with.

First and foremost, I owe a great debt of gratitude to my mentor and supervisor for the past four years, Prof. Martin Maier, who has always fascinated me by his dedication, intelligence, and professionalism. He showed my how be a big picture thinker yet a detail-oriented researcher. He always guided me away from incremental research and showed me how to explore the unknown, how to structure innovative research, and how to set milestones and achieve them efficiently.

The past four years at INRS have been a great experience both educationally and socially. I would like to take this opportunity to appreciate the staff members of INRS. In particular, I thank Hélèn Sabourin, Tatiana Brahmi, and Sylvain Fauvel for their kind support during my PhD studies. Special thanks go to my group mate Abdeljalil Beniiche, who kindly helped me translate the French resume of my thesis. I also thank the incredible friends I've made during the past four years at INRS.

8,942 km away from home, I would like to express my deep gratitude to all my mentors, from whom I had learned a lot throughout my educational journey back in my beloved country, Iran. Unfortunately, these lines are too short and my memory is too limited to remember and name all and put in words how important their role has been. Nevertheless, I would like thank Prof. Akbar Ghaffarpour Rahbar, Prof. Behrooz Alizadeh, Prof. Javad Musevi Niya, Prof. Behzad Mozaffari, and Mr. Ghanbari. Also, I would like to thank my friends from my hometown, Tabriz, for their true friendship and encouragement.

Back in 2016 during the first two semesters of my doctoral studies I took courses at McGill University and Concordia University, where I was fortunate enough to learn a lot from Prof. Yousef R. Shayan, Prof. Chadi Assi, and Prof. Mark Coates. I was truly fascinated by their deep knowledge and great lectures. I am particularly grateful to Prof. Jane M. Simmons, the EiC of IEEE/OSA JOCN, who saw value in my work right when nobody else seemed to be interested in it, and who gave us the opportunity to write an invited paper, which has received quite some attention thus far. I would also like to thank Prof. Eckehard Steinbach, Dr. Claudio Pacchierotti, and Dr. Leonardo Meli for providing us with the teleoperation traces. I appreciate the government of Québec for the financial support during my PhD studies. This work was supported by NSERC Discovery Grant

2016-04521 and the Fonds de Recherche du Québec-Nature et Technologies (FRQNT) B2X doctoral scholarship programme.

My sincere gratitude goes to the jury members of my Ph.D thesis, Prof. Roch H. Glitho, Prof. Vincent Duchaine, and Prof. Thiago H. Falk for their time and effort in evaluating my thesis.

Last but not least, special thanks goes to my parents, my family, and my love, Atefeh, for their endless, unconditional love and support. I am blessed beyond words to have had their love, encouragement, and support in ups and downs of this journey.

Amin Ebrahimzadeh
Montréal, Canada
September 2019

Abstract

Today's telecommunication networks enable people and devices to exchange a tremendous amount of audiovisual and data content.With the advent of commercially available haptic/tactile sensory and display devices and conventional triple-play (i.e., audio, video, and data) content communication now extends to encompass the real-time exchange of haptic information (i.e., touch and actuation) for the remote control of physical and/or virtual objects through the Internet. This paves the way towards realizing the so-called *Tactile Internet*. Through human-machine interaction, the Tactile Internet is expected to convert today's content delivery networks into skillset/labor delivery networks. The Tactile Internet holds great promise to have a profound socio-economic impact on a broad array of applications in our everyday life, ranging from industry automation and transport systems to healthcare, telesurgery, and education. In most of these industry verticals, very low latency and ultra-high reliability are key for realizing immersive applications such as robotic teleoperation. While necessary, though, the design of ultra-reliable and low-latency converged communication network infrastructures is not sufficient to unleash the full potential of the Tactile Internet. In this thesis, we put forward the idea that the Tactile Internet may be the harbinger of human augmentation and human-machine symbiosis envisioned by contemporary and early-day Internet pioneers. In search for synergies between humans and machines/robots, we explore the idea of treating the human as a "member" of a team of intelligent machines rather than keep viewing him as a conventional "user" while putting a particular focus on developing systems that are human-aware and help advance the human condition, e.g., economic inequality. After describing the Tactile Internet's human-in-the-loop-centric design principles and haptic communications and traffic models, we elaborate on the development of decentralized cooperative dynamic bandwidth allocation algorithms for end-to-end resource coordination in fiber-wireless (FiWi) access networks. We then use machine learning to decouple haptic feedback from the impact of extensive propagation delays. Next, we propose a context- and self-aware allocation scheme for both physical and digital tasks to coordinate the automation and augmentation of mutually beneficial human-machine coactivities while spreading ownership of robots across users. In addition to realizing collective context-awareness via task coordination, we aim to exploit local self-awareness in order to improve the energy-delay performance of mobile robots. Further, we study the problem of joint prioritized scheduling and assignment of delay-constrained teleoperation tasks to human operators. Finally, this doctoral thesis investigates the performance gains of cooperative computation offloading for multi-access edge computing (MEC) enabled FiWi enhanced heterogenous networks (HetNets) with capacity-limited backhaul links. After presenting the envisioned two-tier MEC architecture for a FiWi based networking infrastructure, a simple but efficient offloading strategy is proposed, which relies on the flexible trilateral cooperation between end-devices, edge servers, and the remote cloud.

Keywords: AI, Computation Offloading, Context-awareness, FiWi enhanced HetNets, Motion Planning, Multi-access Edge Computing, OPEX, Self-awareness, Teleoperation, Task Allocation.

Statement of Originality

I hereby certify that this thesis contains original work of the author. Some techniques employed from other authors are properly referenced herein.

Amin Ebrahimzadeh
INRS, ÉMT, Montréal, QC
Université du Québec
Date: 25 November 2019

Résumé

Introduction

Tandis que l'exploitation commerciale de l'Internet mobile permet aux utilisateurs d'échanger le trafic traditionnel à triple play (i.e., audio, vidéo et donné), le nouveau réseau *Internet Tactile* émergeant envisage de réaliser des *communications haptiques*, permettant aux utilisateurs de non seulement voir et entendre, mais aussi toucher et manipuler des objets physiques et/ou virtuels distants via Internet. L'Internet Tactile promet de créer de nouvelles opportunités entrepreneuriales et nouvelles emplois, qui devraient avoir un impact socio-économique profond sur presque chaque segment de notre vie quotidienne, avec des cas d'utilisation allant de la réalité augmentée/virtuelle (AR/VR) à la conduite autonome, et les réseaux électriques intelligents. Un grand nombre de ces secteurs de l'industrie nécessitent une très faible latence et une fiabilité extrême pour la réalisation d'applications interactives ultra-réactives telles que la télé-opération/télé-présence bilatérale. Notez cependant que certains cas d'utilisation ne nécessitant pas nécessairement une mobilité constante peuvent être réalisés sur des réseaux à large bande fixes. Cela suggère que les futurs réseaux cellulaires doivent être entièrement convergés, ce qui permet une sélection flexible de différentes technologies d'accès fixes et mobiles tout en partageant les fonctionnalités du réseau principales. Les systèmes interactifs, notamment AR/VR et la télé-opération, exigent une latence aller-retour ultra-faible de 1 à 10 ms et une grande fiabilité. La haute disponibilité et la sécurité, les temps de réponse ultra-rapides et extrêmement fiables, ainsi que la fiabilité de l'Internet Tactile de niveau opérateur, ajoutera une nouvelle dimension à l'interaction des humains avec les machines/robots.

Objectifs

Le premier objectif de la thèse est d'examiner les principes de conception centrés sur l'homme (HITL) qui ajoutent une nouvelle dimension à l'interaction homme-machine via Internet, et démarquer Internet Tactile de l'Internet des objets (IoT) plus centré sur la machine. Un autre objectif de la thèse est d'explorer comment nous pouvons nous assurer que le potentiel de l'Internet Tactile soit libéré pour une course avec (plutôt que contre) des machines. L'un des principaux objectifs de cette thèse est de comprendre le trafic Internet Tactile en matière de débit de paquets, de taille de paquets, de temps d'inter-arrivage de paquets et d'auto-corrélation d'échantillon. Le délai décisif pour le déploiement réussi de la télé-opération Internet Tactile est le temps de latence de bout en bout, qui est soit imposé par la vitesse de la lumière (également appelée délai de propagation), soit par la ou les files d'attente intermédiaires. Concevoir des solutions appropriées de contrôle d'accès au support (MAC) au niveau des réseaux d'accès, ainsi que rapprocher les points d'interaction de télé-opération peut réduire en partie le dernier composant de délai. Néanmoins, le délai de propagation pose toujours de fortes limites quant à la possibilité de réaliser une télé-opération transparente et stable. Vers une conception de la télé-opération Internet Tactile réactive, à très faible latence et extrêmement fiable, afin de libérer pleinement le potentiel des réseaux améliorés par la fibre-sans fil (FiWi), le rôle de l'intelligence de bord, déployé au sein de l'informatique de périphérie à accès multiples (MEC) serveurs situés à proximité des utilisateurs finaux, nécessite une enquête approfondie. Un objectif important de la thèse est de fournir des informations sur l'utilisation de l'apprentissage automatique pour compenser les échantillons haptiques retardés via une approche de conception centrée sur l'homme. Bien que nécessaires, un temps d'exécution réduit et une connectivité homme-robot ultra-fiable ne suffisent pas pour libérer tout le potentiel des applications HART (Human-Agent-Robot Teamwork) résultantes. L'un des objectifs de la thèse est d'étudier les possibilités d'étendre davantage les capacités des HetNets LTE-A améliorés par la FiWi, en accordant une attention particulière à la dichotomie entre l'automatisation et l'augmentation (c'est-à-dire, l'extension des capacités) de l'homme via Internet Tactile. L'un des objectifs de la thèse est de concevoir une solution algorithmique pour résoudre le problème de l'ordonnancement conjoint et de l'attribution de tâches de télé-opération à des opérateurs humains tout en tenant compte des considérations économiques. Outre les considérations économiques et les exigences ultra-fiables en matière de communications à faible temps de latence (URLLC), un autre aspect important de la vision de la 5G est la décentralisation. Un objectif important de la thèse est donc d'étudier le rôle de la MEC coopérative dans un scénario informatique hiérarchique en nuage, en périphérie et au niveau local.

Méthodologie

La méthode de recherche appliquée dans cette thèse comprend la modélisation du trafic, la conception de l'architecture de réseau, la conception des mécanismes et l'analyse de la performance, comme résumé à la fig. R.1. Dans cette thèse, une modélisation statistique du trafic de télé-opération basée sur les traces est réalisée, qui étudie le rôle du codage en bande morte pour réduire le débit de paquets haptique. La caractérisation du trafic haptique présentée vise à modéliser la taille ainsi que le processus d'arriver des paquets haptiques dans les chemins de commande et de retour, en tenant compte du codage perceptuel. De plus, les modèles de corrélation au sein des échantillons de retour sont identifiés et modélisés en matière d'auto-corrélation des échantillons. Du point de vue de la conception architecturale, de nouvelles architectures de réseau basées sur FiWi sont développées pour permettre l'application des applications Internet immersives à faible temps de latence. Les architectures développées incluent notamment l'intelligence artificielle (AI) intégrée à MEC sur des réseaux HetNets améliorés par le FiWi, ce qui permet de prévoir des échantillons haptiques. En outre, une architecture informatique de bord hiérarchique est développée pour permettre un déchargement informatique coopératif afin d'aider les utilisateurs mobiles à accélérer l'exécution des tâches tout en économisant de l'énergie. De nouvelles solutions algorithmiques sont présentées tout au long de la thèse pour répondre aux objectifs susmentionnés. Ces algorithmes/mécanismes comprennent la prévision d'échantillons haptiques basée sur l'IA, la coordination de tâches conscientes du contexte et de soi, la planification hiérarchisée et l'attribution de tâches de télé-opération, ainsi qu'une stratégie de déchargement informatique coopératif. Dans la thèse, la théorie des files d'attente, la théorie des probabilités, l'optimisation multi-objectifs et l'apprentissage automatique ont été mis à profit pour développer les cadres d'analyse et mener des évaluations de performance. En particulier, un cadre analytique complet basé sur la théorie des probabilités et la théorie de la file d'attente est développé.

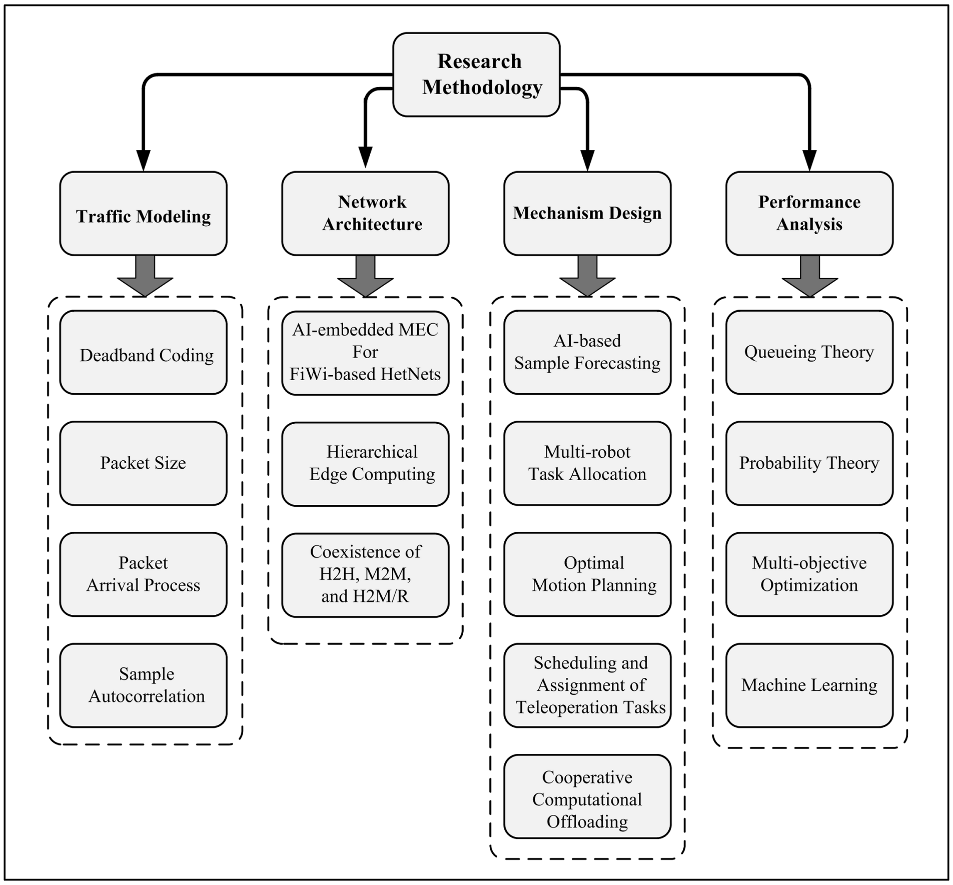

Research Methodology
Traffic Modeling
Network Architecture
Mechanism Design
Performance Analysis
Deadband Coding
Packet Size
Packet Arrival Process
Sample Autocorrelation
AI-embedded MEC For FiWi-based HetNets
Hierarchical Edge Computing
Coexistence of H2H, M2M, and H2M/R
AI-based Sample Forecasting
Multi-robot Task Allocation
Optimal Motion Planning
Scheduling and Assignment of Teleoperation Tasks
Cooperative Computational Offloading
Queueing Theory
Probability Theory
Multi-objective Optimization
Machine Learning

Figure R.1: Méthodologie de recherche.

Contributions de la thèse

Cette thèse est une compilation de dix publications, qui sont publiées ou acceptées pour publication dans des revues de haut calibre IEEE et OSA ainsi qu'à des éditeurs de livres renommés. Les principales contributions de la thèse sont brièvement décrites ci-après.

Expériences Internet Tactile immersives via Edge Intelligence

Nous portons notre attention sur la télé-opération bilatérale à titre d'exemple d'application HITL et présentons une étude approfondie de la caractérisation et de la modélisation du trafic haptique en matière d'arrivée de paquets et d'auto-corrélation d'échantillons. Notre objectif est de développer de nouveaux modèles de description des temps inter-arrivés de paquets ainsi que de l'auto-corrélation d'échantillons tridimensionnels. Nous explorons ensuite comment MEC en général et l'intelligence périphérique, en particulier, peuvent être utilisés pour aider à réaliser une expérience de télé-opération immersive et fiable sur des infrastructures de réseau basées sur FiWi.

Un exemple intéressant d'expérience Internet Tactile permettant une immersion à distance est le cas d'utilisation de la télé-opération centré sur HITL basé sur les *communications haptiques*. La figure R.2 illustre une description d'un système de télé-opération typique basé sur une communication haptique bidirectionnelle entre un opérateur humain (HO) et un télé-opérateur robot (TOR)

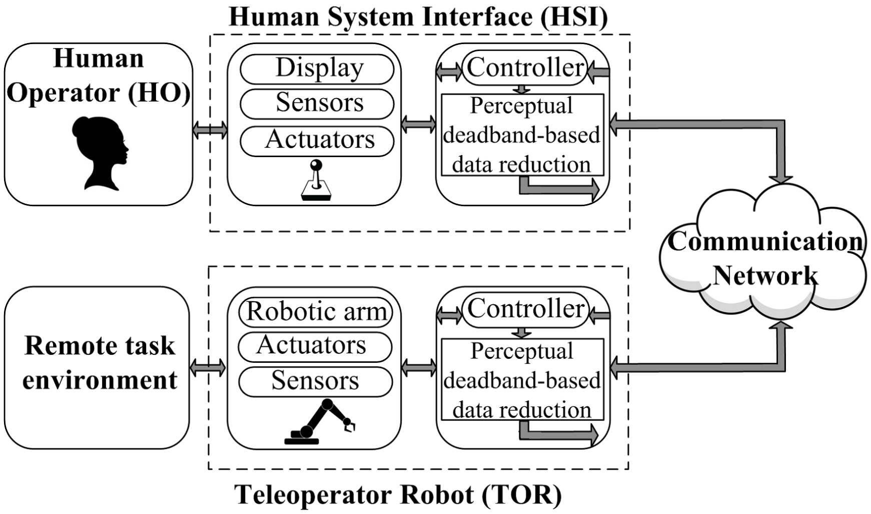
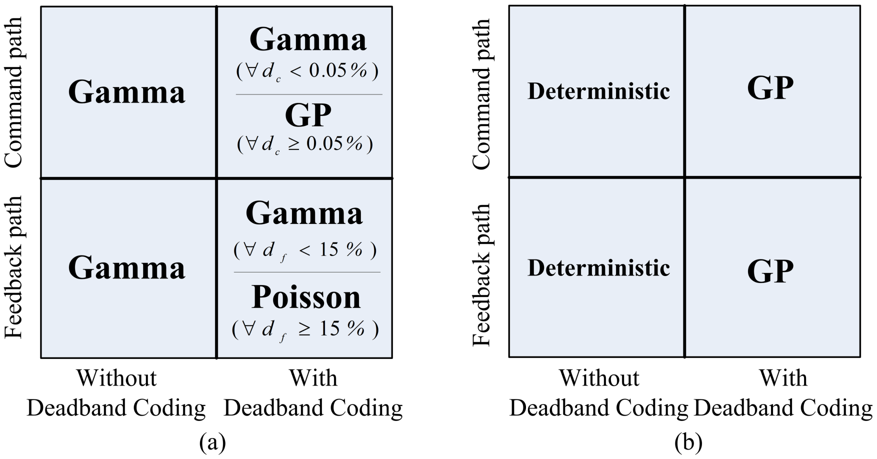


Figure R.2: Système de télé-opération basé sur des communications haptiques bidirectionnelles entre HO et TOR dans un environnement de tâches distant.

dans un environnement de tâche distant. Le HO s'interface avec le réseau de communication via le dispositif d'interface homme-système (HSI), utilisé pour afficher une interaction haptique avec le TOR distant vers le HO. Les contrôleurs situés aux deux extrémités du système de télé-opération assurent les performances de suivi et la stabilité du HSI et du TOR. Une réduction de données basées sur une bande morte perceptuelle peut être utilisée comme mécanisme de compression avec perte en exploitant le fait que les utilisateurs finaux humains ne sont pas en mesure de discriminer des différences arbitrairement petites entre les stimuli haptiques. La perception humaine de l'haptique peut être exploitée pour réduire le débit de paquets haptiques. Plus précisément, la loi de Weber, bien connue, détermine la différence juste perceptible (JND), c'est-à-dire le changement minimal de l'ampleur d'un stimulus pouvant être détecté par l'être humain. La loi de Weber donne lieu à la technique dite "*codage de la bande morte*", selon laquelle un échantillon haptique n'est transmis que si sa modification par rapport à l'échantillon haptique précédemment transmis dépasse un paramètre de bande morte donné $d \geq 0$. Notez que malgré l'intérêt croissant pour l'Internet Tactile, la compréhension des caractéristiques du trafic haptique réel reste limitée, notamment au niveau des paquets. Dans cette thèse, nous étudions deux séries de traces haptiques obtenues à partir d'expériences de télé-opération impliquant des TOR à différents degrés de liberté (DoF). Les deux expériences de télé-opération envisagées impliquent des TOR avec 1 et 6 DoF. De plus, nos traces haptiques comprennent des mesures avec différentes valeurs de paramètre de bande morte d.

La figure R.3 résume nos conclusions sur les différentes distributions de temps inter-arrivage de paquets les mieux adaptées pour les chemins de commande et de retour avec et sans codage en bande morte dans les deux scénarios de télé-opération. Nous observons qu'en général, les chemins de commande et de retour peuvent être modélisés conjointement par la distribution généralisée de Pareto (GP), la distribution de gamma ou par la distribution temporelle inter-arrivale des paquets, en fonction de la valeur donnée des paramètres de bande morte d_c et d_f, respectivement, comme indiqué dans Figure R.3.

La figure R.4 décrit l'architecture de réseau générique des réseaux HetNets LTE-A améliorés par le FiW. Le backhaul de la fibre consiste en un EPON IEEE 802.3ah/1/10 Gb/s multiplexage temps/longueur d'onde (TDM/WDM) avec une distance de fibre typique de 20 km entre le terminal



(a)

	Without Deadband Coding	With Deadband Coding
Command path	Gamma	Gamma $(\forall d_c < 0.05\%)$ GP $(\forall d_c \geq 0.05\%)$
Feedback path	Gamma	Gamma $(\forall d_f < 15\%)$ Poisson $(\forall d_f \geq 15\%)$

(b)

	Without Deadband Coding	With Deadband Coding
Command path	Deterministic	GP
Feedback path	Deterministic	GP

Figure R.3: Récapitulatif de la distribution temporelle optimale entre paquets pour les chemins de commande et de retour avec et sans codage en bande morte: (a) télé-opération à 6 DoF et (b) télé-opération à 1 DoF.

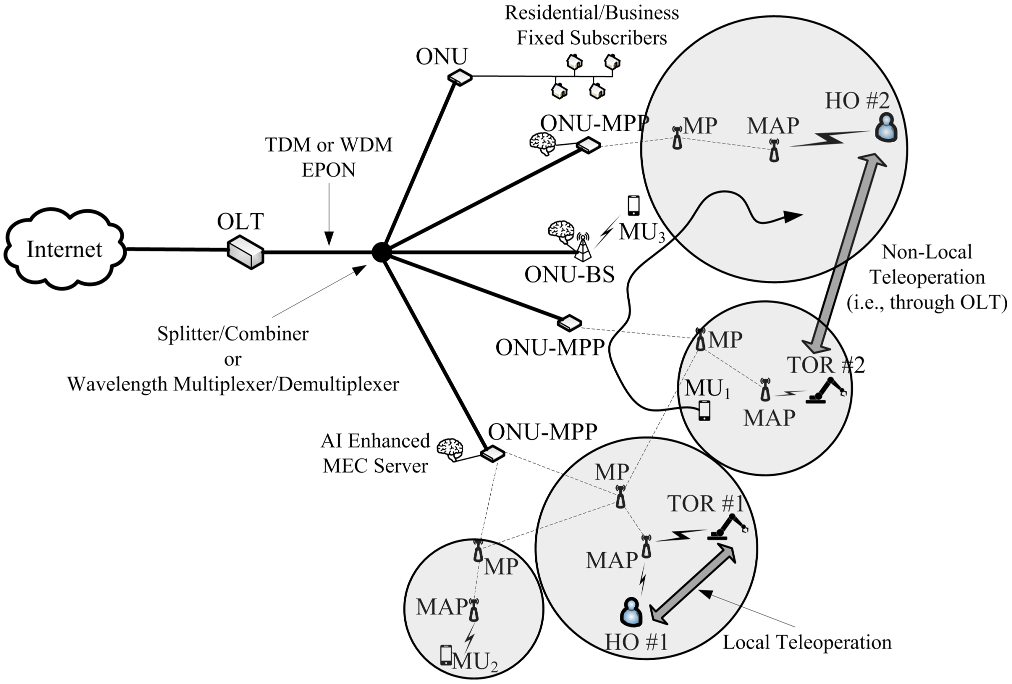


Figure R.4: Télé-opération locale et non locale dans les réseaux HetNets LTE-A améliorés par le FiWi avec fonctionnalités MEC basées sur l'IA.

de ligne optique (OLT) centrale et les unités de réseau optique (ONU) distantes. L'EPON peut comprendre plusieurs étapes, chaque étape étant séparée par un séparateur/combineur de diffusion en longueur d'onde ou un multiplexeur/démultiplexeur en longueur d'onde. Il existe trois différents sous-ensembles d'ONU. Une unité ONU peut desservir des abonnés fixes (câblés). Une unité ONU peut également se connecter à une station de base de réseau cellulaire (BS) ou à un point portail maille (MPP) WLAN IEEE 802.11n/ac/s, donnant lieu à une unité ONU-BS ou ONU-MPP co-localisée, respectivement. Selon son positionnement, un utilisateur mobile (UM) peut communiquer

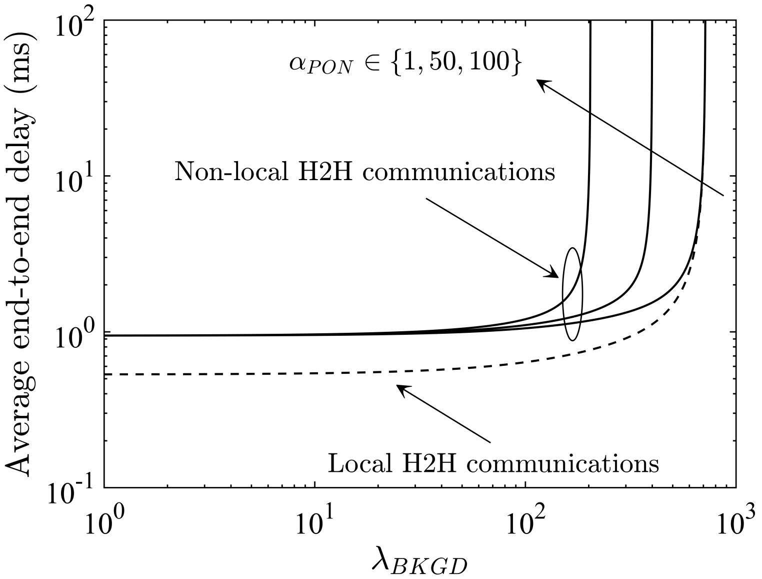


Figure R.5: Retard moyen de bout en bout des utilisateurs mobiles (UMs) par rapport au taux de trafic de fond moyen λ_{BKGD}(paquets/seconde) pour les communications H2H locales et non locales.

via le réseau cellulaire et/ou le maillage de réseau maillé WLAN, composé d'ONU-MPP, de point de maille (MP) intermédiaires et de point d'accès maille (MAP). Notez que la connexion de ces trois différents ensembles d'unités ONU via une infrastructure de backhaul à fibre partagée EPON permet d'atteindre l'important objectif de gain de convergence fixe-mobile de la stratégie d'opérateur de réseau actuelle. Dans cette thèse, contrairement aux études précédentes qui étudiaient uniquement la communication humain-à-humain (H2H) conventionnelle entre les MU, nous étudions le potentiel et les limites de la *télé-exploitation coexistante* dans les réseaux HetNets LTE-A améliorés par le FiWi. La figure R.5 illustre le délai moyen de bout en bout des UM par rapport au taux de trafic de fond moyen λ_{BKGD} pour les communications H2H locales et non locales dans les réseaux FiWi améliorés LTE-A. La figure montre qu'un délai moyen de bout en bout de $10^0 = 1$ ms peut-être atteint pour les communications H2H non locales pour une large gamme de charge de trafic en arrière-plan.

Malgré l'intérêt récent porté à l'exploitation de l'apprentissage automatique pour les communications optiques et les réseaux, l'intelligence périphérique permettant aux opérateurs humains de vivre une expérience de télé-opération immersive et transparente n'a pas encore été explorée. Dans cette thèse, nous présentons l'apprentissage automatique à la périphérie de notre réseau de communication considéré pour la réalisation d'expériences Internet immersives et sans friction. Pour réaliser l'intelligence périphérique, certaines ONU-BS/MPP sont équipées de serveurs MEC améliorés par IA. Ces serveurs s'appuient sur les capacités de calcul des cloudlets co-implantés à l'interface optique sans fil (voir Fig. R.4) pour prévoir les échantillons haptiques retardés dans le chemin de retour. La figure R.6 montre clairement la précision de prévision supérieure de notre schéma de prévision d'échantillons de bord (ESF) basé sur le perceptron multicouche (MLP) en matière d'erreur quadratique moyenne sur une large plage de valeurs deλ_{BKGD} pour scénarios de télé-opération locaux et non locaux, dans lesquels une erreur quadratique moyenne faible est réalisable dans le scénario précédent. En particulier, pour la télé-opération non locale, notre schéma ESF basé sur MLP réduit l'erreur quadratique moyenne d'environ 0.9 à 0.65 $\times 10^{-3}$, ce qui se traduit par

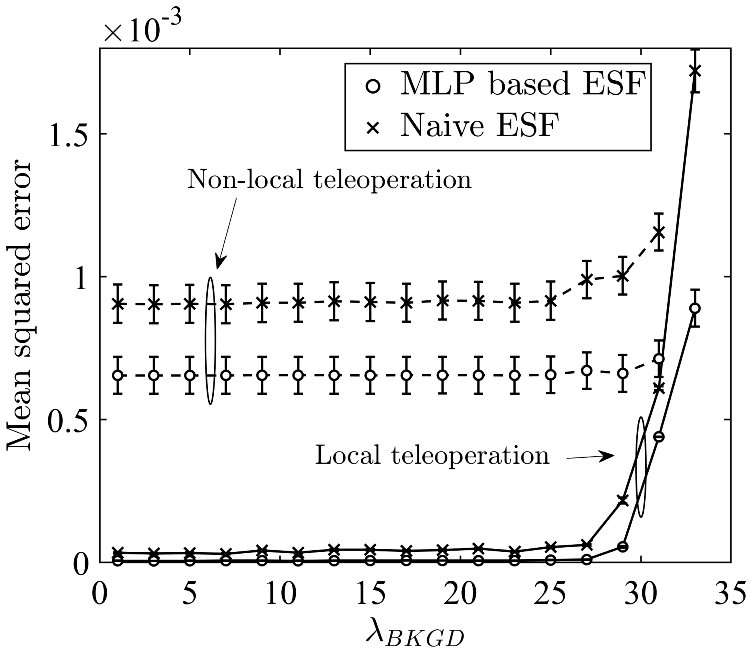


Figure R.6: Comparaison de la précision de prévision entre les schémas ESF naïfs et à base de MLP proposés pour la télé-opération locale et non locale sans codage en bande morte dans le chemin de retour($d_f = 0$).

une amélioration de 27.8%. Pour la télé-opération locale, il est possible de garder l'erreur quadratique moyenne proche de zéro entre 0.006 et 0.007 $\times 10^{-3}$ à une charge de trafic d'arrière-plan faible à moyenne λ_{BKGD}.

Contexte et conscience de soi pour la coordination des tâches

Ensuite, la connaissance du contexte est utilisée pour développer un algorithme de coordination de tâches centré sur un HART qui minimise le temps d'exécution des tâches physiques/numériques ainsi que les dépenses opérationnelles (OPEX) en répartissant la propriété des robots sur les utilisateurs mobiles. En outre, la conscience de soi est mise à profit pour améliorer les performances d'un robot donné en identifiant ses capacités ainsi que les exigences objectives par le biais d'une planification de mouvement optimale afin de minimiser sa consommation d'énergie ainsi que son temps de parcours. Le schéma d'allocation proposé, centré sur HART et tenant compte du contexte, pour les tâches physiques et numériques est utilisé pour coordonner l'automatisation et l'augmentation de co-activités homme-machine mutuellement bénéfiques dans une infrastructure Internet Tactile basée sur le FiWi. Les contributions de ce chapitre sont notamment les suivantes: (*i*) nous formulons un problème d'optimisation à objectifs multiples afin de minimiser le temps d'achèvement des tâches, la consommation d'énergie et OPEX pour l'attribution de tâches multi-robots dans Internet Tactile sur Réseaux améliorés par le FiWi, (*ii*) nous développons un algorithme de coordination de tâches centré sur HART tenant compte du contexte, qui minimise le temps de réalisation des tâches physiques/numériques, tout en portant une attention particulière à la réduction des OPEX en répartissant la propriété des robots sur des utilisateurs mobiles, (*iii*) nous proposons un algorithme de planification de mouvement optimal auto-conscient, qui s'exécute localement sur les robots mobiles, avec pour objectif de trouver le meilleur compromis entre temps de traversée et consommation d'énergie en s'appuyant sur la *conscience de soi locale* des robots mobiles pour identifier leurs limites et capacités respectives ainsi que les exigences objectives pour accomplir les tâches attribuées, et (*iv*) nous fournissons un cadre analytique pour le calcul du délai de connexion et la fiabilité da la connexion homme-robot, deux attributs clés de l'Internet Tactile.

Dans notre formulation du problème, nous supposons que les membres HART sont conscients de leurs objectifs, besoins, applications et contraintes respectifs. En outre, par le biais de la communication, ils établissent une prise de conscience collective du contexte dans le but de minimiser le temps d'exécution des tâches par les robots mobiles (MRs), qui peuvent être détenus par l'utilisateur ou par le réseau. Notre algorithme de coordination de tâches multi-robots vise à minimiser le temps d'achèvement de la tâche $T(\cdot)$ ainsi que la consommation énergétique $E(\cdot)$ et OPEX $C(\cdot)$ de l'exécution de tâches physique/numérique par les MRs. Clairement, $T(\cdot)$, $C(\cdot)$ et $E(\cdot)$ peuvent être des objectifs contradictoires, car minimiser $T(\cdot)$ et $E(\cdot)$ ne peut pas nécessairement minimiser $C(\cdot)$. La raison en est que, pour certaines tâches, la sélection de MR appartenant à un réseau peut réduire considérablement le temps d'exécution de la tâche, ce qui entraîne une augmentation des frais d'exploitation (OPEX) en raison de la tarification plus élevée des MRs appartenant au réseau par rapport à ceux des utilisateurs. Nous notons également que la consommation d'énergie d'un MR est fonction de ses paramètres locaux, par exemple ses paramètres de moteur et de mouvement, entre autres, qui peuvent de préférence ne pas être partagés par les MRs, car ils sont considérés comme des informations privées. De plus, le coordonnateur de tâches doit prendre des décisions sans connaître a priori les instants d'heure d'arrivée des tâches à venir, rendant ainsi impossible l'exploitation des méthodes d'optimisation conventionnelles pour obtenir la solution optimale du problème qui nous intéresse. Par conséquent, afin de trouver un compromis approprié entre les trois objectifs et de parvenir à une solution satisfaisante, nous hiérarchisons les objectifs du problème par ordre décroissant de $T(\cdot)$, $C(\cdot)$ et $E(\cdot)$. Plus spécifiquement, nous découplons le problème en deux sous-problèmes, à savoir la coordination de tâches multi-robots et la planification de mouvements, le premier visant à minimiser $T(\cdot)$ et $C(\cdot)$, tandis que le second minimise $E(\cdot)$.

Notre algorithme proposé de coordination de tâches multi-robots dynamiques (CADMRTC) tenant compte du contexte assigne la tâche donnée au MR le plus proche disponible appartenant à l'utilisateur, le cas échéant. Sinon, il essaie de trouver le MR le plus ancien disponible appartenant à l'utilisateur jusqu'à une échéance maximale donnée, $D \geq 0$ secondes, avant de retomber sur les MRs appartenant au réseau. Dans ce cas, la tâche est affectée au MR le plus proche disponible appartenant au réseau ou au plus ancien disponible, s'il n'y en a pas. Notez que notre schéma tenant compte du contexte donne la priorité aux MRs appartenant à l'utilisateur, réduisant ainsi considérablement les coûts d'exploitation. Il est intéressant de mentionner que notre objectif est de minimiser la consommation d'énergie du MR attribué en utilisant notre planification de mouvement auto-conscient proposée, qui vise à trouver le profil de vitesse optimal en énergie d'un MR pour un chemin donné à parcourir.

La figure R.7 illustre le ratio OPEX moyen par tâche par rapport au ratio de coût utilisateur/réseau, $r_{U2N} = \frac{\varphi_U}{varphi_N} \leq 1$. Fait intéressant, nous observons que si la pleine propriété de l'utilisateur (c'est-à-dire $\gamma_O = 100\%$) est toujours bénéfique pour les UMs en termes d'économies OPEX, la propriété partielle par l'utilisateur (c'est-à-dire $\gamma_O < 100\%$) ne l'est pas nécessairement. La figure R.7 montre que la propriété de l'utilisateur de $\gamma_O = 25\%$ est moins coûteuse que celle d'un réseau complet (c'est-à-dire que $\gamma_O = 0\%$) uniquement pour $r_{U2N} < 0.39$. En revanche, pour $r_{U2N} > 0,39$, les unités centrales sont confrontées à un OPEX inférieur par tâche avec une propriété réseau complète ($\gamma_O = 0\%$) par rapport à une propriété utilisateur partielle de $\gamma_O = 25\%$. La raison

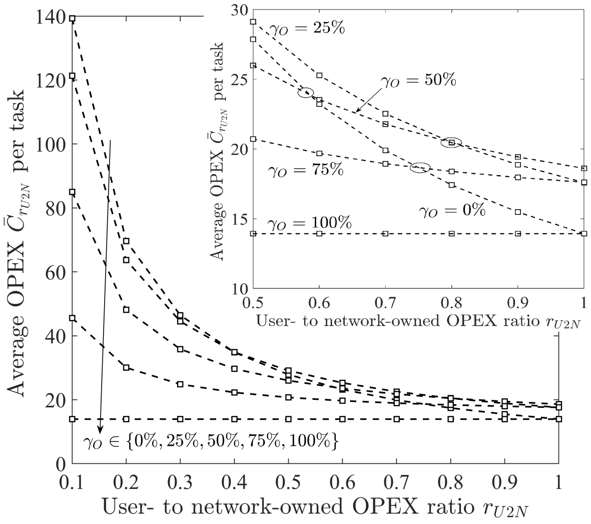


Figure R.7: Coût moyen, $\bar{C}$, par tâche exécutée par rapport au ratio OPEX de l'utilisateur sur le réseau r_{U2N} ($D = 0$ fixée).

en est que pour $\gamma_O = 0\%$, notre algorithme de coordination des tâches attribue des tâches aux MRs préférés de l'utilisateur, qui sont chargés de fournir le service à toutes les unités centrales de la zone. Ceci, à son tour, augmente la distance moyenne parcourue par les MRs appartenant à l'utilisateur, augmentant ainsi le temps de parcours moyen et la consommation d'énergie. Lorsque $\frac{\varphi_U}{\varphi_N}$ devient supérieur à 0.39, la propriété intégrale du réseau s'avère donc moins coûteuse. De plus, notez que pour une propriété partielle de l'utilisateur, $\gamma_O = 50\%$ et $\gamma_O = 75\%$ soit moins coûteuse que la propriété du réseau complet ($\gamma_O = 0\%$), r_{U2N} ne doit pas dépasser 0,58 et 0,73, respectivement (voir aussi Fig. R.7). De plus, nous observons que comme $\frac{\varphi_U}{\varphi_N} \rightarrow 0$, l'impact bénéfique de la propriété des utilisateurs sur l'épargne OPEX est plus prononcé, alors que pour $\frac{\varphi_U}{\varphi_N} \rightarrow 1$, la moyenne des OPEX $\bar{C}_{r_{U2N}}$ par tâche pour différentes valeurs de γ_O convergent vers celle de $\gamma_O = 100\%$. En effet, comme $\frac{\varphi_U}{\varphi_N} \rightarrow 1$ nous avons $\varphi_U \approx \varphi_N$, ainsi la propriété de l'utilisateur ne révèle pas un gain OPEX notable par rapport à la propriété du réseau à part entière.

Les résultats obtenus sur le front de Pareto en deux dimensions de notre algorithme CADMRTC proposé sont illustrés à la figure R.8, qui caractérise le compromis entre l'OPEX moyen par tâche et le temps moyen d'achèvement de la tâche. La Fig. R.8 révèle qu'aucun des résultats obtenus pour un γ_0 donné n'est dominant. Le décideur peut donc offrir un compromis souple entre les deux objectifs du problème en définissant de manière appropriée le délai d'attente D.

Planification et attribution de tâches de télé-opération à délai limité

Contrairement à leurs homologues totalement autonomes, les systèmes robotiques semi-autonomes reposent sur une assistance humaine de temps à autre via la télé-opération et/ou la télé-présence lorsqu'une expertise de domaine est nécessaire pour accomplir une tâche spécifique, permettant ainsi une approche de conception centrée sur HITL. Etant donné que ces robots auront besoin de demander une assistance humaine via la télé-opération/présence, ces demandes adressées aux opérateurs humains eux-mêmes constituent un problème difficile d'optimisation multi-critères avec

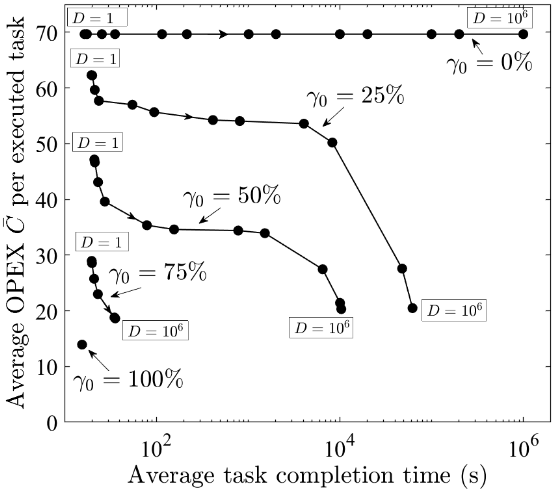


Figure R.8: Diagramme Pareto-front 2D de notre algorithme CADMRTC proposé pour différentes valeurs du facteur de dispersion de propriété γ_0 (la valeur de l'attente de l'échéance D augmente le long de la flèche indiquée sur chaque courbe).

l'objectif de minimiser le temps de réalisation moyen pondéré, le retard maximal et l'OPEX moyen par tâche. Dans ce chapitre, après avoir exposé en détail notre infrastructure de réseau bimodale FiWi considéré et son rôle dans la réalisation de la télé-opération dans l'Internet Tactile, nous formulons et résolvons le problème de la planification conjointe par ordre de priorité et de l'attribution de tâches de télé-opération soumises à des délais excessifs aux opérateurs afin de minimiser la durée moyenne pondérée des tâches, retard maximal, et le moyen OPEX par tâche. Les contributions de ce chapitre sont notamment les suivantes: (i) nous expliquons le rôle des réseaux améliorés par le FiWi en tant qu'infrastructures de communication sous-jacentes permettant de nouvelles applications Internet Tactile sensibles aux délais, dans le but de réaliser des télé-opérations locales et/ou non locale sur des réseaux améliorés FiWi, en exploitant les technologies Ethernet à faible coût axées sur les données (fibre optique et sans fil) en fronthaul et backhaul, (ii) nous définissons le problème de la planification conjointe prioritaire et de l'attribution de retard des tâches de télé-opération sur des opérateurs humains qualifiés disponibles, et après avoir formulé notre problème d'optimisation multi-objectifs, nous proposons notre algorithme appelé "ordonnancement prioritaire et attribution de tâches sensible au contexte" (CAPSTA) pour obtenir des résultats satisfaisants en faisant des compromis appropriés entre les objectifs contradictoires du problème, et (iii) nous développons notre cadre analytique pour estimer le temps de transmission de bout en bout des paquets de télé-opération locale et non locale sur les réseaux améliorés par le FiWi .

Nous considérons le problème d'assignation et d'ordonnancement conjoints de tâches de télé-opération à N retardées sur un nombre fixe M de HO, comme suit. Soit $\boldsymbol{\mathcal{M}} = \{O_1, O_2, ..., O_M\}$ et $\boldsymbol{\mathcal{J}} = \{J_1, J_2, ..., J_N\}$ dénotant l'ensemble des M HO disponibles et des N tâches attribuées, respectivement. Soit T_j le temps d'opération de la tâche $J_j \in \boldsymbol{\mathcal{J}}$. Chaque tâche $J_j \in \boldsymbol{\mathcal{J}}$ a une échéance D_j et est associée à un poids Ω_j. Les poids plus importants correspondent aux niveaux de priorité les plus élevés. Bien que les tâches soient censées être accomplies dans les délais impartis, tout retard encouru est soumis à une pénalité de coût. Nous considérons un scénario de planification *hors ligne*, dans lequel toutes les tâches sont disponibles à l'instant zéro et le sont continuellement par la suite. Chaque tâche ne peut être exécutée que par un seul opérateur humain à la fois et chaque opérateur humain ne peut exécuter qu'une seule tâche à la fois. Nous supposons également que la préemption n'est pas autorisée, ce qui signifie que les tâches ne peuvent pas être scindées. En effet, si les tâches étaient divisées et planifiées sur des périodes non continues, la préemption engendrerait un temps système supplémentaire de reconfiguration/configuration, ce qui est significatif lorsque la durée de configuration est non négligeable. De plus, nous supposons que $N \gg M$. Pour la tâche J_j, les heures de début et d'achèvement sont respectivement notées S_j et C_j. Une affectation/un calendrier réalisable spécifie quand et par quel opérateur humain une tâche donnée est exécutée. Avec un calendrier réalisable, on peut calculer le retard de la tâche J_j comme $\max\{0, C_j - D_j\}$. L'objectif est d'attribuer les tâches aux SH de manière à ce que les contraintes suivantes soient satisfaites: (1) pas plus d'une tâche n'est assignée à un HO à la fois, (2) aucune tâche n'est assignée à plus d'un HO, (3) les tâches ne sont pas pré-emptées, et (4) le délai moyen de transmission de bout en bout d'une télé-opération planifiée ne dépasse pas un seuil de délai donné.

Après avoir formulé notre problème multi-objectif d'ordonnancement et d'attribution de tâches de télé-opération, nous constatons que pour des moments de problème de grande taille de la formulation développée, les difficultés de calcul associées à la recherche de la solution satisfaisante augmentent considérablement. Par conséquent, afin de trouver un compromis approprié entre les objectifs contradictoires, nous proposons notre algorithme CAPSTA sensible au contexte. Lors de la conception de l'algorithme CAPSTA proposé, nous adoptons deux règles de tri, à la fois en affectation et en phases de planification, afin de privilégier les tâches hautement prioritaires dont les délais sont plus courts. Dans un premier temps, l'algorithme CAPSTA proposé vise à partitionner l'ensemble de tâches donné $\mathcal{J}$ en M sous-ensembles. À cette fin, notre politique de tri indique que les tâches données sont triées dans un ordre décroissant de $\frac{\Omega_j}{T_j}$. Ensuite, les tâches sont sélectionnées à partir de l'ensemble trié, puis attribuées aux HO d'une manière alternée. Nous notons cependant que l'affectation de la tâche J_n à l'opérateur humain O_m n'est valide que si les délais moyens estimés de bout en bout dans les chemins de commande et de retour satisfont aux contraintes de délai. Sinon, nous sélectionnons le HO qui correspond au délai moyen minimal de bout en bout avec la tâche J_n dans les chemins de commande et de retour. Cela résout le sous-problème d'affectation. Ensuite, l'algorithme proposé CAPSTA s'attaque au sous-problème d'ordonnancement pour les HO. À cette fin, parmi les tâches non planifiées, nous sélectionnons d'abord la tâche avec le montant minimum de $\frac{D_j}{\Omega_i}$, puis la planifions lorsque le HO devient disponible pour la première fois. De ce fait, les tâches ayant des poids plus importants et des délais plus courts sont privilégiées.

Nous comparons les performances de notre algorithme CAPSTA proposé avec un algorithme de référence *assignation aléatoire et ordonnancement* (RAS), dans lequel, pour une tâche donnée, un HO est sélectionné de manière aléatoire à partir du pool d'opérateurs disponibles. Premièrement, nous présentons le temps moyen pondéré d'exécution (AWCT) et le retard maximal par rapport au nombre total d'opérateurs humains disponibles, en M, sur les figures R.9 et R.10, respectivement. Nous observons dans la Fig. R.9 que l'augmentation de M entraîne une diminution exponentielle de AWCT dans les algorithmes RAS et CAPSTA proposé. Plus précisément, dans l'algorithme proposé CAPSTA, l'augmentation de M de 1 à 3 entraîne une réduction de 67%, tandis que l'augmentation de M de 3 à 5 ne donne qu'une réduction de 41% de la valeur AWTC. En outre, nous notons que l'algorithme CAPSTA proposé permet une réduction de 15 à 27% du AWTC par rapport à l'algorithme RAS. Bien que l'AWCT proposé soit moins performant que l'algorithme RAS, l'effet bénéfique de celui-ci est plus prononcé en termes de retard maximal, comme le montre la figure R.10. Nous observons que l'algorithme CAPSTA proposé permet une réduction de 49 à 56% du retard maximal. Plus précisément, pour $N = 300$, pour que le retard maximal soit inférieur à 25 minutes, un nombre total de 5 HOs est nécessaire dans l'algorithme RAS, alors que dans l'algorithme CAPSTA proposé, seuls 2 HOs suffisent pour obtenir le même résultat au niveau de performance.

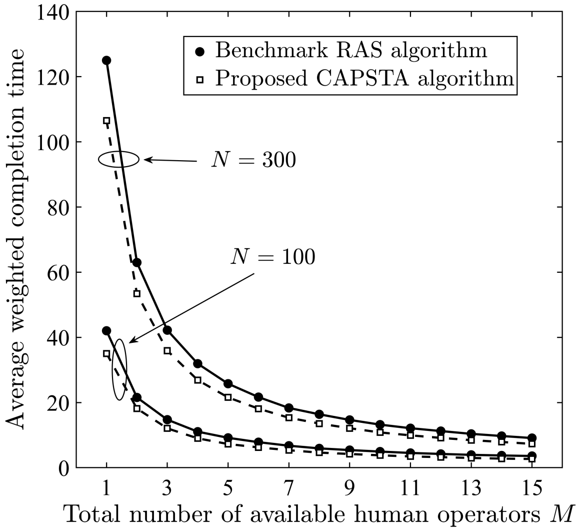


Figure R.9: Temps moyen pondéré d'exécution des tâches par rapport au nombre total d'opérateurs humains disponibles: M ($\alpha = 1$ fixée).

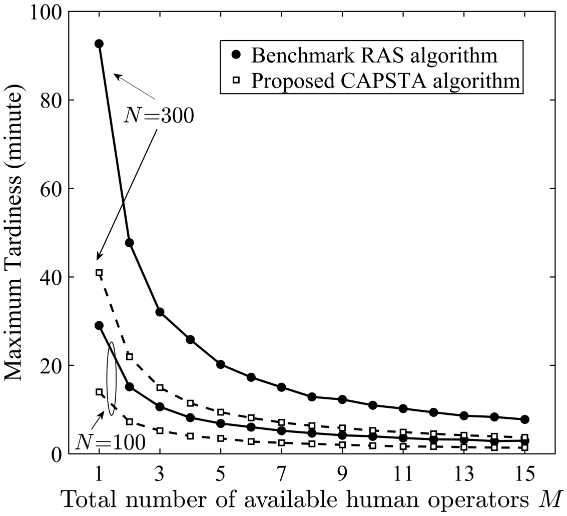


Figure R.10: Retard maximal des tâches par rapport au nombre total d'opérateurs humains disponibles M ($\alpha = 1$ fixée).

En outre, si le décideur préfère maintenir le retard maximal au-dessous de 10 minutes, le nombre d'opérateurs humains requis est respectivement de 5 et 12 dans les algorithmes CAPSTA proposé et RAS, ce qui permet de réaliser une économie notable en OPEX.

Déchargement coopératif des calculs dans les réseaux FiWi améliorés 4G HetNets

Pour faire face à la contradiction entre l'augmentation rapide des applications gourmandes en temps et à calcul intensif et les appareils mobiles intelligents dotés de ressources limitées, l'informatique en nuage mobile (MCC) a émergé pour réduire la charge de calcul des appareils mobiles et élargir leurs capacités en élargissant le concept d'informatique en nuage à l'environnement mobile par le déchargement complet et/ou partiel des calculs. Bien qu'un nuage (distant) conventionnel offre des capacités de stockage et de calcul élevées, il peut entraîner une latence importante en raison des communications, car généralement il est physiquement distant des utilisateurs mobiles. D'autre part, MEC peut offrir une latence réduite induite par la communication, mais elle peut également entraîner une latence de traitement excessive en raison de capacités de calcul limitées. Dans une vision plus large, les serveurs cloud et MEC distants peuvent ainsi coexister et se compléter, ce qui donne lieu à *déchargement coopératif de calcul*. En fait, le but ultime de MEC est d'obtenir un temps de réponse extrêmement bas, défini comme étant l'intervalle de temps entre le moment où une tâche est libérée d'un périphérique mobile et son traitement (localement ou à distance) et le résultat est reçu par le périphérique.

Dans cette thèse, nous examinons les gains de performances obtenus par le déchargement coopératif des calculs dans des réseaux HetNets améliorés par le FiWi compatible avec MEC, qui s'appuie non seulement sur les capacités de calcul des serveurs Edge/Cloud, mais également sur les ressources informatiques locales limitées du côté des périphériques. Plus spécifiquement, nous visons à concevoir une architecture MEC compatible FiWi améliorés HetNets à deux niveaux, dans laquelle les périphériques mobiles ainsi que les serveurs de périphérie délèguent de manière coopérative leurs tâches de calcul de manière à réduire le temps de réponse moyen. Nous prenons en compte les aspects cruciaux des limitations liées aux communications et au calcul dans notre approche de conception via une modélisation précise des serveurs fronthaul/backhaul ainsi que des serveurs Edge/Cloud, tout en accordant une attention particulière au processus de prise de décision entre utilisateurs mobiles et serveurs Edge en tant que serveurs de périphérie et le cloud distant.

Un autre aspect important de MEC est de faire face à la complexité supplémentaire qui peut survenir dans un tel scénario en s'appuyant entièrement ou partiellement sur les ressources informatiques locales limitées des utilisateurs mobiles lorsque celles-ci sont le plus nécessaires. La nature inhérente aux réseaux HetNets améliorés par le FiWi, qui est une conséquence directe de la mobilité des utilisateurs, implique d'exploiter une fonction qui ajuste en permanence les capacités de calcul locales des périphériques mobiles afin d'assurer une qualité d'expérience améliorée (QoE). Ceci peut être réalisé via une reconfiguration adaptative d'un utilisateur mobile en fonction de ses objectifs, de ses capacités et de ses contraintes, via une approche de conception communément appelée conscience de soi. Pour contribuer à cet effort, nous tirons parti de la conscience de soi des utilisateurs de téléphonie mobile en appliquant la technique de dimensionnement dynamique de la tension (DVS) pour soumettre les compromis de délai d'énergie appropriés à des contraintes d'énergie et de délai données.

La figure R.11 décrit l'architecture générique du réseau HetNets LTE-A amélioré par le FiWi considéré. Nous équipons certaines ONU-BS/MPP de serveurs MEC (ou simplement appelés *serveurs périphériques*) co-localisés à l'interface optique sans fil. Les unités centrales peuvent décharger entièrement ou une partie de leurs tâches de calcul entrantes sur des serveurs Edge à proximité. Outre les serveurs de périphérie, la terminaison OLT est équipée d'installations d'informatique en nuage, composées de plusieurs serveurs dédiés au traitement des tâches mobiles. Chaque UM utilise un planificateur de tâches qui décide de décharger une tâche sur un serveur Edge ou de l'exécuter localement dans son CPU. Nous modélisons le planificateur de tâches dans chaque UM par un système de mise en file d'attente, comme illustré à la Fig. R.12. Nous supposons que dans chaque appareil mobile, il y a deux serveurs, à savoir le processeur et l'interface sans fil (c'est-à-dire WiFi ou LTE-A). Le premier serveur est utilisé pour modéliser l'exécution de la tâche locale au niveau de

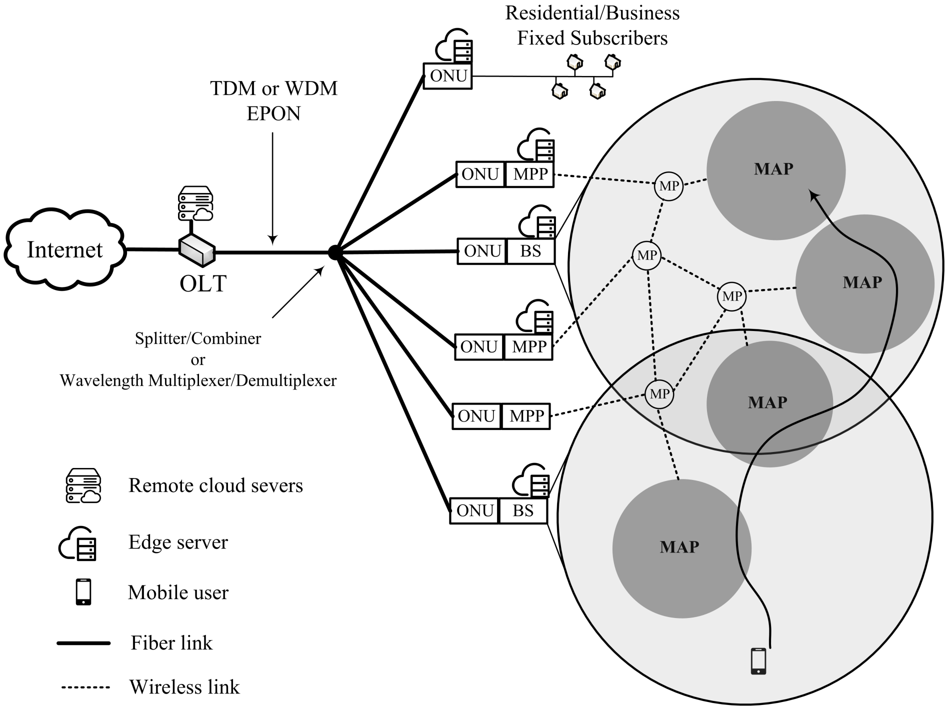


Figure R.11: Architecture générique du réseau HetNets LTE-A amélioré par le FiWi compatible avec MEC.

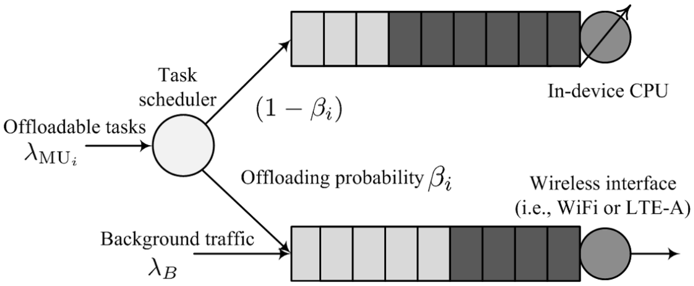
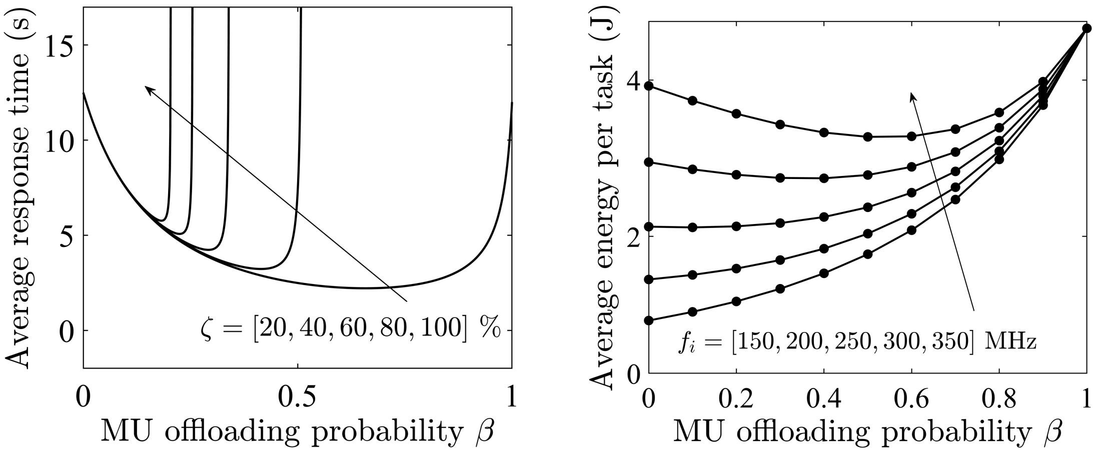


Figure R.12: Schéma du planificateur de tâches et du système de mise en file d'attente pour MU i, qui comprend deux files d'attente séparées desservies par un processeur local et une interface sans fil WiFi/LTE-A.

la CPU de la UM, tandis que le second est responsable du déchargement des tâches sur un serveur de périphérie situé à proximité. Nous supposons que les UM génèrent du trafic Poisson de fond au débit de paquets moyen λ_B (en paquets/seconde) (voir la Fig. R.12). Nous supposons également que les tâches arrivent au planificateur de UM i au tarif λ_{UM_i}. Le planificateur de tâches situé dans UM i prend sa décision en fonction de la valeur de la prétendue probabilité de déchargement, β_i, définie comme la probabilité qu'une tâche entrante soit déchargée sur le serveur de bord. Les tâches générées par UM i sont caractérisées par B_i^l et D_i^l, indiquant la taille moyenne des données d'entrée de calcul (par exemple, codes de programme et entrée paramètres) et le nombre moyen de cycles de processeur requis, respectivement. Les tâches de calcul sont supposées être atomiques et ne peuvent donc pas être divisées en sous-tâches. Nous supposons également que chaque serveur de périphérie est équipé d'un planificateur de tâches, qui décide s'il convient d'exécuter une tâche



(a)

(b)

Figure R.13: (a) Temps de réponse moyen vs probabilité de déchargement de UM β pour différentes valeurs de ζ ($\alpha = 0$ et f_i=150 MHz); (b) moyenne d'énergie par tâche vs probabilité de déchargement de UM β pour différentes valeurs de la fréquence d'horloge locale $f_i(\zeta = 20\%)$.

entrante ou de la décharger davantage sur le nuage distant. Comme pour les UM, une tâche arrivant sur le serveur de périphérie j est ensuite déchargée dans le nuage distant avec la probabilité α_j ou exécutée localement avec la probabilité $(1 - \alpha_j)$.

Selon notre analyse, une QoE améliorée n'est obtenue que lorsqu'un réglage optimal des probabilités de déchargement est effectué à la fois du côté périphérique et du côté serveur. Tout écart par rapport à ce paramètre optimal peut entraîner une dégradation des performances. En raison de la nature inhérente du statut du réseau, qui varie dans le temps et qui est une conséquence directe de la mobilité de l'utilisateur et des fluctuations du trafic, un tel réglage optimal peut ne pas être obtenu et maintenu facilement. Pour faire face à ce problème, nous donnons aux utilisateurs mobiles une conscience de soi qui leur permet de s'appuyer, le cas échéant, sur leurs ressources informatiques locales. Ce faisant, nous développons un cadre d'optimisation des critères pour permettre aux UMs d'utiliser leurs informations locales et de minimiser le temps de réponse ainsi que leur consommation d'énergie en ajustant de manière dynamique leur probabilité de déchargement ainsi que la fréquence d'horloge de CPU à l'aide de la technique DVS.

Premièrement, considérons le scénario d'informatique de périphérie avec α défini sur zéro. Les figures R.13(a) et (b) décrivent les performances en termes de retard en énergie du déchargement partiel assisté par MEC. Le temps de réponse moyen en fonction de la probabilité de déchargement β pour différentes valeurs de ζ est présenté à la Fig. R.13(a). Les résultats indiquent que le temps de réponse moyen est une fonction convexe de β. Pour $\zeta = 20\%$, définir $\beta = 0.66$ entraîne une réduction de 82% du temps de réponse moyen par rapport au schéma de calcul entièrement local (c'est-à-dire, $\beta = 0$). Nous notons que la valeur optimale de β dépend en grande partie de ζ. Plus précisément, lorsque ζ augmente, la valeur optimale de β diminue, comme le montre la Fig. R.13(a). La figure R.13(b) décrit la consommation d'énergie moyenne par tâche par rapport à β pour différentes valeurs de la fréquence d'horloge du processeur f_i. Les courbes du bas des figures R.13(a) et (b) mettent en évidence le compromis qu'une UM peut faire entre le temps de réponse moyen et l'énergie par tâche pour $\zeta = 20\%$. Nous observons également d'après la figure R.13(b) que, pour

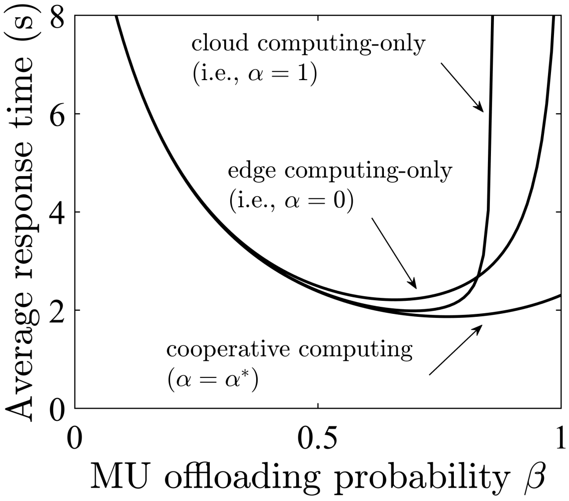


Figure R.14: Comparaison des performances en termes de temps de réponse moyen de edge uniquement, en cloud uniquement et l'informatique coopérative ($\zeta = 20\%$).

des valeurs plus grandes de f_i, le déchargement partiel réduit non seulement le temps de réponse moyen, mais il aide également les UMs à réduire leur consommation d'énergie.

Ensuite, nous examinons les gains de performances obtenus grâce à notre informatique coopérative trilatérale dispositif-périphérique-nuage (device-edge-cloud) proposé. La figure R.14 illustre le temps de réponse moyen par rapport à β pour les trois scénarios suivants: (i) uniquement en périphérie ($\alpha = 0$), (ii) en nuage uniquement ($\alpha = 1$), et (iii) l'informatique coopérative ($\alpha = \alpha^*$), où α^* indique la valeur optimale de α définie par le réseau pour minimiser le temps moyen d'exécution des tâches des serveurs de périphérie. La figure R.14 montre que le schéma de calcul coopératif proposé offre de meilleures performances en terme de retard que le schéma à bord seul ou en nuage, en particulier pour $\beta > 0.64$. Alors que les schémas Edge et Cloud uniquement peuvent tous deux prendre un temps de réponse plus long en raison d'un délai de mise en file d'attente excessif pour des valeurs élevées de β, la coopération trilatérale entre le processeur, le serveur Edge et le cloud distant réduit le temps de réponse de définir β et α sur leurs valeurs optimales (voir la courbe inférieure de la Fig. R.14).

Conclusion

L'Internet a constamment évolué de l'Internet mobile dominé par le trafic H2H vers l'IdO émergent avec ses communications M2M sous-jacentes. L'avènement de la robotique avancée, associé aux infrastructures de réseau émergentes ultra-réactives, permettra de transmettre la modalité tactile (également appelée sensation haptique) en plus du trafic traditionnel triple play (voix, vidéo et données) sous le terme communément appelé Internet Tactile. L'IdO, sans aucune implication humaine dans les communications machine-à-machine (M2M) sous-jacentes, est utile pour l'automatisation de processus industriels et autres processus centrés sur les machines, tout en préservant largement l'humain. En revanche, l'Internet Tactile permet de piloter et de contrôler de manière tactile non seulement des objets virtuels, mais aussi des objets réels via des robots télé-opérés, sera centré sur les communications humain-à-robot/-machine (H2R/M), appelant ainsi une approche centrée sur l'homme.

Cette thèse de doctorat est construite sur les réseaux HetNets LTE-A améliorés par le FiWi, en tant qu'infrastructures de réseau sous-jacentes prometteuses sur lesquelles l'Internet Tactile émergent devrait s'appuyer. En particulier, nous avons étudié différents aspects de l'Internet Tactile émergent et présenté des informations techniques approfondies sur la réalisation de réseaux améliorés de télé-opération centrée sur la technologie HITL, notamment la modélisation du trafic haptique à base de traces, la prévision d'échantillons haptiques, la coordination des tâches via la connaissance du contexte et de soi, l'attribution de tâches de télé-opération et le déchargement informatique coopératif. Nous nous sommes concentrés sur l'Internet Tactile en train de devenir l'une des applications les plus intéressantes à faible temps de latence pour la création de nouvelles expériences immersives. Nous avons passé en revue les principes de conception centrés sur HITL qui ajoutent une nouvelle dimension à l'interaction homme-machine via Internet et distinguent l'Internet Tactile de l'IdO plus centré sur la machine. Exploitant la perception humaine de l'haptique pour réduire le débit de paquets haptiques au moyen du codage en bande morte, nous avons déduit modèles de trafic haptique à partir d'expériences de télé-opération. Notre analyse des traces haptiques a montré que supposer que le trafic Internet Tactile était distribué par Pareto n'était pas valable pour le trafic analysé, tout en supposant qu'il s'agisse d'un trafic de Poisson, n'était valable que dans un cas particulier. En général, nous avons observé que les chemins de commande et de rétroaction des systèmes de télé-opération peuvent être conjointement modélisés par la distribution de Pareto généralisée, la distribution gamma ou les distributions de temps inter-arrivée de paquet déterministe, en fonction de la valeur donnée des paramètres de bande morte respectifs. Nous avons ensuite utilisé l'apprentissage automatique pour implémenter un prévisionniste d'échantillons multi-échantillons, capable de dissocier la rétroaction haptique de l'impact de retards de propagation importants. Cela permet aux humains de percevoir des environnements de tâches distants dans le temps avec une granularité de 1 ms. Nous avons ensuite étudié les performances de l'allocation de tâches multi-robots centrée sur HART, auto-consciente et contextuelle, sur des infrastructures Internet Tactile basées sur FiWi. Nous avons mis en lumière à quel moment, de quelle manière et dans quelles circonstances la propriété des MRs par les utilisateurs devenait bénéfique en termes de OPEX par tâche exécutée. En outre, nous avons évalué les performances de notre algorithme CADMRTC proposé en termes de temps moyen d'achèvement des tâches, d'OPEX par tâche exécutée et de rapport entre le nombre de tâches exécutées par les MRs propriétaires et le nombre total de tâches. Nous avons ensuite examiné les performances de notre algorithme CAPSTA proposé pour résoudre l'attribution prioritaire et la planification des tâches de télé-opération à délai dans les infrastructures de Internet Tactile amélioré par le FiWi. Les résultats obtenus montrent que l'algorithme proposé réduit le temps moyen d'achèvement des tâches pondéré, le retard maximal et l'OPEX moyen par rapport à l'algorithme de référence RSA.

Finalement, nous avons étudié le déchargement coopératif des calculs dans réseaux HetNets LTE-A améliorés par le FiWi compatible avec MEC du point de vue de la conception de l'architecture de réseau et du mécanisme de déchargement. Outre la conception de réseaux HetNets LTE-A améliorés par le FiWi compatible avec MEC à faible temps de latence, nous avons présenté une stratégie de déchargement simple mais efficace qui exploite la coopération trilatérale entre périphériques, serveurs de périphérie et cloud distant. Nous avons développé un cadre analytique pour estimer le temps de réponse moyen et la consommation d'énergie des utilisateurs mobiles dans une infrastructure de réseau compatible MEC basée sur FiWi. Nos résultats démontrent les performances supérieures du schéma de calcul coopératif proposé par rapport aux schémas de type Edge ou Cloud. De plus, nous avons montré qu'en définissant de manière optimale les probabilités de déchargement, les UMs pouvaient réduire le temps de réponse moyen jusqu'à 81%. Afin de faire face à la complexité engendrée, nous avons également conçu un mécanisme basé sur l'auto-organisation, qui permet à un UM, utilisant des informations locales, de faire des compromis de délai d'énergie appropriés et de minimiser conjointement le temps d'exécution moyen et la consommation d'énergie en ajustant de manière dynamique la probabilité de déchargement et sa fréquence d'horloge CPU locale en utilisant la technique DVS.

List of Acronyms

AI	Artificial Intelligence
ANN	Artificial Neural Network
AR	Augmented Reality
ART	Audi Robotic Telepresence
AWCT	Average Weighted Completion Time
BS	Base Station
CADMRTC	Context-Aware Dynamic Multi-Robot Task Coordination
CAPSTA	Context-Aware Prioritized Scheduling and Task Allocation
CCDF	Complementary Cumulative Distribution Function
CDF	Cumulative Distribution Function
CPS	Cyber-Physical Systems
CPU	Central Processing Unit
CNRS	Centre National de la Recherche Scientifique
co-DBA	cooperative Dynamic Bandwidth Allocation
DC	Direct Current
DCF	Distributed Coordination Function
DFR	Decreasing Failure Rate
DoF	Degrees of Freedom
DVS	Dynamic Voltage Scaling
EPON	Ethernet Passive Optical Network
ETSI	European Telecommunications Standards Institute
ESF	Edge Sample Forecast
FiWi	Fiber-Wireless
FRF	Failure Rate Function
GP	Generalized Preto
H2H	Human-to-Human
H2M	Human-to-Machine
H2R	Human-to-Robot
HABA/MABA	Humans-Are-Better-At/Machines-Are-Better-At
HART	Human-Agent-Robot Teamwork
HCCA	Hybrid Coordination function controlled Channel Access
HetNets	Heterogenous Networks
HITL	Human-In-The-Loop
HO	Human Operator
HSI	Human System Interface
IFR	Increasing Failure Rate
IoT	Internet of Things

IP	Internet Protocol
JND	Just Noticeable Difference
LTE-A	LTE-Advanced
M2M	Machine-to-Machine
MAC	Medium Access Control
MAP	Mesh Access Point
MCC	Mobile Cloud Computing
MEC	Multi-access Edge Computing
MLE	Maximum Likelihood Estimation
MLP	Multi-Layer Perceptron
MP	Mesh Point
MPCP	Multi-Point Control Protocol
MPP	Mesh Portal Point
MR	Mobile Robot
MU	Mobile User
NFV	Network Function Virtualization
NG-PON	Next-Generation Passive Optical Network
NR	New Radio
OLT	Optical Line Terminal
ONU	Optical Network Unit
OPEX	Operational Expenditures
PDF	Probability Distribution Function
PON	Passive Optical Networks
QoE	Quality-of-Experience
QoS	Quality-of-Service
RAS	Random Assignment and Scheduling
RTP	Real-time Transport Protocol
SAOMP	Self-Aware Optimal Robot Motion Planning
SDN	Software Defined Networking
SDON	Software Defined Optical Networks
TDM	Time Division Multiplexing
TOR	Teleoperator Robot
UDP	User Datagram Protocol
URLLC	Ultra-Reliable Low-Latency Communications
WDM	Wavelength Division Multiplexing
WLAN	Wireless Local Area Network

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background and Motivation

While the commercial exploitation of the mobile Internet is enabling users to exchange traditional triple-play (i.e., audio, video, and data) traffic, the emerging *Tactile Internet* envisages to realize *haptic communications*, enabling users to not only see and hear but also touch and manipulate remote physical and/or virtual objects through the Internet [1] [2]. The Tactile Internet, which is driven by recent advancements in computerization, automation, and robotization, is expected to significantly augment human-machine interaction, thereby converting today's content delivery networks into skillset/labour delivery networks [3], [4], [5]. The Tactile Internet holds promise to create new entrepreneurial opportunities and jobs, which are expected to have a profound socioeconomic impact on almost every segment of our everyday life with use cases ranging from augmented/virtual reality (AR/VR) and autonomous driving to healthcare and smart grid. Many of these industry verticals require very low latency and ultra-high reliability for realizing ultra-responsive interactive applications such as bilateral teleoperation/telepresence. Note, however, that some use cases which do not necessarily require mobility all the time can be realized over fixed broadband networks. This suggests that future cellular networks need to be fully converged networks, allowing for a flexible selection of different fixed and mobile access technologies while sharing core network functionalities [6].

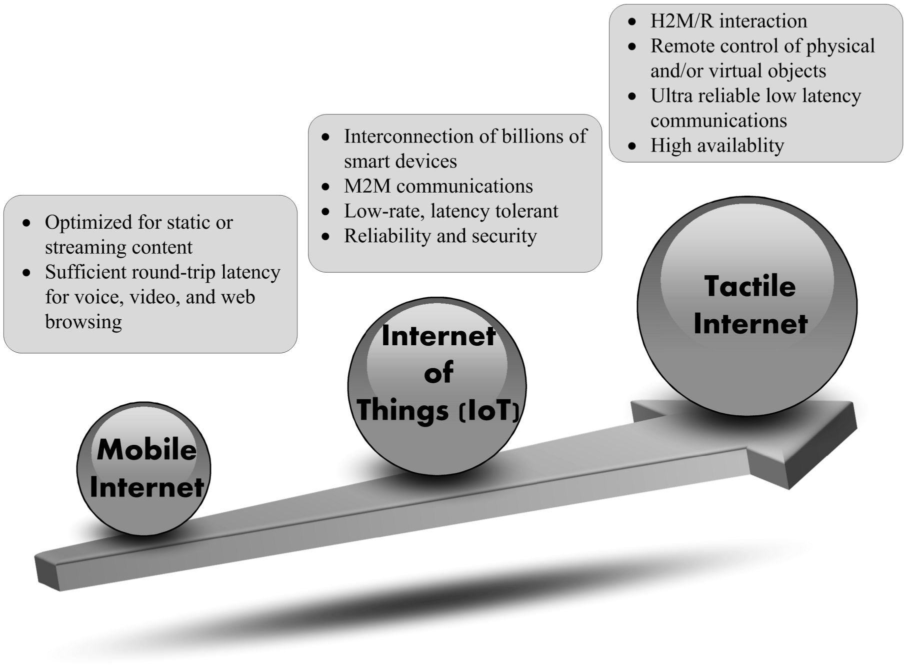

H2M/R interaction
Remote control of physical and/or virtual objects
Ultra reliable low latency communications
High availablity
Interconnection of billions of smart devices
M2M communications
Low-rate, latency tolerant
Reliability and security
Optimized for static or streaming content
Sufficient round-trip latency for voice, video, and web browsing
Tactile Internet
Internet of Things (IoT)
Mobile Internet

Figure 1.1 – Revolutionary leap of the Tactile Internet in compliance with ITU-T Technology Watch Report.

Interactive systems, including in particular AR/VR and teleoperation, demand an ultra low round-trip latency of 1-10 ms together with high reliability. The high availability and security, ultra-fast and highly-reliable response times, and carrier-grade reliability of the Tactile Internet will add a new dimension to interaction of humans with machines/robots. To gain a more profound understanding of the Tactile Internet, it may be helpful to compare it to the emerging Internet of Things (IoT) and 5G mobile networks. While the concept of IoT is far from novel and goes back to 1995, it is only recently that we are experiencing a rapidly increasing growth of interest in IoT from both industry and academia. Figure 1.1 depicts the revolutionary leap of the Tactile Internet in compliance with the ITU-T[1] Technology Watch Report on the Tactile Internet [7]. While the ultra-fast response time and carrier grade reliability of the Tactile Internet will add a new dimension to human-machine interaction, emerging 5G networks have to handle an unprecedented growth of mobile data traffic as well as an enormous volume of data from smart sensors and actuators, the empowering elements of the IoT. It is evident that unlike the previous four generations, future mobile networks will lead to a seamless integration of WiFi and cellular technologies and standards, paving the way towards realizing so-called heterogenous networks (HetNets) based on small-cell technologies, which mandates the need for addressing the backhaul bottleneck challenge.

[1] International Telecommunication Union - Telecommunication Standardization Sector (ITU-T)

The subtle difference between the Tactile Internet and IoT may be best expressed in terms of underlying communications paradigms and enabling end-devices. The Tactile Internet involves the inherent human-in-the-loop (HITL) nature of human-to-machine interaction, whereas the IoT is centered around autonomous machine-to-machine (M2M) communications without any interaction with humans. The Tactile Internet relies on human-to-machine/robot (H2M/R) interaction and thus allows for a human-centric design approach towards creating novel immersive experiences, expanding humans' capabilities through the Internet. A popular misinterpretation about robotics is that intelligent robotic systems will eventually substitute humans in one job after another. This argument may be true for some jobs, but we note that even though advanced robotics can be deployed to automate certain jobs, its greater potential, yet to be further explored, is to complement and augment human capabilities. New jobs and innovative business opportunities that arise from human-machine symbiosis emerge in the so-called *missing middle* that refers to the new ways that have to bridge the gap between human-only and machine-only activities. This paves the way toward the so-called *third wave of business transformation*, which will be centred around human+machine hybrid activities [8]. An important requirement of developing the missing middle is to fully understand how humans can help machines and how machines can help humans. A recent example of identifying the relative strengths of humans and machines and leveraging them to fill the missing middle can be found at automobile manufacturer Audi. Having deployed a fleet of Audi Robotic Telepresence (ART) systems, Audi has set forth employee augmentation that not only helps train technicians in diagnostics and repair, but also accelerates delivery of service to customers [9].

By augmenting traditional audiovisual and data communications by the haptic modality[2], touching, sensing, and physically interacting with remote objects becomes a reality. This, in turn, substantially improves human-to-human (H2H) as well as H2R interaction. For illustration, Fig. 1.2 depicts a typical teleoperation system based on bidirectional haptic communications between a human operator (HO) at one end and a teleoperator robot (TOR) at the other. The ultimate goal of a bilateral teleoperation system is to provide the HO with the impression of feeling immersed and being present in the remote environment. This can be achieved by providing the HO with multi-modal sensory feedback including visual, auditory, and haptic signals [11]. In a teleoperation

[2] The term "*haptics*" refers to both kinesthetic perception (information of forces, torques, position, velocity, etc., which are sensed by the muscles, joints, and tendons of a human body) and tactile perception (information of surface texture, friction, etc. sensed by different types of mechanoreceptors in the skin) [10].

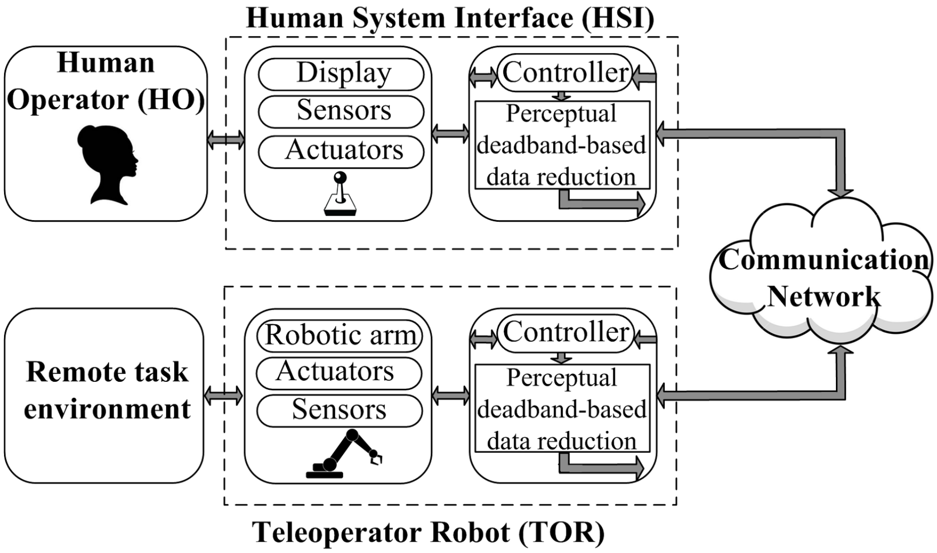


Figure 1.2 – Teleoperation system based on bidirectional haptic communications between human operator (HO) and teleoperator robot (TOR) in a remote task environment.

system, the term degrees-of-freedom (DoF)[3] refers to the number of independent coordinates required to completely specify and control/steer the position, orientation, and velocity of the TOR. Further, a local human system interface (HSI) device is used to display haptic interaction with the remote TOR to the HO. The local control loops on both ends of the teleoperation system ensure the stability and tracking performance of the HSI and TOR. Furthermore, as shown in Fig. 1.2, perceptual deadband based data reduction may be deployed as a lossy compression technique by exploiting the fact that human end-users are not able to discriminate relatively small changes in haptic stimuli. In this thesis, we particularly focus on the communication network aspects of teleoperation, paying particular attention to the notion of average end-to-end delay, given its importance in the Tactile Internet. The average end-to-end delay induced by the communication network is the time elapsed between the time instant a haptic packet arrives at the medium access control (MAC) queue of a given source HO/TOR and the time instant when it is successfully received by the corresponding destination TOR/HO.

The discussion above indicates that improving quality-of-experience (QoE) for immersive bilateral teleoperation applications mandates the need for ultra-reliable low-latency communications (URLLC), which not only provides a low average end-to-end latency, but also ensures an upper-bound delay experienced by haptic packets. From a communication network perspective, ef-

[3] Currently available teleoperation systems range from 1-DoF to >20-DoF TORs. For instance, a 6-DoF TOR allows for both translational motion (in 3D space) via force and rotational motion (pitch, yaw, and roll) via torque.

forts towards realizing an immersive ultra-responsive teleoperation experience have been pursued in the following domains: (*i*) communication network and (*ii*) intelligent edge-computing/-processing. First, we focus on the communication network for networked teleoperation coexistent with conventional H2H and M2M traffic. Note that delay requirements of HITL-centric bilateral teleoperation systems range from 1 ms to hundreds of milliseconds, depending on the application scenarios and dynamicity of the remote environment. If the remote environment is less dynamic, the interaction between the HO and TOR is increased. Conversely, highly dynamic environments may place high demands for ultra-fast response times of as low as 1 ms. Current cellular networks (e.g., 3G or 4G LTE-A) miss this target by at least one order of magnitude. End-to-end latency measurements in an LTE network in a dense urban area for a low-mobility scenario with a proprietary application running on an Android smartphone showed an average end-to-end delay of roughly 47 ms and 54 ms for low and high cell load scenarios [12]. More recently, the authors of [6] have reported that the achieved average round-trip delay of 3G and 4G networks are 63 and 53 ms, respectively, according to the traces released by UK regulator *Ofcom*. According to carrier WiFi vendor Aptilo Networks, despite the ongoing competition between LTE and WiFi, the two technologies are in fact complementary, as the latency being roughly ten times less in WiFi leads to a higher user experience, whereas LTE provides long-range outdoor coverage (see http://www.aptilo.com for further details).

With the wide deployment of passive optical networks (PONs) providing high capacity and reliability and wireless networks offering ubiquitous and flexible connectivity, interest has been growing in bimodal fiber-wireless (FiWi) networks that leverage the complementary benefits of optical fiber and wireless technologies. FiWi enhanced LTE-A HetNets represent a compelling solution to enable 4G cellular networks to meet the key requirements of low-latency and high-availability [13]. Recently, the authors of [13] have evaluated the maximum aggregate throughput, offloading efficiency, and delay performance of FiWi enhanced LTE-A HetNets and have shown that via WiFi offloading and fiber backhaul sharing, an ultra-low latency of 1-10 millisecond and highly reliable network connectivity can be achieved, especially at low to moderate traffic loads.

Next, we turn our attention to intelligent methods to compensate for the communication induced latency. To bridge the gap between the rapid increase of computation-intensive, delay-sensitive applications (e.g., Tactile Internet, AR/VR, and interactive gaming) and resource-limited smart mobile devices, mobile cloud computing (MCC) has emerged to reduce the computational burden of mobile devices and widen their capabilities by extending the notion of cloud computing to the mobile environment via computation offloading. MCC allows mobile devices to benefit from powerful computing resources to save their battery power and accelerate task execution, though it raises several technical challenges due to additional communication overhead and poor reliability that remote computation offloading may introduce. To overcome these limitations, mobile edge computing has recently emerged to provide cloud computing capabilities at the edge of access networks, leveraging the physical proximity of edge servers and mobile users to achieve a reduced communication latency and increased reliability. More recently, the European Telecommunications Standards Institute (ETSI) has dropped the word "mobile" and introduced the term *multi-access edge computing* (MEC) in order to broaden its applicability to heterogeneous networks, including WiFi and fixed access technologies (e.g., fiber) [14]. There is now a growing interest among industry players in extending the cloud to decentralized levels of self-managed entities. This trend toward decentralization has led to the new paradigm of MEC, in which computing and storage resources, variously referred to as cloudlets, mircro datacenters, or fog nodes, are placed at the Internet's edge in proximity to wireless end devices in order to achieve low end-to-end latency.

1.2 Tactile Internet: Prior Art and Open Challenges

In this section, we review prior research work related to the Tactile Internet and, after classifying it into four separate yet interdependent categories, we discuss each one in greater detail. The main branches of our classification are URLLC, multi-robot task allocation, HITL-centric teleoperation, and MEC.

1.2.1 URLLC

For a comprehensive and up-to-date survey on methodologies and technologies for enabling URLLC infrastructures for haptic communications, we refer the interested reader to [10] and the references therein. Ref. [15] provided an overview of communication features and capabilities of 5G and its role as a distributed computing platform to support Tactile Internet services. Recently, there have been numerous research efforts in the area of 5G enabled Tactile Internet. The authors of [15] provided an overview of communication features and capabilities of the 5G and its role as a distributed computing platform to address the Tactile Internet services. The authors of [16] proposed a resource allocation framework in a single-cell sparse code multiple access wireless network with multiple users with the objective of maximizing throughput subject to a given transmit power and delay constraint for haptic users. In [17], a proactive packet dropping mechanism was proposed for time division duplexing cellular systems, which, together with optimizing the queue state information and channel state information dependent transmission policy, was proven to satisfy given quality-of-service requirements with finite transmit power. To achieve low per-packet in-order delivery delays for 5G and Tactile Internet applications, the authors of [18] investigated multipath transport protocols across cellular and/or public WiFi networks. More recently, the authors of [19] studied the Tactile Internet from a physical layer and medium-access control perspective and then presented novel system design of URLLC for new radio (NR) LTE technologies. We note, however, that the above mentioned works focused on designing sophisticated channel access mechanisms for cellular networks only and they considered neither backward compatibility issues nor the widely-deployed WiFi technology.

In order to meet the stringent latency and reliability requirements of the Tactile Internet in a more efficient manner, network operators will rely on softwarization and virtualization via the concepts of software-defined networking (SDN) and network function virtualization (NFV). While softwarization allows for timely and cost-efficient deployment of new services, virtualization will be instrumental in service migration, upgrade, and scalability for mobile users. In this context, SDN enables to exploit the so-called method of *network slicing*, which allows for allocating network resources more efficiently to satisfy the often heterogenous QoS requirements of different applications. For instance, the authors of [20] presented a game theory based flexible dynamic network slicing strategy for Tactile Internet applications, where incoming traffic from users can be temporarily offloaded from operator-provided networks to user-provided networks, if needed. The authors of [21] developed an SDN-based low-latency management framework for distributed service function chains. A softwarized 5G architecture for end-to-end reliability of the mission-critical traffic was proposed in [22].

Clearly, a cost-efficient realization of the URLLC for the Tactile Internet requires leveraging not only the cellular networks but also widely deployed, low-cost Ethernet-based technologies, e.g., WiFi. Toward this end, the authors of [23] investigated a time and wavelength division multiplexing passive optical LAN and demonstrated the benefits of using a predictive resource allocation algorithm for future Tactile Internet based e-health applications. Note, however, that the main focus of this study was only on the optical backhaul without considering any wireless access mechanism. In contrast, [24] studied downlink transmission mechanisms for SmartBANs based on IEEE 802.15.6 technology. In [25], the authors developed an analytical framework using an M/G/1 queueing model for estimating the wireless transmission latency from tactile body-worn devices to the wireless access point, whereby the hybrid coordination function controlled channel access (HCCA) MAC protocol is used. The authors of [26] took a different approach and focused on the latency and reliability requirements of Tactile Internet based cooperative automated driving, relying on recent advancements in vehicular networking and automated driving. Note that backhaul was not discussed in these studies. Further, none of these works presented a unified delay analysis of teleoperation over a bimodal FiWi based Tactile Internet networking infrastructure.

1.2.2 Multi-robot task allocation

Clearly, while URLCC is necessary to meet the very low latency and ultra-high reliability requirements of the Tactile Internet, it does not address the proper task assignment nor does it provide suitable mechanisms to orchestrate the mutually beneficial cooperation of humans and machines. Beside URLLC networking infrastructures, which lay the groundwork for the Tactile Internet, another important aspect and key challenge of the Tactile Internet vision is related to economic considerations, addressing the problem of assignment of tasks in a networked multi-robot system. Ref. [27] presented an overview of open research challenges in cloud robotics and explored the potential benefits of big data, cloud computing, collective robot learning, and human computation in cloud-based robotic and automation systems. The authors of [28] presented a comprehensive comparison between stand-alone, networked, and cloud robotic systems and then put forward the idea of sharing the processing resources across a group of individual robots in a cloud-assisted cooperative robotic scenario.

Despite an existing great body of literature on development of advanced intelligent stand-alone robots, coordination of (tele)robots in a system of networked robots has received only little attention. In this context, a task allocation strategy that combines suitable host robot selection and computation task offloading onto collaborative nodes was presented in [29]. To mitigate failures during task execution, a neighboring robot-assisted failure reporting mechanism was introduced in [30]. We note that both [29] and [30] addressed a static task allocation problem, where the task coordinator has a priori knowledge of the given task demands, robot locations, and failure probability. A hierarchical auction-based resource allocation scheme for cloud robotics systems in ad-hoc networks was presented in [31]. The authors of [32] aimed at modeling and solving the problem of mobile robot task planning by means of mixed-integer programming and constraint programming. None of the existing studies in multi-robot task coordination addresses both self- and context-awareness in a unified manner.

1.2.3 HITL-centric teleoperation

After designing an URLLC networking infrastructure that lays the groundwork for the Tactile Internet, the next important aspect and key challenge is enhancing the quality-of-experience/immersion of users/human operators, while taking into account economic considerations. Typical bilateral teleoperation systems provide the human operator with haptic feedback, which can be categorized as kinesthetic force or cutaneous tactile feedback. Kinesthetic feedback provides forces and torques perceived by a human operator's hand/joints, whereas tactile feedback provides stimulation to her skin. In such systems, haptic feedback is instrumental for precise manipulation of the remote environment. Unlike the transmission of audio and video signals, in bilateral teleoperation systems haptic information are exchanged bidirectionally over the underlying networking infrastructure. It involves a human operator on one end and closes a global control loop between her and the actuators/teleoperators on the other. In such systems, even minor communication induced time delays and packet losses may destabilize the haptic communications system. For a comprehensive survey on methodologies for stabilizing bilateral teleoperation systems over delayed communications networks we refer the interested reader to [10] and the references therein.

A true immersion into the remote environment becomes possible only when the human operator is provided with enough multimodal sensory feedback, including in particular the sense of touch. This is realized via the so-called *robotic embodiment*. Toward this end, the authors of [33] studied the feasibility of displaying interaction force information to the human operator through fingerpad tactile skin deformation feedback. In [34], the authors explored the embodiment enhancement of teleoperated industrial tasks by means of employing AR feedback to the human operator and showed that full AR overlay can lead to an improved task execution in terms of accuracy as well as completion time. The authors of [35] introduced the vision of the Tactile Robots (i.e., avatars) as the next evolutionary step in rapidly developing robotic systems and elaborated on the enabling technologies for embodiment of the Tactile Internet via smart wearables.

To the best of the author's knowledge, none of the existing studies in machine scheduling addresses the multi-objective optimization of joint prioritized scheduling and assignment of delay-constrained teleoperation tasks to human operators, taking into account economic considerations from a system point of view. The joint assignment and scheduling of delay-constrained teleoperation tasks on multiple human operators can be viewed as a variation of the well-studied parallel machine scheduling problem [36]. Mapping the problem of assignment and scheduling of delay-constrained teleoperation tasks to the parallel machine scheduling problem allows leveraging on the existing scheduling strategies, algorithms, and mathematical derivations for developing and evaluating new solutions.

1.2.4 MEC

While computation offloading for mobile computing systems has been around for almost a decade, the edge/fog computing paradigm has emerged only recently. The authors of [37] proposed a novel cloudlet cellular network architecture to enable mobile users to offload their computation workload to nearby cloudlets and then designed a latency-aware workload offloading strategy to allocate the offloaded workload to suitable cloudlets. The work in [37] was extended in [38] by incorporating a two-tier hierarchical architecture for cloudlets. In [39], a multiobjective optimization problem for fog computing systems was formulated with the joint objective of minimizing energy consumption, execution delay, and payment cost by finding the optimal offloading probability and transmit power of mobile devices. Note, however, that in [39] the cooperation between edge servers and the remote cloud is limited to the case when the edge servers are overloaded with the offloaded tasks from mobile devices, thus not fully reaping the benefits of the two-tier hierarchical edge computing architecture. The authors of [40] aimed to minimize the response time in a scenario with two MEC servers by focusing on both computation and communication latencies through virtual machine migration and transmission power control, respectively. The work in [40] was extended in [41] to further reduce the average response time of users by using a particle swarm optimization approach to balance the workload between MEC servers. Note that these studies assumed non-adjustable computational capabilities of the CPUs at the device side. Besides, they considered only the cellular access mode without any computation offloading through WiFi.

To achieve the desired energy-delay performance, the so-called dynamic voltage scaling (DVS) is a promising technique that varies the supply voltage and clock frequency based on the computation load to achieve a suitable tradeoff between task execution time and energy consumption [42]. In DVS, the local execution time of delay-sensitive tasks as well as the energy consumption of mobile devices can be further minimized by controlling the local CPU frequency. In [43], a Lyapunov optimization-based dynamic computation offloading algorithm was proposed for MEC systems with mobile devices powered by renewable energy, which jointly determines the offloading decision, CPU cycle frequencies, and transmit power for offloading. Note that neither the role of backhaul nor conventional cloud were discussed in [43], nor was WiFi considered therein. In [42], the authors investigated partial computation offloading by jointly optimizing the computational speed, transmission power, and offloading ratio of mobile devices in a single-hop edge computing scenario. For femto-cloud computing systems, where the cloud server is formed by a set of femto access points, the transmit power, precoder, and computation load distribution were jointly optimized in [44]. The work in [45] was one of the first to address the joint computation offloading and resource scheduling problem with task dependencies for mobile cloud computing.

To the best of the author's knowledge, while many studies have addressed the problem of MEC-based offloading, the performance-limiting impact of backhaul latency and user mobility has not been examined in sufficient detail previously. Further, none of the existing studies have addressed dynamic cooperative offloading for MEC enabled FiWi enhanced HetNets from an architecture design and offloading orchestration perspective, while considering both WiFi and cellular access networks. Moreover, in all previous studies that aimed to realize multi-tier hierarchical MEC, the incorporation of self-organization via DVS has not been investigated yet.

1.3 Objectives

The objectives of this thesis are as follows.

- The Tactile Internet is one of the most interesting low-latency applications for creating novel immersive experiences. The first objective of the thesis is to investigate the HITL-centric design principles that add a new dimension to the human-to-machine interaction via the Internet and set the Tactile Internet aside from the more machine-centric IoT. Another objective of the thesis is to explore how we can make sure that the potential of the Tactile Internet be unleashed for a race with (rather than against) machines.

- For realizing successful deployment of immersive Tactile Internet applications, it is crucial to identify the salient features and unique characteristics of haptic traffic. Despite growing interest in the Tactile Internet, there is still limited understanding of real haptic traffic. As future access networks will be highly integrative, designing suitable MAC solutions for coexistent H2H, M2M, and H2R traffic classes requires proper stochastic models that closely describe the packet arrival and packet size of the Tactile Internet traffic. One of the main objectives of this thesis is to understand the Tactile Internet traffic in terms of packet rate, packet size, packet interarrival time, and sample autocorrelation.

- The crucial roadblock toward successful deployment of Tactile Internet teleoperation is end-to-end latency, which is either imposed by speed of light –also known as propagation delay– or intermediate queue(s), if any. Designing proper MAC solutions at access networks as well as bringing the teleoperation points of interaction closer to each other may partly reduce the latter delay component. Nevertheless, the propagation delay still poses strong limitations on the extent to which transparent and stable teleoperation can be realized. Towards designing responsive, ultra-low latency, and highly reliable Tactile Internet teleoperation and in order to fully unleash the potential of FiWi enhanced networks, the role of edge-intelligence, deployed at MEC servers in close proximity to end-users, needs careful investigation. An important objective of the thesis is to provide insights into leveraging machine learning to compensate for delayed haptic samples via a human-centric design approach.

- With the advent of safe collaborative robots, their seamless integration into human teams as teammates is starting to gain steam as part of the vision of the emerging Tactile Internet. While necessary, low task execution time and ultra-reliable human-robot connectivity are not sufficient to unleash the full potential of the resultant HART applications. One of the objectives of the thesis is to inquire into possibilities to further extend the capabilities of FiWi enhanced LTE-A HetNets, paying particular attention to the dichotomy between automation and augmentation (i.e., extension of capabilities) of the human through the Tactile Internet.

- An interesting example of human+machine hybrid activities is semiautonomous robotic systems. Unlike their fully-autonomous counterparts, semiautonomous robotic systems rely on human assistance from time to time via teleoperation and/or telepresence when domain expertise is needed to accomplish a specific task. As these robots will need to request human assistance via teleoperation/presence, mapping these requests to the human operators themselves stands as a difficult optimization problem. One of the objectives of the thesis is to design an algorithmic solution to solve the problem of joint scheduling and assignment of teleoperation tasks to human operators while taking into account economic considerations.

- In addition to economic considerations and URLLC requirements, another important aspect of the 5G vision is decentralization. An important objective of the thesis is therefore to investigate the role of cooperative MEC in a hierarchical cloud-, edge-, and local computing scenario. This thesis aims to design proper offloading strategies to decentralize the computing across the edge servers as well as the mobile users by leveraging self-organization for realizing suitable energy-delay trade-offs for mobile users.

1.4 Research Methodology

The research methodology applied in this thesis includes traffic modeling, network architecture design, mechanism design, and performance analysis, as summarized in Fig. 1.3 and briefly described in the following.

- *Traffic Modeling:* A statistical, trace-driven modeling of teleoperation traffic is conducted, which investigates the role of deadband coding to reduce the haptic packet rate. The specific characteristics of Tactile Internet traffic are identified by studying two sets of haptic traces obtained from real-world teleoperation experiments with a different number of DoF. The presented haptic traffic characterization aims to model the size as well as the arrival process of haptic packets in both command and feedback paths, taking TORs with different DoF and perceptual coding into account. In addition, correlation patterns within the feedback samples are identified and modeled in terms of sample autocorrelation.

- *Network architecture:* Novel FiWi-based network architectures are developed to enable emerging immersive low-latency Tactile Internet applications. In particular, the developed architec-

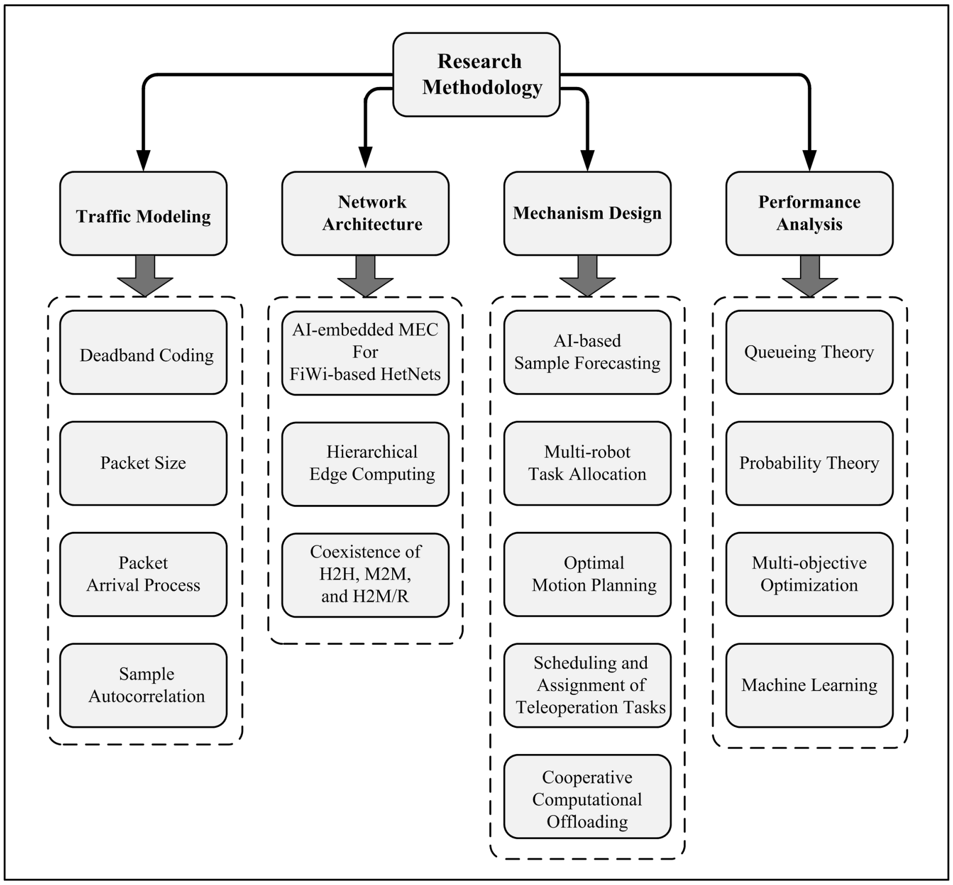

Research Methodology
Traffic Modeling
Network Architecture
Mechanism Design
Performance Analysis
Deadband Coding
Packet Size
Packet Arrival Process
Sample Autocorrelation
AI-embedded MEC For FiWi-based HetNets
Hierarchical Edge Computing
Coexistence of H2H, M2M, and H2M/R
AI-based Sample Forecasting
Multi-robot Task Allocation
Optimal Motion Planning
Scheduling and Assignment of Teleoperation Tasks
Cooperative Computational Offloading
Queueing Theory
Probability Theory
Multi-objective Optimization
Machine Learning

Figure 1.3 – Research methodology.

tures include AI-embedded MEC over FiWi enhanced HetNets, which allows for forecasting haptic samples. Further, hierarchical edge computing architecture is developed to enable cooperative computational offloading to help mobile users accelerate task execution while saving energy.

- *Mechanism design:* Novel algorithmic solutions are presented throughout the thesis to address the above mentioned objectives. These algorithms/mechanisms include AI-based haptic sample forecasting, context- and self-aware task coordination, prioritized scheduling and assignment of teleoperation tasks, and cooperative computational offloading strategy.

- *Performance analysis:* In the thesis, queueing theory, probability theory, multi-objective optimization, and machine learning have been leveraged to develop the analytical frameworks and conduct performance evaluations. In particular, a comprehensive analytical framework based on probability theory and queueing theory is developed. By means of comprehensive analysis and simulations the performance of the proposed algorithmic solutions is investigated for a variety of network and robot configurations.

1.5 Contributions of the Thesis

This thesis is a compilation of ten publications (six journal articles and four book chapters), which are published or accepted for publication in high-calibre IEEE and OSA journals as well as renowned book publishers (e.g., Wiley and CRC). The key contributions of the thesis are briefly discussed in the following.

1.5.1 Immersive Tactile Internet Experiences via Edge Intelligence

The outcome of this research has been published in the following journals/book chapters and the main contributions of this work are summarized below:

[J1] M. Maier and A. Ebrahimzadeh. Towards Immersive Tactile Internet Experiences: Low-Latency FiWi Enhanced Mobile Networks With Edge Intelligence [Invited]. *IEEE/OSA Journal of Optical Communications and Networking, Special Issue on Latency in Edge Optical Networks*, vol. 11, no. 4, pp. B10-B25, Apr. 2019.

[J2] M. Maier, A. Ebrahimzadeh, and M. Chowdhury. The Tactile Internet: Automation or Augmentation of the Human?. *IEEE Access*, vol. 6, pp. 41607-41618, 2018.

[BC1] A. Ebrahimzadeh, M. Chowdhury, and M. Maier. The Tactile Internet over 5G FiWi Architectures. *John Wiley & Sons*: Optical and Wireless Convergence for 5G Networks, pp. 197-223, Aug. 2019.

[BC2] A. Ebrahimzadeh and M. Maier. Toward HITL-Centric Edge Computing: MEC in Fiber-Wireless Tactile Internet Infrastructures. *IET*: Edge Computing: Models, Technologies and Applications, accepted for publication.

This part of the thesis focuses on the emerging Tactile Internet as one of the most interesting low-latency applications enabling novel immersive experiences. After describing the Tactile Internet's human-in-the-loop-centric design principles and haptic communications models, the development of decentralized cooperative dynamic bandwidth allocation algorithms are elaborated on for end-to-end resource coordination in FiWi access networks. Further, a thorough haptic trace-driven, statistical characterization of teleoperation traffic is conducted, which gives insights into modeling the Tactile Internet traffic in terms of packet rate, packet size, packet arrival process, and sample autocorrelation. Machine learning is then employed in the context of FiWi enhanced heterogeneous networks to decouple haptic feedback from the impact of extensive propagation delays. This enables humans to perceive remote task environments in time at a 1-ms granularity [46], [47], [48], and [49]. We note that machine learning is just one of the several techniques under consideration, e.g., haptic data compression via deadband coding, etc.

1.5.2 Context- and Self-Awareness for Task Coordination

The outcome of this research has been published in the following journal article and the main contributions of this work are summarized below:

[J3] A. Ebrahimzadeh, M. Chowdhury, and M. Maier. Human-Agent-Robot Task Coordination in FiWi-based Tactile Internet Infrastructures Using Context- and Self-Awareness. *IEEE Transactions on Network and Service Management*, vol. 16, no. 3, pp. 1127-1142, Sept. 2019.

Context-awareness is used to develop a HART-centric task coordination algorithm that minimizes the completion time of physical/digital tasks as well as average operational expenditure (OPEX) by spreading ownership of robots across mobile users. In addition, self-awareness is capitalized on to improve the performance of a given robot by identifying its capabilities as well as the objective requirements by means of optimal motion planning to minimize its energy consumption as well as traverse time. The proposed self- and context-aware HART-centric allocation scheme for both physical and digital tasks is used to coordinate the automation and augmentation of mutually beneficial human-machine coactivities across a FiWi based Tactile Internet infrastructure [50].

1.5.3 Scheduling and Assignment of Delay-Constrained Teleoperation Tasks

The outcome of this research has been published in the following journal and book chapter and the main contributions of this work are summarized below:

[J4] A. Ebrahimzadeh and M. Maier. Delay-Constrained Teleoperation Task Scheduling and Assignment for Human+Machine Hybrid Activities over FiWi Enhanced Networks. *IEEE Transactions on Network and Service Management*, accepted for publication.

[BC3] A. Ebrahimzadeh and M. Maier. Tactile Internet over FiWi enhanced LTE-A Het-Nets via Artificial Intelligence Embedded Multi-Access Edge-Computing. *CRC Press*: 5G-Enabled Internet of Things, accepted for publication.

After elaborating on the envisioned bimodal FiWi network infrastructure and its role in realizing teleoperation in the Tactile Internet, the problem of joint prioritized scheduling and assignment of delay-constrained teleoperation tasks to human operators is formulated and solved with the objective to minimize the average weighted task completion time, maximum tardiness, and OPEX per task. Further, the end-to-end packet delay of both local and non-local teleoperation over FiWi enhanced networks is estimated. The presented analysis flexibly allows for the coexistence of both conventional H2H and haptic H2M traffic, while focusing on the human operators and teleoperator robots involved in either local or non-local teleoperation [51] [52].

1.5.4 Cooperative Computation Offloading for FiWi Enhanced 4G HetNets

The outcome of this research has been published in the following journals and book chapter and the main contributions of this work are summarized below:

[J5] A. Ebrahimzadeh and M. Maier. Distributed Cooperative Computation Offloading in Multi-Access Edge Computing Fiber-Wireless Networks (Invited paper). *Elsevier Optics Communications Special Issue on Photonics for 5G Mobile Networks and Beyond*, vol. 452, pp. 130-139, Dec. 2019.

[J6] A. Ebrahimzadeh and M. Maier. Cooperative Computation Offloading in FiWi Enhanced 4G HetNets Using Self-Organizing MEC. *IEEE Transactions on Wireless Communications*, in revision.

[BC4] A. Ebrahimzadeh and M. Maier. Next Generation Multi-Access Edge-Computing Fiber-Wireless Enhanced HetNets for Low-Latency Immersive Applications. *IGI Global*: Design, Implementation, and Analysis of Next Generation Optical Networks, accepted for publication.

The performance gains of cooperative computation offloading are investigated for MEC enabled FiWi enhanced HetNets with capacity-limited backhaul links. After presenting the envisioned two-tier MEC architecture for a FiWi based networking infrastructure, a simple but efficient offloading strategy is proposed, which relies on the flexible trilateral cooperation between end-devices, edge servers, and the remote cloud. An analytical framework is then presented to estimate the average response time and energy consumption of mobile users for various offloading scenarios with different wireless access modes (i.e., WiFi and 4G LTE-A). The presented analysis flexibly allows for incorporating both offloaded and conventional H2H traffic of mobile users as well as fixed (wired) subscribers. Finally, a self-organization based mechanism is proposed, which enables mobile users to make suitable energy-delay trade-offs by jointly minimizing the local task execution time and energy consumption, using only their local information [53], [54], [55].

1.6 List of Publications

Publications included in this thesis

In summary, this thesis includes materials extracted from the following publications:

Journals

[J1] M. Maier and A. Ebrahimzadeh. Towards Immersive Tactile Internet Experiences: Low-Latency FiWi Enhanced Mobile Networks With Edge Intelligence [Invited]. *IEEE/OSA Journal of Optical Communications and Networking, Special Issue on Latency in Edge Optical Networks*, vol. 11, no. 4, pp. B10-B25, Apr. 2019.

[J2] M. Maier, A. Ebrahimzadeh, and M. Chowdhury. The Tactile Internet: Automation or Augmentation of the Human?. *IEEE Access*, vol. 6, pp. 41607-41618, 2018.

[J3] A. Ebrahimzadeh, M. Chowdhury, and M. Maier. Human-Agent-Robot Task Coordination in FiWi-based Tactile Internet Infrastructures Using Context- and Self-Awareness. *IEEE Transactions on Network and Service Management*, vol. 16, no. 3, pp. 1127-1142, Sept. 2019.

[J4] A. Ebrahimzadeh and M. Maier. Delay-Constrained Teleoperation Task Scheduling and Assignment for Human+Machine Hybrid Activities over FiWi Enhanced Networks. *IEEE Transactions on Network and Service Management*, IEEE Xplore Early Access.

[J5] A. Ebrahimzadeh and M. Maier. Distributed Cooperative Computation Offloading in Multi-Access Edge Computing Fiber-Wireless Networks (Invited paper). *Elsevier Optics Communications Special Issue on Photonics for 5G Mobile Networks and Beyond*, vol. 452, pp. 130-139, Dec. 2019.

[J6] A. Ebrahimzadeh and M. Maier. Cooperative Computation Offloading in FiWi Enhanced 4G HetNets Using Self-Organizing MEC. *IEEE Transactions on Wireless Communications*, in revision.

Book Chapters

[BC1] A. Ebrahimzadeh, M. Chowdhury, and M. Maier. The Tactile Internet over 5G FiWi Architectures. *John Wiley & Sons*: Optical and Wireless Convergence for 5G Networks, pp. 197-223, Aug. 2019.

[BC2] A. Ebrahimzadeh and M. Maier. Human-In-The-Loop Models for Multi-Access Edge Computing. *IET Press*: Edge Computing: Models, Technologies and Applications, accepted for publication.

[BC3] A. Ebrahimzadeh and M. Maier. Tactile Internet over FiWi enhanced LTE-A Het-Nets via Artificial Intelligence Embedded Multi-Access Edge-Computing. *CRC Press*: 5G-Enabled Internet of Things, accepted for publication.

[BC4] A. Ebrahimzadeh and M. Maier. Next Generation Multi-Access Edge-Computing Fiber-Wireless Enhanced HetNets for Low-Latency Immersive Applications. *IGI Global*: Design, Implementation, and Analysis of Next Generation Optical Networks, pp. 40-68, July 2019.

Publications not included in this thesis

For completeness, we mention that the following publications have also been published or submitted during my doctoral studies, though their content is not included in this thesis.

Book

[B1] A.Ebrahimzadeh and M. Maier. Toward 6G: A New Era of Convergence. *Wiley-IEEE*, accepted proposal, in preparation.

Journals

[J7] A. Beniiche, A. Ebrahimzadeh, and M. Maier. The Way of The DAO: Towards Decentralizing the Tactile Internet. *IEEE Communications Magazine*, in revision.

[J8] M. Maier and A. Ebrahimzadeh. Toward the Internet of No Things: The Role of O2O Communications and Extended Reality. *Submitted to IEEE Communications Magazine*.

Book Chapter

[BC5] A. Beniiche, A. Ebrahimzadeh, and M. Maier. From Blockchain Internet of Things (B-IoT) Towards Decentralizing the Tactile Internet. *CRC Press*: Blockchain-enabled Fog and Edge Computing: Concepts, Architectures and Applications, accepted for publication.

Conferences

[C1] A. Ebrahimzadeh and M. Maier. Coordination of Robotic Teams over Tactile Internet FiWi Enhanced Mobile Networks. *Submitted to IEEE INFOCOM 2020.*

[C2] G. O. Pérez, A. Ebrahimzadeh, M. Maier, J. A. Hernández, D. Larrabeiti, M. F. Veiga. Decentralized Coordination in MEC-Enabled FiWi Enhanced H-CRAN Tactile Internet Infrastructures. *Submitted to IEEE INFOCOM 2020.*

1.7 Thesis Outline

The thesis is organized into six chapters to present a consistent overview of the research conducted during the doctoral studies. The remainder of this thesis is structured as follows.

Chapter 2 aims to explore the HITL-centric design principles and haptic communications models of the Tactile Internet. In an effort to understand the unique characteristics of the Tactile Internet traffic, a statistical, trace-driven modeling of teleoperation traffic is conducted. Machine learning is then employed in the context of FiWi enhanced heterogeneous networks to decouple haptic feedback from the impact of extensive communication induced delays.

Chapter 3 presents a context- and self-aware HART-centric allocation scheme for both physical and digital tasks to coordinate the automation and augmentation of mutually beneficial human-machine coactivities while spreading ownership of robots across users over integrated FiWi Tactile Internet infrastructures. In addition to realizing collective context-awareness via HART-centric task coordination, this chapter aims to exploit local self-awareness in order to improve the energy-delay performance of robots. Further, an analytical framework is developed to estimate the packet transmission delay and human-robot connection reliability.

Chapter 4 studies the problem of joint prioritized scheduling and assignment of delay-constrained teleoperation tasks to human operators with the objective to minimize the average weighted task completion time, maximum tardiness, and average OPEX per task. After presenting the motivation and an illustrative case example, this chapter presents an efficient algorithm to solve the aforementioned problem and develops an analytical framework to estimate the H2R delay.

Chapter 5 investigates the performance gains obtained by cooperative computation offloading for MEC enabled FiWi enhanced HetNets with capacity-limited backhaul links. After presenting the envisioned two-tier MEC architecture for a FiWi based networking infrastructure, this chapter proposes a simple but efficient offloading strategy, which relies on the flexible trilateral cooperation between end-devices, edge servers, and the remote cloud. Further, an analytical framework is presented to estimate the average response time and energy consumption of mobile users for various offloading scenarios with different wireless access networks. Further, the chapter presents a self-organization based mechanism, which enables mobile users to make suitable energy-delay trade-offs by jointly minimizing the local task execution time and energy consumption, using only their local information.

Finally, Chapter 6 concludes the thesis by discussing the major findings of the thesis and briefly outlines potential avenues for future research that may build upon this work.

Chapter 2

Immersive Tactile Internet Experiences via Edge Intelligence

This chapter contains material extracted from the following publications:

[46] M. Maier and A. Ebrahimzadeh. Towards Immersive Tactile Internet Experiences: Low-Latency FiWi Enhanced Mobile Networks With Edge Intelligence [Invited]. *IEEE/OSA Journal of Optical Communications and Networking, Special Issue on Latency in Edge Optical Networks*, vol. 11, no. 4, pp. B10-B25, Apr. 2019.

[47] M. Maier, A. Ebrahimzadeh, and M. Chowdhury. The Tactile Internet: Automation or Augmentation of the Human? *IEEE Access*, vol. 6, pp. 41607-41618, 2018.

[48] A. Ebrahimzadeh, M. Chowdhury, and M. Maier. The Tactile Internet over 5G FiWi Architectures. *John Wiley & Sons*: Optical and Wireless Convergence for 5G Networks, pp. 197-223, Aug. 2019.

[49] A. Ebrahimzadeh and M. Maier. Human-In-The-Loop Models for Multi-Access Edge Computing. *IET Press*: Edge Computing: Models, Technologies, and Applications, accepted for publication.

In the following, my key contributions in each of the aforementioned publications are explained in greater detail.

[46] (1) I conducted the trace-driven characterization of haptic traffic in Section II.B, (2) I developed the delay analysis in Appendix B, (3) I proposed the idea of AI-embedded MEC, developed the algorithms, and implemented/verified the AI-based edge sample forecasting scheme in Section IV.B and Appendix A, (4) I ran the simulations in MATLAB, and (5) I was involved in writing the manuscript.

[47] (1) I contributed to Section III, (2) I contributed to Section IV.A, (3) I partially contributed to Section IV.B, (4) I ran the simulations in MATLAB, and (5) I was involved in writing the manuscript.

[48] (1) I largely contributed to writing the whole manuscript, including in particular Sections I-IV and (2) I ran the simulations of the algorithm presented in Section IV.

[49] (1) I conducted the trace-driven characterization of haptic traffic in Section 3, (2) I developed the algorithms in Section 3, (3) I ran all the simulations in MATLAB, and (4) I was involved in writing the manuscript.

2.1 Introduction

Beside conventional audiovisual and data traffic, the emerging *Tactile Internet* envisions the real-time transmission of haptic information (i.e., touch and actuation) for the remote control of physical and/or virtual objects through the Internet [2]. The Tactile Internet holds promise to provide a paradigm shift in how skills and labor are digitally delivered globally, thereby converting today's content-delivery networks into skillset/labor-delivery networks [56]. The Tactile Internet is expected to have a profound socio-economic impact on a broad array of applications in our everyday life, ranging from industry automation and transport systems to healthcare, telesurgery, and education. Towards this end, at the core of the design of the Tactile Internet is realizing the so-called <10ms-challenge (i.e., achieving a round-trip latency of <10 millisecond) with carrier-grade reliability.

The term Tactile Internet was first coined by G. P. Fettweis in 2014. In his seminal paper [1], the Tactile Internet was defined as a breakthrough enabling unprecedented mobile application for tactile steering and control of real and virtual objects by requiring a round-trip latency of 1-10 milliseconds. Later in 2014, ITU-T published a Technology Watch Report on the Tactile Internet, which emphasized that scaling up research in the area of wired and wireless access networks will be essential, ushering in new ideas and concepts to boost access networks' redundancy and diversity to meet the stringent latency as well as carrier-grade reliability requirements of Tactile Internet applications [7].

To give it a more 5G-centric flavor, the Tactile Internet has been more recently also referred to as the 5G-enabled Tactile Internet [56, 57]. Unlike the previous four cellular generations, future 5G networks will lead to an increasing integration of cellular and WiFi technologies and standards [58]. Furthermore, the importance of the so-called *backhaul bottleneck* needs to be recognized as well, calling for an end-to-end design approach leveraging both wireless front-end and wired backhaul technologies. Or, as eloquently put by J. G. Andrews, the lead author of [58], “placing base stations all over the place is great for providing the mobile stations high-speed access, but does this not just pass the buck to the base stations, which must now somehow get this data to and from the wired core network?” [59].

This mandatory end-to-end design approach is fully reflected in the key principles of the reference architecture within the emerging IEEE P1918.1 standards working group (formed in March 2016), which aims to define a framework for the Tactile Internet [60]. Among others, the key principles envision to (*i*) develop a generic Tactile Internet reference architecture, (*ii*) support local area as well as wide area connectivity through wireless (e.g., cellular, WiFi) or hybrid wireless/wired networking, and (*iii*) leverage computing resources from cloud variants at the edge of the network. The working group defines the Tactile Internet as follows: “*A network, or a network of networks, for remotely accessing, perceiving, manipulating or controlling real and virtual objects or processes in perceived real-time.*” Some of the key use cases considered in IEEE P1918.1 include teleoperation, haptic communications, immersive virtual reality, and automotive control.

Clearly, the Tactile Internet opens up a plethora of exciting research directions towards adding a new dimension to the human-to-machine interaction via the Internet. According to the aforementioned ITU-T Technology Watch Report, the Tactile Internet is supposed to be the next leap in the evolution of today's IoT, though there is a significant overlap among 5G, IoT, and the Tactile Internet. Despite their differences, all three share an intersecting set of design goals [3]:

- Very low latency on the order of < 10 milliseconds

- Ultra-high reliability with an almost guaranteed availability of 99.999 percent

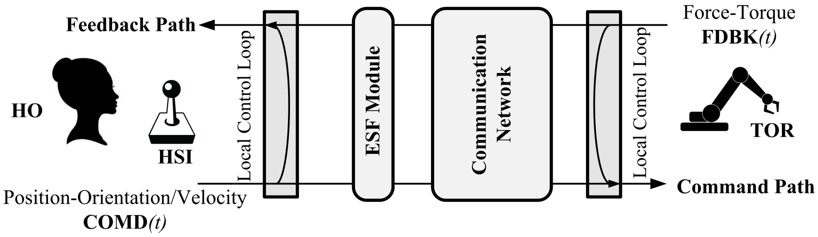


Figure 2.1 – Edge sample forecast (ESF) module at the edge of a general communication network with arbitrary propagation delays.

- H2H/M2M coexistence
- Integration of data-centric technologies with a particular focus on WiFi
- Security

For illustration, Fig. 2.1 depicts a typical teleoperation system based on bidirectional haptic communications between an HO and a TOR. Note that the number of independent coordinates required to completely specify and control/steer the position, orientation, and velocity of the TOR is defined by its DoF.[1] Further, a local HSI device is used to display haptic interaction with the remote TOR to the HO. The local control loops on both ends of the teleoperation system ensure the tracking performance and stability of the HSI and TOR.

In [3], the authors elaborated on the subtle differences between the Tactile Internet and the IoT and 5G visions, which may be best expressed in terms of underlying communications paradigms and enabling end-devices. Importantly, the Tactile Internet involves the inherent HITL nature of human-to-machine interaction, as opposed to the emerging IoT without any human involvement in its underlying M2M communications. While M2M communications is useful for the automation of industrial and other machine-centric processes, the Tactile Internet will be centered around H2M/R communication and thus allows for a human-centric design approach towards creating novel immersive experiences and extending the capabilities of the human through the Internet, i.e., augmentation rather than automation of the human [47].

In this chapter, we pay attention to bilateral teleoperation as an example of HITL applications and present an in-depth study of haptic traffic characterization and modeling in terms of packet

[1] Currently available teleoperation systems range from 1-DoF to >20-DoF TORs. For instance, a 6-DoF TOR allows for both translational motion (in 3D space) via force and rotational motion (pitch, yaw, and roll) via torque.

arrival and sample autocorrelation. We develop new models of describing packet interarrival times as well as three-dimensional sample autocorrelation. We then explore how MEC in general and edge intelligence in particular may be leveraged to help realize an immersive, reliable teleoperation experience over FiWi-based networking infrastructures. More specifically, we focus on the communication network in Fig. 2.1 and its role in realizing the Tactile Internet vision, thereby paying particular attention to the unique characteristics of haptic traffic. According to [11], even minor communication-induced time delays and packet losses may destabilize the haptic communications system. Emphasizing on its HILT-centric aspect, the Tactile Internet allows for a human-centric design approach by exploiting the properties of human haptic perception via advanced perceptual coding techniques in order to substantially reduce the haptic packet rate, as explained in technically greater detail shortly. The contributions of this chapter are threefold:

(*i*) First, we model Tactile Internet traffic by means of extensive haptic traces, taking TORs with different DoF and perceptual coding into account. As shown in Fig. 2.1, in a typical teleoperation system the position-orientation/velocity samples are transmitted from the HO through the HSI in the command path, whereas the force-torque samples are sent back to the HO in the feedback path. In teleoperation, haptic feedback plays a crucial role in providing the HO with transparency, immersion, and togetherness with the remote environment [11]. Note, however, that despite growing interest in the Tactile Internet, there is still limited understanding of the characteristics of real haptic traffic, especially at the packet level. For simplicity and analytical tractability, Tactile Internet traffic has been assumed to be Pareto or Poisson distributed in recent studies, e.g., [23]. Our Tactile Internet traffic models reveal which haptic packet interarrival time distributions best fit different types of teleoperation systems, while the assumption of Poisson traffic is found valid only for a very special case.

(*ii*) Second, we build on the recently proposed concept of so-called *FiWi enhanced LTE-Advanced (LTE-A) HetNets* [13], which were shown to achieve the 5G and Tactile Internet key requirements of very low latency and ultra-high reliability by unifying coverage-centric 4G mobile networks and capacity-centric FiWi broadband access networks based on data-centric Ethernet technologies. By means of probabilistic analysis and verifying simulations based on recent and comprehensive smartphone traces the authors of [13] showed that an average end-to-end latency of 1 millisecond can be achieved for a wide range of traffic loads and that mobile users can be provided with highly fault-tolerant FiWi connectivity for reliable low-latency fiber backhaul sharing and WiFi offloading. Note, however, that only conventional H2H communications was considered in [13]. In this chapter, we investigate the coexistence of mobile users and HOs/TORs and explore HITL-centric teleoperation techniques that achieve the aforementioned Tactile Internet target of 1 millisecond under different haptic traffic scenarios.

(*iii*) Third, for enhanced Tactile Internet reliability performance we present our proposed *edge sample forecast (ESF)* module, which is inserted at the edge of our communication network in close proximity to the HO, as shown in Fig. 2.1. According to [61], edge computing is a new paradigm, in which computing and storage resources—variously referred to as cloudlets, micro datacenters, or fog nodes—are placed at the Internet's edge in proximity to wireless end devices in order to achieve low end-to-end latency, low jitter, and scalability. A similar concept, originally known as mobile edge computing, has been standardized by ETSI for 5G networks. Note that since September 2016, ETSI has dropped the 'mobile' out of MEC and renamed it *multi-access edge computing (MEC)* in order to broaden its applicability to heterogeneous networks, including WiFi and fixed access technologies (e.g., fiber) [62]. Our proposed ESF module leverages MEC servers with embedded artificial intelligence (AI) capabilities that are placed at the optical-wireless interface of FiWi enhanced LTE-A HetNets to compensate for delayed haptic samples in the feedback path by means of multiple-sample-ahead-of-time forecasting. In doing so, the response time of the HO can be kept small, resulting in a tighter togetherness with and thereby an improved safety in the remote TOR environment.

The remainder of the chapter is structured as follows. Section 2.2 derives Tactile Internet traffic models from haptic traces by studying teleoperation as an example of an immersive Tactile Internet experience. Section 2.3 introduces the concept of low-latency FiWi enhanced LTE-A HetNets using advanced MEC with embedded AI capabilities. In Section 2.4 we develop our analytical framework to estimate end-to-end delay in teleoperation over FiWi networks. Section 2.5 presents analytical latency results verified by haptic trace driven simulations. In Section 2.6, we elaborate on the potential benefits and limitations of the proposed scheme. Finally, Section 2.7 concludes the chapter.

2.2 Trace-Based Haptic Traffic Characterization

An interesting example of a Tactile Internet experience that allows for remote immersion is the HITL-centric use case of teleoperation based on *haptic communications*. As mentioned earlier, the Tactile Internet envisions the real-time transmission of haptic information for the remote control of physical and/or virtual objects through the Internet [63]. Recall from Chapter 1 that in a typical bilateral teleoperation system, the HO interfaces with the communication network (to be described in greater detail in Section 2.3) via the HSI device, which is used to display haptic interaction with the remote TOR to the HO. The controllers on both ends of the teleoperation system ensure the tracking performance and stability of the HSI and TOR (see Fig 1.2). A perceptual deadband-based (i.e., zero output if changes in consecutive samples are minimal) data reduction may be deployed as a lossy compression mechanism by exploiting the fact that human end-users are not able to discriminate arbitrarily small differences in haptic stimuli. The human perception of haptics can be exploited to reduce the haptic packet rate. Specifically, the well-known Weber's law determines the just noticeable difference (JND), i.e., the minimum change in the magnitude of a stimulus that can be detected by humans [64]. Weber's law gives rise to the so-called *deadband coding* technique, whereby a haptic sample is transmitted only if its change with respect to the previously transmitted haptic sample exceeds a given deadband parameter $d \geq 0$ (given in percent) [11].

In the following, we take a closer look at the specific characteristics of Tactile Internet traffic by studying the use case of teleoperation. Specifically, we study two sets of haptic traces obtained from teleoperation experiments involving TORs with different DoF. The two considered teleoperation experiments involve TORs with 1 and 6 DoF. Furthermore, our haptic traces comprise measurements with different values of deadband parameter d.

2.2.1 Teleoperation Experiments

(a) 6-DoF Teleoperation without Deadband Coding The first set of our traces for a haptics-enabled telesurgery system were provided by the authors of [65] from the Centre National de la Recherche Scientifique (CNRS) at IRISA, Rennes, France. Note that telesurgery represents a well-known type of teleoperation in the healthcare sector. The system consists of a 6-DoF haptic interface at the HO side, a 6-DoF manipulator, and a six-axes force/torque sensor at the TOR side. Update samples containing the position and orientation signals from the HO are transmitted at every refresh time instant. Similarly, the HO receives force-torque feedback samples from the remote TOR. The local HO and remote TOR environment were put back-to-back during the experiments, i.e., there were no communication-induced artifacts such as latency. Note that deadband coding was not applied in this 6-DoF telesurgery experiment, i.e., $d = 0$.

(b) 1-DoF Teleoperation with Deadband Coding The second set of haptic traces were obtained from the 1-DoF teleoperation experiments at the Technical University of Munich, Germany [66]. Two Phantom Omni[2] devices were used as master (i.e., HO) and slave (i.e., TOR) devices to create a 1-DoF bilateral teleoperation scenario. The communication channel between HO and TOR was emulated by using a variable queueing system to generate constant or time-varying delays. The velocity signal at the HO side was sampled before being transmitted to the TOR, which in turn fed the force signal back to the HOR. The experiments were run with different deadband values set to $d \in \{0, 5\%, 10\%, 15\%, 20\%\}$ in both the command and feedback paths.

(c) Packetization Typically, haptic samples are packetized and transmitted immediately once new sensor readings are available to help minimize the end-to-end delay, implying a real-time transport protocol (RTP), user datagram protocol (UDP), and Internet protocol (IP) header of 12, 8, and 20 bytes, respectively [11]. Additionally, for each DoF the haptic sample of the aforementioned experimental sensor readings comprises 8 bytes. Note that N_{DoF} haptic samples are encapsulated into one RTP/UDP/IP packet, where N_{DoF} denotes the number of DoF in either experiment (i.e., 6 or 1 in our case). Thus, the packet size is equal to $40 + 8 \cdot N_{DoF}$ bytes.

2.2.2 Packet Interarrival Times

We begin by investigating the packet interarrival times of both teleoperation traces. For a given deadband parameter d, let $\lambda^c(d_c)$ and $\lambda^f(d_f)$ denote the mean packet rate at which packets arrive at the MAC layer of the wireless interface in the command path and feedback path, respectively. In the following, we discuss both teleoperation traces separately, first without deadband coding ($d = 0$) and then with deadband coding ($d > 0$).

[2] Phantom Omni is a widely used HSI device that enables HOs to interact with and manipulate objects by adding 3D navigation to a broad range of applications, e.g., games, entertainment, visualization, among others.

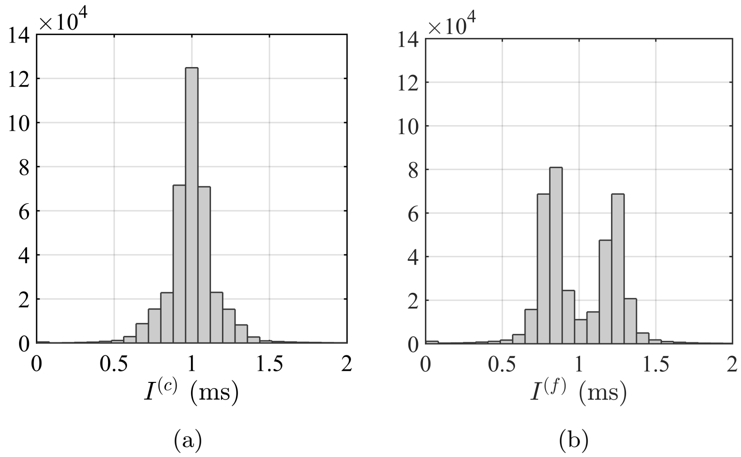


Figure 2.2 – Histogram of experimental 6-DoF teleoperation packet interarrival times: (a) command path and (b) feedback path.

Note that in our 1-DoF teleoperation traces without deadband, packet interarrival times are deterministic with a constant packet arrival rate of $\lambda^c(d_c)|_{d_c=0} = \lambda^f(d_f)|_{d_f=0} = 1000$ packets/sec in both command and feedback paths due to the fixed haptic sampling rate of 1 kHz. Conversely, in our 6-DoF teleoperation traces without any deadband coding, haptic samples are immediately packetized and transmitted at varying (i.e., non-deterministic) refresh time instants. In the following, we examine the command and feedback paths of our 6-DoF teleoperation traces separately and try to find the best fitting distribution for the respective packet interarrival times.

First, let us focus on the position and orientation samples in the command path (from HO to TOR), which are measured as a triplet and quaternion (i.e., quadruple) at each refresh time instant, respectively. Let $\mathbf{COMD}_i$ denote the resultant position-orientation sample i, which is transmitted as packet in the command path at time instant $T_i^{(c)}$. Thus, the corresponding packet interarrival times $I_i^{(c)} = T_i^{(c)} - T_{i-1}^{(c)}$, $i = 2, 3, \ldots$, represent realizations of the random variable $I^{(c)}$. Figure 2.2(a) depicts the histogram of the packet interarrival times $I^{(c)}$ in the command path obtained from the 6-DoF teleoperation traces, with the most frequent packet interarrival time expectedly being centered at 1 ms due to the default haptic sampling rate of 1 kHz.

The histogram of the packet interarrival times $I^{(f)}$ in the feedback path (from TOR to HO) is shown in Fig. 2.2(b). Interestingly, the feedback path differs from the command path in that it exhibits two peaks at approximately 0.75 ms and another one at 1.25 ms. Upon examining the force/torque traces stemming from the TOR side, we found that the two peaks exist because the

Table 2.1 – Summary of the estimated parameters of fitted PDFs using MLE method.

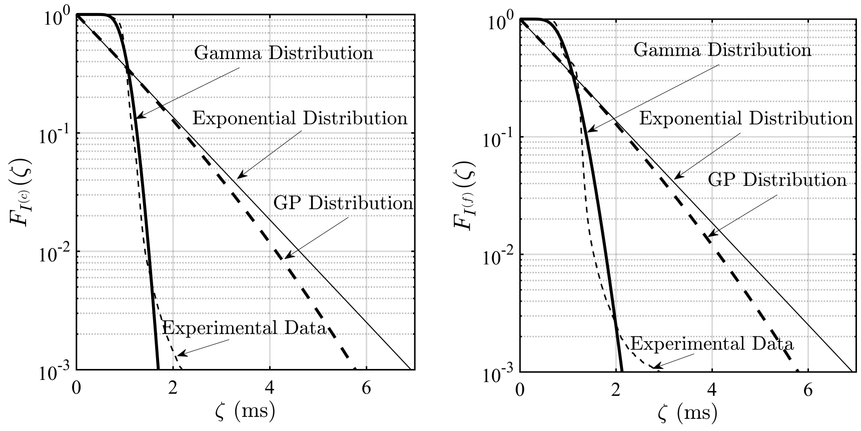
		Exponential		Generalized Pareto				Gamma		
		$f_I(x) = \frac{1}{\mu} e^{-\frac{x}{\mu}}, x \geq 0$		$f_I(x) = \frac{1}{\sigma}\left(1 + k\frac{x-\theta}{\sigma}\right)^{-1-\frac{1}{k}}, x \geq \theta$				$f_I(x) = \frac{r^a}{\Gamma(a)} x^a \frac{e^{-rx}}{x}, x \geq 0$		
	d (%)	μ	D^*	k	σ	θ	D^*	a	r	D^*
6-DoF (Command Path)	0	0.0010	0.47	-0.065	1.0×10^{-3}	3.7×10^{-6}	0.46	27	27620	**0.14**
	0.05	0.0028	0.23	0.12	2.4×10^{-3}	4.7×10^{-6}	**0.17**	1.51	540	0.19
	0.10	0.0047	0.16	0.16	3.8×10^{-3}	9.6×10^{-6}	**0.14**	1.29	270	0.19
	0.20	0.0087	0.18	0.19	6.6×10^{-3}	7.15×10^{-6}	**0.13**	0.65	27	0.15
6-DoF (Feedback Path)	0	0.001	0.45	-0.064	1.0×10^{-3}	2.6×10^{-6}	0.45	12	12166	**0.10**
	5	0.0012	0.41	-0.02	1.2×10^{-3}	2.8×10^{-6}	0.40	4.85	4121	**0.19**
	10	0.0014	0.36	0.05	1.3×10^{-3}	4.1×10^{-6}	0.37	2.56	1877	**0.23**
	20	0.0017	**0.21**	0.13	1.3×10^{-3}	2.8×10^{-6}	0.27	1.54	931	0.31
1-DoF (Command Path)	5	0.0022	0.36	0.63	5.7×10^{-4}	7.3×10^{-4}	**0.34**	1.46	663	0.38
	10	0.0027	0.38	0.81	4.9×10^{-4}	7.5×10^{-4}	**0.34**	1.04	386	0.39
	15	0.0038	0.41	0.88	7.8×10^{-4}	6.4×10^{-4}	**0.32**	0.79	208	0.36
1-DoF (Feedback Path)	5	0.0075	0.16	0.46	3.7×10^{-3}	5.8×10^{-4}	**0.10**	0.91	120	0.14
	10	0.0024	0.20	0.57	11.5×10^{-3}	3.4×10^{-4}	**0.05**	0.69	28	0.13
	15	0.0036	0.12	0.32	24.2×10^{-3}	11.2×10^{-4}	**0.04**	0.91	25	0.10

force and torque sensors of the TOR operate at two slightly different sampling rates above and below 1 kHz.

In an effort to find a probability distribution function (PDF) that best fits the experimental packet interarrival times in Fig. 2.2(a), we considered a variety of well-known distributions. Our preliminary evaluations narrowed our choice down to three candidate PDFs, namely, exponential, generalized Pareto (GP), and gamma distributions. Our method of selecting the best fitting PDF comprised the following three steps. First, we used the maximum likelihood estimation (MLE) method to estimate the parameters of each PDF. Second, the estimates of the first step were verified by computing the complementary cumulative distribution function (CCDF) $F_{I^{(c)}}(\zeta) = P(I^{(c)} > \zeta)$. Third, to compare the goodness-of-fit among the three PDFs under consideration, we used the maximum difference D^* between the fitted and experimental CCDFs, which is given by

$$D^* = \sup_{\zeta} \left| \hat{F}_{I^{(c)}}(\zeta) - F_{I^{(c)}}(\zeta) \right|, \tag{2.1}$$

whereby $\hat{F}_{I^{(c)}}(\zeta)$ denotes the experimental CCDF. The estimated parameters as well as the calculated D^* of the fitting PDFs are listed in Table 2.1, where we observe that the gamma distribution matches the experimental data reasonably well, as opposed to the exponential and GP distributions. Next, we proceed by fitting the best PDF to the 6-DoF experimental packet interarrival times in the feedback path. Similar to the command path, we observe from Table 2.1 that in the feedback path, the gamma distribution again fits the experimental data best.



(a)

(b)

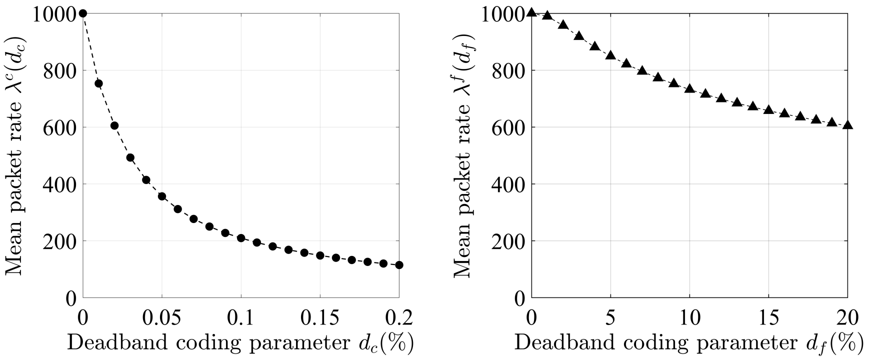
Figure 2.3 – CCDF of fitted PDFs and experimental 6-DoF teleoperation packet interarrival times: (a) command path and (b) feedback path.

Figure 2.3b(a) shows the CCDF of the three fitted PDFs and experimental 6-DoF teleoperation packet interarrival times in the command path. We observe from the figure that the gamma distribution matches the experimental data reasonably well, as opposed to the exponential and GP distributions. Similar to the command path, we observe from Fig. 2.3b(b) and Table 2.1 that for the CCDF in the feedback path, $F_{I(f)}(\zeta) = P(I^{(f)} > \zeta)$, the gamma distribution again best fits the experimental data.

Next, we study the 1-DoF teleoperation experiment, which included deadband coding unlike its 6-DoF counterpart. For fair comparison of the two sets of haptic traces, we post-processed the original 6-DoF traces and applied deadband coding for a variety of different deadband parameter values in the command path (d_c) and feedback path (d_f), as explained in the following. To begin with, we model the position signal with a 3D vector-valued function of time denoted by $\mathbf{p}(t)$. Further, let $\mathbf{o}(t)$ denote the orientation signal, which is modeled by a quaternion[3]. Similar to the position signal in the command path, we model the force and torque signals in the feedback path by 3D vector-valued function $\mathbf{f}(t)$ and $\mathbf{t}(t)$, respectively. We apply the deadband coding as follows. In the command path of 6-DoF teleoperation with deadband coding a position-orientation sample $\mathbf{comd}(t)$ is transmitted only if the proportional change with respect to the previously transmitted sample $\mathbf{COMD}_{i-1}$ exceeds a given d_c, i.e., $\mathbf{COMD}_i = \mathbf{comd}(t)$ only if

$$\max\{\Delta_p, \Delta_o\} > d_c,$$

[3] Quaternion representation of orientation is characterized by $\hat{\mathbf{v}} = (\hat{v}_x, \hat{v}_y, \hat{v}_z)$ and $\hat{\theta}$, where $\hat{\theta}$ is the angle of rotation and $\hat{\mathbf{v}}$ is the unit vector about which rotation is performed, i.e., the axis of rotation.



(a) command path

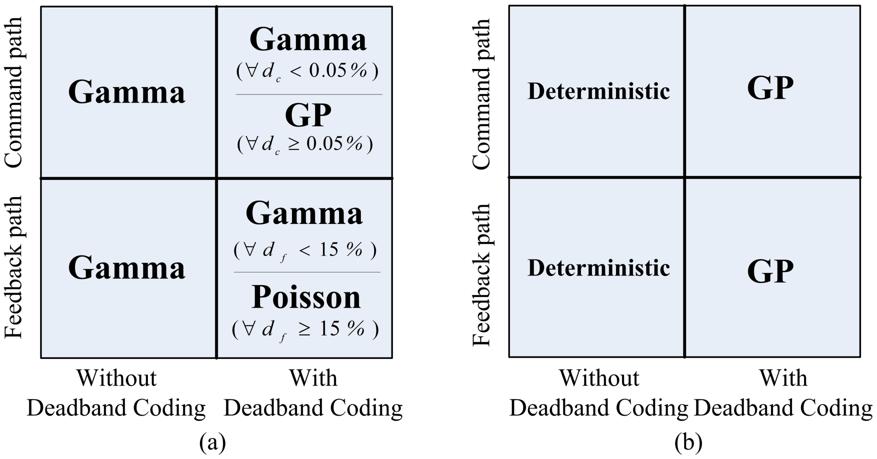
(b) feedback path

Figure 2.4 – Mean packet rate (in packets/second) vs. d for 6-DoF teleoperation.

where $\Delta_p = \frac{\|\mathbf{p}(t)-\mathbf{p}_{i-1}\|}{\|\mathbf{p}_{i-1}\|}$ and $\Delta_o = \max\left\{\Delta_{\hat{v}}, \Delta_{\hat{\theta}}\right\}$, whereby $\Delta_{\hat{v}} = \frac{\|\hat{\mathbf{v}}(t)-\hat{\mathbf{v}}_{i-1}\|}{\|\hat{\mathbf{v}}_{i-1}\|}$ and $\Delta_{\hat{\theta}} = \frac{\left\|\hat{\theta}(t)-\hat{\theta}_{i-1}\right\|}{\left\|\hat{\theta}_{i-1}\right\|}$. Note that $\|\cdot\|$ denotes the Euclidean norm function. In the feedback path of 6-DoF teleoperation, an update force-torque sample is transmitted, only if $\max\left\{\Delta_f, \Delta_t\right\} > d_f$, where $\Delta_f = \frac{\|\mathbf{f}(t)-\mathbf{f}_{i-1}\|}{\|\mathbf{f}_{i-1}\|}$ and $\Delta_t = \frac{\|\mathbf{t}(t)-\mathbf{t}_{i-1}\|}{\|\mathbf{t}_{i-1}\|}$.

Figure 2.4 illustrates the beneficial impact of deadband coding on reducing the haptic packet rate in the feedback path and in particular the command path. More specifically, note that in the command path a deadband parameter of only $d_c = 0.02\%$ decreases the mean packet rate $\lambda^c(d_c)$ to roughly 600 packets/second, translating into a haptic packet rate reduction of 39.5% compared to the case without deadband (i.e., $d_c = 0$). As shown in Fig. 2.4(a), $\lambda^c(d_c)$ further decreases for increasing d_c and levels off for $d_c > 0.1\%$. We observe from Fig. 2.4(b) that deadband coding is less effective in the feedback path, where a deadband parameter of as high as $d_f = 20\%$ (compared to $d_c = 0.02\%$ above) is needed to reduce the mean packet rate $\lambda^f(d_f)$ to roughly 600 packets/second.

We again determined the best fitting PDFs for the packet interarrival times with the different deadband parameter values by following the same approach as described above. Table 2.1 comprehensively summarizes our findings on the different best fitting packet interarrival time distributions for command and feedback paths with and without deadband coding in both teleoperation scenarios under consideration. Note that in Table 2.1, the goodness-of-fit of the best fitting PDF for each teleoperation scenario is shown in bold. For completeness, Fig. 2.5 comprehensively summarizes our findings on the different best fitting packet interarrival time distributions for command and feedback paths with and without deadband coding in both of the teleoperation scenarios. We observe that in general command and feedback paths can be jointly modeled by the GP, gamma, or deterministic



	Without Deadband Coding	With Deadband Coding
Command path	Gamma	Gamma ($\forall d_c < 0.05\%$) GP ($\forall d_c \geq 0.05\%$)
Feedback path	Gamma	Gamma ($\forall d_f < 15\%$) Poisson ($\forall d_f \geq 15\%$)

(a)

	Without Deadband Coding	With Deadband Coding
Command path	Deterministic	GP
Feedback path	Deterministic	GP

(b)

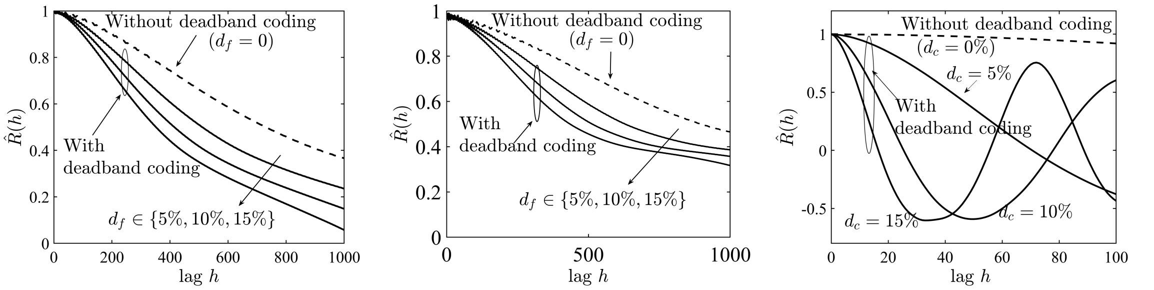
Figure 2.5 – Summary of best fitting packet interarrival time distributions for command and feedback paths with and without deadband coding: (a) 6-DoF teleoperation and (b) 1-DoF teleoperation.

packet interarrival time distribution, depending on the given value of deadband parameters d_c and d_f, as shown in Fig. 2.5.

Importantly, our haptic trace analysis indicates that the assumption made in recent studies that Tactile Internet traffic is Pareto distributed is not valid for the analyzed traffic. Furthermore, the assumption of Poisson traffic (e.g., [23]) with exponentially distributed packet interarrival times was found valid only for 6-DoF teleoperation in the feedback path with deadband parameter values of $d_f \geq 15\%$. We note that our trace analysis provides important yet preliminary insights into the statistics of Tactile Internet traffic. Clearly, a more systematic approach looking at additional haptic traces of different types of teleoperation experiments will be instrumental in accurately validating the packet interarrival time distributions reported above.

2.2.3 Sample Autocorrelation

After modeling haptic traffic arrival, we take a closer look at our available traces in the feedback path to identify possible correlation patterns in haptic samples. Such correlation patterns can be useful in developing sample forecasting techniques, leveraging AI capabilities at the network edge to compensate for delayed feedback samples by making accurate forecasts [46], to be discussed in technically great detail later on. In the following, we are going to answer the following questions: (*i*) how deep the feedback samples are correlated with their own lagged samples? and (*ii*) what is the impact of deadband coding on the autocorrelation of the feedback samples?



(a) force samples in 6-DoF teleoperation (b) torque samples in 6-DoF teleoperation (c) force samples in 1-DoF teleoperation

Figure 2.6 – Estimation of the autocorrelation of the haptic samples in the feedback path of 1- and 6-DoF teleoperation.

To answer these questions, we devise the autocorrelation function. We note, however, that haptic packets transmitted in a typical teleoperation system contain the samples taken from continuous signals, which are either 1D (i.e., force signal in 1-DoF teleoperation) or 3D (i.e., force/torque signal in 6-DoF teleoperation) vector-valued functions of time. Unlike 1D signals, estimating the autocorrelation function of a multi-dimensional vector-valued function is not straightforward. We thus present our method of estimating autocorrelation function of a given multi-dimensional discrete vector-valued function. For the sake of argument, let us consider $\mathbf{x}(t)$ as a 3D vector-valued function evaluated at time t, which is characterized by $\mathbf{x}(t) = x_1(t)\mathbf{i} + x_2(t)\mathbf{j} + x_3(t)\mathbf{k}$, where $x_1(t)$, $x_2(t)$, and $x_3(t)$ are the corresponding $x-$, $y-$, and $z-$coordinates of $\mathbf{x}(t)$, respectively. Note that $\mathbf{i}$, $\mathbf{j}$, and $\mathbf{k}$ are unit vectors representing the axes of the Cartesian coordinate system. We estimate the sample mean $\bar{\mathbf{x}} \in \mathbb{R}^3$ of a given vector-valued function $\mathbf{x}(t)$ by $\frac{1}{N_s}\sum_{i=1}^{N_s} \mathbf{x}(i\bar{T}_s)$, where N_s and $\bar{T}_s$ denote the total number of samples and inter-sample time, respectively. We then estimate the sample variance $\sigma^2_{\mathbf{x}} \in \mathbb{R}^+ \cup \{0\}$ by $\frac{1}{N_s-1}\sum_{i=1}^{N_s} \left\|\mathbf{x}(i\bar{T}_s) - \bar{\mathbf{x}}\right\|^2$, which can be generalized to estimate the autocorrelation function $\hat{R}_{\mathbf{x}}(h)$ by $C(h)/\sigma^2_{\mathbf{x}}$, where $C(h)$, $\forall h \ll N_s$, is given by

$$C(h) = \frac{1}{N_s - 1}\sum_{i=1}^{N_s-h} \ll (\mathbf{x}(i\bar{T}_s) - \bar{\mathbf{x}}) \cdot (\mathbf{x}((i+h)\bar{T}_s) - \bar{\mathbf{x}}) \gg, \tag{2.2}$$

where $\ll \cdot \gg$ denotes the inner product[4].

Figure 2.6 depicts the autocorrelation function of the force/torque feedback samples of both available sets of traces for different teleoperation scenarios with and without deadband coding. To

[4] Note that our defined autocorrelation function is based on the notion of the inner product of vectors $\mathbf{x_1}$ and $\mathbf{x_2}$, which is given by $\|\mathbf{x_1}\| \cdot \|\mathbf{x_2}\| \cdot \cos\theta$, where θ is the angle between $\mathbf{x_1}$ and $\mathbf{x_2}$ in a multi-dimensional space. As θ deviates from zero, the two vectors are less correlated and vice versa.

cope with irregular sampling intervals, which occur after performing deadband coding, we have used a zero-order hold interpolator at the rate of 1 kHz. We observe that the force/torque samples represent a quite deep correlation with their own lagged samples. Let correlation depth h^*_α denote the maximum time lag such that, for $h < h^*_\alpha$, force/torque sample autocorrelation $\hat{R}(h)$ is greater than $\alpha\%$. Note that deadband coding, in general, decreases the autocorrelation of feedback samples for a given time lag h, thus decreasing the correlation depth, see Fig. 2.6(a)-(c). For instance, the force feedback signal in 6-DoF teleoperation without deadband coding exhibits a correlation depth $h^*_{90\%}$ of 202. Deadband coding, in turn, reduces the correlation depth to 142, 116, and 98 for $d = 5\%$, $d = 10\%$, and $d = 15\%$, respectively. Further, we find that the torque samples show a slightly higher autocorrelation compared with that of the force samples in 6-DoF teleoperation. Also note that the 1-DoF force samples with deadband coding are associated with less autocorrelation, compared to the 6-DoF forece samples. This is mainly due to the fact that in 6-DoF teleoperation, 3D force samples are transmitted, as opposed to 1-DoF teleoperation, where only one dimensional force samples are fed back, thus being more susceptible to deadband coding.

2.3 Low-latency FiWi Enhanced LTE-A HetNets with Edge Intelligence

Recall from Section 2.1 that FiWi access networks provide a promising approach to offload mobile data from cellular networks by means of WiFi offloading. Recent backhaul-aware 4G studies have begun to investigate the performance-limiting impact of backhaul links in small-cell networks, though most of them did not take fiber link failures into account and assumed the reliability of the backhaul to be ideal (i.e., offering an availability of $\backsim$100%).

To meet the URLLC requirements of 5G networks, the authors of [13] recently explored the performance gains obtained from enhancing coverage-centric 4G LTE-A HetNets with capacity-centric FiWi access networks based on low-cost, data-centric Ethernet NG-PON and Gigabit-class WLAN technologies. Clearly, by unifying LTE-A HetNets and FiWi access networks, low-cost high-speed mobile data offloading is achievable via high-capacity fiber backhaul (e.g., IEEE 802.3av 10G-EPON) and Gigabit-class WLAN that has been able to consistently provide data rates 100 times higher than cellular networks [67], thus helping reach the envisioned 1000-fold gains in area capacity and 10 Gb/s peak data rates of 5G.

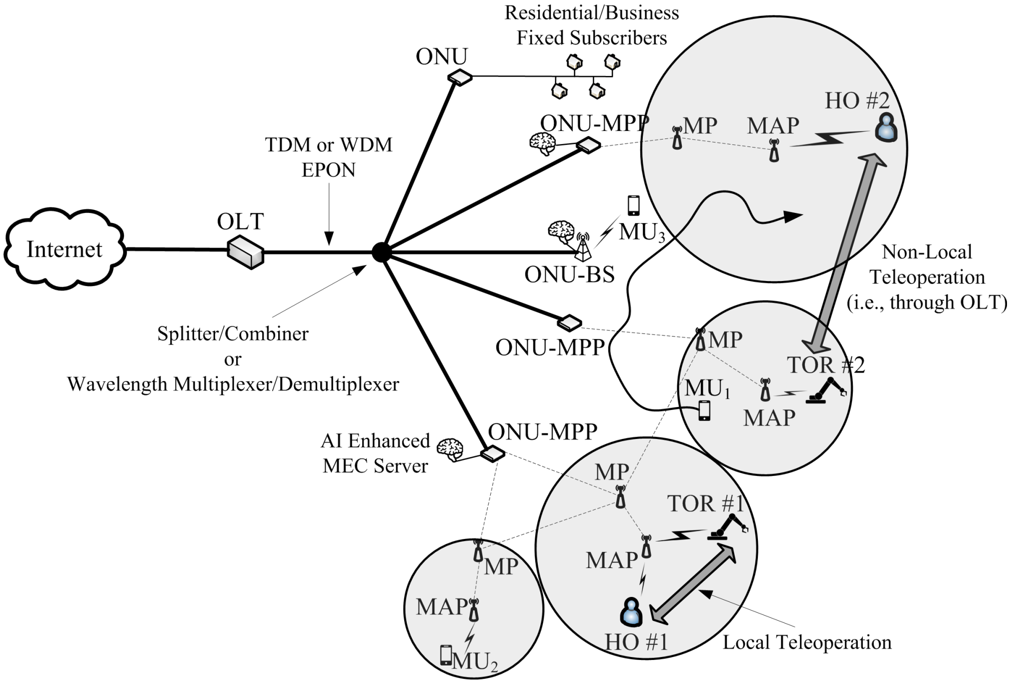


Figure 2.7 – Local and non-local teleoperation in FiWi enhanced LTE-A HetNets with AI-based MEC capabilities.

In the following, we extend the concept of FiWi enhanced LTE-A HetNets in order to enable both local and non-local teleoperation by exploiting AI-based MEC capabilities. Note that neither teleoperation nor edge intelligence were addressed in [13].

2.3.1 Low-Latency FiWi Enhanced LTE-A HetNets

Figure 2.7 depicts the generic network architecture of FiWi enhanced LTE-A HetNets. The fiber backhaul consists of a time/wavelength division multiplexing (TDM/WDM) IEEE 802.3ah/av 1/10 Gb/s EPON with a typical fiber range of 20 km between the central OLT and remote ONUs. The EPON may comprise multiple stages, each stage separated by a wavelength-broadcasting splitter/combiner or wavelength multiplexer/demultiplexer. There are three different subsets of ONUs. An ONU may either serve fixed (wired) subscribers. Alternatively, an ONU may connect to either a cellular network base station (BS) or an IEEE 802.11n/ac/s WLAN mesh portal point (MPP), giving rise to a collocated ONU-BS or ONU-MPP, respectively. Depending on positioning, a mobile user (MU) may communicate through the cellular network and/or WLAN mesh front-end, which consists of ONU-MPPs, intermediate mesh points (MPs), and mesh access points (MAPs). Note that connecting these three different sets of ONUs via a common shared EPON fiber backhaul infrastructure helps achieve the important goal of fixed-mobile convergence gain of today's network operator strategy, as discussed in Section 2.1.

In [13], various advanced fiber-lean backhaul redundancy strategies (not shown in Fig. 2.7), were proposed, which can be used to realize a local Fx fronthaul (Fx-FH) with direct inter-ONU communication. Specifically, the following three strategies can be considered: (*i*) interconnection fiber links between pairs of neighboring ONUs, (*ii*) small-scale fiber protection rings among multiple nearby ONUs, and (*iii*) wireless bypassing of backhaul fiber faults via the WLAN front-end. The results showed that the localized protection techniques proposed in [13] are instrumental in providing fixed wired and mobile users with highly fault-tolerant FiWi connectivity. Recall from Section 2.1 that Fx-FH solutions also help reduce latency by forming local clusters of ONUs as well as ONU-MPPs, thereby increasing the diversity of network connections. Our analytical results verified by recent comprehensive smartphone traces showed that the presented interconnection fiber, protection ring, and wireless protection techniques are able to keep the FiWi connectivity probability of MUs essentially flat for a wide range of EPON fiber-link failure probabilities while decreasing the average end-to-end delay to 1 millisecond for a wide range of traffic loads.

To better understand the reason behind the low delay performance of FiWi enhanced LTE-A HetNets, we note that LTE systems themselves cannot guarantee low latency due to the fact that the transmission time interval is 1 millisecond. Thus, both uplink and downlink transmissions take at least 1 millisecond, translating into an end-to-end delay being lower bounded by 2 milliseconds. In real-world deployment scenarios, the latency in LTE networks may increase by an order of magnitude. On the other hand, low-latency WiFi technology can bring 5G level of service today if the network is properly set up to mitigate interference, given that distributed coordination function (DCF) per se does not impose inherent latency limitations in that it allows users to immediately access (after a short DIFS of 50 μs) the idle wireless channel in a decentralized manner.[5]

In this work, unlike [13] which studied only conventional H2H communication between MUs, we investigate the potential and limits of *coexistent teleoperation* in FiWi enhanced LTE-A HetNets. Given the typical WiFi-only operation of state-of-the-art robots [3], HOs and TORs are assumed to communicate only via WLAN, as opposed to MUs who use dual-mode 4G/WiFi smartphones. Teleoperation is done either locally or non-locally, depending on the proximity of the involved

[5] Aptilo Networks, "Why wait for 5G? Carrier Wi-Fi is here today," Dec. 22, 2016. Online: `www.wifinowevents.com`

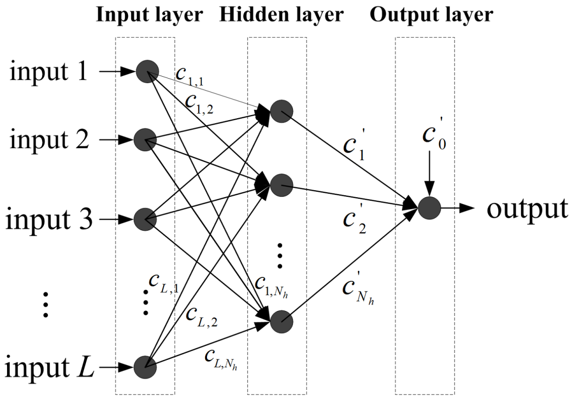


Figure 2.8 – Generic architecture of an MLP-ANN model with L inputs and one output.

HO and TOR, as illustrated in Fig. 2.7. In local teleoperation, the HO and corresponding TOR are associated with the same MAP and exchange their command and feedback samples through this MAP without traversing the fiber backhaul. Conversely, if HO and TOR are associated with different MAPs, non-local teleoperation is generally done by communicating via the backhaul EPON and central OLT. For simplicity, in this work we focus on the generic network architecture of FiWi enhanced LTE-A HetNets, shown in Fig. 2.7, without leveraging direct inter-ONU communication.

2.3.2 Edge Intelligence

Despite recent interest in exploiting machine learning for optical communications and networking, edge intelligence for enabling an immersive and transparent teleoperation experience for human operators has not been explored yet. In the following, we introduce machine learning at the edge of our considered communication network for realizing immersive and frictionless Tactile Internet experiences.

To realize edge intelligence, selected ONU-BSs/MPPs are equipped with AI-enhanced MEC servers. These servers rely on the computational capabilities of cloudlets collocated at the optical-wireless interface (see Fig. 2.7) to forecast delayed haptic samples in the feedback path. Towards this end, we deploy a type of parameterized artificial neural network (ANN) known as *multi-layer perceptron (MLP)*, which is capable of approximating any linear/non-linear function to an arbitrary degree of accuracy [68]. Figure 2.8 illustrates the generic architecture of an MLP-ANN model. Note that an MLP with N_h hidden neurons represents a linear combination of N_h parameterized non-linear functions called neurons. Furthermore, note that a neuron is a nonlinear function $\mathcal{G}(\cdot)$ of a linear combination of its input variables. In this work, the ANN is an MLP with L input variables and one output variable. More specifically, Ξ denotes the set of $L \cdot N_h + N_h + 1$ weights of the model, i.e., $\Xi = \{c_{i,j} : i = 1, \ldots, N_h, j = 1, \ldots, L\} \cup \{c_j^{'} : j = 0, 1, \ldots, N_h\}$, which are estimated during the training phase. The MLP yields the following output:

$$\Psi\left(\boldsymbol{\mathcal{A}}\ , \Xi\right) = \sum_{j=1}^{N_h} c_j^{'} \mathcal{G}\left(\sum_{i=1}^{L} c_{i,j} \boldsymbol{\mathcal{A}}(i)\right) + c_0^{'}, \tag{2.3}$$

where $\boldsymbol{\mathcal{A}} \in \mathbb{R}^L$ represents the input vector (see Fig. 2.8). We note that the weights Ξ of the ANN are computed by the corresponding MEC server and are subsequently sent to the HO in close proximity.

Recall from Section 2.2.2 that deadband coding is less effective in the feedback path (see also Fig. 2.3b(b)). In this section, leveraging the notable amount of correlation between the haptic samples, as observed in our teleoperation traces in Section 2.2.3, we elaborate on our proposed ESF module as an interesting alternative to deadband coding in the feedback path, using the MLP described above instead. To do so, we present an *ESF* module based on the aforementioned MLP to compensate for delayed haptic feedback samples by means of multiple-sample-ahead-of-time forecasting. As a result, the response time of the HO can be kept small, which in turn leads to a tighter togetherness with the remote TOR and an enhanced immersion. In a nutshell, our developed MLP based ESF module forecasts the force samples in the feedback path in real-time. More specifically, instead of waiting for the force samples that are delayed by more than a given waiting deadline T_{thr}, the module locally generates and delivers the forecast feedback samples to the HO. Let us refer to the feedback signal to be forecast as the target signal $X(\cdot)$, i.e., the force feedback samples in our case. Our objective is to generate at any time instant t a forecast sample denoted by θ^* for time instant $t_0 = t - T_{thr}$, whereby T_{thr} is the maximum period of time that the HO can wait until receiving the actual sample $\theta = X(t_0)$. More precisely, at any time t, if the sample for time instant t_0 is not received, a forecast sample is generated and immediately delivered to the HO. This procedure is repeated every 1 millisecond, which equals the typical intersample time of teleoperation systems. Note that the proposed MLP predicts θ from the past observations of the target signal. A technically more detailed description of our proposed ESF module is presented in the following.

Algorithm 1 Edge Sample Forecast

```
Input: $\boldsymbol{\mathcal{T}}, \boldsymbol{\mathcal{S}}, t_0, \Xi$
Output: $\theta^*$
 1: $\delta = 1/F_s$
 2: $\boldsymbol{\mathcal{T}}^{\boldsymbol{\delta}}, \boldsymbol{\mathcal{S}}^{\boldsymbol{\delta}} = \text{SAMPLE\_ALIGNER}(\boldsymbol{\mathcal{T}}, \boldsymbol{\mathcal{S}}, \delta)$
 3: $\Delta \leftarrow \left\lceil \frac{t_0 - \boldsymbol{\mathcal{T}}^{\boldsymbol{\delta}}(L)}{\delta} \right\rceil$
 4: $\boldsymbol{\mathcal{A}_0} \leftarrow \left(s_1^{\delta}, ..., s_L^{\delta}\right) \in \mathbb{R}^L$
 5: for $i = 1$ to $\Delta$ do
 6:     $t_i^* \leftarrow t_L^{\delta} + i \times \delta$
 7:     $\theta_i = \Psi\left(\boldsymbol{\mathcal{A}_{i-1}}, \Xi\right)$
 8:     $\boldsymbol{\mathcal{A}_i} = \left(\boldsymbol{\mathcal{A}_{i-1}}(2), \boldsymbol{\mathcal{A}_{i-1}}(3), ..., \boldsymbol{\mathcal{A}_{i-1}}(L), \theta_i\right)$
 9: end for
10: $\theta^* \leftarrow \frac{\theta_{\Delta} - \theta_{\Delta-1}}{t^*_{\Delta} - t^*_{\Delta-1}} \left(t_0 - t^*_{\Delta-1}\right) + \theta_{\Delta-1}$
11: return $\theta^*$
```

Algorithm 2 SAMPLE_ALIGNER()

```
Input: $\boldsymbol{\mathcal{T}}, \boldsymbol{\mathcal{S}}, \delta$
Output: $\boldsymbol{\mathcal{T}}^{\boldsymbol{\delta}}, \boldsymbol{\mathcal{S}}^{\boldsymbol{\delta}}$
 1: $L \leftarrow \left\lceil \frac{t_K - t_1}{\delta} \right\rceil$
 2: for $i = 1$ to $L$ do
 3:     $t_i^{\delta} \leftarrow t_1 + (i-1)\delta$
 4: end for
 5: $s_1^{\delta} \leftarrow s_1$
 6: for $i = 2$ to $L$ do
 7:     $s_i^{\delta} \leftarrow \frac{s_j - s_{j-1}}{t_j - t_{j-1}} \left(t_i^{\delta} - t_{j-1}\right) + s_{j-1}, \forall j : t_{j-1} < t_i^{\delta} < t_j$
 8: end for
 9: return $\boldsymbol{\mathcal{T}}^{\boldsymbol{\delta}}, \boldsymbol{\mathcal{S}}^{\boldsymbol{\delta}}$
```

Our objective is to generate at any time t a forecast sample θ^* for time instant $t_0 = t - T_{thr}$, where T_{thr} is the waiting threshold until which the HO can wait to receive the actual sample $\theta = X(t_0)$. Let $\boldsymbol{\mathcal{S}}, \boldsymbol{\mathcal{T}} \in \mathbb{R}^K$ denote the last K samples $\{s_1, s_2, \ldots, s_K\}$ at time stamps $\{t_1, t_2, \ldots, t_K\}$. Note that $\boldsymbol{\mathcal{S}}, \boldsymbol{\mathcal{T}}$ are used to forecast sample θ^* at any time instant $t_0 > t_K$. The feedback sample is forecast by Algorithm 1 with input $\boldsymbol{\mathcal{S}}, \boldsymbol{\mathcal{T}}, t_0, \Xi$ and output θ^*. We define δ as the intersample time step in our sample forecaster and set it to $1/F_s$, where F_s denotes the sampling frequency of 1 kHz (line 1 in Algorithm 1). To align the received samples in time, we call the SAMPLE_ALIGNER() procedure (Algorithm 2) with input $\boldsymbol{\mathcal{S}} \in \mathbb{R}^K, \boldsymbol{\mathcal{T}} \in \mathbb{R}^K$, and δ and output consisting of the aligned sample set $\boldsymbol{\mathcal{S}}^{\boldsymbol{\delta}} \in \mathbb{R}^L$ and time stamp set $\boldsymbol{\mathcal{T}}^{\boldsymbol{\delta}} \in \mathbb{R}^L$.

Next, we calculate the forecast horizon Δ at time t, which denotes the estimated number of samples during time interval $\boldsymbol{\mathcal{T}}^{\boldsymbol{\delta}}(L)$ between the last observed sample and target time t_0 (line 3

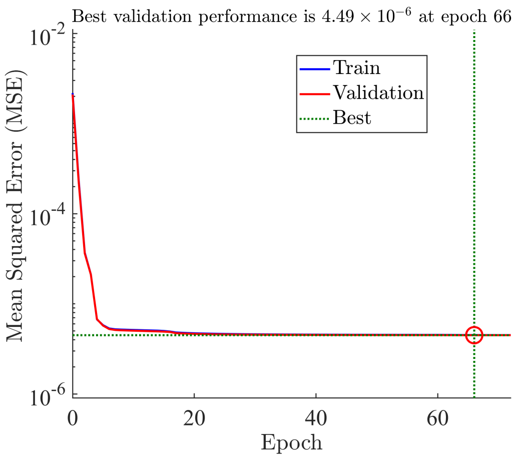


Figure 2.9 – Mean squared error vs. epoch for both train and validation data sets.

in Algorithm 1). Our objective is to forecast sample set $\boldsymbol{\Theta} = \{\theta_1, \theta_2, \ldots, \theta_\Delta\}$ for time stamp set $\{t_1^*, t_2^*, \ldots, t_\Delta^*\}$ to finally estimate sample θ^* at time t_0. Specifically, sample $\theta_i \in \boldsymbol{\Theta}$ is forecast by feeding our MLP with input vector $\boldsymbol{\mathcal{A}}_{i-1} \in \mathbb{R}^L$, where $\boldsymbol{\mathcal{A}_0} = \left(s_1^\delta, \ldots, s_L^\delta\right) \in \mathbb{R}^L$ and $\boldsymbol{\mathcal{A}_i} = (\boldsymbol{\mathcal{A}_{i-1}}(2), \boldsymbol{\mathcal{A}_{i-1}}(3), \ldots, \boldsymbol{\mathcal{A}_{i-1}}(L), \theta_i)$, i.e., each sample is forecast based on the preceding L samples. To further improve the forecasting accuracy, we estimate θ^* by performing a 2-point linear interpolation between $(t_{\Delta-1}^*, \theta_{\Delta-1})^T$ and $(t_\Delta^*, \theta_\Delta)^T$ (line 10 in Algorithm 1).

To create our training dataset, we used the available 6-DoF teleoperation traces. We used MATLAB to build and train a one-hidden-layer ANN. Our training data set comprised 59,710 force feedback samples with the waiting deadline set to $T_{thr} = 1$ millisecond. We used the so-called Levenberg-Marquardt training method for adjusting the weights until a desired input/output relationship was obtained. Prior to simulations, we applied brute force for determining the optimal value of the number of neurons in the hidden layer, which led us to set it to 5. Note that a so-called method of *early stopping* was applied to avoid overfitting. Overfitting happens when an ANN is highly specialized to the training data, while lacking the capability of generalizing to the validation data. For completeness, Fig. 2.9 illustrates the mean squared error vs. epoch for our training phase, which specifies how the underlying method of early stopping can avoid overfitting by detecting the epoch beyond which the training error still decreases while the validation error starts to increase. More specifically, according to Fig. 2.9, the training is stopped at epoch 66 where the validation mean squared error becomes 4.44×10^{-6}, thus avoiding the ANN from overfitting. After training, we used a new data set comprising 1,000 samples (different from the training data set) to evaluate the performance of our proposed sample forecaster in terms of mean squared error between the actual and forecast samples. It is worthwhile to mention that once the training was complete, the ANN was run on the HO side to provide the HO with forecast samples. Moreover, we note that the processing/running delay of the developed ANN on the order of microseconds was relatively small compared to the communication induced packet delays.

For completeness, we note that a one-hidden-layer MLP is also known as universal approximator. We decided to use a one-hidden-layer MLP since it is simple (i.e., easy to implement and train) yet achieves an accuracy that is good enough to approximate a wide variety of linear and/or non-linear functions. Beside longer training times, note that increasing the number of hidden layers in our considered one-hidden-layer MLP may result in over-fitting, which in turn may have a detrimental impact on its forecasting accuracy.

Note that in our considered FiWi enhanced LTE-A HetNets architecture in Fig. 2.7, all HOs and TORs are connected through a shared fiber backhaul, whose fiber reach does not exceed the typical 20 km of an IEEE 802.3ah EPON or up to 100 km in case of long-reach PONs. The limited fiber reach keeps the propagation delay below 0.1 millisecond and 0.5 millisecond, respectively. Thus, in a conventional EPON and in long-reach PONs the fiber propagation delay does not pose a challenge to meeting the 1 millisecond latency requirement of the Tactile Internet. However, an interesting question is how the 1-millisecond challenge of the Tactile Internet can be addressed for significantly larger geographical distances, e.g., connecting HOs in North America with TORs in Europe and/or Asia. This is where our proposed ESF module offers a potentially promising solution in that it decouples haptic feedback from the impact of extensive propagation delays, as typically encountered in wide area optical fiber networks. To see this, Fig. 2.1 illustrates our ESF module for the general case of a communication network with arbitrary propagation delays. The ESF module may be inserted at the edge of the communication network in close proximity to the HO. Rather than waiting for delayed haptic feedback samples that exceed the waiting deadline of 1 millisecond, the ESF module generates forecast samples and delivers them to the HO. Hence, the HO is enabled to perceive the remote task environment in real-time at a 1-millisecond granularity, resulting in a tighter togetherness, improved safety control, and increased reliability of the teleoperation systems. It should be noted, however, that a more rigorous experimental investigation would be needed to validate the viability of our proposed ESF module for real-word deployment scenarios with various wide area network propagation delays.

Clearly, the capability of our proposed ESF module to enable HOs to perceive the remote task environment in real-time at a 1-millisecond granularity requires a sufficiently high forecasting accuracy of haptic feedback samples, as discussed in more detail later on.

2.4 Delay Analysis

In this section, we develop our analytical framework to compute the average end-to-end delay and its distribution for local and non-local teleoperation with coexistent H2H traffic.

2.4.1 Assumptions

In our analysis, we make the following assumptions:

- *Single-hop WLAN:* MUs, HOs, and TORs are directly associated with an ONU-AP via a wireless single hop, whereby ONU-MPPs serve as ONU-APs (i.e., no MPs).

- *WiFi channel access:* Similar to [13], [69], [70], [71], [72], [73], [74], and [75], the WiFi channel access time governed by the IEEE 802.11 DCF is assumed to be exponentially distributed. This is justified by the DCF channel access mechanism, which includes carrier sensing, binary exponential back-off(s), and reattempts (if any) due to collisions and erroneous transmissions.

- *WiFi connectivity and WiFi offloading:* The WiFi connection and interconnection times of MUs are assumed to fit a truncated Pareto distribution, as validated via recent smartphone traces in [13]. The probability $P_{temporal}^{MU}$ that an MU is temporarily connected to an ONU-AP is estimated as $\bar{T}_{on}/(\bar{T}_{on}+\bar{T}_{off})$, whereby $\bar{T}_{on}$ and $\bar{T}_{off}$ denote the average WiFi connection and interconnection time, respectively. In this chapter, we assume that $\bar{T}_{on} \gg \bar{T}_{off}$ based on the fact that the recent smartphone traces reported in [13] indicate that the ratio $\bar{T}_{on}/(\bar{T}_{on}+\bar{T}_{off})$ has been constantly increasing. Hence, we assume that $P_{temporal}^{MU} \approx 1$ for MUs as well as HOs and TORs. Further, we assume that MUs offload their mobile traffic onto WiFi within the coverage area of an ONU-AP.

- *Traffic model:* MUs generate background Poisson traffic at mean packet rate λ_{BKGD} (in packets/second). Background traffic coming from ONUs with attached fixed (wired) subscribers is set to $\alpha_{PON} \cdot \lambda_{BKGD}$, where $\alpha_{PON} \geq 1$ is a traffic scaling factor for fixed subscribers that are directly connected to the backhaul EPON. Note that HOs and TORs generate traffic according to the different best fitting packet interarrival time distributions in Fig. 2.5.

2.4.2 Local Teleoperation

For notational convenience, let us use the term 'WiFi user' for all MUs, HOs, and TORs within the coverage area of an ONU-AP. We model each WiFi user as a GI/M/1 queue to account for the different packet interarrival time distributions under consideration. Let random variable D denote the delay experienced by any packet generated by a WiFi user, where D comprises queueing delay D_Q and service time D_S.

Suppose that packets arrive at rate λ at time instants $T_1, T_2, \ldots$, and assume that the interarrival times $T_{k+1} - T_k, k = 0, 1, \ldots$, are mutually independent, identically distributed random variables with distribution function $G(t) = P(T_{k+1} - T_k \leq t)$. Let N_k denote the number of packets in the system (i.e., queue and server) just prior to the arrival of packet k. By applying the theorem of total probability, we have

$$P\left(N_{k+1} = j\right) = \sum_{i=0}^{\infty} P\left(N_{k+1} = j \mid N_k = i\ \right) P\left(N_k = i\right), j = 0, 1, 2, \ldots; k = 0, 1, 2, \ldots \tag{2.4}$$

We define π_j as the probability that an arriving packet finds j packets in the system. A unique stationary distribution $\pi_j = \lim_{k\to\infty} P\left(N_k = j\right)$, $j = 0, 1, 2, \ldots$, exists if and only if $\rho = \frac{\lambda}{\mu} < 1$, where μ denotes the service rate, which is equal to $1/\mathbb{E}[D_S]$. By taking limits on both sides of Eq. (2.4) we obtain

$$\pi_j = \sum_{i=0}^{\infty} p_{ij}\pi_i \quad ; j = 0, 1, 2, \ldots, \tag{2.5}$$

where $p_{ij} = P\left(N_{k+1} = j \mid N_k = i\ \right)$ represent the state transition probabilities and $\sum_{j=0}^{\infty} \pi_j = 1$. Clearly, we have $p_{ij} = 0$, $\forall j > i + 1$, because an arriving packet can find at most one more packet in the system than was found by the preceding packet. The remaining state transition probabilities can be computed by considering the following three cases:

Case 1: The server is busy between T_k and T_{k+1}, i.e., $i \geq 0$ and $1 \leq j \leq i + 1$. The probability that arriving packet $k+1$ finds exactly j packets, given that the preceding packet k found i packets, is equal to the probability that exactly $i + 1 - j$ packets depart during interarrival time x. Thus, we have

$$P\left(N_{k+1}=j \mid N_k=i, T_{k+1}-T_k=x\ \right)=\frac{(\mu x)^{i+1-j}}{(i+1-j)!}e^{-\mu x}, i \geq 0, 1 \leq j \leq i+1. \tag{2.6}$$

Using interarrival time distribution function $G(x)$, we obtain

$$p_{ij}=\int_0^{\infty}\frac{(\mu x)^{i+1-j}}{(i+1-j)!}e^{-\mu x}dG(x), i \geq 0, 1 \leq j \leq i+1. \tag{2.7}$$

Case 2: The server becomes idle between two consecutive arrivals and arriving packet $k+1$ and preceding packet k find the system empty, i.e., $i=j=0$. This occurs if the service time of packet k is smaller than $T_{k+1}-T_k$. Hence, we have

$$p_{00}=\int_0^{\infty}\left(1-e^{-\mu x}\right)dG(x). \tag{2.8}$$

Case 3: This case is like case 2 except that preceding packet k found $i \geq 1$ packets. The time y, $T_k < y < T_{k+1}$, until $(i+1-s)$th service completion leaving one packet in the system has an Erlang distribution with density function $f(y)=\frac{(\mu y\ )^{i-s}}{(i-s)!}e^{-\mu y}\mu$. For $y<x$, the probability of service completion during the remaining interarrival time interval of length $x-y$ equals $1-e^{-\mu(x-y)}$. Thus, we have

$$p_{i0}=\int_0^{\infty}\int_0^{x}\left(1-e^{-\mu(x-y)}\right)\frac{(\mu y\ )^{i-1}}{(i-1)!}e^{-\mu y}\mu dy dG(x), i=1,2,\ldots \tag{2.9}$$

Lemma 2.1: *The stationary state probabilities π_j have a geometric distribution given by*

$$\pi_j=(1-\omega\ )\,\omega^j. \tag{2.10}$$

Proof. Let us consider the equilibrium probability state equations in Eq. (2.5) for $j \geq 1$. Substituting Eqs. (2.7) and (2.10) into Eq. (2.5) yields $\omega=\int_0^{\infty}e^{-(1-\omega)\mu x}dG(x)$, which can be rewritten as

$$\omega=\Phi\left(z\right)|_{z=(1-\omega)\mu}, \tag{2.11}$$

where $\Phi\left(z\right)$ is the Laplace-Stieltjes transform of $G(x)$. The equation has a unique root in $(0,1)$ if the queue is stable, i.e., $\rho<\lambda/\mu$. Substituting Eqs. (2.8)-(2.10) into Eq. (2.5) verifies Eq. (2.10) for $j=0$. □

Next, we compute the distribution of D_Q. Given N packets currently in the system, the probability $P\left(D_Q > t\right)$ is equal to $\sum_{i=1}^{\infty} \pi_i P\left(D_Q > t \mid N = i\right)$, which can be rewritten as

$$\sum_{j=0}^{\infty} (1-\omega)\, \omega^{j+1} P\left(D_Q > t \mid N = j+1\right).$$

The probability that a packet waits for longer than t in the queue given that $N = j+1$ is equivalent to the probability that the number of departures during time interval t is smaller than or equal to j, which is given by

$$P\left(D_Q > t \mid N = j+1\right) = \sum_{i=0}^{j} \frac{(\mu t)^i}{i!} e^{-\mu t}. \tag{2.12}$$

Thus, we have

$$P\left(D_Q > t\right) = (1-\omega)\omega \sum_{j=0}^{\infty} \omega^j \sum_{i=0}^{j} \frac{(\mu t)^i}{i!} e^{-\mu t}, \tag{2.13}$$

which reduces to $P\left(D_Q > t\right) = \omega e^{-(1-\omega)\mu t}$. The cumulative distribution function (CDF) of D_Q is then given by

$$F_{D_Q}(t) = P(D_Q \leq t) = 1 - \omega e^{-(1-\omega)\mu t}. \tag{2.14}$$

Next, let us consider D_S, whose CDF is given by

$$F_{D_S}(t) = P(D_S \leq t) = 1 - e^{-\mu t}. \tag{2.15}$$

To compute the service rate μ in Eq. (2.15), we defined the two-dimensional Markov process $(s(t), b(t))$ shown in Fig. 2.10 under unsaturated non-Poisson traffic conditions and estimated the average service time $\mathbb{E}(D_S)$ in a WLAN using the IEEE 802.11 DCF for access control, whereby $b(t)$ and $s(t)$ denote the random backoff counter and size of contention window at time t, respectively. Let P_f and W_i denote the probability of a failed transmission attempt (i.e., collision or erroneous transmission) and contention window size at the back-off stage i, respectively. Note that the back-off stage i is incremented after each failed attempt up to the maximum value m, while the contention window is doubled at each stage, i.e., $W_i = 2^i W_0$.

A WiFi user is in idle state if: (i) a successfully transmitted packet leaves the system without any waiting packet in the queue, and (ii) no packet arrives during the current time slot given that the user was in idle state in the preceding time slot. We note that for non-Poisson arrival, these two

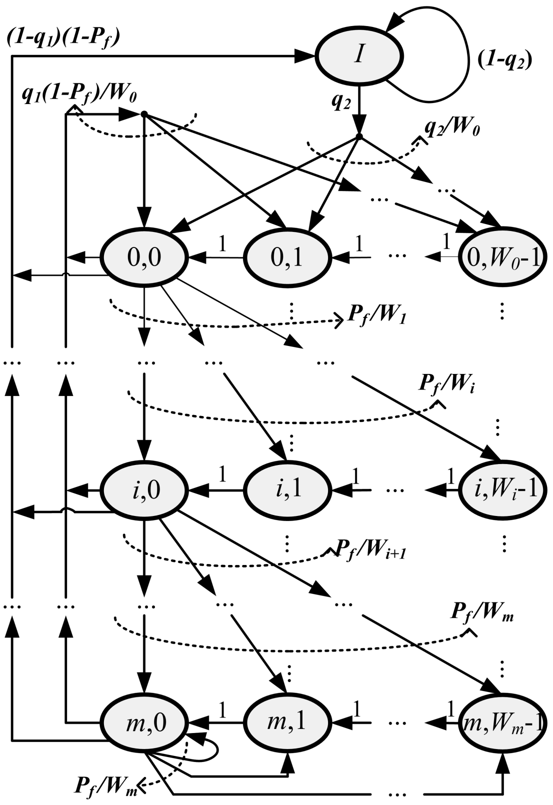


Figure 2.10 – Two-dimensional Markov process.

events are not identical and should be calculated separately. Define π^*_j and $\hat{\pi}_j$ as the probability that a departing packet leaves j packets in the queue (i.e., from the viewpoint of the departing packet), and the fraction of time during which j packets are present in the queue (i.e., from the viewpoint of an outside observer), respectively. According to Fig. 2.10, $1 - q_1$ is equal to the probability that a departing packet leaves the queue without any waiting packet, thus $1 - q_1 = \pi^*_0$. On the other hand, q_2 is the probability that at least one packet arrives during the current time slot given that the user was in idle state in the preceding time slot. This, however, does depend on the time interval during which the system has been in idle state so far. Nevertheless, for a slot duration being much smaller than the mean interarrival time, it is reasonable to estimate q_2 by $1 - \hat{\pi}_0 \approx \frac{\lambda}{\mu}$. Note that, according to Burke's theorem, $\pi^*_j = \pi_j$ holds for any arrival model, whereas $\hat{\pi}_j = \pi_j$ is valid only for Poisson arrival.

After finding the stationary distributions

$$b_{i,k} = \lim_{k \to \infty} P\left(s(t) = i, b(t) = k\right), \forall k \in [0, W_i - 1], i \in [0, m],$$

the probability τ that a WiFi user attempts to transmit in a given time slot is obtained as

$$\tau = \sum_{i=0}^{m} b_{i,0} = \frac{\frac{2\left(1-2P_f\right)q_2}{2(1-q_1)(1-P_f)(1-2P_f)}}{\frac{q_2\left[(W_0+1)(1-2P_f)+W_0 P_{eq}(1-(2P_f)^m)\right]}{2(1-q_1)(1-P_f)(1-2P_f)}+1}. \tag{2.16}$$

The probability of a failed transmission attempt $P_{f,i}$ by WiFi user i is given by

$$1 - P_{f,i} = \left(1 - p_{e,i}\right)\left(1 - p_{c,i}\right), \tag{2.17}$$

where $p_{e,i}$ and $p_{c,i}$ denote the probability of an erroneous transmission and the probability of a collision, respectively. Note that WiFi subscriber i does not experience a collision if the remaining users don't attempt to transmit, thus $1 - p_{c,i} = \prod_{v:v\neq i}\left(1 - \tau_v\ \right)$. Moreover, $p_{e,i}$ is estimated by $1 - \left(1 - p_b\right)^{\bar{L}_i}$, where $\bar{L}_i$ and p_b is the average length of a packet transmitted by WiFi user i and the bit error probability, respectively.

The probability of a collision-free packet transmission P_s given that there is at least one transmission attempt is given by $\frac{1}{P_{tr}}\left(\sum_i \tau_i \prod_{v,v\neq i}\left(1-\tau_v\ \right)\right)$, whereby the probability P_{tr} that there is at least one transmission attempt is equal to $1 - \prod_i\left(1 - \tau_i\ \right)$. The average slot duration E_s is then obtained as

$$E_s = \left(1 - P_{tr}\right)\epsilon + P_{tr}\left(1 - P_s\right)T_c + P_{tr}P_sP_eT_e + P_{tr}P_s\left(1 - P_e\right)T_s, \tag{2.18}$$

where T_c, T_e, and T_s are given in [75]. We then obtain $\mathbb{E}(D_S)$ as

$$\mathbb{E}(D_S) = \frac{1}{\mu} = \sum_{k=0}^{\infty} p_e^k\left(1 - p_e\right)\left[\sum_{j=0}^{\infty} p_c^j\left(1 - p_c\right)\cdot\left(\left(\sum_{b=0}^{k+j}\frac{2^{min(b,m)}W_0 - 1}{2}E_s\right) + jT_c + kT_e + T_s\right)\right]. \tag{2.19}$$

In order to obtain the steady-state values of q_1, q_2, P_f, τ, and μ, we numerically solve the system of non-linear equations (2.19), (2.17), (2.16), and (2.10).

The CDFs of D_Q (2.14) and D_S (2.15) are used to calculate the CDF of $D = D_Q + D_S$ at a WiFi user as follows

$$F_D(t) = P(D \leq t) = \int_0^t F_{D_S}(t-u)dF_{D_Q}(u). \tag{2.20}$$

The end-to-end delay of local teleoperation $D^{E2E}_{LT(i \to j)}$ between WiFi users i and j communicating via ONU-AP$_z$ is obtained as $D_{i \to \text{ONU-AP}_z} + D_{\text{ONU-AP}_z \to j}$, whose CDF is given by

$$F_{D^{E2E}_{LT(i \to j)}}(t) \;=\; P(D^{E2E}_{LT(i \to j)} \leq t) = \int_0^t F_{D_{i \to \text{ONU-AP}_z}}(t-\zeta) dF_{D_{\text{ONU-AP}_z \to j}}(\zeta), \tag{2.21}$$

where the CDFs $F_{D_{i \to \text{ONU-AP}_z}}(t)$ and $F_{D_{\text{ONU-AP}_z \to j}}(t)$ are calculated similar to Eq. (2.20).

2.4.3 Non-Local Teleoperation

The average end-to-end delay of non-local teleoperation between WiFi user i and WiFi user j associated with ONU-AP$_m$ and ONU-AP$_n$, respectively, is given by

$$\bar{D}^{E2E}_{NLT(i \to j)} = \bar{D}_{i \to \text{ONU-AP}_m} + \bar{D}^u_{PON} + \bar{D}^d_{PON} + \bar{D}_{\text{ONU-AP}_n \to j}, \tag{2.22}$$

where $\bar{D}_{i \to \text{ONU-AP}_m}$ and $\bar{D}_{\text{ONU-AP}_n \to j}$ denote the expected values of $D_{i \to \text{ONU-AP}_m}$ and $D_{\text{ONU-AP}_n \to j}$, respectively. Both expected values can be obtained from Eq. (2.20). Note that $\bar{D}^u_{PON}$ and $\bar{D}^d_{PON}$ denote the average delay of the backhaul EPON in the upstream and downstream direction, respectively, which are given by $\Phi\left(\rho^u, \bar{L}, \varsigma^2_L\ , c_{PON}\right) + \bar{L}/c_{PON} + 2\tau_{PON}\frac{2-\rho^u}{1-\rho^u} - B^u$ and $\Phi\left(\rho^d, \bar{L}, \varsigma^2_L\ , c_{PON}\right) + \bar{L}/c_{PON} + \tau_{PON} - B^d$, respectively; whereas ρ^u is the traffic intensity in upstream, ρ^d is the traffic intensity in downstream, τ_{PON} is the propagation delay between ONUs and OLT, c_{PON} is the EPON data rate, $\Phi(\cdot)$ denotes the well-known Pollaczek-Khintchine formula, and B^u and B^d are obtained as $\Phi\left(\frac{\bar{L}}{\Lambda c_{PON}} \sum_{i=1}^{O} \sum_{q=1}^{O} \Gamma^{PON}_{iq}, \bar{L}, \varsigma^2_L\ , c_{PON}\right)$, where O is the number of ONUs and Γ^{PON}_{iq} is the traffic emanating from ONU$_i$ to ONU$_q$, and Λ denotes the number of wavelengths in the WDM PON [13].

2.5 Results

In this section, we present the trace-driven simulation results along with the numerical results derived from the analysis. Note that the obtained simulation results include confidence intervals at 95% confidence level. The following results were obtained by using the FiWi network parameter settings listed in Table 2.2. We assume that MUs, HOs, and TORs are directly connected to their associated MPPs, i.e., MPPs serve as conventional WLAN APs. By default, let us consider 4

Table 2.2 – FiWi network parameters & Default values

Parameter	Value
Minimum contention window W_0	16
Maximum back-off stage H	6
Empty slot duration ϵ	9 μs
DIFS	34 μs
SIFS	16 μs
PHY Header	20 μs
MAC Header	36 bytes
RTS	20 bytes
CTS	14 bytes
ACK	14 bytes
Line rate r in wireless fronthaul	600 Mbps
Uplink and Downlink data rate r_{PON} in PON	1 Gbps
Propagation delay δ_p in WMN	3.33 μs
l_{BKGD}	1500 Bytes
p_b	10^{-6}
N_{DoF}	6
l_{PON}	20 km

ONU-APs, each with 4 associated MUs, whereby two MUs communicate with each other via their associated ONU-AP using an IEEE 802.11n WLAN (i.e., local H2H communications) while the two remaining MUs communicate with two uniformly randomly selected MUs associated with a different ONU-AP by using a backhaul IEEE 802.3ah 1Gb/s EPON with a typical fiber range of 20 km (i.e., non-local H2H communications). Furthermore, let us consider four conventional ONUs, serving fixed (wired) subscribers that are all involved in non-local H2H communications among each other. The MUs and fixed subscribers generate background traffic at a mean rate of λ_{BKGD} and $\alpha_{PON} \cdot \lambda_{BKGD}$, respectively. Note that $\alpha_{PON} \geq 1$ is a traffic scaling factor for fixed subscribers that are directly connected to the backhaul EPON. Figure 2.11 depicts the average end-to-end delay of MUs vs. mean background traffic rate λ_{BKGD} with different $\alpha_{PON} \in \{1, 50, 100\}$ for both local and non-local H2H communications in FiWi enhanced LTE-A HetNets. The figure shows that an average end-to-end delay of $10^0 = 1$ ms can be achieved for non-local H2H communications for a wide range of background traffic loads.

Next, we include teleoperation and investigate the interplay between Tactile Internet traffic and the above H2H background traffic. Towards this end, we consider the above scenario and replace two MUs with a pair of HO and TOR in the coverage area of each ONU-AP for local teleoperation

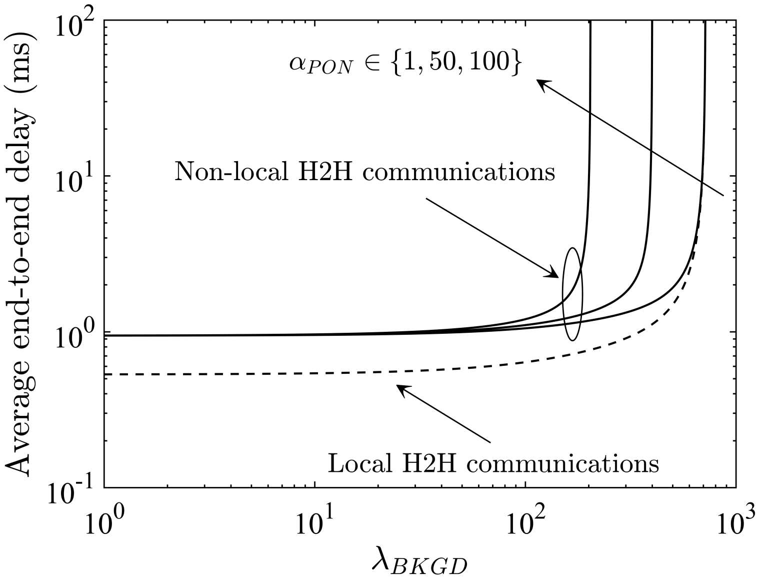


Figure 2.11 – Average end-to-end delay of mobile users (MUs) vs. mean background traffic rate λ_{BKGD} (packets/second) for local and non-local H2H communications with different $\alpha_{PON} \in \{1, 50, 100\}$.

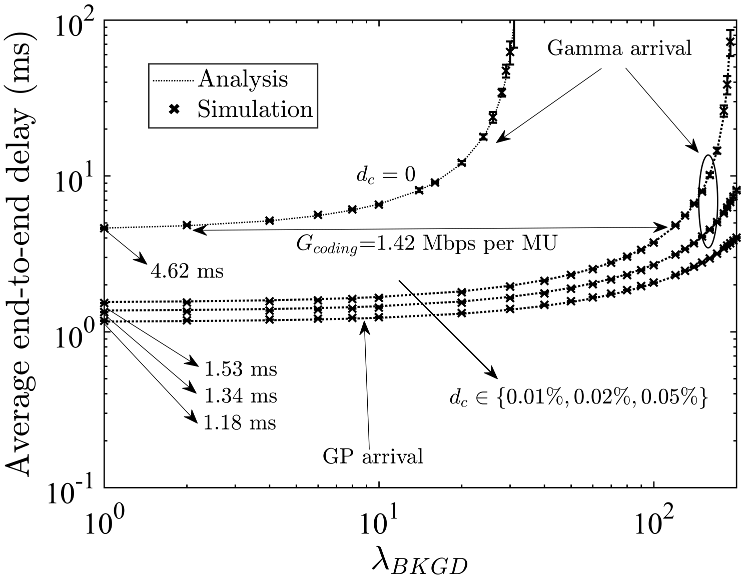


Figure 2.12 – Average end-to-end delay of human operators (HOs) vs. mean background traffic rate λ_{BKGD} (packets/second) for local teleoperation with and without deadband coding in the command path for different $d_c \in \{0, 0.01\%, 0.02\%, 0.05\%\}$ ($\alpha_{PON} = 100$ fixed).

with and without deadband coding in the command path. Specifically, we consider our findings on 6-DoF teleoperation in Fig. 2.5(a) and accordingly assume gamma and GP distributed haptic packet arrivals for $d_c \in \{0, 0.01\%, 0.02\%\}$ and $d_c = 0.05\%$, respectively. Fig. 2.12 depicts the

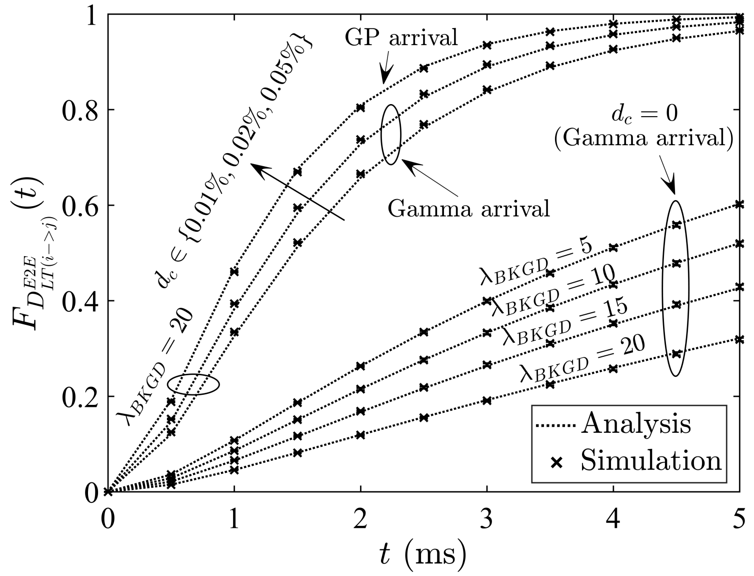


Figure 2.13 – End-to-end delay CDF $F_{D^{E2E}_{LT(i\to j)}}(t)$ of local teleoperation.

average end-to-end delay of HOs vs. mean background traffic rate λ_{BKGD} along with verifying trace-driven simulations based on our 6-DoF haptic traces and packetization procedure described in Section 2.2.1. We observe from Fig. 2.12 that without deadband coding ($d_c = 0$) the minimum achievable average end-to-end delay experienced by HOs equals 4.62 ms, thus missing the Tactile Internet target of 1 ms. However, note that this target can be achieved with deadband coding for increasing d_c. For illustration, Figure 2.12 shows that we achieve a minimum average end-to-end delay of 1.18 ms for $d_c = 0.05\%$. In addition to decreasing the latency of HOs, note that deadband coding also has a beneficial impact on the admissible background traffic load of MUs due to the reduced haptic packet rates. To see this, let us define the coding gain G_{coding} as the difference between the maximum admissible throughput of MUs in teleoperation with and without deadband coding, while not violating a certain upper average end-to-end delay limit. For instance, for a given upper limit of 4.8 ms a coding gain of $G_{coding} = 1.42$ Mbps per MU can be achieved in our teleoperation scenario by increasing d_c from 0 to 0.01%, as depicted in Fig. 2.12. Note that overall the presented analytical results and verifying trace-driven simulation results (shown with 95% confidence interval) match very well.

Figure 2.13 provides useful insights into the upper end-to-end delay bounds by showing its CDF $F_{D^{E2E}_{LT(i\to j)}}(t)$ for the scenario of Fig. 2.12. Notably, we observe that for $d_c = 0.05\%$ and a high

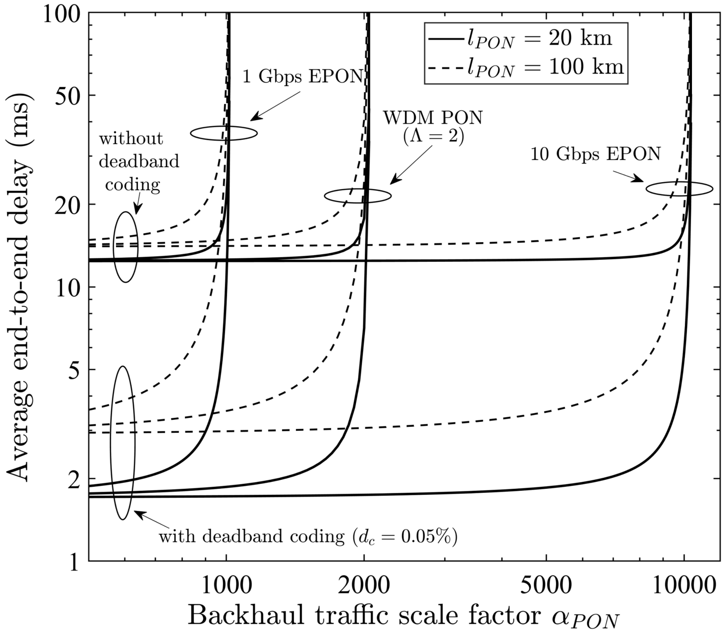


Figure 2.14 – Average end-to-end delay of human operators (HOs) vs. backhaul traffic scale factor α_{PON} of fixed subscribers ($\lambda_{BKGD} = 20$ packets/second fixed) for non-local teleoperation across different NG-PON backhaul infrastructures.

background traffic rate of $\lambda_{BKGD} = 20$ packets/second (top curve), the end-to-end delay stays below 2 ms with a probability as high as 0.8.

To provide insights into the impact of different NG-PON backhaul infrastructures in the case of non-local teleoperation, Fig. 2.14 depicts the average end-to-end delay performance of HOs vs. backhaul traffic scale factor α_{PON} of fixed subscribers with the mean background traffic rate set to $\lambda_{BKGD} = 20$ packets/second. For comparison, we consider a conventional 1 Gbps EPON, a high-speed 10 Gbps EPON, and a WDM PON with $\Lambda = 2$ wavelength channels, each operating at 1 Gb/s. Note that for all three considered NG-PONs we include a conventional fiber reach of $l_{PON} = 20$ km as well as its respective long-reach counterpart with an extended fiber reach of $l_{PON} = 100$ km. We observe from Fig. 2.14 that the use of deadband coding ($d_c = 0.05\%$) is instrumental in lowering the average end-to-end delay below 10 ms for all NG-PON backhaul infrastructures under consideration. The figure also confirms previous findings (see Section 2.1) that 10G PON and WDM technologies represent cost-effective solutions to support 5G low-latency applications over a wide range of backhaul traffic loads by sharing a common optical transport platform among fixed subscribers, MUs, and HOs.

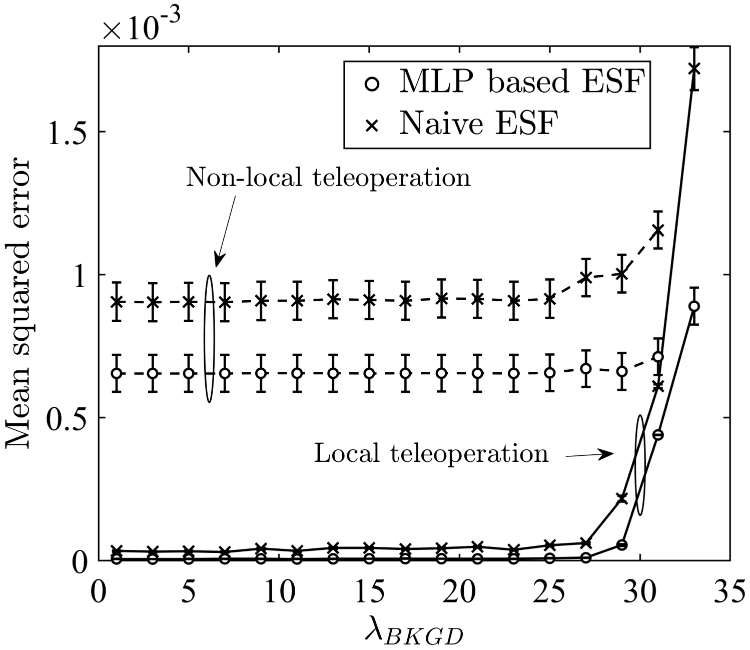


Figure 2.15 – Comparison of forecasting accuracy between proposed MLP based and naive ESF schemes for local and non-local teleoperation without deadband coding in the feedback path ($d_f = 0$).

We have seen in the results above that deadband coding is effective in decreasing the average end-to-end delay by reducing the haptic packet rate. Nevertheless, some haptic packets may still experience an instantaneous delay that exceeds the desired waiting deadline on the order of 1 millisecond until their reception due to varying traffic conditions and MAC layer queueing times. To ensure that the HO receives expected haptic packets before the deadline, our proposed MLP based ESF module may be used as a complementary technique to deadband coding in the feedback path. Figure 2.15 compares the forecasting accuracy of our proposed MLP based ESF scheme with a naive ESF scheme, where the forecast sample is simply set to the last received sample. In our simulation, we used our 6-DoF teleoperation traces to train a one-hidden-layer MLP by using 59,710 force feedback samples with the waiting deadline set to $T_{thr} = 1$ ms. Figure 2.15 clearly shows the superior forecasting accuracy of our proposed MLP based ESF scheme in terms of mean squared error over a wide range of λ_{BKGD} for both local and non-local teleoperation scenarios, whereby a low mean squared error is achievable in the former scenario. Specifically, for non-local teleoperation, our MLP based ESF scheme decreases the mean squared error from roughly 0.9 to 0.65, translating into an improvement of 27.8%. For local teleoperation, it is able to keep the mean squared error close to zero between 0.006 and 0.007 $\times 10^{-3}$ at low to medium background traffic load λ_{BKGD}. Note that the observed performance improvement is due to the relatively high autocorrelation in the haptic feedback samples that allows our proposed MLP based ESF module to achieve a more accurate forecast compared to that of the naive ESF scheme.

2.6 Discussion

Motivated by the notable amount of correlation depth seen within the haptic samples in the feedback path of our available traces, the proposed multiple-sample-ahead-of-time forecasting scheme was shown to be capable of providing accurate forecast samples, to be delivered to the HO immediately. We note, however, that our available traces are restricted to only 1-DoF teleoperation and 6-DoF telesurgery systems under consideration, thus posing limitations to our reported results above due to being application- and/or device-specific, especially given that the emerging advanced robotic teleoperation systems with relatively larger number of DoF may rely on the exchange of a wider variety of haptic feedback samples, rather than simply 6-DoF force-torque samples. We note that this study may be considered as a starting point to suggest methods of investigating the haptic traffic characteristics by means of a more comprehensive trace-driven study. Hence, in order to validate our findings, it is necessary to look into a more diverse set of traces gathered from various teleoperation systems. Moreover, further investigations are needed to generalize the proposed ANN-based ESF module to other deployment scenarios apart from the studied 1- and 6-DoF teleoperation use cases. Toward this end, the applicability of more sophisticated machine learning models such as deep neural networks and/or long-short-term-memory neural networks may be further investigated for detecting the patterns within highly complicated multi-dimensional data, thus providing the HO with more accurate forecasts. Last but not least, we note that although achieving a very low-latency on the order of < 10 ms and providing the HO with accurate forecasting capabilities may be necessary to realize immersive Tactile Internet experiences, they are not sufficient to guarantee a desired QoE. This suggests for conducting subjective and objective studies to allow for both qualitative and quantitative measuring of how closely an HO is coupled with the involved experience.

2.7 Conclusions

We have seen that there is a significant overlap among 5G, IoT, and the Tactile Internet in that they share various important design goals, including very low latency, ultra-high reliability, and integration of data-centric technologies. This chapter described how FiWi enhanced LTE-A HetNets leveraging low-cost data-centric EPON and WiFi technologies for fiber backhaul sharing and WiFi offloading may help realize not only the aforementioned shared design goals but also the key attributes of end-to-end co-DBA of both PON and wireless network resources, decentralization, and edge intelligence in support of future 5G low-latency applications over a common optical transport platform.

Our focus was on the emerging Tactile Internet as one of the most interesting 5G low-latency applications for creating novel immersive experiences. We reviewed the HITL-centric design principles that add a new dimension to the human-to-machine interaction via the Internet and set the Tactile Internet aside from the more machine-centric IoT. Exploiting the human perception of haptics to reduce the haptic packet rate by means of deadband coding, we derived haptic traffic models from teleoperation experiments. Our haptic trace analysis showed that assuming Tactile Internet traffic to be Pareto distributed was not valid for the analyzed traffic, while assuming it to be Poisson traffic was valid only in a special case. In general, we observed that command and feedback paths of teleoperation systems can be jointly modeled by generalized Pareto, gamma, or deterministic packet interarrival time distributions, depending on the given value of the respective deadband parameters.

We elaborated on the importance of the decentralized nature of WLAN's access protocol DCF to realize low-latency FiWi enhanced LTE-A HetNets. Furthermore, by exploiting their inherent distributed processing and storage capabilities, we investigated the potential of enabling immersive teleoperation experiences for human operators by introducing machine learning at the optical-wireless interface of FiWi enhanced LTE-A HetNets. Our proposed MLP based ESF module compensates for delayed haptic feedback samples by means of multiple-sample-ahead-of-time forecasting for a tighter togetherness, improved safety control, and increased reliability.

Chapter 3

Context- and Self-Awareness for Human-Agent-Robot Task Coordination

This chapter contains material extracted from the following publication:

[[50]] A. Ebrahimzadeh, M. Chowdhury, and M. Maier. Human-Agent-Robot Task Coordination in FiWi-based Tactile Internet Infrastructures Using Context- and Self-Awareness. *IEEE Transactions on Network and Service Management*, vol. 16, no. 3, pp. 1127-1142, Sept. 2019.

In the following, my key contributions in the aforementioned publication are explained in greater detail: (1) I largely contributed to writing the whole manuscript, (2) I developed the analytical framework, (3) I conducted the algorithmic work, and (4) I ran the simulations.

3.1 Introduction

Today's telecommunication networks enable us to connect devices and people for an unprecedented exchange of audiovisual and data content. With the advent of commercially available haptic/tactile sensory and display devices, conventional triple-play (i.e., audio, video, and data) content communication now extends to encompass the real-time exchange of haptic information (i.e., touch and actuation) for the remote control of physical and/or virtual objects through the Internet. This paves the way towards realizing the so-called *Tactile Internet* [3], whereby human-machine interactions will convert today's content-delivery networks into skillset/labor delivery networks [56]. The Tactile Internet holds great promise to have a profound socio-economic impact on a broad array of applications in our everyday life, ranging from industry automation and transport systems to healthcare, telesurgery, and education [46].

Beside the design of low-latency/jitter and highly reliable networking infrastructures, a key challenge little discussed in the existent Tactile Internet literature is how we can make sure that the potential of the Tactile Internet be unleashed for a race with (rather than against) machines. The overarching goal of the Tactile Internet should be the production of new goods and services by means of empowering rather than automating machines that complement humans rather than substitute for them. In the future, coworking with robots will require human expertise in the coordination of the human-robot symbiosis. A promising approach toward achieving advanced human-machine coordination by means of a superior process for fluidly orchestrating human and machine coactivity may be found in the still young field of human-agent-robot-teamwork (HART) research [76]. In HART, the dynamic allocation of functions and tasks between humans and machines, which may vary over time or be unpredictable in different situations, plays a central role. In particular, with the rise of increasingly smarter machines, the historical humans-are-better-at/machines-are-better-at (HABA/MABA) approach to decide which tasks are best performed by people or machines rather than working in concert has become obsolete [77]. In addition, we note that any technological advance can be labor-saving or capital-saving. In either case, regardless of the speed with which robots approach or even exceed human skill sets, the key to the effect of the new technologies on human society is who owns the technologies. We would lose our jobs if other persons owned our replacement technologies. By contrast, if users owned them, humans would have their current earnings and their time freed from labor to explore other productive activities [78].

In this chapter, we leverage on our recently proposed concept of FiWi enhanced LTE-A HetNets, which were shown to achieve the 5G and Tactile Internet key requirements of very low latency on the order of 1-10 milliseconds and ultra-high reliability by unifying coverage-centric 4G mobile networks and capacity-centric FiWi broadband access networks based on low-cost, data-centric Ethernet NG-PON and Gigabit-class WLAN technologies [13]. While necessary, though, the design of reliable low-latency converged communication network infrastructures is not sufficient to realize the full potential of the Tactile Internet.

Depending on the context-awareness of future Tactile Internet applications, tasks may be classified into three different categories: (*i*) location-dependent physical-only tasks (e.g., lifting an object), (*ii*) location-independent digital-only tasks (e.g., object classification from a captured image, which might be offloaded for computation at a remote cloud or nearby cloudlet), or (*iii*) location-dependent physical/digital tasks that include both types of tasks (e.g., assemblage followed by a unit test). As users will need to request robot assistance from time to time, mapping these requests to the robots stands as an optimization problem, whose objective is to minimize not only the task completion time but also the OPEX and robot energy consumption. The difficulty of solving such a problem lies in the following reasons. First, it is clear that we are dealing with different conflicting objectives, which makes it challenging to obtain a satisfactory result, especially for large-sized problems. Second, in real-world scenarios, there is no a priori knowledge of the task arrival times, making it almost impossible to obtain the optimal solution. Third, and more importantly, to minimize the energy consumption of mobile robots, the task coordinator requires global knowledge of the all system parameters, including in particular the local parameters of the mobile robots, which may not be willing to share such private information. In this chapter, we use context-awareness to develop a HART-centric task coordination algorithm that minimizes the completion time of physical/digital tasks as well as the OPEX by spreading ownership of robots across mobile users. In addition, we capitalize on self-awareness to improve the performance of a given robot by identifying its capabilities as well as the objective requirements by means of optimal motion planning to minimize its energy consumption as well as traverse time. Our proposed self- and context-aware HART-centric allocation scheme for both physical and digital tasks is used to coordinate the automation and augmentation of mutually beneficial human-machine coactivities across a FiWi based Tactile Internet infrastructure. In particular, the contributions of this chapter are as follows:

- We formulate a multi-objective optimization problem to minimize the task completion time, energy consumption, and OPEX for multi-robot task allocation in the Tactile Internet over FiWi enhanced networks.

- We develop a context-aware HART-centric task coordination algorithm that minimizes the completion time of physical/digital tasks, while paying particular attention to reducing OPEX by spreading ownership of robots across mobile users.
- We propose a self-aware optimal motion planning algorithm, which runs locally at the mobile robots, with the objective to find the best trade-off between traverse time and energy consumption by leveraging on *local self-awareness* of the mobile robots to identify their respective limitations and capabilities as well as objective requirements for accomplishing the allocated tasks.
- We provide an analytical framework to calculate the average packet transmission delay and human-robot connection reliability, two key attributes of the Tactile Internet.

The remainder of the chapter is structured as follows. Section 3.2 describes our considered FiWi based Tactile Internet infrastructures for HART-centric task coordination in greater detail, followed by motion and energy consumption models for mobile robots. In Section 3.4, we present our self-aware optimal motion planning algorithm. In Section 3.3, we develop our multi-objective optimization problem considering characteristics and key parameters of mobile robots and tasks, which is subsequently solved by our proposed HART-centric context-aware multi-robot task coordination algorithm. Our delay and reliability analysis is presented in Section 3.5. In Section 3.6, we report on our obtained results and findings. Section 3.7 concludes the chapter.

3.2 System Model

3.2.1 Network Architecture

Figure 3.1 illustrates the generic network architecture of our considered FiWi enhanced LTE-A HetNets. The optical backhaul consists of a TDM/WDM IEEE 802.3ah/av 1/10 Gb/s EPON with a typical fiber length of 20 km between the central OLT and remote ONUs. The EPON may comprise multiple stages, each stage separated by a wavelength-broadcasting splitter/combiner or a wavelength multiplexer/demultiplexer. There are three different subsets of ONUs. An ONU may either serve fixed (wired) subscribers. Alternatively, it may connect to a cellular network BS or an IEEE 802.11n/ac/s WLAN MPP, giving rise to collocated ONU-BS or ONU-MPP, respectively.

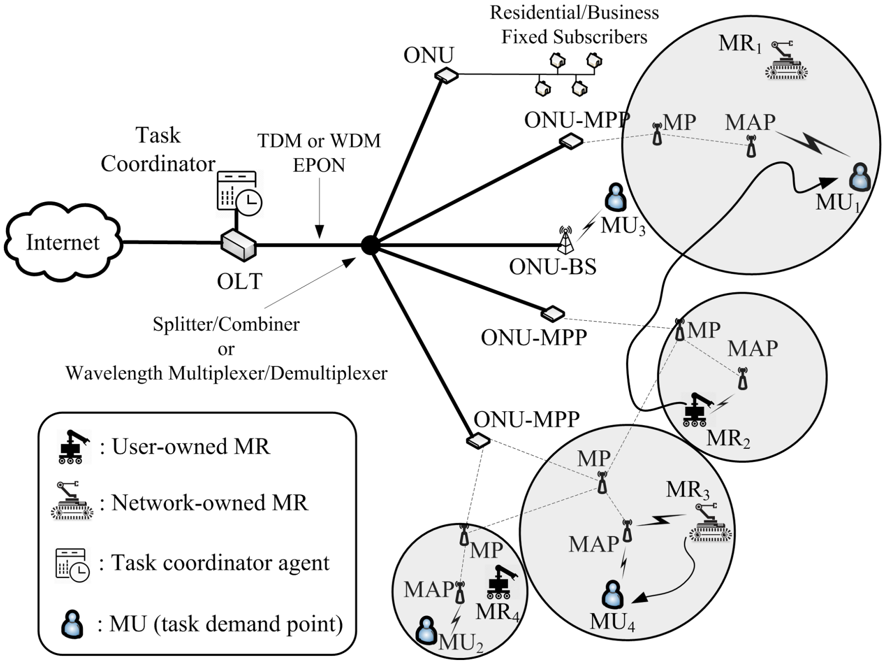


Figure 3.1 – Generic architecture of FiWi based Tactile Internet network infrastructure for multi-robot task coordination.

Depending on her trajectory, an MU may communicate through the cellular network and/or WLAN mesh front-end, which consists of ONU-MPPs, intermediate MPs, and MAPs.

Note that tasks arrive at random time instants at the MUs, which act as the demand points. The MUs then send their demands upstream to the task coordinator agent, which is co-located with the OLT (see also Fig. 3.1), via the wireless front-end and EPON backhaul until they reach the OLT. The task coordinator agent is responsible for allocating the incoming tasks to mobile robots (MRs), which may be owned by either the users or the network operator. After receiving the task allocation from the OLT, the selected MR moves toward the demand point and collaboratively executes the physical and/or digital tasks. After successfully executing the tasks, the MR transmits the results/output of the physical and/or digital task to the task demand point (i.e., MU).

3.2.2 Energy and Motion Models of Mobile Robots

For an MR with forward translational velocity powered by a direct current (DC) motor, we use the detailed model presented in [79]. Specifically, let $v(t)$ and $a(t)$ denote the velocity and acceleration profile of the MR, respectively. The energy consumption of the brushed DC motor deployed at the

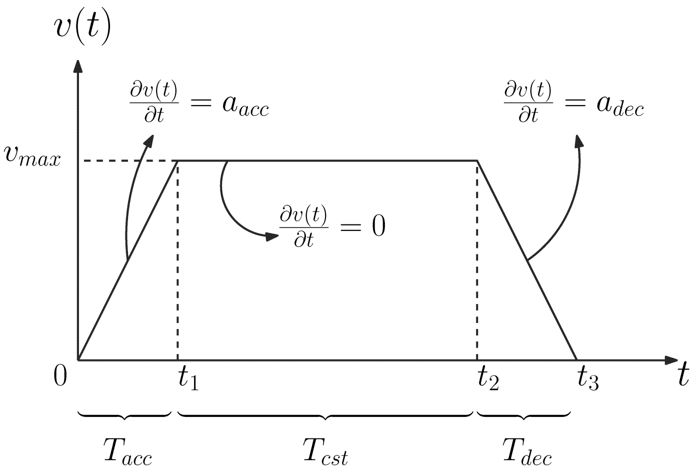


Figure 3.2 – Trapezoidal velocity profile of MRs.

MR is given by

$$E = \int_0^{t_3} e(t)i(t).dt = \int_0^{t_3} \left[c_1a^2(t) + c_2v^2(t) + c_3v(t) + c_4 + c_5a(t) + c_6v(t)a(t)\right].dt, \tag{3.1}$$

where $i(t)$ and $e(t)$ denote the instantaneous current and voltage of the DC motor, respectively, while the constants $\{c_i\}_{i=1}^{6}$, given in [79], are combinations of the motor parameters and depend on the design of the motor and surface on which the robot traverses.

Next, let us consider the trapezoidal velocity profile of the MR, shown in Fig. 3.2. The profile indicates that along a given path the MR accelerates from rest during T_{acc}, traverses with constant velocity v_{max} during T_{cst}, and then decelerates during T_{dec} until it returns to rest. Based on the considered velocity profile, the distance Δd traversed by the MR is given by

$$\Delta d = \int_0^{T_{trav}} v(t).dt = (T_{acc} + 2T_{cst} + T_{dec})\frac{v_{max}}{2}, \tag{3.2}$$

which yields

$$T_{acc} + T_{cst} = \frac{\Delta d}{v_{max}} \tag{3.3}$$

by assuming $a_{acc} = -a_{dec}$. Having

$$T_{cst} = \omega_d \frac{\Delta d}{v_{max}}, \tag{3.4a}$$

$$T_{acc} = T_{dec} = (1-\omega_d)\frac{\Delta d}{v_{max}}, \tag{3.4b}$$

the traverse time T_{trav} is then equal to

$$\begin{aligned} T_{trav} &= (1-\omega_d)\frac{\Delta d}{v_{max}} + \omega_d\frac{\Delta d}{v_{max}} + (1-\omega_d)\frac{\Delta d}{v_{max}} \\ &= (2-\omega_d)\frac{\Delta d}{v_{max}}, \end{aligned} \tag{3.5}$$

with $\omega_d \in [0, 1)$. Clearly, T_{trav} is a monotonically decreasing function of ω_d with the upper and lower bounds given by

$$T^U_{trav} = \lim_{\omega_d \to 0} T_{trav} = 2\frac{\Delta d}{v_{max}}, \tag{3.6a}$$

$$T^L_{trav} = \lim_{\omega_d \to 1} T_{trav} = \frac{\Delta d}{v_{max}}. \tag{3.6b}$$

Lemma 3.1: For the velocity profile shown in Fig. 3.2, the energy consumption E_{trav} of the MR to traverse a given distance Δd is given by

$$E_{trav} = E(\omega_d) = \frac{2c_1 v^3_{max}}{(1-\omega_d)\Delta d} + \frac{c_2}{3}(\omega_d + 2)v_{max}\Delta d + c_3\Delta d + (2-\omega_d)\frac{c_4 \Delta d}{v_{max}}. \tag{3.7}$$

Proof. See Appendix A.1.

Note that $E(\omega_d)$ is a convex function of ω_d, as $\frac{\partial^2 E(\omega_d)}{\partial \omega_d^2} = \frac{4c_1 v^3_{max}}{\Delta d(1-\omega_d)^3} > 0$ for $\omega_d \in [0, 1)$.

Lemma 3.2: $E(\omega_d)$ has a local minimum in interval $(0, 1)$ if and only if

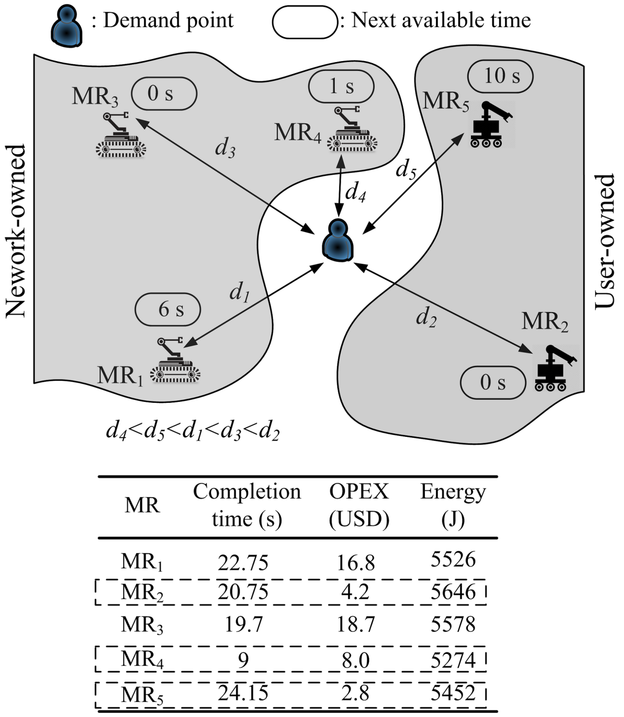
$$v_{max} < \sqrt{\frac{-c_2 + \sqrt{c_2^2 - 4\left(\frac{6c_1}{\Delta d^2}\right)(-3c_4)}}{2\left(\frac{6c_1}{\Delta d^2}\right)}}; \tag{3.8}$$

otherwise, $E(\omega_d)$ is a monotonically increasing function of ω_d with a minimum at $\omega_d = 0$.

Proof. See Appendix A.2.

3.3 Context-Aware Multi-Robot Task Coordination

In this section, we study the problem of task allocation to MRs in multi-robot FiWi-based infrastructures in greater detail. We note that the automation of various physical and digital tasks with



MR	Completion time (s)	OPEX (USD)	Energy (J)
MR$_1$	22.75	16.8	5526
MR$_2$	20.75	4.2	5646
MR$_3$	19.7	18.7	5578
MR$_4$	9	8.0	5274
MR$_5$	24.15	2.8	5452

Figure 3.3 – An illustrative case study demonstrating the trade-off between delay, OPEX, and energy performance of the multi-robot task allocation problem at a given task arrival time instant.

context-aware requirements is doable by state-of-the-art agents and robots. We start by presenting an illustrative use case as a simplified example of our optimization problem of interest. Next, we develop the multi-objective formulation of our problem. We then develop a context-aware allocation algorithm of physical/digital tasks for the HART-centric multi-robot task coordination based on the shared use of user- and network-owned robots.

3.3.1 Illustrative case study

For illustration, we present a case study to better understand the impact of different coordination strategies on the delay/cost/energy performance from the viewpoint of both users and network operator. Let us consider 2 user- and 3 network-owned MRs (i.e., a 40% user ownership), as shown in Fig. 3.3, where a task has arrived at the demand point to be allocated to one of the MRs. The next available time of each MR is also shown. Assume that receiving service from a user- and network-owned MR is subject to an incurred OPEX of 0.2 and 1 USD per second, respectively. The outcome of the allocation of the given task to each of the MRs is shown in Fig. 3.3, which demonstrates the trade-off between task completion time, OPEX, and energy consumption. The results indicate that allocating the task to MR$_2$, MR$_4$, or MR$_5$ is a Pareto optimal solution with respect to the three objectives of task completion time, OPEX, and energy consumption, i.e., one cannot find any other solution whose performance in terms of all the three objectives is better than MR$_2$, MR$_4$, or MR$_5$. Note that any allocation decision made for a given task updates the next available time of the allocated MR, thus having a direct impact on the performance results for upcoming tasks, whose arrival time instants are not known in advance.

3.3.2 Problem formulation

We assume that HART members are self-aware about their respective goals, application needs, capabilities, and constraints, to be elaborated on in Section 3.4. Further, through communication they establish a collective context-awareness with the objective of minimizing the completion time of tasks by MRs, which may be either user-owned or network-owned. Let the ownership spreading factor γ_O denote the percentage of robots that are jointly owned by MUs, whereas the remaining robots are owned by the network operator. More specifically, our multi-robot task coordination algorithm aims to minimize the task completion time along with the energy consumption and OPEX of physical/digital task execution by MRs. In the following, after introducing the decision variables and parameters, we develop a multi-objective formulation of the dynamic task allocation problem.

Given:

- J_i: Task i, $(i = 1, 2, ...)$.
- t_i^a: Arrival time of task demand i.
- W_i^p: Physical workload (in Joules) generated by J_i.
- W_i^d: Digital workload (in required CPU cycles) generated by J_i.
- l_i^{task}: Demand location of task J_i.
- $\mathcal{R}_N$: Set of network-owned MRs.
- $\mathcal{R}_U$: Set of user-owned MRs.
- $\mathcal{R}_U^A$: Set of available user-owned MRs.

- $\mathcal{R}_U^B$: Set of busy user-owned MRs.
- $\mathcal{R}_N^A$: Set of available user-owned MRs.
- $\mathcal{R}_N^B$: Set of busy network-owned MRs.
- $\mathcal{R}$: Set of all user- or network-owned MRs.
- N: Total number of MRs.
- l_j^r: Location of MR$_j$.
- t_j^{av}: Next available time of MR$_j$.
- v_{max}^j: Maximum speed of robot MR$_j$.
- $a_{acc}^{max,j}$: Maximum acceleration of robot MR$_j$.
- C_j^p: Physical task processing capacity (in Watts) of MR$_j$.
- C_j^d: Digital task processing capacity (in CPU cycles per time unit) of MR$_j$.
- D: Maximum scheduling deadline.
- $d(l_j^r, l_i^{task})$: Euclidean distance between the demand location of task J_i and MR$_j$.

Parameters:

- φ_U: Operational cost per time unit of user-owned MRs.
- φ_N: Operational cost per time unit of network-owned MRs.
- ϵ_d: Energy (in Joules) per CPU cycle.

Decision variables:

- X_i^j: A binary variable set to 1 if task J_i is assigned to MR$_j$.

Objectives:

- $T(\mathbf{X})$: Task completion time.

- $C(\mathbf{X})$: Operational expenditures (OPEX).
- $E(\mathbf{X})$: Energy consumption.

Multi-objective formulation:

$$\begin{aligned} \underset{\mathbf{X}}{\text{minimize}} \quad & T(\mathbf{X}), C(\mathbf{X}), E(\mathbf{X}), \qquad \forall i = 1, 2, 3, ..., \\ \text{subject to} \quad & \sum_{j \in \mathcal{R}_U} \max\{t_j^{av} - t_i^a, 0\} X_i^j < D, \\ & \sum_{j=1}^{N} X_i^j = 1, \\ & X_i^j \in \{0, 1\}, \forall j = 1, 2, ..., N, \end{aligned} \tag{P1}$$

where $T(\mathbf{X})$, $C(\mathbf{X})$, and $E(\mathbf{X})$ are obtained as follows. The total task completion time comprises the following delay components: (i) transmission delay T_{trs}^{dmd} of allocation demand from a given MU to the OLT, (ii) scheduling delay $T_{sch}^{i,j}$, which is the elapsed time between arrival time t_j^a of task J_i until MR$_j$ becomes available, (iii) transmission delay T_{trs}^{alc} of allocation from the OLT to the allocated MR, (iv) traverse time $T_{trav}^{i,j}$, which is the amount of time that takes MR$_j$ to traverse to the demand location of task J_i, (v) execution time $T_{exc}^{i,j}$, which is the amount of time that takes MR$_j$ to execute physical/digital task J_i, and (vi) transmission delay T_{trs}^{o} to transmit the output/result of digital/physical task from the MR to the MU. $T(\mathbf{X})$ is then given by

$$\begin{aligned} T(\mathbf{X}) = \sum_{j=1}^{N} X_i^j (T_{trs}^{dmd} + T_{sch}^{i,j} + T_{trs}^{alc} \\ + T_{trav}^{i,j} + T_{exc}^{i,j} + T_{trs}^{o}), \forall i = 1, 2, ..., \end{aligned} \tag{3.9}$$

where the scheduling delay $T_{sch}^{i,j}$ is obtained as

$$T_{sch}^{i,j} = \max\{t_j^{av} - t_i^a, 0\}, \forall i = 1, 2, \ldots, \forall j = 1, 2, \ldots, N. \tag{3.10}$$

We note that after rearranging and considering $\sum_{N}^{j=1} X_i^j = 1$, $\forall i = 1, 2, \ldots$, Eq. (3.9) reduces to

$$T(\mathbf{X}) = T_{trs}^{dmd} + T_{trs}^{alc} + T_{trs}^{o} + \sum_{j=1}^{N} X_i^j (T_{sch}^{j} + T_{trav}^{j} + T_{exc}^{j}), \forall i = 1, 2, \ldots \tag{3.11}$$

Before estimating the task execution time, let incoming task J_i consist of both physical and digital workloads denoted by W_i^p (in Joules) and W_i^d (in required CPU cycles), respectively. We then estimate the task execution time $T_{exc}^{i,j}$ by

$$T_{exc}^{i,j} = \underbrace{\frac{W_i^p}{C_j^p}}_{\text{physical sub-task}} + \underbrace{\frac{W_i^d}{C_j^d}}_{\text{digital sub-task}}, \forall i = 1, 2, \ldots \tag{3.12}$$

The traverse time $T_{trav}^{i,j}$ is given in Eq. (3.5), whereas the other delay components that are related to packet transmission delay will be computed shortly in Section 3.5.1.

Next, we estimate the OPEX of task execution by user- and/or network-owned MRs. To do so, we assume a flat-rate pricing policy that charges MUs from the time instant when the MR becomes available and is allocated to the task until it successfully accomplishes task execution. Let φ_U and φ_N denote the operating cost per time unit for user- and network-owned MRs, respectively, whereby $\frac{\varphi_U}{\varphi_N} \leq 1$. We then estimate the OPEX, $C(\mathbf{X})$, of task execution as follows:

$$C(\mathbf{X}) = \sum_{j \in S_U} \varphi_U X_j^i \left(T_{trav}^{i,j} + T_{exc}^{i,j}\right) + \sum_{j \in S_N} \varphi_N X_j^i \left(T_{trav}^{i,j} + T_{exc}^{i,j}\right), \forall i = 1, 2, \ldots \tag{3.13}$$

Next, let us calculate the total energy consumption. We note that the energy consumed to transmit the task demand/allocation/output/result is negligible compared to the energy consumption of an MR to traverse and execute the task. Therefore, we consider only the energy consumption of MRs to traverse to the demand location and execute the physical/digital task. Accordingly, we model the total energy consumption, $E(\mathbf{X})$, as follows:

$$E(\mathbf{X}) = \sum_{j=1}^{N} X_j^i \left(E_{trav}^{i,j} + E_{exc}^{i}\right), \tag{3.14}$$

where $E_{trav}^{i,j}$ is given in Eq. (A.1) and execution energy E_{exc}^i of task J_i (which is independent of the MR selection) is given by

$$E_{exc}^{i} = \underbrace{W_i^p}_{\text{physical sub-task}} + \underbrace{\epsilon_d W_i^d}_{\text{digital sub-task}}, \forall i = 1, 2, \ldots, \tag{3.15}$$

where ϵ_d denotes the energy (in Joules) per CPU cycle. Note that $E^{i,j}_{trav}$ depends on the MR selection, while E^{i}_{exc} does not. Thus, Eq. (3.14) reduces to

$$E(\mathbf{X}) = E^{i}_{exc} + \sum_{j=1}^{N} X^{i}_{j} E^{i,j}_{trav}, \forall i = 1, 2, \ldots \tag{3.16}$$

3.3.3 The proposed algorithm

Clearly, $T(\mathbf{X})$, $C(\mathbf{X})$, and $E(\mathbf{X})$ may be conflicting objectives, as minimizing $T(\mathbf{X})$ and $E(\mathbf{X})$ may not necessarily minimize $C(\mathbf{X})$ (see Fig. 3.3). The reason for this is that for some tasks selecting network-owned MRs can significantly reduce the task completion time, resulting in increased OPEX due to higher pricing of network-owned MRs compared to that of user-owned ones. We also note that the energy consumption of an MR is a function of its local parameters, e.g., motor and motion parameters, among others, which may preferably not be shared by the MRs, as they are considered private information. Furthermore, the task coordinator has to make decisions without a priori knowledge of the arrival time instants of upcoming tasks, thus making it impossible to exploit conventional optimization methods to obtain the optimal solution of the problem of interest. Therefore, in order to make a suitable trade-off between the three objectives and achieve a satisfactory solution, we prioritize the objectives of the problem in descending order of $T(\mathbf{X})$, $C(\mathbf{X})$, and $E(\mathbf{X})$. More specifically, we decouple the problem into two subproblems namely multi-robot task coordination and motion planning, where the former aims to minimize $T(\mathbf{X})$ and $C(\mathbf{X})$, whereas the latter minimizes $E(\mathbf{X})$ (to be discussed in Section 3.4).

As shown in Algorithm 3, our proposed context-aware dynamic multi-robot task coordination (CADMRTC) algorithm assigns the given task to the nearest available user-owned MR, if there is any (see line 2 of Alg. 3). Otherwise, it tries to find the earliest available user-owned MR up to a given maximum scheduling deadline $D \geq 0$ seconds before falling back onto network-owned MRs (see lines 6-15 of Alg. 3). In this case, the task is assigned to the nearest available network-owned MR (see line 10 of Alg. 3) or the earliest available one, if there is not any (see line 12 of Alg. 3). Note that our context-aware scheme gives priority to user-owned MRs, thus substantially reducing OPEX. It is worthwhile to mention that we aim at minimizing the energy consumption of the assigned MR by using our proposed self-aware motion planning (see line 19 of Alg. 3), to be elaborated on in technically greater detail in Section 3.4.

Algorithm 3 CADMRTC Algorithm

```
Input: $J_i, t_i^a, \mathcal{R}_U^A, \mathcal{R}_U^B, \mathcal{R}_N^A, \mathcal{R}_N^B, \mathcal{R}, l_j^r, t_j^{av}, D$
Output: $X_i^j, t_j^{av}, l_j^r, \forall j = 1, 2, ..., N$
if $\mathcal{R}_U^A \neq \emptyset$ then
    $j^* \leftarrow \underset{j \in \mathcal{R}_U^A}{\text{argmin}}\, d\left(l_j^r, l_i^{task}\right)$
else
    if $\mathcal{R}_U^B \neq \emptyset$ then
        $W_{min} \leftarrow \min_{j \in \mathcal{R}_U^B}(t_j^{av} - t_i^a)$
        if $W_{min} < D$ then
            $j^* \leftarrow \underset{j \in \mathcal{R}_U^B}{\text{argmin}}(t_j^{av} - t_i^a)$
        else
            if $\mathcal{R}_N^A \neq \emptyset$ then
                $j^* \leftarrow \underset{j \in \mathcal{R}_N^A}{\text{argmin}}\, d\left(l_j^r, l_i^{task}\right)$
            else
                $j^* \leftarrow \underset{j \in S}{\text{argmin}}(t_j^{av} - t_i^a)$
            end if
        end if
    end if
end if
$X_i^{j^*} \leftarrow 1$
$\Delta d \leftarrow d\left(l_{j^*}^r, l_i^{task}\right)$
$(T_{trav}^{i,j^*}, E_{trav}^{i,j^*})$=SAOMP $(\Delta d, j^*)$ (call Algorithm 4)
$t_{j^*}^{av} \leftarrow t_{j^*}^{av} + T_{trav}^{i,j^*} + T_{exc}^{i,j^*}$
return $X_i^j, t_j^{av}, l_j^r, \forall j = 1, 2, ..., N$
```

Next, we present a complexity analysis of our proposed CADMRTC algorithm. We note that the best and worst case time complexity of our proposed algorithm are $\mathcal{O}(\left|\mathcal{R}_U^A\right| + n)$ and $\mathcal{O}\left(\left|\mathcal{R}_U^A\right| + \left|\mathcal{R}_U^B\right| + \left|\mathcal{R}_N^A\right| + \left|\mathcal{R}_N^B\right| + n\right)$, respectively, where n is the number of operations performed by the self-aware SAOMP algorithm (see Alg. 4). We note that n is a constant number that depends on the number of local parameters of a given robot and doesn't scale with growing numbers of MR. This suggests that the total time complexity of our CADMRTC algorithm is $\mathcal{O}\left(\left|\mathcal{R}_U^A\right| + \left|\mathcal{R}_U^B\right| + \left|\mathcal{R}_N^A\right| + \left|\mathcal{R}_N^B\right|\right)$.

3.4 Self-Aware Optimal Motion Planning

Battery-powered MRs typically operate for long periods of time. Therefore, it is necessary to optimize their motion by minimizing not only their traverse time but also their energy consumption. In this section, we aim to find the energy-optimal velocity profile of an MR for a given path to traverse.

So far, we have derived traverse time T_{trav} of an MR for a given distance Δd and velocity profile $v(t)$. We have shown that an increasing ω_d decreases traverse time T_{trav}. Moreover, our derived closed-form formula for energy consumption of the MR demonstrates that under certain conditions, there exists an $\omega_d^* \in (0,1)$ for which the energy consumption is minimized. Otherwise, the energy consumption increases for increasing ω_d. Nevertheless, we note that the choice of ω_d is constrained by the maximum achievable acceleration, which in turn depends on the physical design of the motor deployed at the MR. Hence, a_{acc} is given by

$$a_{acc} = -a_{dec} = \frac{v_{max}}{T_{acc}} = \frac{v_{max}}{(1-\omega_d)\frac{\Delta d}{v_{max}}}, \tag{3.17}$$

which implies that a_{acc} is a monotonically increasing function of ω_d with lower and upper bounds given by $\frac{v_{max}^2}{\Delta_d}$ and ∞, which are reached for $\omega_d \to 0$ and $\omega_d \to 1$, respectively. Thus, for a given maximum achievable acceleration a_{acc}^{max} the feasible range for ω_d is obtained as

$$\frac{v_{max}}{(1-\omega_d)\frac{\Delta d}{v_{max}}} \leq a_{acc}^{max} \Rightarrow \omega_d \leq \overbrace{1 - \frac{v_{max}^2}{a_{acc}^{max}\Delta d}}^{\omega_d^m}. \tag{3.18}$$

Next, we aim to minimize both traverse time T_{trav} and $E(\omega_d)$, which under certain conditions may become conflicting objectives. Therefore, to find a compromise between these two conflicting objectives, we aim to solve the following multi-objective optimization problem:

$$\min_{\omega_d} \quad f(\omega_d) = \frac{E(\omega_d)}{E_m} + \frac{T_{trav}(\omega_d)}{T_{trav}^U} \tag{3.19a}$$

$$\text{s.t.} \quad \omega_d \leq \omega_d^m, \tag{3.19b}$$

$$0 \leq \omega_d < 1, \tag{3.19c}$$

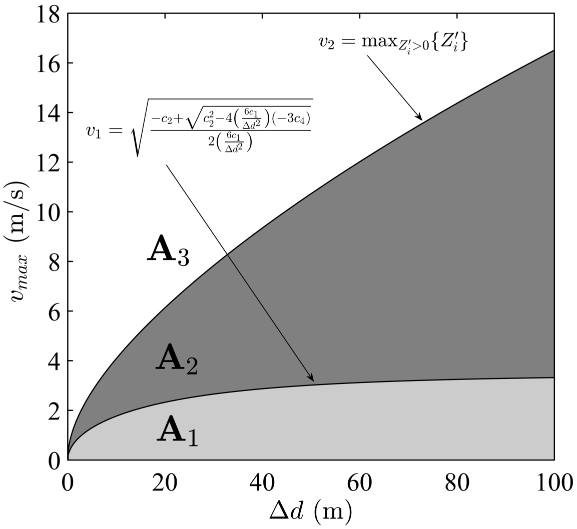


Figure 3.4 – Different MR operational regions represented by $\mathbf{A}_1$, $\mathbf{A}_2$, and $\mathbf{A}_3$ on $\Delta d - v_{max}$ plane, which the proposed self-aware optimal motion planning strategy relies on ($a_{acc}^{max} = 2$ m/s^2 fixed).

where T_{trav}^U is given in Eq. (3.6a) and E_m, the upper bound of $E(\omega_d)$, is equal to

$$E_m = \max\{E(0), E(\omega_d^m)\}. \tag{3.20}$$

We note that $f(\omega_d)$ is a convex function of ω_d, as it is the sum of two convex functions. For now, we relax the constraint (5.28b) and then solve the relaxed optimization problem by letting $\frac{\partial f(\omega_d)}{\partial \omega_d} = 0$ for the following two cases.

Case 1: In this case, $E(\omega_d)$ does not have a local minimum for $\omega_d \in (0, 1)$, i.e., $(\Delta d, v_{max}) \notin \mathbf{A}_1$ in Fig. 3.4. Since $E(\omega_d)$ is a monotonically increasing function of ω_d, its upper bound, E_m, is obtained for $\omega_d = 1 - \frac{v_{max}^2}{a_{acc}^{max}\Delta d}$. Thus, we have

$$E_m = 2c_1 v_{max} a_{acc}^{max} + \frac{c_2}{3}\left(3v_{max}\Delta d - \frac{v_{max}^3}{a_{acc}^{max}}\right) + c_3\Delta d + \left(1 + \frac{v_{max}^2}{a_{acc}^{max}\Delta d}\right)\frac{c_4\Delta d}{v_{max}}. \tag{3.21}$$

By substituting Eqs. (3.21) and (3.6a) into Eq. (5.28a), we obtain $f(\omega_d)$ as

$$f(\omega_d) = \frac{1}{E_m}E(\omega_d) + \frac{2-\omega_d}{2}. \tag{3.22}$$

Let ω_d^* denote the optimal value of $\omega_d \in (0,1)$, for which $f(\omega_d)$ is minimized. We then obtain ω_d^* by solving $\frac{\partial f(\omega_d)}{\partial \omega_d} = 0$, where $\frac{\partial f(\omega_d)}{\partial \omega_d}$ is given by

$$\frac{\partial f(\omega_d)}{\partial \omega_d} = \frac{1}{E_m}\frac{\partial E(\omega_d)}{\partial \omega_d} - \frac{1}{2}. \tag{3.23}$$

By substituting Eq. (A.6) into Eq. (3.23), we obtain

$$\frac{\partial f(\omega_d)}{\partial \omega_d} = \frac{v_{max} c_2 \Delta d}{3E_m} + \frac{2c_1 v_{max}^3}{\Delta d E_m (1-\omega_d)^2} - \frac{c_4 \Delta d}{v_{max} E_m} - \frac{1}{2}. \tag{3.24}$$

Solving $\frac{\partial f(\omega_d)}{\partial \omega_d} = 0$ gives us ω_d^* as

$$\omega_d^* = 1 \pm \sqrt{K'}, \tag{3.25}$$

where

$$K' = \frac{2c_1 v_{max}^3}{\Delta d \left(\frac{\Delta d c_4}{v_{max}} - \frac{\Delta d v_{max} c_2}{3} + \frac{E_m}{2} \right)}. \tag{3.26}$$

We note that for $K' > 0$ we have $1 + \sqrt{K'} \notin (0,1)$ and thus it is not acceptable. Whereas $1 - \sqrt{K'}$ lies in interval $(0,1)$ for a particular range of v_{max}, as specified in the following Lemma.

Lemma 3.3: ω_d^* lies in interval $(0,1)$ if and only if the following inequality holds:

$$v_{max} < \overbrace{\max_{Z_i' > 0 : \mathfrak{Im}[Z_i'] = 0} \{Z_i'\}}^{v_2}, \tag{3.27}$$

where $\{Z_i'\}_{i=1}^4$ are the roots of the quartic equation given by

$$(A_1 \Delta d - 2c_1)\, v_{max}^4 + (\Delta d B_1)\, v_{max}^2 + \Delta d C_1 v_{max} + \Delta d D_1 = 0, \tag{3.28}$$

where

$$\begin{aligned} A_1 &= -\frac{c_2}{3a_{acc}^{max}}, \\ B_1 &= \left(\frac{2c_2}{3} + c_1 a_{acc}^{max} + \frac{c_4}{a_{acc}^{max}} \right), \\ C_1 &= \frac{c_3 \Delta d}{2}, \\ D_1 &= \frac{3\Delta d c_4}{2}. \end{aligned} \tag{3.29}$$

Proof: See Appendix A.3.

Algorithm 4 SAOMP Algorithm

```
Input: $v_{max}, \Delta d, a_{acc}^{max}, c_1, c_2, c_3, c_4$
Output: $E_{trav}, T_{trav}$
Use $v_1$ and $v_2$ given by Eq. (A.10) and (3.27), respectively, to determine $\mathbf{A}_1$, $\mathbf{A}_2$, and $\mathbf{A}_3$
$\omega_d^m \leftarrow 1 - \frac{v_{max}^2}{a_{acc}^{max}}$
if $(\Delta d, v_{max}) \in \mathbf{A}_1 \cup \mathbf{A}_2$ then
    $\omega_d^* \leftarrow 1 - \sqrt{K'}$, where $K'$ is given by Eq. (3.26)
    if $\omega_d^* < \omega_d^m$ then
        $T_{trav} \leftarrow (2 - \omega_d^*) \frac{\Delta d}{v_{max}}$
        $E_{trav} \leftarrow E(\omega_d^*)$ given by Eq. (3.7)
        Update the MR velocity profile using Eqs. (3.4) and (3.17)
    else
        $T_{trav} \leftarrow (2 - \omega_d^m) \frac{\Delta d}{v_{max}}$
        $E_{trav} \leftarrow E(\omega_d^m)$ given by Eq. (3.7)
        Update the MR velocity profile using Eqs. (3.4) and (3.17)
    end if
end if
if $(\Delta d, v_{max}) \in \mathbf{A}_3$ then
    $T_{trav} \leftarrow \frac{2\Delta d}{v_{max}}$
    $E_{trav} \leftarrow E(0)$ given by Eq. (3.7)
    Update the MR velocity profile using Eqs. (3.4) and (3.17)
end if
return $E_{trav}, T_{trav}$
```

We conclude that for $v_1 < v_{max} < v_2$ (i.e., $(\Delta d, v_{max}) \in \mathbf{A}_2$ shown in Fig. 3.4), the optimal value of optimization problem (3.19) is obtained as

$$f^* = \begin{cases} f(\omega_d^*) = f(1 - \sqrt{K'}), & 0 < \omega_d < \omega_d^m \\ f(\omega_d^c), & \text{otherwise.} \end{cases} \tag{3.30}$$

For $v_{max} > v_2$ (i.e., $(\Delta d, v_{max}) \in \mathbf{A}_3$ shown in Fig. 3.4), on the other hand, $g(\omega_d) = 0$ does not have any root in interval $(0, 1)$. Since $f(\omega_d) > 0$ and $\frac{\partial^2 f(\omega_d)}{\partial \omega_d^2} > 0$ for $\omega_d \in [0, 1)$, and $\lim_{\omega_d \to 1} g(\omega_d) = +\infty$, the optimal value f^* of optimization problem (3.19) is equal to $f(0)$.

Case 2: In this case, which is illustrated by $(\Delta d, v_{max}) \in \mathbf{A}_1$ in Fig. 3.4, both $f(\omega_d)$ and $E(\omega_d)$ have a local minimum for $\omega_d \in (0, 1)$. Similarly to Case 1, the optimal value of optimization problem (3.19) is obtained by using Eq. (3.30).

In summary, Algorithm 4 shows the pseudo-code of our proposed self-aware optimal robot motion planning (SAOMP) algorithm, which runs locally in the MRs. Given the local parameters of v_{max},

Δd, a_{acc}^{max}, c_1, c_2, c_3, and c_4 of the assigned MR, our proposed self-aware algorithm makes a trade-off between the traversing time T_{trav} and energy consumption E_{trav} by means of optimally planning its motion.

3.5 Delay and Reliability Analysis

In this section, we develop our analytical framework to calculate the average packet transmission delay as well as the human-robot connection reliability in FiWi based Tactile Internet infrastructures. In our analysis, we make the following assumptions:

- *Single-hop WLAN*: MUs and MRs are directly associated with an ONU-AP via a wireless single hop, whereby ONU-MPPs serve as ONU-APs.
- *WiFi connectivity and WiFi offloading*: The WiFi connection and interconnection times of MUs are assumed to fit a truncated Pareto distribution, as validated via recent real-world smartphone traces in [13]. The probability $P_{temporal}^{MU}$ that an MU is temporarily connected to an ONU-AP is estimated as $\bar{T}_{on}/(\bar{T}_{on}+\bar{T}_{off})$, whereby $\bar{T}_{on}$ and $\bar{T}_{off}$ denote the average WiFi connection and interconnection time, respectively.
- *Task arrival model*: MUs act as service demand points, where tasks arrive at random time instants following a Poisson distribution.
- *Traffic model*: The background traffic rate generated by ONUs with attached fixed (wired) subscribers that are directly connected to the backhaul EPON is set to λ_{ONU}.

3.5.1 Delay Analysis

Recall from Section 3.3 that we estimated the scheduling, traversing, and execution delay components of the total task completion time. In this section, we proceed to develop an analytical framework to estimate the packet transmission related delay components of multi-robot task execution over FiWi-based Tactile Internet infrastructures.

We build on the analytical frameworks presented in [13] and [75]. We first define the backhaul downstream traffic intensity ρ^u and ρ^d for a TDM PON ($\Lambda = 1$) and a WDM PON ($\Lambda > 1$) as

$$\rho^u = \frac{\bar{L}}{\Lambda . c_{PON}} \sum_{q=1}^{O} \sum_{i=0}^{O} \Gamma_{qi}^{PON} < 1, \tag{3.31a}$$

$$\rho^d = \frac{\bar{L}}{\Lambda . c_{PON}} \sum_{q=0}^{O} \sum_{i=1}^{O} \Gamma_{qi}^{PON} < 1, \tag{3.31b}$$

where c_{PON} denotes the PON data rate, O denotes the number of ONUs, and Γ_{qi}^{PON} represents the traffic rate (in packet/second) between PON nodes q and i (with $q = 0$ denoting the OLT).

Similar to [75], upstream delay, D_{PON}^u, and downstream delay, D_{PON}^d, of both TDM and WDM PONs are obtained as

$$D_{PON}^u = \Phi(\rho^u, \bar{L}, \varsigma^2, c_{PON}) + \frac{\bar{L}}{c_{PON}} + 2\tau_{PON} \frac{2 - \rho^u}{1 - \rho^u} - B^u, \tag{3.32}$$

$$D_{PON}^d = \Phi(\rho^u, \bar{L}, \varsigma^2, c_{PON}) + \frac{\bar{L}}{c_{PON}} + \tau_{PON} - B^u, \tag{3.33}$$

where τ_{PON} denotes the average propagation delay between ONUs and OLT, $\Phi(\cdot)$ is the average queueing delay of an M/G/1 queue characterized by the Pollaczek-Khintchine formula as

$$\Phi(\rho, \bar{L}, \varsigma^2, c) = \frac{\rho}{2c(1 - \rho)} \left(\frac{\varsigma^2}{\bar{L}} + \bar{L} \right), \tag{3.34}$$

and

$$B^d = B^u = \Phi \left(\frac{\bar{L}}{\Lambda . c_{PON}} \sum_{q=1}^{O} \sum_{i=1}^{O} \Gamma_{qi}^{PON}, \bar{L}, \varsigma^2, c_{PON} \right). \tag{3.35}$$

Next, we calculate the average delay experienced by an arriving packet at wireless subscribers. Let $D_{z,i}^{e2e}$ denote the average packet delay of wireless subscriber i that resides within the coverage area of $\mathrm{ONU} - \mathrm{AP}_z$. The set of MUs, $\mathcal{U}_z^{MU}$, and MRs, $\mathcal{U}_z^{MR}$, along with their associated ONU-AP$_z$ constitute $\mathcal{U}_z = \mathcal{U}_z^{MU} \cup \mathcal{U}_z^{MR} \cup \{0\}$ with $i = 0$ representing $\mathrm{ONU} - \mathrm{AP}_z$. We then obtain $D_{z,i}^{e2e}$ as

$$D_{z,i}^{e2e} = \frac{1}{\frac{1}{\Delta_{z,i}} - \sigma_{z,i}}, \qquad \Delta_{z,i}\sigma_{z,i} < 1, \forall i \in \boldsymbol{\mathcal{U}_z}, z = 1, 2, ..., N_{AP}, \tag{3.36}$$

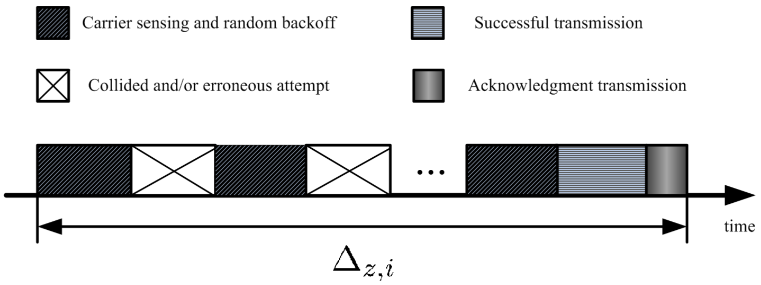

Carrier sensing and random backoff
Successful transmission
Collided and/or erroneous attempt
Acknowledgment transmission
...
time
$\Delta_{z,i}$

Figure 3.5 – Delay components of average channel access delay in IEEE 802.11 DCF with random back-offs.

where $\Delta_{z,i}$ and $\sigma_{z,i}$ denote the average channel access delay and traffic rate, respectively, and N_{AP} is the total number of ONU-APs. Note that Eq. (3.36) accounts for both queueing delay as well as channel access (service) delay of wireless subscriber $i \in \boldsymbol{\mathcal{U}_z}$. We also note that the average access delay $\Delta_{z,i}$ consists of time delays due to carrier sensing, exponential back-offs, collided and erroneous (if any) attempts, successful transmission, and acknowledgement transmission, as illustrated in Fig. 3.5.

To compute the average channel access delay, we define a two-dimensional Markov process $(s(t), b(t))$ under unsaturated conditions (see Fig. 2.10) and estimate the average service time $\Delta_{z,i}$ in a WLAN using IEEE 802.11 DCF for access control, whereby $b(t)$ and $s(t)$ denote the random back-off counter and size of the contention window at time t, respectively. We note that $\Delta_{z,i}$ is obtained as

$$\Delta_{z,i} = \sum_{k=0}^{\infty} p_{e,i}^{k} \left(1 - p_{e,i}\right) \left[\sum_{j=0}^{\infty} p_{c,i}^{j} \left(1 - p_{c,i}\right) \left(\left(\sum_{b=0}^{k+j} \frac{2^{min(b,m)} W_0 - 1}{2} E_s \right) + jT_{c,i} + kT_{e,i} + T_{s,i} \right) \right],$$
$$\forall i \in \boldsymbol{\mathcal{U}_z}, z = 1, 2, ..., N_{AP}. \tag{3.37}$$

In the following, we proceed to evaluate transmission delay T_{trs}^{dmd} from a given MU to the OLT, transmission delay T_{trs}^{alc} from the OLT to the allocated MR, and transmission delay T_{trs}^{o} from the MR to the MU.

Transmission Delay from MU to OLT

The routing path of an allocation demand transmitted by an MU consists of a single wireless hop and subsequent upstream transmission across the backhaul EPON. Therefore, the average packet transmission delay T_{trs}^{dmd} of an MU to the OLT is estimated as

$$T_{trs}^{dmd} = \underbrace{\mathbb{E}\left(D_{z,i}^{e2e}\right)}_{\text{MU to ONU-AP}} + \underbrace{D_{PON}^{u}}_{\text{ONU-AP to OLT}} , \tag{3.38}$$

where $\mathbb{E}\left(D_{z,i}^{e2e}\right)$ is computed for $\forall i \in \boldsymbol{\mathcal{U}_z^{MU}}, z = 1, 2, ..., N_{AP}$.

Transmission Delay from OLT to MR

After scheduling, the task coordinator collocated at the OLT transmits the task allocation to the selected MR. Therefore, the average transmission delay T_{trs}^{alc} from the OLT to an MR is given by

$$T_{trs}^{alc} = \underbrace{D_{PON}^{d}}_{\text{OLT to ONU-AP}} + \underbrace{\mathbb{E}\left(D_{z,i}^{e2e}\right)}_{\text{ONU-AP to MR}} , \tag{3.39}$$

where $\mathbb{E}\left(D_{z,i}^{e2e}\right)$ is computed for $\forall i = 0, z = 1, 2, ..., N_{AP}$.

End-to-End Delay from MR to MU

After successfully accomplishing the task, the MR transmits the task output/result to the corresponding MU via the associated ONU-AP. The average transmission delay T_{trs}^{o} from an MR to an MU is then obtained as

$$T_{trs}^{alc} = \underbrace{\mathbb{E}\left(D_{z,i}^{e2e}\right)}_{\text{MR to ONU-AP}} + \underbrace{\mathbb{E}\left(D_{z,i}^{e2e}\right)}_{\text{ONU-AP to MU}} , \tag{3.40}$$

where the first term is averaged over $\forall i \in \boldsymbol{\mathcal{U}_z^{MR}}, z = 1, 2, ..., N_{AP}$, whereas the second term is averaged over $\forall i = 0, z = 1, 2, ..., N_{AP}$.

3.5.2 Reliability Analysis

Recall from above that according to recent real-world smartphone traces, WiFi connection and interconnection times follow a truncated Pareto distribution [13]. The stationary probability that an MU/MR temporarily resides within the coverage area of an ONU-AP is given by

$$P_{temp} = \frac{\bar{T}_{on}}{\bar{T}_{on} + \bar{T}_{off}} = \frac{\bar{T}_{on}/\bar{T}_{off}}{1 + \bar{T}_{on}/\bar{T}_{off}}. \tag{3.41}$$

In order for an MR to successfully perform the task requested by an MU, both MU and MR have to be connected to the associated ONU-AP. Let us define the human-to-robot (HR) connectivity probability P_{HR} as the probability that both MU and MR are connected to the associated ONU-APs, which is given by

$$P_{HR} = P_{temp}^{MU} \cdot (1 - P_{MU \to AP}^{drop}) \cdot (1 - P_{AP \to MR}^{drop}) \cdot P_{temp}^{MR} \cdot (1 - P_{MR \to AP}^{drop}) \cdot (1 - P_{AP \to MU}^{drop}), \tag{3.42}$$

where P^{drop} denotes the packet dropping probability.

Furthermore, let us define the HR connection reliability function, $R_{HR}(t)$, as the probability that the HR connection time lasts longer than t seconds. First, let random variables T_{MU} and T_{MR} denote the WiFi connection lifetime of the MU and MR, respectively. Recall that according to our mobility model, T_{MU} and T_{MR} follow a truncated Pareto distribution, whose PDF is given by

$$f(t) = \frac{\alpha\gamma^{\alpha}}{1 - (\frac{\gamma}{\nu})^{\alpha}} t^{-(\alpha+1)}, 0 < \gamma \leq t \leq \nu. \tag{3.43}$$

For notational convenience, we use subscripts H and R to denote the MU and MR, respectively. Note that HR connection time T_{HR} can be computed as

$$T_{HR} = \min\{T_{MU}, T_{MR}\}. \tag{3.44}$$

The probability that the HR connection time T_{HR} is greater than t translates to the joint probability

$$P\left(T_{MU} > t, T_{MR} > t\right),$$

which in turn gives HR connection reliability $R_{HR}(t)$ as follows:

$$R_{HR}(t) = P(T_{MU} > t) \cdot P(T_{MR} > t). \tag{3.45}$$

We note that this is due the fact that T_{MU} and T_{MR} are two independent random variables since the mobility of an MU does not depend on that of an MR, thus rendering the MU-AP and MR-AP connections completely independent from each other. Consequently, Eq. (3.45) is then equal to

$$R_{HR}(t) = \left(1 - \underbrace{\int_0^t f_H(t).dt}_{P(T_H)<t}\right)\left(1 - \underbrace{\int_0^t f_R(t).dt}_{P(T_R)<t}\right) \overset{\text{Eq. (3.43)}}{=} \left(1 - \frac{1-(\frac{\gamma_H}{t})^{\alpha_H}}{1-(\frac{\gamma_H}{\nu_H})^{\alpha_H}}\right)\left(1 - \frac{1-(\frac{\gamma_R}{t})^{\alpha_R}}{1-(\frac{\gamma_R}{\nu_R})^{\alpha_R}}\right). \tag{3.46}$$

Next, we proceed to estimate the conditional probability of an HR connection failure during time interval $[t, t+\xi]$, given that MU and MR have been connected for the last t seconds:

$$P(t < T_{HR} < t+\xi \mid T_{HR} > t) = \frac{P(\text{connection failure happens in } [t, t+\xi])}{P(T_{HR} > t)}. \tag{3.47}$$

As $\xi \to 0$, Eq. (3.47) reduces to

$$P(t < T_{HR} < t+\xi \mid T_{HR} > t) = \frac{dF_{T_{HR}}(t)}{1 - F_{T_{HR}}(t)}. \tag{3.48}$$

The right-hand side of Eq. (3.48), which is commonly referred to as *failure rate function (FRF)* denoted by $h_{HR}(t)$, represents the conditional probability intensity that an HR connection fails, given that it has lasted up to time t [80]. Hence, $h_{HR}(t)$ is then obtained as

$$h_{HR}(t) = \frac{\frac{\partial}{\partial t}(1 - R_{HR}(t))}{R_{HR}(t)}. \tag{3.49}$$

Substituting Eq. (3.46) into Eq. (3.49) and then differentiating with respect to t finally yields

$$h_{HR}(t) = \frac{\frac{\alpha_H \gamma_H^{\alpha_H}}{1-\left(\frac{\gamma_H}{\nu_H}\right)^{\alpha_H}} t^{-(\alpha_H+1)}}{\frac{1-(\frac{\gamma_H}{t})^{\alpha_H}}{1-(\frac{\gamma_H}{\nu_H})^{\alpha_H}}} + \frac{\frac{\alpha_R \gamma_R^{\alpha_R}}{1-\left(\frac{\gamma_R}{\nu_R}\right)^{\alpha_R}} t^{-(\alpha_R+1)}}{\frac{1-(\frac{\gamma_R}{t})^{\alpha_R}}{1-(\frac{\gamma_R}{\nu_R})^{\alpha_R}}}. \tag{3.50}$$

Note that $F_{T_{HR}}(t)$ is an increasing/decreasing failure rate (IFR/DFR) distribution, if $h_{HR}(t)$ is an increasing/decreasing function of t. We can easily verify that $h_{HR}(t)$ in Eq. (3.50) is a convex function of t. Thus, there exists a $t^* > 0$ such that $\frac{\partial}{\partial t}h_{HR}(t) \mid_{t^*} = 0$, whereby $\frac{\partial}{\partial t}h_{HR}(t)$ is negative for $t < t^*$ (i.e., DFR) and positive for $t > t^*$ (i.e., IFR). We note that IFR renders an intuitive concept in that the probability of an HR connection failure increases over time. DFR, on the other hand, implies that the probability of losing an HR connection decreases over time, which happens for $t < t^*$.

Table 3.1 – MR and FiWi Network Parameters & Default Values

Parameter	Value	Parameter	Value
c_1	17.75	DIFS	34 μsec
c_2	1.16	SIFS	16 μsec
c_3	10.46	PHY Header	20 μsec
c_4	4.70	W_0	16 slots
γ_0	$[0, 25, 50, 75, 100]\%$	H	6
v_{max}	2 m/s	ϵ	9 μsec
a_{max}^{acc}	2 m/s^2	RTS	20 bytes
$\mathbb{E}(t_i^a)$	8 s	CTS	14 bytes
W_i^p	1-10 KJ	ACK	14 bytes
W_i^d	100-500 MHz (cycles/s)	r in WMN	300 Mbps
N	4	c_{PON} in PON	10 Gbps
N_{MU}	8	l_{PON}	20 km
C^p	1 KW	$\bar{L}$	1500 bytes
C^d	100-500 GHz (cycles/s)	ς_L^2	0
ϵ_d	10×10^{-10} J	ONU-AP radius	$10\sqrt{2}$ m

3.6 Results

In this section, we investigate the performance of our proposed task allocation scheme. For convenience, we have summarized the key parameters and their assigned default values in Table 3.1, which lists the parameter values of the considered FiWi network taken from [13], those of the mobile robots in compliance with [79], and those of the physical and digital tasks in consistency with [81] and [77][1]. We consider 4 ONU-APs as well as 4 ONUs serving fixed wired subscribers. Associated with each ONU-AP are 2 MUs along with an MR, with a total of 8 MUs and 4 MRs. We use a Poisson point process to generate the random locations of MUs and MRs in an 80×80 m^2 area. We assume that task demands arrive at the task coordinator with exponential interarrival times.

Fig. 3.6 depicts the average OPEX per task vs. user- to network-ownership cost ratio, $r_{U2N} = \frac{\varphi_U}{\varphi_N} \leq 1$. Interestingly, we observe that while full user-ownership (i.e., $\gamma_O = 100\%$) is always beneficial for MUs in terms of OPEX savings, partial user-ownership (i.e., $\gamma_O < 100\%$) isn't necessarily so. From Fig. 3.6 we observe that a user-ownership of $\gamma_O = 25\%$ is less costly than full network ownership (i.e., $\gamma_O = 0\%$) only for $r_{U2N} < 0.39$. For $r_{U2N} > 0.39$, on the other hand, MUs face lower OPEX per task with full network-ownership ($\gamma_O = 0\%$) compared with a partial user-ownership of $\gamma_O = 25\%$. The reason for this is that for $\gamma_O = 0\%$, our task coordination algorithm allocates tasks

[1] It is worthwhile to mention that these parameters are either based on real-world experiments/measurements or in compliance with well-known standards (e.g., IEEE 802.11n/ac and IEEE 802.3ah).

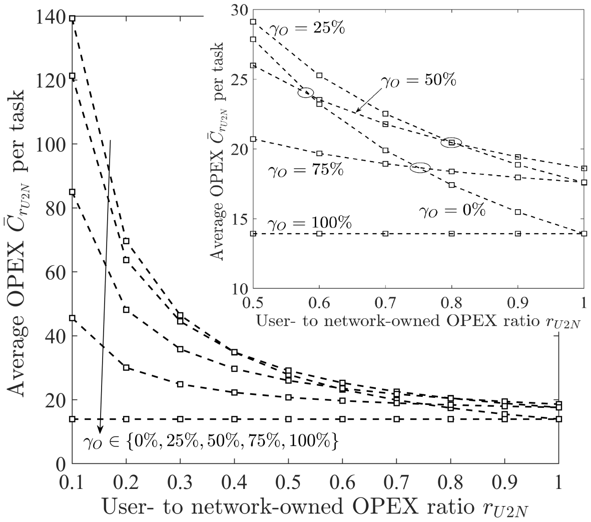


Figure 3.6 – Average cost, $\bar{C}$, per executed task vs. user- to network-owned OPEX ratio r_{U2N} ($D = 0$ fixed).

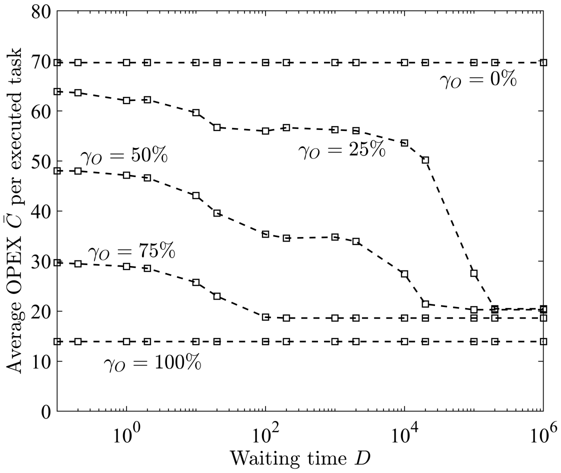


Figure 3.7 – Average OPEX, $\bar{C}$, per executed task vs. waiting deadline D ($r_{U2N} = 0.2$ fixed).

to the preferred user-owned MR(s), which are responsible for giving service to all MUs in the area. This, in turn, increases the average distance traversed by user-owned MRs, thus increasing average traverse time and energy consumption. As $\frac{\varphi_U}{\varphi_N}$ becomes greater than 0.39, full network-ownership therefore proves less costly. Further, note that in order for a partial user-ownership of $\gamma_O = 50\%$ and $\gamma_O = 75\%$ to be less costly than full network-ownership ($\gamma_O = 0\%$), r_{U2N} must not exceed 0.58 and 0.73, respectively (see also Fig. 3.6). Moreover, we observe that as $\frac{\varphi_U}{\varphi_N} \to 0$, the beneficial impact of user-ownership on OPEX savings is more pronounced, whereas for $\frac{\varphi_U}{\varphi_N} \to 1$, the average

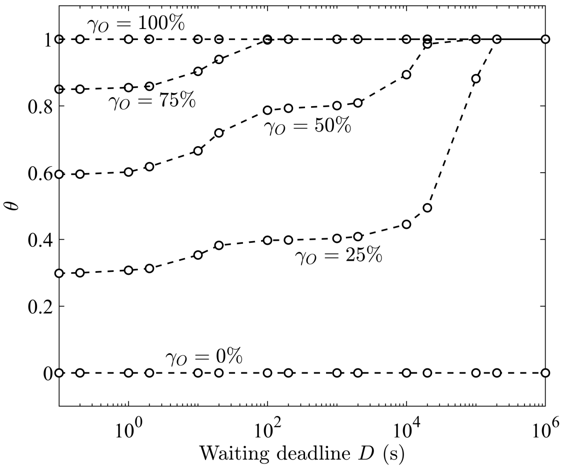


Figure 3.8 – θ **vs. waiting deadline** D **(**$r_{U2N} = 0.2$ **fixed).**

OPEX $\bar{C}_{r_{U2N}}$ per task for different values of γ_O converge to that of $\gamma_O = 100\%$. This is because as $\frac{\varphi_U}{\varphi_N} \to 1$ we have $\varphi_U \approx \varphi_N$, thus user-ownership does not reveal a notable OPEX gain compared with full network-ownership.

Next, we explore the impact of increasing waiting deadline[2] D on OPEX savings in Fig. 3.7, which depicts the average OPEX, $\bar{C}$ per executed task, vs. waiting deadline D. We find that an increasing D reduces $\bar{C}$ only for partial user-ownership (i.e., γ_O = 25, 50, and 75%). To better understand this, let θ denote the ratio of the number of executed tasks by user-owned MRs to the total number of tasks. Fig. 3.8 depicts θ vs. waiting deadline D for the same fixed $r_{U2N} = 0.2$. We observe that an increasing D has no impact on $\gamma_O = 0\%$ and 100%, whereas it increases θ for $\gamma_O = 25, 50$, and 75%. This is due to the fact that for increasing D more tasks are executed by user-owned MRs rather than their network-owned counterparts, resulting in a decreased $\bar{C}$ (see also Fig. 3.7).

Next, we plot the average task completion time vs. waiting deadline in Fig. 3.9. The figure shows that for γ_0 the average task completion time increases linearly with D. We note that Fig. 3.9 along with Figs. 3.7 and 3.8 demonstrate that for $\gamma_O = 0\%$, setting $D = 0$ achieves the best performance in terms of not only OPEX but also average task completion time. For $\gamma_0 = 25, 50$, and 75%, on the other hand, the average task completion time increases for increasing D until it hits a plateau.

[2] We note that by setting $D = 0$ for $\gamma_0 = 0\%$ and 100%, our proposed CADMRTC algorithm may be viewed as the nearest available robot allocation scheme, which stands as a baseline for fair comparison.

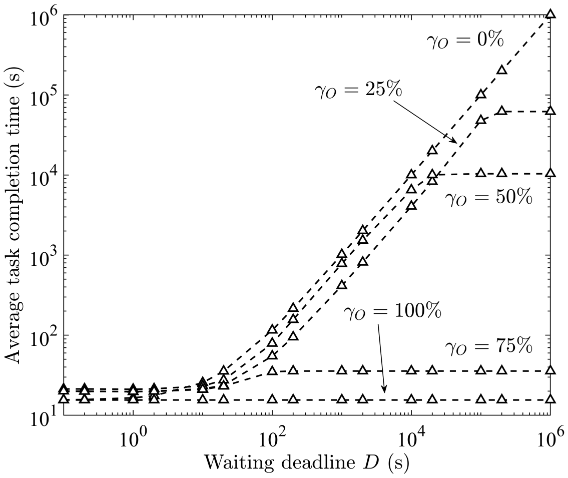


Figure 3.9 – Average task completion time vs. waiting deadline D.

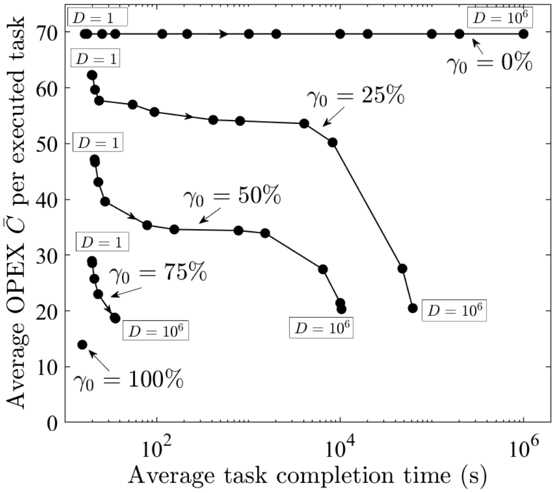


Figure 3.10 – 2-D Pareto-front of our proposed CADMRTC algorithm for different values of ownership spreading factor γ_0 (waiting deadline D increases along the arrow shown on each curve).

The values of D above, with the average task completion time remaining constant, are obtained as 10^5, 2×10^4, and 100 seconds for $\gamma_0 = 25, 50$, and 75%, respectively.

The obtained 2-D Pareto front results of our proposed CADMRTC algorithm are depicted in Fig. 3.10, which characterizes the trade-off between the average OPEX per task and average task completion time. Fig. 3.10 reveals that none of the obtained results for a given γ_0 is dominant, thus the decision maker can yield a flexible trade-off between the two objectives of the problem by appropriately setting the waiting deadline D. Fig. 3.11 depicts the average task completion time vs.

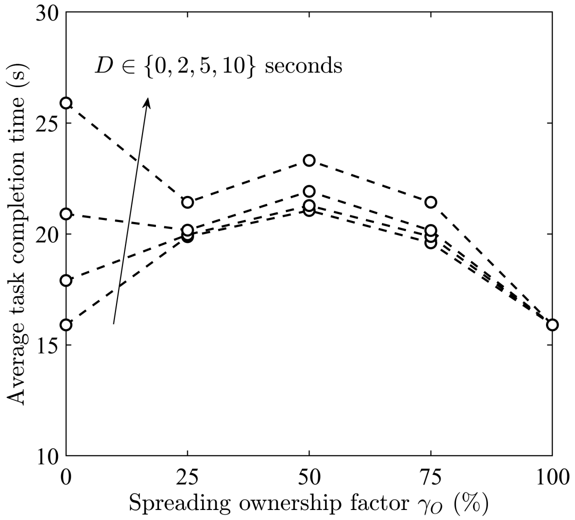


Figure 3.11 – Average task completion time vs. ownership spreading factor.

ownership spreading factor γ_O for different deadline $D \in \{0, 2, 5, 10\}$ (given in seconds). Generally, we observe a trend of decreasing average task completion time for increasing ownership spreading factor, whereby the impact of varying D becomes negligible for an ownership spreading factor of 75% and higher. Note that the lowest average task completion time of roughly 16 seconds can be achieved for $D = 0$ with either 0% or 100% ownership spreading. This is due to the fact that in both cases all MRs are eligible for immediate task allocation. More interestingly, for $D = 0$ and to a lesser extent also for $D = 2$ seconds the average task completion time increases for an ownership spreading factor of up to 50%, as opposed to the aforementioned general trend. This observation stems from the unbalanced task allocation between a few over-utilized user-owned MRs and the rest of under-utilized network-owned MRs (see also Fig. 3.8).

Figure. 3.12 illustrates the probability of HR connectivity vs. $\bar{T}_{on}^{MU}/\bar{T}_{off}^{MU}$ for different values of $\bar{T}_{on}^{MR}/\bar{T}_{off}^{MR}$. For $\bar{T}_{on}/\bar{T}_{off} = 2.73$, which is obtained from real-world measurements in [13], we achieve a maximum of 53.57% H2R connectivity probability. Note that for an increasing $\bar{T}_{on}^{MU}/\bar{T}_{off}^{MU}$ of up to 30, P_{HR} increases until it it levels off. Conversely, for $\bar{T}_{on}^{MU}/\bar{T}_{off}^{MU} > 30$, P_{HR} highly depends on the temporal availability of MR, P_{temp}^{MR}.

Finally, Fig. 3.13 shows the HR connection reliability function $R_{HR}(t)$ and HR connection failure rate $h_{HR}(t)$. Note that for $t < 615.4$ minutes, the reliability function is DFR, whereas it is IFR for $t > 615.4$. This implies that the connection failure rate $h_{HR}(t)$ decreases as time t increases up to

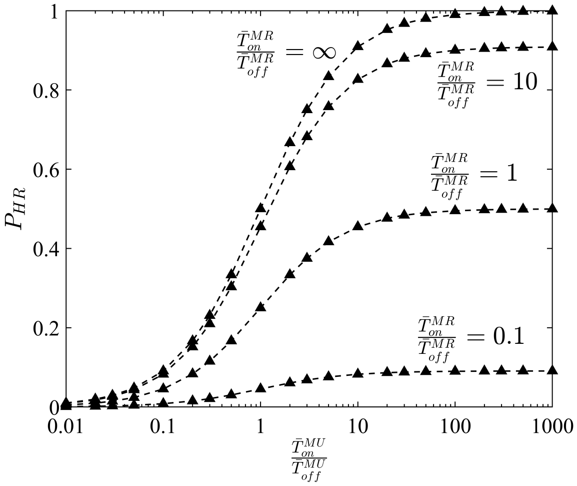


Figure 3.12 – HR connectivity probability vs. $\bar{T}_{on}^{MU}/\bar{T}_{off}^{MU}$ for different values of $\bar{T}_{on}^{MR}/\bar{T}_{off}^{MR}$.

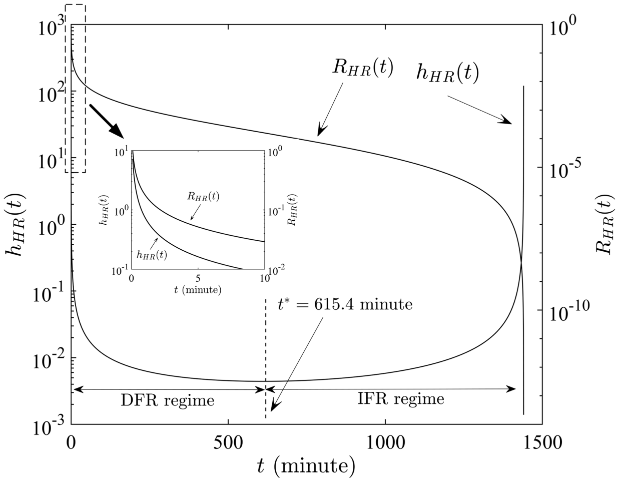


Figure 3.13 – HR connection reliability function $R_{HR}(t)$ and failure rate function $h_{HR(t)}$ vs. time.

$t^* = 615$ minute, given that it has not failed by time t. At $t = t^*$, the minimum value of 0.0044 is achieved. For $t > t^*$, on the other hand, the HR connection failure rate increases up to roughly 120 per minute (i.e., average inter-failure time becomes 500 ms).

3.7 Conclusions

We investigated the performance of our proposed context- and self-aware HART centric multi-robot task allocation over FiWi based Tactile Internet infrastructures. We shed light on when, how, and under which circumstances user-ownership of mobile robots (MRs) becomes beneficial in terms of

OPEX per executed task. Further, we evaluated the performance of our proposed CADMRTC algorithm in terms of average task completion time, OPEX per executed task, and ratio of the number of executed tasks by user-owned MRs and the total number of tasks. By leveraging on the low-latency and reliable fiber backhaul and distributed WiFi-based fronthaul, we showed that a human-robot connectivity probability of $> 90\%$ is achievable for $\bar{T}_{on}^{MR}/\bar{T}_{off}^{MR} > 10$. In addition, our obtained results show that our proposed self-aware scheme plays a key role in minimizing the traverse time as well as energy consumption of MRs in a distributed manner, whereas our context-aware task coordination is instrumental in minimizing the task completion time, while paying particular attention to reducing OPEX of user-/network-ownership of MRs.

Importantly, our obtained results show that from a performance perspective (in terms of average task completion time) almost no deterioration occurs if the ownership is shifted entirely from network operators to mobile end-users ($D = 0$), though such a shift in ownership of robots has significant implications on sharing the profits and collaborative business opportunities arising from the emerging Tactile Internet in a more equitable fashion. As a result, this may open up new opportunities for synergies between humans and machines/robots, while spurring the symbiotic human-machine/robot development envisaged by early-day Internet pioneers and imagining entirely new categories of abundance for a low entry cost economy. Among others, one future research direction is to further explore the synergies between the aforementioned HART membership and the complementary strengths of robots to facilitate local human-machine coactivity clusters by decentralizing the Tactile Internet. Another interesting open research problem is how human *crowdsourcing* can help decrease task completion time in the event of unreliable connectivity and/or network failures. Note that our presented spreading ownership of robots across mobile users may be an important stepping stone to collaborative business relationships that function more like localized share-economy ecosystems than markets.

Chapter 4

Delay-Constrained Teleoperation Task Scheduling and Assignment

This chapter contains material extracted from the following publication:

[51] A. Ebrahimzadeh and M. Maier. Delay-Constrained Teleoperation Task Scheduling and Assignment for Human+Machine Hybrid Activities over FiWi Enhanced Networks. *IEEE Transactions on Network and Service Management*, IEEE Xplore Early Access.

[52] A. Ebrahimzadeh and M. Maier. Tactile Internet over FiWi enhanced LTE-A Het-Nets via Artificial Intelligence Embedded Multi-Access Edge-Computing. *CRC Press*: 5G-Enabled Internet of Things, accepted for publication.

In the following, my key contributions in both of the aforementioned publications are explained in greater detail: (1) I largely contributed to writing the whole manuscript, (2) I developed the analytical framework, (3) I conducted the algorithmic work, and (4) I ran the simulations.

4.1 Introduction

A popular misinterpretation about robotics is that intelligent systems, ranging from advanced robots to digital bots, will gradually substitute humans in one job after another. This argument may be true for some jobs, but we note that even though advanced robotics can be deployed to automate certain jobs, its greater potential, yet to be unleashed, is to *complement* and *augment* human capabilities. The cutting-edge jobs and innovative businesses that arise from human-machine symbiosis are happening in the so-called *missing middle* that refers to the new ways that have to bridge the gap between human-only and machine-only activities. This gives way to the so-called *third wave of business transformation*, which will be centred around human+machine hybrid activities [8]. Key toward developing the missing middle is to understand the ways humans help machines and the ways machines help humans. An interesting example of recognizing the relative strengths of humans and machines and leveraging on them to fill the missing middle can be found at automobile manufacturer Audi. Having deployed a fleet of audi robotic telepresence (ART) systems, Audi has set forth toward employee augmentation that not only helps train technicians in diagnostics and repair, but also accelerates delivery of service to customers [9].

The advent of semi-autonomous robotic assistance systems is becoming a part of the vision of the Tactile Internet. An early example is the European research project Robot-Era, which recently concluded the world's largest real-life trial of robot aids for the ageing population. With their small-stage deployment proven successful, robotic helpers will need to request human assistance every now and then, as stated recently by automobile manufacturer Nissan to augment their autonomous vehicle technology with a crew of on-call remote human operators acting as "mobility managers", who can remotely take control in unexpected situations [82].

While the Tactile Internet has been more recently also referred to as the 5G-enabled Tactile Internet, the importance of the so-called *backhaul bottleneck* needs to be recognized as well, calling for an end-to-end design approach leveraging both wireless frontend and wired backhaul technologies [46]. This mandatory end-to-end design approach is fully reflected in the key principles of the reference architecture within the emerging IEEE P1918.1 standards working group (formed in March 2016), which aims to define a framework for the Tactile Internet [60]. These key principles aim to develop a generic Tactile Internet reference architecture, supporting local area as well as wide area connectivity through wireless (e.g., cellular, WiFi) or hybrid wireless/wired networking. The importance of such design approach is more highlighted for the Tactile Internet applications that may not always require mobility, e.g., remote healthcare. With the wide deployment of PONs providing high capacity and reliability and wireless networks offering ubiquitous and flexible connectivity, interest has been growing in bimodal FiWi networks that leverage the complementary benefits of optical fiber and wireless technologies (see [83] for a detailed survey of network architectures, algorithms, and standardization). FiWi enhanced LTE-A HetNets represent a compelling solution to enable 4G cellular networks to meet the key requirements of low-latency and high-availability [84]. Recently, the authors of [13] have evaluated the maximum aggregate throughput, offloading efficiency, and delay performance of FiWi enhanced LTE-A HetNets and have shown that via WiFi offloading and fiber backhaul sharing, an ultra-low latency of 1-10 millisecond and highly reliable network connectivity can be achieved for a wide range of traffic loads.

Unlike their fully-autonomous counterparts, semi-autonomous robotic systems rely on human assistance from time to time via teleoperation and/or telepresence when domain expertise is needed to accomplish a specific task, thus allowing for an HITL centric design approach. As these robots will need to request human assistance via teleoperation/presence, mapping these requests to the human operators themselves stands as a difficult multicriteria optimization problem with the objectives of minimizing the average weighted task completion time, maximum tardiness, and average OPEX per task. The difficulty of solving such a problem lies in the following reasons. First, it is clear that we are dealing with different conflicting objectives, which makes it challenging to obtain a satisfactory result, especially for large-sized problem instants. Second, the assignment of a given task to a human operator is subject to strict end-to-end packet delay constraints, thus calling for a cross-layer approach, taking into account the delay experienced by packets in both command and feedback paths (to be discussed later on).

In this chapter, after elaborating on our envisioned bimodal FiWi network infrastructure and its role in realizing teleoperation in the Tactile Internet, we formulate and solve the problem of joint prioritized scheduling and assignment of delay-constrained teleoperation tasks to human operators so as to minimize the average weighted task completion time, maximum tardiness, and average OPEX per task. In particular, the contributions of this chapter are as follows:

- We elaborate on the role of FiWi enhanced networks as the underlying communications infrastructure for enabling emerging delay-sensitive Tactile Internet applications. In particular, trying to build on our findings in [13] and [46], we aim to realize local and/or non-local teleoperation over FiWi enhanced networks, leveraging on low-cost data-centric (optical fiber and wireless) Ethernet technologies in both fronthaul and backhaul.

- We define the problem of joint prioritized scheduling and assignment of delay-constrained teleoperation tasks onto available skilled human operators. After formulating our multi-objective optimization problem, we propose our so-called context-aware prioritized scheduling and task assignment (CAPSTA) algorithm to achieve satisfactory results by making suitable trade-offs between the contradicting objectives of the problem.

- We develop our analytical framework to estimate the end-to-end packet delay of both local and non-local teleoperation over FiWi enhanced networks. Our analysis flexibly allows for the coexistence of both conventional H2H and haptic H2M traffic, while focusing on the human operators and teleoperator robots involved in either local or non-local teleoperation. The results of our delay analysis are then fed into the proposed CAPSTA algorithm.

The remainder of the chapter is structured as follows. Section 4.2 describes FiWi based Tactile Internet infrastructures for HITL-centric teleoperation-based task coordination. In Section 4.3, we present our problem formulation, which is then solved by proposing our context-aware task coordination algorithm in Section 4.4. Our end-to-end packet delay analysis is presented in Section 4.5. In Section 4.6, we present our obtained results and findings. In section 4.7, we present a complementary discussion on our findings and point to some interesting future research avenues. Section 4.8 concludes the chapter.

4.2 System Model

4.2.1 Teleoperation

Figure 4.1 depicts a typical bilateral teleoperation system based on bidirectional haptic communications between an HO and a TOR, which are both connected via a communication network. In a typical teleoperation system, the position-orientation samples are transmitted from the HO through the HSI in the command path, whereas the force-torque samples are fed back to the HO in the feedback path. By interfacing with the HSI, the HO commands the motion of the TOR in the remote environment. This couples the HO closely with the remote environment and thereby creates a more realistic feeling of remote presence.

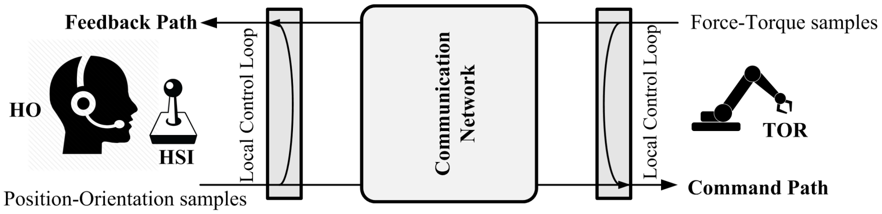


Figure 4.1 – Bilateral teleoperation system based on bidirectional haptic communications between HO and TOR.

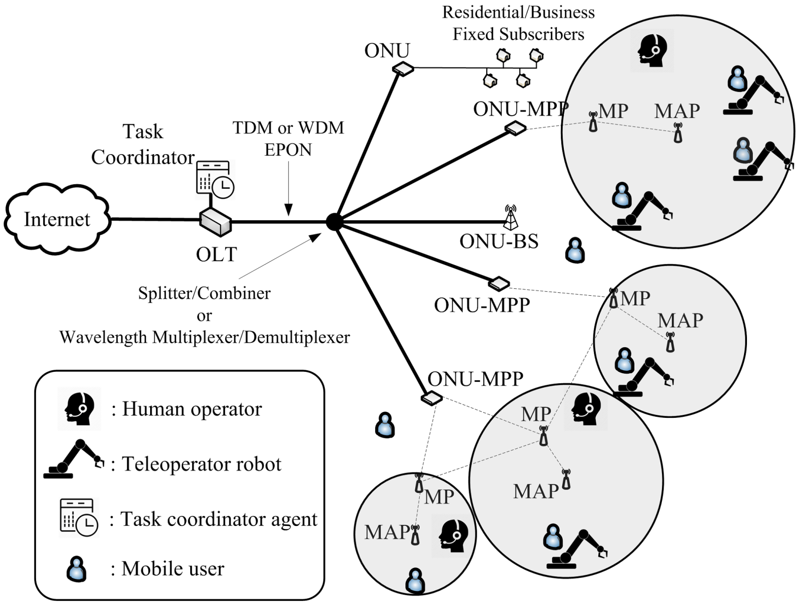


Figure 4.2 – Generic architecture of FiWi based Tactile Internet network infrastructure for teleoperation task coordination.

We let N_{DoF} denote the number of DoFs in the teleoperation system in use. Typically, N_{DoF} haptic samples coming from the application layer are encapsulated in a single segment with a prepended RTP/UDP/IP header. Each DoF haptic sample typically consists of 8 bytes, thus translating into a total of $8N_{DoF}$ bytes of payload. The haptic samples are packetized using RTP, UDP, and IP with a header size of 12, 8, and 20 bytes, respectively. Hence, the size of the resultant packet at the MAC layer SAP is equal to $8N_{DoF} + 40$ bytes.

4.2.2 Network architecture

Figure 4.2 illustrates the generic network architecture of our considered FiWi enhanced LTE-A HetNets. The optical backhaul consists of a TDM/WDM IEEE 802.3ah/av 1/10 Gb/s EPON with a typical fiber length of 20 km between the central OLT and remote ONUs, which may be extended up to 100 km to account for the long-reach PON deployment scenarios. The EPON may comprise multiple stages, each stage separated by a wavelength-broadcasting splitter/combiner or a wavelength multiplexer/demultiplexer. There are three different subsets of ONUs. An ONU may either serve fixed (wired) subscribers. Alternatively, it may connect to a cellular network BS or an IEEE 802.11n/ac/s WLAN MPP, giving rise to collocated ONU-BS or ONU-MPP, respectively. Depending on her trajectory, an MU may communicate through the cellular network and/or WLAN mesh front-end, which consists of ONU-MPPs, intermediate MPs, and mesh access points MAPs.

As shown in Fig. 4.2, selected MUs are equipped with TORs, which are capable of performing physical tasks (simply referred to as tasks hereafter) by establishing haptic communications with HOs. The MUs that are collocated with the TORs act as task demand points. Typically, the number of task demands is greater than that of available skilled HOs. This necessitates a suitable mapping of tasks to the available HOs. Given the set of tasks and available skilled HOs, the task coordinator agent is responsible for the assignment of tasks and scheduling them on the HOs (see Fig. 4.2). Note that teleoperation-based tasks arrive at the demand points. The corresponding MUs then send their demands upstream to the task coordinator agent, which is collocated with the OLT (see Fig. 4.2), via the wireless front-end and EPON backhaul until they reach the OLT. The task coordinator agent then transmits the schedule to the HOs as well as demand points. According to the schedule received from the task coordinator agent, an HO may be involved in either local or non-local teleoperation with the corresponding TOR, depending on the proximity of the involved HO and TOR, as illustrated in Fig. 4.2. In local teleoperation, the HO and corresponding TOR are associated with the same MAP and exchange their command and feedback samples through this MAP without traversing the fiber backhaul. Conversely, if HO and TOR are associated with different MAPs, non-local teleoperation is generally done by communicating via the backhaul EPON and central OLT.

4.3 Problem Statement

We consider the problem of joint assignment and scheduling of N delay-constrained teleoperation tasks on any fixed number M of HOs as follows. Let $\boldsymbol{\mathcal{M}} = \{O_1, O_2, ..., O_M\}$ and $\boldsymbol{\mathcal{J}} = \{J_1, J_2, ..., J_N\}$ denote the set of M available HOs and N given tasks, respectively. Let T_j denote the operation time of task $J_j \in \mathcal{J}$. Note that operation time T_j is given by

$$T_j = s_j + w_j, \tag{4.1}$$

where s_j and w_j is the teleoperation session setup time and workload (both in seconds) of task J_j, respectively. Each task $J_j \in \mathcal{J}$ has a due time D_j and is associated with weight Ω_j. Larger weights correspond to higher priority levels. Although the tasks are expected to be accomplished by the given due time, any incurred tardiness is subject to a cost penalty (to be elaborated on in technically greater detail shortly).

We consider an *offline* scheduling scenario, where all tasks are available at time zero and remain available continuously thereafter. Each task can be operated by only one human operator at a time and each human operator can operate only one task at a time. We also assume that preemption is not allowed, meaning that tasks cannot be split. This is because if tasks were divided and scheduled in noncontinuous time periods, preemption would incur extra reconfiguration/setup overhead, which is significant when the setup time is non-negligible. For simplicity, we assume, without loss of generality, that operation times, due times, and priority weights are all integers. Further, we assume $N \gg M$. For task J_j, the start and completion times are denoted by S_j and C_j, respectively. A feasible assignment/schedule specifies when and by which human operator a given task is operated. Given a feasible schedule, one can compute the tardiness of task J_j as $\max\{0, C_j - D_j\}$. The goal is to assign the tasks to the HOs such that the following constraints are satisfied: (1) no more than one task is assigned to an HO at a time, (2) no task is assigned to more than one HO, (3) tasks are not preempted, and (4) the average end-to-end packet delay of a scheduled teleoperation doesn't exceed a given delay threshold.

4.3.1 Problem formulation

We formulate our mixed integer programming (MIP) problem of joint prioritized scheduling and assignment of delay-constrained teleoperation tasks onto HOs as follows:

Given:

- $\mathcal{J}$: Set of tasks
- $\mathcal{M}$: Set of available human operators

- J_j: Task j, $j = 1, 2, ..., N$
- T_j: Operation time of task J_j, $j = 1, 2, ..., N$
- Ω_j: Weight of task J_j, $j = 1, 2, ..., N$
- D_j: Due time of task J_j, $j = 1, 2, ..., N$
- O_k: Human operator k, $k = 1, 2, ..., M$
- $\mathbf{D}_c$: Average end-to-end packet delay matrix of teleoperation pairs in the command path
- $\mathbf{D}_f$: Average end-to-end packet delay matrix of teleoperation pairs in the feedback path

Parameters:

- ϵ_h: Operational cost per time unit of tardiness
- ϵ_m: Operational cost of activating a teleoperation session
- ϵ_k: Operational cost per time unit of performing a teleoperation task by human operator O_k

Decision variables:

- δ_{ij}: A binary variable, which equals 0 unless task J_i precedes task J_j
- z_{jk}: A binary variable, which equals 0 unless task J_j is assigned to human operator O_k
- y_{ij}: A binary variable, which equals 0 unless tasks J_i and J_j are not assigned to the same human operator
- S_j: Operation start time associated with task J_j, $j = 1, 2, ..., N$
- C_j: Operation completion time associated with task J_j, $j = 1, 2, ..., N$
- $\boldsymbol{X}$: Set of total decision variables of the problem represented by $(\{\delta_{ij}\}, \{y_{ij}\}, \{z_{jk}\}, \{S_j\}, \{C_j\})$

Objective functions:

- $L(\boldsymbol{X})$: Average weighted task completion time

- $T(\boldsymbol{X})$: Maximum tardiness
- $C(\boldsymbol{X})$: Operational expenditure

Multi-objective formulation:

$$\underset{\boldsymbol{X}}{\text{minimize}} \qquad L(\boldsymbol{X}), T(\boldsymbol{X}), C(\boldsymbol{X}) \tag{4.2}$$

subject to

$$\delta_{ij} + \delta_{ji} + y_{ij} = 1; \qquad i, j \in \boldsymbol{\mathcal{J}}, i < j \tag{4.3a}$$

$$\delta_{ij} + \delta_{jl} + \delta_{lj} \leq 2; \qquad i, j, l \in \boldsymbol{\mathcal{J}}, i < j < l \tag{4.3b}$$

$$z_{ik} + z_{jk} + y_{ij} \leq 2; \qquad i, j \in \boldsymbol{\mathcal{J}}, i < j, k \in \boldsymbol{\mathcal{M}} \tag{4.3c}$$

$$\sum_{k=1}^{M} z_{jk} = 1; \qquad \forall j \in \boldsymbol{\mathcal{J}} \tag{4.3d}$$

$$C_j \geq T_j z_{jk}; \qquad j \in \boldsymbol{\mathcal{J}}, i < j, k \in \boldsymbol{\mathcal{M}} \tag{4.3e}$$

$$\begin{aligned} &C_j \geq C_i + T_j(\delta_{ij} + z_{ik} + z_{jk} - 2) - K(1 - \delta_{ij}); \\ &\qquad\qquad i, j \in \boldsymbol{\mathcal{J}}, k \in \boldsymbol{\mathcal{M}} \end{aligned} \tag{4.3f}$$

$$\sum_{k \in \boldsymbol{\mathcal{M}}} D^{c}_{kj} z_{jk} \leq D_0; \qquad j \in \boldsymbol{\mathcal{J}} \tag{4.3g}$$

$$\sum_{k \in \boldsymbol{\mathcal{M}}} D^{f}_{jk} z_{jk} \leq D_0; \qquad j \in \boldsymbol{\mathcal{J}} \tag{4.3h}$$

$$\delta_{ij}, \delta_{ji}, y_{ij}, z_{jk} \in \{0, 1\} \qquad i, j \in \boldsymbol{\mathcal{J}}, k \in \boldsymbol{\mathcal{M}} \tag{4.3i}$$

$$C_j \in \mathbb{R}^{+}, \qquad j \in \boldsymbol{\mathcal{J}} \tag{4.3j}$$

where $L(\boldsymbol{X})$, $T(\boldsymbol{X})$, $C(\boldsymbol{X})$ are given as follows. Our first objective is to minimize the average weighted task completion time $L(\boldsymbol{X})$, which is given by

$$L(\boldsymbol{X}) = \frac{1}{N} \sum_{j \in \boldsymbol{\mathcal{J}}} \Omega_j C_j. \tag{4.4}$$

The second objective is to minimize the maximum tardiness $T(\boldsymbol{X})$, which is given by

$$T(\boldsymbol{X}) = \max_{j \in \boldsymbol{\mathcal{J}}} \overbrace{\max\{C_j - D_j, 0\}}^{\text{tardiness of task } j}, \tag{4.5}$$

which stands as a non-linear objective function of the decision variables. The third objective is to minimize OPEX, $C(\boldsymbol{X})$, which is estimated as

$$\begin{aligned} C(\boldsymbol{X}) = M \cdot \epsilon_m + \sum_{j \in \mathcal{J}} \epsilon_h \Omega_j \max\{C_j - D_j, 0\} \\ + \sum_{k \in \boldsymbol{\mathcal{M}}} \sum_{j \in \mathcal{J}} z_{jk} \epsilon_k T_j, \end{aligned} \tag{4.6}$$

where the first term represents the cost of activating M teleoperation sessions, the second term penalizes the tardy tasks according to their priority levels, i.e., the tardy tasks with higher priorities are subject to higher incurred cost penalty, and the third term models the total cost of performing tasks by HOs. The aforementioned definitions clearly indicate that these objectives are independent and often conflicting optimization targets.

In our aforementioned MIP formulation, constraint set (4.3a) ensures that if tasks J_i and J_j are assigned to the same HO (i.e., $y_{ij} = 0$), one of them should precede the other, thus either δ_{ij} or δ_{ji} must equal 1. On the other hand, if the tasks are assigned to different HOs (i.e., $y_{ij} = 1$), both δ_{ij} and δ_{ji} must equal zero. Constraint (4.3b) ensures a linear ordering of the tasks. According to constraint (4.3c), when tasks J_i and J_j are assigned to human operator O_k, then y_{ij} must equal zero. Constraint (4.3d) ensures that each task is assigned to one of the available HOs. Constraints (4.3e) and (4.3f) represent the completion time of the scheduled tasks. We note that in constraint set (4.3f), K is a relatively large number, which is set to $\sum_{j=1}^{N} T_j$ in our problem. Constraints (4.3g) and (4.3h) ensure that the average end-to-end packet delay of any scheduled teleoperation pair HO-TOR in both command and feedback paths is kept below a given threshold D_0. To be more specific, among all the possible HO assignments, the teleoperation pairs that are incurred with an excessive amount of connection latency are excluded from the feasible set.

Note that the average end-to-end packet delays of any possible HO-TOR pair are characterized by two matrixes, one of which represents the command path whereas the other accounts for the feedback path. More specifically, the command delay matrix $\mathbf{D}_c$ is given by

$$\mathbf{D}_c = \begin{bmatrix} D^c_{11} & D^c_{12} & \dots & D^c_{1N} \\ D^c_{21} & D^c_{22} & \dots & D^c_{2N} \\ \vdots & & \ddots & \vdots \\ D^c_{M1} & D^c_{M2} & \dots & D^c_{MN} \end{bmatrix}_{M \times N}, \tag{4.7}$$

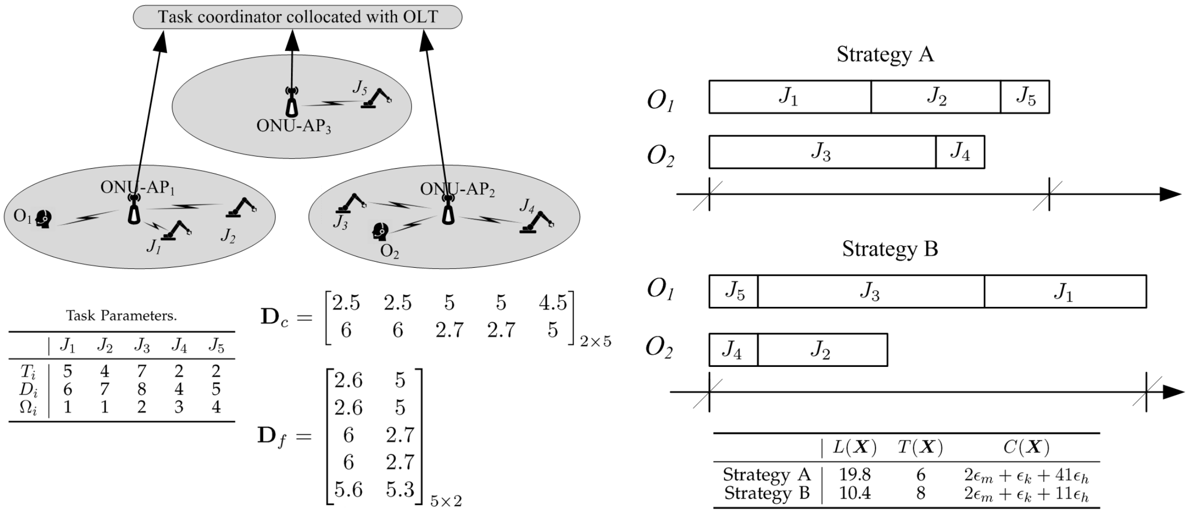
where element D^c_{kj} in row k ($k = 1, ..., M$) and column j ($j = 1, ..., N$) denotes the average end-to-end packet delay between human operator k and teleoperator robot j. Similarly, the feedback delay matrix $\mathbf{D}_f$ is given by

$$\mathbf{D}_f = \begin{bmatrix} D^f_{11} & D^f_{12} & \dots & D^f_{1M} \\ D^f_{21} & D^f_{22} & \dots & D^f_{2M} \\ \vdots & & \ddots & \vdots \\ D^f_{N1} & D^f_{N2} & \dots & D^f_{NM} \end{bmatrix}_{N\times M}, \tag{4.8}$$

where element D^f_{jk} in row j ($j = 1, ..., N$) and column k ($k = 1, ..., M$) denotes the average end-to-end packet delay between teleoperator robot j and human operator k. Note that the elements of the delay matrixes $\mathbf{D}_c$ and $\mathbf{D}_f$, which depend on the state of the underlying network, are estimated by using our delay analysis presented in Section 4.5.

4.3.2 Model scalability

Recall from above that the problem of assigning and scheduling of tasks to the human operators is subject to strict end-to-end packet delay constraints, which limits the feasible set. If we consider a special case where there are tasks to be mapped to human operators without any end-to-end packet delay constraint and with only one objective of minimizing the average weighted task completion time, then the problem reduces to parallel machine scheduling problem, which is known to be $\mathcal{NP}$-hard [36]. Given that the single-criterion parallel machine scheduling without any end-to-end delay constraint is a special case of the multi-criteria delay-constrained teleoperation task scheduling and assignment, this makes the latter also $\mathcal{NP}$-hard by restriction. Given a set of N tasks and M human operators, the developed formulation has $2N^2 + 2N + N \cdot M$ variables and $\frac{N(N-1)}{2} + \frac{N(N-1)(N-2)}{6} + \frac{2N \cdot M(N-1)}{2} + 2N^2 + N \cdot M + 4N$ constraints, which, along with the conflicting objectives, drastically restrict the scalability of the model even for small-sized problems, therefore calling for algorithmic solutions.



Task Parameters.

	J_1	J_2	J_3	J_4	J_5
T_i	5	4	7	2	2
D_i	6	7	8	4	5
Ω_i	1	1	2	3	4

$$\mathbf{D}_c = \begin{bmatrix} 2.5 & 2.5 & 5 & 5 & 4.5 \\ 6 & 6 & 2.7 & 2.7 & 5 \end{bmatrix}_{2\times 5}$$

$$\mathbf{D}_f = \begin{bmatrix} 2.6 & 5 \\ 2.6 & 5 \\ 6 & 2.7 \\ 6 & 2.7 \\ 5.6 & 5.3 \end{bmatrix}_{5\times 2}$$

	$L(\boldsymbol{X})$	$T(\boldsymbol{X})$	$C(\boldsymbol{X})$
Strategy A	19.8	6	$2\epsilon_m + \epsilon_k + 41\epsilon_h$
Strategy B	10.4	8	$2\epsilon_m + \epsilon_k + 11\epsilon_h$

Figure 4.3 – An illustrative case study of the delay/cost performance of two different task coordination strategies.

4.4 Algorithmic Solution

4.4.1 Illustrative case study

For illustration, we present a case study in order to better understand the impact of different prioritized and non-prioritized coordination strategies on the delay/cost performance from the viewpoint of both users and network operator. Let us consider two human operators and five tasks, as shown in Fig. 4.3, where the task parameters (i.e., operation times, due times, and weights[1]) as well as the command/feedback delay matrixes (in millisecond) are illustrated.

Strategy **A**, regardless of task weights, assigns the tasks to the nearest HO that resides within the coverage area of the same access point, thus giving preference to realize local teleoperation sessions. Therefore, in strategy **A**, among the feasible solutions that meet the delay constraints specified by Eqs. (4.3g) and (4.3h), tasks J_1, J_2, and J_5 are assigned to O_1, whereas J_3 and J_4 are assigned to O_2. In contrast, Strategy **B** relies on giving preference to high-priority tasks with shorter due times, thus J_5, J_3, and J_1 are assigned to O_1 whereas J_2 and $J4$ are assigned to O_2. The results indicate that strategy **B** yields a lower average weighted task completion time and smaller OPEX compared to strategy **A**. We note, however, that such superior performance is achieved at the expense of a 20% increase in maximum tardiness (see also Fig. 4.3).

[1] Weight is usually related to the importance, while due time is associated with the urgency of a given task and a prioritized scheduler must prepare a sequence able to first perform high-priority tasks.

4.4.2 Proposed task coordination algorithm

We note that while the first objective function, $L(\boldsymbol{X})$, aims to minimize the average weighted task completion time without considering the due times, the second and third objective functions (i.e., $T(\boldsymbol{X})$ and $C(\boldsymbol{X})$) deal with the tardiness incurred by overdue completion of tasks, thus they do consider the task due times. Also note that the second objective, $T(\boldsymbol{X})$, represents the maximum task tardiness, which is preferred to be minimized from a user standpoint. In addition, the third objective, $C(\boldsymbol{X})$, which addresses the operator revenue, tries to push the task completion times towards minimizing the incurred OPEX, thus implicitly minimizing the average weighted tardiness. This justifies the selection of the three different objectives in our problem formulation.

Clearly, a so-called optimum with respect to one objective may perform extremely bad with respect to other criteria (see example in Fig. 4.3). Therefore, a non-optimal solution with satisfactory performance in terms of other measures might be considered a better alternative by the decision maker. For large-sized problem instants of the developed formulation, the computational difficulties associated with finding a satisfactory solution increase dramatically. Therefore, in order to find a suitable trade-off between the conflicting objectives, we propose our so-called context-aware prioritized scheduling and task assignment (CAPSTA) algorithm, which is illustrated in Algorithm 5. The suitable performance of the proposed algorithm relies on an accurate estimation of the context parameters (e.g., task parameters, delay matrixes in both command and feedback paths, location of MUs/HOs/TORs, incoming H2H/H2M traffic pattern). In the design of the proposed CAPSTA algorithm, we adopt two sorting policies (to be elaborated on shortly), in both assignment and scheduling phases, in order to perform in favor of high-priority tasks with shorter due times.

As a first step, the proposed CAPSTA algorithm aims to partition the given task set $\boldsymbol{\mathcal{J}}$ into M subsets. Towards this end, our sorting policy indicates that the given tasks are sorted in a decreasing order of $\frac{\Omega_j}{T_j}$ (see line 1 in Algorithm 5). Next, the tasks are selected from the sorted set and then are assigned to the HOs in a round-robin fashion (see lines 3-15 in Algorithm 5). We note, however, that the assignment of task J_n to the human operator O_m is valid only if the estimated average end-to-end delays in both command and feedback paths satisfy the delay constraints in Eqs. (4.3g) and (4.3h). Otherwise, we select the HO that corresponds to the minimum average end-to-end delay with task J_n in both command and feedback paths (see lines 7-14 in Algorithm 5). This solves the assignment sub-problem. Next, the proposed CAPSTA algorithm tackles the scheduling

Algorithm 5 CAPSTA Algorithm

```
**Input:** $\boldsymbol{\mathcal{J}}, \boldsymbol{\mathcal{M}}, T_j, \Omega_j, D_j; \forall j \in \boldsymbol{\mathcal{J}}, \mathbf{D}_c, \mathbf{D}_f$
**Output:** $S_j, C_j, z_{jk}; \forall j \in \boldsymbol{\mathcal{J}}, \forall k \in \boldsymbol{\mathcal{M}}$
 1: Sort $\boldsymbol{\mathcal{J}}$ in a decreasing order of $\frac{\Omega_j}{T_j}, \forall j \in \boldsymbol{\mathcal{J}}$
 2: $k \leftarrow 0$
 3: for $j = 1$ to $N$ do
 4:     $k \leftarrow k + 1$
 5:     $k^* \leftarrow \begin{cases} \text{mod}(k, M) & \text{if mod}(k, M) \neq 0 \\ M & \text{otherwise} \end{cases}$
 6:     $D^c_{k^*j} \leftarrow$ Use Eq. (4.7) to estimate the average end-to-end packet delay in the command path
 7:     $D^f_{jk^*} \leftarrow$ Use Eq. (4.8) to estimate the average end-to-end packet delay in the feedback path
 8:     if $\max\{D^c_{k^*j}, D^f_{jk^*}\} \leq D_0$ then
 9:         $z_{jk^*} \leftarrow 1$
10:         $\boldsymbol{\mathcal{S}}_{k^*} \leftarrow \boldsymbol{\mathcal{S}}_{k^*} \cup \{J_j\}$
11:     else
12:         $k^* \leftarrow \underset{O_k \in \boldsymbol{\mathcal{M}}}{\text{argmin}} \left\{ \max\{D^c_{kj}, D^f_{jk}\} | k = 1, 2, ..., M \right\}$
13:         $z_{jk^*} \leftarrow 1$
14:         $\boldsymbol{\mathcal{S}}_{k^*} \leftarrow \boldsymbol{\mathcal{S}}_{k^*} \cup \{J_j\}$
15:     end if
16: end for
17: for $k = 1$ to $M$ do
18:     $t \leftarrow 0$
19:     while $\boldsymbol{\mathcal{S}}_k \neq \varnothing$ do
20:         $J_{j^*} = \underset{J_j \in \boldsymbol{\mathcal{S}}_k}{\text{argmin}} \left\{ \frac{D_j}{\Omega_j} \right\}$
21:         $S_{j^*} \leftarrow t$
22:         $C_{j^*} \leftarrow S_{j^*} + T_{j^*}$
23:         $t \leftarrow C_{j^*}$
24:         $\boldsymbol{\mathcal{S}}_k \leftarrow \boldsymbol{\mathcal{S}}_k \setminus \{J_{j^*}\}$
25:     end while
26: end for
27: return $S_j, C_j, z_{jk}, \forall j = 1, ..., N, k = 1, ..., M$
```

sub-problem to HOs. Toward this end, among unscheduled tasks, we first select the task with the minimum amount of $\frac{D_j}{\Omega_i}$ and then schedule it when the HO first becomes available (see lines 16-25 in Algorithm 5). This, as a result, gives preference to the tasks with larger weights and shorter due times.

4.4.3 Complexity analysis

In the proposed CAPSTA algorithm, partitioning the given task set $\boldsymbol{\mathcal{J}}$ into M subset returns a solution with complexity $\mathcal{O}(N \log N) + \mathcal{O}(N) = \mathcal{O}(N \log N)$. Next, CAPSTA solves the scheduling sub-problem with time complexity $\mathcal{O}(\left\lceil \frac{N}{M} \right\rceil \log \left\lceil \frac{N}{M} \right\rceil) + \mathcal{O}(N.M)$. The overall time complexity is thus calculated as $\mathcal{O}(N \log N) + \mathcal{O}(M.N)$, which reduces to $\mathcal{O}(N \log N) + \mathcal{O}(N^2)$ since $M \ll N$.

4.5 Delay Analysis

Recall from above that in order to ensure the quality-of-control of local/non-local teleoperation loops, the average end-to-end delay of HO-TOR pairs should not exceed a given threshold. Thus, in order to ensure the proper performance of our proposed CAPSTA algorithm, it is of vital importance to estimate the connection delay between any given TOR and the available HOs in both command and feedback paths. Toward this end, we develop our analytical framework to estimate the average end-to-end packet delay of local and non-local teleoperation in FiWi-based Tactile Internet infrastructures. In our analysis, we make the following assumptions:

- *Single-hop WLAN*: MUs, HOs, and TORs are directly associated with an ONU-AP via a wireless single hop, whereby ONU-MPPs serve as ONU-APs.
- *Haptic traffic model*: In both command and feedback paths, HOs and TORs transmit their update packets at a rate of 1000 packets/second with fixed deterministic interarrival times set to 1 ms [11].
- *Background traffic model* : MUs generate background Poisson traffic with mean packet rate λ_B (in packets/second). In addition, the background traffic rate generated by ONUs with attached fixed (wired) subscribers that are directly connected to the backhaul EPON is set to $\lambda_{ONU} = \alpha_{PON} \lambda_B$, where α_{PON} is a traffic scale factor.

For notational convenience, let us use the term “WiFi user” for all MUs, HOs, and TORs within the coverage area of an ONU-AP. We model each WiFi user as a GI/G/1 queue to account for the different packet interarrival time distributions under consideration (i.e., Poisson for background traffic and deterministic for haptic traffic). While the GI/G/1 queueing model requires the fewest assumptions among other models, it yields quite conservative results in that we can obtain only an upper bound for the average delay experienced by any packet. An accurate analysis of GI/G/1 queues can be done by solving the Lindley’s integral equation in [85]. Closed-form solutions, however, are difficult to obtain, except for some known distributions. Therefore, we use the approximation method presented in [86] to estimate the upper bound of the average packet delay.

Let the delay experienced by any packet generated by a WiFi user be denoted by random variable D, which is the sum of the queueing delay D_Q and service time (channel access delay) D_S. To begin with, let the number of packets in the system (i.e., queue and server) be denoted by N_t, which is approximated as

$$\mathbb{E}(N_t) \approx \left(\frac{\rho^2\left(1+C_s^2\right)}{1+\rho^2 C_s^2}\right)\left(\frac{C_a^2+\rho^2 C_s^2}{2(1-\rho)}\right)+\rho, \tag{4.9}$$

where C_s and C_a denote the coefficient of variation of service and interarrival times, and ρ denotes the server utilization. According to Little's law, the average delay experienced by an arbitrary packet since the time it arrives in the queue until it successfully departs service is then calculated as

$$\mathbb{E}(D) = \frac{\mathbb{E}(N_t)}{\lambda} = \overbrace{\frac{1}{\lambda}\left(\frac{\rho^2\left(1+C_s^2\right)}{1+\rho^2 C_s^2}\right)\left(\frac{C_a^2+\rho^2 C_s^2}{2(1-\rho)}\right)}^{\text{average queueing delay } \mathbb{E}(D_Q)} + \mathbb{E}(D_S). \tag{4.10}$$

Clearly, in order to obtain $\mathbb{E}(D)$, we need to calculate the mean service time and coefficient of variation of service time. This requires to obtain the first and second moments of service time D_S.

To compute the first and second moments of channel access delay D_S in Eq. (4.10), we defined the two-dimensional Markov chain $(s(t), b(t))$ shown in Fig. 2.10 under unsaturated traffic conditions and estimated the average service time $\mathbb{E}(D_S)$ and service time variance $\mathbb{VAR}(D_S)$ in a WLAN using IEEE 802.11 DCF for access control (see Chapter 2 for further details).

We then obtain the first and second moments of the packet delay as follows:

$$\mathbb{E}(D_S) = \sum_{k=0}^{\infty} p_e^k \cdot (1-p_e) \cdot \left[\sum_{j=0}^{\infty} p_c^j\,(1-p_c)\left(\left(\sum_{b=0}^{k+j} \frac{2^{min(b,m)}W_0-1}{2} E_s\right) + jT_c + kT_e + T_s\right)\right],$$

$$\mathbb{VAR}(D_S) = \sum_{k=0}^{\infty} p_e^k\,(1-p_e) \sum_{j=0}^{\infty} p_c^j\,(1-p_c)\,Q^2(j,k) - \mathbb{E}^2(D_S),$$

with

$$Q(j,k) = \left(\sum_{b=0}^{k+j} \frac{2^{min(b,m)}W_0-1}{2} E_s\right) + jT_c + kT_e + T_s. \tag{4.11}$$

After finding $\mathbb{E}(D_S)$ and $\mathbb{VAR}(D_S)$ and given the incoming traffic rate λ, we can now compute the average delay experienced by any packet for a given WiFi subscriber using Eq. (4.10).

Next, we calculate the average delay experienced by an arriving packet at the EPON backhaul. In doing so, we build on the analytical frameworks presented in [13] and [75]. We first define the backhaul downstream traffic intensity ρ^u and ρ^d for a TDM PON ($\Lambda = 1$) and a WDM PON ($\Lambda > 1$) as

$$\rho^u = \frac{\bar{L}}{\Lambda \cdot c_{PON}} \sum_{q=1}^{O} \sum_{i=0}^{O} \Gamma_{qi}^{PON} < 1, \tag{4.12a}$$

$$\rho^d = \frac{\bar{L}}{\Lambda \cdot c_{PON}} \sum_{q=0}^{O} \sum_{i=1}^{O} \Gamma_{qi}^{PON} < 1, \tag{4.12b}$$

where c_{PON} denotes the PON data rate, O denotes the number of ONUs, and Γ_{qi}^{PON} represents the traffic rate (in packets/second) between PON nodes q and i (with $q = 0$ denoting the OLT).

Similar to [75], the upstream delay, D_{PON}^u, and downstream delay, D_{PON}^d, of both TDM and WDM PONs are obtained as

$$D_{PON}^u = \Phi(\rho^u, \bar{L}, \varsigma^2, c_{PON}) + \frac{\bar{L}}{c_{PON}} + 2\tau_{PON}\frac{2-\rho^u}{1-\rho^u} - B^u, \tag{4.13}$$

$$D_{PON}^d = \Phi(\rho^u, \bar{L}, \varsigma^2, c_{PON}) + \frac{\bar{L}}{c_{PON}} + \tau_{PON} - B^u, \tag{4.14}$$

where τ_{PON} denotes the average propagation delay between ONUs and OLT, $\Phi(\cdot)$ is the average queueing delay of an M/G/1 queue characterized by the Pollaczek-Khintchine formula as

$$\Phi(\rho, \bar{L}, \varsigma^2, c) = \frac{\rho}{2c(1-\rho)} \left(\frac{\varsigma^2}{\bar{L}} + \bar{L} \right), \tag{4.15}$$

and

$$B^d = B^u = \Phi\left(\frac{\bar{L}}{\Lambda \cdot c_{PON}} \sum_{q=1}^{O} \sum_{i=1}^{O} \Gamma_{qi}^{PON}, \bar{L}, \varsigma^2, c_{PON} \right). \tag{4.16}$$

In the following, we proceed to estimate the elements of the command delay matrix $\mathbf{D}_c$ and feedback delay matrix $\mathbf{D}_f$, accounting for both local and non-local teleoperation scenarios.

Local teleoperation: If human operator O_k and the teleoperator robot that is collocated with task J_j are both associated with the same ONU-AP, the average end-to-end packet delay D_{kj}^c,

$\forall k = 1, 2, ..., M$ and $j = 1, 2, ..., N$, in the command path is estimated as

$$D^c_{kj} = \mathbb{E}(D_{O_k}) + \mathbb{E}(D_{\mathrm{ONU-AP}_n}), \tag{4.17}$$

where $\mathbb{E}(D_X)$ for a given WiFi subscriber X is obtained from Eq. (4.10) and $\mathrm{ONU} - \mathrm{AP}_n$ denotes the ONU-AP, which human operator k and teleoperator robot j are connected to.

The average end-to-end packet delay D^f_{jk}, $\forall j = 1, 2, ..., N$ and $k = 1, 2, ..., M$, in the feedback path is then estimated as

$$D^f_{jk} = \mathbb{E}(D_{\mathrm{TOR}_j}) + \mathbb{E}(D_{\mathrm{ONU-AP}_n}). \tag{4.18}$$

Note that in local teleoperation, the average end-to-end delay in command and feedback paths may, in general, be different due to different traffic patterns/rates, bit error probabilities, and MAC settings, among others.

Non-local teleoperation: Unlike local teleoperation, non-local teleoperation is carried out, if human operator O_k and the teleoperator robot that is collocated with task J_j are associated with different ONU-APs. The average end-to-end packet delay D^c_{kj}, $\forall k = 1, 2, ..., M$ and $j = 1, 2, ..., N$, in the command path is therefore estimated as

$$D^c_{kj} = \mathbb{E}(D_{O_k}) + D^u_{PON} + D^d_{PON} + \mathbb{E}(D_{\mathrm{ONU-AP}_{n'}}), \tag{4.19}$$

which accounts for the average upstream delay D^u_{PON} and downstream delay D^u_{PON} in the backhaul EPON given in Eqs. (4.13) and (4.14), respectively. Also note that $\mathrm{ONU} - \mathrm{AP}_{n'}$ denotes the ONU-AP with which human operator O_k is associated.

The average end-to-end packet delay D^f_{jk}, $\forall j = 1, 2, ..., N$ and $k = 1, 2, ..., M$, in the feedback path is then estimated as

$$D^f_{jk} = \mathbb{E}(D_{\mathrm{TOR}_j}) + D^u_{PON} + D^d_{PON} + \mathbb{E}(D_{\mathrm{ONU-AP}_{n'}}). \tag{4.20}$$

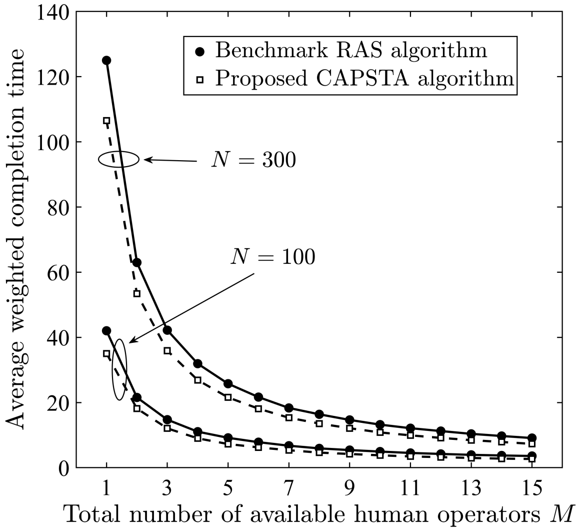


Figure 4.4 – Average weighted completion time of tasks vs. total number of available human operators M ($\alpha = 1$ fixed).

4.6 Results

In this section, we examine our proposed CAPSTA. In our simulations, the task operation time T_j is sampled from a discrete uniform distribution over the range of $[10, 30]$ seconds. The delay threshold D_0 is set to 10 ms. The weight Ω_j is randomly chosen from $\{1, 2, 3, 4\}$ (i.e., four different classes). The due times are randomly chosen from $\alpha \cdot [1, \lceil \frac{1}{M} \sum_{j=1}^{N} T_j \rceil]$, where $\lceil x \rceil$ denotes the smallest integer greater than or equal to x. Each point shown in the following results is averaged over 50 randomly generated problem instants and falls within the 95% confidence interval. We compare the performance of our proposed CAPSTA algorithm with a benchmark *random assignment and scheduling* (RAS) algorithm, where for a given task a human operator is randomly selected from the pool of available ones [36].

First, we present the average weighted completion time (AWCT) and maximum tardiness vs. total number of available human operators M in Figs. 4.4 and 4.5, respectively. We observe from Fig. 4.4 that increasing M results in an exponential decrease of AWCT in both RAS and proposed CAPSTA algorithms. Specifically, in the proposed CAPSTA algorithm, increasing M from 1 to 3 results in a 67% reduction, whereas increasing M from 3 to 5 results in only 41% reduction of AWTC. Further, we note that the proposed CAPSTA algorithm achieves a 15-27% reduction of AWTC compared to the RAS algorithm. Although achieving a lower AWCT, the beneficial impact

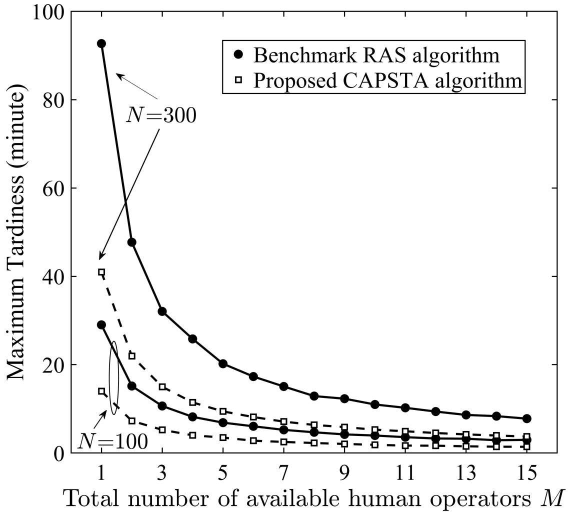


Figure 4.5 – Maximum tardiness of tasks vs. total number of available human operators M ($\alpha = 1$ **fixed).**

of the proposed CAPSTA algorithm compared to the RAS algorithm is more pronounced in terms of the maximum tardiness, as shown in Fig. 4.5. We observe that the proposed CAPSTA algorithm achieves a 49-56% reduction of maximum tardiness. Specifically, for $N = 300$, in order to keep the maximum tardiness below 25 minutes, a total number of 5 human operators is needed in the RAS algorithm, whereas in the proposed CAPSTA algorithm, only 2 human operators are sufficient to achieve the same performance level. Further, if the decision maker likes keep the maximum tardiness below 10 minutes, then the number of required human operators is 5 and 12 in the proposed CAPSTA and RAS algorithms, respectively, thus achieving a notable saving in OPEX, to be examined shortly.

Next, we investigate the impact of increasing due time on the portion of the total tasks of different classes that are subject to tardiness. Toward this end, let us define R_Ω as the rate of tardy tasks with weight Ω to the total number of tasks in the same class. We have considered four different priority classes **A**, **B**, **C**, and **D**, which are associated with weight W equal to 1, 2, 3, and 4, respectively. The results of R_Ω vs. α are shown in Fig. 4.6. First, for small average task due times, 94% of class **D** tasks cannot be accomplished within the expected due times, thus most of the tasks are regarded as tardy tasks. This figure is >98% for class **A**-**C** tasks. Nevertheless, as the average given due time increases, the portion of class **D** tasks that are subject to tardiness decreases exponentially. Specifically, for $\alpha = 1$, R_Ω drops below 2%. Second, we find that the

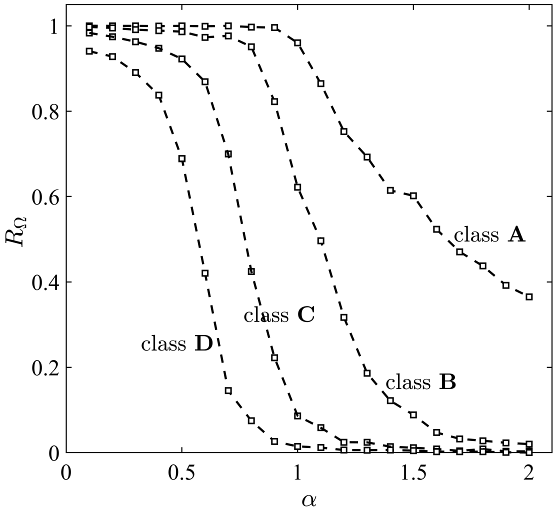


Figure 4.6 – Rate R_Ω of tardy tasks vs. α for different task classes ($N = 300$ and $M = 5$ fixed).

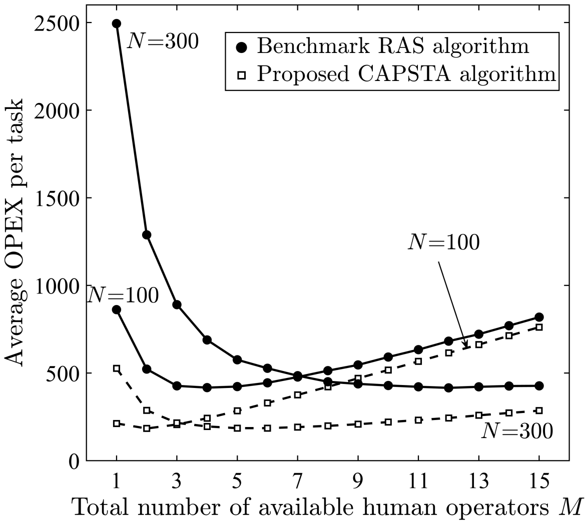


Figure 4.7 – Average OPEX per task vs. total number of available human operators M ($\epsilon_h = 1$, $\epsilon_m = 5000$, and $\alpha = 1$ fixed).

proposed CAPSTA algorithm schedules the tasks in favor of high-priority ones, especially for α greater than 0.5, as shown in Fig. 4.6. We note that for α equal to 2, R_Ω converges to < 2% for classes **B**-**D**, whereas 36% of class **A** tasks (i.e., low-priority tasks) are still subject to tardiness.

Fig. 4.7 depicts the average OPEX per task vs. total number of available human operators M for both the proposed CAPSTA and benchmark RAS algorithms. Overall, the proposed CAPSTA algorithm outperforms the RAS algorithm in terms of average OPEX per task, especially when the number of tasks is large, i.e., $N = 300$. For $M = 1$, comparing the performance of the proposed CAPSTA algorithm with the benchmark RAS algorithm, we observe a 75.3% and 78.9% reduction of average OPEX per task for $N = 100$ and $N = 300$, respectively. As M increases, the OPEX savings of the CAPSTA algorithm with respect to RAS algorithm decreases until both curves converge. The reason for this is that when the total number of available human operators M is small, the incurred OPEX is mainly due to tardy tasks, which are penalized proportional to the weighted amount of tardiness. Fig. 4.7 demonstrates that the efficient scheduling of the proposed CAPSTA algorithm reduces the number of high-priority task that are subject to tardiness, thus achieving a significant reduction of the average OPEX per task, compared to that of the benchmark RAS algorithm.

More importantly, Fig. 4.7 gives us further insights into selecting an optimal number of human operators, which should be, on one hand, large enough to reduce the number of high-priority tardy tasks, and, on the other hand, small enough to avoid incurring excessive OPEX due to activating new teleoperation sessions. For the proposed CAPSTA algorithm, the optimal number of available human operators $M^\star$ that minimizes $C(\mathbf{X})$ is 2 and 5 for $N = 100$ and $N = 300$, respectively. We note that for the proposed CAPSTA algorithm with $M < 3$, the average OPEX per task for $N = 100$ is less compared to that of $N = 300$. Both curves meet at $M = 4$ and then the OPEX per task for $N = 100$ grows larger than that of $N = 300$. The reason for this is that for $N = 100$, while increasing M doesn't result in a further decrease of tardiness, it does result in an excessive increase of OPEX due to the incurred activation costs of new teleoperation sessions. In contrast, for $N = 300$, a large portion of the tasks are subject to tardiness, thus increasing M reduces the incurred OPEX due to tardiness, which in turn partly compensates for the incurred OPEX due to activating teleoperation sessions.

The average OPEX per task vs. M for different $\alpha \in \{0.1, 0.5, 1, 2\}$ for a fixed $N = 100$ is illustrated in Fig. 4.8, where we examine the impact of increasing average task due times on the OPEX performance of our proposed CAPSTA algorithm. We find that for $\alpha = 0.1, 0.5$ and 1, the average OPEX, $C(\mathbf{X})$, is a convex function of M, having a minimum at $M^\star = 6$, 4, and 2, respectively, compared to that of $\alpha = 2$, where $C(\mathbf{X})$ increases linearly for increasing M, as explained above. We note that for relatively relaxed due times (i.e., $\alpha = 2$), the contribution of the incurred penalty due to task tardiness is negligible compared to that of activating excessive teleoperation sessions. Hence, the average OPEX grows linearly as M increases. Therefore, when

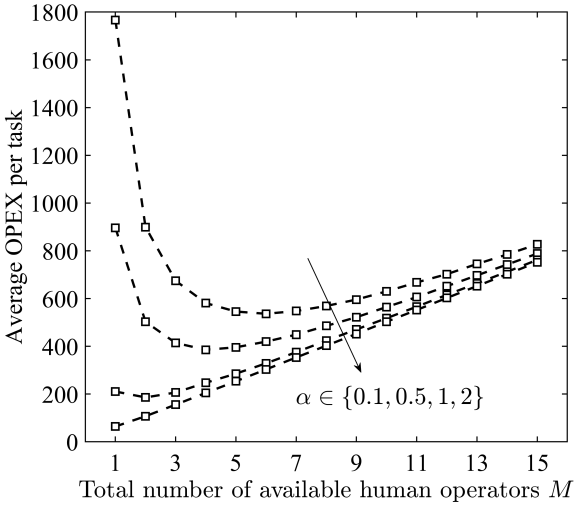


Figure 4.8 – Average OPEX per task vs. total number of available human operators M ($\epsilon_h = 1$, $\epsilon_m = 5000$, and $N = 100$ fixed).

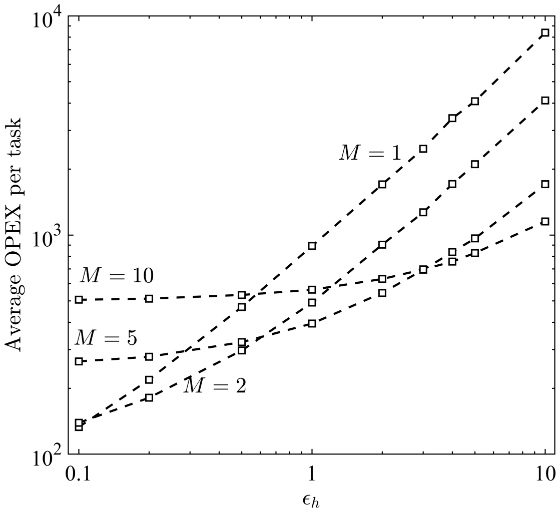


Figure 4.9 – Average OPEX per task vs. ϵ_h for different number of available human operators $M \in \{1, 2, 5, 10\}$ ($\alpha = 0.5$, $\epsilon_m = 5000$, and $N = 100$ fixed).

the average due time is large, it is beneficial to perform the teleoperation tasks by only one human operator, provided that proper scheduling is fulfilled (see Fig. 4.8).

Next, the average OPEX vs. operational cost, ϵ_h, per time unit of tardiness is shown in Fig. 4.9, which renders the following interesting insights. First, Fig. 4.9 specifies the range of ϵ_h for which the proposed CAPSTA algorithm achieves more beneficial results in terms of OPEX for two given $M = M_1$ and $M = M_2$. For instance, while $M = 2$ always leads to a smaller OPEX per task

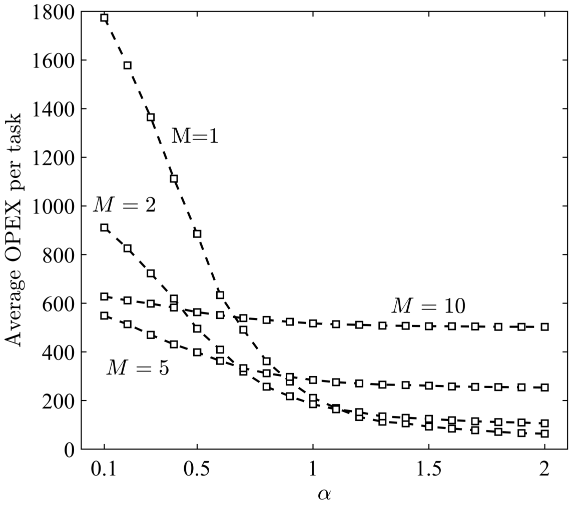


Figure 4.10 – Average OPEX per task vs. α for different number of available human operators $M \in \{1, 2, 5, 10\}$ ($\epsilon_h = 1$, $\epsilon_m = 5000$, and $N = 100$ fixed).

compared with that of $M = 1$, decreasing the number of HOs from $M = 5$ to $M = 2$ does not achieve such reduction. To be more specific, $M = 5$ is more OPEX-beneficial than $M = 2$ only if ϵ_h is greater than 0.5. This, however, is a quite counterintuitive observation whether or not increasing M results in OPEX savings depends not only on the average task due times (as explained before) but also the operational cost, ϵ_h, per time unit of tardiness, as increasing M from 2 to 10 only incurs an additional OPEX due to activating new teleoperation sessions. Further, we also note that for $M = 10$, the rate at which $C(\mathbf{X})$ increases degrades as ϵ_h grows. This is due to the fact that for a large M (e.g., $M = 10$) OPEX is less likely due to task tardiness, thus increasing ϵ_h does not increase $C(\mathbf{X})$ significantly, as opposed to small values of M (e.g., $M = 1$), where increasing ϵ_h results in a significant increase of $C(\mathbf{X})$. As a result, Fig. 4.9 together with Fig. 4.8 are instrumental in helping optimize OPEX for a given set of system parameter values.

Next, we examine the impact of average task due times on the OPEX performance of the proposed CAPSTA algorithm. Fig. 4.10 presents the average OPEX per task vs. α, which reflects the amount of average task due time. For $M = 1$, 2, and 5, the average OPEX per task decreases for increasing α and levels off for $\alpha > 1$. This is due to the fact that for smaller values of M, OPEX is mainly due to penalizing the tardy tasks. Therefore, increasing α translates into a reduced average tardiness, thus alleviating the average OPEX per task. On the other hand, for large values of M

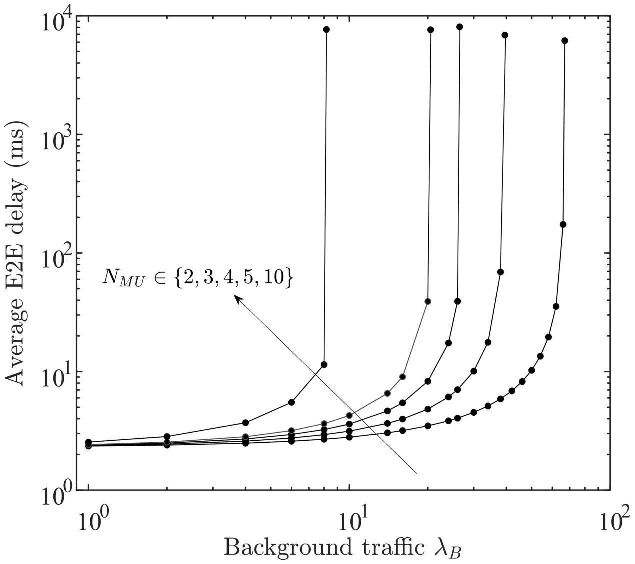


Figure 4.11 – Average end-to-end packet delay of local teleoperation vs. background traffic rate λ_B for different $N_{MU} \in \{2, 3, 4, 5, 10\}$.

(e.g., $M = 10$), the contribution of the first term of Eq. (4.6) to OPEX is greater than the second term. For this reason, we do not observe a notable decrease of the OPEX as α increases.

Finally, we evaluate the end-to-end delay performance of local and non-local teleoperation. We apply the same default parameter settings of IEEE 802.11n DCF as listed in Table I in [13]. We consider four ONU-APs and four conventional ONUs, each serving fixed (wired) subscribers that are all involved in non-local H2H communications among each other. MUs and fixed subscribers generate background traffic at a mean rate of λ_B and $\alpha_{PON} \cdot \lambda_B$, respectively, whereas HOs and TORs generate haptic traffic at a fixed rate of 1000 packets/second. We consider 6-DoF TORs.

Figure 4.11 depicts the average end-to-end packet delay of local teleoperation vs. mean background traffic λ_B for different $N_{MU} \in \{2, 3, 4, 5, 10\}$, where N_{MU} denotes the number of MUs that reside within the coverage of each ONU-AP. We find that an average end-to-end delay of 2.5 ms is achievable for local teleoperation involved HO-TOR pairs. It is worthwhile to mention that the amount of time waited by packets in the second hop (i.e., MAC queue of the ONU-AP) is notably larger than that of the first hop (i.e., MAC queue of the HO/TOR), which is a direct consequence of the high incoming packet rate at the ONU-AP. Second, we observe that for $N_{MU} \in \{2, 3, 4, 5\}$, the average end-to-end delay remains under 10 ms for a wide range of background traffic rate λ_B.

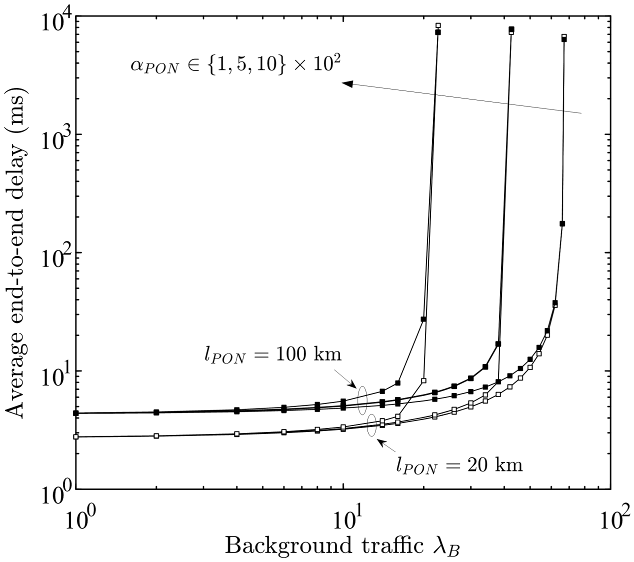


Figure 4.12 – Average end-to-end packet delay of non-local teleoperation vs. background traffic rate λ_B for different $\alpha_{PON} \in \{100, 200, 500\}$ ($N_{MU} = 2$ fixed).

Figure 4.12 illustrates the average end-to-end packet delay of non-local teleoperation vs. background traffic rate λ_B for different values of $\alpha_{PON} \in \{100, 200, 500\}$ and $N_{MU} = 2$. In non-local teleoperation, the obtained end-to-end delay is as low as 2.8 ms and 4.5 ms for l_{PON} =20 and 100 km (compared to 2.5 ms in local teleoperation). Further, we observe that for a given background traffic load, say $\lambda_B = 10$ packets/second, increasing l_{PON} from 20 km to 10 km results in a 1.6 ms increase of the end-to-end delay from 3.5 ms to 5.1 ms. This is counterintuitive in that the 80 km increase of backhaul fiber length accounts for only 267 μs in propagation delay, which is much smaller than 1.6 ms. The reason for this lies in the impact of increasing l_{PON} on the delay performance of the multi-point control protocol (MPCP) protocol used in the backhaul EPON.

4.7 Discussion

4.7.1 SDN/NFV: Potential and Benefits

As the Tactile Internet emerges, the flows generated by different applications become more diverse, each requiring a different QoS/E. To overcome the issues arising from traditional network management models, including limited reconfigurability and complex per-flow traffic management, SDN/NFV is a promising solution, where a clear distinction is made between the control and data planes. This as a result can provide the task coordinator with a logically centralized overview of the whole network, gather application-dependent requirements (teleoperation in our studied scenario), and reconfigure network parameters to achieve the desired QoS/E. In this context, NFV is a promising technique, which can be used not only to further reduce the CAPEX and OPEX issues of teleoperation over FiWi networks, but also to support a wider variety of HSI and TOR types (see Fig. 1). More importantly, given that FiWi networks have to cope with seamless integration of both optical and wireless sub-networks, the role of SDN is even more pronounced in alleviating the difficulties of network design, control, and management, especially with the co-existence of different types of traffic [83]. In this context, [87] presents a thorough review of the studies that examine the SDN paradigm in optical networks, also referred to as software-defined optical networks (SDONs). While the concept of sotfwarization of network protocols realized via SDN enables the study of new ideas and optimization models, thereby significantly reducing the deployment costs and speeding up the upgrade process, virtualization facilitates service migration, thus allowing for location-aware service provisioning in a cost-efficient manner [88].

4.7.2 Deployment Considerations

Recall from Section II that an immersive bilateral teleoperation experience relies on the low-latency, reliable communication between HO and TOR (see Fig. 4.1). In the presence of communication-induced artifacts such as latency and jitter, the role of local control loops becomes more important, especially given that excessive latencies, which may occur in non-local teleoperation, may decouple the HO from the remote environment. Further, the predictive bandwidth allocation proposed in [23] may be used to reduce the average end-to-end latency of haptic traffic. Complementary to [23], the multi-sample-ahead-of-time forecasting scheme proposed in [46] may also be leveraged to ensure the reliable delivery of feedback samples at a 1-ms granularity, thus helping the HO maintain connectivity with the remote environment.

4.8 Conclusions

We investigated the performance of our proposed CAPSTA algorithm in solving the prioritized assignment and scheduling of delay-constrained teleoperation tasks in FiWi enhanced Tactile Internet network infrastructures. The obtained results show that the proposed algorithm reduces the average weighted task completion time, maximum tardiness, and average OPEX, compared to the benchmark RSA algorithm. Specifically, the proposed CAPSTA algorithm achieves a 15-27% reduction of average weighted task completion time and a 49-56% reduction of maximum tardiness. In addition, compared to the benchmark RAS algorithm, the proposed CAPSTA algorithm achieves a 75.3% and 78.9% reduction of average OPEX per task for $N = 100$ and $N = 300$, respectively. Our results also give insights into finding the optimal number of HOs to minimize the average OPEX per completed task for different deployment scenarios. More precisely, we have shown that for the proposed CAPSTA algorithm, the optimal number of available human operators $M^\star$ that minimizes OPEX is 2 and 5 for $N = 100$ and $N = 300$, respectively. Finally, we have shown that the considered solution is able to achieve an average end-to-end packet delay of < 10 ms for both local and non-local teleoperation for a wide range of background traffic rates. An interesting future research avenue is to investigate the role of virtualization in FiWi networks to eliminate the physical layer interaction of the often heterogenous Tactile Internet applications, thus realizing a infrastructure/technology independent architecture.

Chapter 5

Cooperative Computation Offloading Using Self-Organizing MEC

This chapter contains material extracted from the following publication:

[53] A. Ebrahimzadeh and M. Maier. Distributed Cooperative Computation Offloading in Multi-Access Edge Computing Fiber-Wireless Networks (Invited paper). *Elsevier Optics Communications Special Issue on Photonics for 5G Mobile Networks and Beyond*, vol. 452, pp. 130-139, Dec. 2019.

[54] A. Ebrahimzadeh and M. Maier. Cooperative Computation Offloading in FiWi Enhanced 4G HetNets Using Self-Organizing MEC. *IEEE Transactions on Wireless Communications*, in revision.

[55] A. Ebrahimzadeh and M. Maier. Next Generation Multi-Access Edge-Computing Fiber-Wireless Enhanced HetNets for Low-Latency Immersive Applications. *IGI Global*: Design, Implementation, and Analysis of Next Generation Optical Networks, accepted for publication.

In the following, my key contributions in the aforementioned publications are explained in greater detail: (1) I largely contributed to writing the whole manuscripts, (2) I developed the analytical frameworks, (3) I conducted the algorithmic works, and (4) I ran the simulations.

5.1 Introduction

To address the contradiction between the rapid increase of computation-intensive, delay-sensitive applications (e.g., Tactile Internet, AR/VR, and interactive gaming) and resource-limited smart mobile devices, MCC has emerged to reduce the computational burden of mobile devices and broaden their capabilities by extending the concept of cloud computing to the mobile environment via full and/or partial computation offloading. Even though MCC allows mobile devices to benefit from powerful computing resources to save battery power and accelerating task execution, it raises several technical challenges due to additional communication overhead and poor reliability that remote computation offloading may introduce. To overcome these limitations, mobile edge computing has recently emerged to provide cloud computing capabilities in the edge of access networks, leveraging the physical proximity of edge servers and mobile users to achieve a reduced communication latency and increased reliability [89]. More recently, the ETSI has dropped the word “mobile” and introduced the term *multi-access edge computing* in order to broaden its applicability to heterogeneous networks, including WiFi and fixed access technologies (e.g., fiber) [14].

While a conventional (remote) cloud provides high storage and computational capabilities, it may pose large latency due to communications, as it is usually physically distant from the mobile users. On the other hand, MEC may offer a reduced communication-induced latency, but it may pose an excessive processing latency due to limited computational capabilities. In a broader vision, remote cloud and MEC servers can thus coexist and be complementary to each other, giving rise to *cooperative computation offloading*. The ultimate goal of MEC, in fact, is to achieve an ultra-low response time, which is defined as the time interval between the time instant at which a task is released from a mobile device until it is processed (either locally or remotely) and the result is received by the device. This time interval may include the waiting (queueing) and processing times in either the local central processing unit (CPU) or edge/remote server as well as the communication latency between the mobile device and edge/remote cloud. Given the additional communication overhead that offloading introduces, a key technical challenge is to find a tradeoff between the cost of computation and communication to enhance user experience in terms of lower latency and energy consumption. In this chapter, motivated by [90], we focus on the QoE of mobile users measured by the average response time that can be influenced by the queueing/processing and transmission delay components, including those between mobile users and MEC servers and also between MEC servers and the remote cloud.

To achieve the desired energy-delay performance, the so-called dynamic voltage scaling (DVS) is a promising technique that varies the supply voltage and clock frequency based on the computation load to achieve a suitable tradeoff between task execution time and energy consumption [42]. While computation offloading mainly relies on the computational capabilities of the edge/remote servers, the DVS technique enables the mobile users to adaptively adjust their computational speed to reduce energy consumption or shorten task execution time. Therefore, incorporating the DVS technique into computation offloading offers more flexibility at the device side, enabling mobile users to achieve self-awareness via a design approach commonly known as *self-organization* to further improve their QoE under different scenarios [91].

It is evident that future 5G mobile networks will lead to an increasing integration of cellular and WiFi technologies and standards, giving rise to so-called HetNets, which mandates the need for addressing the backhaul bottleneck challenge [46]. Recently, we have explored the performance gains obtained from unifying coverage-centric 4G LTE-A HetNets and capacity-centric FiWi access networks based on data-centric Ethernet technologies with resulting fiber backhaul sharing and WiFi offloading capabilities towards realizing future 5G networks [13]. By means of probabilistic analysis and verifying simulations based on recent and comprehensive smartphone traces, we showed that an average end-to-end latency of < 10 ms can be achieved for a wide range of traffic loads and that mobile users can be provided with highly fault-tolerant FiWi connectivity for reliable low-latency fiber backhaul sharing and WiFi offloading. Note, however, that only data offloading was considered in [13] without any computation offloading via MEC. Furthermore, the feasibility of implementing conventional cloud and MEC in FiWi access networks was investigated in [92], where the main objective was to design a unified resource management scheme to integrate offloading activities with the underlying FiWi operations. While much of the effort in these papers has been devoted to the management of networking resources, cooperation between mobile devices, MEC servers, the remote cloud and the problem of offloading decision making have not been investigated. In [93], a scalable online algorithm for task scheduling in an edge-cloud system was proposed, which was verified by simulations using real-world traces from Google. A hierarchical MEC-based architecture was presented in [94] with a focus on the workload placement problem. In [95], an optimization framework was presented for solving the problem of joint offloading decision and allocation of computation and communication resources with the aim of minimizing a weighted sum of the costs of energy, cost of computation, and the delay for all users. More recently, the authors of [90] studied the computation offloading problem for cooperative fog computing networks and investigated the fundamental tradeoff between QoE of mobile users and power efficiency of fog nodes. In [96], a collaborative computation offloading scheme for MEC over FiWi networks was presented. All mentioned papers, however, mainly focused on the management of computing resources without further investigating the impact of the capacity-limited backhaul.

In this chapter, we examine the performance gains obtained by cooperative computation offloading in MEC enabled FiWi enhanced HetNets, which relies on not only the computational capabilities of edge/cloud servers but also the limited local computing resources at the device side. More specifically, we aim to design a two-tier MEC enabled FiWi enhanced HetNet architecture, where the mobile devices as well as the edge servers cooperatively offload their computation tasks towards achieving a reduced average response time. We take into account both crucial aspects of limitations stemming from communications and computation in our design approach via accurate modeling of the fronthaul/backhaul as well as edge/cloud servers, while paying particular attention to offloading decision making between mobile users and edge servers as well as edge servers and the remote cloud. Another important aspect of MEC is to cope with the additional complexity that may arise in such a scenario by relying, fully or partially, on the limited local computing resources of mobile users when they are most needed. The inherent time-varying nature of FiWi enhanced HetNets, which is a direct consequence of user mobility, entails exploiting a function that continuously tune the local computational capabilities of mobile devices in order to ensure an improved QoE. This can be achieved via adaptive reconfiguration of a mobile user given its goals, capabilities, and constraints via a design approach commonly known as self-awareness. Contributing to this effort, we leverage on the self-awareness of mobile users by applying the DVS technique for making appropriate energy-delay tradeoffs subject to given energy and delay constraints. In particular, the contributions of this chapter are as follows:

- We design a two-tier hierarchical MEC enabled FiWi enhanced HetNet-based architecture for computation offloading, which leverages both local and nonlocal computing resources to achieve low response time and energy consumption for mobile users. We also propose a simple but efficient offloading orchestration mechanism to achieve an improved QoE for mobile users.

- We develop an analytical framework to examine the performance of our proposed FiWi based cooperative offloading scheme coexistent with conventional H2H traffic (i.e., voice, video, and data) in terms of average response time as well as energy consumption of mobile users. In our analysis, we develop detailed models of both communication and computation, incorporating WiFi/LTE-A wireless access and capacity-limited backhaul fiber links as well as resource-limited edge/remote cloud servers.

- Given the additional complexity incurred by integrating the cooperative computation offloading strategy in a FiWi enhanced HetNet architecture, any deviation from optimal delay performance is inevitable. To cope with this and in order to allow mobile users to flexibly rely on their local computing resources by means of reconfiguration, we propose a self-organization framework to allow mobile devices to adaptively tune their offloading probability as well as computational capabilities via the DVS technology. The proposed self-organizing design results in a Pareto frontier characterization of the tradeoff between average task execution time and energy consumption.

The remainder of the chapter is structured as follows. In Section 5.2, we present our proposed architecture of MEC enabled FiWi enhanced HetNets and cooperative offloading mechanism. In Section 5.3, we present our analytical framework for estimating the energy-delay performance of our proposed cooperative task offloading scheme. The proposed self-organization scheme is presented in Section 5.4. Section 5.5 presents numerical results. Finally, Section 5.6 concludes the chapter.

5.2 Network Architecture and System Model

Figure 5.1 depicts the generic architecture of the considered FiWi enhanced LTE-A HetNets. The fiber backhaul consists of a TDM/WDM IEEE 802.3ah/av 1/10 Gbps EPON with a typical fiber range of 20 km between the central OLT and remote ONUs. The EPON may comprise multiple stages, each separated by a wavelength broadcasting splitter/combiner or a wavelength multiplexer/demultiplexer. There are three different subsets of ONUs. An ONU may either serve fixed (wired) subscribers. Alternatively, it may connect to a cellular network BS or an IEEE 802.11n/ac/s WLAN MPP, giving rise to a collocated ONU-BS or ONU-MPP, respectively. Depending on her

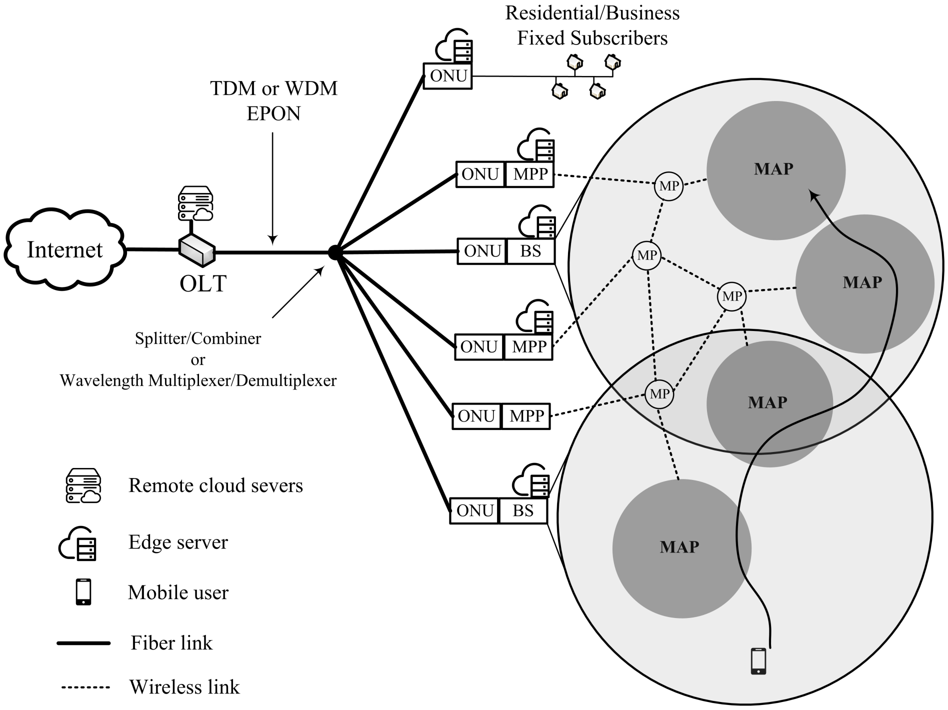


Figure 5.1 – Generic MEC-enabled FiWi enhanced LTE-A HetNets architecture.

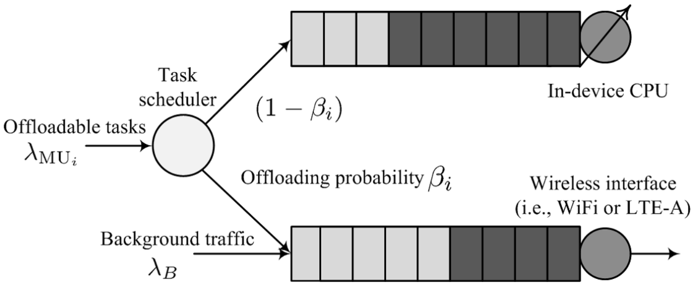


Figure 5.2 – Schematic of task scheduler and queueing system for MU i, which includes two disjoint queues served by local CPU and WiFi/LTE-A wireless interface.

trajectory, an MU may communicate through the cellular network and/or WLAN mesh front-end, which consists of ONU-MPPs, intermediate MPs, and MAPs.

We equip selected ONU-BSs/MPPs with MEC servers (or simply called *edge servers* hereafter) collocated at the optical-wireless interface. MUs may offload fully or portion of their incoming computational tasks to nearby edge servers. In addition to edge servers, the OLT is equipped with cloud computing facilities, which consist of multiple servers dedicated to processing mobile tasks. Each MU uses a task scheduler that decides whether to offload a task to an edge server or execute it locally in its local CPU. We model the task scheduler in each MU by a queuing system, as illustrated in Fig. 5.2. We assume that in each mobile device there are two servers, namely, the CPU and the wireless interface (i.e., WiFi or LTE-A). The former server is used to model the local task execution at the MU's CPU, whereas the latter is responsible for offloading tasks to an edge server in proximity. We assume MUs generate background Poisson traffic at mean packet rate λ_B (in packets/second) (see Fig. 5.2). We also assume that tasks arrive at MU i's scheduler at rate λ_{MU_i}. The task scheduler at MU i makes its decision based on the value of the so-called *offloading probability*, β_i, which is defined as the probability that an incoming task is offloaded to the edge server. Tasks generated by MU i are characterized by B_i^l and D_i^l, which denote the average size of computation input data (e.g., program codes and input parameters) and average number of CPU cycles required, respectively. Computation tasks are assumed to be atomic and thus cannot be divided into sub-tasks. We also assume that each edge server is equipped with a task scheduler, which decides whether to execute an incoming task or further offload it to the remote cloud. Similar to MUs, a task arriving at edge server j is further offloaded to the remote cloud with probability α_j or executed locally with probability $(1 - \alpha_j)$.

5.3 Energy-Delay Analysis of the Proposed Cooperative Offloading

In this section, we analyze the performance of our proposed cooperative MEC enabled FiWi enhanced LTE-A HetNets in terms of average response time and energy consumption for task offloading coexistent with conventional H2H traffic. Many related recent studies (e.g., [90], [89], [37]-[41]) assumed a Poisson task arrival model and an exponentially distributed number of required CPU cycles for task execution. In this chapter, we follow the same research line and build our analysis on these assumptions. Further, tasks are assumed to be computationally intensive, mutually independent, and can be executed either locally or remotely on an edge server or the remote cloud via computation offloading. Each edge server has a limited computational capability and can serve a single task at a time [90][37]. Besides, the remote cloud comprises a limited number of high-performance computing servers, each of which can serve a single task at a time.

5.3.1 Average response time

In the proposed cooperative offloading scheme, both computation and communication induced latencies may contribute to the resultant average response time experienced by MUs. First, we estimate the latencies due to computation for both local and nonlocal computing. For a given MU i, who is involved in task offloading, assuming i.i.d exponentially distributed task interarrival times and given the offloading probability β_i, the tasks arriving at the CPU queue for local computing follow a Poisson process with rate $(1-\beta_i)\cdot\lambda_{\mathrm{MU}_i}$, whereas the offloaded tasks arriving at the wireless interface queue follow a Poisson process with rate $\beta_i\cdot\lambda_{\mathrm{MU}_i}$. This is because thinning a Poisson process with a fixed probability results in another Poisson process. Let D_i^l be the average number of required CPU cycles to execute a task arriving at MU i. The average local task execution time τ_i^l at MU i is given by

$$\mathcal{D}(f_i)=\tau_i^l=\frac{D_i^l}{f_i}, \tag{5.1}$$

where f_i is the clock frequency (in CPU cycles per second) of MU i. Assuming that the number of required CPU cycles per task follows an exponential distribution, we can model the local CPU server of MU i as an M/M/1 queue with mean arrival rate $(1-\beta_i)\,\lambda_{\mathrm{MU}_i}$ and mean task execution time τ_i^l. The average delay Δ_{MU_i} of local task execution (which includes both queueing and service times) at MU i's CPU is then given by

$$\Delta_{\mathrm{MU}_i}=\frac{1}{\mu_i^l-(1-\beta_i)\,\lambda_{\mathrm{MU}_i}}, \tag{5.2}$$

where μ_i^l, which is equal to $1/\tau_i^l$, is the rate at which the executed tasks depart from MU i's CPU. We note that Eq. (5.2) is valid only if $(1-\beta_i)\,\lambda_{\mathrm{MU}_i}\tau_i^l<1$.

Let $\mathcal{R}_j$ denote the set of MUs that are served by edge server j. Further, let $\lambda_{0,j}^e$ be the mean arrival rate and $D_{0,j}^e$ denote the required number of CPU cycles of offloaded tasks from the fixed (wired) subscribers, if any, which may be directly connected to edge server j. Given the offloading probabilities β_i, $\forall \mathrm{MU}_i\in\mathcal{R}_j$, the mean arrival rate λ_{MEC_j} at the task scheduler of edge server j is computed as follows:

$$\lambda_{\mathrm{MEC}_j}=\lambda_{0,j}^e+\sum_{\mathrm{MU}_i\in\mathcal{R}_j}\beta_i\cdot\lambda_{\mathrm{MU}_i}. \tag{5.3}$$

Let τ_j^e denote the average task execution time at edge server j. For estimating τ_j^e, we compute the average number $\bar{D}_j^e$ of CPU cycles required to execute a task at edge sever j as follows:

$$\bar{D}_j^e = \frac{\lambda_{0,j}^e D_{0,j}^e + \sum_{\mathrm{MU}_i \in \mathcal{R}_j} \beta_i \cdot \lambda_{\mathrm{MU}_i} \cdot D_i^l}{\lambda_{0,j}^e + \sum_{\mathrm{MU}_i \in \mathcal{R}_j} \beta_i \cdot \lambda_{\mathrm{MU}_i}}, \tag{5.4}$$

which is then used to calculate τ_j^e, which is given by

$$\tau_j^e = \frac{\bar{D}_j^e}{f_j^e}, \tag{5.5}$$

where f_j^e is the computational capability (in CPU cycles per second) of edge server j. Modeling edge server j as an M/M/1 queue with mean arrival rate $(1-\alpha_j)\,\lambda_{\mathrm{MEC}_j}$ and mean service time τ_j^e, the average delay Δ_{MEC_j} of task execution at edge server j is calculated as follows[1]:

$$\Delta_{\mathrm{MEC}_j} = \frac{1}{\mu_j^e - (1-\alpha_j)\,\lambda_{\mathrm{MEC}_j}}, \tag{5.6}$$

whereby $\mu_j^e = 1/\tau_j^e$. Substituting Eq. (5.3) in Eq. (5.6) provides the following expression:

$$\Delta_{\mathrm{MEC}_j} = \frac{1}{\mu_j^e - (1-\alpha_j) \cdot \left(\lambda_{0,j}^e + \sum_{\mathrm{MU}_i \in \mathcal{R}_j} \beta_i \cdot \lambda_{\mathrm{MU}_i}\right)}, \tag{5.7}$$

which is valid only if

$$\tau_j^e \cdot (1-\alpha_j) \cdot \left(\lambda_{0,j}^e + \sum_{\mathrm{MU}_i \in \mathcal{R}_j} \beta_i \cdot \lambda_{\mathrm{MU}_i}\right) < 1.$$

Next, we proceed to estimate the task execution delay at the remote cloud. Let $\mathcal{R}$ denote the set of edge servers that are connected to the remote cloud. The mean arrival rate λ_c at the remote cloud is obtained as follows:

$$\lambda_c = \lambda_0^c + \sum_{\mathrm{MEC}_j \in \mathcal{R}} \alpha_j \cdot \lambda_{\mathrm{MEC}_j}. \tag{5.8}$$

Let λ_0^c and D_0^c denote the arrival rate and number of CPU cycles required to execute the background tasks[2] at the remote cloud, respectively.

[1] Although an edge server may comprise a number of physical and/or virtual machines to process the incoming tasks, we are focusing on a coarse-grained scenario, thus modeling an edge server as a single entity, as in many related works, e.g., [90], [37], and [38].

[2] Some of the fixed subscribers may not be connected to any edge server in proximity and thus offload their tasks directly to the remote cloud via the backhaul EPON. We refer to such tasks as cloud background tasks.

Further, let τ_c denote the average task execution time at the remote cloud. In order to estimate τ_c, we first calculate the average number $\bar{D}_c$ of CPU cycles required to execute a task at the remote cloud, which is given by

$$\bar{D}_c = \frac{\lambda_0^c D_0^c + \sum_{\mathrm{MEC}_j \in \mathcal{R}} \alpha_j \cdot \lambda_{\mathrm{MEC}_j} \cdot \bar{D}_j^e}{\lambda_0^c + \sum_{\mathrm{MEC}_j \in \mathcal{R}} \alpha_j \cdot \lambda_{\mathrm{MEC}_j}}, \tag{5.9}$$

which is then used to estimate τ_c as follows:

$$\tau_c = \frac{\bar{D}_c}{f^c}, \tag{5.10}$$

where f^c is the computational capability of each of the s homogeneous servers deployed at the remote cloud. We can thus model the remote cloud as an M/M/s queue with mean arrival rate λ_c (given by Eq. (5.8)) and mean service time τ_c (given by Eq. (5.10)). The average delay Δ_c experienced by an arbitrary task in the remote cloud is then estimated by the well-known Erlang-C formula:

$$\Delta_c = \frac{\mathcal{C}(s,a) \cdot \tau_c}{s-a} + \tau_c, \tag{5.11}$$

where the carried load a is equal to $\lambda_c \cdot \tau_c$ and $\mathcal{C}(s,a)$ is given by

$$\mathcal{C}(s,a) = \frac{\frac{a^s \cdot s}{s!(s-a)}}{\sum_{k=0}^{s-1} \frac{a^k}{k!} + \frac{a^s \cdot s}{s!(s-a)}}, \tag{5.12}$$

which is the probability that an arriving task finds all the s servers busy. Note that Eq. (5.12) is valid only if the carried load does not exceed the number of servers (i.e., $\lambda_c \cdot \tau_c < s$).

Next, we turn our attention to calculating the communication induced latency in our cooperative task offloading scheme. Recall from above that the offloaded tasks arrive at the wireless interface of MU i with rate $\beta_i \cdot \lambda_{\mathrm{MU}_i}$. With L_m denoting the maximum payload size of a single packet, the number of packets per task is equal to $\left\lceil \frac{B_i^l}{L_m} \right\rceil$. We can then estimate the rate Γ_{MU_i} at which packets arrive at the wireless interface of MU i as follows:

$$\Gamma_{\mathrm{MU}_i} = \lambda_B + \left\lceil \frac{B_i^l}{L_m} \right\rceil \cdot \beta_i \cdot \lambda_{\mathrm{MU}_i}, \tag{5.13}$$

where λ_B denotes the background H2H traffic (see also Fig. 5.2).

Delay analysis of WiFi users

First, we calculate the average packet delay Θ_i^{WiFi} in the uplink for MU i, who is associated with an ONU-AP through WiFi. For a given set of network model parameters, we can estimate Θ_i^{WiFi} as in [75]:

$$\Theta_i^{\text{WiFi}} = \frac{1}{\frac{1}{\Delta_i} - \Gamma_{\text{MU}_i}}; \quad \Delta_i \cdot \Gamma_{\text{MU}_i} < 1, \tag{5.14}$$

where Δ_i denotes the average channel access delay and Γ_{MU_i} is given by Eq. (5.13). Note that Eq. (5.14) accounts for both queueing and channel access (service) delay[3].

Lemma 5.1: *The average channel access delay Δ_i of MU i is obtained as follows*

$$\Delta_i = \sum_{k=0}^{\infty} p_{e,i}^{k} \left(1 - p_{e,i}\right) \left[\sum_{j=0}^{\infty} p_{c,i}^{j} \left(1 - p_{c,i}\right) \left(\left(\sum_{b=0}^{k+j} \frac{2^{\min(b,m)} W_0 - 1}{2} E_s \right) + jT_{c,i} + kT_{e,i} + T_{s,i} \right) \right], \tag{5.15}$$

where $p_{e,i}$ is the probability of an erroneous transmission, $p_{c,i}$ is the probability of a collision, W_0 is the initial contention window size, E_s is the expected time-slot duration, and $T_{c,i}$, $T_{e,i}$, and $T_{s,i}$ denote the average duration of a collided, erroneous, and successful transmission of MU i, respectively.

Proof: See Appendix A.4.

We note that the average access delay Δ_i consists of time delays due to carrier sensing, exponential back-offs, collided and erroneous (if any) attempts, successful transmission, and acknowledgement. It is also worthwhile to mention that the presence of interfering users may increase the collision probability, $p_{c,i}$, of MU i, thus increasing its average channel access delay (see Eq. (5.15) above).

Delay analysis of 4G LTE-A users

Next, we assume a 4G LTE-A cellular network and estimate its uplink delay. Let p_i^{tx} denote the transmission power of MU i. We use the Shannon-Hartley Theorem to estimate the uplink data

[3] Similar to [13], [75], [97], [98], [99], [100], and [101], the WiFi channel access time governed by the IEEE 802.11 DCF is assumed to be exponentially distributed. This is justified by the DCF channel access mechanism, which includes carrier sensing, binary exponential back-off(s), and reattempts (if any) due to collisions and erroneous transmissions.

rate r_i^{LTE} of MU i transmitting to BS k via 4G LTE-A cellular network as follows:

$$r_i^{\mathrm{LTE}} = \omega \log_2 \left(1 + \frac{p_i^{tx} G_{i,k}}{\bar{\omega}_0^2 + \sum_{j \neq i} p_j^{tx} G_{j,k}} \right), \tag{5.16}$$

where ω and $\bar{\omega}_0^2$ are the channel bandwidth and background noise power, respectively; $G_{i,k}$ denotes the channel gain between MU i and BS k. We use [13, Eq. (37)] to estimate the uplink delay of LTE-A users, which is given by

$$\Theta_i^{\mathrm{LTE}} = \frac{\rho_{BS}^u}{2 r_i^{\mathrm{LTE}} (1 - \rho_{BS}^u)} \left(\frac{\varsigma_L^2}{\bar{L}} + \bar{L} \right) + \frac{\bar{L}}{r_i^{\mathrm{LTE}}} + D_{RA}^{up} + D_{setup} + \tau_{BS}, \tag{5.17}$$

where D_{RA}^{up} is the initial random access delay (given by [13, Eq. (38)]), D_{setup} denotes the connection setup delay after passing the random access process successfully, ρ_{BS}^u denotes the uplink traffic intensity, τ_{BS} is the propagation delay in the cellular network, and $\bar{L}$ and ς_L^2 denote the mean and variance of the packet length, respectively. We note that, according to Eqs. (5.16)-(5.17), the achievable uplink data rate for MU i is decreased as a larger number of users is connected to the cellular BS, thereby increasing the packet delay experienced by MU i.

Each MU is directly associated with an ONU-AP or a cellular BS via a wireless single hop, whereby ONU-MPPs serve as ONU-APs. The WiFi connection and interconnection times of MUs are assumed to fit a truncated Pareto distribution, as validated via recent smartphone traces in [13]. The probability P_{temp}^{MU} that an MU is temporarily connected to an ONU-AP is estimated as $\bar{T}_{on}/(\bar{T}_{on} + \bar{T}_{off})$, whereby $\bar{T}_{on}$ and $\bar{T}_{off}$ denote the average WiFi connection and interconnection time, respectively. In this chapter, we assume that $\bar{T}_{on} = 28.1$ minute and $\bar{T}_{off} = 10.3$ minute, which are consistent with the measurements of PhoneLab traces (see [13] for further details). With these considerations, MU i is either connected to an ONU-AP through WiFi with probability P_{temp}^{MU} or an ONU-BS through cellular network with probability $(1 - P_{temp}^{\mathrm{MU}})$. The average task transmission delay Θ_i^{UL} in the uplink is then computed as follows:

$$\Theta_i^{\mathrm{UL}} = \left(P_{temp}^{\mathrm{MU}} \cdot \Theta_i^{\mathrm{WiFi}} + (1 - P_{temp}^{\mathrm{MU}}) \cdot \Theta_i^{\mathrm{LTE}} \right) \left\lceil \frac{B_i^l}{L_m} \right\rceil. \tag{5.18}$$

Delay analysis of backhaul EPON

Let D^u_{PON} denote the average packet delay in the backhaul EPON in the upstream direction. The average task transmission delay Θ^{PON} in the backhaul is then equal to $D^u_{\text{PON}} \cdot \left\lceil \frac{B^l_i}{L_m} \right\rceil$, where D^u_{PON} is given by [13]:

$$D^u_{\text{PON}} = \Phi\left(\rho^u, \bar{L}, \varsigma^2_L \ , c_{PON}\right) + \frac{\bar{L}}{c_{PON}} + 2\tau_{PON}\frac{2-\rho^u}{1-\rho^u} - B^u, \tag{5.19}$$

whereby ρ^u is the upstream traffic intensity, τ_{PON} is the propagation delay between ONUs and OLT, c_{PON} is the EPON data rate, $\Phi(\cdot)$ denotes the well-known Pollaczek-Khintchine formula, and B^u is obtained as $\Phi\left(\frac{\bar{L}}{\Lambda c_{PON}} \sum_{i=1}^{O} \sum_{q=1}^{O} \Gamma^{PON}_{iq}, \bar{L}, \varsigma^2_L \ , c_{PON}\right)$, where O is the number of ONUs and Γ^{PON}_{iq} is the traffic coming from ONU$_i$ to ONU$_q$, and Λ denotes the number of wavelengths in the WDM PON.

After calculating the computation and communication delay components, we proceed to compute the total average response time Υ_i of MU i as follows[4]:

$$\Upsilon_i = (1-\beta_i) \cdot \overbrace{\Delta_{\text{MU}_i}}^{\text{local response time } D^r_{L,i}} + \beta_i \cdot \overbrace{\left(\Theta^{\text{UL}}_i + (1-\alpha_j)\,\Delta_{\text{MEC}_j} + \alpha_j\left(\Theta^{\text{PON}} + \Delta_c\right)\right)}^{\text{non-local response time } D^r_{NL,i}}, \tag{5.20}$$

where the terms denoted by Δ and Θ represent the latency components of computation and communication, respectively. Note that the communication-induced latency terms Θ^{UL}_i and Θ^{PON} depend on the offloading probabilities β_i and α_j, respectively. More specifically, if MUs decide to offload a large portion of their incoming tasks to the edge servers, the average task transmission delay in the uplink as well as the waiting times in the edge server may increase significantly. On the other hand, if the edge servers also decide to further offload a large portion of their tasks arriving from MUs and fixed subscribers to the remote cloud, the backhaul upstream delay as well as waiting delay at the cloud servers may increase as a result. Therefore, in order for the MUs to benefit from the powerful computational capabilities of the edge/remote servers and experience a low response time, it is important for both device and edge-server schedulers to optimally adjust their offloading probabilities.

[4] Similar to [89], [96], [39], and [45] we neglect the time overhead for sending the computation result back to the mobile users due to the fact that for many applications (e.g., face/object recognition) the size of the computation result is generally smaller than that of the computation input data.

5.3.2 Average energy consumption per task

Similar to [90], we model the power consumption of MU i's CPU as κf_i^3, where κ is the effective switched capacitance related to the chip architecture [42]. The energy consumption per CPU cycle is thus equal to κf_i^2, as f_i represents the number of CPU cycles per second. The average energy consumption E_i^l for local execution of a task at MU i is then given by

$$E_i^l = \kappa \cdot f_i^2 \cdot D_i^l. \tag{5.21}$$

Recall from above that an incoming task at MU i is either executed locally with probability $(1 - \beta_i)$ or it is offloaded for nonlocal execution with probability β_i. The energy consumption, E_i^o, of MU i for offloading an incoming task is given by

$$E_i^o = E_i^{\text{UL}} + E_i^{\text{DL}}, \tag{5.22}$$

where E_i^{UL} and E_i^{DL} are the average energy consumptions of MU i to offload an incoming task in the uplink direction and receive its output in the downlink direction, respectively. In the uplink, E_i^{UL} is calculated as follows:

$$E_i^{\text{UL}} = \left(k_1^{tx} + k_2^{tx} \cdot p_i^{tx}\right) \cdot \Theta_i^{\text{UL}}, \tag{5.23}$$

whereby k_1^{tx} represents the static power consumption for having the radio frequency (RF) transmission circuitries switched on and k_2^{tx} measures the linear increase of the transmitter power consumption with radiated power p_i^{tx}. In the downlink, E_i^{DL} of MU i is estimated by

$$E_i^{\text{DL}} = \left(k_1^{rx} + k_2^{rx} \cdot r_i^{\text{DL}}\right) \cdot \Theta_i^{\text{DL}}, \tag{5.24}$$

where k_1^{rx} represents the extra power consumption for having the receiver circuit switched on, k_2^{rx} (measured in W/Mbps) is the power consumption per Mbps in the downlink direction, and r_i^{DL} is the downlink rate, which is given by

$$r_i^{\text{DL}} = P_{temp}^{\text{MU}} \cdot r^{\text{WiFi}} + (1 - P_{temp}^{\text{MU}}) \cdot r^{\text{LTE}}, \tag{5.25}$$

where r^{WiFi} and r^{LTE} are the average transmission rates of the WiFi access point and LTE BS, respectively. Further, we note that the transmission time, Θ_i^{DL}, of the task output in the downlink direction is estimated by $B_i^o / r_i^{\mathrm{DL}}$, where B_i^o is the task output size. Note that unlike Θ_i^{UL}, Θ_i^{DL} does not depend on the offloading probability β_i. The average energy consumption E_i (for either executing a task locally or transmitting its input data to an edge server) of MU i is then estimated as

$$E_i = (1 - \beta_i) E_i^l + \beta_i E_i^o. \tag{5.26}$$

By substituting Eqs. (5.21) and (5.22) into Eq. (5.26), we have

$$E_i = (1 - \beta_i) \left(\kappa \cdot f_i^2 \cdot D_i^l \right) + \beta_i [\left(k_1^{tx} + k_2^{tx} \cdot p_i^{tx} \right) \cdot \Theta_i^{\mathrm{UL}} + \left(k_1^{rx} + k_2^{rx} \cdot r_i^{\mathrm{DL}} \right) \cdot \Theta_i^{\mathrm{DL}}]. \tag{5.27}$$

5.4 Energy-delay Trade-off via Self-organization

According to our analysis above, an improved QoE is only achieved when an optimal setting of the offloading probabilities is done at both device and edge server sides. Any deviation from this optimal setting may result in performance deterioration. Due to the inherent time-varying nature of the network state, which is a direct consequence of user mobility and traffic fluctuation, such an optimal setting may not be obtained and maintained easily. To cope with this issue, we enable mobile users with self-awareness such that they can rely on their local computing resources, when needed.

In the following, we develop a bicriteria optimization framework to enable MUs to use their local information and minimize the response time as well as their energy consumption by dynamically adjusting their offloading probability as well as CPU clock frequency using the aforementioned DVS technique. For notational simplicity, we consider a tagged user and drop the subscript i hereafter. Similar to [42] and [45], we assume that the CPU clock frequency f of the tagged MU is restricted to a continuous interval of $[f_{min}, f_{max}]$. We formulate the bicriteria energy-delay self-organization problem as follows:

$$
\begin{aligned}
(\mathcal{P}_1): \quad & \min_{f,\beta} && \Upsilon(f,\beta), E(f,\beta) && \text{(5.28a)}\\
& \text{s.t.} && f_{min} \leq f \leq f_{max}, && \text{(5.28b)}\\
& && 0 \leq \beta \leq 1, && \text{(5.28c)}
\end{aligned}
$$

where Υ and E are given in Eqs. (5.20) and (5.27), respectively. To assess the developed model and characterize the tradeoff between the two objectives of the formulation above, we apply the Pareto front analysis. To obtain the Pareto front solutions, a common approach is to transform the original problem into an optimization problem by transferring one of the objectives into the constraints and solving it iteratively. In doing so, the problem $(\mathcal{P}_1)$ is transformed into a single-objective nonlinear optimization problem as follows:

$$
\begin{aligned}
(\mathcal{P}_2): \quad & \min_{f,\beta} && \Upsilon(f,\beta) && \text{(5.29a)}\\
& \text{s.t.} && f_{min} \leq f \leq f_{max}, && \text{(5.29b)}\\
& && 0 \leq \beta \leq 1, && \text{(5.29c)}\\
& && E(\beta, f) \leq \mathcal{E}_{thr}, && \text{(5.29d)}
\end{aligned}
$$

where $\mathcal{E}_{thr}$ denotes the given energy budget.

Lemma 5.2: *Problem (5.29) is a convex optimization problem.*

Proof. $\Upsilon(\beta, f)$ is a continuous twice-differentiable convex function of f and β, which can be verified by the fact that its Hessian matrix is positive definite. Besides, constraints (5.29b) and (5.29d) are affine functions of f and β, respectively; and constraint (5.29d) is a convex function of f and β. Therefore, the feasible set of the problem is a convex set. □

Lemma 5.3: *Necessary condition for optimality: The optimal solution (f^*, β^*) of problem $(\mathcal{P}_2)$ satisfies the following equation:*

$$
f^* = \mathcal{L}(\beta^*) = \min\{\mathcal{G}(\beta), f_{max}\}, \qquad \forall \beta \in [0, \beta_{max}], \tag{5.30}
$$

where

$$\mathcal{G}(\beta) = \sqrt{\frac{\mathcal{E}_{thr} - \beta[(k_1^{tx} + k_2^{tx} p_t)\Theta^{\mathrm{UL}} + (k_1^{rx} + k_2^{rx} r^{\mathrm{DL}})\,\Theta^{\mathrm{DL}}]}{(1-\beta)\kappa D^l}}, \tag{5.31}$$

and β_{max} *is obtained by solving the following equation*:

$$\mathcal{G}(\beta) - f_{min} = 0. \tag{5.32}$$

Proof. First, we show that $f = \mathcal{L}(\beta)$ (given by Eq. (5.30)) determines the upper limit of CPU clock frequency f for a given β. In doing so, we take the energy constraint given by Constraint (5.29d) and calculate f as a function of β for a fixed $\mathcal{E}_{thr}$ as follows:

$$f \leq \mathcal{G}(\beta). \tag{5.33}$$

Taking into account constraint (5.29b), inequality (5.33) becomes:

$$f \leq \overbrace{\min\{\mathcal{G}(\beta), f_{max}\}}^{\mathcal{L}(\beta)}. \tag{5.34}$$

Clearly, $f = \mathcal{L}(\beta)$ gives the upper limit of f for a given β (see Fig. 5.3).

Next, we prove Lemma 5.3 by contradiction. Let (β_1, f_1) satisfy $f_1 = \mathcal{L}(\beta_1)$. Assume (β_1, f_2), $\forall f_2 < f_1$, achieves a smaller response time, thus:

$$\Upsilon(\beta_1, f_2) < \Upsilon(\beta_1, f_1). \tag{5.35}$$

Obviously, since $f = \mathcal{L}(\beta)$ gives the upper limit of the CPU clock frequency f for a given (fixed) offloading probability β, an MU can experience a smaller response time by increasing its CPU clock frequency from f_2 to f_1. This happens in light of the fact that for a fixed offloading probability β_1, as CPU clock frequency f increases, the first term on the right-hand side of Eq. (5.20) decreases, whereas the second term remains unchanged. Hence, $\Upsilon(\beta_1, f_1) < \Upsilon(\beta_1, f_2)$, which is in contradiction with (5.35). We also note that for a given energy budget $\mathcal{E}_{thr}$, the maximum offloading probability β_{max} is obtained by solving $f_{min} = \mathcal{G}(\beta)$, as illustrated in Fig. 5.3. Therefore, the optimal solution (f^*, β^*) of problem $(\mathcal{P}_2)$ satisfies Eq. (5.30). This completes the proof. □

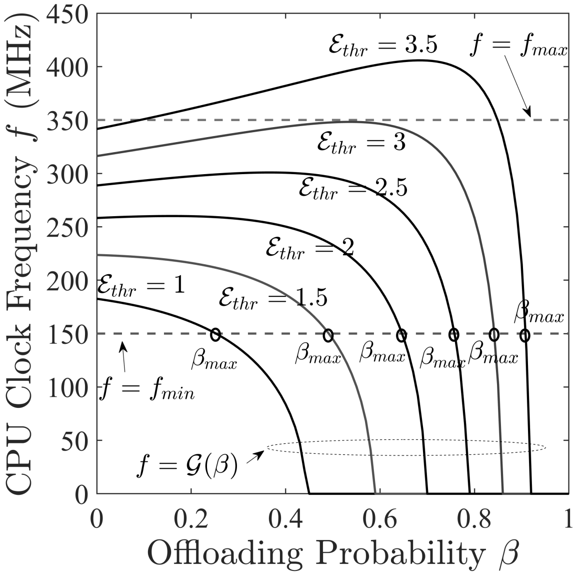


Figure 5.3 – Illustration of the search space for problem $(\mathcal{P}_2)$ for different values of $\mathcal{E}_{thr}$.

Algorithm 6 Joint Offloading and DVS Procedure ()

```
Input: $\mathcal{E}_{thr}, f_{min}, f_{max}$, energy and task parameters
Output: $\beta^*$ and $f^*$
Initialize: $\beta_0 \leftarrow 0$ and $f_0 \leftarrow \mathcal{L}(0)$
Solve Eq. (5.32) and obtain $\beta_{max}$
$\Delta \leftarrow M$
while $(\Delta > \epsilon_1)$ & $(\beta_0 \leq \beta_{max})$ do
    $\beta_1 \leftarrow \beta_0 + \epsilon_2$
    $f_1 \leftarrow \mathcal{L}(\beta_1)$ using Eq. (5.30)
    $\Delta \leftarrow \Upsilon(f_0, \beta_0) - \Upsilon(f_1, \beta_1)$ using Eq. (5.20)
    if $\Delta > 0$ then
        $\beta^* \leftarrow \beta_1$ and $f^* \leftarrow f_1$
    else
        $\beta^* \leftarrow \beta_0$ and $f^* \leftarrow f_0$
    end if
    $\beta_0 \leftarrow \beta_1$
end while
return $\beta^*, f^*$
```

Since *Lemma 2* and *Lemma 3* hold, we use a standard constrained convex optimization approach and obtain the optimal solution (f^*, β^*) of problem $(\mathcal{P}_2)$ for a given MU using only its local information/parameters, as described in Algorithm 6, where ϵ_2 is the minimum step size for searching the optimal solution (β^*, f^*), M is a big number, and ϵ_1 is a small number. In Algorithm 6, the optimal offloading probability β^* is obtained via a one-dimensional search method, whereas the optimal CPU clock frequency f^* is calculated accordingly using the closed-form relation given by Eq. (5.30). It is worthwhile to mention that the non-local processing latency term, D^r_{NL}, in

Table 5.1 – MEC-enabled FiWi enhanced HetNet Parameters & Default Values

Parameter	Value	Parameter	Value
Traffic Model Parameters			
L_m	1500 Bytes	λ_B	30 packet/s
α_{PON}	100	$\bar{L}, \varsigma_L^2$	1500 Bytes, 0
Backhaul EPON			
l_{PON}	20 km	c_{PON}	10 Gbps
N_{ONU}	$\{12, 16, 20, 24\}$	Λ	1
WiFi Parameters			
DIFS	34 μsec	SIFS	16 μsec
PHY Header	20 μsec	W_0, H	16 slots, 6
ϵ	9 μsec	RTS	20 bytes
CTS	14 bytes	ACK	14 bytes
r in WMN	300 Mbps	ONU-AP radius	15 m
LTE-A Parameters			
p^{tx}	100 mW	ω	5 MHz
$\bar{\varpi}_0^2$	-100 dBm	k_1^{tx}	0.4 W
k_2^{tx}	18	ONU-BS radius	50 m
p^{rx}	200 mW	k_1^{rx}	0.4 W
k_2^{rx}	2.86 W/Mbps		
Task and Edge/Cloud Server Parameters			
λ_{MU}	25 task/min	f_i	$[150, 450]$ MHz
λ_0^e	30 task/min	f_j^e	1.44 GHz
λ_0^c	240 task/min	s	6
f^c	1.44 GHz	B^l	66 KB
D^l, D_0^e, D_0^c	300 Mcycles	κ	10^{-26}
ϵ_1, ϵ_2	10^{-3}	M	10^2
B^o	1 KB		

Eq. (5.20) can be calculated using only local information as follows:

$$D_{NL}^r = \frac{\hat{\Upsilon} - (1-\beta) \cdot D_L^r - \beta \cdot \hat{\Theta}^{\text{UL}}}{\beta}, \tag{5.36}$$

where $\hat{\Upsilon}$ and $\hat{\Theta}^{\text{UL}}$ can be obtained via measurements. Given that the number of arithmetic operations within each iteration is upper-bounded, the complexity of Algorithm 6 is $\mathcal{O}(K)$, where $K = \frac{\beta_{max}}{\epsilon_2}$ is the number of iterations for searching the optimal solution.

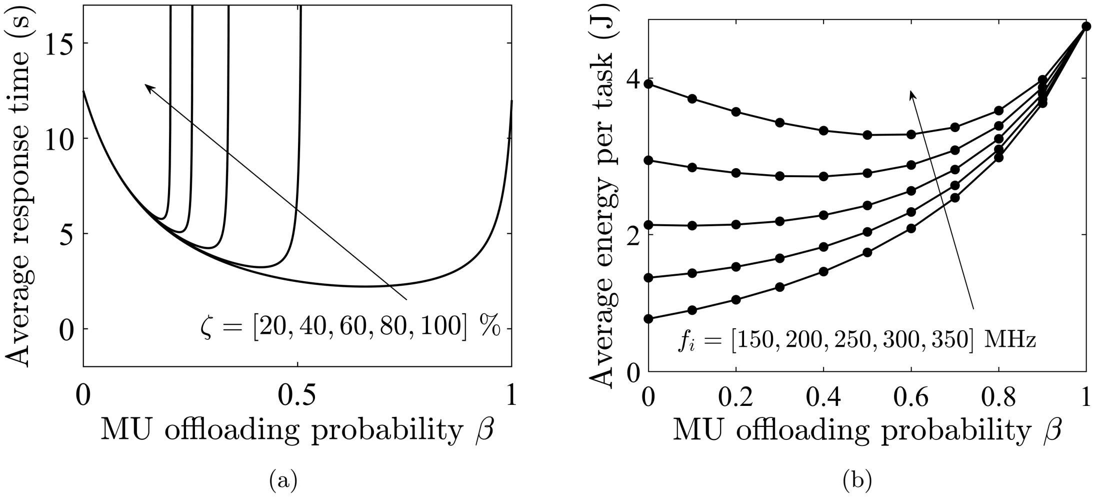


Figure 5.4 – (a) Average response time vs. MU offloading probability β for different values of ζ ($\alpha = 0$ and f_i=150 MHz); (b) average energy per task vs. MU offloading probability β for different values of local clock frequency f_i ($\zeta = 20\%$).

5.5 Results

The following numerical results were obtained by using the LTE-A and FiWi network and traffic parameter settings listed in Table 5.1, which are consistent with those in [13], [75], [42], [45], [44], and [102]. In our considered scenario, 50 MUs are scattered randomly within the range of 50 m from each ONU-BS. Besides, we consider four MUs within the coverage area of each ONU-AP. In the cellular access mode, we set the channel gain to $G_{i,k} = d_{i,k}^{-\xi}$ between MU i and BS k, where $d_{i,k}$ is the distance between MU i and BS k, and $\xi = 4$ is the path loss factor. Further, we set $\beta_i = \beta$ ($\forall i = 1, 2, ...$) and $\alpha_j = \alpha$ ($\forall j = 1, 2, ...$). A portion ζ of the number of MUs that reside within the coverage area of an ONU-AP or cellular BS is involved in task offloading, while the remaining portion $(1 - \zeta)$ generates the conventional Poisson H2H traffic at mean packet rate λ_B. Background traffic coming from ONUs with attached fixed (wired) subscribers is set to $\alpha_{PON} \cdot \lambda_B$, where $\alpha_{PON} \geq 1$ is a traffic scaling factor for fixed subscribers that are directly connected to the backhaul EPON. Also, the user mobility parameters in our simulations are tuned such that the WiFi connection and interconnection times fit a truncated Pareto distribution with $P_{temp}^{\mathrm{MU}} = 73.18\%$, which is compliant with the measurements in [13].

First, we consider the edge computing-only scenario with α being set to zero. Figures 5.4a and 5.4b depict the energy-delay performance of MEC-assisted partial offloading. The average response time vs. offloading probability β for different values of ζ is shown in Fig. 5.4a. The results indicate

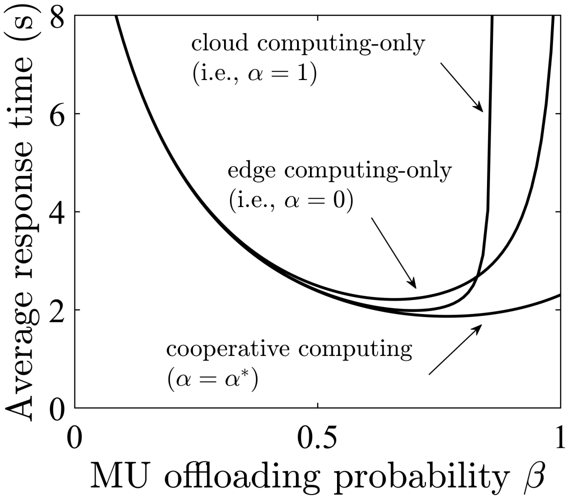


Figure 5.5 – Comparison of average response time performance of edge-only, cloud-only, and cooperative computing ($\zeta = 20\%$).

that the average response time is a convex function of β. For $\zeta = 20\%$, setting $\beta = 0.66$ leads to an 82% reduction of the average response time compared to the fully local computing scheme (i.e., $\beta = 0$). We note that the optimal value of β largely depends on ζ. More specifically, as ζ increases, the optimal value of β decreases, as shown in Fig. 5.4a. Figure 5.4b depicts the average energy consumption per task vs. β for different values of f_i. The bottom curves in Figs. 5.4a and 5.4b highlight the trade-off an MU can make between the average response time and energy per task for $\zeta = 20\%$. We also observe from Fig. 5.4b that for larger values of f_i, partial offloading not only reduces the average response time but it also helps MUs reduce their energy consumption.

Next, we examine the performance gains obtained from our proposed trilateral device-edge-cloud cooperative computing. Figure 5.5 depicts the average response time vs. β for the following three different scenarios: (i) edge-only ($\alpha = 0$), (ii) cloud-only ($\alpha = 1$), and (iii) cooperative computing ($\alpha = \alpha^*$), where α^* denotes the optimal value of α set by the network operator to minimize the average task execution time experienced by the edge servers. Figure 5.5 shows that the proposed cooperative computing scheme yields a better delay performance compared to either the edge-only or cloud-only scheme, especially for $\beta > 0.64$. While the edge- and cloud-only schemes may both pose a longer response time due to an excessive queueing delay for large values of β, the trilateral cooperation between the CPU, edge server, and remote cloud yields a reduced response time by setting β and α to their optimal values (see bottom curve in Fig. 5.5).

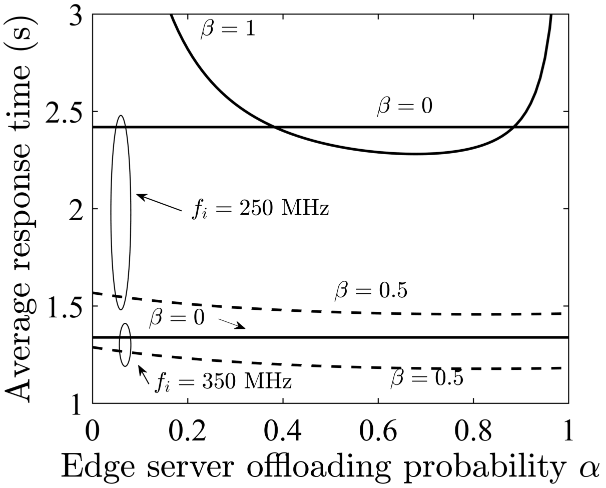


Figure 5.6 – Average response time vs. edge-server offloading probability α for different values of β and f_i ($\zeta = 20\%$).

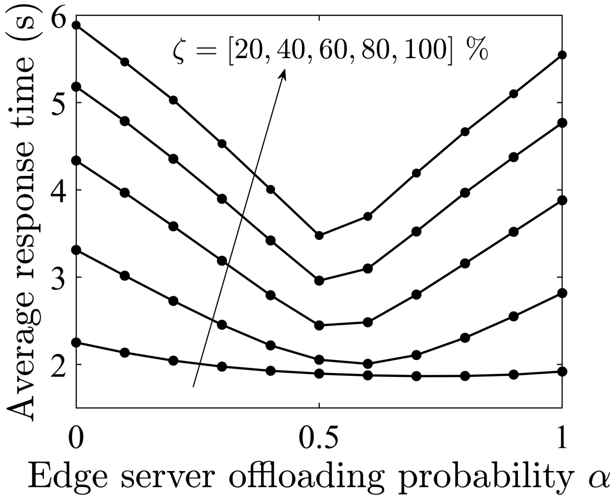


Figure 5.7 – Average response time vs. edge-server offloading probability α for different values of ζ ($N_{ONU} = 12$ and $\beta = \beta^*$).

Figure 5.6 shows the impact of edge server offloading probability α on the delay performance of our cooperative computing scheme. For $f_i = 250$ MHz, when MUs operate in the full offloading mode (i.e., $\beta = 1$), the delay of the cooperative computing scheme equals 2.28 s by setting $\alpha = 0.68$, compared to ~ 12 s of the edge- or cloud-only schemes, which translates into an 81% reduction of the average response time. More interestingly, the reduction is achieved by appropriately setting the edge server offloading probability α and without incurring any additional energy per task, which remains unchanged for a given β. We note that the average response time as well as the energy

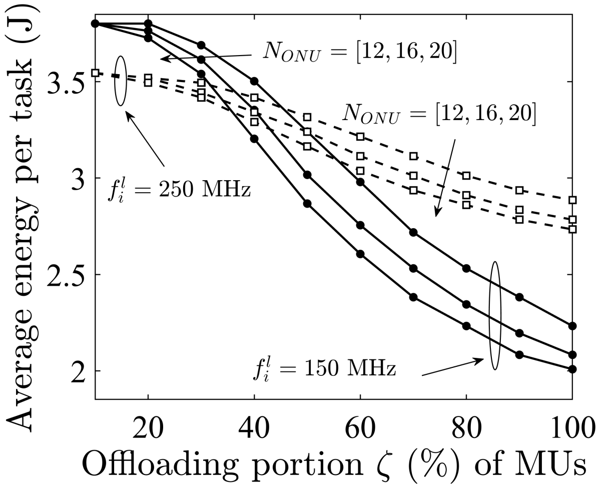


Figure 5.8 – Energy per task vs. ζ for different values of local clock frequency f_i and number N_{ONU} of ONUs ($\alpha = \alpha^*$ and $\beta = \beta^*$).

per task can be further reduced by setting $\beta = 0.5$. Moreover, for $f_i = 350$ MHz, the minimum response time is achieved by setting $\beta = 0.5$ and $\alpha = 0.86$. In doing so, setting $\beta = 0.5$ also yields a near-optimal energy performance according to Fig. 5.4b. We proceed by discussing the results of the average response time vs. α for different values of ζ in Fig. 5.7. We observe that the beneficial impact of edge-cloud cooperation through the backhaul on the average response time becomes even more pronounced for larger values of ζ. Specifically, we observe that the average response time decreases from 5.89 s for $\alpha = 0$ to 3.48 s for $\alpha = \alpha^* = 0.5$ and $\zeta = 100\%$, as opposed to only a slight decrease from 2.25 s for $\alpha = 0$ to 1.87 s for $\alpha = \alpha^* = 0.7$ and $\zeta = 20\%$.

Next, we examine the energy performance of our proposed cooperative offloading scheme in Fig. 5.8, where α and β are set to their optimal values of α^* and β^*. For a given f_i and an increasing ζ or N_{ONU}, we observe a generally decreasing trend in the energy consumption. This occurs because for a larger ζ or N_{ONU}, the optimal delay performance is achieved by relying more on local rather than nonlocal computing resources by setting β to smaller values, which in turn results in a reduced energy consumption (see Fig. 5.4b). Importantly, we also observe that while increasing the local clock frequency f_i always leads to a decreased average response time (see Fig. 5.6), it may not necessarily increase the energy consumption. For instance, we observe that the energy consumption for $f_i = 250$ MHz, unexpectedly, is lower than that of $f_i = 150$ MHz, provided that $\zeta < 45\%$. This is because the delay-optimal setting β^* in the former is smaller than that of the latter,

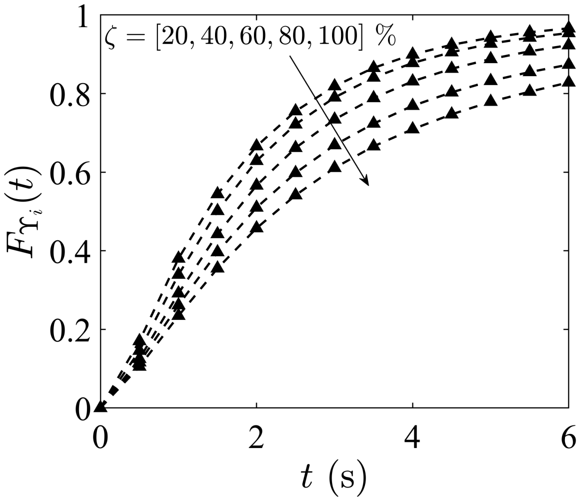


Figure 5.9 – CDF of response time for different values of ζ ($\alpha = \alpha^*$ and $\beta = \beta^*$).

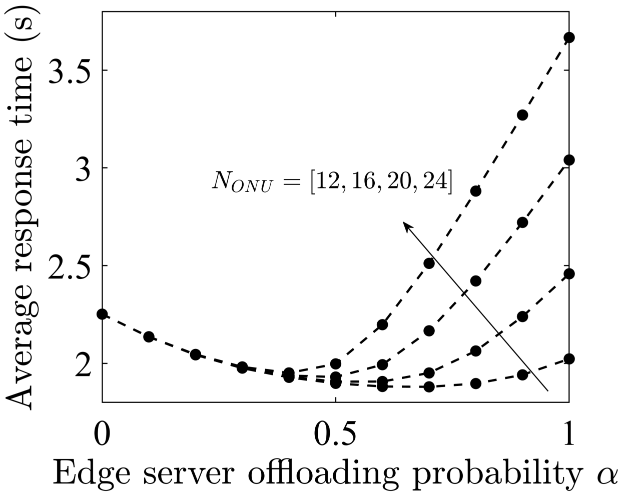


Figure 5.10 – Average response time vs. edge server offloading probability α for different values of N_{ONU} ($\beta = \beta^*$, $\zeta = 20\%$).

thus revealing a better energy performance (see also Fig. 5.4b). Note, however, that for $\zeta > 45\%$, setting $f_i = 150$ MHz can achieve an energy saving of up to 38% compared to $f_i = 250$ MHz.

Next, we present the cumulative distribution function (CDF), $F_{\Upsilon_i}(t)$, of the response time Υ_i (i.e., $\Pr(\Upsilon_i \leq t)$) in Fig. 5.9 for different values of ζ. We find that our proposed cooperative computing scheme ensures a lower bound probability of 80% that an incoming task is executed (either locally or nonlocally) and returned to the MU within 5.5 s. The average response time vs. α for different values of N_{ONU} is depicted in Fig. 5.10, where we evaluate the delay performance of

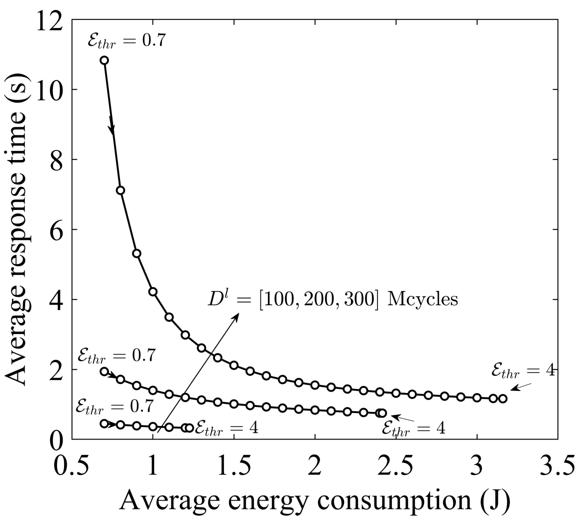
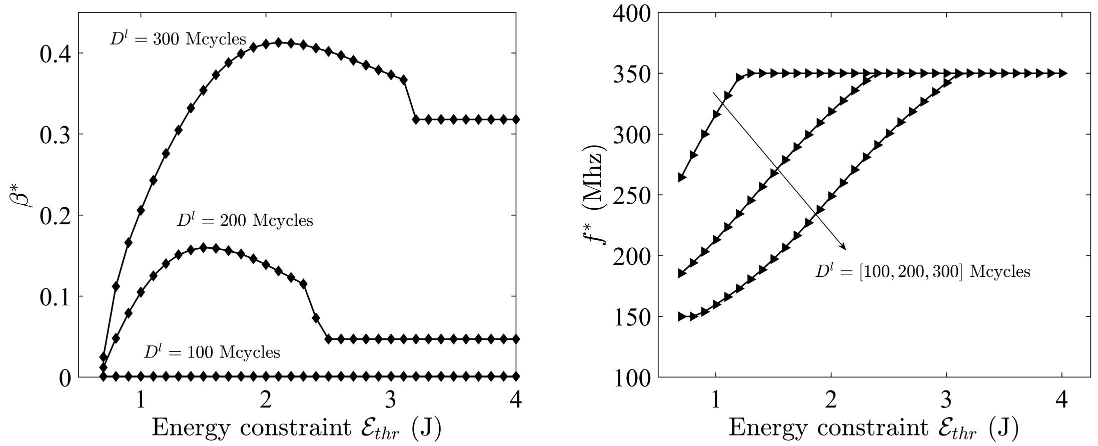


Figure 5.11 – Pareto front solutions of self-organization problem ($\mathcal{P}_2$) for $D^l \in [100, 200, 300]$ Mcycles (the value of energy constraint $\mathcal{E}_{thr}$ increases along the arrow shown on each curve).



(a) (b)

Figure 5.12 – (a) Optimal offloading probability β^* vs. energy constraint $\mathcal{E}_{thr}$ and (b) optimal CPU clock frequency f^* vs. energy constraint $\mathcal{E}_{thr}$ for different values of $D^l = [100, 200, 300]$ Mcycles.

our proposed edge-cloud cooperation through the backhaul. Interestingly, we find that by doubling the number N_{ONU} of ONUs from 12 to 24, the average response time of MUs only increases from 1.88 s to 1.95 s, provided that an optimal setting of both α and β is carried out.

Finally, Fig. 5.11 illustrates the Pareto frontier characterization of the energy-delay trade-off an MU can make via dynamic reconfiguration by using our proposed self-organization scheme for different values of D^l. For a given energy constraint $\mathcal{E}_{thr}$, which is determined by the decision maker according to a given energy budget of the battery as well as delay requirement of incoming tasks, an MU can optimally adjust its offloading probability and CPU clock frequency using only its local information to achieve the desired energy-delay performance. For instance, Fig. 5.11 shows that for $D^l = 300$ Mcycles, by increasing $\mathcal{E}_{thr}$ from 0.7 J to 1.8 J a 84% reduction of the average response time is achieved. We also observe from Fig. 5.11 that any further increase of the energy budget may not lead to a significant reduction of the average response time, especially for $D^l = 100$ and 200 Mcycles. Moreover, the results of the optimal offloading probability β^* and CPU clock frequency f^* vs. the energy constraint $\mathcal{E}_{thr}$ for different values of D^l are shown in Fig. 5.12. Interestingly, Figs. 5.12a and 5.12b, along with Fig. 5.11, illustrate the impact of increasing D^l and $\mathcal{E}_{thr}$ on the optimal decision made by the MU. For instance, we observe that for $D^l = 100$ Mcycles, offloading doesn't have any benefit in terms of the average response time, thus $\beta^* = 0$, $\forall \mathcal{E}_{thr} \in [0.7, 3]$ J (see Fig. 5.12a). Instead, MUs can reduce their response time by increasing f (see Fig. 5.12b).

5.6 Conclusions

This chapter studied the cooperative computation offloading in MEC enabled FiWi enhanced HetNets from both network architecture and offlading mechanism design perspectives. Beside the design of reliable low-latency MEC enabled FiWi enhanced LTE-A HetNets, we presented a simple but efficient offloading strategy that leverages trilateral cooperation among device, edge server, and remote cloud. We developed an analytical framework to estimate the average response time and energy consumption of mobile users in a FiWi based MEC enabled network infrastructure. Our results demonstrate the superior performance of the proposed cooperative computing scheme compared to edge- or cloud-only schemes. Further, we showed that by optimally setting the offloading probabilities, MUs can achieve a reduction of the average response time of up to 81%. In order to cope with the incurred complexity, we also designed a self-organization based mechanism, which enables an MU, using local information, to make suitable energy-delay trade-offs and jointly minimize the average execution time and energy consumption by dynamically adjusting the offloading probability and its local CPU clock frequency using the DVS technique.

Chapter 6

Conclusions

This chapter summarizes the important contributions of the dissertation and outlines possible directions for future work.

6.1 Summary

The Internet has constantly evolved from the mobile Internet dominated by H2H traffic to the emerging IoT with its underlying M2M communications. The advent of advanced robotics, along with the emerging ultra responsive networking infrastructures, will allow for transmitting the modality of touch (also known as haptic sensation) in addition to the traditional triple-play traffic (i.e., voice, video, and data) under the commonly known term Tactile Internet. The IoT without any human involvement in its underlying M2M communications is useful for the automation of industrial and other machine-centric processes while keeping the human largely out of the loop. In contrast, the Tactile Internet allows for the tactile steering and control of not only virtual but also real objects via teleoperated robots, and will be centered around H2R/M communications, thus calling for a human-centric design approach.

This doctoral thesis is built on FiWi enhanced LTE-A HetNets as promising underlying networking infrastructures on which the emerging Tactile Internet is envisioned to rely. In particular, we studied different aspects of the emerging Tactile Internet and presented in-depth technical insights into realizing HITL-centric teleoperation Tactile Internet over FiWi enhanced networks, including trace-based haptic traffic modeling, perceptual deadband coding, haptic sample forecasting, task coordination via context- and self-awareness, teleoperation task allocation, and cooperative computation offloading. In the following, a more detailed summary of each chapter is presented.

We have seen that there is a significant overlap among 5G, IoT, and the Tactile Internet in that they share various important design goals, including very low latency, ultra-high reliability, and integration of data-centric technologies. In Chapter 2, we described how FiWi enhanced LTE-A HetNets leveraging low-cost data-centric EPON and WiFi technologies for fiber backhaul sharing and WiFi offloading may help realize not only the aforementioned shared design goals but also the key attributes of end-to-end co-DBA of both PON and wireless network resources, decentralization, and edge intelligence in support of future low-latency applications over a common optical transport platform.

Our focus was on the emerging Tactile Internet as one of the most interesting low-latency applications for creating novel immersive experiences. We reviewed the HITL-centric design principles that add a new dimension to the human-to-machine interaction via the Internet and set the Tactile Internet aside from the more machine-centric IoT. Exploiting the human perception of haptics to reduce the haptic packet rate by means of deadband coding, we derived haptic traffic models from teleoperation experiments. Our haptic trace analysis showed that assuming Tactile Internet traffic to be Pareto distributed was not valid for the analyzed traffic, while assuming it to be Poisson traffic was valid only in a special case. In general, we observed that command and feedback paths of teleoperation systems can be jointly modeled by generalized Pareto, gamma, or deterministic packet interarrival time distributions, depending on the given value of the respective deadband parameters. We then used machine learning to implement an MLP-based multi-sample-ahead-of-time sample forecaster, which is capable of decoupling haptic feedback from the impact of extensive propagation delays. This enables humans to perceive remote task environments in time at a 1-ms granularity.

In Chapter 3, we investigated the performance of our proposed context- and self-aware HART centric multi-robot task allocation over FiWi based Tactile Internet infrastructures. We shed light on when, how, and under which circumstances user-ownership of MRs becomes beneficial in terms of OPEX per executed task. Further, we evaluated the performance of our proposed CADMRTC algorithm in terms of average task completion time, OPEX per executed task, and ratio of the number of executed tasks by user-owned MRs and the total number of tasks. By leveraging on the low-latency and reliable fiber backhaul and distributed WiFi-based fronthaul, we showed that a human-robot connectivity probability of $> 90\%$ is achievable for $\bar{T}_{on}^{MR}/\bar{T}_{off}^{MR} > 10$. In addition, our obtained results show that our proposed self-aware scheme plays a key role in minimizing the traverse time as well as energy consumption of MRs in a distributed manner, whereas our context-aware task coordination is instrumental in minimizing the task completion time, while paying particular attention to reducing OPEX of user-/network-ownership of MRs.

In Chapter 4, we investigated the performance of our proposed CAPSTA algorithm in solving the prioritized assignment and scheduling of delay-constrained teleoperation tasks in FiWi enhanced Tactile Internet network infrastructures. The obtained results show that the proposed algorithm reduces the average weighted task completion time, maximum tardiness, and average OPEX, compared to the benchmark RSA algorithm. Specifically, the proposed CAPSTA algorithm achieves a 15-27% reduction of average weighted task completion time and a 49-56% reduction of maximum tardiness. In addition, compared to the benchmark RAS algorithm, the proposed CAPSTA algorithm achieves a 75.3% and 78.9% reduction of average OPEX per task for $N = 100$ and $N = 300$, respectively. Our results also give insights into finding the optimal number of HOs to minimize the average OPEX per completed task for different deployment scenarios. More precisely, we have shown that for the proposed CAPSTA algorithm, the optimal number of available human operators $M^{\star}$ that minimizes OPEX is 2 and 5 for $N = 100$ and $N = 300$, respectively. Finally, we have shown that the considered solution is able to achieve an average end-to-end packet delay of < 10 ms for both local and non-local teleoperation for a wide range of background traffic rates.

In Chapter 5, we studied the cooperative computation offloading in MEC enabled FiWi enhanced HetNets from both network architecture and offlading mechanism design perspectives. Beside the design of reliable low-latency MEC enabled FiWi enhanced LTE-A HetNets, we presented a simple but efficient offloading strategy that leverages trilateral cooperation among device, edge server, and remote cloud. We developed an analytical framework to estimate the average response time and energy consumption of mobile users in a FiWi based MEC enabled network infrastructure. Our results demonstrate the superior performance of the proposed cooperative computing scheme compared to edge- or cloud-only schemes. Further, we showed that by optimally setting the offloading probabilities, MUs can achieve a reduction of the average response time of up to 81%. In order to cope with the incurred complexity, we also designed a self-organization based mechanism, which enables an MU, using local information, to make suitable energy-delay trade-offs and jointly minimize the average execution time and energy consumption by dynamically adjusting the offloading probability and its local CPU clock frequency using the DVS technique.

6.2 Future Work

In Chapter 2, we elaborated on the importance of the decentralized nature of WLAN's access protocol DCF to realize low-latency FiWi enhanced LTEA HetNets. Furthermore, by exploiting their inherent distributed processing and storage capabilities, we investigated the potential of enabling immersive teleoperation experiences for human operators by introducing machine learning at the optical-wireless interface of FiWi enhanced LTE-A HetNets. Our proposed MLP-based ESF module compensates for delayed haptic feedback samples by means of multiple-sample-ahead-of-time forecasting for a tighter togetherness, improved safety control, and increased reliability. Future work will investigate the applicability of this technique in networks with arbitrary delays using sophisticated deep learning models.

Our obtained results in Chapter 3 show that from a performance perspective (in terms of average task completion time) almost no deterioration occurs if the ownership is shifted entirely from network operators to mobile end-users, though such a shift in ownership of robots has significant implications on sharing the profits and collaborative business opportunities arising from the emerging Tactile Internet in a more equitable fashion. As a result, this may open up new opportunities for synergies between humans and machines/robots, while spurring the symbiotic human-machine/robot development envisaged by early-day Internet pioneers and imagining entirely new categories of abundance for a low entry cost economy. Among others, one future research direction is to further explore the synergies between the aforementioned HART membership and the complementary strengths of robots to facilitate local human-machine coactivity clusters by decentralizing the Tactile Internet. Another interesting open research problem is how human *crowdsourcing* can help decrease task completion time in the event of unreliable connectivity and/or network failures. Note that our presented spreading ownership of robots across mobile users may be an important stepping stone to collaborative business relationships that function more like localized share-economy ecosystems than markets.

Last but not least, an interesting future research avenue is to investigate the role of virtualization in FiWi networks to eliminate the physical layer interaction of the often heterogenous Tactile Internet applications, thus realizing an infrastructure/technology independent architecture.

Bibliography

[1] G. P. Fettweis, "The Tactile Internet: Applications and Challenges," *IEEE Vehicular Technology Magazine*, vol. 9, no. 1, pp. 64–70, Mar. 2014.

[2] M. Simsek, A. Aijaz, M. Dohler, J. Sachs, and G. Fettweis, "5G-Enabled Tactile Internet," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 3, pp. 460–473, Mar. 2016.

[3] M. Maier, M. Chowdhury, B. P. Rimal, and D. P. Van, "The Tactile Internet: Vision, Recent Progress, and Open Challenges," *IEEE Communications Magazine*, vol. 54, no. 5, pp. 138–145, May 2016.

[4] M. Dohler, T. Mahmoodi, M. A. Lema, M. Condoluci, F. Sardis, K. Antonakoglou, and H. Aghvami, "Internet of Skills, Where Robotics Meets AI, 5G and the Tactile Internet," in *Proc. European Conference on Networks and Communications (EuCNC)*, Jun. 2017, pp. 1–5.

[5] M. Dohler, "The Future and Challenges of Communications–Toward a World Where 5G Enables Synchronized Reality and an Internet of Skills," *Internet Technology Letters*, vol. 1, no. 2, pp. 1–3, Apr. 2018.

[6] M. A. Lema, A. Laya, T. Mahmoodi, M. Cuevas, J. Sachs, J. Markendahl, and M. Dohler, "Business Case and Technology Analysis for 5G Low Latency Applications," *IEEE Access*, vol. 5, pp. 5917–5935, 2017.

[7] *ITU-T Technology Watch Report*, "The Tactile Internet," Aug. 2014.

[8] P. R. Daugherty and H. J. Wilson, *Human + Machine: Reimagining Work in the Age of AI*. Harvard Business Review Press, 2018.

[9] "Audi and VGo," 2014, [Online]. Available: http://www.vgocom.com/audi.

[10] K. Antonakoglou, X. Xu, E. Steinbach, T. Mahmoodi, and M. Dohler, "Towards Haptic Communications over the 5G Tactile Internet," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 4, pp. 3034–3059, Fourth quarter 2018.

[11] E. Steinbach, S. Hirche, M. Ernst, F. Brandi, R. Chaudhari, J. Kammerl, and I. Vittorias, "Haptic Communications," *Proceedings of the IEEE*, vol. 100, no. 4, pp. 937–956, Apr. 2012.

[12] P. Schulz, M. Matthe, H. Klessig, M. Simsek, G. Fettweis, J. Ansari, S. A. Ashraf, B. Almeroth, J. Voigt, I. Riedel, A. Puschmann, A. Mitschele-Thiel, M. Muller, T. Elste, and M. Windisch, "Latency Critical IoT Applications in 5G: Perspective on the Design of Radio Interface and Network Architecture," *IEEE Communications Magazine*, vol. 55, no. 2, pp. 70–78, Feb. 2017.

[13] H. Beyranvand, M. Lévesque, M. Maier, J. A. Salehi, C. Verikoukis, and D. Tipper, "Toward 5G: FiWi Enhanced LTE-A HetNets With Reliable Low-Latency Fiber Backhaul Sharing and WiFi Offloading," *IEEE/ACM Transactions on Networking*, vol. 25, no. 2, pp. 690–707, Apr. 2017.

[14] T. Taleb, K. Samdanis, B. Mada, H. Flinck, S. Dutta, and D. Sabella, "On Multi-Access Edge Computing: A Survey of the Emerging 5G Network Edge Cloud Architecture and Orchestration," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 3, pp. 1657–1681, Third quarter 2017.

[15] J. Sachs, L. A. A. Andersson, J. Araújo, C. Curescu, J. Lundsjö, G. Rune, E. Steinbach, and G. Wikström, "Adaptive 5G Low-Latency Communication for Tactile Internet Services," *Proceedings of the IEEE*, vol. 107, no. 2, pp. 325–349, Feb. 2018.

[16] N. Gholipoor, H. Saeedi, and N. Mokari, "Cross-layer Resource Allocation for Mixed Tactile Internet and Traditional Data in SCMA based Wireless Networks," in *Proc. IEEE Wireless Communications and Networking Conference Workshops (WCNCW)*, Apr. 2018, pp. 356–361.

[17] C. She, C. Yang, and T. Q. S. Quek, "Cross-Layer Transmission Design for Tactile Internet," in *Proc. IEEE GLOBECOM*, Dec. 2016, pp. 1–6.

[18] A. Garcia-Saavedra, M. Karzand, and D. J. Leith, "Low Delay Random Linear Coding and Scheduling Over Multiple Interfaces," *IEEE Transactions on Mobile Computing*, vol. 16, no. 11, pp. 3100–3114, Nov. 2017.

[19] C. Li, C. Li, K. Hosseini, S. B. Lee, J. Jiang, W. Chen, G. Horn, T. Ji, J. E. Smee, and J. Li, "5G-Based Systems Design for Tactile Internet," *Proceedings of the IEEE*, vol. 107, no. 2, pp. 307–324, Feb. 2019.

[20] A. S. Shafigh, S. Glisic, and B. Lorenzo, "Dynamic Network Slicing for Flexible Radio Access in Tactile Internet," in *Proc. IEEE GLOBECOM*, Dec. 2017, pp. 1–7.

[21] Z. Xiang, F. Gabriel, E. Urbano, G. T. Nguyen, M. Reisslein, and F. H. P. Fitzek, "Reducing Latency in Virtual Machines: Enabling Tactile Internet for Human-Machine Co-Working," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 5, pp. 1098–1116, May 2019.

[22] V. Petrov, M. A. Lema, M. Gapeyenko, K. Antonakoglou, D. Moltchanov, F. Sardis, A. Samuylov, S. Andreev, Y. Koucheryavy, and M. Dohler, "Achieving End-to-End Reliability of Mission-Critical Traffic in Softwarized 5G Networks," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 3, pp. 485–501, March 2018.

[23] E. Wong, M. P. I. Dias, and L. Ruan, "Predictive Resource Allocation for Tactile Internet Capable Passive Optical LANs," *IEEE/OSA Journal of Lightwave Technology*, vol. 35, no. 13, pp. 2629–2641, July 2017.

[24] L. Ruan, M. P. I. Dias, and E. Wong, "Towards Tactile Internet Capable E-Health: A Delay Performance Study of Downlink-Dominated SmartBANs," in *Proc. IEEE GLOBECOM*, Dec. 2017, pp. 1–6.

[25] Y. Feng, C. Jayasundara, A. Nirmalathas, and E. Wong, "Hybrid Coordination Function Controlled Channel Access for Latency-Sensitive Tactile Applications," in *Proc. IEEE GLOBECOM*, Dec. 2017, pp. 1–6.

[26] F. Dressler, F. Klingler, M. Segata, and R. L. Cigno, "Cooperative Driving and the Tactile Internet," *Proceedings of the IEEE*, vol. 107, no. 2, pp. 436–446, Feb. 2019.

[27] B. Kehoe, S. Patil, P. Abbeel, and K. Goldberg, "A Survey of Research on Cloud Robotics and Automation," *IEEE Transactions on Automation Science and Engineering*, vol. 12, no. 2, pp. 398–409, Apr. 2015.

[28] W. Chen, Y. Yaguchi, K. Naruse, Y. Watanobe, K. Nakamura, and J. Ogawa, "A Study of Robotic Cooperation in Cloud Robotics: Architecture and Challenges," *IEEE Access*, vol. 6, pp. 36 662–36 682, 2018.

[29] M. Chowdhury and M. Maier, "Collaborative Computing for Advanced Tactile Internet Human-to-Robot (H2R) Communications in Integrated FiWi Multirobot Infrastructures," *IEEE Internet of Things Journal*, vol. 4, no. 6, pp. 2142–2158, Dec. 2017.

[30] M. Chowdhury and M. Maier, "Local and Nonlocal Human-to-Robot Task Allocation in Fiber-Wireless Multi-Robot Networks," *IEEE Systems Journal*, vol. 12, no. 3, pp. 2250–2260, Sep. 2017.

[31] L. Wang, M. Liu, and M. Q. H. Meng, "A Hierarchical Auction-Based Mechanism for Real-Time Resource Allocation in Cloud Robotic Systems," *IEEE Transactions on Cybernetics*, vol. 47, no. 2, pp. 473–484, Feb. 2017.

[32] K. E. C. Booth, T. T. Tran, G. Nejat, and J. C. Beck, "Mixed-Integer and Constraint Programming Techniques for Mobile Robot Task Planning," *IEEE Robotics and Automation Letters*, vol. 1, no. 1, pp. 500–507, Jan. 2016.

[33] Z. F. Quek, W. Provancher, and A. Okamura, "Evaluation of Skin Deformation Tactile Feedback for Teleoperated Surgical Tasks," *IEEE Transactions on Haptics*, vol. 12, no. 2, pp. 102–113, Apr. 2019.

[34] F. Brizzi, L. Peppoloni, A. Graziano, E. D. Stefano, C. A. Avizzano, and E. Ruffaldi, "Effects of Augmented Reality on the Performance of Teleoperated Industrial Assembly Tasks in a Robotic Embodiment," *IEEE Transactions on Human-Machine Systems*, vol. 48, no. 2, pp. 197–206, Apr. 2018.

[35] S. Haddadin, L. Johannsmeier, and F. D. Ledezma, "Tactile Robots as a Central Embodiment of the Tactile Internet," *Proceedings of the IEEE*, vol. 107, no. 2, pp. 471–487, Feb. 2019.

[36] P. Brucker, *Scheduling Algorithms.* Springer, 2007.

[37] X. Sun and N. Ansari, "Latency Aware Workload Offloading in the Cloudlet Network," *IEEE Communications Letters*, vol. 21, no. 7, pp. 1481–1484, Jul. 2017.

[38] Q. Fan and N. Ansari, "Workload Allocation in Hierarchical Cloudlet Networks," *IEEE Communications Letters*, vol. 22, no. 4, pp. 820–823, Apr. 2018.

[39] L. Liu, Z. Chang, X. Guo, S. Mao, and T. Ristaniemi, "Multiobjective Optimization for Computation Offloading in Fog Computing," *IEEE Internet of Things Journal*, vol. 5, no. 1, pp. 283–294, Feb. 2018.

[40] T. G. Rodrigues, K. Suto, H. Nishiyama, and N. Kato, "Hybrid Method for Minimizing Service Delay in Edge Cloud Computing Through VM Migration and Transmission Power Control," *IEEE Transactions on Computers*, vol. 66, no. 5, pp. 810–819, May 2017.

[41] T. G. Rodrigues, K. Suto, H. Nishiyama, N. Kato, and K. Temma, “Cloudlets Activation Scheme for Scalable Mobile Edge Computing with Transmission Power Control and Virtual Machine Migration,” *IEEE Transactions on Computers*, vol. 67, no. 9, pp. 1287–1300, Sept. 2018.

[42] Y. Wang, M. Sheng, X. Wang, L. Wang, and J. Li, “Mobile-Edge Computing: Partial Computation Offloading Using Dynamic Voltage Scaling,” *IEEE Transactions on Communications*, vol. 64, no. 10, pp. 4268–4282, Oct. 2016.

[43] Y. Mao, J. Zhang, and K. B. Letaief, “Dynamic Computation Offloading for Mobile-Edge Computing With Energy Harvesting Devices,” *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 12, pp. 3590–3605, Dec. 2016.

[44] O. Muñoz, A. Pascual-Iserte, and J. Vidal, “Optimization of Radio and Computational Resources for Energy Efficiency in Latency-Constrained Application Offloading,” *IEEE Transactions on Vehicular Technology*, vol. 64, no. 10, pp. 4738–4755, Oct. 2015.

[45] S. Guo, B. Xiao, Y. Yang, and Y. Yang, “Energy-Efficient Dynamic Offloading and Resource Scheduling in Mobile Cloud Computing,” in *Proc. IEEE INFOCOM*, Apr. 2016, pp. 1–9.

[46] M. Maier and A. Ebrahimzadeh, “Towards Immersive Tactile Internet Experiences: Low-Latency FiWi Enhanced Mobile Networks With Edge Intelligence [Invited],” *IEEE/OSA Journal of Optical Communications and Networking, Special Issue on Latency in Edge Optical Networks*, vol. 11, no. 4, pp. B10–B25, Apr. 2019.

[47] M. Maier, A. Ebrahimzadeh, and M. Chowdhury, “The Tactile Internet: Automation or Augmentation of the Human?” *IEEE Access*, vol. 6, pp. 41 607–41 618, 2018.

[48] M. C. A. Ebrahimzadeh and M. Maier, “The Tactile Internet over 5G FiWi Architectures,” *John Wiley & Sons:* Optical and Wireless Convergence for 5G Networks, to appear.

[49] A. Ebrahimzadeh and M. Maier, “Human-In-The-Loop Models for Multi-Access Edge Computing,” *IET Press:* Edge Computing: Models, Technologies, and Applications, to appear.

[50] A. Ebrahimzadeh, M. Chowdhury, and M. Maier, “Human-Agent-Robot Task Coordination in FiWi-based Tactile Internet Infrastructures Using Context-and Self-Awareness,” *IEEE Transactions on Network and Service Management*, vol. 16, no. 3, pp. 1127–1142, Sep. 2019.

[51] A. Ebrahimzadeh and M. Maier, “Delay-Constrained Teleoperation Task Scheduling and Assignment for Human+Machine Hybrid Activities over FiWi Enhanced Networks,” *IEEE Transactions on Network and Service Management*, IEEE Xplore Early Access.

[52] A. Ebrahimzadeh and M. Maier, “Tactile Internet over Fiber-WirelessŰEnhanced LTE-A HetNets via Artificial Intelligence-Embedded Multi-Access Edge Computing,” *CRC Press:* 5G-Enabled Internet of Things, to appear.

[53] Amin Ebrahimzadeh and Martin Maier, “Distributed Cooperative computation Offloading in Multi-Access Edge Computing Fiber-Wireless Networks (Invited paper),” *Elsevier Optics Communications Special Issue on Photonics for 5G Mobile Networks and Beyond*, vol. 452, pp. 130 – 139, Dec. 2019.

[54] A. Ebrahimzadeh and M. Maier, "Cooperative Computation Offloading in FiWi Enhanced 4G HetNets Using Self-Organizing MEC," *Submitted to IEEE Transactions on Wireless Communications*, in revision.

[55] A. Ebrahimzadeh and M. Maier, "Next Generation Multi-Access Edge-Computing Fiber-Wireless Enhanced HetNets for Low-Latency Immersive Applications," *IGI Global:* Design, Implementation, and Analysis of Next Generation Optical Networks, to appear.

[56] A. Aijaz, M. Dohler, A. H. Aghvami, V. Friderikos, and M. Frodigh, "Realizing the Tactile Internet: Haptic Communications over Next Generation 5G Cellular Networks," *IEEE Wireless Communications*, vol. 24, no. 2, pp. 82–89, Apr. 2017.

[57] M. Simsek, A. Aijaz, M. Dohler, J. Sachs, and G. Fettweis, "5G-Enabled Tactile Internet," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 3, pp. 460–473, Mar. 2016.

[58] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. K. Soong, and J. C. Zhang, "What Will 5G Be?" *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1065–1082, June 2014.

[59] J. G. Andrews, "Seven ways that HetNets are a cellular paradigm shift," *IEEE Communications Magazine*, vol. 51, no. 3, pp. 136–144, Mar. 2013.

[60] A. Aijaz, Z. Dawy, N. Pappas, M. Simsek, S. Oteafy, and O. Holland, "Toward a tactile Internet reference architecture: Vision and progress of the IEEE P1918. 1 standard," *arXiv preprint arXiv:1807.11915*, 2018.

[61] M. Satyanarayanan, "The Emergence of Edge Computing," *IEEE Computer*, vol. 50, no. 1, pp. 30–39, Jan. 2017.

[62] T. Taleb, K. Samdanis, B. Mada, H. Flinck, S. Dutta, and D. Sabella, "On Multi-Access Edge Computing: A Survey of the Emerging 5G Network Edge Cloud Architecture and Orchestration," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 3, pp. 1657–1681, Third quarter 2017.

[63] K. Antonakoglou, X. Xu, E. Steinbach, T. Mahmoodi, and M. Dohler, "Toward Haptic Communications Over the 5G Tactile Internet," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 4, pp. 3034–3059, Fourth quarter 2018.

[64] E. Weber, "Die Lehre vom Tastsinn und Gemeingefuehl, auf Versuche gegruendet," *London, UK: Verlag Friedrich Vieweg und Sohn*, 1978.

[65] C. P. L. Meli and D. Prattichizzo, "Experimental Evaluation of Magnified Haptic Feedback for Robot-assisted Needle Insertion and Palpation," *The International Journal of Medical Robotics and Computer Assisted Surgery*, vol. 13, no. 4, p. e1809, Feb. 2017.

[66] X. Xu, C. Schuwerk, B. Cizmeci, and E. Steinbach, "Energy Prediction for Teleoperation Systems That Combine the Time Domain Passivity Approach with Perceptual Deadband-Based Haptic Data Reduction," *IEEE Transactions on Haptics*, vol. 9, no. 4, pp. 560–573, Oct. 2016.

[67] G. Fettweis and S. Alamouti, "5G: Personal Mobile Internet Beyond what Cellular Did to Telephony," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 140–145, Feb. 2014.

[68] K. Hornik, M. Stinchcombe, and H. White, “Multilayer Feedforward Networks are Universal Approximators,” *Neural Networks*, vol. 2, no. 5, pp. 359–366, 1989.

[69] S. Kafaie, M. H. Ahmed, Y. Chen, and O. A. Dobre, “Performance Analysis of Network Coding with IEEE 802.11 DCF in Multi-Hop Wireless Networks,” *IEEE Transactions on Mobile Computing*, vol. 17, no. 5, pp. 1148–1161, May 2018.

[70] K. Medepalli and F. A. Tobagi, “Towards Performance Modeling of IEEE 802.11 Based Wireless Networks: A Unified Framework and Its Applications,” in *Proc. IEEE INFOCOM*, Apr. 2006, pp. 1–12.

[71] Y. h. Zhu, H. c. Lu, and V. C. M. Leung, “Access Point Buffer Management for Power Saving in IEEE 802.11 WLANs,” *IEEE Transactions on Network and Service Management*, vol. 9, no. 4, pp. 473–486, Dec. 2012.

[72] Y. Liu, M. Liu, and J. Deng, “Evaluating Opportunistic Multi-Channel MAC: Is Diversity Gain Worth the Pain?” *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 11, pp. 2301–2311, Nov. 2013.

[73] Y. S. Han, J. Deng, and Z. J. Haas, “Analyzing Multi-Channel Medium Access Control Schemes with ALOHA Reservation,” *IEEE Transactions on Wireless Communications*, vol. 5, no. 8, pp. 2143–2152, Aug. 2006.

[74] P. P. Pham, S. Perreau, and A. Jayasuriya, “New Cross-Layer Design Approach to Ad Hoc Networks Under Rayleigh Fading,” *IEEE Journal on Selected Areas in Communications*, vol. 23, no. 1, pp. 28–39, Jan. 2005.

[75] F. Aurzada, M. Lévesque, M. Maier, and M. Reisslein, “FiWi Access Networks Based on Next-Generation PON and Gigabit-Class WLAN Technologies: A Capacity and Delay Analysis,” *IEEE/ACM Transactions on Networking*, vol. 22, no. 4, pp. 1176–1189, Aug. 2014.

[76] J. M. Bradshaw, V. Dignum, C. M. Jonker, and M. Sierhuis, “Human-Agent-Robot Teamwork,” in *Proc. ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Mar. 2012, pp. 487–487.

[77] M. Chowdhury, E. Steinbach, W. Kellerer, and M. Maier, “Context-Aware Task Migration for HART-Centric Collaboration over FiWi Based Tactile Internet Infrastructures,” *IEEE Transactions on Parallel and Distributed Systems*, vol. 29, no. 6, pp. 1231–1246, June 2018.

[78] R. B. Freeman, “Who Owns the Robots Rules the World,” *Harvard Magazine*, vol. 118, no. 5, pp. 37–39, May/Jun 2016.

[79] P. Tokekar, N. Karnad, and V. Isler, “Energy-optimal Trajectory Planning For Car-like Robots,” *Autonomous Robots*, vol. 37, no. 3, pp. 279–300, Oct. 2014.

[80] S. M. Ross, *Introduction to Probability Models*. Academic press, 2014.

[81] C. You, K. Huang, H. Chae, and B. H. Kim, “Energy-Efficient Resource Allocation for Mobile-Edge Computation Offloading,” *IEEE Transactions on Wireless Communications*, vol. 16, no. 3, pp. 1397–1411, Mar. 2017.

[82] P. Nowak, “Nissan Uses NASA Rover Tech to Remotely Oversee Autonomous Car,” New Scientist, [Online; accessed 2018-04-12].

[83] J. Liu, H. Guo, H. Nishiyama, H. Ujikawa, K. Suzuki, and N. Kato, “New Perspectives on Future Smart FiWi Networks: Scalability, Reliability, and Energy Efficiency,” *IEEE Communications Surveys & Tutorials*, vol. 18, no. 2, pp. 1045–1072, Second quarter 2016.

[84] M. Maier, “Towards 5G: Decentralized routing in FiWi enhanced LTE-A HetNets,” in *Proc. IEEE International Conference on High Performance Switching and Routing (HPSR)*, Jul. 2015, pp. 1–6.

[85] D. V. Lindley, “The Theory of Queues with a Single Server,” *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 48, no. 2, pp. 277–289, Apr. 1952.

[86] J. A. Buzacott, “Commonalities in Reengineered Business Processes: Models and Issues,” *Management Science*, vol. 42, no. 5, pp. 768–782, May 1996.

[87] A. S. Thyagaturu, A. Mercian, M. P. McGarry, M. Reisslein, and W. Kellerer, “Software Defined Optical Networks (SDONs): A Comprehensive Survey,” *IEEE Communications Surveys & Tutorials*, vol. 18, no. 4, pp. 2738–2786, Fourth quarter 2016.

[88] J. A. Cabrera, R. Schmoll, G. T. Nguyen, S. Pandi, and F. H. P. Fitzek, “Softwarization and Network Coding in the Mobile Edge Cloud for the Tactile Internet,” *Proceedings of the IEEE*, vol. 107, no. 2, pp. 350–363, Feb. 2019.

[89] X. Chen, L. Jiao, W. Li, and X. Fu, “Efficient Multi-User Computation Offloading for Mobile-Edge Cloud Computing,” *IEEE/ACM Transactions on Networking*, vol. 24, no. 5, pp. 2795–2808, Oct. 2016.

[90] Y. Xiao and M. Krunz, “Distributed Optimization for Energy-Efficient Fog Computing in the Tactile Internet,” *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 11, pp. 2390–2400, Nov. 2018.

[91] P. V. Klaine, M. A. Imran, O. Onireti, and R. D. Souza, “A Survey of Machine Learning Techniques Applied to Self-Organizing Cellular Networks,” *IEEE Communications Surveys & Tutorials*, vol. 19, no. 4, pp. 2392–2431, Fourth quarter 2017.

[92] B. P. Rimal, D. P. Van, and M. Maier, “Mobile Edge Computing Empowered Fiber-Wireless Access Networks in the 5G Era,” *IEEE Communications Magazine*, vol. 55, no. 2, pp. 192–200, Feb. 2017.

[93] H. Tan, Z. Han, X. Y. Li, and F. C. M. Lau, “Online Job Dispatching and Scheduling in Edge-Clouds,” in *Proc. IEEE INFOCOM*, May 2017, pp. 1–9.

[94] L. Tong, Y. Li, and W. Gao, “A Hierarchical Edge Cloud Architecture for Mobile Computing,” in *Proc. IEEE INFOCOM*, Apr. 2016, pp. 1–9.

[95] M. H. Chen, B. Liang, and M. Dong, “Joint Offloading and Resource Allocation for Computation and Communication in Mobile Cloud with Computing Access Point,” in *Proc. IEEE INFOCOM*, May 2017, pp. 1–9.

[96] H. Guo and J. Liu, “Collaborative Computation Offloading for Multiaccess Edge Computing Over Fiber-Wireless Networks,” *IEEE Transactions on Vehicular Technology*, vol. 67, no. 5, pp. 4514–4526, May 2018.

[97] S. Kafaie, M. H. Ahmed, Y. Chen, and O. A. Dobre, "Performance Analysis of Network Coding with IEEE 802.11 DCF in Multi-Hop Wireless Networks," *IEEE Transactions on Mobile Computing*, vol. 17, no. 5, pp. 1148–1161, May 2018.

[98] Y. Zhu, H. Lu, and V. C. M. Leung, "Access Point Buffer Management for Power Saving in IEEE 802.11 WLANs," *IEEE Transactions on Network and Service Management*, vol. 9, no. 4, pp. 473–486, Dec. 2012.

[99] Y. Liu, M. Liu, and J. Deng, "Evaluating Opportunistic Multi-Channel MAC: Is Diversity Gain Worth the Pain?" *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 11, pp. 2301–2311, Nov. 2013.

[100] Y. S. Han, Jing Deng, and Z. J. Haas, "Analyzing Multi-Channel Medium Access Control Schemes with ALOHA Reservation," *IEEE Transactions on Wireless Communications*, vol. 5, no. 8, pp. 2143–2152, Aug. 2006.

[101] P. P. Pham, S. Perreau, and A. Jayasuriya, "New Cross-Layer Design Approach to Ad Hoc Networks under Rayleigh Fading," *IEEE Journal on Selected Areas in Communications*, vol. 23, no. 1, pp. 28–39, Jan. 2005.

[102] A. P. Miettinen and J. K. Nurminen, "Energy Efficiency of Mobile Clients in Cloud Computing," in *Proc. 2nd USENIX Conference on Hot Topics in Cloud Computing*, Berkeley, CA, USA, June 2010, pp. 1–7.

Appendix A

Appendix

A.1 Proof of Lemma 3.1

Proof. Using Eq. (3.1) along with the velocity profile shown in Fig. 3.2, the energy consumption E_{trav} of the MR to traverse a given distance Δd is calculated as follows:

$$E_{trav} = E(\omega_d) = \underbrace{\int_0^{T_{trav}} c_1 a^2(t).dt}_{E_1} + \underbrace{\int_0^{T_{trav}} c_2 v^2(t).dt}_{E_2} + \underbrace{\int_0^{T_{trav}} c_3 v(t).dt}_{E_3} + \underbrace{\int_0^{T_{trav}} c_4.dt}_{E_4}. \tag{A.1}$$

Note that for the considered velocity profile $v(t)$ in Fig. 3.2, the contributions of both fifth and sixth terms in Eq. (3.1) are equal to zero, as $\int_0^{T_{trav}} a(t).dt = 0$ and $\int_0^{T_{trav}} v(t)a(t).dt = 0$. We obtain E_1, E_2, E_3, and E_4 as follows:

$$\begin{aligned} E_1 &= c_1 \left(\int_0^{t_1} a_{acc}^2.dt + 0 + \int_{t_2}^{t_3} a_{dec}^2.dt\right) = c_1 \left(\int_0^{T_{acc}} (\frac{v_{max}}{T_{acc}})^2.dt + \int_{T_{acc}+T_{cst}}^{T_{trav}} (-\frac{v_{max}}{T_{dec}})^2.dt\right) \\ &\overset{\text{Eq. (3.4)}}{=} 2c_1(1-\omega_d)\frac{\Delta d}{v_{max}} \left(\frac{v_{max}^2}{(1-\omega_d)\,\Delta d}\right)^2 = \frac{2c_1 v_{max}^3}{(1-\omega_d)\Delta d}, \end{aligned} \tag{A.2}$$

$$\begin{aligned} E_2 &= c_2(\int_0^{t_1} \left(\frac{v_{max}}{T_{acc}}t\right)^2.dt + \int_{t_1}^{t_2} \left(v_{max}\right)^2.dt + \int_{t_2}^{t_3} \left(-\frac{v_{max}}{T_{acc}}(t-T_{trav})\right)^2.dt) \\ &\overset{\text{Eq. (3.4)}}{=} c_2 \left(\frac{2}{3}(1-\omega_d)v_{max}\Delta d + \omega_d v_{max}\Delta d\right), \end{aligned} \tag{A.3}$$

$$E_3 = c_3 \int_0^{T_{trav}} v(t).dt \overset{\text{Eq. (3.3)}}{=} c_3\Delta d, \tag{A.4}$$

$$E_4 = c_4 \int_0^{T_{trav}} 1.dt \overset{\text{Eq. (3.5)}}{=} c_4(2-\omega_d)\frac{\Delta d}{v_{max}}. \tag{A.5}$$

Substituting Eqs. (A.2)-(A.5) into Eq. (A.1), completes the proof. □

A.2 Proof of Lemma 3.2

Proof. Given $\frac{\partial^2 E(\omega_d)}{\partial \omega_d^2} > 0$ for $\omega_d \in (0,1)$, in order for $E(\omega_d)$ to have a local minimum in interval $(0,1)$, $\frac{\partial E(\omega_d)}{\partial \omega_d}$ has to be zero. Therefore, we have

$$\frac{\partial E(\omega_d)}{\partial \omega_d} = \frac{v_{max} c_2 \Delta d}{3} + \frac{2c_1 v_{max}^3}{\Delta d (1-\omega_d)^2} - \frac{c_4 \Delta d}{v_{max}} = 0, \tag{A.6}$$

which gives $\hat{\omega}_d$ as follows:

$$\hat{\omega}_d = 1 - \sqrt{\underbrace{\frac{\frac{6c_1 v_{max}^4}{\Delta d^2}}{3c_4 - v_{max}^2 c_2}}_{M'}}. \tag{A.7}$$

We note that $\hat{\omega}_d$ has to lie in interval $(0,1)$, thus implying that

$$0 < M' < 1. \tag{A.8}$$

The left-hand inequality holds for $v_{max} < v'_1$, where $v'_1 = \sqrt{\frac{3c_4}{c_2}}$. Whereas the right-hand inequality translates into

$$\overbrace{\frac{6c_1}{\Delta d} v_{max}^4 + c_2 v_{max}^2 - 3c_4}^{Q_E(v_{max})} < 0. \tag{A.9}$$

To evaluate the range of v_{max}, for which inequality $Q_E(v_{max}) < 0$ holds, we first determine the roots of $Q_E(v_{max}) = 0$. In doing so, we develop the auxiliary equation $Q'_E(v_{max}) = 0$ by replacing $v'_{max} = v_{max}^2$. We note that as the discriminant of equation $Q'_E(v_{max}) = 0$ is equal to $72c_1c_4/\Delta d^2$, which is greater than zero for Δd, $Q'_E(v_{max}) = 0$ has two distinct roots, one of which is positive and the other one is negative. Clearly, as $v_{max} = \pm\sqrt{v'_{max}}$, the negative root does not give a valid real value for v_{max}, whereas the positive one does. Therefore, $Q_E(v_{max})$ has only one positive root. Note that $Q_E(v_{max}) < 0$ holds for $v_{max} < v'_2$, where v'_2 is the positive root of $Q_E(v_{max}) = 0$. The reason for this is that $Q_E(0) < 0$ and $\frac{\partial^2 Q_E(v_{max})}{\partial v_{max}^2} > 0$, thus implying that $Q_E(0)$ is negative for $0 < v_{max} < v'_2$.Therefore, inequality (A.9) holds if

$$v_{max} < \overbrace{\sqrt{\frac{-c_2 + \sqrt{c_2^2 - 4\left(\frac{6c_1}{\Delta d^2}\right)(-3c_4)}}{2\left(\frac{6c_1}{\Delta d^2}\right)}}}^{v'_2}. \tag{A.10}$$

Subsequently, in order to satisfy Eq. (A.8), we have

$$v_{max} < \overbrace{\min\{v'_1, v'_2\}}^{v_1}, \tag{A.11}$$

for which the right-hand is equal to v'_2 because it is straightforward to show that $v'_2 < v'_1$ for $\Delta d > 0$ given the experiment-driven values of $\{c_i\}_{i=1}^6$ taken from [79]. For illustration, $\mathbf{A}_1 \in \mathbb{R}_+^2$ depicts the region that satisfies inequality (A.11), thus representing the values of $(\Delta d, v_{max})$, for which $E(\omega_d)$ has a local minimum $\forall \omega_d \in (0,1)$ (see Fig. 3.4). □

A.3 Proof of Lemma 3.3

Proof. $g(\omega_d) = \frac{\partial f(\omega_d)}{\partial \omega_d}$ is a continuous function of ω_d^*, thereby having a root in interval (0,1) if and only if $g(0)g(1) < 0$, which implies that

$$(1 - K')K' < 0 \Rightarrow 0 < K' < 1. \tag{A.12}$$

The left-hand inequality, $0 < K'$, reduces to

$$\overbrace{A_1 v_{max}^4 + B_1 v_{max}^2 + C_1 v_{max} + D_1}^{Q_1(v_{max})} > 0, \tag{A.13}$$

where A_1, B_1, C_1, and D_1 are given in Eq. (3.29). Note that in order to evaluate the range of v_{max} for which inequality (A.12) holds, we have to determine the location of the roots (i.e., zeros) of equation $Q_1(v_{max}) = 0$. We also note that $Q_1(v_{max}) = 0$ is a quartic equation that has four roots, two of which are real while the other two are complex, as its discriminant is negative. Further, as $Q_1(0) > 0$ holds and $\lim_{v_{max}\to+\infty} Q_1(v_{max}) = -\infty$, we conclude that one of the real roots is positive while the other one is negative. Therefore, inequality (A.12) holds for $v_{max} < z_m$, where z_m is the (only) positive root of $Q_1(v_{max}) = 0$.

Next, we turn our attention to the right-hand side inequality, $K' < 1$, which reduces to

$$\begin{aligned} 2c_1 v_{max}^3 &< \Delta d \left(\frac{\Delta d c_4}{v_{max}} - \frac{\Delta d v_{max} c_2}{3} + \frac{E_m}{2} \right) \\ &\overset{v_{max}>0}{\Rightarrow} \overbrace{A_2 v_{max}^4 + B_2 v_{max}^2 + C_2 v_{max} + D_2}^{Q_2(v_{max})} > 0, \end{aligned} \tag{A.14}$$

where

$$\begin{aligned} A_2 &= A_1 \Delta d - 2c_1, \\ B_2 &= B_1 \Delta d, \\ C_2 &= C_1 \Delta d, \\ D_2 &= D_1 \Delta d. \end{aligned} \tag{A.15}$$

We note that $Q_2(v_{max})$ is greater than zero only for $v_{max} < z'_m$, where z'_m is the (only) positive root of $Q_2(v_{max}) = 0$. The reason for this is that as $Q_2(v_{max}) = 0$ has two real roots, one of which is positive and the other one is negative, and $Q_2(0) > 0$, $Q_2(v_{max})$ is greater than zero for $v_{max} < z'_m$. Clearly, in order for both right- and left-hand inequalities in (A.12) to hold, v_{max} has to be smaller than $\min\{z_m, z'_m\}$. We note that for $\Delta d > 0$ we have $z'_m < z_m$, therefore $\min\{z_m, z'_m\} = z'_m$. Standing as the only positive root of $Q_2(v_{max})$, z'_m is $\max_{Z'_i > 0 : \mathfrak{Im}[Z'_i]=0}\{Z'_i\}$. Then, $g(\omega_d) = 0$ has a root ω_d^* in interval $(0, 1)$ if and only if $v_{max} < \max_{Z'_i > 0 : \mathfrak{Im}[Z'_i]=0}\{Z'_i\}$, $\forall \Delta d > 0$. □

A.4 Proof of Lemma 5.1

Proof. To compute the average channel access delay, we define a two-dimensional Markov process $(s(t), b(t))$ shown in Fig. A.1 under unsaturated conditions and estimate the average service time Δ_i of MU i in a WLAN using the IEEE 802.11 distributed coordination function (DCF) for access

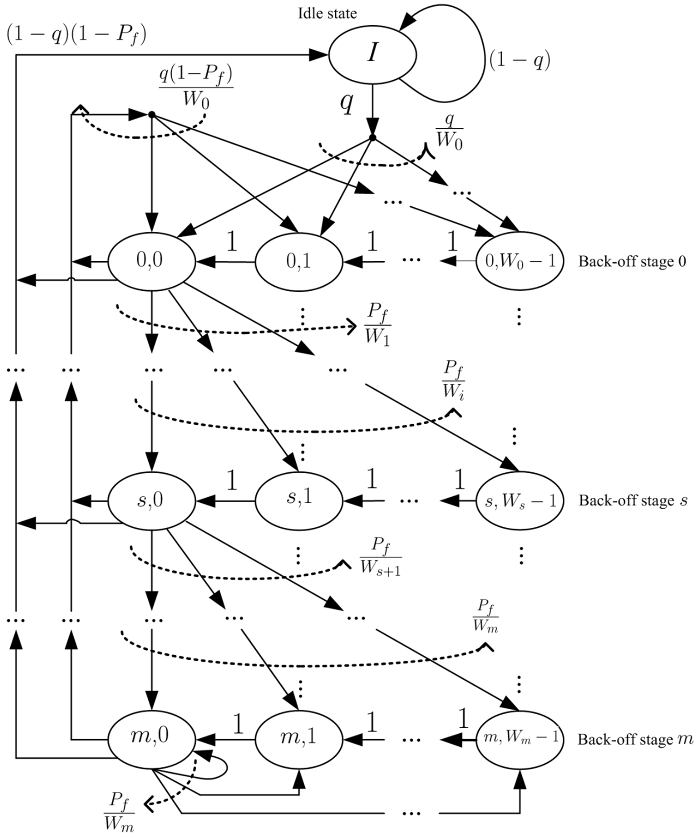


Figure A.1 – Two-dimensional Markov chain for DCF contention model under unsaturated traffic conditions.

control, whereby $b(t)$ and $s(t)$ denote the random back-off counter and size of the contention window at time t, respectively. Without loss of generality, let us focus on a tagged user and drop the subscript i for now. Let P_f and W_s denote the probability of a failed transmission attempt (i.e., collision or erroneous transmission) and contention window size at back-off stage s, respectively. Note that the back-off stage s is incremented after each failed attempt up to the maximum value m, while the contention window is doubled at each stage, i.e., $W_s = 2^s W_0$.

From the viewpoint of a WiFi user, collisions may occur with probability p_c on transmitted packets, while erroneous transmission attempts may happen with probability p_e. Assuming that the collided and erroneous transmission events are statistically independent, a packet is successfully transmitted after a collision-free attempt followed by an error-free transmission. The probability of a successful transmission is therefore equal to $(1-p_e)(1-p_c)$, from which we infer that the probability P_f of a failed transmission attempt is computed as follows:

$$P_f = 1 - (1-p_c)\,(1-p_e)\,. \tag{A.16}$$

A WiFi user is in idle state, if (i) a successfully transmitted packet leaves the system without any waiting packet in the queue and (ii) no packet arrives during the current time slot given that the user was in idle state in the preceding time slot. We note that for Poisson arrival these two events are identical and equal to $1-q$.

With these considerations, we move on to analyze the Markov model in Fig. A.1, where $m+1$ different back-off stages are considered. The maximum contention window size is $2^m W_0$. Transmissions are attempted only in $(s,0)$ states $(s=0,1,...,m)$. Upon a failed transmission attempt in state $(s,0)$, there will be a transition to new state $(s+1,k)$, where k is uniformly selected from $[0,W_{s+1}]$. From state $(s,0)$, we enter the initial back-off stage, again given that the transmission is successful and the buffer is still nonempty; otherwise, we transit in the idle state I and wait for an incoming packet.

The transition probabilities of the two-dimensional Markov chain in Fig. A.1 are computed as follows:

$$P_{(s,k)|(s,k+1)} = 1; \qquad \forall k \in [0, W_s - 2], s \in [0,m] \tag{A.17a}$$

$$P_{(0,k)|(s,0)} = \frac{q(1-P_f)}{W_0}; \qquad \forall k \in [0, W_0 - 1], s \in [0,m] \tag{A.17b}$$

$$P_{(s,k)|(s-1,0)} = \frac{P_f}{W_s}; \qquad \forall k \in [0, W_0 - 1], s \in [1,m] \tag{A.17c}$$

$$P_{(m,k)|(m,0)} = \frac{P_f}{W_m}; \qquad \forall k \in [0, W_m - 1] \tag{A.17d}$$

$$P_{I|(s,0)} = (1-q)\left(1-P_f\right); \quad \forall s \in [0,m] \tag{A.17e}$$

$$P_{(0,k)|I} = \frac{q}{W_0}; \qquad \forall k \in [0, W_0 - 1] \tag{A.17f}$$

$$P_{I|I} = 1-q, \tag{A.17g}$$

where $P_{(a,b)|(c,d)}$ denotes the transition probability from state $(s(t)=c, b(t)=d)$ at time t to state $(s(t+1)=a, b(t+1)=b)$ at time $t+1$.

In order to find the stationary distributions

$$b_{s,k} = \lim_{k\to\infty} P\left(s(t)=s, b(t)=k\right), \forall k \in [0, W_s - 1], s \in [0,m],$$

we consider Eqs. (A.17) together with the normalization equation

$$b_I + \sum_s \sum_k b_{s,k} = 1,$$

where b_I denote the stationary probability that the WiFi user is in idle state. After finding the stationary distributions, the probability τ that a WiFi user attempts to transmit in a given time slot is then obtained as

$$\tau = \sum_{s=0}^{m} b_{s,0} = \frac{1}{\frac{W_0+1}{2} + \frac{W_0 P_f\left(1-\left(2P_f\right)^m\right)}{2\left(1-2P_f\right)q} + \frac{(1-q)\left(1-P_f\right)}{q}}. \tag{A.18}$$

From a system point of view, WiFi subscriber i does not experience a collision if the remaining users do not attempt to transmit, thus $1 - p_{c_i} = \prod_{v:v\neq i}\left(1-\tau_v\ \right)$. Moreover, $p_{e,i}$ is estimated by

$1-(1-p_b)^{\bar{L}_i}$, where $\bar{L}_i$ and p_b is the average length of a packet transmitted by WiFi user i and bit error probability, respectively.

The probability of a collision-free packet transmission P_s provided that there is at least one transmission attempt is given by $\frac{1}{P_{tr}}\left(\sum_i \tau_i \prod_{v,v\neq i}(1-\tau_v\ )\right)$, whereby the probability P_{tr} that there is at least one transmission attempt is equal to $1-\prod_i(1-\tau_i\ )$. The average slot duration E_s is then obtained as

$$E_s=(1-P_{tr})\,\epsilon+P_{tr}\,(1-P_s)\,T_c+P_{tr}P_sP_eT_e+P_{tr}P_s\,(1-P_e)\,T_s, \tag{A.19}$$

where T_c, T_e, and T_s are given in [75]. We also note that q can be approximated as follows

$$q=1-e^{-\lambda E_s}, \tag{A.20}$$

whereby E_s is given in Eq. (A.19). In order to obtain the steady-state values of q, P_f, τ, and E_s, and Δ_i, we numerically solve the system of non-linear equations (5.15), (A.18), (A.19), and (A.20).

□