Pricing design and resource allocation for 5G services

Naresh Modina

AVIGNON UNIVERSITÉ

THESIS

presented at the Avignon Université
in Partial Fulfillment
of the Requirements for the Degree of
DOCTOR OF PHILOSOPHY
in

Graduate School 536
Agrosciences & Sciences
Sciences, Technologies, Santé
Speciality : *Computer Science*

By

Naresh Modina

Pricing design and resource allocation for 5G services

Research Unity : EA 4128 LIA – Laboratoire Informatique d'Avignon

Jury members :

Reviewers :	M. Mohamad ASSAAD	Professor, CentraleSupelec, France
	M. Tijani CHAHED	Professor, Telecom SudParis, France
Examiners :	M. Balakrishna J. PRABHU	CRCN CNRS, LAAS, France
	M^{me}. Jahanne COHEN	DR CNRS, Université Paris-Saclay, France
Advisor :	M. Rachid EL-AZOUZI	Professor, Avignon Université, France
Co-Advisor :	M. Francesco DE PELLEGRINI	Professor, Avignon Université, France

LABORATOIRE INFORMATIQUE D'AVIGNON

Acknowledgement

First and foremost, I would like to thank my thesis advisor Rachid Elazouzi and co-advisor Francesco de Pellegrini for their invaluable guidance and incredible support. I have been fortunate to work with you for nearly four years and learned the art of conducting research, your knowledge and vision helped me to lay the foundations for my research career. I would like to thank my collaborators Rosa Figueiredo, Daniel Sadoc Menasche, and Stefano Secci for your support, you helped me extend my knowledge and improve as a researcher.

I would like to extend my gratitude towards the jury members Tijani Chahed, Mohamad Assaad, Balakrishna J Prabhu, and Johanne Cohen for spending your valuable time to review the thesis and providing excellent suggestions towards the future research.

To all my colleagues and friends Afaf Arfaoui, Cedric Richier, Mandar Datar, Nejat Arinik, Sarkis Moussa, Samira Habli, Julio cesar Perez, Sadaf Ul Zuhra, Trung Nguyen, Olivier Tsemogne, Vibhav Gupta, Priyanka, Harish, Kamalesh, Vinay, Daniel, your presence brought joy to my life both at work and outside. You have been kind and provided exceptional support to me, I thank you for your friendship.

Finally, I thank my dear wife Santhoshi for her enormous love and support during our journey in Avignon, you have been the backbone of this endeavour. My family has played an exceptional role in realizing my dream of pursuing Ph.D., Thanks is small word for your role in my life.

This work received financial support from the project MAESTRO5G, I would like to thank all the partners of the project for your valuable inputs.

Title: Pricing design and resource allocation for 5G services

Keywords: 5G Network Slicing, Kelly mechanism, Coupled constrained game, Normalized Nash equlibrium, Fisher market, aging control, traffic migration, Markov decision process, age of information, data services

Abstract: The widespread adoption of 5G cellular technology will evolve as one of the major drivers for the growth of IoT-based applications. In the first part of this thesis, we consider a service provider that launches a smart city service based on IoT data readings: to serve IoT data collected across different locations, the SP dynamically negotiates and re-scales bandwidth and service functions. Network slicing is becoming the platform of choice for several applications and services. Nowadays, most applications are virtualized to gain flexibility and portability. With network slicing, operators can create multiple network slices, which can be used for different applications with specific requirements. Behind the network slicing, a slice expresses the need to access a precise service type, under a fully qualified set of computing and network requirements. Also, different infrastructure providers charge slicing services depending on specific access technology supported across sites and IoT data collection patterns.

In the first part of this work, we introduce a pricing mechanism based on the age of information to reduce the cost of service providers. This provides incentives for devices to smooth traffic by shifting part of the traffic load from highly congested and more expensive locations to locations with cheaper prices while meeting the quality of service requirements of the IoT service. The proposed optimal pricing scheme comprises a two-stage decision process, where the SP determines the pricing of each location and devices schedule uploads of collected data, based on the optimal uploading policy. First, the upload of collected data to reduce the costs of the SPs is considered to be a decision problem. By employing a Markov decision process framework, we determine threshold-based optimal policies to achieve the primary objective using dynamic programming. We establish that the pricing of the locations can be reduced to finding appropriate thresholds respectively for each location, which shifts part of traffic from the highly congested locations to locations with lower congestion. Given the nature of the problem, we propose an algorithm based on simulated annealing to find the best combination of the thresholds. Then, we modify the algorithm to perform parallel computation using a well-known coloring technique that exploits the neighborhood structure of the locations to reduce the convergence time twofold.

One of the key contributors to the service provider cost is the cost of leasing a network slice. For this reason, we study the resource allocation for network slices in 5G wireless networks in the later part of the thesis. Resource allocation encompasses a combination of different resource types (e.g., radio resource, CPU, memory, bandwidth). In this work, we explore a differential pricing scheme that maximizes social welfare among slices as well as among end-users. To do so, we propose a pricing mechanism that makes fairness at multiple levels: fairness among slices and fairness among slice locations. Therefore, the proposed scheme is beneficial for both the slices and the end-users independent of their location. In addition, we study the case where slices can manipulate their preferences to improve their utility. We show that the Fisher market game always has a pure Nash equilibrium and we prove Price of Anarchy is $1/N$, where N is the number of slices.

A major drawback of resource allocation with a centralized approach is the privacy concerns of the service providers and infrastructure providers. In general, infrastructure providers do not prefer to reveal information related to the available resource quantity. On the other hand, service providers do not prefer to reveal their utility functions. In the final part of this thesis, we study a decentralized resource allocation mechanism inspired by the Kelly Mechanism that preserves multi-level fairness. In addition, we show that each infrastructure provider can implement its own allocation rule independent of the other. With the proposed mechanism, we establish that the resulting allocation is a social optimum. Each theoretical finding in this work is validated by numerical simulations in respective chapters.

Titre : Tarification et allocation des ressources pour les services 5G

Mot clés : 5G Network Slicing, Kelly mechanism, Coupled constrained game, Normalized Nash equlibrium, Fisher market, Trading post mechanism, aging control, traffic migration, markove decision process, age of information, data services

Résumé : L'adoption généralisée des réseaux cellulaires de cinquième génération (5G) deviendra l'un des principaux moteurs de la croissance des applications basées sur l'Internet des objets (IoT). En effet, la 5G offre non seulement ces services classiques (de façon améliorée),mais également de nouveaux services tel que l'Internet des Objets (IoT) ou l'Internet Tactile.

Dans la première partie de cette thèse, nous considérons un fournisseur de services (SP) qui lance une application nécessitant la récolte de données à partir d'objets connectés distribués dans différentes cellules. Cependant, l'objectif du SP est de minimiser le coût de cette récolte permanente de données. En raison de ces cas extrêmes d'usage, la 5G donne la possibilité de traiter de façon adaptée chaque trafic ou application. Pour cela, les techniques de virtualisation ont été introduites dans la 5G pour traiter les applications par des ressources en couches (network slicing) de façon à s'adapter à chaque besoin de façon efficace. Cependant, le découpage du réseau permet aux opérateurs de créer plusieurs tranches de réseau, qui peuvent être utilisées pour différentes applications avec des exigences spécifiques. Une tranche exprime le besoin d'accéder à un type de service précis, dans le cadre d'un ensemble complet d'exigences pour respecter le niveau de qualité de service (SLA : Service Level Agreement). En outre, différents fournisseurs d'infrastructure facturent des services de découpage en tranches en fonction à la fois de la technologie spécifique d'accès prise en charge sur les sites et des modèles de collecte de données IoT.

Dans la première partie de ce travail, afin de réduire le coût des fournisseurs de services, nous proposons un mécanisme de tarification basé sur l'âge de l'information et la tarification. Ce mécanisme incite les mobiles à lisser leur trafic en déplaçant une partie de la charge de trafic des cellules très chargées et plus chère vers des cellules à prix plus bas, tout en respectant les exigences de qualité de service (SLA). Le schéma de tarification optimale proposé, comprend un processus de décision en deux étapes : le SP détermine la tarification pour chaque cellule et les mobiles déterminent la stratégie à mettre en place pour l'envoi des données en fonction de l'âge de l'information et sa localisation. Nous présentons ce problème comme un processus de décision markovien et nous déterminons les politiques de seuil optimales qui permettent d'atteindre l'objectif principal. Nous établissons que la tarification de l'emplacement ou cellule peut être réduite à la recherche de seuils appropriés pour chaque cellule. Compte tenu de la nature du problème, nous proposons un algorithme pour trouver la meilleure combinaison de seuils. Ensuite, nous modifions l'algorithme pour effectuer un calcul parallèle en utilisant une technique de coloration qui exploite l'interconnexion des cellules pour réduire le temps de convergence.

L'un des principaux facteurs du coût du fournisseur de services est le coût de location d'une tranche de réseau. Pour cette raison, dans la dernière partie de la thèse, nous étudions l'allocation des ressources aux tranches de réseau, en ce qui concerne les réseaux sans fil 5G. L'allocation de ressources englobe une combinaison de divers types de ressources (par exemple, ressource radio, CPU, mémoire, bande passante). Dans ce travail, nous explorons un système de tarification différentielle qui maximise le bien-être social parmi les tranches ainsi que parmi les utilisateurs finaux. Pour ce faire, nous proposons un mécanisme de tarification qui aboutit à une tarification équitable à plusieurs niveaux : équité entre les tranches et équité entre les emplacements des tranches. Par conséquent, le schéma proposé est bénéfique à la fois pour les tranches et les utilisateurs finaux, indépendamment de leurs emplacements. De plus, nous étudions le cas où les tranches peuvent manipuler leurs préférences pour améliorer leur utilité, nous montrons que le jeu de marché de Fisher a toujours un équilibre de Nash en stratégies pures et nous prouvons que le coût de l'anarchie est de $1/N$, où N est le nombre de tranches.

Une insuffisance majeure de l'approche centralisée de l'allocation des ressources porte sur

la confidentialité des données des fournisseurs de services et des fournisseurs d'infrastructures. En général, les fournisseurs d'infrastructures ne préfèrent pas révéler les informations relatives à la quantité des ressources disponibles. En revanche, les fournisseurs de services ne préfèrent pas dévoiler leurs fonctions d'utilité. Dans la dernière partie de cette thèse, nous étudions un mécanisme décentralisé d'allocation des ressources inspiré du mécanisme de Kelly qui préserve l'équité à plusieurs niveaux. De plus, nous montrons que chaque fournisseur d'infrastructure peut implémenter sa propre règle d'allocation indépendamment de l'autre fournisseur. Avec le mécanisme proposé, nous établissons que l'allocation qui en résulte est un optimum social. Chaque découverte théorique de ce travail est validée par des simulations numériques dans les chapitres respectifs.

TABLE OF CONTENTS

List of Figures

List of Tables

Chapter 1

Introduction

Wireless networks have been playing a key role in breaking barriers and enabling new services in various fields such as medical, agriculture, domestic services, and any field that involves communication from a distance. This enables the service sector, large enterprises to small scale businesses to perform their daily operations. Starting from 2^{nd} generation, the network architecture has been constantly modified and adapted to the needs of the users. 2G networks introduced General Packet Radio Service (GPRS) [27] laying the foundation for packet-based communication. Although bit rates were significantly lower in the order of a few Kbps, given the timeline, it was a revolutionary idea as it provided wireless access to the internet. With the introduction of High-Speed Packet Access (HSPA) [42] in 3^{rd} generation networks, data speeds were improved up to 14.4 Mbps for downlink and 5.8 Mbps for uplink. 3G improved the data rates by using Code Division Multiple Access (CDMA) [95] instead of the traditional time/frequency division multiplexing that was used in 2G. Despite these improvements, the data rates were relatively slow and expensive for the users. In parallel, different wireless services were developed such as ZigBee, Bluetooth, WIFI, etc., that provide wireless access from a few meters to hundreds of meters with data rates ranging from a few Kbps to a few Mbps.

The next generation networks referred to as Long-Term Evolution (LTE) have completely changed the user experience by deploying Orthogonal Frequency Division Multiple Access (OFDMA) [92]. This 4^{th} generation networks offered a significant increase in data rates and a decrease in cost per gigabyte of data. The data rates ranged from 10 Mbps to 600 Mbps in LTE-Advanced. Along with OFDMA different technologies such as Multi Input Multi Output (MIMO) are used to achieve the said data rates. The network architecture has seen a notable change, the traffic is now completely packet based including the call service. The network is primarily composed of two fundamental networks called Radio Access Network (RAN) and Core Network (CN), where the RAN is responsible for providing access and performing baseband signal processing. On the other hand, the CN is responsible for the management of user traffic. 4G has improved the latency to 20ms - 30ms and provided better connections for users with faster mobility such as users in high-speed trains. This has given the necessary boost for businesses to implement digital services more effectively. Smartphones became more and more popular as the number of services improved the quality of experience for the users. Various services such as video streaming, online gaming, and video conferencing have seen enormous growth that was not possible with the previous generation of networks.

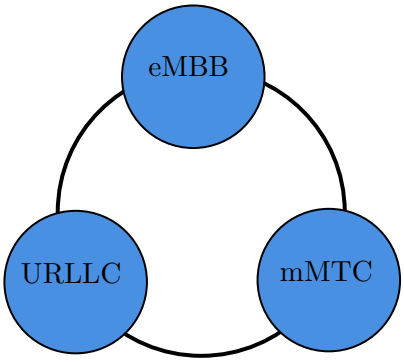


Figure 1.1: 5G use case scenarios.

So far the sole focus of the telecommunication industry has been focused on improving mobile broadband. But the new innovative ideas of smart homes, smart factories, smart cities, autonomous driving, and remote health care with heterogeneous access and resource requirements need diverse network capabilities that can support the ever-growing business ideas. This led to the most important technology thus far, it is the 5^{th} generation networks. First, we briefly describe 5G networks on an abstract level, then the network slicing which is responsible for providing tailor-made resources to the Service Providers (SP). Later, we explain the importance of IoT data services which is in the scope of this thesis, and the related metric named age of information which is very important for data services. Finally, this chapter ends with the description of mathematical models that were employed in this thesis and the organization of the remaining chapters.

1.1 5G Networks

5G, also referred to as New Radio (NR), is designed to accommodate evolved user requirements. The primary focus of the NR is to provide network access to a wide variety of services. The services that 5G can support are three broad categories named enhanced Bobile BroadBand (eMBB)[4], massive machine type communication (mMTC)[26], and Ultra Reliable Low Latency Communication (URLLC)[21] as shown in figure 1.1. eMBB is considered for regular mobile communication with higher data rates and better latency requirements to support the end-user needs for various applications such as online gaming, streaming services such as Spotify, Netflix, Prime Video, etc, and better audio and video calls. Other use cases like autonomous driving and remote health care require very high reliability and low latency to run the service without failures that can lead to disastrous outcomes, URLLC is specifically designed for these types of scenarios. With the rise of IoT, more and more devices are being connected to the internet for remote control and monitoring. This leads to massive connection requirements that can be categorized under mMTC. There are a few fundamental technological considerations that allow 5G to successfully achieve its main targets,

they are:

- millimeter (mm) waves: There are many limiting factors in enabling new technologies, one such limiting factor for telecommunications is available bandwidth. The available bandwidth for any major carrier is around 200MHz across various bands as reported in [87]. The number of users seeking internet connection is growing enormously, on the other hand, the number of devices per user is growing as well. So it is evident that at some point, we should look for a frequency band to increase the available bandwidth. As specified in [87], 28 and 38 GHz along with other bands have been explored for 5G. The waves in this frequency range are referred to as millimeter waves, extensive research is going on, to find a way to utilize these waves as they are not suitable in dense scenarios.

- Beamforming and Massive MIMO: millimeter waves are prone to absorption by the objects such as trees, hence normal methods of signal transmission may not work efficiently. The idea of concentrating the energy in a specific direction by adjusting the phases of the signals from multiple antennas, referred to as beamforming, is a well-known concept. Beamforming has been proven effective in steering the beams in a specific direction referred to as a look direction [33]. Beamforming is not only effective in steering the signals but also in the reception of the signals, referred to as spatial filtering [105]. In 5G, the number of antennas used at base stations is in the order 100s [68]. Implementing this in the cellular systems is a challenge due to the space constraints but current 5 G-enabled smartphones are equipped with multiple antennas in a limited capacity to handle the mm-waves.

Further discussion of these technologies is not in the scope of this thesis. With earlier mentioned technologies, 5G is a very powerful and flexible technology that enables the industry to reach new frontiers. Another key technology that provides the flexibility required for 5G to provide network access to heterogeneous services is called network slicing. This allows Infrastructure Providers (InP) to allocate tailor-made resources to the SPs.

1.2 Network Slicing

Network slicing is a scheme by which a logical network is created on top of the existing physical infrastructure. Then the logical network is logically separated to create network slices [46]. Such a slice contains the resources in required proportions to support the use cases mentioned earlier that has different requirements to run the services as depicted in figure 1.2. Thanks to technologies such as virtualization and Software Defined Networks (SDN), the realization of network slicing is achieved. Various SPs try to procure resources such as bandwidth, memory, storage, etc., over micro, macro, and small cells. These resources are found typically in the core cloud and edge

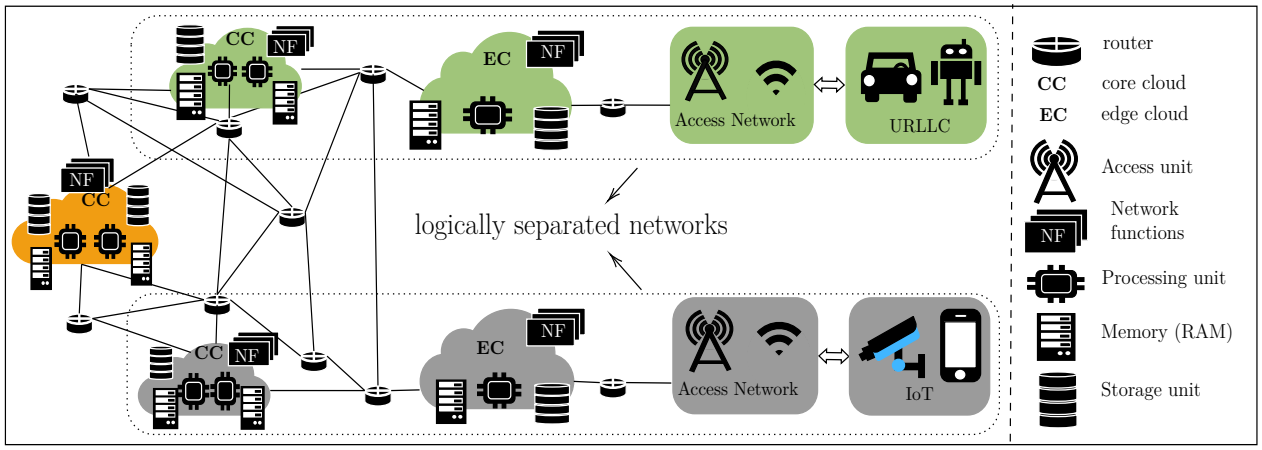


Figure 1.2: Schematic representation of Network slicing in 5G (inspired from [81]).

cloud. The core cloud consists of several data centers with incredible computing power and storage capacity, so the majority of the processing and storage is achieved with the cloud as it allows for flexible resource sharing.

However, the core cloud is often at a considerable distance compared to the geographic location at which the SP operates, this can lead to higher latency that is not desirable for services that rely on low latency. On the other hand, transferring large amounts of data is not appropriate as it creates enormous bandwidth requirements that can create a bottleneck in the network. Often this data can be pre-processed and then transfer the resulting data for further processing. This can be achieved by placing these resources closer to the operating locations in a limited capacity to achieve the objectives. The corresponding computing paradigm is referred to as edge cloud and this enables edge computing, an idea that recently gained traction.

Part of this work is focused on the operational aspect of the service providers. We assume that an IoT network slice has been allocated to the SPs and they run the data collection operation using mobile IoT devices. And the reminder of the thesis is focused on resource allocation for service providers with network slicing.

1.3 IoT data services and Age of Information

Data has been playing a decisive role in many applications ranging from health to daily weather and traffic updates. With the use of machine learning and deep learning, artificial intelligence has been widely deployed in many applications to predict or analyze the behavior of the subject of interest in many fields. Proper collection of data that helps to solve many real world problems has been of pivotal importance. Wide deployment of sensors to collect a wide range of data from agriculture to health has enabled the idea of the Internet of Things (IoT) that connects all these devices embed-

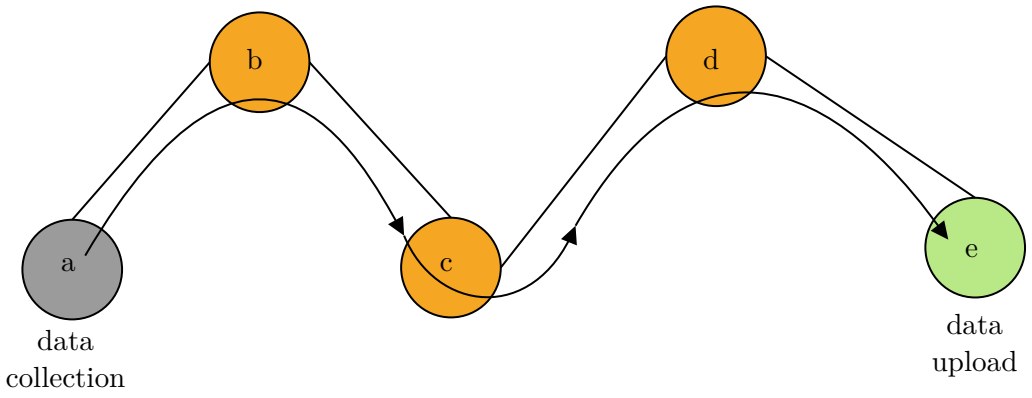


Figure 1.3: Age of information.

ded with various sensors to the internet. The predicted importance of IoT and the data collection concerning future technologies allow for a study of operational aspects of such services.

In this thesis, we study the upload mechanism for collecting data that address two key issues found in IoT data services: first is the metric that is associated with the data named age of information, second is the better management of traffic associated with uploading the collected data. Age of information is the time lapsed between the data collection and data upload as shown in figure 1.3. In this thesis, we consider that a fleet of mobile IoT devices is deployed to collect necessary data from diverse locations. A mobile vehicle carrying such a device moves across the region to collect enough data that represent different locations adequately. We consider that each location is associated with a price to upload the data depending on the upload traffic. General idea is that locations with heavy traffic are pricier in comparison to the locations with moderate to light loads. Hence for some applications that do not require data to be uploaded immediately after collection, the collected data can be differed anticipating a future incentive in terms of lower upload costs. For example: assume that there are five locations $a,\ b,\ c,\ d, and\ e$ out of which location e is a cheaper option to upload the information, and consider that it takes 2 minutes to reach each location on an average. Hence as shown in the figure 1.3, data is collected at a location a and the device waits until reaching the location e before uploading the data as the previously visited locations are expensive in comparison. In this way, part of the traffic can be moved from highly congested locations to less congested locations without exceeding the preset constrain on the average age of information.

This is the general idea behind the first part of the thesis that is described in chapter 2 and chapter 3 with extensive detail. There have been prior works that address the two facets of this problem separately, but to the best of our knowledge, this is the first proposed work that bridges these two aspects under a single framework. Given that the upload of data is associated with the cost that plays a key role in the decision-making process, it is important to study the costs associated

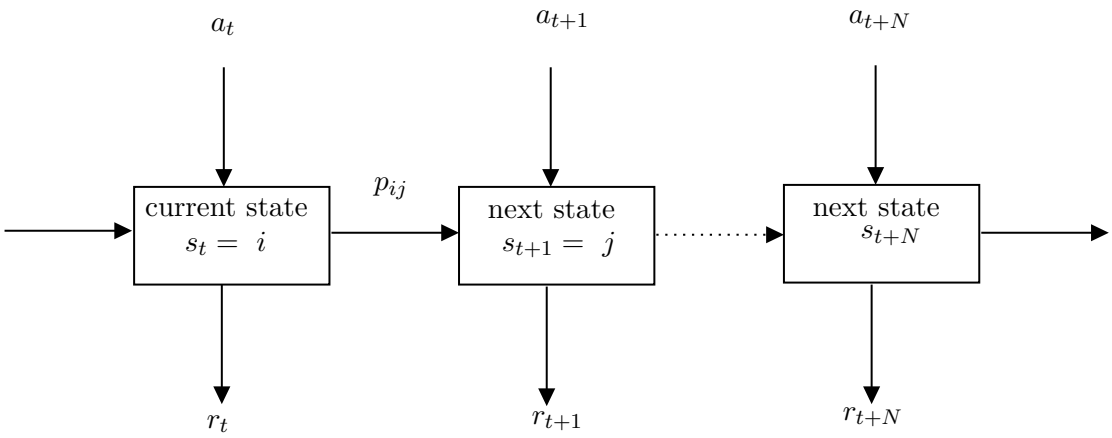


Figure 1.4: Sequential decision process (inspired from [85]).

with the lease of a slice. With this attempt, we study the resource allocation for network slices in the second part of this thesis work. In this part we comprehensively study the multi-resource allocation models for multi-users. We study two different approaches, first based on the Fisher market[17], we propose a centralized multi-resource allocation scheme with non-linear pricing, and the second work is based on the Kelly mechanism[60], where we proposed a mechanism for decentralized resource allocation to preserve the user privacy.

1.4 Mathematical tools

In this section we describe the basic mathematical models that are in use to provide comprehensive analysis of the problems that are in the scope of this work.

1.4.1 Markov Decision Process

Decision making is a very important process that decides the overall outcome of a system in the long run. A sequential decision making is a process where an agent takes an action by observing the state of the system at particular point in time. This results in an immediate reward (or cost) depending on the action and the system evolves to a new state as shown in figure 1.4. Following are the key components of a sequential decision theory [85].

- A set of system states $\mathcal{S}$.
- A set of decision epochs $\mathcal{N}$.
- A set of available actions $\mathcal{A}_s$.
- A set of immediate rewards that depend on state and action $\mathcal{R}_s$.

- A set of transition probabilities that depend on state and actions $p_t(.|s,a)$.

In this work, we employ a particular type of sequential decision model called the Markov Decision Process (MDP). In which the set of available actions, transition probabilities, and rewards depend only on the current state but not on the previous history. A $decision\ rule$ defines the action to be chosen at a specific time, and a $policy(\mu)$ is defined as the sequence of decision rules at all epochs. The decision process problem aims at choosing a policy before the first decision epoch to maximize a function of the sequence of rewards. One such function is the long-run average reward, often referred to as the expected total reward. It is defined as follows,

$$\mathbb{E}[r,\mu] = \lim_{\eta\to\infty} \frac{1}{\eta} \sum_{t=0}^{\eta-1} \mathbb{E}[r_t(x_t,a_t);\mu] \tag{1.1}$$

where $r_t(x_t,a_t)$ is an instantaneous reward, it is defined as follows,

$$r_t(x_t,a_t) = \sum_{j\in S} r_t(s,a,j)p_t(j|s,a) \tag{1.2}$$

The objective is to find an $optimal\ policy\ \mu$ that maximizes the average expected $\mathbb{E}[r,\mu]$. Further details of MDP can be found in [85].

Aging control: As described in [110], a new field of network research has emerged referred to as the timeliness of status updates. In any system with utility based on the status updates, sending more and more updates can increase the utility, however, this can lead to congestion among the communication channels which causes delayed updates. On the contrary, sending fewer updates may not be a good strategy as well, since the received updates can be outdated. In such cases, a simple optimization of delay is not sufficient, this has led to the interest in a performance metric called **age of information**. This metric has been proposed to quantify the freshness of status information of a physical process [6, 59]. This has been seen as an end-to-end metric to characterize the latency in status updating systems and applications [110]. In the current generation, IoT sensing services that rely on 5G technology pose their own challenges. The informative content of sensed data changes over time depending on the profile of the IoT sensing service. Information on traffic mobility, for instance, will retain its value on the timescale of the tenths of seconds, whereas temperature and pollution measurements will change in the timescale of the hours. Clearly, managing IoT devices requires a mechanism to control information freshness, the latter being also referred to as the *age of information*. Such a mechanism, known as **aging control**, determines when IoT readings should be uploaded to avoid stale information. Perhaps the earliest use of freshness can be seen in periodic updates in real-time databases, where the concurrency of the computations was enforced by using the age of an update [45], the time-stamped fresh measurements were written into the real-time database by the sensors. Page refresh policies have been adapted to minimize the AoI of cached pages [22]. The work in [58] focused on minimizing the age of safety messages

for connected cars. In contrast to the prior work [45, 22, 48] based on the status update age, [59] looked at the impact of random service times on the age of delivered updates. The focus of the initial work was on evaluating the time-average AoI

$$\langle \Delta \rangle_T = \frac{1}{T} \int_0^T \Delta(t) dt \tag{1.3}$$

As shown in [110], despite the simplicity, the exact analysis of the age can be challenging, which prompted an alternate age metric called peak age of information (PAoI) [25]. Another work [109] on aging control based on a stochastic hybrid system (SHS) [41] provided an alternate approach for average age analysis. In this approach a hybrid state $[q(t), \boldsymbol{x}(t)]$ was considered, where $\boldsymbol{x}(t) \in \mathbb{R}^{1\times n}$ is an age vector and $q(t) \in \mathcal{Q} = \{0, 1, ..., M\}$ describes the discrete state of a network and $\mathcal{Q}$ is a continuous-time Markov chain. As mentioned in [119], there are two broad categories of works, the first category [59, 57, 44, 55, 43, 51, 50] consider that the status packets stochastically arrive at the source node and model this generation process as a queuing system. The works based on first come first serve (FCFS) [59], last come first serve (LCFS) [57] used Queuing theory to analyze and optimize average AoI. Whereas the works in [44, 55, 43], propose scheduling schemes that aim to minimize the average AoI. In the remaining two works [51, 50] propose decentralized scheduling policies with near-optimal performance. The second category of works [101, 12, 106, 10, 29, 19] consider that the status packets can be generated at any time at the source node. Optimal updating policies are proposed to minimize the average AoI in [101, 12] with single and multiple sources, whereas the authors in [106, 10, 29] propose updating schemes for an energy harvesting source to minimize the average AoI. Minimizing the average AoI under resource constraints is the focus of work in [19]. In our work in the first two chapters, we consider a similar approach to that of the second category of works where a fleet of mobile IoT devices collect information among a given number of locations and the information can be collected by the devices as they move from one location to another location. Controlling the age of information dynamics of data carried by IoT devices permits a trade-off between the value of IoT device readings and the cost of uploading them with the 5G IoT slicing service. Hence the problem boils down to making a decision of whether to upload the data or not at a given location. This is a classic sequential decision problem, we use the previously described MDP model to design optimal policies that maximize the service provider's expected average reward. Another aspect of sensing services is the traffic encountered at various locations. There may be a few highly congested locations and some may not have such a level of congestion. So migrating part of the traffic from highly congested locations to other locations may alleviate the burden on infrastructure provider resource capacities.

The two control actions considered in this work are traffic offloading control and aging control. Traffic offloading is a standard networking technique to perform load balancing and avoid traffic congestion. However, in 5G networks, it must work on a per-slice basis and must be made available to SPs in a transparent fashion with respect to InP traffic management tools. Aging control on the

other hand is a key requirement for sensing applications in IoT systems. These two problems have been addressed separately [7, 67, 96, 86, 62], but no prior work has considered the two problems at once to the best of our knowledge. The work presented in chapter 2 and chapter 3 aims to connect these two research lines within the same control framework, resulting in a scheme for the cost-efficient brokerage of IoT data using 5G slicing. While IoT data offloading techniques have been proposed in the context of vehicular networks [86] or sensor networks [62], the proposed solution is tailored specifically to the case of 5G slicing since the SP can stimulate IoT data offload towards less congested areas using a distributed and location-aware scheme which operates at the sensing application level. Furthermore, by means of flexible pricing control, we minimize the cost incurred by the SP in order to lease slice resources from InPs. Finally, the proposed framework includes inherently a notion of service level agreement (SLA) since it is rooted in the concept of AoI which captures latency requirements of IoT data readings.

1.4.2 Market models for resource allocation

Resource allocation is a common problem that appears in a variety of fields such as economics, computers, etc. This problem has been well studied in the field of economics with many works laying the foundation for modern studies. Market models provide a standard technique to study resource allocation under various scenarios. Buyers want to buy the listed goods in exchange for money. Now the problem is to design a mechanism that adequately allocates the goods. The mentioned allocation should be efficient and fair among the buyers.

FISHER MARKET: The general equilibrium model proposed by Irving Fisher is a comprehensive model for computing equilibrium prices [16]. It is one of the fundamental model which has been extensively applied in multi-resource allocation for wireless networks. In this work, we follow the Fiher market model defined in [17]. The general idea behind this model is that there are N buyers, each with a budget B_i, and they try to buy the available resources from a seller. The market $\mathcal{M} := \langle \mathcal{N}, (B_i)_{i\in\mathcal{N}}, \mathcal{R}, (u_i)_{i\in\mathcal{N}}, \boldsymbol{p} \rangle$ is defined as follows,

- Player set: set of buyers $\mathcal{N}$ that are competing to procure resources from the seller.
- Budgets: Budget B_i associated to each buyer i.
- Resource set: Available resources $\mathcal{R}$ for provisioning by the seller.
- Utility: Each user i obtains a utility u_i for procuring an amount of resource from the seller that indicates the level of satisfaction by the user for obtaining a certain resource r.
- Prices: Each resource is assigned with a price p_r depending on the competition.

The objective of this model is that upon receiving the budgets from all the competing users, the seller has to compute a price that adequately distributes the resources among the users. Hence,

the tuple $(\boldsymbol{x}_i, p_r)$ of resource bundle $\boldsymbol{x}_i$ and price play a crucial role in obtaining an optimal resource allocation. The budget associated with each user plays a key role as the allocation of the resources can be impacted by B_i. The budget has multiple interpretations, it can be seen as the amount that a user posses to spend on the resources or it can be seen as market share of a user i. Users with higher budget can procure higher amount of resource, where as the users with lower budget get lower resource in comparison. Here, resource bundle $\boldsymbol{x}_i$ is a vector of resources allocated to user i. One way to obtain an adequate resource allocation is to obtain an optimal tuple $(\boldsymbol{x}^*, \boldsymbol{p}^*)$ that corresponds to a market equilibrium (ME), which is defined as follows,

Definition 1.1. *Allocation and price vector $(\boldsymbol{x}^*, \boldsymbol{p}^*)$ is called as Market Equilibrium (ME) of market $\mathcal{M}$ if the following conditions are satisfied.*

C1 Each $i \in \mathcal{N}$ SP gets his favourite bundle x_i^*,where

$$x_i^* : \underset{x_i \geq 0; C(\boldsymbol{x}_i) \leq B_i}{argmax} \; u_i(x_i) \tag{C1}$$

C2 The demand $\boldsymbol{x}^*$ meets the supply or the market is cleared, i.e.,

$$\sum_{i \in \mathcal{N}} x_{ir}^* \leq C_r \quad \forall r \in \mathcal{R} \tag{C2}$$

and the inequality (C2) is saturated if $p_r > 0$.

Where, $C(\boldsymbol{x}_i)$ is the cost for obtaining the resource bundle $\boldsymbol{x}_i$ and c_r is the available capacity for resource r. Condition (C1) ensures that the allocated resource bundle provides each user with maximum utility where the cost of such bundle does not exceed the available budget. On the other hand condition (C2) warrants that all the users get their requested resource or the resource is completely sold.

With this set of conditions for allocation, previous works employ [17, 79] a centralized approach that propose an optimization problem to obtain an optimal allocation among the competing buyers.

$$P_{SW} : \quad \underset{\boldsymbol{x}}{\text{Maximize}} : \quad \sum_{i \in \mathcal{N}} B_i U(x_i) \tag{1.4}$$

$$\sum_{i \in \mathcal{N}} x_{ir} \leq c_r, \quad \forall r \in \mathcal{R}. \tag{1.5}$$

Where the objective function (1.4) is an aggregated utility functions weighted by the budget of a

user i, and the utility function is considered to be α-fair allocation rule and it is defined as follows

$$U(y) = \begin{cases} \frac{(y)^{1-\alpha}}{(1-\alpha)} & \text{if } \alpha \neq 1, \\ \log(y) & \text{if } \alpha = 1. \end{cases} \tag{1.6}$$

Depending on the value of the parameter α, different fairness can be achieved among the competing users. Given that each user is associated with a budget B_i, when $\alpha = 0$, it may be possible that the user with slightly higher budget can result in allocating the entire resource to that user and there may be starvation for other users. For this reason, it is preferred to implement some level of fairness to prevent any user from starvation. When α = 1, the corresponding optimization problem is known as Eisenberg-Gale program [28] . And it has been proven that the optimal solution for this problem is an exact ME[17]. Given that the objective function is a concave function, gradient projection method can be used to find an optimal solution[15]. It is shown that setting α to 1 leads to the fairness function in [77] referred to as proportional fairness, which is an intermediate choice between the aforementioned extreme cases. Assuming that $\boldsymbol{x}^*$ is a feasible solution (5.1)-(5.2), an allocation $\boldsymbol{x}^*$ is said to be a proportionally fair allocation, if the aggregate of proportional change with respect to any other feasible allocation $\boldsymbol{x}$ is negative, i.e.,

$$\sum_{i \in \mathcal{N}} \frac{x_i^* - x_i}{x_i^*} \leq 0. \tag{1.7}$$

However, one of the limiting factors of proportional fairness is that this leads to linear pricing which is often not a practical approach in the real world applications. For this reason non-linear pricing scheme has been proposed by [34]. Other fariness schemes that are covered under α-fair allocation rule are detailed in chapter 4. In this work, we use this fisher market approach to propose non-linear pricing scheme for an optimal resource allocation among service providers competing for multiple resources. Resource allocation based on this central approach has been significantly employed in wireless networks, however it has severe drawback as the central entity requires that service providers to reveal their utility functions which are sensitive for them. For this reason, decentralized resource allocation methods have been investigated. One of prominent works that propose local algorithms for service provider to obtain an exact optimal allocation achieved by the central approach has been proposed by Frank Kelly [60].

KELLY MECHANISM: Frank Kelly proposed an alternate model [60] that aims in achieving the optimal resource allocation for a single resource in decentralized manner to comply with the privacy aspect of the SPs. In this model, instead of traditional centralized approach that require the SPs to reveal their utility functions, a decentralized allocation mechanism is proposed by dis-aggregating the Main optimization problem into two sub optimization problem solved by the InP and SPs respectively.

The objective Kelly mechanism [60] is to find an optimal rate allocation for communication net-

works, which is defined by following optimization problem.

$$\underset{\boldsymbol{x}}{\text{maximize}} \quad \sum_{i\in\mathcal{N}} U_i(x_{ir}), \tag{1.8}$$

$$\text{subject to} \quad \sum_{i\in\mathcal{N}} x_{ir} \leq c_r \quad \forall r \in \mathcal{R}, \tag{1.9}$$

$$x_{ir} \geq 0; \quad \forall i \in \mathcal{N}, \quad \forall r \in \mathcal{R}. \tag{1.10}$$

It can be observed that solving above concave optimization problem require that the SPs reveal the utility functions, which is a sensitive information. For this reason, two sub-optimization problems are defined for InP and SP respectively that are solved concurrently with exchange of non sensitive information.

Each user solves the following optimization problem:

$$\underset{\boldsymbol{b}}{\text{maximize}} \quad U_i\left(\frac{b_{ir}}{\phi_r}\right) - \sum_{r\in\mathcal{R}} b_{ir}, \tag{1.11}$$

$$\text{over} \quad b_{ir} \geq 0; \quad \forall i \in \mathcal{N}, \quad \forall r \in \mathcal{R}. \tag{1.12}$$

Where b_{ir} is the bid value of SP i with an interpretation of market power or monitory value depending on the scenario, ϕ_r is the price associated with the resource. User computes an optimal bid and communicates it to the InP. Then, InP solves the following optimization problem:

$$\underset{\boldsymbol{x}}{\text{maximize}} \quad \sum_{i\in\mathcal{N}} b_i \log(x_{ir}) \tag{1.13}$$

$$\text{subject to} \quad \sum_{i\in\mathcal{N}} x_i \leq c_r, \tag{1.14}$$

$$x_{ir} \geq 0; \quad \forall i \in \mathcal{N}. \tag{1.15}$$

Now, InP sends the price ϕ_r and allocation vector $\boldsymbol{x}$ to the users, now the users compute new bid b_{ir} and communicates it to the InP. This iterative process converges to an equilibrium beyond which no SP has no benefit in changing the bid. It is shown in [60] that the resulting price ϕ_r and allocation vector $\boldsymbol{x}$ solves the system optimization problem defined in (1.11)-(1.12). R. Johari extended this mechanism for multi-resource context for the communication networks in [53]. We extend this work to the context of multiple resource types, each supported by a different InP with multiple SPs submitting a request for available resource types. We employ an α-fair allocation rule at the system level and allow each InP to implement its own allocation rule. In addition, we propose two separate

algorithms to obtain the desired results.

1.5 Chapters organization

As mentioned earlier, this thesis consists of two major works, the first part focuses on the operational aspect of the data collection services, and the second part focuses on the resource allocation aspect.

- We begin with chapter 2 that focuses on the primary work that consists of the decision-making problem that is concerned with the age of information. This is an MDP problem, we propose a threshold-based policy to address this problem. Then, we focus on traffic offloading by formulating an optimization problem that accounts for traffic offloading as well as the age of information.
- This problem turns out to be an NP-hard problem, we propose a heuristic solution based on simulated annealing in chapter 3. We describe several algorithms that solve the given combinatorial optimization problem and speed up the convergence by exploiting the neighborhood structure of the locations.
- Chapter 4 describes the central resource allocation mechanism. In this chapter, we focus on an allocation model that fairness across service providers and locations as well. This multi-level allocation model focuses on more general pricing that is non-linear pricing instead of the linear pricing mechanism described in many works. We consider α - fairness to formulate the optimization problem. We study the strategic aspect of the SPs at the end to understand the impact of the selfish behavior of SPs on social welfare.
- There are several drawbacks to centralized solutions such as a single point of failure and the necessity of revealing private information by both the SPs and InPs. In chapter 5, we propose another mechanism that is based on a distributed approach where each SP locally solves an optimization problem to increase the value of obtained resources. There is no need for either the InP or the SP to reveal sensitive information to obtain a solution. The optimal solution is achieved with an iterative bidding mechanism. It turns out that this optimal solution solves the social welfare function as well. And finally, this manuscript ends with a conclusion and some insights into future work.

Chapter 2

Aging control

2.1 Introduction

Data collection at scale represents the key signature of future IoT applications, posing significant challenges in the integration of emerging 5G networks and IoT technologies as identified in early studies [71]. In fact, pervasive object readings will play a decisive role in the context of smart cities for both process monitoring and management [63]. Using IoT, a whole new set of applications will be able to feed local information generated by both objects and mobile devices into their databases. Such information streams are consumed for management and prediction purposes by services such as city air management, smart waste management or traffic management, and demand-response schemes [113]. Data brokerage is thus emerging as one of the most interesting business opportunities: new Service Providers (SP) in 5G networks can seize the opportunity to mediate between companies purchasing IoT data and device owners. This is considered a cornerstone in creating a marketplace for IoT data [72, 82, 11, 89] which is essential for the uptake of smart city services.

The architecture of IoT networks must be able to support local data streams from highly heterogeneous information sources, including e.g., meters for water and electricity management, outdoor and indoor positioning data, parking presence sensors, and a whole new set of user-generated contents related to mobile application-specific data. Indeed, long-standing problem of integrated architectures and protocols to support IoT data collection appears finally solved by the uptake of 5G connectivity [23]. Slicing techniques offered by 5G technology allow Infrastructure Providers (InP) to offer differentiated services to their customers using shared resource pools. A slice for IoT services, in this context, is a share of mobile network infrastructure obtained by forming a logical network on top of the physical one connecting IoT devices (Fig.2.1). More generally, traffic differentiation in 5G systems can be obtained by isolating specific traffic categories within slices, which in turn can be dedicated to serving target verticals under specific service isolation guarantees [114, 88]. Smart city services, where SPs support IoT data readings from mobile sensing devices are a key use case of *slicing service*. In this context, the role of the SP is to lease resources (radio, processing, storage, etc.) in the form of one or more dedicated slices and from one or multiple InPs; the leased slice will support the connectivity of a fleet of devices taking part in the IoT sensing services with a cost for the upload of sensed data.

The costs incurred by sensing services depend on a number of factors, including the business

model and the ownership of the sensing devices. It is possible that the sensing devices are owned by the SP, whereas the sensing services are designed and run by third parties and offered, e.g., as a smartphone application. In this case, the sensing services involve payments to the SP [49]. If the SP is in charge of the sensing services and also the slicing services, in turn, non-monetary costs – similar to the shadow prices defined in [60] – can be used effectively as a penalty to avoid hot-spot phenomena by deterring the upload of sensed data in congested areas.

IoT sensing services relying on 5G technology pose their own challenges. The informative content of sensed data changes over time depending on the profile of the IoT sensing service. Information on traffic mobility, for instance, will retain its value on the timescale of the tenths of seconds, whereas temperature and pollution measurements will change in the timescale of the hours. Clearly, managing IoT devices requires a mechanism to control information freshness, the latter being also referred to as the *age of information* (AoI). Such mechanism, known as aging control, determines when IoT readings should be uploaded to avoid stale information.

Controlling the AoI dynamics of data carried by IoT devices permits a trade-off between the value of IoT device readings – indeed specific to a tagged service – and the cost for uploading them with the 5G IoT slicing service. Motivated by the aging control problem intrinsic to IoT devices, and by the traffic offloading capabilities enabled by 5G technology, we investigate the following two questions in chapter 2 and chapter 3 respectively:

1. given the requirements of a tagged *IoT sensing service* and the SP charging rates, what is the optimal upload strategy to control information freshness at the device level?

2. how should the SP incentivize users to offload IoT data in order to reduce the costs to lease the resource slice?

In this chapter, we address the first question via the control of AoI at the device level, and an optimal upload strategy is derived. Sensing devices trigger the upload of sensed data depending on two factors: the application profile and the price for the IoT sensing service. It is the application profile that determines for how long sensed data retains their value, whereas location-dependent prices determine the unit cost of sensed data uploads performed using the IoT slice. The problem is formulated as a Markov decision process (MDP). The optimal stationary policy solving the problem has the multi-threshold structure: the upload of information occurs depending on the upload prices available to a tagged device, i.e., prices available in the cell it is connected to, and on the AoI relative to the data stored in the device memory. The next chapter addresses the minimization of slicing service costs: the SP optimizes the vector of prices that are exposed to devices with the aim to minimize the cost paid to the InP for leasing the slice while satisfying the applications' delay target.

Following tables summarize all the notations, variables that are used in chapter 1 and chapter 2.

Table 2.1: Table of notations: Basic parameters

Notation	*Description*
$\mathcal{L}$	set of regional locations; $\mathcal{L} = \{1, \ldots, L\}$
$\mathcal{P}$	set of upload unit prices; $\mathcal{P} = \{p_1, \ldots, p_K\}$
$\mathcal{M}$	set of information age values; $\mathcal{M} = \{1, \ldots, M\}$
D_i	amount of data generated at location i during a time slot
Π_{ij}	transition probability from location i to j
${}_A\lambda_{ij}^n$	probability of moving from location i to j in n steps without entering taboo set A
π_i	occupation probability of location i
B_i	maximum bandwidth at location i
$\mathbf{B}$	Maximum bandwidth vector
C_i	monetary unit cost to lease bandwidth at location i
d	target latency
ϵ_i	tolerance factor at location i
N	number of IoT devices
F	average size of the collected data
κ	timeslot duration (seconds)

Table 2.2: Table of notations: States, actions, transitions and rewards

Notation	*Description*
t	current timeslot
x_t	age of information at time t
l_t	location at time t
s_t	state at time t; $s_t = (x_t, l_t)$
a_t	action at time t, where 1 means upload, and 0 defer
$\mu(x, l)$	function expressing the probability that the device performs action $a = 1$ in state $s = (x, l)$
$\Gamma_{s,a,s'}$	transition probability from s to s' under action a
$r_t(s_t, a_t)$	instantaneous reward under state action pair (s_t, a_t) at time t

Table 2.3: Table of variables

$\Delta_i(\boldsymbol{p})$	random variable characterizing age of information at upload time for data collected at location i
F_{Δ_i}	probability distribution (CCDF) for the age of information at upload time for data collected at location i
$Y_{ij}(\boldsymbol{p})$	average traffic rate for data collected at location i and uploaded at location j
y_{ij}	average traffic rate for data collected at location i and uploaded at location j, per device
$f(i,j,t;\boldsymbol{p})$	probability that a device collects data from location i and upload it at time t in location j
$\mathcal{U}$	set of prices corresponding to locations wherein the optimal policy is to upload
K	number of threshold values in the current multi-threshold policy
$\tau^{(j)}$	j^{th} AoI threshold value, $\tau^{(1)} = 0$, $\tau^{(j)} \leq \tau^{(j+1)}$, and $\tau^{(K)} \leq M$ (for convenience, $\tau^{(K+1)} = M$)
τ_l	AoI threshold for data collected at location l
$\boldsymbol{\tau}$	AoI threshold vector (one threshold per location); $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_L)$
$\boldsymbol{\tau}_{max}$	AoI threshold vector with all values equal τ_{max} (maximum achievable AoI)

Prior art and main contribution. The two control actions considered in this work are traffic offloading control and aging control. Traffic offloading is a standard networking technique to perform load balancing and avoid traffic congestion. However, in 5G networks, it must work on a per slice basis, and must be made available to SPs in a transparent fashion with respect to InP traffic management tools. Aging control on the other hand is a key requirement for the sensing applications in IoT systems. These two problems have been addressed separately [7, 67, 96, 86, 62], but no prior work has considered the two problems at once to the best of the authors' knowledge. The work presented in chapter 2 and chapter 3 aims to connect these two research lines within the same control framework, resulting in a scheme for the cost-efficient brokerage of IoT data using 5G slicing. While IoT data offloading techniques have been proposed in the context of vehicular networks [86] or sensor networks [62], the proposed solution is tailored specifically to the case of 5G slicing since the SP can stimulate IoT data offload towards less congested areas using a distributed and location-aware scheme which operates at the sensing application level. Furthermore, by means of flexible pricing control, we minimize the cost incurred by the SP in order to lease slice resources from InPs. Finally, the proposed framework includes inherently a notion of service level agreement (SLA) since it is rooted in the concept of AoI which captures latency requirements of IoT data readings.

Chapter organization. The remainder of this work is organized as follows. A detailed system description is reported in Sec. 2.3. Then, the two main control actions, namely, traffic offloading control and aging control are discussed in Sec. 2.4 and in Sec. 2.5, respectively. Sec. 2.6 bridges the two pillars with a unified framework. We describe the numerical results in Sec. 2.7 to validate the theoretical findings. The algorithms based on simulated annealing to control the said price are described in the next chapter.

2.2 Related work

Most related works tackle either the control of AoI in IoT networks or traffic offloading for 5G networks.

Aging control for IoT. Aging control is at the core of IoT sensing applications, as it captures the trade-off between data staleness and resources utilization. Given the increasing demand for IoT systems, the literature on control of AoI is correspondingly growing. Most of the work on information aging control focus on users' standpoint, accounting for costs as perceived by the devices whose AoI is under control. Connections of AoI with traffic offloading are typically analyzed in the literature as a downstream effect of aging control. The potential relationship between AoI and traffic offloading has been signaled in [7]. With previous works mostly focusing on computation and task offloading [65, 67, 96] rather than traffic offloading. In this paper, in contrast, we have considered jointly aging control and traffic offloading as first class citizens of an ecosystem wherein users and providers interact. In our scheme SPs influence users via pricing mechanisms able to couple aging control and traffic offloading in a unified framework.

Traffic offloading and slicing in 5G networks. Utility service providers have long performed IoT data collection to reduce operational costs. Such traditional schemes are typically based on M2M to match the requirements of proprietary SCADA systems and charged per message. Nowadays, they appear inadequate for emerging IoT systems. In fact, the second major driver of the 5G technology, beyond multimedia traffic, is the current growth of mobile IoT connections [23]. Actually, with both new LTE-M radio interfaces and the new suite of architectural paradigms, 5G introduces key infrastructural assets able to ease both IoT access to radio resources and computing at the edge of the network. Traffic offloading is enabled by 5G technology through slicing. Technical aspects such as slice insulation and fair slice allocation are still under development to upgrade LTE technology towards 5G, with large effort by the research community to overcome such technical issues [90][52][117]. Nevertheless, slicing techniques are currently under standardization: specifications of the 5G system's slicing architecture and its requirements are available [3]. In future 5G networks virtual private networks for IoT data collection will be shipped to SPs on top of the existing mobile network infrastructure with InP dedicated customer support. The traffic offloading mechanisms proposed in this work can be used by any such SP at the slice level for cost minimization purposes.

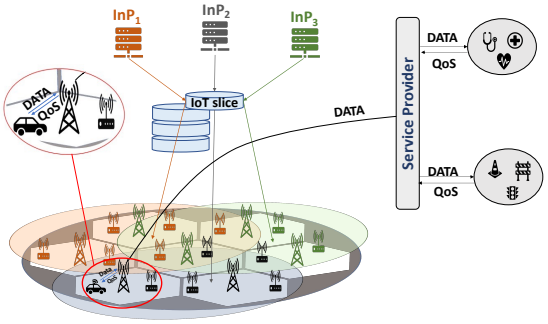


Figure 2.1: Cutting slices of resources across multiple locations and across multiple InPs.

2.3 System description

A *SP* offers internet connectivity to heterogeneous IoT devices over a physical region (Fig.2.1). The SP can act as a data broker, i.e., collects data from device owners or from mobile IoT devices deployed across the region and sell the data to interested parties under Data as a Service (DaaS) scheme or use the collected data for own service. To that aim, a single SP can aggregate resources leased from various available InPs at different locations. Each InP provides dedicated 5G slices for IoT data collection at certain cost. In practice, sensed data is relayed using a fleet of mobile devices uploading them at the need while mobile relays are served through resources across a pool of base stations covered by the selected InPs infrastructures.

Because sensed data belongs to a variety of categories, e.g., healthcare data, environmental monitoring data, road traffic data, etc., it has different time sensitivity. SP customers will require brokered IoT data to comply with certain QoS requirements. Throughout this work, the latency of delivered IoT data is the reference SLA metric (indeed it is a fundamental parameter for, e.g., industrial automation, intelligent transport systems, and healthcare monitoring applications). Latency, in turn, is impacted by the locations from which mobile devices upload sensed data. Note that aggregated traffic may vary significantly across regional locations, e.g., due to the presence of hotspots. Ultimately, the SPs need to grant target QoS figures for a given IoT application and obey standardized SLA. To this aim, the key enabler is 5G network slicing by which the SP negotiates and adjusts the scale of bandwidth and service functions. In practice, this entails orchestrating slicing functionality across heterogeneous access technologies (5G, LTE, 3G, and WI-FI), over different site types (macro, micro, and pico base stations) and over multiple InPs. The cost of leased infrastructures depends on chosen InPs, specific access technology supported across regional sites, and IoT data collection patterns. For the sake of clarity, only bandwidth costs are considered, but the whole framework may include other costs for local computation and/or storage [3] as well.

In order to comply with SLA agreements for IoT data collection, the SP dynamically determines the resources per slice required to match the current demand. Due to scarcity of resources, higher

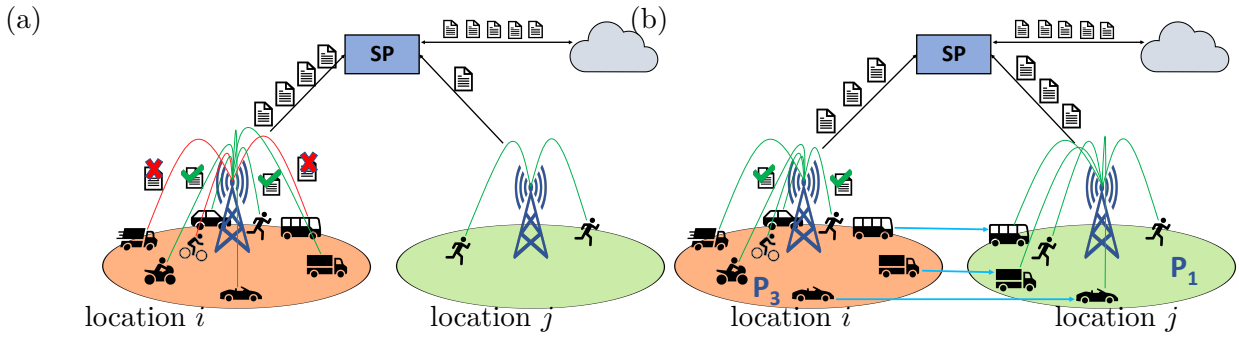


Figure 2.2: (a) Without traffic offloading mechanism, (b) With traffic offloading mechanism.

costs will be incurred in crowded and congested locations. Hence, the SP designs an *IoT data collection policy* by which freshness of IoT data is traded off against costs. In fact, the upload of non-critical data can be deferred to occur at a location with smaller costs yet complying with SLAs, i.e., target latency figures.

The key mechanism detailed in the next section is a price-based load balancing scheme where the SP incentivizes users not to upload data from congested locations. Prices are dynamically set, e.g., based on the congestion levels. Different locations are tagged by a price to upload a unit of IoT data. The whole scheme takes advantage of user mobility: while IoT devices are carried by users appliances and move across the regions, data upload can be diverted towards less congested locations. Through pricing, the SP can shift the IoT traffic generated by mobile IoT devices from locations where data are sensed to less crowded ones, where leased slice resources are relatively cheaper. Following example provides the intuition behind the traffic offloading mechanism,

Fig. 2.2(a) shows a scenario that has no control mechanism for traffic upload. As a result, up-link traffic is high in location i and SP may lose part of the data due to outage. In addition, upload costs would be higher as well. In Fig. 2.2(b), in contrast, upload traffic is well distributed. A traffic offloading mechanism is adopted, and there is no overburden on any particular access point, hence the data is not lost due to outage. This avoids re-transmissions further congesting the area.

In the following sections, a mechanism that combines the aging control and traffic offloading is introduced. This mechanism, deployed at each device, is designed based on the aggregated mobility of the devices. Nonetheless, if the individual mobility of devices is available, the mechanism can also leverage this information, noting that the objective of the device is to find an optimal strategy to upload the collected data. If decisions are made per device, using individual mobility patterns, heterogeneity across devices does not impact the upload decisions of each device.

2.4 Traffic Offloading

In this section, the SP pricing scheme which is used to minimize the total cost incurred by the SP is formulated. The resulting control problem accounts for the mobility pattern of devices and the delay requirements of the IoT data collection service. At each location, the SP selects the corresponding InPs. Let $\mathcal{L} = \{1, 2, .., L\}$ be the set of locations. Let B_i be the maximum bandwidth available at location i, $1 \leq i \leq L$, resulting in a maximum bandwidth vector $\boldsymbol{B}$. Each location is tagged with a unit price, resulting in a price vector $\boldsymbol{p} = (p(i), i \in \mathcal{L}) \in \mathbb{R}^L$. The price vector induces a set of K different prices denoted by $\mathcal{P}_K$, $\mathcal{P}_K = \{P_1, .., P_K\}$, where $P_1 < \ldots < P_K$. Prices impact location-dependent upload policies which determine when a mobile IoT device should upload sensed data, based on the current age of information. The age of information represents the time elapsed from sensor reading until upload. Let $\Delta_i(p)$ be the random variable representing the age of information based on the price vector – at upload time – for data collected at location i. Let C_i be the monetary unit cost to lease bandwidth at location i. Let D_i be the amount of data generated by devices at location i during a time slot. Finally, the upload control is represented by variable $Y_{ij}(\boldsymbol{p})$ which is the average traffic rate for data collected at location i and uploaded at location j.

Each SP face with the following optimization problem, named here as TRAFFIC OFFLOADING:

$$\underset{\boldsymbol{p}}{\text{minimize}} \sum_{i \in \mathcal{L}} \sum_{j \in \mathcal{L}} C_j \, Y_{ij}(\boldsymbol{p})$$

subject to

$$\sum_{j \in \mathcal{L}} Y_{ji}(\boldsymbol{p}) \leq B_i, \ \forall i \in \mathcal{L} \tag{2.1}$$

$$\sum_{j \in \mathcal{L}} Y_{ij}(\boldsymbol{p}) = D_i, \ \forall i \in \mathcal{L} \tag{2.2}$$

$$\mathbb{P}(\Delta_i(\boldsymbol{p}) > d) \leq \epsilon, \ \forall i \in \mathcal{L} \tag{2.3}$$

$$Y_{ij} \geq 0, \ \forall i, \forall j \in \mathcal{L} \tag{2.4}$$

In this problem, eq. (2.1) is the per location constraint on the available bandwidth for the IoT slice, and eq. (2.2) is concerned with the flow conservation constraint. Constraint (2.3) provides a tunable SLA constraint on the age of information collected at specific location $i \in \mathcal{L}$, depending on a target latency value $d > 0$ and on tolerance $\epsilon > 0$.

The main challenge to solve the TRAFFIC OFFLOADING problem is to account for the mobility pattern of devices. In fact, they collect data at some tagged location, and they upload it according to the chosen policy, in order to meet QoS requirements. In practice, once a sensing device is associated with a tagged location, it will be informed of a price available for the IoT slicing service, so that the decision to upload or not can be implemented onboard of sensing devices in a fully

distributed fashion. An algorithm that is able to determine the optimal price vector $\boldsymbol{p}$ solving the traffic offloading problem is described in the next chapter. First optimal upload control at the device for a given price vector $\boldsymbol{p}$ should be analyzed properly to study the algorithms that solve the combined problem.

2.5 Aging Control

Each device decides to upload data or defer based on its actual location, the vector of prices, and the age of information. Let x_t be the age of information for data collected at time t by a tagged device: $x_t = 1$ when the device collects it, and increases by one at every time slot, except when the device uploads data or the collected data reaches the maximum age, denoted by M. Note that M is a design parameter, assumed to be fixed and given. Let $\mathcal{M} = \{1, \ldots, M\}$ so that $x_t \in \mathcal{M}$. Let $U(x)$ be the utility corresponding to uploading data with age of information x, where $U(\cdot)$ is a non-increasing function. The selection of the utility function is up to the SP. For example, if the SP wants to collect data concerning traffic updates, the value of the information may decrease exponentially fast. For pollution level updates, in contrast, the value may not decrease as fast.

The state of a tagged device at time t is denoted $s_t = (x_t, l_t)$, where x_t is the AoI as described in the above paragraph and l_t is the device's location at time t. Location $l_t \in \mathcal{L}$ is the state of a finite, discrete, ergodic Markov chain, whose dynamics determine the mobility pattern. Let the transition probability between location l and k denoted by λ_{lk}; $\Lambda = \{\lambda_{lk}\}$, is the corresponding transition probability matrix. Finally, let $\boldsymbol{\pi} = [\pi_1, \pi_2, ..\pi_L]$ be the steady state probability distribution.

The action set available at each device is to upload or defer, i.e., $A = \{0, 1\}$, where 0 means "defer" and 1 "upload"; the action taken at time t is denoted by a_t. Hence the dynamics of the age of information at a tagged device is given by

$$x_{t+1} = \begin{cases} 1, & \text{if } a_t = 1, \\ \min(x_t + 1, M), & \text{if } a_t = 0. \end{cases}$$

Next, the transition probability of the resulting MDP is characterized. Let $s = (x, l)$ be the current state of the device and let $s' = (x', l')$ be its next state under action a. The transition probability from s to s', under action a, is given by

$$\Gamma_{s,a,s'} = \begin{cases} \lambda_{l,l'}, & \text{if } x' = \min(x+1, M) \text{ and } a = 0 \\ & \text{or } x' = 1 \text{ and } a = 1, \\ 0, & \text{otherwise.} \end{cases} \tag{2.5}$$

Instantaneous reward. The instantaneous reward under the state action pair (s_t, a_t) at time t, $r_t(s_t, a_t)$, is

$$r_t(s_t, a_t) = U(x_t) - p(l_t) \cdot a_t. \tag{2.6}$$

Upload policy. The upload policy μ for a tagged device is a probability distribution over the action space. The rest of the discussion is restricted to stationary policies; since the action space is a binary set, a policy simplifies into function $\mu = \mu(s)$ expressing the probability the device performs action $a = 1$ in state s.

Problem statement: The objective of each device is to maximize the expected average reward:

$$\text{AGING CONTROL: } \max_{\mu} \mathbb{E}[r, \mu] \tag{2.7}$$

$$\mathbb{E}[r, \mu] = \lim_{\eta \to \infty} \frac{1}{\eta} \sum_{t=0}^{\eta-1} \mathbb{E}[r_t(x_t, l_t, a_t); \mu]$$

Since the service provider (SP) aims to promote uploads as much data as possible at locations with smaller cost, it is natural to consider the smallest price to be $P_1 = 0$. As a consequence, in any optimal strategy the devices will upload immediately their collected data at locations with price $P_1 = 0$. In [75], All the results are extended for $P_1 > 0$. For this scenario, the instantaneous reward can be expressed as

$$r_t(s_t, a_t) = U(x_t) - (p(l_t) - P_1) \cdot a_t - P_1 a_t \tag{2.8}$$

Note that the value P_1 can be interpreted as the energy cost of each uploaded message. In the remainder of this work, and without loss of generality, we assume $P_1 = 0$.

The optimal control policy is characterized as follows that solves (2.7). A special type of strategy is introduced, referred to as a *multi-threshold strategy*.

Definition 2.1 (Multi-threshold strategy)**.** *A multi-threshold strategy is such that there exists K and threshold values $\tau^{(j)}$, $j = 0, \ldots, K-1$ such that $\tau^{(1)} \leq \tau^{(2)} \leq \ldots \leq \tau^{(K)} \leq M$ and*

$$\mu(x, l) = \begin{cases} 1 & \text{if} \quad x \geq \tau^{(j)} \text{ and } p(l) \leq P_j \\ 0 & \text{otherwise} \end{cases}$$

Note that K is the number of thresholds, and $\tau^{(1)}$ and $\tau^{(K)}$ are the minimum and maximum threshold values.

A device using this multi-threshold strategy uploads the collected data at its current location l at price $p(l) = P_j$ if the age of information exceeds $\tau^{(j-1)}$. The following theorem reduces the problem of finding the optimal strategy for the AGING CONTROL problem to the one of finding the K thresholds $\tau^{(j)}$, $j = 1, \ldots, K$.

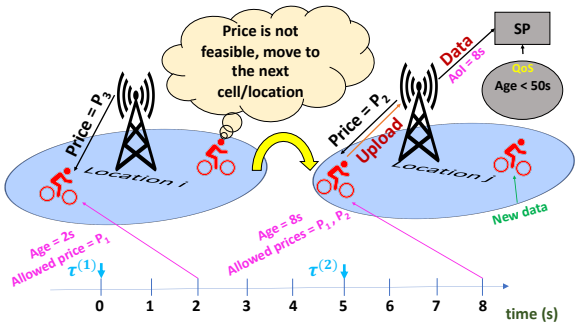


Figure 2.3: The upload mechanism: depending on the AoI, the upload decision is taken based on the shadow price value at the current location.

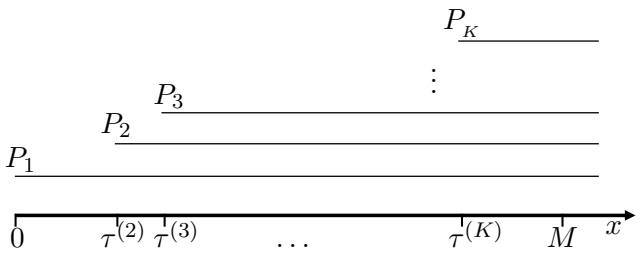


Figure 2.4: Structure of the multi-threshold policy; at the increase of the age of information upload action is optimal for an increasingly larger set of prices.

Theorem 2.2. *The optimization problem* (2.7) *admits a unique deterministic optimal multi-threshold strategy.*

Proof. See Appendix B □

Some further properties of the optimal thresholds are characterized as follows. A qualitative description of the behavior of the optimal policy is depicted in Figs. 2.3 and 2.4. Observe that a multi-threshold strategy is a simple procedure to implement the distributed IoT upload control. In practice, when data is stored on a device, the AoI is one. Thus the device at the beginning will start by uploading only at locations where the price is $P_1 = 0$, noting that $\tau^{(1)} = 0$. Whenever AoI reaches $\tau^{(2)}$, i.e., $x \geq \tau^{(2)}$, the device switches to a second phase wherein an upload occurs if prices are less than or equal to P_2, that is either in locations with corresponding prices P_1 or P_2. Similarly, once a new threshold is reached, say $\tau^{(j)}$, the device will upload the collected data at locations with a price less than or equal to P_j.

Illustrative example. Figure 2.3 displays a simple illustration of the multi-threshold policy, wherein a device attached to a bicycle enters a location i where the price is P_3 and the age of the information is such that it can only upload if price is P_1. After displacement, the device enters a new location j with a tagged price of P_2. By this time, the age is higher than the threshold $\tau^{(2)}$, allowing the device to upload information with either P_1 or P_2. Now, the device can upload the information

as the price is acceptable.

As an immediate consequence of the proof of the previous theorem, the following corollary is obtained.

Corollary 2.3. *At any location l, $\mu(x,l)=1$ if $x \geq \tau^{(K)}$.*

The above corollary implies that the maximum age that can be reached by a message is $\tau^{(K)}$, where $\tau^{(K)} \leq M$.

In general, the set of locations where a device is allowed to upload data, as well as the age of information when the upload action is performed, depends on the distribution of the prices across the set of locations $\mathcal{L}$ used for the slice leased by the SP. Such distribution can be optimized to reduce the cost of infrastructure utilization and yet satisfy the QoS requirements of the IoT service. In the next section, the dynamics of AoI, the structure of the multi-threshold strategy, and the distribution of the prices are connected. Before that further characterization of additional properties of the multi-threshold policy is needed. In particular, a key step is to characterize the number of prices that the optimal threshold strategy uses with positive probability.

Let $\mathcal{L}_i = \{l \in \mathcal{L} | p(l) \leq P_i\}$ and $K_{lP_i} = \sum_{l' \in \mathcal{L}_i} \lambda_{ll'}$. In addition,

$$\mathcal{S}(i) = \sum_{x=2}^{M} (U(x) - U(M))(1 - K_{lP_i}) + U(1) - U(M) \tag{2.9}$$

and

$$\bar{K}_{lP_i} = K_{lP_i} - K_{lP_{i-1}}. \tag{2.10}$$

Theorem 2.4. *Let $\mathcal{U}$ be the set of prices corresponding to locations wherein the optimal policy is to upload. Then,*

- *$\mathcal{U} = \{P_1\}$, with $P_1 = 0$, if and only if*

$$\mathcal{S}(1) < p(l), \quad \forall l \in \mathcal{L}/\mathcal{L}_1, \tag{2.11}$$

- *$\{P_i\} \subseteq \mathcal{U}$, with $P_i \neq 0$, if and only if*

$$U(1) - (\bar{K}_{lP_i} U(2) + (1 - \bar{K}_{lP_i}) U(M)) > P_i \tag{2.12}$$

$$\mathcal{S}(i) < p(l), \quad \forall l \in \mathcal{L}/\mathcal{L}_i, \tag{2.13}$$

- *$\{P_i, P_{i+1}, .., P_{i+k}\} \subseteq \mathcal{U}$, if and only if condition (2.12) is met and*

$$\mathcal{S}(i+k) < p(l), \quad \forall l \in \mathcal{L}/\mathcal{L}_{i+k}. \tag{2.14}$$

Proof. See Appendix B □

Theorem 2.4 establishes conditions under which devices upload data if and only if they are found in a given finite set of locations. An optimal pricing assignments minimizing SP costs while still satisfying users QoS requirements is derived in the following section, for a given assignment of price to locations.

2.6 Joint Aging Control and Traffic Offloading

Given that the optimal distributed upload control is determined properly, the next task is to address the TRAFFIC OFFLOADING problem introduced in Sec. 2.4.

2.6.1 Pricing as a tool for joint aging control and offloading

Recall that the SP aims at setting optimally the value of the shadow prices to reduce the total cost to lease resources from different InPs. Assume that N IoT devices spread over the set of locations $\mathcal{L}$. Each device generates data to be collected and sent to the IoT server located in the core network every κ seconds. Let π_j be the ergodic probability of a device collecting data at location j – which in turn depends on the mobility profile of devices. Hence, the total rate of collected data by devices in location j is given by

$$D_j = N\pi_j F/\kappa, \tag{2.15}$$

where F is the average size of the collected data.

First, observe that if shadow prices are constant over locations, i.e., $p(l) = P_1$, for $l \in \mathcal{L}$, each device will transmit immediately the collected data and the total cost for SP is

$$\sum_{j\in\mathcal{L}} C_j D_j = \frac{NF}{\kappa} \sum_{j\in\mathcal{L}} C_j \pi_j. \tag{2.16}$$

The primary interest for the distributed upload control via shadow pricing is to perform load balancing by shifting part of the traffic load from highly congested locations, which are expected indeed to be more expensive to lease, compared to lesser charged locations. At the same time, the aim is to ensure that the QoS requirements of the IoT service are satisfied. Under shadow pricing vector $\boldsymbol{p}$, the total rate uploaded at location j under the optimal threshold strategy is given by

$$Y_j(\boldsymbol{p}) = \sum_{i\in\mathcal{L}} Y_{ij}(\boldsymbol{p}) = \sum_{i\in\mathcal{L}} D_i y_{ij}(\boldsymbol{p}) = \frac{NF}{\kappa} \sum_{i\in\mathcal{L}} \pi_i y_{ij}(\boldsymbol{p}). \tag{2.17}$$

Hence the total cost writes

$$\sum_{j\in\mathcal{L}} C_j Y_j(\boldsymbol{p}) = \frac{NF}{\kappa} \sum_{j\in\mathcal{L}} C_j \left(\sum_{i\in\mathcal{L}} \pi_i y_{ij}(\boldsymbol{p}) \right). \tag{2.18}$$

In what follows, the above equation is leveraged as the objective of the optimization problem.

2.6.2 Formulation of joint offloading and aging control problem

Next, we account for the AGING CONTROL problem introduced in Sec. 2.5 under the TRAFFIC OFFLOADING problem introduced in Sec. 2.4. The resulting joint problem is posed as follows.

JOINT OFFLOADING AND AGING CONTROL (JOAC) :

$$\underset{\boldsymbol{p}}{\text{minimize}} \sum_{i\in\mathcal{L}} \pi_i \sum_{j\in\mathcal{L}} y_{ij}(\boldsymbol{p}) C_j \tag{2.19}$$

subject to

$$\sum_{i\in\mathcal{L}} \pi_i y_{ij}(\boldsymbol{p}) \leq B_j, \ \forall j \in \mathcal{L} \tag{2.20}$$

$$\sum_{j\in\mathcal{L}} y_{ij}(\boldsymbol{p}) = D_i, \ \forall i \in \mathcal{L} \tag{2.21}$$

$$\mathbb{P}(\Delta_i(\boldsymbol{p}) > d) \leq \epsilon, \ \forall i \in \mathcal{L} \tag{2.22}$$

$$y_{ij}(\boldsymbol{p}) \geq 0, \ \forall i, \forall j \in \mathcal{L} \tag{2.23}$$

where y_{ij} is the expected per device upload rate for data collected at location i and uploaded at location j.

$y_{ij}(\boldsymbol{p})$ can be calculated based on the threshold strategy from section 2.5: the probability that a device collects data at location i and uploads it at location j needs to be computed. The calculation is performed by determining $f(i,j,t)$, namely the probability that a device collects data from location i and uploads it at time t in location j. Such computation involves the use of taboo probability, defined as follows:

$$_A\lambda_{ij}^n = \mathbb{P}\left(l_1, .., l_{n-1} \notin A, l_n = j | l_0 = i\right).$$

This is the probability of moving from location i to location j in n steps without entering the taboo set A; such transition probabilities are calculated in the standard way by considering the n-th power of the taboo matrix, which is obtained by zeroing the columns and the rows of the transition probability matrix corresponding to the taboo states, i.e., the states in A. Based on the optimal threshold strategy, if a device collects data from a location $i \in \mathcal{L}_1$, it will immediately upload it. Thus for $i \in \mathcal{L}_1$,

we have

$$f(i,z,t;\boldsymbol{p}) = \begin{cases} 1, & \text{if } z = i \text{ and } t = 1, \\ 0, & \text{otherwise.} \end{cases}$$

For $i \notin \mathcal{L}_1$ and $z \in \mathcal{L}_j$, let us consider $\widehat{\tau}^{(t)} = \max(\tau^{(j)} | \tau^{(j)} < t)$.

The explicit expression can be derived as follows

$$f(i,z,t;\boldsymbol{p}) = 0, \text{ for } t < \tau^{(j)} \tag{2.24}$$

$$f(i,z,t;\boldsymbol{p}) = \sum_{l_1 \notin \mathcal{L}_1} \sum_{l_2 \notin \mathcal{L}_2} \cdots \sum_{l_{t-1} \notin \mathcal{L}_{\widehat{\tau}^{(t)}}} \mathcal{L}_1 \lambda_{il_1}^{n_1} \cdot \mathcal{L}_2 \lambda_{l_1 l_2}^{n_2} \cdots \mathcal{L}_{\widehat{\tau}^{(t)}} \lambda_{l_{t-2} l_{t-1}}^{t-\widehat{\tau}^{(t)}-1} \lambda_{l_{t-1} j}$$

$$\text{for } \tau^{(j)} \leq t \leq \tau^{(K)} \tag{2.25}$$

$$f(i,z,t) = 0, \text{ for } \tau^{(K)} < t \leq M \tag{2.26}$$

where

$$n_k = \tau^{(k+1)} - \tau^{(k)}, \quad k = 1, \ldots, K$$

with $\tau^{(K+1)} = M$. The expression of y_{iz} for $z \in \mathcal{L}_j$ yields

$$y_{iz} = \sum_{t=\tau^{(j)}}^{\tau^{(K)}} f(i,z,t;\boldsymbol{p}). \tag{2.27}$$

Once the values of $f(i,z,t)$ are obtained, the stationary probability distribution for the age of information can be derived– at the upload time – for the data collected at location i, namely $\Delta_i(\boldsymbol{p})$,

$$F_{\Delta_i}(d) := \mathbb{P}(\Delta_i(\boldsymbol{p}) > d) = \sum_{t=d+1}^{\tau^{(K)}} \sum_{z \in \mathcal{L}} f(i,z,t;\boldsymbol{p}). \tag{2.28}$$

Relation (2.28) provides an important measure for SP: it is the probability that an input shadow price vector can meet the requirements for the IoT data collected at a tagged location. Furthermore starting from $F_{\Delta_i}(d)$, it is possible to evaluate the deviation of the age of collected data from its average value, e.g., by using Chebyshev inequality.

Finally, the expected age of collected data from location $i \in \mathcal{L}$ is given by

$$\mathbb{E}[\Delta_i(\boldsymbol{p})] = \sum_{t=0}^{\tau^{(K)}} \sum_{z \in \mathcal{L}} t \cdot f(i,z,t;\boldsymbol{p}) \tag{2.29}$$

The JOAC problem is a constrained non-linear integer valued optimization problem defined over the set of multi-threshold policies. Finding a solution is made difficult because the structure of function Y is not convex over the shadow price vectors $\boldsymbol{p}$. A heuristic algorithm which utilizes the structure

of the devices' optimal strategy to solve the problem are proposed in the next chapter.

2.7 Numerical evaluation

2.7.1 Experimental setup

We use a vehicular mobility traces of the city of Cologne (Germany), covering a region of 400 km^2 in a period of 24 hours in a typical work day, involving more than $700,000$ individual vehicles to validate our theoretical findings. The data set is available at [104]. It comprises the list of users' position records, each record including a sampling timestamp, the user ID, and her position in (x, y) Cartesian coordinates. Positions are sampled each second. User mobility is spanned across 230 (macro) cells and the coverage of each cell is determined a posteriori according to the Voronoi tessellation shown in Fig. 2.5. In order to generate the transition probability across cells, it is restricted to a subset of records corresponding to one hour of trace data. Then, the data set is resampled at 2 second intervals to discretize the process: within such a time step the probability for a user to cross two cells is bounded below by 0.05.

2.7.2 Aging control analysis

In the first set of experiments, we validate the aging control policy on the real-world traces. The computation of the optimal policy using the proposed model requires estimating the transitions of the Markov chain Γ, $U(x)$ and price p_l of each location $l \in \mathcal{L}$. The reference setting comprises of the utility of the message that decays linearly over time, and remains zero afterward, $U(x) = \max(M - x, 0)$. Assume that the device collects data as soon as the existing data is uploaded. Performance of the policy is evaluated over 67 epochs corresponding to a total duration of 134s for a device: at each epoch, a device either uploads or defers based on the multi-threshold policy obtained from the model. All 230 locations are divided into 3 effective groups, namely range 1, range 2, and range 3, each corresponding to a specific cost C_1, C_2, and C_3, respectively, sorted in ascending order. The grouping of locations is based on the congestion level, corresponding to low congestion, medium congestion, and high congestion respectively. The intuition is that the location with a higher cost should be configured with higher price according to the previous analysis. The results displayed in this section are configured with three prices (P_1, P_2, P_3), $\tau^{(1)} = 0$, two effective threshold values $(\tau^{(2)}, \tau^{(3)})$ and M = 10. Locations that belong to range 1, 2 and 3 are assigned $P_1 = 0$, P_2 and P_3, respectively.

Note that we consider the initial price to be zero throughout our model description and numerical experiments. In addition, in these experiments, the maximum price $P_3 = 9$, and the intermediate price $P_2 = 6$. In particular, the values of P_2 and P_3 are chosen according to the experimental goals, namely, to illustrate a threshold policy which is not degenerate to either always transmit or never

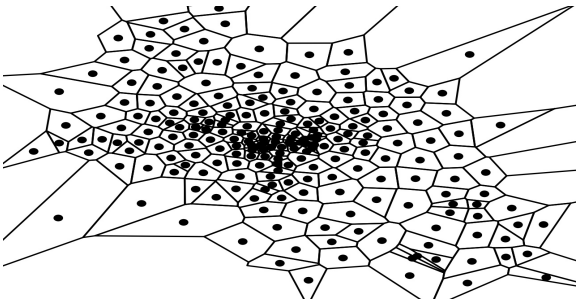


Figure 2.5: Voronoi tessellation for the Cologne mobility trace.

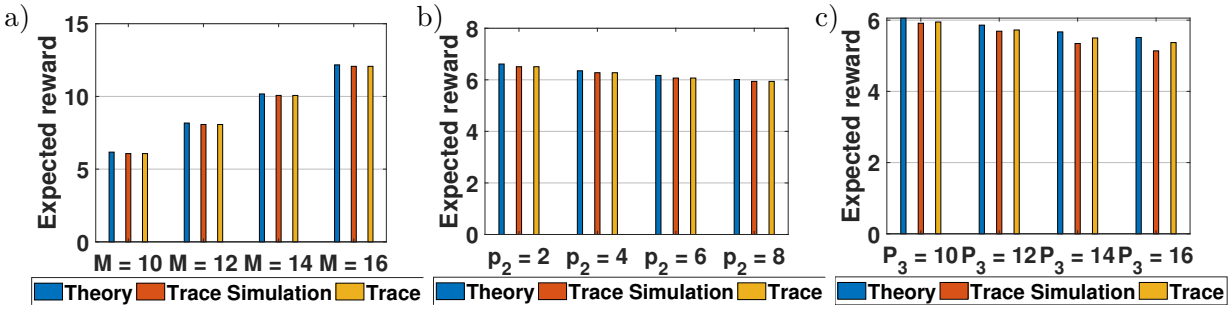


Figure 2.6: (a) Expected average rewards for i. theoretical model, ii. simulation and iii. exhaustive search; $P_1 = 0$, $P_2 = 6$ and $P_3 = 9$, (b) Effect of price P_2 on the average reward; $P_1 = 0$ and $P_3 = 9$; $M = 10$, (c) Effect of price P_3 on the average reward $P_1 = 0$ and $P_2 = 6$; $M = 10$.

transmit – P_2 and P_3 are set with enough separation to illustrate their impact, while remaining in the same order of magnitude. Since the transition matrix Γ is derived from traces, it is interesting to compare how the optimal policy obtained by the model compares against its alternatives. In particular, note that the mobility in the traces is neither stationary nor memoryless. Therefore, one of the goals is to assess to what extent these results still hold if the considered assumptions are removed.

Fig. 2.6(a), compares 1) the theoretical optimal reward predicted by the model – i.e., using the empirical transition matrix Γ, 2) the average reward obtained simulating data collection and upload using the original real traces under the optimal policy predicted by the model and, finally, 3) the optimal reward obtained by using the multi-threshold policy obtained by exhaustive search on the real traces. The match appears quite tight and also rather insensitive to the variation of M.

2.7.3 Offloading under unconstrained aging control

The effects of prices on the freshness of information delivered by each device are explored here. To this aim, Fig. 2.6(b) and 2.6(c) illustrate how the average reward changes as function of the price. Fig. 2.6(b) is obtained by fixing P_1 and P_3 and varying P_2 from 2 to 8. It shows that the average reward decreases with price P_2. Indeed, as P_2 becomes larger, the device has more incentive to upload

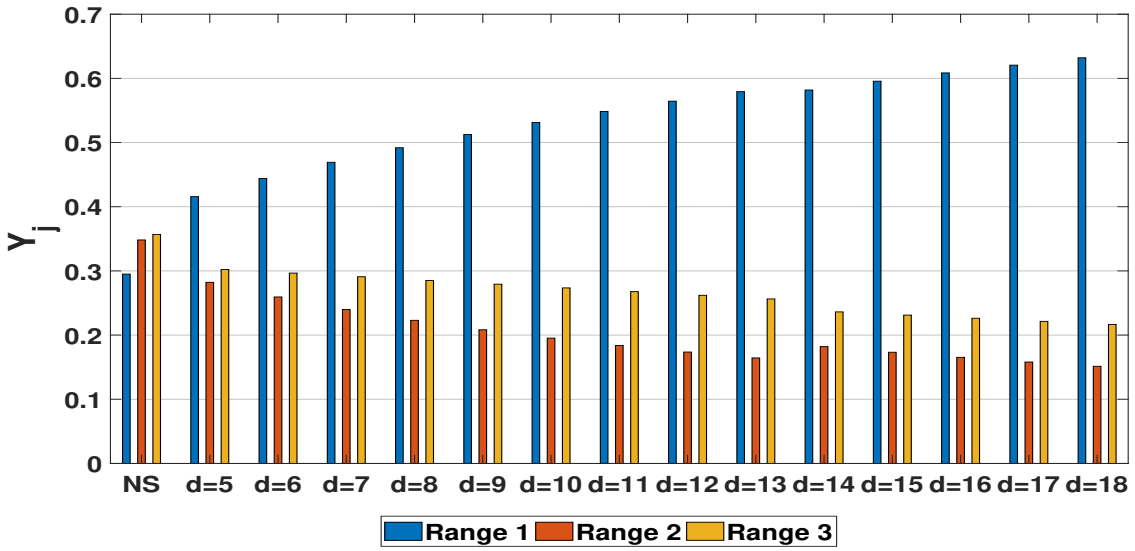


Figure 2.7: Effect of pricing for various values of d; $\epsilon = 0.01$.

the collected data to locations in range 1. It is possible that the age of collected data becomes higher – i.e. upload occurs farther from the origin site – which explains why the average reward decreases with P_2. Same behavior is observed by changing the price P_3 and making P_1 and P_2 fixed, as depicted in Fig. 2.6(c). In summary, if the QoS constraints are ignored, the SP should indeed increase the price for locations in range 2 and range 3: the effect is to shift all collected traffic to locations in range 1, which are less expensive to lease.

2.7.4 Offloading under aging control with AoI constraints

Fig. 2.7 shows the relative volume of traffic uploaded in range 1, 2 and 3 under optimal pricing for increasing values of the AoI constraint d. From these simulation results, some useful insights can be obtained on the load balancing operated by the proposed pricing scheme. The first point on the x-axis corresponds to the profile of traffic obtained under uniform flat price, that is No Strategy (NS), meaning that all IoT traffic is uploaded where it is produced. The remaining points correspond to the optimal prices for different values of d. Observe that it is possible to shift an important part of traffic (13%-32%) towards locations in range 1 (which corresponds to price $P_1 = 0$). A smaller part of traffic (1%-12%) is instead shifted to locations in range 3. When d increases, observe that more traffic is shifted since devices have more opportunities to upload collected data in locations in range 1.

Fig. 2.8 shows how the price impacts the cost incurred by the SP. A larger value of d for data generated at a given location means lower sensitivity to delay, which in turn increases the probability to upload at locations in range 1, which explains why the total cost is ultimately decreasing with the

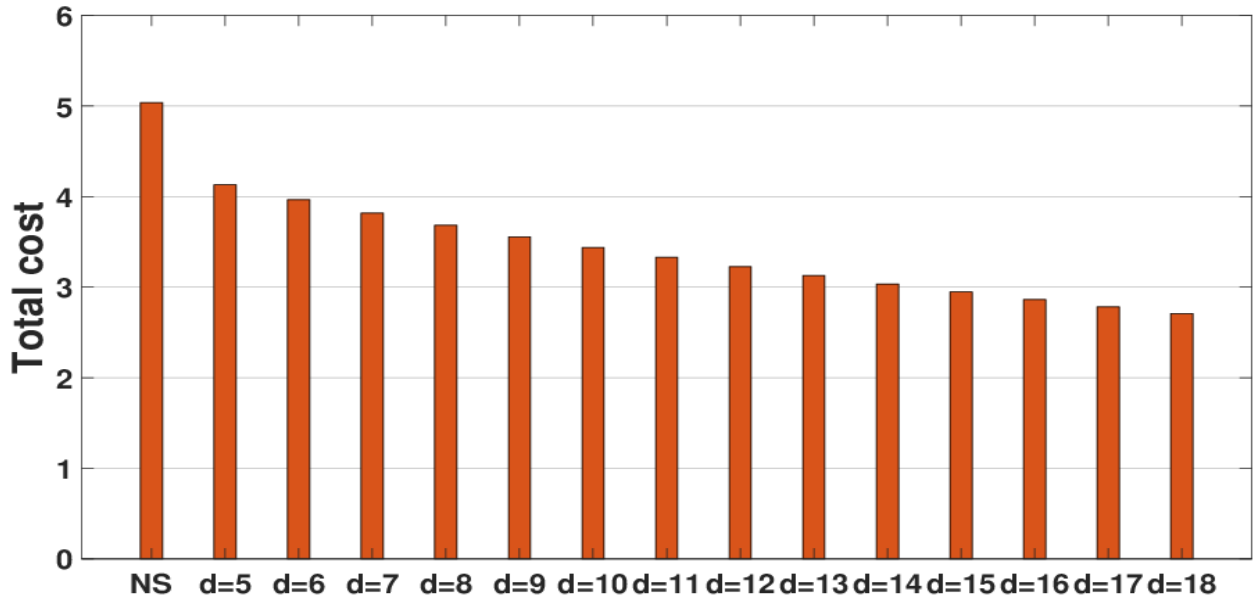


Figure 2.8: cost incurred by the MSP for various values of d.

QoS constraint d (total cost reduced by approximately 50% when $d = 18$).

2.8 Conclusion

Age of information has been a crucial metric in data related services driven by the growth of the IoT. In addition, the operational cost is of vital importance for the SPs. In this chapter, we studied a framework that leverages the fact that some data can be applied at a later time for a lower price. We provide a model for controlling the age of information called aging control which provides an optimal threshold based policy for the employed MDP problem of sequential upload decision making, and proved the existance of such policy. The inherent traffic management influences the overall cost for the SPs and also deals with the resource capacity issues for the InPs. By migrating part of the traffic to a low traffic locations, traffic demands can be eased on the highly congested locations. We provided a formulation to combine these two factors under one unified framework for obtaining the optimal shadow prices to jointly control the aging and traffic offloading issues. This problem is an NP-hard problem, hence it is opted to use heuristics for faster convergence. In the next chapter, we describe the proposed algorithms based on a heuristic approach named simulated annealing, and further enhance it by leveraging the inherent neighbourhood structure of the locations.

Chapter 3

Simulated annealing algorithms

Having established the optimization problem to jointly control aging and traffic offloading in the previous chapter, we introduce efficient algorithms in chapter to find the optimal pricing. The algorithms are driven by the rationale according to which a shadow price vector should permit to offload as much traffic as possible towards locations with smaller costs. In order to obtain the optimal pricing, we begin by showing that the optimal prices correspond to the optimal thresholds, allowing to simplify the analysis through the control of thresholds rather than prices (Section 3.1). Then, we consider a Markov Chain Monte Carlo (MCMC) approach to find the optimal thresholds (Section 3.2), followed by its simulated annealing (SA) extension – a standard technique for constrained combinatorial optimization problems [83, 24] (Section 3.3). The special nature of the problem allows for further refinement of the SA solution leveraging the independence of nodes that are geographically far apart (Section 3.4 and 3.5).

3.1 From prices to thresholds

In the JOAC problem introduced in the previous chapter, shadow prices set by SP are the control variables. Next, we argue that thresholds set by users can alternatively be taken as our controls. Indeed, SP prices impact users thresholds, and users thresholds impact load at different locations. Hence, framing the problem exclusively based on users thresholds rather than prices simplifies the analysis.

Let τ_l be the AoI threshold corresponding to location l. A threshold $\tau_l = t$ means a device uploads data collected at location l only if its age exceeds t. Then, the threshold vector $\boldsymbol{\tau}$ is an L dimensional vector given by $\boldsymbol{\tau} = (\tau_1, \tau_2, \ldots, \tau_L)$, comprising one threshold per location.

Let τ_{max} be the maximum threshold induced from all pricing vectors in $\mathbb{R}^L$. Then, τ_{max} is given by

$$\tau_{max} = \max\{t \in \mathbb{N} \mid \mathbb{P}(\Delta_i(t, \mathbf{0}_{-i}) > d) < \epsilon, \forall i \in \mathcal{L}\} \tag{3.1}$$

where $\Delta_i(t, \mathbf{0}_{-i})$ is the age of data collected at location i, in a setup wherein all locations except i correspond to threshold $\tau_i = 0$. Let S be the set of feasible threshold values, i.e., $S = \{0, 1, \ldots, \tau_{max}\}$.

Finally, the threshold-based JOAC is given as follows:

THRESHOLD-BASED JOAC (T-JOAC):

$$\min_{\boldsymbol{\tau}\in\boldsymbol{\tau}_{\epsilon,d}} W(\boldsymbol{\tau}) := \sum_{i\in\mathcal{L}} Y_j(\boldsymbol{\tau})C_j \text{ł} \tag{3.2}$$

$$y_{ij}(\boldsymbol{\tau}) \geq 0,\ \forall i, \forall j \in \mathcal{L} \tag{3.3}$$

where

$$\boldsymbol{\tau}_{\epsilon,d} = \{\boldsymbol{\tau} \in S^L \mid \mathbb{P}(\Delta_i(\boldsymbol{\tau}) > d) < \epsilon,\ Y_i(\boldsymbol{\tau}) \leq B_i, \forall i\}. \tag{3.4}$$

In the above formulation, the objective function corresponds to (2.18)-(2.19) in JOAC. The constraints (3.3) and (3.4) capture (2.23) and (2.19)-(C2), respectively.

Let $\boldsymbol{\tau}^*_{\epsilon,d}$ be the set of optimal threshold vectors,

$$\boldsymbol{\tau}^*_{\epsilon,d} = \{\boldsymbol{\tau} \in \boldsymbol{\tau}_{\epsilon,d} \mid W(\boldsymbol{\tau}) = \min_{\boldsymbol{\tau}'\in\boldsymbol{\tau}_{\epsilon,d}} W(\boldsymbol{\tau}')\}. \tag{3.5}$$

Next, efficient algorithms to find elements in $\boldsymbol{\tau}^*_{\epsilon,d}$ are provided.

3.2 Markov Chain Monte Carlo (MCMC)

MCMC starts from a feasible solution and attempts to improve it by performing random perturbations. A key feature of MCMC is the use of trial and error to avoid being trapped at local minima. Furthermore, it is simple to implement in a distributed way.

Given the current state, the procedure generates a trial state at random and evaluates the objective function at that state. If the trial state improves the objective function, i.e., if the objective function evaluated at the trial state is better than at the current state, the system jumps to this new state. Otherwise, the trial is accepted or rejected based on a certain probabilistic criterion. The main feature of the procedure is that a worse off solution may be accepted as a new solution with a certain probability.

Next, we introduce the Boltzmann-Gibbs distribution corresponding to T-JOAC,

$$\pi_T(\boldsymbol{\tau}) = \frac{1}{Z}\exp^{-W(\boldsymbol{\tau})/T}, \tag{3.6}$$

where Z is a normalization constant

$$Z = \sum_{\boldsymbol{\tau}\in S^L} \exp^{-W(\boldsymbol{\tau})/T}, \tag{3.7}$$

and T is a constant, referred to as the temperature, and whose discussion is deferred to the upcom-

ing section.

MCMC has multiple flavors. Next, the most common MCMC method, namely Metropolis–Hastings (MH)[40] is considered. One of the ingredients of MH is a transition matrix Q^* for *any* irreducible discrete time Markov chain (DTMC). The states of the DTMC are given by the reachable T-JOAC threshold vectors (see (3.4)). Chain Q^* is the *proposal chain*, as samples collected from Q^* are the proposal threshold vectors. Then, based on those proposals, the MH algorithm decides whether or not they will be accepted.

Although the algorithm works for any irreducible proposal chain, the choice of the chain impacts its time to convergence. Consider the simplest chain, whose transition matrix is uniform and symmetric. Noting that the threshold vector is an L-dimensional vector $\boldsymbol{\tau}$, in the simplest setting, this allows every single component of the vector to be updated conditional on the other $L-1$ components being fixed and given. In this case, the algorithm is also known as Gibbs sampler, and is a special case of the MH algorithm. Sections 3.4 and 3.5 describe the scheme on how to leverage spacial information to refine the proposal matrix and reduce convergence time by allowing multiple dimensions of vector $\boldsymbol{\tau}$ to be updated simultaneously.

The proposal chain is given by Q^*. Let $q^*(\boldsymbol{\tau}, \boldsymbol{\tau}')$ be the entry at position $(\boldsymbol{\tau}, \boldsymbol{\tau}')$ of the corresponding transition matrix:

$$q^*(\boldsymbol{\tau}, \boldsymbol{\tau}') = \begin{cases} \frac{1}{L\,(\tau_{max}-1)}, & \exists i : \tau_i \neq \tau_i' \text{ and } \tau_j = \tau_j', \forall j \neq i, \\ 0, & \text{otherwise.} \end{cases} \tag{3.8}$$

Clearly, $\sum_{\boldsymbol{\tau}'} q^*(\boldsymbol{\tau}, \boldsymbol{\tau}') = 1$ as each transition corresponds to the change of one of the L thresholds to one of the distinct $\tau_{max} - 1$ values.

Let $\delta(\boldsymbol{\tau}, \boldsymbol{\tau}')$ be the change in the objective function when going from $\boldsymbol{\tau}$ to $\boldsymbol{\tau}'$,

$$\delta(\boldsymbol{\tau}, \boldsymbol{\tau}') = W(\boldsymbol{\tau}') - W(\boldsymbol{\tau}) = \sum_j \delta_j(\boldsymbol{\tau}, \boldsymbol{\tau}') \tag{3.9}$$

where

$$\delta_j(\boldsymbol{\tau}, \boldsymbol{\tau}') = (Y_j(\boldsymbol{\tau}') - Y_j(\boldsymbol{\tau}))C_j. \tag{3.10}$$

To describe the Markov chain $\boldsymbol{\tau}(0), \boldsymbol{\tau}(1), \ldots$, assume that the threshold vector at iteration t is given by $\boldsymbol{\tau}$. Then, the threshold vector is determined as follows

1. choose threshold vector $\boldsymbol{\tau}'$ according to Q^*, i.e., choose $\boldsymbol{\tau}'$ with probability given by (3.8). Threshold vector $\boldsymbol{\tau}'$ is the *proposal threshold vector*;

2. let the *acceptance function* be given as follows,

$$\tilde{a}(\boldsymbol{\tau}, \boldsymbol{\tau}') = \frac{\pi_T(\boldsymbol{\tau}')}{\pi_T(\boldsymbol{\tau})} = e^{-\delta(\boldsymbol{\tau}, \boldsymbol{\tau}')/T}. \tag{3.11}$$

If $\tilde{a}(\boldsymbol{\tau}, \boldsymbol{\tau}') \geq 1$, i.e., if $\delta(\boldsymbol{\tau}, \boldsymbol{\tau}') \leq 0$, then $\boldsymbol{\tau}'$ is accepted, $\boldsymbol{\tau}(t+1) \leftarrow \boldsymbol{\tau}'$. Otherwise, it is accepted with probability $\tilde{a}(\boldsymbol{\tau}, \boldsymbol{\tau}')$ and rejected otherwise. If rejected, the threshold vector remains unchanged, $\boldsymbol{\tau}(t+1) \leftarrow \boldsymbol{\tau}(t)$.

A.

Lemma 3.1. *The Markov chain produced by the above algorithm has stationary distribution* π_T.

Proof. See Appendix B □

In what follows, the above algorithm is extended through a simulated annealing approach.

3.3 Simulated Annealing

Next, we consider simulated annealing, that allows T to decrease in time, in order to guarantee the convergence to an optimal threshold vector (and corresponding pricing). Let T_t be the temperature at iteration t. For a temperature T_t, an inhomogeneous Markov chain $(Y_t)_{t\in\mathcal{N}}$ is defined with transition kernel Q_{T_t} at time t. If T_t decays to zero sufficiently slowly, the Markov chain Q_{T_t} will reach a sufficiently small neighborhood of the target equilibrium, π_{T_t}. For this reason T_t is called the cooling schedule of SA. A standard cooling schedule in the form $T_t = \frac{\hat{a}}{\log(1+t)}$ is used in this work, where $\hat{a} > 0$ is a constant that determines the cooling rate order.

Theorem 3.2. *If* T_t *assumes the parametric form*

$$T_t = \frac{\hat{a}}{\log(t+1)} \tag{3.12}$$

where

$$\hat{a} = \frac{NF}{\kappa} \max_{j\in\mathcal{L}} C_j, \tag{3.13}$$

then

$$\lim_{t\to\infty} \pi_{T_t}(\{\boldsymbol{\tau} \in \boldsymbol{\tau}^*_{\epsilon,d}\}) = 1 \tag{3.14}$$

The above theorem shows that the Markov chain with transition matrix Q_{T_t} converges to an optimal threshold $s^* \in S^*$, where S^* is the set of the optimal solutions of the T-JOAC problem (see (3.5)).

Proof. This proof is based on the technique introduced in [36]. Note that the objective function W is nonnegative and its maximum value is attained when data is collected by all devices at locations where cost is maximal. Thus $\hat{a} > W(\boldsymbol{\tau})$ for all states $\boldsymbol{\tau}$. In particular, letting d^* denote the maximum value of $W(\boldsymbol{\tau})$ at all states $\boldsymbol{\tau}$ which correspond to a local but not global minima, and $\hat{a} > d^*$. Then,

$$\sum_{t=1}^{\infty} \exp^{-d^*/T_t} = \sum_{t=1}^{\infty} \exp^{(-\frac{d^*}{\hat{a}}\log(t+1))} \tag{3.15}$$

$$> \sum_{t=1}^{\infty} \frac{1}{t+1} = +\infty \tag{3.16}$$

Theorem 1 in [36] ensures that under the above condition the limit (3.14) holds, which completes the proof. □

Algorithm 1: Simulated Annealing (SA) algorithm for T-JOAC at time t, at the neighborhood of location i

Input: $\mathcal{L}, S, \boldsymbol{\tau}(t-1), i$ ▷ `i is the candidate location for AoI threshold change`

1 **Assignment Phase**
2 Set temperature $T_t = \hat{a}/\log(1+t)$
3 Set new threshold τ_i' uniformly at random, $\tau_i' \in S \backslash \{\tau_i(t-1)\}$
4 Set $\tau_j' \leftarrow \tau_j(t-1)$, $\forall j \in \mathcal{L} \backslash \{i\}$
5 **Test Phase**
6 At each $j \in \mathcal{N}_i$ locally measure δ_j (see (3.10)) and $\mathbb{P}(\Delta_j(\boldsymbol{\tau}') > d)$
7 Each location $j \in \mathcal{N}_i$ sends measurements to location i
8 **Decision Phase**
9 **if** *$\boldsymbol{\tau}'$ does not satisfy constraints for all locations in $\mathcal{N}_i$* **then**
10 go back to selection phase (line 3)
11 Set $\delta = \sum_{j \in \mathcal{N}_i \cup \{i\}} \delta_j(\boldsymbol{\tau}(t-1), \boldsymbol{\tau}')$ (see (3.9))
12 $\tau_i(t) \leftarrow \tau_i(t-1)$
13 **if** $\delta \leq 0$ **then**
14 $\tau_i(t) \leftarrow \tau_i'$
15 **else**
16 $\tau_i(t) \leftarrow \tau_i'$ with probability $e^{-\delta/T_t}$ (see (3.11))

Output: $\tau_i(t)$

3.4 Simulated annealing leveraging neighborhoods

3.4.1 Neighborhood structure

Next, the neighborhood structure between locations is leveraged to specialize SA to our T-JOAC problem. Let the neighborhood set $\mathcal{N}_i$ for location i be defined as follows: a location j belongs to $\mathcal{N}_i$ if j is located within a given radius such that data offloaded to j can be impacted by traffic generated at location i.

Let $G(V, E)$ be the location neighborhood graph, where V is the set of vertices representing the locations and E is the set of edges, where an edge is a link between two vertices indicating that the

two corresponding locations are neighbors.

Note that the neighborhood structure depends primarily on the geographic position of the locations, mobility of the devices, and the maximum time that a device can wait before uploading the data. Indeed, let $\boldsymbol{\tau}_{max}$ be a threshold vector wherein all elements equal τ_{max}. Such threshold vector corresponds to nodes that defer transmissions as much as possible. Then, the neighborhood of location i is defined as follows,

$$\mathcal{N}_i = \{j \in \mathcal{L} \mid Y_{ij}(\boldsymbol{\tau}_{max}) > 0 \ \text{ or } \ Y_{ji}(\boldsymbol{\tau}_{max}) > 0\}.$$

Indeed, if traffic at locations i and j does not interfere with each other under the extreme scenario where all thresholds are set to their maximum values, one can safely assume that locations i and j are not neighbors.

3.4.2 Simulated annealing leverage neighborhoods

Hereafter, the implementation of the simulated annealing algorithm for solving the T-JOAC problem (see Algorithm 1) is described in detail. Time is divided into discrete slots. At the first slot, it begins by initializing the thresholds of all locations to zero, which corresponds to a price $p = 0$. Then, at each time slot t, it lets $T_t = \hat{a}/\log(1+t)$ (the initial temperature T_t should be large enough to allow all candidate solutions to be accepted uniformly at random), and the system goes through three phases: assignment, testing, and decision. The SP selects a location $i \in \mathcal{L}$ uniformly at random and run Algorithm 1. During the assignment phase, the threshold of location i is modified, while letting all other thresholds unchanged (lines $1-4$). At the test phase, the SP receives measurements from all locations which are possibly affected by a change in the threshold of location i, i.e., from all $j \in \mathcal{N}_i$, and checks whether the newly generated threshold vector $\boldsymbol{\tau}'$ is feasible (lines $5-10$). If it isn't feasible, the algorithm returns to the selection phase. Otherwise, it continues in the decision phase, by assessing the change in the objective function, δ, again using data from $j \in \mathcal{N}_i \cup \{i\}$ (line 11). If the change is negative, the new threshold vector is accepted (lines $13-14$). Otherwise, it is accepted with probability $\exp(-\delta/T_t)$. SP repeats this procedure until the established stopping conditions are satisfied. i.e., either threshold vector is not changed for two successive time slots or $T_t < \varepsilon$.

3.4.3 Independent sets

Independent sets of locations can be exploited to accelerate the SA algorithm, i.e., a partition of locations into sets where locations within each set are not affected by a change of threshold that may occur in other locations of the same set. The following paragraph indicates how the proposal chain can be adapted to account for independent sets of locations, under a serial implementation.

In practice, the speedup is obtained since the algorithm can be run in parallel for all the locations that belong to the same independent set, as indicated in Section 3.5. The experiments conducted demonstrate that this parallelization can attain a two-fold speedup of the run time with respect to the basic implementation of the algorithm.

3.4.4 Proposal chain leveraging the neighborhood structure

Given the neighborhood structure, the proposal chain introduced in (3.8) is adapted in order to allow for multiple threshold adjustments at the same iteration. The new proposal chain $\tilde{Q}^*$, whose $(\boldsymbol{\tau}, \boldsymbol{\tau}')$ entry is denoted by $\tilde{q}^*(\boldsymbol{\tau}, \boldsymbol{\tau}')$, is given as follows:

$$\tilde{q}^*(\boldsymbol{\tau}, \boldsymbol{\tau}') = \begin{cases} \frac{1}{|\mathcal{M}_{\boldsymbol{\tau}}|}, & \text{if } \boldsymbol{\tau} \sim \boldsymbol{\tau}' \\ 0, & \text{otherwise} \end{cases} \tag{3.17}$$

where

$$\mathcal{M}_{\boldsymbol{\tau}} = \{\boldsymbol{\tau}' | \boldsymbol{\tau} \sim \boldsymbol{\tau}'\} \tag{3.18}$$

and $\boldsymbol{\tau} \sim \boldsymbol{\tau}'$ denotes that threshold vectors $\boldsymbol{\tau}$ and $\boldsymbol{\tau}'$ are adjacent. Two threshold vectors are adjacent if they differ in at least one position and, in addition, all positions that differ across the two threshold vectors correspond to locations that belong to the same independent set, i.e., in the location neighborhood of graph $G(V, E)$ there is no edge between the locations whose thresholds differ. If each location corresponds to its own independent set, i.e., if there are L independent sets, then $|\mathcal{M}_{\boldsymbol{\tau}}| = L(\tau_{max} - 1)$ and the above proposal chain reduces back to (3.8).

Note that the above proposal chain produces proposal threshold vectors wherein multiple thresholds may change concomitantly with respect to the current threshold vector. Then, a straightforward adaptation of Algorithm 1 accepts or rejects the proposal threshold vector as a whole, treated as a single entity.

In the following section, in contrast, each of the neighborhoods is treated independently. In particular, at each step of the algorithm, there are multiple new proposal thresholds, which are evaluated in parallel and may be independently accepted or rejected. Even if one neighborhood rejects a particular proposal for a new threshold, other independent neighborhoods may accept their proposals.

Table 3.1: Table of notation for coloring algorithm and T-JOAC

Variable	*Description*
$H = \{h_1, \ldots, h_L\}$	set of colors that can be assigned to a location
$c_l \in H$	color of location l
$\phi(\boldsymbol{c})$	set of colors used by $\boldsymbol{c}$
$\boldsymbol{c}^{(n)}$	current coloring vector in Algorithm 2
$\boldsymbol{c}^\star$	best coloring vector so far (broadcast from Algorithm 2 to Algorithm 3)
$\boldsymbol{c}(t)$	current coloring vector in Algorithm 3

3.5 Accelerated simulated annealing with parallel computations: a coloring approach

Algorithm 2: SA-based coloring agorithm

Input: $\mathcal{L}, H, b, \boldsymbol{c}^{(0)}$

1 $n \leftarrow 1$
2 $\boldsymbol{c}^\star \leftarrow \boldsymbol{c}^{(0)}$
3 **while** *true* **do**
4 Set $\tilde{T}_n = b/\log(1+n)$
5 Select location $l \in \mathcal{L}$ and color $c'_l \in H \backslash \{c_l^{(n-1)}\}$
6 Set $c'_j \leftarrow c_j^{(n-1)} \quad \forall j \in \mathcal{L} \backslash \{l\}$
7 **if** $\boldsymbol{c}'$ *is not feasible* **then**
8 go to step 5
9 Set $\beta = |\phi(\boldsymbol{c}')| - |\phi(\boldsymbol{c})|$
10 $\boldsymbol{c}^{(n)} \leftarrow \boldsymbol{c}^{(n-1)}$
11 **if** $\beta \leq 0$ **then**
12 $\boldsymbol{c}^{(n)} \leftarrow \boldsymbol{c}'$
13 **else**
14 $\boldsymbol{c}^{(n)} \leftarrow \boldsymbol{c}'$ with probability $e^{-\beta/\tilde{T}_n}$
15 **if** $|\phi(\boldsymbol{c}^{(n)})| < |\phi(\boldsymbol{c}^\star)|$ **then**
16 $\boldsymbol{c}^\star \leftarrow \boldsymbol{c}^{(n)}$
17 broadcast new coloring $\boldsymbol{c}^\star$ to all locations
18 $n \leftarrow n+1$

3.5.1 Colorings

Fewer independent sets correspond to more opportunities for concomitant threshold adjustments. A partition of the locations into independent sets can be obtained as the result of a graph coloring procedure. The coloring of a graph is a function that assigns different colors to adjacent vertices of a graph.

Given a graph G, the Graph Coloring Problem (GCP) seeks the minimum number of colors $\chi(G)$ which can be used to color G. Such a number is called the chromatic number. An upper bound for the chromatic number is given by the maximum vertex degree plus one, and is attained by a greedy coloring procedure. Nonetheless, such upper bound (may be loose) and optimal coloring is an $\mathcal{NP}$-hard problem, motivating heuristic solutions. It is observed that the simulated annealing can be used as a heuristic to solve GCP [54] to obtain a near optimal solution. Thus this can account for the coloring process and the resulting parallelization to T-JOAC in an extension to Algorithm 1.

3.5.2 SA for colorings

The location coloring is obtained by using a specialized simulated annealing algorithm. Let $H = \{h_1, h_2, ..., h_L\}$ be the set of available colors. Then, the coloring $\boldsymbol{c}$ (vector) indicates, for each location, its corresponding color, i.e., $c_l = h_m$ if the color of location l equals h_m.

A coloring is *feasible* if, for any pair of locations i, j such that $i \in \mathcal{N}_j$, and $c_j \neq c_i$. Let $\phi(\boldsymbol{c})$ be the set of colors used by coloring $\boldsymbol{c} = (c_1, c_2, .., c_L)$. Thus the objective of the coloring problem is to find a feasible solution that minimizes the cardinality of set $\phi(\boldsymbol{c})$. The algorithm begins with an initial feasible coloring scheme, then improve it further using simulated annealing. The initial coloring schemes could be achieved with a random allocation of one color to each location, where the number of colors are equal to the number of locations or an intermediate solution is obtained by greedy algorithm. The temperature $\tilde{T}_n$ is structurally similar to T_t used in Algorithm 1, noting that now b plays the role of $\hat{a}$, and corresponds to an upper bound on the number of colors to be adopted. In the simplest setting, we let $b = L$.

Algorithm 2 is used to continuously search for better colorings. In lines 5 and 6 the algorithm chooses a location l uniformly at random, and a color from $H \backslash \{c_l^{(n-1)}\}$. Lines 7 and 8 produce the proposal coloring vector $\boldsymbol{c}'$. Lines $9 - 15$ test if the proposal is accepted or not. Finally, if the new coloring vector uses fewer colors than the current best candidate, the new coloring vector is broadcasted to all locations.

In summary, Algorithm 2 is continuously run by the SP, e.g., at a fast time scale, and as improvements are found they are broadcasted to the locations which locally run Algorithm 1. In particular, the time scale at which Algorithm 2 is executed, whose iterations are denoted by n, is decoupled from the scale of the time slots considered in Algorithm 1, denoted by t. The integrated solution involving Algorithms 1 and 2 is presented in Algorithm 3, and is described in the sequel.

3.5.3 SA for T-JOAC with colorings

The accelerated simulated annealing for T-JOAC is obtained by leveraging the colorings produced by Algorithm 2, integrating them with Algorithm 1 as shown in Algorithm 3. As explained previously, the SP continuously broadcasts colorings $\boldsymbol{c}^\star$ obtained using Algorithm 2. At line 5 of Algorithm 3,

Algorithm 3: Accelerated SA algorithm (at time slot t)

Input: $\mathcal{L}, H, S, \varepsilon, \boldsymbol{\tau}(t-1), \boldsymbol{c}(t-1), \boldsymbol{\tau}^{\star}$

1 $\boldsymbol{c}(t) \leftarrow \boldsymbol{c}(t-1)$; $\boldsymbol{\tau}(t) \leftarrow \boldsymbol{\tau}(t-1)$
2 **if** *there exists untreated broadcasted* $\boldsymbol{c}^{\star}$ **then**
3 update colorings at all locations
4 $\boldsymbol{c}(t) \leftarrow \boldsymbol{c}^{\star}$
5 Choose uniformly at random a color $h(t) \in \phi(\boldsymbol{c})$
6 **Optimal threshold search**
7 **run in parallel**
8 run Algorithm 1 at neighborhood of first location with color $h(t)$
9 $\vdots$
10 run Algorithm 1 at neighborhood of last location with color $h(t)$
11 Update $\boldsymbol{\tau}(t)$ given outputs from parallel runs of Algorithm 1
12 **if** $W(\boldsymbol{\tau}^{\star}) > W(\boldsymbol{\tau}(t))$ **then**
13 $\boldsymbol{\tau}^{\star} \leftarrow \boldsymbol{\tau}(t)$
14 **if** *threshold vector is not changed for two successive time slots or* $T_t < \varepsilon$ **then**
15 stop

Output: $\boldsymbol{\tau}(t)$, $\boldsymbol{\tau}^{\star}$ and $\boldsymbol{c}(t)$

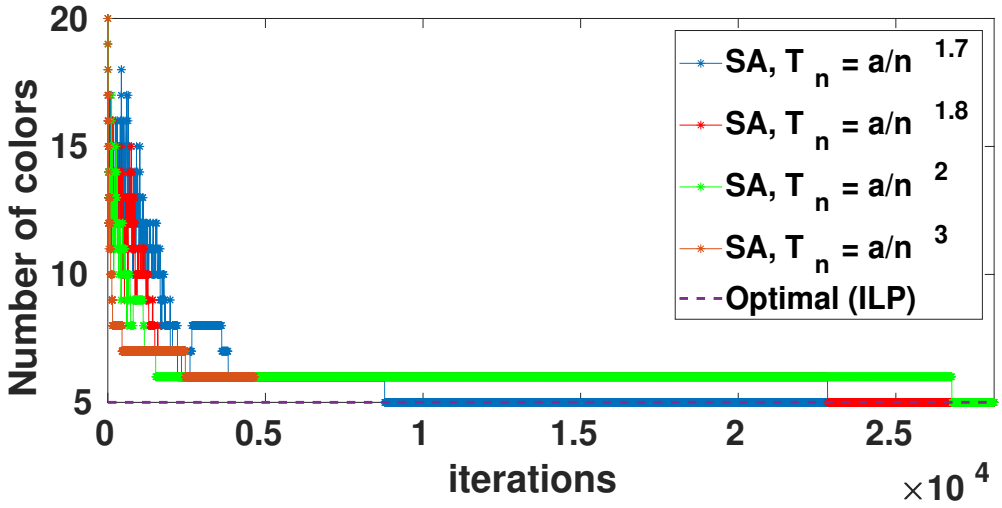
the SP selects a color $h(t)$ from set of colors $\phi(c)$ currently in use. Then, Algorithm 1 is run locally (in parallel) at locations colored with the same color $h(t)$ (lines $7-10$ in Algorithm 3). As in the basic implementation, such locations generate new threshold values which are then combined into a new threshold vector $\boldsymbol{\tau}(t)$ (line 11 in Algorithm 3). By locally searching for optimal thresholds at multiple locations, we improve the running time for computing optimal thresholds, as shown in the evaluations that follow.

3.6 Numerical evaluation

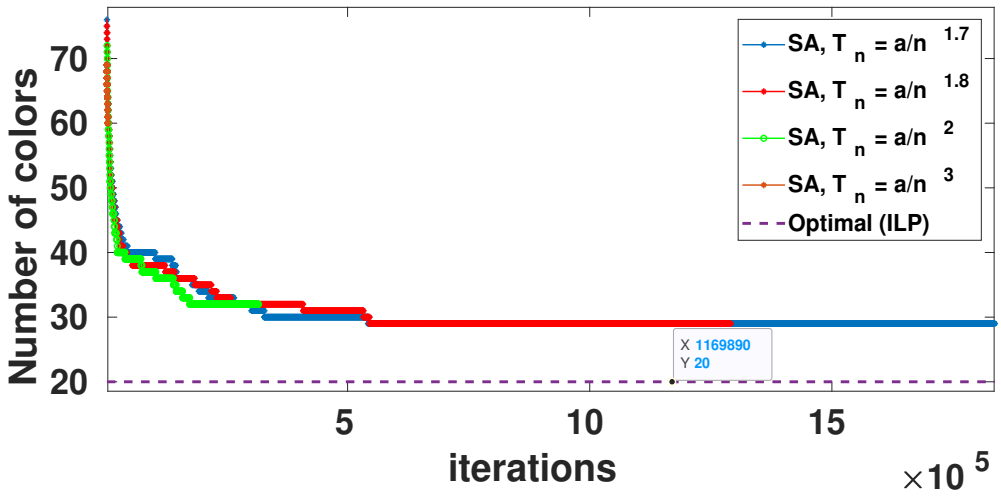
We use the same experimental setup used in chapter 2 to validate the theoretical findings reported in the previous sections. Following numerical results show the convergence of the coloring solution and compare the algorithm 1 with algorithm 3.

3.6.1 Leveraging coloring for joint offloading and aging control

Results concerning location coloring are depicted in Fig. 3.1, where, $d = 7$, ϵ = 0.01 (factor which determines the neighborhood of locations in the SA algorithm) and we use the same cell topology as mentioned in chapter 2. Fig. 3.1 shows the convergence of the minimum number of colors for a region of 20 locations and 230 locations. Different temperatures are considered to test the convergence of the algorithm. The results obtained using Algorithm 2 are compared to the integer linear programming ILP (branch-and-bound) [39] which is more conventional and can be considered as



(a) 20 locations



(b) 230 locations

Figure 3.1: Convergence of SA algorithm for coloring of locations.

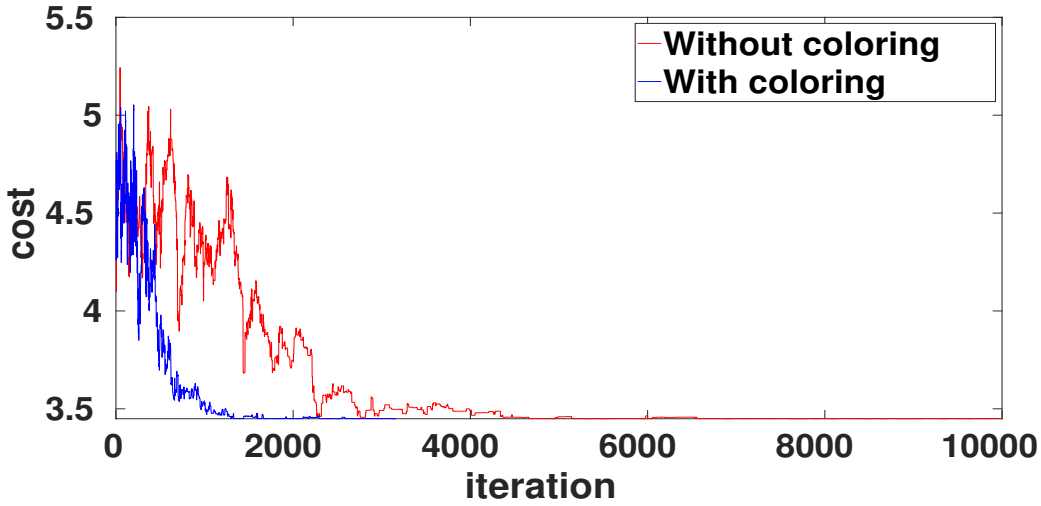


Figure 3.2: Comparing convergence of two algorithms for $d = 7$.

a baseline solution for coloring algorithms. As shown in Fig. 3.1(a) the SA algorithm has optimal results for the region with less locations, However, when tested on the region with full 230 locations the SA algorithm lags behind the ILP as depicted in Fig. 3.1(b). However, for the purposes of traffic offloading, the near-optimal solution to the coloring problem provided by the SA algorithm is acceptable as shown in the next set of experiments.

For the results shown in Fig. 3.2-3.4, a 20 macro cell topology is utilized; with no loss of generality, the traffic generation is normalized as $\frac{NF}{T} = 1$. Consider $d = 7$, $\epsilon = 0.01$, and the maximum threshold $t_{max} = d + 3$. The temperature used to control the number of iteration was derived by trying different configurations; the basic trade-off is to perform a number of iterations large enough for the algorithm to converge to the minimum value and yet bound it to a maximum value for the sake of computation time. It is discovered by numerical exploration that use of $T_t = \hat{a}/\log(1 + t)$ or $T_t = \hat{a}/t^{2.8}$, where $\hat{a} = 10^6$, produces the same optimum value where the latter has faster convergence. Hence, the temperature setting that is best suited for the algorithm is selected. Fig. 3.2 to Fig .3.4 show the difference in the convergence of average cost using Algo. 1 (without coloring) and Algo. 3 (with coloring). As it can be observed, the proposed coloring method significantly speeds up the convergence compared to the case where the coloring method is not deployed. The gain in performance holds for all the three considered values of d as shown in Fig. 3.2 for $d = 7$, Fig. 3.3 for $d = 9$,Fig. 3.4 for $d = 12$. The improvement in convergence time with respect to the SA without coloring is approximately 50% .

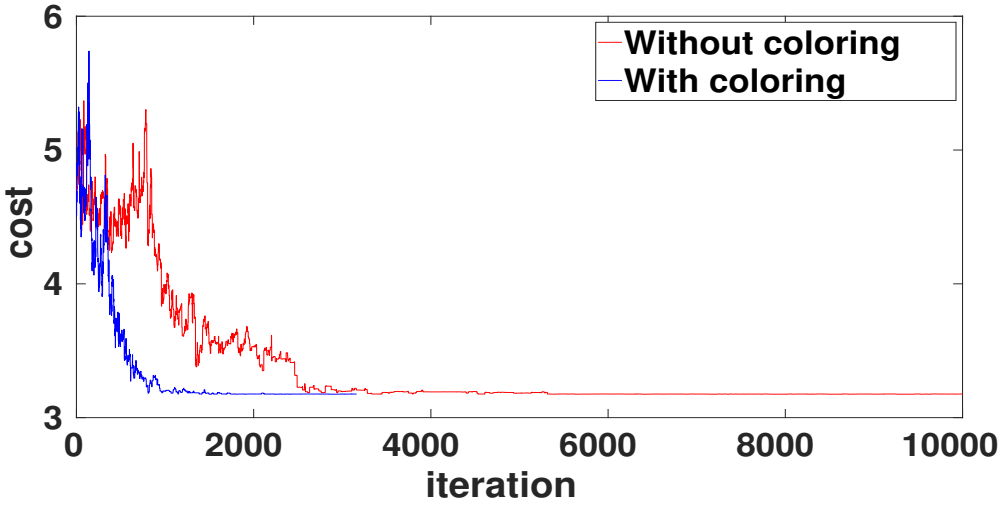


Figure 3.3: Comparing convergence of two algorithms for $d = 9$.

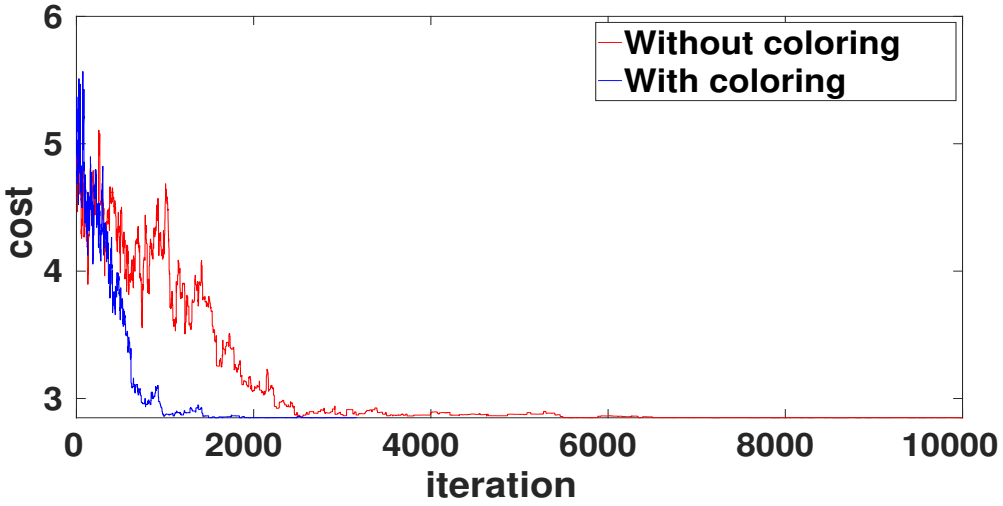


Figure 3.4: Comparing convergence of two algorithms for $d = 12$.

3.7 conclusion

Future IoT service providers will need ubiquitous IoT data collection, mandating in turn the support of IoT access at scale over the 5G infrastructure. At the same time, new schemes to control data generation and upload should allow to SPs to perform IoT data brokerage across diverse access resources made available by concurrent infrastructure providers at different costs, in the form of 5G IoT resource slices.

This work introduces a new framework to connect two fundamental aspects: the AoI of IoT data to be uploaded and the cost of 5G resources leased in order to obtain network access services. The upload control can be performed in a distributed way at the device level using optimal dynamic multi-threshold policies. Such policies have been showed to outperform their static counterparts. At same time, a SP can control prices to match optimal multi-threshold policies to service requirements while minimizing operational costs. It does so at the slice level by incentivizing users to perform IoT data uploads where resources leased from InPs are cheaper. This work opens new directions at the bridge between IoT and 5G research, by describing on a quantitative basis how to trade-off IoT data freshness and load balancing, as supported by the 5G slicing paradigm. In particular, we envision real testbed deployments and the investigation of strategic mobility patterns to reduce costs as interesting areas for future exploration. In chapter 2 and chapter 3, we assume that the cost incurred due to resource procurement from InPs is given. However, this is such an important factor in deciding the total cost faced by the SPs and the resource prices are crucial for the InPs in order to increase their profits. We study this aspect in the following chapters.

Chapter 4

Fisher market for multi-resource allocation

4.1 Introduction

5G networks promise to enable new paradigms such as edge and fog computing by deploying virtualized resources in a multi-tenant and multi-service scenario, capable of fulfilling the dynamic and demanding requirements of numerous applications. With network slicing, the Infrastructure Providers (InPs) can offer differentiated services using shared resource pools. A slice, in this context, is a share of the mobile network operator infrastructure obtained via virtualization with the help of Software-Defined Networking (SDN) and Network Function Virtualization (NFV) technologies. A slice forms a logical network on top of the physical one [115, 13]. Evolving from previous mobile technology, the 5G core network architecture integrates data-centers into their architectures to support network function virtualization and computation offloading. Thus, a slice will typically encompass different resource types, such as radio access capacity, edge storage memory, and computing power available [115]. Hence, with network slicing, InPs can create multiple virtual networks or "Slice", which can be used for a specific application or service with particular requirements. Currently, most often these services are virtualized to capitalize on the inherent scaling flexibility of virtualization.

Efficient resource allocation helps to improve resource utilization, provides high quality of service for the end-users. But in the context of network slicing, the resource allocation problem is more challenging when there are multiple resource types and competing Service Providers (SP) with diverse characteristics and preferences. Hence, the main goal here is how to enable efficient creation and fair multi-resource allocation for slices as well as end-users. To address this problem, we propose a new market-based framework to harmonize the needs of each SP while the system strikes a balance between fairness and efficiency.

To address the aforementioned challenges, we formulate the multi-resource allocation for network slices as a Fisher Market where the SPs act as buyers and a set of resources are divisible goods available at different locations. Unlike the previous work, using the fisher market, we propose a generalized α-fairness resource allocation applied to SPs that allows to adapt the degree of fairness as a function of $\alpha \in [0, \infty)$. It controls the trade-off between fairness and efficiency. In our

model, each SP has a certain budget for resource procurement, which represents the market power of the SP. Given the resource prices, each SP buys an optimal multi-resources to maximize its utility under budget constraints. Hence the market equilibrium is to compute a vector price of resources that ensures market cleaning, i.e., the demand of a resource equal its supply. In this work, we show that the market equilibrium corresponds to the allocations maximizing α-fair utility, which is obtained under non-linear pricing. Furthermore, we obtain a closed-form of the pricing as a function of α and resources purchased by SP. In particular, we show that the marginal price increases as more of the resource is purchased when $\alpha > 1$ and it decreases when $\alpha < 1$. Indeed, by choosing $\alpha > 1$, the InP may choose to impose an increasing marginal cost to ensure no single SP can monopolize all resources when its budget is higher compared to other network slices.

Over time resource allocation problem has been well-studied [38] in wireless networks, this problem resurfaces with the evolution of such networks, namely the introduction of cloud computing and assigning resources from virtual machines to compute the uploaded user tasks [112, 70] and recent development of network slicing [118, 18] and edge computing [107, 116] to capture the different use case scenarios. There has been a constant challenge to adapt the existing methods or to develop new methods to address these changes in the wireless networks and numerous solutions were proposed for edge computing and network slicing. Even with these changes, the underlying mechanism remains the same, i.e., agents with specific budgets try to obtain a set of goods that creates a market. One of the market models that received rigorous treatment in the network economics is the Fisher Market model [17], which still receives great attention in resource allocation, majority of the works that employ this model for resource allocation [79, 76] focus on obtaining Market Equilibrium (ME) by solving Eisenberg-Gale [28] convex programming. This produces linear prices that do not fit well in the real world scenarios, prices, in general, are non-linear, [34] address this issue very well but the price curves are constructed with a condition on the agent preferences, that leads to same price curves for all the agents.

An important aspect that seizes most of the attention in the literature is the fairness and efficiency [61, 60, 5], most of the auction-based models [69] prefer to allocate a higher share of resources to SPs with better marginal utilities, this is similar to the utilitarian approach, whereas the Max-min [31] allocates more resources to the SPs with weaker marginal utilities to make the resource allocation extremely fair. There is another approach that allocates resources proportionally [80] to achieve intermediate fairness of the two extremes mentioned earlier, all of the above fairness schemes can be generalized under the so-called α-fair allocations [73],[97], which provides very good flexibility to the InP in terms of the trade-off between the fairness and efficiency. Another issue that was not addressed in these works is the location dependency for the services, as each SP has a user base that varies across the locations. In this case, the utility of the service provider depends on the utility obtained per location. To the best of our knowledge, intra-slice fairness (fairness between locations) was not addressed in combination with inter-slice fairness (fairness between slices). This motivates

us to propose a Fisher market based model to address these aspects of resource allocation.

Our main contributions are as follows,

- We propose a framework to consider both intra-slice and inter-slice fairness at once. This model considers α-fairness at the slice level and proportional fairness at the location level.
- This work extends the theoretical results of non-linear pricing [34] by not imposing any conditions on SP's valuations/preferences, this results in differential price curves.
- We provide numerous numerical results to support our theoretical claims and demonstrate that our model is consistent and effective.

We believe, this novel approach of incorporating both intra-slice, inter-slice fairness, and non-linear pricing under the same framework extends the literature of telecommunication network economics.

This chapter is organized as follows, section 4.2 introduces the system model with some of the key aspects, section 4.3 formulates the allocation problem and discuss market equilibrium and establish two-level fairness mechanism, section 4.3.5 describes the strategic aspect of the Fisher market model. Finally, this article ends with a conclusion.

4.2 System Model

We begin by presenting the system of interest and introducing some basic terminology. Consider a system with a set of Infrastructure Providers (InPs), who own the physical resources such as CPU, memory, radio resource, etc., and lease these resources to a set of SPs, who run different services (IoT, URLLC, eMMB, etc.) to support the subscribed customers at different locations. The SP negotiates and scale resources using 5G $network\ slicing$ by orchestrating slicing functionality across heterogeneous access technologies (5G, LTE, 3G, and WI-FI), over different site types (macro, micro, and pico base stations) as shown in Fig.4.1. Cloud Network Slicing is a mechanism that logically separates the physical resources owned by InPs by virtualizing the physical resources and assigning only the required resources to support specific use case scenarios as depicted in Fig.1.2.

4.2.1 Key aspects of the system model

Development in technologies has boosted the possibility of accommodating various services under 5G, this brings many changes at the infrastructure level and the business level, there are many aspects in the unprecedented growth of the electronics and communication business sector for the foreseeable future. We describe a few of such aspects that are important for our model description. They are as follows,

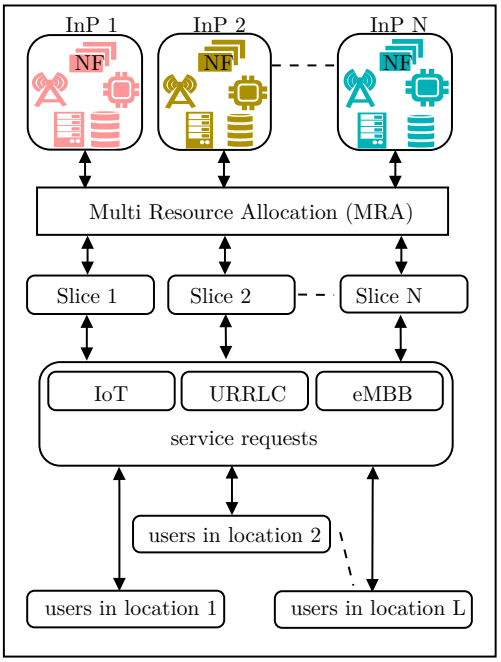


Figure 4.1: MRA for service requests in heterogeneous multi location scenario.

Services and SPs

The emergence of business ideas such as smart city, smart home, smart factories, autonomous driving, remote medical assistance, etc., push service providers to run various services to meet the user demands and create a sustainable business environment. A prime example of such an idea is the exponential growth of the IoT, these services require massive connections to connect the sensors and central entities to successfully run smart home applications. Another dimension to this is the huge data produced by the sensors that must be processed and analyzed, this requires vast storage and computational capacities. Robust and low latency connections are required to enable autonomous driving, intelligent traffic management, robotic assistance in smart factories. This diverse nature of the services requires multiple resource types across heterogeneous access points to meet the quality of service promised to the users. The user base of an SP varies across different locations, this leads to higher or lower resource requirements based on the location. Hence, SPs with a specific budget procure various resource types through network slicing from the InPs based on the service and locations.

Infrastructure and InPs

Network infrastructure plays an essential role in enabling upcoming technologies especially given the rise of various use case scenarios with 5G. The use of SDN and Network function virtualization changes the face of existing wireless networks to be the most flexible and efficient in terms of

resource allocation. Infrastructure provider is one of the key players in the telecom business, perhaps performs a very complex and difficult job of Multi Resource Allocation (MRA) to the SPs due to the diversity of the service requirements. In general, there could be more than one InP who owns the physical resources across different locations. An InP may not fully cover a region consisting of multiple locations, other InP may not have all the resource types needed to support a particular service. As shown in Fig.1.2, there are three key parts in the 5G wireless network architecture, i.e., core cloud that consists a majority of core network and manage the network functionalities, edge network that contains some of the core network functions, computational and storage resources and the radio access network that consists of heterogeneous radio access elements. An InP who owns full or part of resource types covering various locations would lease the resources to SP and benefits from the overall growth in the telecom industry.

Resource scaling with Network Slicing

A key to a successful and sustainable business is to better serve the customers with limited resources, In this context, network slicing facilitates such flexibility and helps both InPs and SPs to control their costs. We envision a scenario in which, a large metropolitan city is furnished with all the use case scenarios mentioned earlier. Each location in the city consists of hundreds of users with various types of service requests. An SP that supports a specific service accepts these requests and processes them with an agreed Quality of Service (QoS). Of course, to run this service across multiple locations, SP has to obtain resources from an InP or group of InPs. However, these resource types can be completely distributed, meaning that the resource is present at each location (e.g., radio resources), or partially distributed, i.e., a particular resource or pool of resources are present at a location and is/are shared with a group of neighboring locations(e.g., resources from edge cloud). Finally, the centralized resources, where the resource is present at a remote location but accessible to all locations to perform the user tasks(e.g., core cloud). Now, multiple SPs with limited budgets compete for a bundle of resources to run their respective services, based on the budgets and base demands, InP or a social planner (in case of multiple InP) allocate resource bundle to SPs, who in turn support the users in processing their service requests. This process is depicted in Fig. 4.1. This is formally explained with a $Fisher\ Market$ analogy in the following sections. Here, each SP is assigned with a dedicated network slice that is tailored to his requirements.*

Given the heterogeneous nature of the services offered by the SP, the slice obtained by the SP contains network and computational resources. However, the service offered to the subscribed users can be limited by any of the resources that belong to the slice, i.e., exhausting a particular resource creates a bottleneck for the slice and limits the SP in offering the service to additional users. Hence, the service capacity of the SP is determined by the bottleneck resource. We elaborate this in the

*we consider that multi-resource allocation to an SP or a slice is alike (both are used depending on the situation to convey the same meaning).

following section where we define a utility function to numerically determine the service capacity of the SP.

4.2.2 Service utility

We confine the discussion to a single InP that provides a set of resource types, namely $\mathcal{R} = \{1, 2, .., R\}$, over a set of locations. Given that each SP is assigned with a dedicated slice by InP, the total number of slices supported by InP is $\mathcal{N} = \{1, 2, .., N\}$[†]. The set of locations $\mathcal{L}_i$ concerns SP i. The capacity of resource type r at a given location l is $c_{l,r}$. The resource requirement across locations varies based on the service type. For example, to process a single IoT request, SP may need 4 units of bandwidth, 2 units of RAM, and 2 units of storage and processing resources, respectively. Let a base resource preference vector $a_{i,r}$ represent the need for each resource type r for SP i, it is the minimum quantity to run the service with certain QoS. Under budget B_i, SP i procures resources from the InP. She obtains a resource bundle $\boldsymbol{x}_i = (x_{il1} \dots x_{ilR})$, where x_{ilr} is the amount of resource type r at location l allocated to SP i. The utility function for the service level of SP i writes

$$u_i(\boldsymbol{x}_i) = \sum_{l \in \mathcal{L}_i} u_{il} \tag{4.1}$$

where, the utility of the SP i at location l u_{il} is defined as,

$$u_{i,l} = \min \Big\{ \frac{x_{il1}}{d_{il} a_{i1}}, \frac{x_{il2}}{d_{il} a_{i2}}, \dots, \frac{x_{ilr}}{d_{il} a_{ir}}, \dots, \frac{x_{ilR}}{d_{il} a_{iR}} \Big\}, \tag{4.2}$$

Where d_{il} represents the overall user demand of SP i at location l. We used $Leontief\ function$ since the resource types are perfect compliments [100], i.e, obtaining a resource type r in excess does not yield higher utility.

Here, service utility u_i depends linearly on location-based utility u_{il}, which can lead to unfair resource allocation between locations, this aspect is discussed in detail in 4.3.4.

4.2.3 Objective of the service providers

The goal of each SP $i \in \mathcal{N}$ is to meet the variable user demand across multiple locations. Hence, it requires multiple resources in sufficient proportions at different locations to meet the SLA. For this reason, SP i with a predefined budget B_i tries to obtain the resource bundle that suffice its service. Thus the utility of each SP i, named F_i writes

$$F_i = \underset{\boldsymbol{x}_i \in R_+^m : C(\boldsymbol{x}_i) \leq B_i}{argmax} \; u_i(\boldsymbol{x}_i) \tag{4.3}$$

[†]Since each SP is assigned with a dedicated slice the term 'slice' or 'SP' or 'tenant' is equivalent.

where $C(\boldsymbol{x}_i)$ is the total cost of the resource bundle that should not exceed budget B_i of SP i.

4.3 Resource Allocation Problem

In this section we present different approaches to the allocation problem in the setting described in section 4.2.

4.3.1 Fisher Market under generalized α-fair resources allocation

The classical optimization framework for the InP is to provide an efficient and fair allocation to all SPs based on their budgets. To capture the situation that the SPs may have different priorities, we consider the weighted version of the social welfare objective. Hence the main aim of the InP is to maximize the total social welfare, leading to the following 5G resource allocation problem (RAP)

$$
\begin{aligned}
P_{SW}: \quad & \underset{\boldsymbol{x}}{\text{Maximize :}} && \sum_{i\in\mathcal{N}} B_i U(u_i) && (4.4)\\
& \text{subject to} && u_i = \sum_{l\in\mathcal{L}_i} u_{il} \quad \forall\, i \in \mathcal{N}, && (4.5)\\
& && u_{i,l} \leq \frac{x_{ilr}}{d_{il}a_{ir}} \quad \forall\, i \in \mathcal{N}, l \in \mathcal{L}_i, r \in \mathcal{R}, && (4.6)\\
& && \sum_{i\in\mathcal{N}} x_{ilr} \leq c_{lr}, \quad \forall\, l \in \mathcal{L}_i, r \in \mathcal{R}. && (4.7)
\end{aligned}
$$

Given that SPs generally have their services across multiple locations, their utility indeed depends upon the utility obtained from each location as mentioned previously, this location dependency is captured by the constraint 4.5. As described under service utility that the resources are perfect compliments, hence the utility usually depends on the bottleneck resource, which is the point of leontief function, 4.6 warrants this functionality. Constraint 4.7 set the seal on the amount of resource that can be allocated to each SP for each resource type at all the locations where SP has a presence, this guarantees that the capacity is not exceeded.

The utility U is assumed to belong to the well-known class of fairness [73] that measures canned α-fairness. Specifically, we have

$$
U(y) = \begin{cases} \frac{(y)^{1-\alpha}}{(1-\alpha)} & \text{if } \alpha \neq 1 \\ \log(y) & \text{if } \alpha = 1 \end{cases} \tag{4.8}
$$

The values of enclosed $\alpha \in [0, \infty)$ give the trade-off between individual fairness and efficiency, the smaller α corresponds to utilitarian welfare where a social planner cares more about societal good

(efficiency). In contrast, larger α corresponds to the egalitarian nature of social a planner, where it cares more about individual equality (fairness). For $\alpha = 1$, for instance, the customary log-based proportional-fair utility will severely penalize serving high utility in a lightly loaded location while starving slice users in another location. This corresponds to the social well fair defined in the Fisher Market solution where the objective function is defined as follows

$$NSW(\boldsymbol{x}) = \Big(\prod_{i \in \mathcal{I}} u_i(\boldsymbol{x}_i)^{B_i} \Big)^{\frac{1}{B}}. \tag{4.9}$$

where B is the total budget for all SPs.

4.3.2 Market Equilibrium

We assume that the InP acts as a social planner or mechanism designer whose goal is to maximize the total social welfare regardless of the budget difference among SPs. Under the α-fair setting, the market is said to be at equilibrium if the supply provided by InP exactly matches SPs' demand, and each SP gets its favorite resource bundle. Even out of markets mentioned in the literature, probably the simplest one is the fisher market, where each SP owns the finite budget and SPs purchase the resources based on the linear pricing. For $\alpha = 1$, Eisenberg and Gale [28] showed that if the SPs utilities in the fisher market with $\alpha = 1$, then the market equilibria solution problem is equivalent to the Nash welfare optimization problem. In other words, resource allocation under fisher market equilibrium achieves optimal Nash welfare. An immediate question arises what if the social planner (InP) wishes to maximize a different welfare function. Motivated by this question, we focus on developing a pricing scheme for the market such that the market equilibrium induced through the proposed pricing scheme achieves various α fairness criteria. Here the pricing is associated with the capacity constraints (P_{SW}) and represents the rate of change of the objective function associated with any change in the capacity. Without any loss of generality, we assume that each resource type r is desired by at least one SP, and SPs have no value over their leftover money. We also consider that the total budgets of SPs (B_i) are normalized to one, i.e.,

$$\sum_{i}^{N} B_i = 1.$$

As discussed above, a fisher market adopts a linear pricing scheme under proportional fairness; in this work, we consider a more general pricing scheme under α-fair, where $\alpha \in [0, +\infty)$.

Definition 4.1. ***Price curves****: Let $\gamma_{ilr}(\boldsymbol{x}_i) : \mathbb{R}_+^{L \times R} \times \mathbb{R}_+ \rightarrow \mathbb{R}_+$ [34], is an increasing function denoting cost for purchasing of x_{ilr} amount of resource of type r at location l given bundle of resources*

$\boldsymbol{s}_i$ *purchased by SP* i*, thus the total cost for purchasing bundle* $\boldsymbol{x}_i$ *of resources*

$$C_{\gamma_i(\boldsymbol{x}_i)} = \sum_{l=1}^{L_i} \sum_{r=1}^{R} \gamma_{ilr}(\boldsymbol{x}_i) \tag{4.10}$$

we define the market
$\mathcal{M} := \left\langle \mathcal{N}, (B_i)_{i\in\mathcal{N}}, \bigcup_{l=1}^{L_i} \mathcal{R}_l, (u_i)_{i\in\mathcal{N}}, \boldsymbol{\gamma} \right\rangle$ *as follows:*

- *Player set: the set of service providers* $\mathcal{N}$
- *Budgets :*B_i
- *Resources set:* $\bigcup_{l=1}^{L_i} \mathcal{R}_l$
- *Utility: The utility of each SP* i *is equal to the* u_i
- *Price curve:* $\gamma_{ilr}(\boldsymbol{x}_i)$

Definition 4.2. *Allocation and price curve vector* $(\boldsymbol{x}, \boldsymbol{\gamma}(\boldsymbol{x}))$ *is called as Market Equilibrium (ME) of market* $\mathcal{M}$ *if the following conditions are satisfied.*

C1 Each $i \in \mathcal{N}$ *SP gets his favourite bundle* $\boldsymbol{x}_i$*,where*

$$\boldsymbol{x}_i : \underset{\boldsymbol{x}_i \geq 0; C_\gamma(\boldsymbol{x}_i) \leq B_i}{argmax} u_i(\boldsymbol{x}_i) \tag{C1}$$

C2 The demand $\boldsymbol{x}$ *meets the supply or the market is cleared,* i.e.,

$$\sum_{i\in\mathcal{N}} x_{ilr} \leq c_{lr} \quad \forall l \in \mathcal{L}_i, \forall r \in \mathcal{R} \tag{C2}$$

and the inequality (C2) is saturated if $\gamma_{ilr} > 0$.

4.3.3 Fisher Market Equilibrium Price

This section contains the main results of the problem defined in 4.4. The resulting resource allocation is fair among the slices, which eliminates the starvation of SPs with lower budgets, and the fairness level can be steered by a factor α. It is important to notice that α = 0 yields worst fairness as the resource allocation has a strong dependency on the budget of SPs, if all the budgets are equal then, resources will be allocated in proportion to the demand vector d_{ilr}. If budgets are not equal then it is likely that more resources are allocated to the slices that belong to SPs with higher budgets, this could cause starvation for other slices. On the contrary, a high α leads to a fair allocation among SPs while degrading efficiency. Hence, InP should strike a balance between fairness and efficiency by selecting α value depending on the scenario.

Theorem 4.3. *There exist a price curve vector $\boldsymbol{\gamma}$ and associated market equilibrium $(\boldsymbol{x}, \boldsymbol{\gamma}(\boldsymbol{x}))$ for the market $\mathcal{M}$ such that the allocation $\boldsymbol{x}$ maximizes the social welfare (4.4) and the price curve for each resource type r at location l for SP i is characterized as*

$$\gamma_{ilr}(\boldsymbol{x}) = p_{lr} x_{ilr} \Big(\sum_{l \in \mathcal{L}_i} \frac{x_{ilr}}{d_{il} a_{ir}} \Big)^{\alpha - 1}, \ \text{if } d_{il} a_{ir} > 0 \tag{4.11}$$

where p_{lr} is the Lagrangian multiplier associated with the capacity constraints in (4.4) and u_i is the utility of each SP i.

Proof. See Appendix B □

We observe that the structure of the price function reflects the goal of α-fair. When α is higher, i.e. $\alpha > 1$, the price of a resource increases faster proportional to the total utility of SP. Therefore, the higher the α is, the more we care about SP with low utility. When $\alpha = 1$, which corresponds to proportional fairness (PS) or Nash bargaining allocation, the price function becomes linear and the resource allocation corresponds to a mix of fairness and efficiency. Indeed, under Under PF, if compared to any other feasible allocation of utilities, the aggregate proportional change is less than or equal to zero. When $\alpha < 1$, we have the opposite behavior since the smaller α is, the more we care about SP with high utility.

4.3.4 Two-level resource allocation

So far, we have focused on achieving the α-fair resource allocation between the SPs. However, we observe that the utility u_i considered in the problem 4.4 is additive in location-based utility and could result in unbalanced resource allocation towards the locations supported by each SP. To overcome this, we propose a slightly modified problem that accounts for fairness among slices and locations supported by each slice. To be more specific, we apply proportional fairness among locations served by each SP. Here, the most important challenge is the resource allocation among locations with a fairness guarantee for end-users.

We replace the linear utility function 4.5 with a slightly tweaked CES utility function 4.13 in the following model to incorporate fairness between different locations of a SP as well, $\beta \in [0, \infty)$

determines the fairness level.

$$\text{Maximize}_{\boldsymbol{x}}: \quad \sum_{i\in\mathcal{N}} B_i U(u_i) \tag{4.12}$$

$$\text{subject to} \quad u_i = \Big(\sum_{l\in\mathcal{L}_i} (u_{il})^{1-\beta}\Big)^{\frac{1}{1-\beta}} \quad \forall i \in \mathcal{N}, \tag{4.13}$$

$$u_{il} \leq \frac{x_{ilr}}{d_{il}a_{ir}} \quad \forall i \in \mathcal{N},\ \forall l \in \mathcal{L}_i,\ \forall r \in \mathcal{R}, \tag{4.14}$$

$$\sum_{i\in\mathcal{N}} x_{ilr} \leq c_{lr}, \quad \forall l \in \mathcal{L}_i,\ \forall r \in \mathcal{R}. \tag{4.15}$$

The market equilibrium of (4.12) subject to the constraints 4.13-4.14 ensures fair allocation among SPs as well as locations of a SP. Now, we study how the new formulation of the fisher market influences the price curves.

Proposition 1. *There exist a price curve vector $\boldsymbol{\gamma}$ and associated market equilibrium $(\boldsymbol{x}, \boldsymbol{\gamma}(\boldsymbol{x}))$ for the market $\mathcal{M}$ such that the allocation $\boldsymbol{x}$ maximizes the social welfare (4.4). and the price curve for each resource type r at location l for SP i is characterized as*

$$\gamma_{ilr}(\boldsymbol{x}) = p_{lr}x_{ilr}\left(\Big(\sum_{l\in\mathcal{L}_i}\Big(\frac{x_{ilr}}{d_{il}a_{ir}}\Big)^{1-\beta}\Big)^{\frac{1}{1-\beta}}\right)^{\alpha-1}, \quad \text{if } d_{il}a_{ir} > 0, \tag{4.16}$$

where p_{lr} is the Lagrangian multiplier associated with the capacity constraints in (4.4).

Proof. See Appendix B □

A fundamental aspect of the price function obtained in Proposition 1 is the systematic exploitation of heterogeneity of traffic in different location.

4.3.5 Strategic service providers

In a strategic setting, SPs are players, and may report a strategy profile s_{ir} instead of reporting the true preferences a_{ir}, in an attempt to gain larger utility. The use of a plain Fisher market mechanism in this situation for resource allocation induces game $\mathcal{G}$. There are N SPs and R resource types; for the sake of simplicity we drop the dependency on location. Now the utility of SP writes

$$u_i' = \min_{r\in[M]}\left\{\frac{1 - s_{2r}\left(\frac{B_2}{s_2^{max}}\right)^{\frac{1}{\alpha}}}{a_{1r}}\right\}. \tag{4.17}$$

By inversion, it is possible to compute the expression for the best response strategy of a SP to other players actions [17]. An explicit formula for a two player market with strategies s_1 and s_2 is the following

$$s_{1r} = 1 - s_{2r} \left(\frac{B_2}{s_2^{\max}} \right)^{\frac{1}{\alpha}}. \tag{4.18}$$

where $s_i^{\max} = \max_{r \in \mathcal{R}} \{s_{ir}\}$.

Let $\boldsymbol{s}^* = (\boldsymbol{s}_1{}^*, \ldots, \boldsymbol{s}_N{}^*)$ be the strategy profile for the market under strategic players and $\mathcal{S}_i$ be the strategy space for player i. The standard definition of equilibrium is as follows,

Definition 4.4. *Strategy profile $\boldsymbol{s}^*$ is a Nash Equilibrium (NE) of the game $\mathcal{G}$ if*

$$\forall i \in \mathcal{N},\, U_i(\boldsymbol{s}_i{}^*, \boldsymbol{s}_{-i}{}^*) \geq U_i(\boldsymbol{s}_i, \boldsymbol{s}_{-i}{}^*), \boldsymbol{s}_i \in \mathcal{S}_i \tag{4.19}$$

Here, $(\boldsymbol{s}_i, \boldsymbol{s}_{-i}{}^*)$ denotes the strategy profile with i^{th} element equals s_i and all other elements equal $s_{i'}{}^*$ (for any $i' \neq i$).

Theorem 4.5. *A uniform strategy ($s_{ij} = \frac{1}{M}\ \forall j \in [M]$) is a Nash Equilibrium for the given Fisher market game with leontief utilities where the player utilities are $u_i = \frac{B_i^{\frac{1}{\alpha}}}{\sum_{k=1}^{N} B_k^{\frac{1}{\alpha}} d_i^{max}}$*

To, see the effect of player's strategic nature, we use *price of anarchy* (POA), which is defined as the ratio between the worst NE social welfare and the optimum social welfare.

Proposition 2. *The PoA of the Fisher Market under α-fair utilities is $\frac{1}{N}$*

Proof.

$$PoA = \frac{\sum_{i=1}^{N} \frac{1}{1-\alpha} B_i (u_i^*)^{1-\alpha}}{\sum_{i=1}^{N} \frac{1}{1-\alpha} B_i (u_i^{NE})^{1-\alpha}},$$

$$= \frac{\sum_{i=1}^{N} B_i (\frac{1}{d_i^{max}})^{(1-\alpha)}}{\sum_{i=1}^{N} B_i \left(\frac{B_i^{\frac{1}{\alpha}}}{\sum_{k=1}^{N} B_k^{\frac{1}{\alpha}} d_i^{max}} \right)^{1-\alpha}}.$$

We consider that the budgets B_i of the SPs are equal, then we have

$$PoA = N. \tag{4.20}$$

□

The above proposition extends the result in [17], which is covering just the case $\alpha = 1$. Here we observe the PoA does not depend on α. Hence, no value of α improves the efficiency when SPs behave strategically.

API Name	Bandwidth (Gbps)	vCPU	Memory (GB)	Instance Type
r4.8xlarge	10.00	32.00	244.00	Memory optimized
r4.16xlarge	25.00	64.00	488.00	Memory optimized
m4.10xlarge	10.00	40.00	160.00	General purpose
m4.16xlarge	25.00	64.00	256.00	General purpose
c5.9xlarge	10.00	36.00	72.00	Compute optimized
c5.18xlarge	25.00	72.00	144.00	Compute optimized
c4.8xlarge	10.00	36.00	60.00	Compute optimized

Table 4.1: API instances from AMAZON EC2

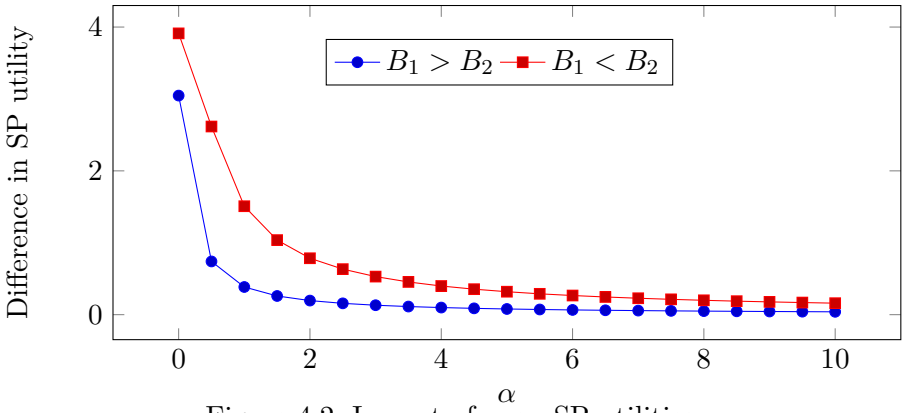


Figure 4.2: Impact of α on SP utilities.

4.4 Numerical evaluations

In this section, we provide numerical results to support the mechanisms that we have described so far. We consider Amazon EC2 instances [8] to compute numerical results, some of these instances are described in table 4.1.

We consider first a simple set up with two slices, both slices cover demands at two locations and each SP needs three resource types to run their services. Assume that each SP provides two application services with the APIs mentioned in the table, SP1 supports API `m4.10xlarge` and `m4.16xlarge`, SP2 supports API `c5.9xlarge`, and `c5.18xlarge`. In order to generate Monte Carlo simulations, we let each SP support services at each location with given probability p_{kl} per generated instance, where, p_{kl} is the probability to support API k at location l. SP i has fixed budget B_i (normalized to 1). Capacity c_{lr} is available at each location l for resounce r, and resource preferences are aggregated for both SPs as they support only one service at the time:

$$a'_{ir} = p_{1l}a^1_{ir} + p_{2l}a^2_{ir}, \quad l \in \mathcal{L}_i \tag{4.21}$$

Budget B_i, α, and demand d_{il} per location l has a great impact on the resource allocation for the slice and in turn, impact the utility of the SP. Hence, we focus on the impact that these factors have on resource allocation. For the remainder of this section, we consider the same preferences evaluated using (4.21) based on the Amazon instances mentioned earlier. First we provide the impact of α in tuning the degree of fairness, then the impact of budget and parameter α. Finally we show the impact of demands by imposing higher fairness at both levels.

4.4.1 Impact of factor α

As described throughout the article, the factor α tunes fairness in the proposed mechanism, imposing different degrees of fairness across the slices. Instead of comparing the utilities U_1 and U_2 side by side, we prefer to observe the change in $|U_1 - U_2|$, which provides a better illustration for observing the fairness variation. Figure 4.2 displays this by varying α over the range 0-10 with an interval of 0.5. Figure 4.2 shows that at the increase of α, the difference between the utility of SP decreases. I.e., the allocation of resources is more and more fairly distributed, thus vanishing the difference between U_1 and U_2. We compare two scenarios to show the consistent behavior of the fairness: first, we let $B_1 > B_2$, uniform demand across locations, and the resource preferences are as mentioned earlier (aggregated for the two APIs). In the second scenario SP 2 has higher budget. In both cases, higher α leads to higher fairness. Though the behaviour appears similar for both scenarios, the curve in the first scenario decreases sharply than the other curve. In fact, the utility depends on resource preferences as well, and not only on the budget. The preference for resources are higher for slice 1 than for slice 2. Based on the numerical results, we observe that the value of β does not play a role because corresponding fairness is at the location level not at the slice level; the objective of the comparison is at slice level.

4.4.2 Impact of SP budget B_i

To show the impact of purely the budget under higher fairness criteria, we consider uniform demands ($d_{il} = 1$), $\alpha = 10$, $\beta = 5$. When α and β are zero, allocations depend for the major part on budget. A trivial solution is for the SP with the higher budget to get all the resources, starving all other SPs' users due to lack of resources, as known in the literature for the utilitarian approach. The proposed scheme, as reported in figure 4.3, avoids starvation effects both at the slice level and at the location level, whereas the SP with higher budget gets relatively higher resources. In the same figure, it can be observed that slice 2 is not allotted with higher resources compared to slice 1 even when the SP2 budget is higher than SP2, the reason being that the allocation depends not only on the budget but also on their resource preferences as described earlier.

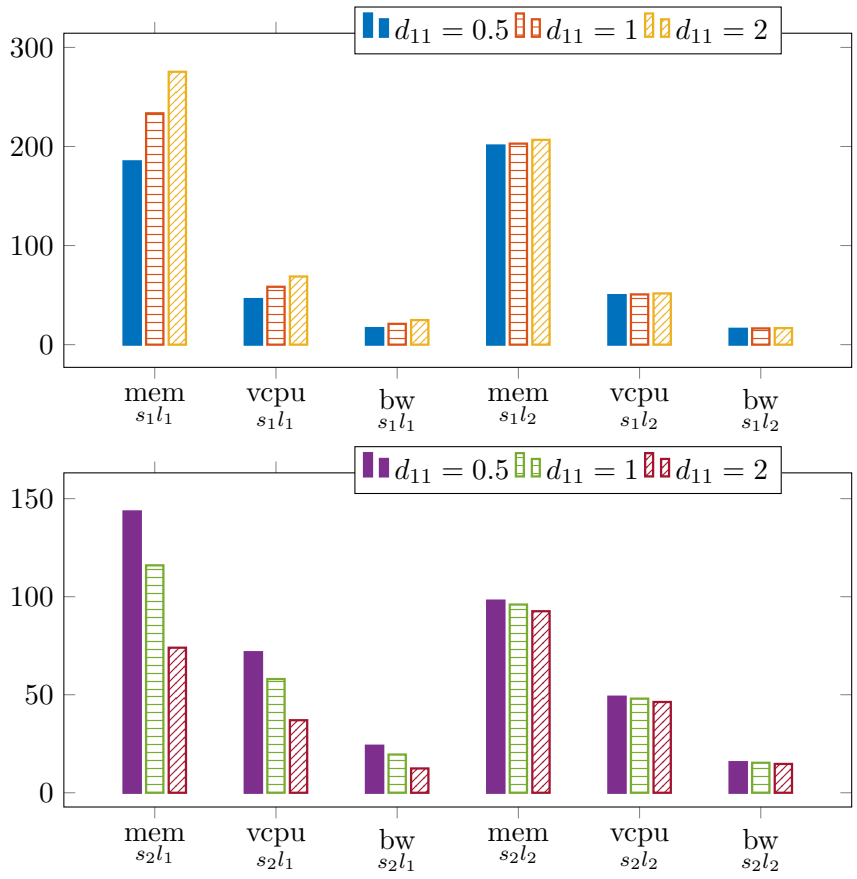


Figure 4.3: Resource allocation for slice 1 and slice 2 respectively, with change in SP budgets.

4.4.3 Insights of user demand d_{il}

Figure 4.4 displays the results for different demands of SP1 at location 1. Here, $B_1 > B_2$, resource preferences as in previous experiments and the user demands d_{12} = 1, d_{21} = 1.5, and d_{22} = 1; α = 10, β = 5. It can be observed that as d_{11} increases the allocation for location 1 served by SP1 increases as well, at the same time resources for location 1 of SP 2 is decreasing to satisfy the capacity constraint 4.7. But the increment in allocation is rather low due to fairness imposed by α and β.

4.5 Conclusion

Resource allocation has always been a very challenging aspect in telecommunication due to the scarcity of resources and high demand from the users, even more so in the context of 5G, which is by far the most diverse and complex system. In this work, we investigated a method to cope up

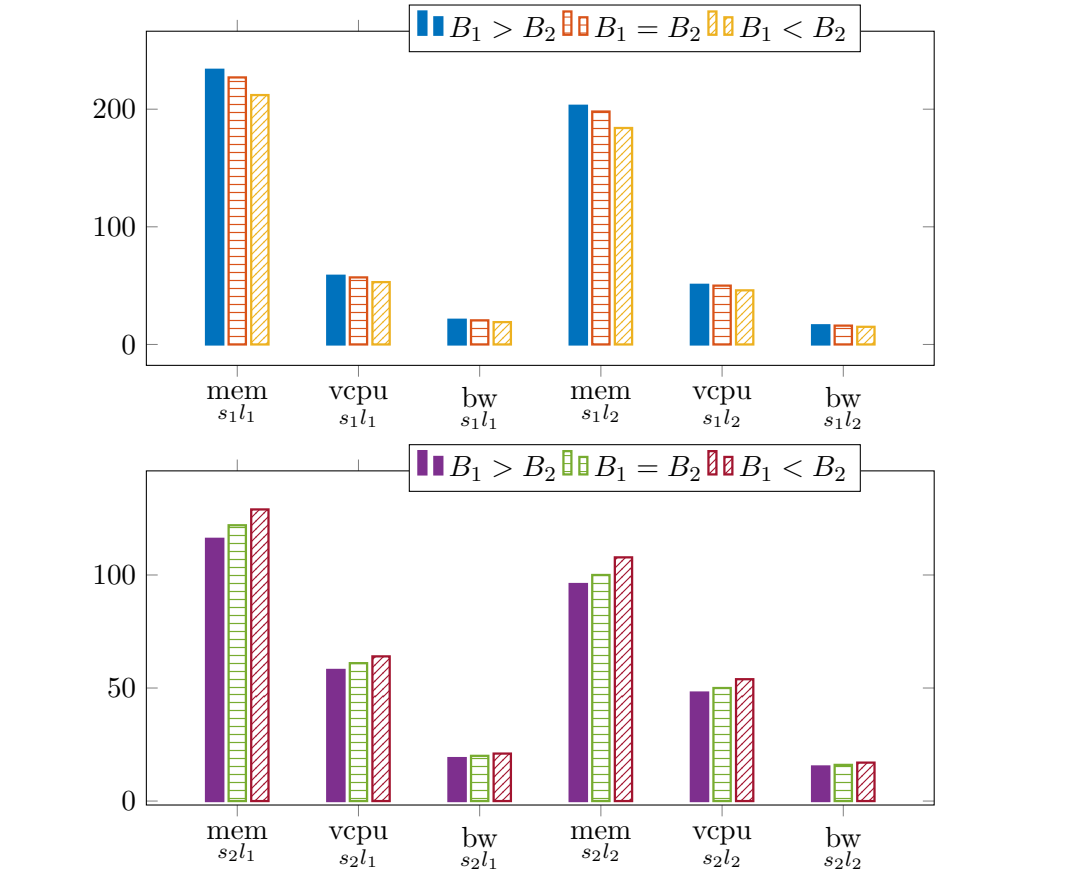


Figure 4.4: Resource allocation for slice 1 and slice 2 respectively, with change in the demand at location 1.

with those challenges by emphasizing the fact that the resources should be fairly allocated to the SPs competing for the resources. We demonstrated that our model is capable of allocating multiple resource types that are fair among slices and among the locations of a slice with help of the factors α and β respectively, which act as control parameters to balance the trade-off between fairness and efficiency. On top of this, by defining POA, we have shown that the social welfare deteriorates significantly when players strategically report false preferences. This unified framework that focuses on a non-linear pricing scheme to allocate resources with multi-level fairness can be applied to various scenarios under the 5G network slicing phenomenon. This centralized approach for resource allocation has several drawbacks: central entity that is responsible for resource allocation needs need information such as resource capacity and utility functions which can be sensitive to InPs and SPs respectively, and it does not allow InPs to choose their own resource allocation rule for allocation. For these reasons we propose a decentralized resource allocation mechanism in the next chapter to address these issues.

Chapter 5

Privacy preserving decentralized resource allocation mechanism

5.1 Introduction

Web and mobile applications have seen exceptional growth in both quality and quantity due to the excellent support provided by companies such as Google, Microsoft, Amazon, and others through cloud services. They use cloud computing to remove the hardware bottlenecks for consumers and enterprises. This idea paved the way for industries to use centralized resources to run their operations. However, recent developments in unnamed areal vehicles (UAV) [47, 32], and autonomous driving [64] to name a few, require high reliable connections with excellent connectivity and very low latency. Edge computing[93] and fog computing [111] have been developed to achieve nearby and assured level computational capacity. In this, most of the resources supported by the core cloud are shifted closer to serving locations. Other services that are data intensive transmit large amounts of data to the central cloud, which can increase the congestion along network paths, thus leading to increased delay and re-transmissions where available bandwidth gets reduced. Instead, edge cloud can pre-process the data locally, eventually send the result to the cloud for computation or storage, hence reducing the strain on the central cloud and routes leading to it. This combination of centralized and distributed placement of resources is being extensively adapted in the current and next generation networks.

Cloud computing also ease the softwarization of the network functions (or Network Function Virtualization, NFV), increasing the flexibility reachable to satisfy current and future needs of network Service Providers (SPs) and end-users. Complementary architectures such Software defined networking (SDN) can further provide flexible routing while moving the control from the devices to a central entity; this eliminates the need for individual update of the routing tables and protocol metrics, which can be prone to errors [99]. Such network softwarization evolutions extend the capabilities of communication networks, starting notably with the 5^{th} generation of cellular networks, and beyond 5G. A significant benefit of incorporating the network softwarization is the novel idea called Network Slicing that logically separates the virtualized resources involved in the end-to-end network provisioning, creating a logical network partition called network slice [46]. Network slicing

enables the independent and flexible scaling of the resources for each SP, in a chain of SPs involved in network service provisioning [66]; this helps the SPs to obtain resources that are tailored to their service needs.

Generally, the cellular network is segregated into Radio Access Network (RAN) and Core Network (CN). RAN provides the access to the User-Equipments (UE), RAN functions such as Base Band Units (BBU) perform digital signal processing on the received signals. Cloud-RAN (C-RAN) allows the placement of BBUs on central or edge cloud [20]; this allows C-RAN to dynamically scale the resources based on the requirements. Other functionalities such as Access and Mobility Management (AM), Session Management (SM), authentication, user and data plane management that are part of the core network can also be performed in the similar approach[2]. Further RAN function disaggregation is envisioned in novel Open-RAN architectures [94]. This architectural desegregation helps the SPs to efficiently manage radio resource, also in coordination with the orchestration layer and other resource controllers.

Such diversity in the service types creates heterogeneous resource requirement which poses severe challenges to the network design; 5G wireless networks have been developed to address these challenges. Three main use-case scenarios, namely, enhanced mobile broad band (eMBB) [4], massive machine type communication (mMTC)[26], and ultra reliable low latency communication (URLLC) [21] have been created in 5G to address the needs of ever increasing business ideas such as smart cities, autonomous driving, smart factories, online gaming, high resolution streaming services, etc [84, 103]; additional services are being introduced with 6G, namely, ubiquitous Mobile UltrabroadBand (uMUB), ultraHigh-Speed-with-Low-Latency Communications (uHSLLC), and ultraHigh Data Density (uHDD), with possible integration of undersea, terrestrial, non terrestrial and satellite networks. Each use case is designed to address the specific needs of the service providers (SP) such as low latency, higher throughput, and ability to handle massive connections to support IoT devices. For instance, 5G supports a modified data frame with introduction of frame numerlogies to enable multiple sub carrier spacings ranging from 15kHz to 960kHz [1]; this further expands the capabilities of the 5G network as it can facilitate higher bandwidths by using millimeter Waves[87].

Each use-case mentioned earlier require multiple resources in a specific proportion to maintain the Quality of Service. To this aim, SPs have to jointly provide various resources namely, CPU, RAM, storage, bandwidth, from the infrastructure providers (InP) that can own one or multiple resources resources. In some cases, an SP may be an InP as well, typically for the radio access domain, simplifying the end-to-end resource provisioning. Nonetheless, given the changes in the data frame structure and network architecture, the resource allocation has never been more complex and challenging, as for the 5G and beyond-5G or 6G environments. Consider that each SP can be allocated with one or multiple dedicated network slices for each communication service offered. From here on, resource allocation to the network slices refers to the resource allocation for the SPs and network slice is referred as slice for brevity.

In this new communication service provisioning landscape, multi-resource provisioning systems are therefore needed. Open-RAN, in conjunction with NFV and SDN systems, is in this respect opening the scheduling logic at different resource controllers to facilitate multi-resource integration. At the state of the art, the majority of the multi-resource allocation rules are based on a centralized approach [74, 14, 31, 80] as shown in figure 5.1: a central slice orchestrator mediates between SPs and InPs to obtain the best possible outcome for both sides; both parties prefer that their interests are better served, in that each InP tries to maximize revenues with the efficient use of the resources, and the SPs prefer slice isolation. As described in the previous chapter, with network slicing, the slice orchestrator has the burden of dynamically allocating the resources that best serve the interests of InPs and SPs. In chapter 4, we introduced one such centralized resource allocation framework with multi-level fairness. However, this approach has several drawbacks [37, 80]:

- To realize the resource allocation through central allocation rules, the orchestrator needs information such as available resource capacity from InPs and utility functions of the SPs. This is private information that they can be reluctant to share.
- With dynamic changes in the network due to varying growth in the number of slices, scalability is a big issue for the orchestrator.
- Finally, as with any centralized solution, a single point of failure is an inevitable drawback.

To address these issues, distributed resource allocation rules were developed [60, 53], where each SP can directly communicate with InPs and obtain the resources as shown in the figure 5.2 without the need to disclose private information. Thus, said distributed multi-resource allocation rules help preserve privacy on multiple ends. In this chapter, we investigate the following question:

1. Does decentralized resource allocation mechanism achieve social optimal where each InP can implement an independent resource allocation scheme?

5.2 Related Work

Resource allocation in wireless networks is a well-studied area of research[38], however, with the constantly evolving network architectures and heterogeneous requirements of the SPs, the models that allow for efficient and fair resource allocation have to be proposed to address these changes. Despite all the complex requirements, the underlying phenomenon remains the same, i.e., infrastructure provider wants to maximize their revenues by leasing the resources and the clients try to procure resources at the best price to minimize their costs. There have been many notable works that address this problem with different approaches, in [79, 76], the authors obtain an optimal resource allocation with the use of Eisenberg - Gale convex programming [28] to solve the proposed

market model that employs proportional fairness. With auction-based models [69], SPs with better marginal utilities obtain a higher share of resources, whereas the authors in [**Ali**] employ max-min fairness that allocates as much resource as possible to the SPs with weaker marginal utilities to make the resource allocation extremely fair. A common observation in all these models is that there is no flexibility to choose an appropriate fairness model depending on the requirement. But authors in [73] propose an α - fairness based model that provides the necessary flexibility.

A common drawback of the above-mentioned works is that they employ central allocation models that lead to privacy concerns for InPs and SPs. F P Kelly [60] proposed a distributed resource allocation model that suppresses the need for resource providers and procurers to reveal sensitive information. R. Johari [53] extended this work to the network with multiple resources. The authors of [30] provide a nice framework to allocate resources in a distributed manner but it does not guarantee particular fairness for the proposed solution. The model proposed in [37] based on the Kelly mechanism, employs an α - fair allocation mechanism that achieves optimal allocation. However, there is no particular control for the individual InP to choose their own allocation model.

Main contributions:

- Given the privacy concerns for SPs and InPs alike, we design mechanism for optimal resource allocation in a decentral approach. This mechanism allows each InP to independently choose an allocation rule for each associated resource by choosing a specific value for parameter β_r. We define an optimization problem for each SP that can be locally solved to obtain the resource allocation in conjunction with InPs who allocate resources based on its allocation rule. We prove that such a mechanism results in an allocation that can optimally solve the allocation scheme employed by the central system.

- We provide two resource control algorithms that can be implemented at InP for obtaining an optimal pricing and resource allocation. These algorithms are easy to implement in the current network architectures or future O-RAN enabled networks. In addition, we provide the proof for the convergence of the algorithms. Finally, We provide numerous simulation results to back our theoretical claims which include the convergence of resource allocation, impact of α on the overall allocation.

Organization of the chapter:

Section 5.3 consists of the system description, utility function definition and detailed interpretation, and the resource allocation rule for each InP. Section 5.5 contains the mechanism designs corresponding to the two prominent pricing schemes with price taking SPs and associated theorems to prove the existence of equilibrium and optimal solutions. Section 5.6 comprises of several simulated results to validate the theoretical claims and it concludes with final remarks and future directions.

Notations	*Description*
i	service provider identifier
m	infrastructure provider identifier
r	resource type indicator
$\mathcal{N}$	set of service providers; $\mathcal{N} = \{1, 2, \ldots, N\}$
$\mathcal{M}$	set of infrastructure providers; $\mathcal{M} = \{1, 2, \ldots, M\}$
$\mathcal{R}$	set of information age values; $\mathcal{R} = \{1, 2, \ldots, R\}$
C_r	capacity of resource type r
x_{ir}	allocated quantity of resource type r to SP i
x_i	resource bundle allocated to SP i
d_{ir}	base demand of SP i for resource type r
u_{ir}	utility of SP i for resource type r (single resource)
u_i	utility of SP i for obtaining multiple resource
b_{ir}	bid submitted by SP i for resource type r
b_i	vector of bids corresponding to SP i
p_r	price of resource type r (uniform price)
p_{ir}	price of resource type r for SP i (differential pricing)

Table 5.1: Table of notation

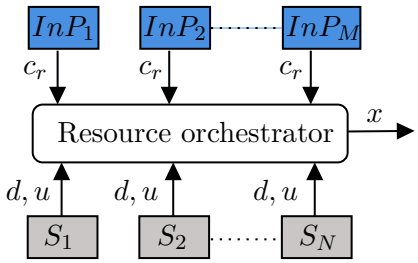


Figure 5.1: Central resource allocation example with N slices and M InPs.

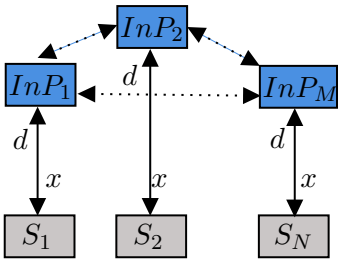


Figure 5.2: Distributed resource allocation with N slices and M InPs.

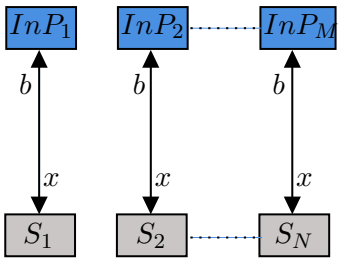


Figure 5.3: Decentralized resource allocation with N slices and M InPs.

5.3 System Model

Consider a 5G system with multiple service providers $\mathcal{N} = \{1, 2, \ldots, n, \ldots N\}$, where each SP requires different resource types $\mathcal{R} = \{1, 2, \ldots, r, \ldots R\}$ to run their services and meet users' demands. There are several InPs $\mathcal{M} = \{1, 2, \ldots, m, \ldots M\}$, each of them owns at least one resource with an available capacity c_r; let us assume that no resource is commonly owned by two InPs. Each InP has the freedom to employ a specific allocation mechanism. Figure 5.4 shows the architecture of the network with the introduction of O-RAN. We can easily extend the general description provided in this section to the new architecture by considering that the data centers in different parts of the network are supported by different InPs. Every SP has a base demand d_{ir} that conveys the SP's preference for a specific resource r to maintain the service rate. The required resources are

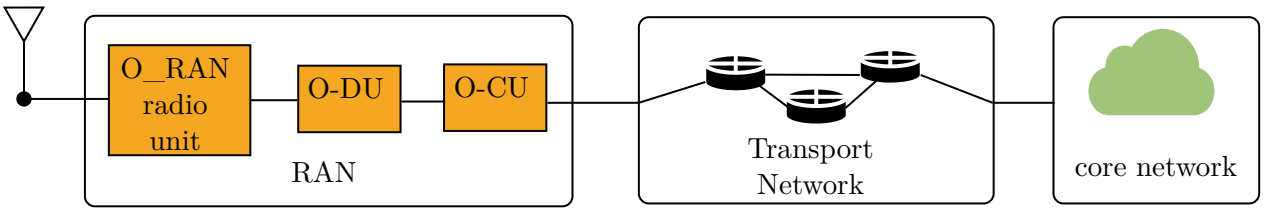


Figure 5.4: Schematic representation of Beyond 5G network architecture with open-RAN.

procured from various InPs depending on the service requirements. Before establishing the problem formulation, it is crucial to qualify an utility function of the SP as it plays a key role in the mechanism design for an optimal resource allocation. In addition, formulating a precise utility function can provide a measure to understand the service rate based on the amount of resource obtained from InPs.

In this work, we consider that a feasible resource allocation vector x satisfies following conditions [102].

- Non negativity: the amount of resource type r allocated to SP i is non-negative.

$$i.e., \quad x_{ir} \geq 0 \quad \forall\, i \in \mathcal{N};\ \forall\, r \in \mathcal{R}, \tag{5.1}$$

- Capacity limitation: the total allocation does not exceed the available capacity.

$$i.e., \sum_{i=1}^{n} x_{ir} \leq c_r. \tag{5.2}$$

5.3.1 service utility function

For any given SP, the service rate is crucial to cope with the user demand. Hence, the SP must procure enough resources to maintain the desired service rate. It is important to define a service utility function that quantitatively determines the service rate based on the resources obtained by the SP and the base demand. From here on, the service utility function is referred to as utility for brevity. The utility of the SP is defined as follows, depending on the number of resources required.

Single resource utility: in this case, it is trivial to see that the ratio of the obtained quantity of resource x_i and the base demand d_i is the utility of the SP i.

$$i.e, u_i = \frac{x_i}{d_i} \tag{5.3}$$

Multi-resource utility: services that are currently offered by the SPs require more than one resource, hence the definition of the utility shall depend on the relation of resource types and the

service. There is a well-known class of utility functions that is referred to as the constant elasticity of substitution (CES) function [9]. It is defined as follows:

$$u_i = \left(\sum_{r=1}^{R} u_{ir}^{\rho} \right)^{\frac{1}{\rho}} \tag{5.4}$$

where ρ ranges from $-\infty$ to 0. The factor ρ determines the utility type of the SP; it originates from the field of economics, used to measure buyer satisfaction based on the amount of goods purchased.

In general, there are three popular cases that are well studied and extensively used in the literature [17]: (i) linear ($\rho \to 1$), (ii) Cobb-Douglas ($\rho \to 0$), and (iii) Leontief ($\rho \to -\infty$) functions. The Leontief one captures the utility of goods that are perfect complements, that is a set of goods lose its value without another set of goods, for example, a pair of shoes. Buying any number of the left shoe has no value for the buyer if the same number of right shoes are not purchased; as resources in wireless networks are known to be perfect complements (to execute a specific vnf, the slice requires resources such as bandwidth, cpu, memory. Without any of the mentioned resources the vnf can not be executed), we use as utility one defined based on the Leontief function, as follows

$$i.e., u_i = min\{u_{ir}\} = min\{\frac{x_{ir}}{d_{ir}}\} \quad \forall\, r \in \mathcal{R} \tag{5.5}$$

So the increment in utility is a function of the proportion of the obtained resource: the utility increases only when all the resources are increased by the same proportion, which is the reason why the utility is defined as the minimum of the ratios of the allocated resource x_{ir} and the SP base demand d_{ir}.

Example: for illustration purposes, let us consider that an SP requires 3 resource types, bandwidth, vCPUs, and memory, to run a service. Assume that the base demand vector is d = (2 Gbps, 4, 200 Mb). The following cases represent 3 possible allocations highlighting the impact of increased/decreased allocation on utility:

- case 1: the allocated resource vector is x = (1 Gbps, 2, 100 Mb).

$$u_i = min\{\frac{1}{2}, \frac{2}{4}, \frac{100}{200}\} = 0.5$$

- case 2: x = (1 Gbps, 4 100 Mb), SP does not get additional benefit despite receiving additional vCPUs as the service rate is limited by the other two resources. This can be observed with the computation of utility,

$$u_i = min\{\frac{1}{2}, \frac{4}{4}, \frac{100}{200}\} = 0.5$$

- case 3: x = (2 Gbps, 4, 200 Mb);

$$u_i = min\{\frac{2}{2}, \frac{4}{4}, \frac{200}{200}\} = 1$$

now the utility of the SP is 1. In this case, all the resources increased in proportion by twofold, hence the utility is increased by twofold.

It is clear to see that the SP utility increase only when all the resources increase in proportion, as the Leontief function takes a minimum of ratios between the base demand and allocated resources. It is not necessary that all of the SPs require every available resource type; to address this, an SP can submit a base demand of zero for any resource that is not desired by the SP. This ensures that the utility is computed based on the resources that are compulsory for SP.

5.3.2 InP allocation rule

We consider that all InPs have the freedom to choose their own allocation rule and based on the information provided by the SPs that does not reveal any private details, InP can allocate resources to each SP that has a non-negative preference for a resource referred to as a bid in this work. In this work, we consider that each InP implement α - fair allocation rule which is discussed in detail in the next section. We use the notation β_r instead of α while referring to the allocation rule for InP, hence the rule remains the same except for the notation. Each InP can propose a specific value for β_r and determines the resource quantity x_{ir} to be allocated to each slice. In this model, we assume that the vnf placement has already been defined. Every slice must evaluate the amount of resource that can be procured from each InP who owns the corresponding resource to meet the SLA of the respective SP. To this aim, each slice sends a signal containing the bid value of the required resource to the corresponding InP. By receiving all the bids from slices, the InP then computes the optimal resource bundle x_i to be allocated to each slice. This information is then communicated back to the corresponding slice.

Let $\boldsymbol{b}_r = [b_{ir}]_{i\in\mathcal{N}}$ be the vector of demands (referred to as bids*) submitted by each SP $i \in \mathcal{N}$ for a resource $r \in \mathcal{R}$, $\boldsymbol{x}_r = [x_{ir}]_{i\in\mathcal{N}}$ is a vector of allocated quantity of resource type $r \in \mathcal{R}$ for every SP $i \in \mathcal{N}$ that submitted a positive bid. The following optimization problem is the mathematical representation of such an allocation method, as considered in this work, for each InP $m \in \mathcal{M}$.

*In this mechanism, bids acts as signaling between the SP and InP to obtain an optimal resource allocation with out the need for SPs to reveal utility function

$$INFRA(\boldsymbol{b}_r, \phi_r, \boldsymbol{x}_r) : \underset{x}{\text{maximize}}\ \mathcal{H}_m : \quad \sum_{i \in \mathcal{N}} b_{ir} \frac{1}{1-\beta_r} x_{ir}^{1-\beta_r} \tag{5.6}$$

$$\text{subject to} \quad \sum_{i \in \mathcal{N}} x_{ir} \leq c_r \tag{5.7}$$

The objective of the above optimization problem is to compute an optimal resource allocation vector $\boldsymbol{x}$ for based on the bid values submitted by the slices. The parameter β_r allows each InP to implement a specific allocation rule for the slices. More details on this are provided in the following section. The constraint (5.7) ensures that the addition of allocated resources x_{ir} does not exceed the available capacity c_r. In the following, we write the closed form expressions for obtaining the resource quantity x_{ir} by each InP and the corresponding price ϕ_r which is a Lagrange multipler associated with the capacity constraint (5.7). In this work, we consider this as a price charged by the InP as it ensures that the aggregated resource allocation $\sum_{i \in \mathcal{N}} x_{ir}$ does not exceed the available capacity c_r of a resource r. The price ϕ_r reaches a significantly high value as aggregated demands of SPs approaches available capacity c_r, on the contrary it reaches zero as the aggregated demand is considerably lower than available capacity.

Now, we write the Lagrangian for INFRA($\boldsymbol{b}_r, \phi_r, \boldsymbol{x}_r$)

$$L_{infra}(\boldsymbol{x}_r, \phi_r) = \sum_{i \in \mathcal{N}} \frac{1}{1-\beta_r} b_{ir} x_{ir}^{1-\beta_r} - \phi_r \left(\sum_{i \in \mathcal{N}} x_{ir} - c_r \right). \tag{5.8}$$

by writing the first order necessary and sufficient conditions, we have

$$\frac{\partial L_{infra}}{\partial x_{ir}} = b_{ir} x_{ir}^{-\beta_r} - \phi_r = 0.$$

From the above equation, we can write the closed form solution for resource allocation x_{ir} by an InP to each SP $i \in \mathcal{N}$

$$x_{ir} = \left(\frac{b_{ir}}{\phi_r} \right)^{\frac{1}{\beta_r}}. \tag{5.9}$$

By applying sum on both sides, we have the following,

$$1 = \sum_{i \in \mathcal{N}} \left(\frac{b_{ir}}{\phi_r} \right)^{\frac{1}{\beta_r}}$$

With that we can write the closed form expression for the price of a resource

$$\phi_r = \left(\sum_{i \in \mathcal{N}} b_{ir}^{\frac{1}{\beta_r}} \right)^{\beta_r}. \tag{5.10}$$

Each InP returns to every SP that submitted a positive bid b_{ir}, a price ϕ_r and x_{ir} based on

eq.(5.10) and (5.9) respectively.

EXAMPLE: Consider that there are 3 InPs, each provide one resource type and 3 SPs procuring the resources from each InP to support the corresponding services. InP_1 provides bandwidth and implements proportional fairness (β_r = 1) while allocating the resources to the SPs. InP_2 provides CPU and implements max-min fairness ($\beta_r \to \infty$) among the competing SPs. InP_3 provides Memory and employ proportional fairness criterion while allocating the resources to SPs. As per this scheme, the InPs are not forced to choose a specific allocation rule. In the following sections, we explain the possibility of obtaining such different allocation rules by selecting an approapriate value for β_r.

5.4 Social welfare function

Each SP shall obtain a slice to run its services with a certain QoS. To obtain the resources, every SP submit their demand d_{ir} to the resource orchestrator (RO), who is responsible for resource provisioning for the associated slice. The slice shall obtain the resource bundle $\boldsymbol{x}_i = [x_{ir}]_{r\in\mathcal{R}}$ that maximizes the SP utility u_i. Hence, the resource orchestrator shall employ an allocation mechanism that computes such resource quantities for each slice corresponding to the submitted demands d_{ir}. The optimization problem shown in (5.11) - (5.13) is associated with the social welfare of all the SPs. Usually, it consists of maximizing the aggregated utility of their concerning utilities. However, we consider the α - fair function that provides the system with needed flexibility in employing a fair and efficient resource allocation. Let $\boldsymbol{x} = [\boldsymbol{x}_i]_{i\in\mathcal{N}}$, then the objective of the RO is to find $\boldsymbol{x}$ that maximizes the following optimization problem.

$$SYSTEM(\boldsymbol{\lambda}, \boldsymbol{x}) : \underset{x}{\text{maximize}}\ \mathcal{G} : \quad \sum_{i=1}^{N} U(u_i) \tag{5.11}$$

$$\text{subject to} \quad \sum_{i=1}^{N} x_{ir} \leq c_r \quad \forall\, r \in \mathcal{R} \tag{5.12}$$

$$u_i \leq \frac{x_{ir}}{d_{ir}} \quad \forall r \in \mathcal{R}, \forall i \in \mathcal{N} \tag{5.13}$$

The first constraint (5.12) ensures that the allocated resources do not exceed the available capacity, and the constraint (5.13) ensures that the resulting utility from the multiple resources obtained by the SP follows the Leontief utility function. Here, $U(u_i)$ is assumed to follow the α - fair allocation rule [74, 98] as it has the flexibility in achieving a trade-off between efficiency and fairness by adjusting the factor α. It is defined as follows,

$$U(u_i) = \begin{cases} \frac{(u_i)^{1-\alpha}}{(1-\alpha)} & \text{if } \alpha \neq 1 \\ \log(u_i) & \text{if } \alpha = 1 \end{cases} \tag{5.14}$$

α defines the level of fairness achieved while allocating the resources to the slices. Every InP chooses a specific α for computing the optimal allocation for each resource type that is associated with InP m. The following are some of the popular fairness rules that are generalized under α-fair allocation.

- $\alpha \to 1$: this leads to the fairness function in [77], which is an intermediate choice between the aforementioned extreme cases. Assuming that $\boldsymbol{x}^*$ is a feasible solution (5.1)-(5.2), an allocation $\boldsymbol{x}^*$ is said to be a proportionally fair allocation, if the aggregate of proportional change with respect to any other feasible allocation $\boldsymbol{x}$ is negative, i.e.,

$$\sum_{i \in \mathcal{N}} \frac{x_i^* - x_i}{x_i^*} \leq 0. \tag{5.15}$$

  This is also referred to as Nash Welfare function [56, 78], where the welfare function is defined as the product of utilities.

$$NSW(\boldsymbol{x}) = \prod_{i \in \mathcal{N}} u_i(x_i) \tag{5.16}$$

- $\alpha \to \infty$: this is a popular fairness rule often referred to as Max-min or bottleneck optimality criterion [14]. A feasible allocation $\boldsymbol{x}$ is said to be Max-min fair, if an increment in x_i leads to a decrement in x_j, where $x_j \leq x_i$. In this approach, the InP is concerned with the equality among the individual utilities of the SPs [91].

In general there are three types of approaches for multi-resource allocation, they are: centralized approach, distributed approach, and decentralized approach as shown in figure 5.1-5.3.

Centralized approach: As described earlier, there have been many works that employ a central entity to compute the optimal resource quantities for each of the competing client using gradient projection method [15]. As shown in the figure 5.1, all the SPs competing for resources with a dedicated slice S_i communicates the resource orchestrator with necessary information like demand vector d and the corresponding utility function u_i, where demand d is a vector of individual resource demand d_{ir}. Every InP communicates the resource capacities c_r associated with the resources they own. Now, the resource orchestrator has the burden of finding an optimal solution that serves the best interest of InPs and SPs as well. However, this approach has several drawbacks [37, 80]:

- To realize the resource allocation through central allocation rules, the orchestrator needs information such as available resource capacity from InPs and utility functions of the SPs. This is private information that they can be reluctant to share.

- With dynamic changes in the network due to varying growth in the number of slices, scalability is a big issue for the orchestrator.

- InPs can not choose their individual allocation rule in this approach.
- Finally, as with any centralized solution, a single point of failure is an inevitable drawback.

Distributed approach: To avoid the issuess encountered in the central approach, solutions based on distributed approach were proposed in [30]. In this approach there is no need for the central resource orchestrator as shown in figure 5.2. Resource allocation is achieved with a minimal information exchange between InPs and SPs. As reported in [30], InPs can independently implement a specific allocation rule. However, there is a need for some coordination among the InPs to allocate resources to the SPs. In addition to that, there is no control over the fairness among the SPs.

Decentralized approach: As opposed to both approaches described before, decentralized mechanisms do not require any information sharing among InPs or there is no need for a central resource orchestrator. On the other hand, there is no necessity for the SPs reveal sensitive information. So, as far as the privacy is concerned, this is by far the suitable approach for multi-resource allocation. Inspired by the works of Kelly in [60], and other works [53, 37], we design a mechanism by defining two sub optimization problems, one for InPs and another for the slices. The mechanism employs a game between InPs and slices to iteratively improve the allocation vector by exchanging non sensitive data and eventually converge to the optimal allocation vector. The optimization problem associated with the InPs has been introduced in section 5.3.2. The optimization problem associated with the slices is defined in the following section along with the associated game.

5.5 Decentralized Resource Allocation Mechanism (DRAM)

Resource allocation under network slicing in 5G wireless networks is a challenging issue due to the involvement of numerous resources and distinct SP requirements. In addition, privacy is an important aspect of the resource allocation process. In general, neither the InPs nor the SPs are willing to share the sensitive information as described earlier. In this section, we propose a decentralized resource allocation mechanism that preserves the sensitive information of the stakeholders and allows for each InP to different allocation rule. In fact, this holds efficient as long as the objective function is concave or strictly concave.

We consider that very SP is aware of the value for β_r for all InPs. Now, we define a local optimization problem for each SP that allows them to obtain an optimal allocation vector x that maximizes the social welfare function $\mathcal{G}$.

$$SP(\boldsymbol{\phi}, \boldsymbol{b}_i) \quad \underset{b}{\text{maximize}}\ Q_i : \qquad \frac{1}{1-\alpha} u_i^{1-\alpha} - \sum_{r \in \mathcal{R}} b_r^{\frac{1}{\beta_r}}, \tag{5.17}$$

$$\text{subject to :} \qquad u_i \leq \left(\frac{b_{ir}}{\phi_r}\right)^{\beta_r} \frac{1}{d_{ir}} \quad \forall r \in \mathcal{R}. \tag{5.18}$$

In this Mechanism, each SP sends a bids $[b_{ir}]_{i\in\mathcal{N}}$ to the InP that owns the corresponding resource r. In return, InPs provide each SP with ϕ by using the following relation that is obtained by solving the optimization problem INFRA($\boldsymbol{b}_r, \phi_r, \boldsymbol{x}_r$) defined in (5.6)-(5.7).

From eq.5.9, we have the closed form expression for price for resource r

$$\phi_r = \left(\sum_{i\in\mathcal{N}} b_{ir}^{\frac{1}{\beta_r}}\right)^{\beta_r}. \tag{5.19}$$

Each SP use the price vector $\boldsymbol{\phi}$ to compute optimal bids that maximize their pay-off function defined in (5.17)-(5.18) with a motivation to obtain an optimal allocation that is a social optimal. The user optimization problem is defined in such a way that the resulting allocation vector for this mechanism is a social optimal.

Theorem 5.1. *Let $\boldsymbol{x}_i^*$ be an optimal allocation for the local optimization problem defined in (5.17)-(5.18) for each SP $i \in \mathcal{N}$ in conjunction with optimization problem defined in (5.6)-(5.7) for each InP $m \in \mathcal{M}$, then $\boldsymbol{x}^*$ solves the social welfare function defined in (5.11)-(5.13).*

Proof. We can prove that above theorem by writing the KKT conditions [15], and finding the similarity between them. First, we write the Lagrngian for SYSTEM($\boldsymbol{\lambda}, \boldsymbol{x}$), where $\boldsymbol{\lambda} = [\lambda_r]_{r\in\mathcal{R}}$, $\boldsymbol{\gamma} = [\gamma_{ir}]_{i\in\mathcal{N},r\in\mathcal{R}}$ are the Lagrange multipliers associated to the capacity constraint (5.12) and Leontief utility constraint (5.13) respectively.

$$L_{sys}(\boldsymbol{u}, \boldsymbol{x}, \boldsymbol{\gamma}, \boldsymbol{\lambda}) = \sum_{i\in\mathcal{N}} \frac{1}{1-\alpha} u_i^{1-\alpha} - \sum_{i\in\mathcal{N}}\sum_{r\in\mathcal{R}} \gamma_{ir}(u_i - \frac{x_{ir}}{d_{ir}}) - \sum_{r\in\mathcal{R}} \lambda_r \left(\sum_{i\in\mathcal{N}} x_{ir} - c_r\right) \tag{5.20}$$

By writing the first order necessary and sufficient conditions, we have

$$\frac{\partial L_{sys}}{\partial u_i} = 0 \implies u_i^{-\alpha} - \sum_{r\in\mathcal{R}} \gamma_{ir} = 0, \tag{5.21}$$

$$\frac{\partial L_{sys}}{\partial x_{ir}} = 0 \implies \frac{\gamma_{ir}}{d_{ir}} - \lambda_r = 0. \tag{5.22}$$

From the above equation, we can write

$$\gamma_{ir} = \lambda_r d_{ir}. \tag{5.23}$$

Now, we write the Lagrangian for INFRA($\boldsymbol{b}_r, \phi_r, \boldsymbol{x}_r$), where ϕ_r is a Lagrange multipler associated with the capacity constraint (5.7).

$$L_{infra}(\boldsymbol{x}_r, \phi_r) = \sum_{i\in\mathcal{N}} \frac{1}{1-\beta_r} b_{ir} x_{ir}^{1-\beta_r} - \phi_r \left(\sum_{i\in\mathcal{N}} x_{ir} - c_r\right). \tag{5.24}$$

by writing the first order necessary and sufficient conditions, we have

$$\frac{\partial L_{infra}}{\partial x_{ir}} = b_{ir} x_{ir}^{-\beta_r} - \phi_r = 0.$$

From the above equation, we can write the closed form solution for allocated resource quantity x_{ir} by an InP

$$x_{ir} = \left(\frac{b_{ir}}{\phi_r}\right)^{\frac{1}{\beta_r}}. \tag{5.25}$$

By applying sum on both sides, we have the following,

$$1 = \sum_{i \in \mathcal{N}} \left(\frac{b_{ir}}{\phi_r}\right)^{\frac{1}{\beta_r}}.$$

With that we have,

$$\phi_r = \left(\sum_{i \in \mathcal{N}} b_{ir}^{\frac{1}{\beta_r}}\right)^{\beta_r}. \tag{5.26}$$

Finally, we write the Lagrangian for SP($\boldsymbol{\phi}, \boldsymbol{b}_i$), where $\boldsymbol{\gamma}_i$ is the Lagrange multipliers associated with the constraint (5.18)

$$L_{sp}(u_i, \boldsymbol{x}_i, \boldsymbol{\phi}, \boldsymbol{\gamma}_i) = \frac{1}{1-\alpha} u_i^{1-\alpha} - \sum_{r \in \mathcal{R}} (b_{ir})^{\frac{1}{\beta_r}} - \sum_{r \in \mathcal{R}} \gamma_{ir} (u_i - \left(\frac{b_{ir}}{\phi_r}\right)^{\beta_r} \frac{1}{d_{ir}}). \tag{5.27}$$

$$\frac{\partial L_{sp}}{\partial u_i} = 0 \implies u_i^{-\alpha} - \sum_{r \in \mathcal{R}} \gamma_{ir} = 0, \tag{5.28}$$

$$\frac{\partial L_{sp}}{\partial b_{ir}} = 0 \implies -\phi_r^{\frac{1}{\beta_r}} + \gamma_{ir} dir = 0.$$

From the above equation, we can write

$$\gamma_{ir} = \phi_r^{\frac{1}{\beta_r}} d_{ir}. \tag{5.29}$$

From eq. (5.28) and eq. (5.29), we can write

$$u_i^{-\alpha} = \sum_{r \in \mathcal{R}} \phi_r^{\frac{1}{\beta_r}} d_{ir}.$$

Given that we have $u_i = \frac{x_{ir}}{d_{ir}}$, we can write

$$x_{ir} = d_{ir} \left(\sum_{r' \in \mathcal{N}} d_{ir'} \phi_{r'}^{\frac{1}{\beta_{r'}}} \right)^{\frac{-1}{\alpha}}. \tag{5.30}$$

Each service provider uses eq.5.32 to compute the resource quantity given the price vector $\boldsymbol{\phi}$ sent by the infrastructure providers.

By comparing the results in eq. (5.21), (5.23) with the results in eq. (5.28) - (5.29), we can conclude that with $\lambda_r = \phi_r^{\frac{1}{\beta_r}}, r \in \mathcal{R}, (\boldsymbol{x}_i, i \in \mathcal{N})$ maximizes the optimization problem defined in eq. (5.11) - (5.13). □

For the reasons mentioned earlier, implementing this mechanism using a centralized RO is difficult. In the following, we provide an algorithm that can be implemented by each InP to determine the optimal pricing vector ϕ with which the SPs can obtain the optimal allocation vector x that is a social optimal.

5.5.1 Online distributed algorithm for multi-resource allocation

We propose an algorithm based on the dual variable ϕ_r associated to the capacity constraint (5.7) to obtain the optimal allocation vector $\boldsymbol{x}$ for the system. This is based on simple gradient projection method with a constant step size δ. By choosing an appropriate choice of step size δ, we prove that the $\boldsymbol{\phi}$ converges to $\boldsymbol{\phi}'$. The resulting allocation vector $\boldsymbol{x}$ is an optimal solution for system.

Now, we define a pricing update mechanism for each InP as follows

$$\phi_r(t+1) = max(0, \phi_r(t) + \delta(\sum_{i \in \mathcal{N}} x_{ir} - c_r)). \tag{5.31}$$

- Each InP computes the price ϕ_r as per pricing update defined in (5.31) and communicates the updated price to the corresponding SPs.
- Then each SP locally computes the resource allocation as per eq.(5.32), which is

$$x_{ir}(\boldsymbol{\phi}) = d_{ir} \left(\sum_{r' \in \mathcal{N}} d_{ir'} \phi_{r'}^{\frac{1}{\beta_{r'}}} \right)^{\frac{-1}{\alpha}}. \tag{5.32}$$

- Then SP communicates the resulting adjusted resource quantity to InP. This iterative improvement continuous till the price vector $\boldsymbol{\phi}$ convergence in $\hat{\boldsymbol{\phi}}$.
- Given that the allocation is computed as a function of $\boldsymbol{\phi}$, when the prices converge to an optimal vector $\hat{\boldsymbol{\phi}}$, according to theorem 5.1, the resulting allocation is an optimal solution for the social welfare function defined in (5.11)-(5.13).

API Name	vCPU	Memory (Gb)	Storage (Gb)	Bandwidth (Gbps)	Instance Type
i3en.large	2	16	1250	25	Storage optimized
i3en.xlarge	4	32	2500	25	Storage optimized
r5d.large	2	16	75	10	Memory optimized
r5d.xlarge	4	32	150	10	Memory optimized
c6gd.large	2	4	118	10	Compute optimized
c6gd.xlarge	4	8	237	10	Compute optimized

Table 5.2: API instances from AMAZON EC2.

- We set bounds on the value of $\phi_r \in [\phi_{min}, \phi_{max}]$ such that $\phi_r = \phi_{max}$ when the aggregated resource allocation $\sum_{i\in\mathcal{N}} x_{ir}$ exceeds the available capacity c_r, this prevents the value of ϕ_r reaching ∞, on the contrary, $\phi_r = \phi_{min}$ when $\sum_{i\in\mathcal{N}} x_{ir}$ is significantly lower in comparison to c_r, this prevents the value of ϕ_r from reaching zero.

Theorem 5.2. *Let* $\{\boldsymbol{\phi}\}$ *be sequence generated by the eq.* (5.31) *such that* $\boldsymbol{\phi}(0) \in \mathbb{R}^{|\mathcal{R}|}$ *and* $\delta \in (0, \frac{2}{\sigma})$ *where*

$$\sigma = R \sum_{i\in\mathcal{N}} \frac{1}{\alpha} \left(\frac{1}{\phi_{min}}\right)^{-(\frac{1}{\alpha}+1)} \max_{r'} K_{r'}$$

the sequence $\boldsymbol{\phi}(t)$ *converges to* $\{\hat{\boldsymbol{\phi}}\}$*, i.e.,* $\lim_{t\to\infty} \boldsymbol{\phi}(t) = \hat{\boldsymbol{\phi}}$.

Proof. See appendix C.1. □

In the following paragraphs, we explain the implementation of our mechanism in the O-ran system envisioned for the beyond 5G networks. However, our algorithm is not limited to just RAN, but can be applied for end-to-end resource allocation for a given slice by considering each data center as an InP proving the resources. This algorithm is suitable for near-real-time scheduling.

Here, we provide a simple example for a high-level description regarding the implementation of the proposed mechanism in Open-RAN to illustrate the operational aspects motivated by [37]. We consider that there are four vendors who provide COTS hardware, two at the distributed unit and two at the centralized unit[†] as shown in figure 5.5. Each vendor provides 3 types of resources such as CPU, $memory(mem)$, and $bandwidth(BW)$, each resource is associated with a controller responsible for computing the resource quantities x_{ir} based on the bids b_{ir} submitted by the slices. Consider that a slice manager pre-determined the placement of vnfs. Now each slice submits a `beginmechrequest` signal to the resource controllers which conveys that the slice wants to begin the resource allocation mechanism, then the resource controller acknowledges with a `beginmechrequestack`. After that slice sends a `initialdemand` signal (this can be a guess by the slice based on the resource demand submitted by the SP), then the controller returns the computed

[†]In this work, we consider that a vendor is simply an infrastructure provider who owns the resources.

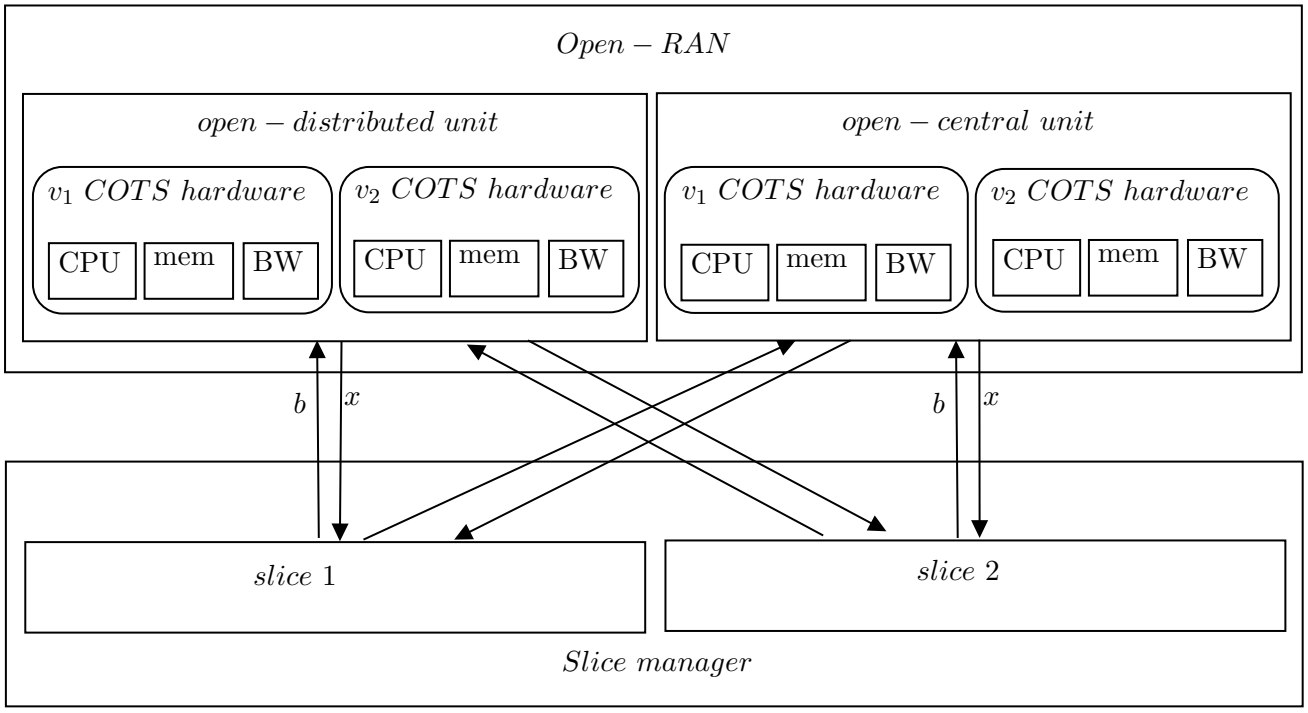


Figure 5.5: Game implementation with multi-vendor COTS hardware in open-RAN.

price per resource quantity `resourceprice`. Now, the slice submits the adjusted demand based on the information received from the controller and replies to the controller as `updateddemand`. This process continues until there is convergence, then slice returns `convergencereached` and controller replies with `stopmech` to end the mechanism. Finally, the slice returns `stopmechack` which concludes the mechanism with each of the slices obtaining the optimal resources. This mechanism is concurrently running between all the controllers and slices.

5.6 Numerical solutions

In this section, we provide numerical results to validate the theoretical findings described in the previous section. We consider AMAZON ec2 instances to generate the demand vector d_i for service providers. We use CVX package [35] for producing the numerical results, it is a matlab software designed for solving convex optimization problems.

The system setting goes as follows: There are 3 SPs running various services each requiring a resource combination that belongs to a particular ec2 instance, and there are 3 InPs supporting 3 resource types (each provides one resource type) such as vCPUs, memory, and bandwidth respectively. Each SP has a non-negative demand for all resource types. Let the capacity of each resource

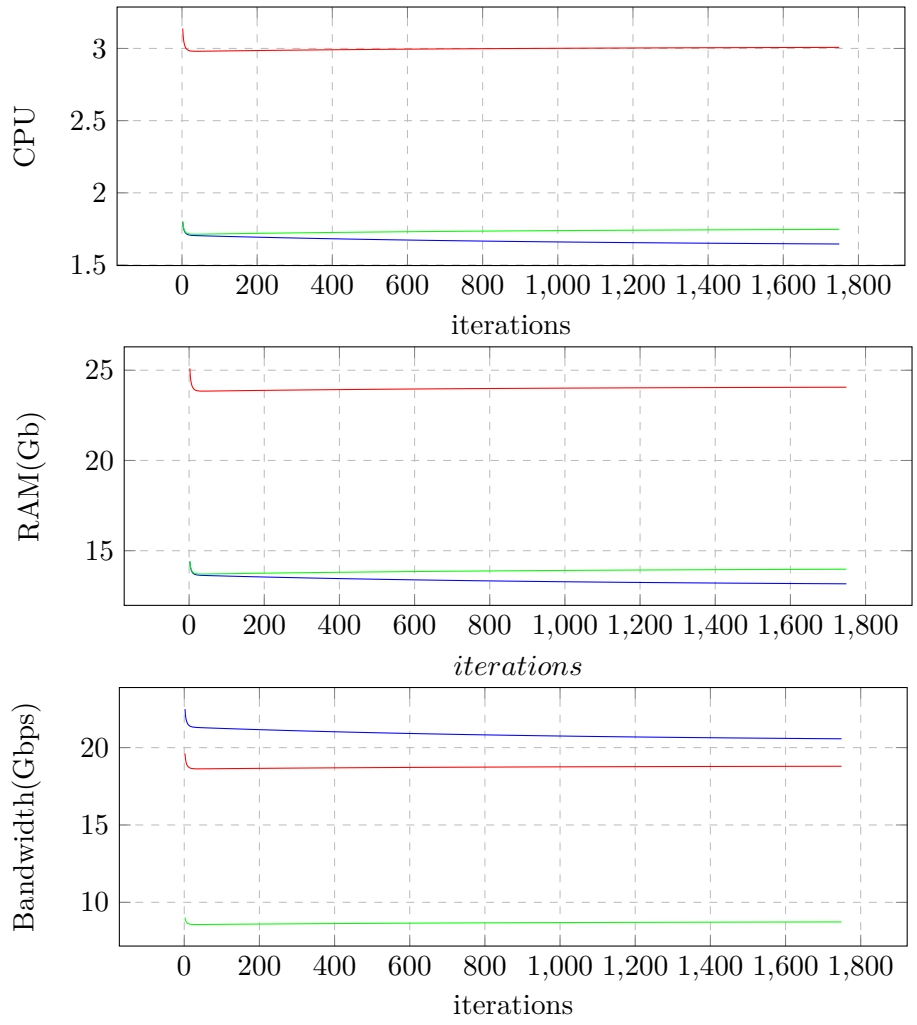


Figure 5.6: Convergence of resource allocation for CPU, Memory, Bandwidth for three users.

type be C_{vCPU} = 14, C_{memory} = 86.4 Gb, and $C_{bandwidth}$ = 72 Gbps. Here, each SP tries to procure as many resources as possible. First, we show the convergence of the algorithms described in the previous section with α = 5 for social welfare function, β_1 = 1, β_2 = 5, β_3 = 1 for 3 InPs respectively. We let Δ to be a difference between the prices for successive iterations, it gives a trade-off between the solution quality and convergence time and we found that $\Delta = 10^{-4}$ results in a good balance. In figure 5.6, we show the convergence of all three resources as it can be observed, they converge in few hundreds of iterations.

In figure 5.7, we compare our decentralized mechanism DRAM labeled with `DRAM (initial)` and `DRAM (converged)`, with the centralized approach labeled as `central`. `DRAM (initial)` shows the allocation of each InP during the initial phase of the algorithm, InP 1 and InP 3 are assigned with proportional allocation fairness, whereas InP 2 is assigned with $\beta_2 = 5$ which imposes higher

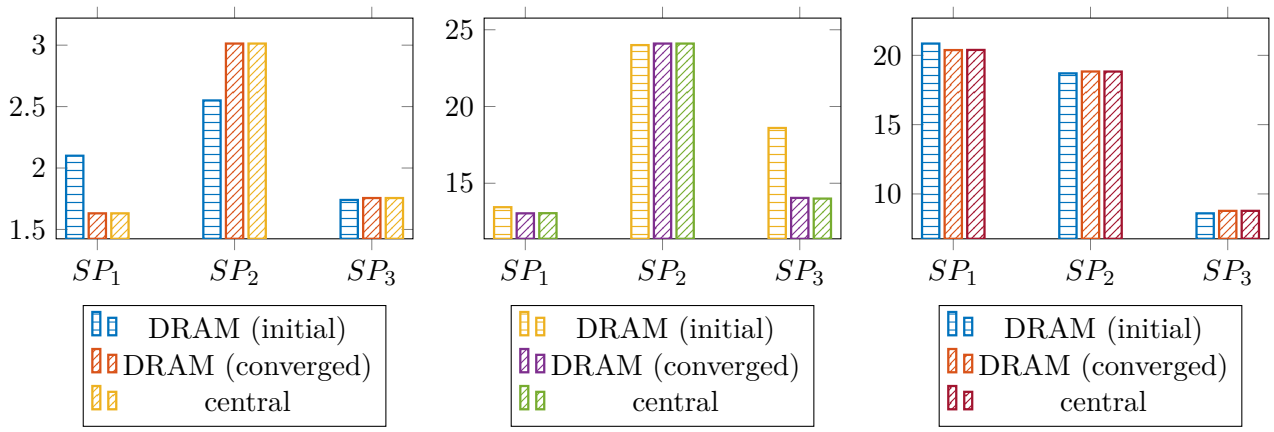


Figure 5.7: Comparison of resource allocation for CPU, Memory, Bandwidth with optimal allocation.

fairness compared to the other two InPs. At this point, we do observe any influence of the fairness α employed at the system level. But this changes in a few iterations and we notice the influence of system level α on the resource allocation. Bars labeled with `DRAM (converged)` in figure 5.7, are associated with the resource allocation after the convergence of the algorithm, we can notice that they match exactly the allocation obtained using the central solution. Now, we can no longer see the impact of the individual allocation rule by InPs as it is overridden by the system-level allocation rule. We compare this pattern for all the resources shown in figure 5.7. We describe the impact of α value on the fairness among the slices in the next paragraphs.

5.6.1 Impact of α on SP utility and allocation

The system setting for the remainder of the section goes as follows: There are 6 SPs running various services each requiring a resource combination that belongs to a particular ec2 instance, and there are 4 InPs supporting 4 resource types such as vCPUs, memory, storage, and bandwidth respectively. Each SP has a non-negative demand for all resource types. The demand vectors for SP1 - SP6 are listed in the table 5.2 with API Name `i2en.large` to `c6gd.xlarge` respectively. Let the capacity of each resource type be C_{vCPU} = 14, C_{memory} = 86.4 Gb, $C_{storage}$ = 3464 Gb, and $C_{bandwidth}$ = 72 Gbps. Here, each SP tries to procure as many resources as possible. First, we compare the utilities of the SPs to describe the impact of α, then we compare the allocations of each SP for different values of α.

The purpose of tuning factor α is to achieve the desired fairness criterion depending on the system requirements. Fig.5.8 compares the utilities of 6 SPs for the values of α ranging between 1 and 10 with an increment of 0.5. Observe that the utility for SP_2 and SP_5 are considerably higher than the rest of the SPs for α = 1. Among all the SPs, SP_2 has the lowest utility. As the value of α is increasing, the difference between the utilities gets smaller and smaller. As α reaches the value of

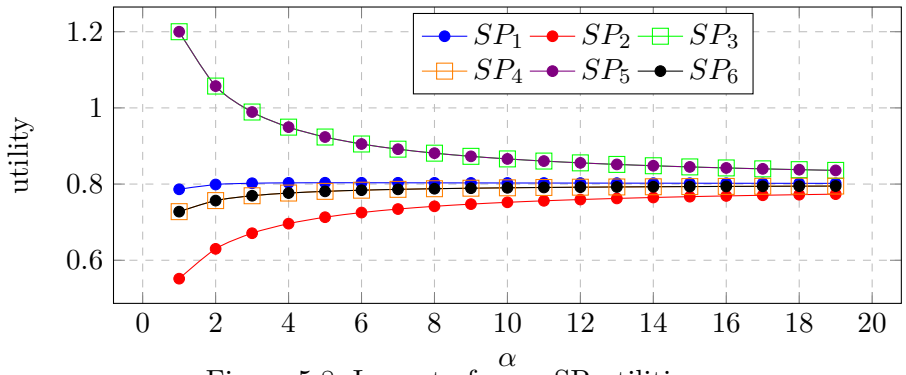


Figure 5.8: Impact of α on SP utilities.

10, the difference between SPs is nominal validating our theoretical claim that a higher value for α leads to better fairness among the SPs.

Fig. 5.9 shows the comparison of resource allocation under three different α values, where each plot displays the allocated resource in comparison to the demanded resource. For α = 1, every SP gets a lower allocation than their base demand except SP_3 and SP_5 who gets higher resource than their base demand. The reason for this behavior is that the base demand of SP_3 and SP_5 for bandwidth as shown in fig. 5.12 is significantly lower than SP_1 and SP_2 and relatively lower than SP_4 and SP_6, i.e., in order for the allocation to be effective for SP_1 and SP_2 all allocations of other resource types should be in proportion to that of bandwidth. As the bandwidth is limited and can not be assigned as desired by the SP_1 and SP_2, allocation of other resources should be reduced as they can not be effective for them while the bandwidth acts as the bottleneck. This additional resource is allocated to SP_3 and SP_5 who have the least amount of demand for bandwidth. This situation is unfair for other SPs as their utilities fall well below the utilities of SP_3 and SP_5 as shown in fig. 5.8. Hence, a fairness mechanism should ensure that the desired fairness is achieved among the competing SPs. As shown in fig. 5.9 - 5.12, the increment in value of α ensures that no SP gets more resource in comparison to other SPs. For $\alpha = 10$ in spite of lower bandwidth demand by SP_3 and SP_5, they do not get higher allocation. This prevents SPs from strategically reporting lower demands and anticipating higher returns.

5.7 Conclusion

For an orchestrator to manage resource allocation using the centralized approach, it is necessary for the InPs reveal the available resource quantity and SPs reveal the utility functions. However, this is private information that both entities do not want to reveal. To account for privacy, In this work, we proposed a decentralized mechanism that addresses multi-resource allocation problems with

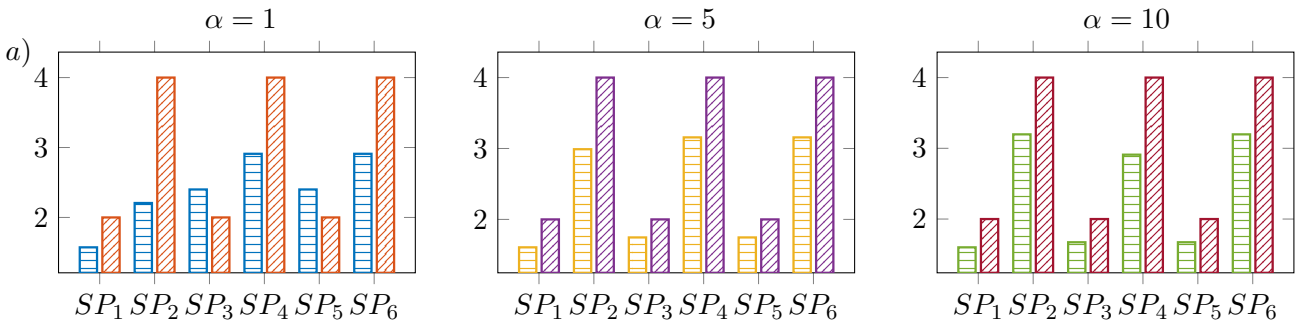


Figure 5.9: Impact of alpha on allocation of CPU.

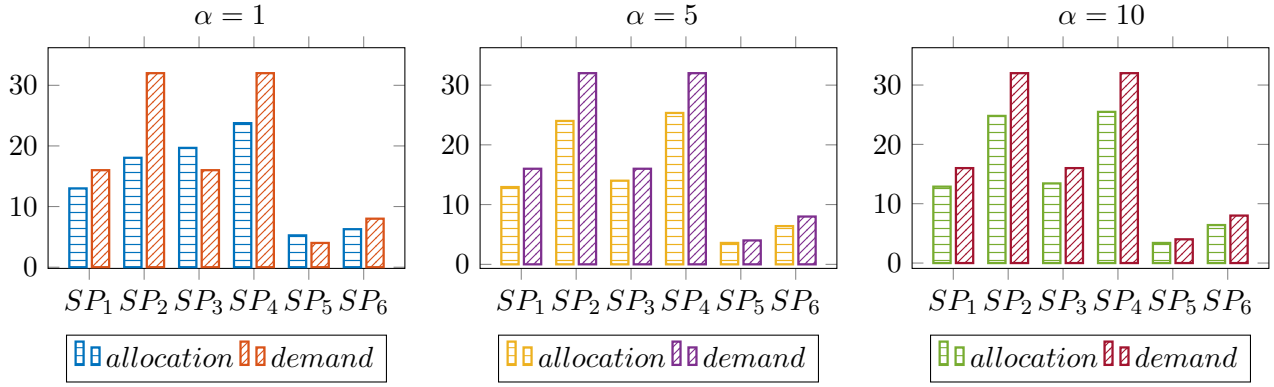


Figure 5.10: Impact of alpha on allocation of Memory (Gb).

heterogeneous requirements. Majority of the models proposed in the literature address this issue either at the SPs level or InP level. Our approach provides a unified mechanism that accounts for multi-level privacy. This mechanism allows each InP to implement its preferred allocation rule. One of the interesting aspects of this problem is to study the impact of the intrinsic strategic nature of the SPs on social welfare. We elaborate on this in the next chapter, where we provide the overall conclusions and future directions.

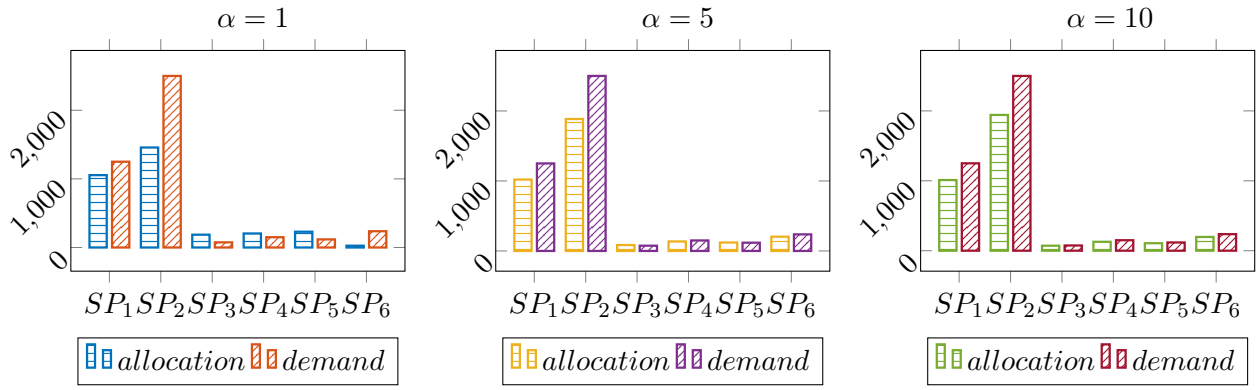


Figure 5.11: Impact of alpha on allocation of Storage (Gb).

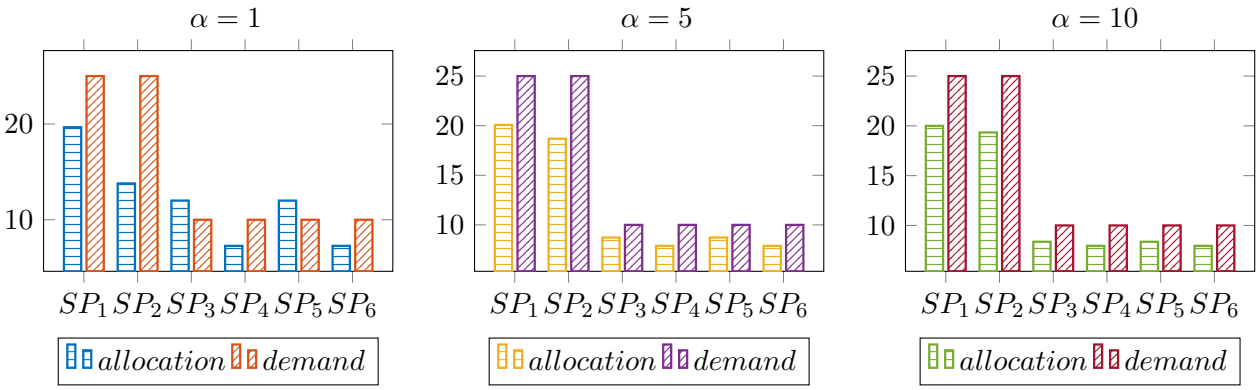


Figure 5.12: Impact of alpha on allocation of Bandwidth (Gbps).

Chapter 6

Conclusion and Future Directions

6.1 Conclusions

Future IoT service providers will need ubiquitous IoT data collection, mandating in turn the support of IoT access at scale over the 5G infrastructure. At the same time, new schemes to control data generation and upload should allow to SPs to perform IoT data brokerage across diverse access resources made available by concurrent infrastructure providers at different costs, in the form of 5G IoT resource slices.

IN chapter 2, we introduced a new framework to connect two fundamental aspects: the AoI of IoT data to be uploaded and the cost of 5G resources leased in order to obtain network access services. The upload control can be performed in a distributed way at the device level using optimal dynamic multi-threshold policies. Such policies have been showed to outperform their static counterparts. At same time, a SP can control shadow prices to match optimal multi-threshold policies to service requirements while minimizing operational costs. It does so at the slice level by incentivizing users to perform IoT data uploads where resources leased from InPs are cheaper. In chapter 3, we proposed low complexity algorithms to determine the best shadow prices to migrate the traffic from highly congested areas with lower congestion. This reduces the operational cost for the SPs and also reduce the congestion issue for the InPs. We proposed a further enhancement to these algorithms to reduce the convergence time in obtaining the best pricing for the locations that account for aging control and balance the traffic load. This work opens new directions at the bridge between IoT and 5G research, by describing on a quantitative basis how to trade-off IoT data freshness and load balancing, as supported by the 5G slicing paradigm.

In chapter 4, we addressed resource allocation that has always been a very challenging aspect in telecommunication due to the scarcity of resources and high demand from the users, even more so in the context of 5G, which is by far the most diverse and complex system. In this chapter, we investigated a method to cope up with those challenges by emphasizing the fact that the resources should be fairly allocated to the SPs competing for the resources. We formulated a model that based on Fisher market that allocate resources based on the budget available for each SP. Our objective is to enforce fairness such that SPs with lower budget are not starved due to the lack of resources. We demonstrated that our model is capable of allocating multiple resource types that are fair among slices and among the locations of a slice with help of the factors α and β respectively, which act as

control parameters to balance the trade-off between fairness and efficiency.

In practice, cost for obtaining resources is not linear, it generally depends on the amount of resource procured. Higher amount of resource leads to lower price per unit of resource. In this work, we address this aspect designing a non-linear pricing for allocating the resources. This also allows the InP to employ different welfare functions other than Nash welfare function. In general, SPs are self centric and may not report the true preferences to the InP. This strategic nature has an impact on the social welfare. We study the loss of efficiency due to strategic nature of the service providers and we provide price of anarchy, which is a measure to indicate the deterioration in social welfare due to selfish behaviour of the SP compared to the social optimal. Thus, this unified framework focuses on a non-linear pricing scheme to allocate resources with multi-level fairness that can be applied to various scenarios under the 5G network slicing phenomenon.

For an orchestrator to manage resource allocation using centralized approach, it necessary that the InPs reveal the available resource quantity and SPs reveal the utility functions. However, this is a private information that both entities do not want to reveal. To account for the privacy, In chapter 5, we proposed a distributed mechanism that address multi-resource allocation problem with heterogeneous requirements. Majority of the models proposed in the literature address this issue either at SPs level or InP level. Our approach provides a unified mechanism that accounts for the multi-level privacy.

6.2 Future directions

- In chapter 2 and 3, we defined threshold based policies based on that leverages device mobility and optimal price design to partially move the traffic away from congested areas. However, it is interesting study if it is possible to divert users to a specific path by providing additional incentives. We would like to explore this idea in future works.

  Recently in many IoT applications, it has been observed that while uploading the data, considering the just factor of the information age is insufficient. The quality of data which is being uploaded also plays an important role. For instance, if a platform is gathering data from a group of sensors, the precision of data or variance of noise that has been added to data is also crucial. In such cases, the end users might be incentivised depending on the quality of information and its age. Thus in future, we would like to design a policy for end users considering the age and quality or value of information.

- In chapter 5, we proposed distributed resource allocation mechanism that preserve the privacy concerns of various entities involved. However, in this case it is interesting study the impact of strategic SPs, where the SP report false preferences to gain additional utility benefit. This can worsen the social welfare, to this aim, we would like to obtain a closed form solution for price

of anarchy to effectively understand the effect such selfish behaviour.

- In chapters 4 and 5, while designing the resource allocation mechanism, we have considered that the user load with service providers is stationary; however, in practice, the users' load might vary with time. Thus it would be interesting to explore how our proposed method can be extended to the scenario where load varies with time. Moreover, end users could have a minimum requirement over their service rates in many cases. Due to limitations over the resource inventory, providing the minimum service requirements of all the users is not always feasible. To handle such situations, admission control over the arrivals of the end-users has to forced. Thus extending current formation with admission control over the users' arrivals would be potential direction for research.

6.3 List of publications

Following are the list of publications that were successfully published at conferences and a journal with the results obtained in the thesis work.

1. Naresh Modina, Rachid El Azouzi, Francesco De Pellegrini, Daniel Sadoc Menasche, "Joint traffic offloading and aging control in 5g IoT networks", 32nd International teletraffic congress (ITC32), 2020.
2. Naresh Modina, Rachid El Azouzi, Francesco De Pellegrini, Daniel Sadoc Menasche, Rosa Figueiredo "Joint traffic offloading and aging control in 5g IoT networks", IEEE Transactions on Mobile Computing, 2022.
3. Naresh Modina, Mandar Datar,Rachid El-Azouzi,Francesco de Pellegrini "Multi Resource Allocation for Network Slices with Multi-Level fairness",IEEE International Conference on Communications (ICC), 2022.

Appendices

Appendix A

A.1 Proof of Theorem 2.2

Next, we present the proof of Thm.2.2. We begin by noting that from [85] there exist a value function $V(x,l)$ and a scalar ρ satisfying the Bellman equation for the average cost MDP problem

$$V(x,l)+\rho=\max\Big(U(x)-p(l)+\sum_{l'\in\mathcal{L}}\lambda_{ll'}V(1,l'),\\ U(x)+\sum_{l'\in\mathcal{L}}\lambda_{ll'}V(\min(x+1,M),l')\Big) \tag{A.1}$$

An optimal policy μ able to select the per-state action maximising the right hand side of (A.1) is an optimal solution (2.7). Moreover, it is known that an unconstrained MDP admit a deterministic optimal policy [85]. Since a multi-threshold strategy belongs to this class of policies, we restrict our discussion to the case of deterministic policies.

In what follows, we consider locations sorted by increasing price order, that is $p(l)\leq p(l+1)$, for $l=1,\dots,L$. Let define the function $H:\mathcal{M}\times\mathcal{L}\times\{0,1\}\to\mathbb{R}$ as follows

$$H(x,l,1)=U(x)-p(l)+\sum_{l'\in\mathcal{L}}\lambda_{ll'}V(1,l') \tag{A.2}$$

$$H(x,l,0)=U(x)+\sum_{l'\in\mathcal{L}}\lambda_{ll'}V(x+1,l') \tag{A.3}$$

$$\Delta H(x,l)=H(x,l,1)-H(x,l,0) \tag{A.4}$$

Hereafter we shall demonstrate that $i)$ the value function is decreasing in the age of information for any given location, $ii)$ that the optimal policy for any given location switches from 0 to 1 at most once and finally $iii)$ that if uploading is optimal for a certain value of the age of information at a given price, it is also optimal for larger prices as well. Such facts are proved formally in the following lemma.

Lemma A.1. *For any optimal policy, for $x=2,\dots,M$ the following facts hold:*
i. $V(x-1,l)\geq V(x,l),\ \forall l\in\mathcal{L}$.
ii. $\Delta H(x-1,l)\geq 0\Rightarrow\Delta H(x,l)\geq 0,\ \forall l\in\mathcal{L}$.

iii. $\Delta H(x,l) \geq 0 \Rightarrow \Delta H(x,l-1) \geq 0,\ \forall l \in \mathcal{L}$.
iv. $V(x,l-1) \geq V(x,l),\ \forall l \in \mathcal{L}$.

Proof. We show each of the four items above in the corresponding order.

i. We can verify the result directly by backward induction on (A.1). For a deterministic policy, we define

$$Z_l(x) := V(x,l) - U(x) + \rho = \tag{A.5}$$

$$= \max\Big(-p(l) + \sum_{l' \in \mathcal{L}} \lambda_{ll'} V(1,l'),$$

$$\sum_{l' \in \mathcal{L}} \lambda_{ll'} V(\min(x+1, M), l')\Big) \tag{A.6}$$

We shall prove that $Z_l(x) \geq Z_l(x+1)$. This implies $V(x,l) \geq V(x+1,l)$ since $V(x,l) - U(x) \geq V(x+1,l) - U(x+1) \geq V(x+1,l) - U(x)$, where the last step holds because U is non increasing. First, we observe that

$$Z(M-1) = \tag{A.7}$$

$$\max\Big(-p(l) + \sum_{l' \in \mathcal{L}} \lambda_{ll'} V(1,l'), \sum_{l' \in \mathcal{L}} \lambda_{ll'} V(M,l')\Big) = Z_l(M)$$

so that the inductive basis holds true.

Now, in the general case we can observe that if the statement is true for $x+1$, that is $Z_l(x+1,l) \geq Z_l(x+2,l)$, it needs to hold for x as well. Using the induction hypothesis, we have $V(x+1,l) \geq V(x+2,l)$ and thus

$$Z_l(x) = \max(-p(l) + \sum_{l' \in \mathcal{L}} \lambda_{ll'} V(1,l'), \sum_{l' \in \mathcal{L}} \lambda_{ll'} V(x+1,l'))$$

$$\geq \max(-p(l) + \sum_{l' \in \mathcal{L}} \lambda_{ll'} V(1,l'), \sum_{l' \in \mathcal{L}} \lambda_{ll'} V(x+2,l')) \tag{A.8}$$

$$= Z_l(x+1) \tag{A.9}$$

which concludes the inductive step.

ii. It is sufficient to write $\Delta H(x-1,l) - \Delta H(x,l) = \sum_{l' \in \mathcal{L}} \lambda_{ll'}[V(x+1,l') - V(x,l')] \leq 0$.

iii. In this case, we can directly verify

$$\Delta H(x,l-1)-\Delta H(x,l)=-p(l-1)+p(l)+\sum_{l'\in\mathcal{L}}(\lambda_{(l-1)l'}-\lambda_{ll'})[V(1,l')-V(x+1,l')]$$
$$\geq\tilde{\kappa}\sum_{l'\in\mathcal{L}}(\lambda_{(l-1)l'}-\lambda_{ll'})=0$$

where

$$\tilde{\kappa}=\sup_{l'\in\mathcal{L}}\{V(1,l')-V(x+1,l')\}. \tag{A.10}$$

iv. Immediate since $p(l+1)\geq p(l)$. □

In what follows, we complete the proof of Theorem 2.2.

Proof. The proof of the multi-threshold structure is a consequence of Lemma A.1. In particular, let us define

$$\tau^{(l)}:=\max\{x|\Delta H(x,l)<0\},\ l=1,\dots,L.$$

so that in location l it is optimal to upload for $x\geq\tau^{(l)}$ for all prices $p\leq P_l$; also, from *ii*., it follows that $\tau^{(1)}\leq\tau^{(2)}\dots\leq\tau^{(L)}$. □

A.2 Proof of Theorem 2.4

Proof. (i) The following condition has to be satisfied for the optimal policy to be always using price $P_1=0$:

$$\Delta H(x,l)<0,\ \forall l\in\mathcal{L}/\mathcal{L}_1,\ x=1,..,M. \tag{A.11}$$

From (A.2) and (A.3), the condition (A.11) yields

$$\sum_{l'\in\mathcal{L}}\lambda_{ll'}[V(1,l')-V(x+1,l')]<p(l),$$
$$\forall l\in\mathcal{L}/\mathcal{L}_1,x=1,..,M-1. \tag{A.12}$$

Since the value function V is non-increasing, the conditions in (A.12) are satisfied if and only if

$$\sum_{l'\in\mathcal{L}_1}\lambda_{ll'}[V(1,l')-V(M,l')]+\sum_{l'\notin\mathcal{L}_1}\lambda_{ll'}[V(1,l')-V(M,l')]<p(l), \tag{A.13}$$

$\forall l\in\mathcal{L}/\mathcal{L}_1$. Let assume the device can upload data only at location $l\in\mathcal{L}_1$: from (A.1) we have

$$V(x,l)-V(x-1,l)=U(x)-U(x-1),\ \forall l\in\mathcal{L}_1 \tag{A.14}$$
$$V(x,l)-V(x-1,l)=U(M)-U(x-1),\ \forall l\in\mathcal{L}/\mathcal{L}_1 \tag{A.15}$$

(A.14) and (A.15) yield, respectively,

$$V(1,l) - V(M,l) = U(1) - U(M),\ \forall l \in \mathcal{L}_1 \tag{A.16}$$

$$V(1,l) - V(M,l) = \sum_{x=1}^{M} (U(x) - U(M)),\ \forall l \in \mathcal{L}/\mathcal{L}_1 \tag{A.17}$$

Plugging these values of $V(1,l) - V(M,l)$ into (A.13) gives the condition (2.11).

The derivations of (ii) and (iii) are similar to the above proof. □

A.3 Proof of Lemma 3.1

If the temperature is fixed, the transition matrix of the resulting time-reversible Markov chain, $Q_T = (q_T(\boldsymbol{\tau}, \boldsymbol{\tau}'))$, $0 \leq \tau_i, \tau_i' \leq \tau_{max}$, is given by:

$$q_T(\boldsymbol{\tau}, \boldsymbol{\tau}') = \tag{A.18}$$

$$\begin{cases} q^*(\boldsymbol{\tau}, \boldsymbol{\tau}')\tilde{a}(\boldsymbol{\tau}, \boldsymbol{\tau}') & \text{if } \tilde{a}(\boldsymbol{\tau}, \boldsymbol{\tau}') < 1 \text{ and } \boldsymbol{\tau} \neq \boldsymbol{\tau}' \\ q^*(\boldsymbol{\tau}, \boldsymbol{\tau}'), & \text{if } \tilde{a}(\boldsymbol{\tau}, \boldsymbol{\tau}') \geq 1 \text{ and } \boldsymbol{\tau} \neq \boldsymbol{\tau}' \\ q^*(\boldsymbol{\tau}, \boldsymbol{\tau}) + \sum_{\mathbf{z}} q^*(\boldsymbol{\tau}, \mathbf{z}) \left(1 - \min\{1, \frac{\pi_T(\mathbf{z})}{\pi_T(\boldsymbol{\tau})}\}\right), & \\ & \text{if } \boldsymbol{\tau} = \boldsymbol{\tau}'. \end{cases} \tag{A.19}$$

where,

$$\tilde{a}(\boldsymbol{\tau}, \boldsymbol{\tau}') = \frac{\pi_T(\boldsymbol{\tau}')}{\pi_T(\boldsymbol{\tau})} \qquad \text{from (3.11)}$$

In what follows, we drop subscript T to simplify presentation.

Proof. The above discrete time Markov chain, with transition probability matrix given by (A.18), has stationary distribution given by $\pi(\boldsymbol{\tau})$ if the following balance equations hold

$$\sum_{\boldsymbol{\tau} \in S} \pi(\boldsymbol{\tau}) q(\boldsymbol{\tau}, \boldsymbol{\tau}') = \pi(\boldsymbol{\tau}'). \tag{A.20}$$

Next, we show that the above equality holds. Indeed,

$$\sum_{\boldsymbol{\tau} \in S} \pi(\boldsymbol{\tau}) q(\boldsymbol{\tau}, \boldsymbol{\tau}') = \tag{A.21}$$

$$= \sum_{\boldsymbol{\tau} \in C_1(\boldsymbol{\tau}')} \pi(\boldsymbol{\tau}) q(\boldsymbol{\tau}, \boldsymbol{\tau}') + \sum_{\boldsymbol{\tau} \in C_2(\boldsymbol{\tau}')} \pi(\boldsymbol{\tau}) q(\boldsymbol{\tau}, \boldsymbol{\tau}') + \\ + \pi(\boldsymbol{\tau}') q(\boldsymbol{\tau}', \boldsymbol{\tau}') \tag{A.22}$$

where

$$C_1(\boldsymbol{\tau}') = \{\boldsymbol{\tau} \in \boldsymbol{\tau}_{\epsilon,d} \setminus \{\boldsymbol{\tau}'\} \mid \pi(\boldsymbol{\tau}') < \pi(\boldsymbol{\tau})\} \tag{A.23}$$

$$C_2(\boldsymbol{\tau}') = \{\boldsymbol{\tau} \in \boldsymbol{\tau}_{\epsilon,d} \setminus \{\boldsymbol{\tau}'\} \mid \pi(\boldsymbol{\tau}') \geq \pi(\boldsymbol{\tau})\} \tag{A.24}$$

Then,

$$\sum_{\boldsymbol{\tau} \in S} \pi(\boldsymbol{\tau}) q(\boldsymbol{\tau}, \boldsymbol{\tau}') = \tag{A.25}$$

$$= \sum_{\boldsymbol{\tau} \in C_1(\boldsymbol{\tau}')} \pi(\boldsymbol{\tau}) \frac{q^*(\boldsymbol{\tau}, \boldsymbol{\tau}')\pi(\boldsymbol{\tau}')}{\pi(\boldsymbol{\tau})} + \sum_{\boldsymbol{\tau} \in C_2(\boldsymbol{\tau}')} \pi(\boldsymbol{\tau}) q^*(\boldsymbol{\tau}, \boldsymbol{\tau}')$$
$$+ \pi(\boldsymbol{\tau}') \left(q^*(\boldsymbol{\tau}', \boldsymbol{\tau}') + \sum_{z \in C_2(\boldsymbol{\tau}')} q^*(\boldsymbol{\tau}', z) \left(1 - \frac{\pi(z)}{\pi(\boldsymbol{\tau}')} \right) \right) \tag{A.26}$$

$$= \sum_{\boldsymbol{\tau} \in C_1(\boldsymbol{\tau}')} \pi(\boldsymbol{\tau}') q^*(\boldsymbol{\tau}, \boldsymbol{\tau}')$$
$$+ \cancel{\sum_{\boldsymbol{\tau} \in C_2(\boldsymbol{\tau}')} \pi(\boldsymbol{\tau}) q^*(\boldsymbol{\tau}, \boldsymbol{\tau}')} + \pi(\boldsymbol{\tau}') q^*(\boldsymbol{\tau}', \boldsymbol{\tau}')$$
$$+ \pi(\boldsymbol{\tau}') \sum_{z \in C_2(\boldsymbol{\tau}')} q^*(\boldsymbol{\tau}', z) - \cancel{\sum_{z \in C_2(\boldsymbol{\tau}')} q^*(\boldsymbol{\tau}', z)\pi(z)} \tag{A.27}$$

Finally, as Q^* is symmetric and stochastic, $q^*(\boldsymbol{\tau}, \boldsymbol{\tau}') = q^*(\boldsymbol{\tau}', \boldsymbol{\tau})$,

$$\sum_{\boldsymbol{\tau} \in S} \pi(\boldsymbol{\tau}) q(\boldsymbol{\tau}, \boldsymbol{\tau}') = \tag{A.28}$$

$$= \pi(\boldsymbol{\tau}') \left(\sum_{\boldsymbol{\tau} \in C_1(\boldsymbol{\tau}')} q^*(\boldsymbol{\tau}, \boldsymbol{\tau}') + \sum_{\boldsymbol{\tau} \in C_2(\boldsymbol{\tau}')} q^*(\boldsymbol{\tau}, \boldsymbol{\tau}') + q^*(\boldsymbol{\tau}', \boldsymbol{\tau}') \right) \tag{A.29}$$

$$= \pi(\boldsymbol{\tau}'), \tag{A.30}$$

which shows that (A.20) holds and concludes the proof. □

Appendix B

B.1 Proof of theorem 4.3

Proof. We write the Lagrangian for the optimization problem defined in (4.4)-(4.7),

$$L_1(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{p}, \boldsymbol{u}) = \sum_{i=1}^{N} B_i \frac{u_i^{1-\alpha}}{1-\alpha} - \sum_{i=1}^{N}\sum_{l=1}^{L_i}\sum_{r=1}^{R} \lambda_{ilr}\left(u_{il} - \frac{x_{ilr}}{d_{ilr}}\right) - \sum_{l=1}^{L_i}\sum_{r=1}^{R} p_{lr}\left(\sum_{i=1}^{N} x_{ilr} - C_{lr}\right) \tag{B.1}$$

By writing first order necessary and sufficient conditions [15], we have

$$\frac{\partial L_1}{\partial u_{il}} = 0 \quad i \in \mathcal{N}, \forall l \in \mathcal{L}_i. \tag{B.2}$$

$$\frac{\partial L_1}{\partial x_{ilr}} = 0 \qquad i \in \mathcal{N}, \forall l \in \mathcal{L}_i, \forall r \in \mathcal{R}. \tag{B.3}$$

From (B.2), we have

$$B_i u_i^{-\alpha} = \sum_{r=1}^{R} \lambda_{ilr}, \forall i \in \mathcal{N}, \forall l \in \mathcal{L}_i \tag{B.4}$$

From (B.3), we have

$$\frac{\lambda_{ilr}}{d_{ilr}} - p_{lr} = 0, \forall i \in \mathcal{N},\ \forall l \in \mathcal{L}_i,\ \forall r \in \mathcal{R} \tag{B.5}$$

$$\lambda_{ilr} = p_{lr} d_{ilr}, \tag{B.6}$$

Further, we have

$$u_{il} = \frac{x_{ilr}}{d_{ilr}} \quad \text{for } \lambda_{ilr} > 0, \forall i \in \mathcal{N},\ \forall l \in \mathcal{L}_i, \forall r \in \mathcal{R}$$

By using relations in eq. (B.6) and eq.(B.1), we can rewrite the eq. (B.4) as follows

$$B_i u_i^{-\alpha} = \frac{1}{u_{il}} \sum_{r=1}^{R} p_{lr} x_{ilr}, \tag{B.7}$$

$$B_i u_i^{-\alpha} u_{il} = \sum_{r=1}^{R} p_{lr} x_{ilr}. \tag{B.8}$$

By summing over $l \in \mathcal{L}_i$ on both sides, we have

$$B_i u_i^{1-\alpha} = \sum_{l=1}^{L_i} \sum_{r=1}^{R} p_{lr} x_{ilr}. \tag{B.9}$$

we can rewrite the above equation as follows

$$B_i = u_i^{\alpha-1} \sum_{l=1}^{L_i} \sum_{r=1}^{R} p_{lr} x_{ilr}. \tag{B.10}$$

Now, we construct the price curve as

$$\gamma_{ilr}(\boldsymbol{x}_i) = p_{lr} x_{ilr} (u_i)^{\alpha-1}, \forall i \in \mathcal{N},\ \forall l \in \mathcal{L}_i\ \forall r \in \mathcal{R}. \tag{B.11}$$

Given that the cost of the resource bundle is

$$C_\gamma(\boldsymbol{x}_i) = \sum_{l=1}^{L_i} \sum_{r=1}^{R} \gamma_{ilr}(\boldsymbol{x}_i).$$

Now to show that $(\boldsymbol{x}, \boldsymbol{\gamma}(\boldsymbol{x}))$ is a price curve market equilibrium, we need to show that $(\boldsymbol{x}, \boldsymbol{\gamma}(\boldsymbol{x}))$ satisfies the conditions C1 and C2 of Definition 4.2

$$C_\gamma(\boldsymbol{x}_i) = \sum_{l=1}^{L_i} \sum_{r=1}^{R} \gamma_{ilr}(\boldsymbol{x}_i)$$

By substituting $\gamma_{ilr}(\boldsymbol{x}_i)$ in the above equation, we have

$$C_\gamma(\boldsymbol{x}_i) = u_i^{\alpha-1} \sum_{l=1}^{L_i} \sum_{r=1}^{R} p_{lr} x_{ilr} C_\gamma(\boldsymbol{x}_i). \tag{B.12}$$

From eq.(B.10), can write

$$C_\gamma(\boldsymbol{x}_i) = B_i. \tag{B.13}$$

Bundle $\boldsymbol{x}_i$ is affordable to user i, and $\boldsymbol{\gamma}(\boldsymbol{x})$ is strictly increasing, that implies bundle x_i belongs to the demand set. Hence $(\boldsymbol{x}_i, \boldsymbol{\gamma}(\boldsymbol{x}))$ satisfies the condition $C1$ and as $\boldsymbol{x}$ satisfies the KKT condition of (4.4) it satisfies the $C2$ as well thus it proves that $(\boldsymbol{x}, \boldsymbol{\gamma}(\boldsymbol{x}))$ is a ME. □

B.2 Proof of proposition 1

Proof. We write the Lagrangian for the optimization problem defined in (4.12)-(4.15)

$$L_2(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{p}, \boldsymbol{u}) = \sum_{i=1}^{N} B_i \frac{u_i^{1-\alpha}}{1-\alpha} - \sum_{i=1}^{N}\sum_{l=1}^{L_i}\sum_{r=1}^{R} \lambda_{ilr}\left(u_{il} - \frac{x_{ilr}}{d_{ilr}}\right) - \sum_{l=1}^{L_i}\sum_{r=1}^{R} p_{lr}\left(\sum_{i=1}^{N} x_{ilr} - C_{lr}\right). \quad \text{(B.14)}$$

And, we already have

$$u_i = \Big(\sum_{l\in\mathcal{L}_i} (u_{il})^{1-\beta}\Big)^{\frac{1}{1-\beta}}. \quad \text{(B.15)}$$

By writing first order necessary and sufficient conditions [15], we have

$$\frac{\partial L_2}{\partial u_{il}} = 0 \quad i \in \mathcal{N}, \forall l \in \mathcal{L}_i. \quad \text{(B.16)}$$

$$\frac{\partial L_2}{\partial x_{ilr}} = 0 \qquad i \in \mathcal{N}, \forall l \in \mathcal{L}_i, \forall r \in \mathcal{R}. \quad \text{(B.17)}$$

From (B.16), we have

$$B_i u_i^{-\alpha} \left(\sum_{l=1}^{L_i} {u_{il}}^{1-\beta}\right)^{(\frac{1}{1-\beta})-1} {u_{il}}^{1-\beta-1} = \sum_{r=1}^{R} \lambda_{ilr}, \forall i \in \mathcal{N},\ l \in \mathcal{L}_i. \quad \text{(B.18)}$$

From (B.17), we have

$$\frac{\lambda_{ilr}}{d_{ilr}} - p_{lr} = 0, \forall i \in \mathcal{N},\ \forall l \in \mathcal{L}_i,\ \forall r \in \mathcal{R}, \quad \text{(B.19)}$$

$$\lambda_{ilr} = p_{lr} d_{ilr}. \quad \text{(B.20)}$$

Further, we have

$$u_i = \frac{x_{ilr}}{d_{ilr}} \quad \text{for } \lambda_{ilr} > 0.$$

From last two equations, we can write

$$\lambda_{ilr} = p_{lr} u_{il}^{-1} x_{ilr}. \tag{B.21}$$

By substituting the above equation, we can rewrite the equation in eq.(B.18) as follows

$$B_i u_i^{-\alpha} \left(\sum_{l=1}^{L_i} {u_{il}}^{1-\beta} \right)^{(\frac{1}{1-\beta})-1} {u_{il}}^{1-\beta-1} = \sum_{r=1}^{R} p_{lr} u_{il}^{-1} x_{ilr}. \tag{B.22}$$

We can rewrite the above equation as follows

$$B_i u_i^{-\alpha} \left(\sum_{l=1}^{L_i} {u_{il}}^{1-\beta} \right)^{(\frac{1}{1-\beta})-1} {u_{il}}^{1-\beta} = \sum_{r=1}^{R} p_{lr} x_{ilr}. \tag{B.23}$$

By summing over $l \in \mathcal{L}_i$ on both sides, we have

$$B_i u_i^{-\alpha} \left(\sum_{l=1}^{L_i} {u_{il}}^{1-\beta} \right)^{(\frac{1}{1-\beta})-1} \sum_{l=1}^{L_i} {u_{il}}^{1-\beta} = \sum_{r=1}^{R} p_{lr} x_{ilr}. \tag{B.24}$$

By simplifying the terms in the above equation, we can write

$$B_i = u_i^{\alpha-1} \sum_{r=1}^{R} p_{lr} x_{ilr}. \tag{B.25}$$

Now, we construct the price curve as

$$\gamma_{ilr}(\boldsymbol{x}_i) = p_{lr} x_{ilr} (u_i)^{\alpha-1}, \forall i \in \mathcal{N},\ \forall l \in \mathcal{L}_i\ \forall r \in \mathcal{R}. \tag{B.26}$$

Given that the cost of the resource bundle $\boldsymbol{x}_i$ is

$$C_\gamma(\boldsymbol{x}_i) = \sum_{l=1}^{L_i} \sum_{r=1}^{R} \gamma_{ilr}(\boldsymbol{x}_i).$$

Now, to show that $(\boldsymbol{x}, \boldsymbol{\gamma}(\boldsymbol{x}))$ is a price curve market equilibrium, we need to show that $(\boldsymbol{x}, \boldsymbol{\gamma}(\boldsymbol{x}))$ satisfies the conditions C1 and C2 of Definition 4.2,

By substituting the value of $\gamma_{ilr}(\boldsymbol{x}_i)$ in the above equation, we have

$$C_\gamma(\boldsymbol{x}_i) = u_i^{\alpha-1} \sum_{l=1}^{L_i} \sum_{r=1}^{R} p_{lr}^* x_{ilr}.$$

From (B.25), we can write

$$C_\gamma(\boldsymbol{x}_i) = B_i. \tag{B.27}$$

Bundle $\boldsymbol{x}_i$ is affordable to user i, and $\boldsymbol{\gamma}(\boldsymbol{x})$ is strictly increasing, that implies bundle x_i belongs to the demand set. Hence $(\boldsymbol{x}_i, \boldsymbol{\gamma}(\boldsymbol{x}))$ satisfies the condition $C1$ and as $\boldsymbol{x}$ satisfies the KKT condition of (4.4) it satisfies the $C2$ as well thus it proves that $(\boldsymbol{x}, \boldsymbol{\gamma}(\boldsymbol{x}))$ is a ME. □

Appendix C

C.1 Proof for theorem 5.2

Proof. The parameter ϕ_r has several interpretations, it is referred as price per resource as it controls the user demands based on the available capacity in market models. In flow control, ϕ_r is seen as a congestion indicator, where it depends on the aggregate flows through each switch. In this work, we considered ϕ_r to be price associated to the resource r. In chapter 5, we defined a pricing mechanism for each InP as follows

$$\phi_r(t+1) = max(0, \phi_r(t) + \delta(\sum_{i \in \mathcal{N}} x_{ir} - c_r)).$$

Given the price vector $\boldsymbol{\phi}$, the SPs can compute the resource allocation $\boldsymbol{x}_{ir}$ for resource $r \in \mathcal{R}$ according to the following formulation,

$$x_{ir}(\boldsymbol{\phi}) = d_{ir} \left(\sum_{r' \in \mathcal{N}} d_{ir'} \phi_{r'}^{\frac{1}{\beta_{r'}}} \right)^{\frac{-1}{\alpha}}. \tag{C.1}$$

Hence, if the price vector $\boldsymbol{\phi}$ is optimal, then the SPs can compute the optimal allocation vector $\boldsymbol{x}^*$ that is a social optimal.

At optimal price vector, we have the following

$$\frac{\partial}{\partial \phi_r}(L_{sys}(\boldsymbol{u}, \boldsymbol{x}, \boldsymbol{\gamma}, \boldsymbol{\phi})) = 0, \quad \forall r \in \mathcal{R}. \tag{C.2}$$

The theorem in [108] shows that if we prove that the gradient $\frac{\partial}{\partial \phi_r}(L_{sys}(\boldsymbol{u}, \boldsymbol{x}, \boldsymbol{\gamma}, \boldsymbol{\phi}))$ is Lipschitz continuous, then with an appropriate selection of step size δ, the price vector $\boldsymbol{\phi}$ will converge in $\hat{\boldsymbol{\phi}}$. At this point the gradient satisfies the relation in (C.2).

Hence the proof of the theorem 5.2 strongly depends on proving that the $\frac{\partial}{\partial \phi_r}(L_{sys}(\boldsymbol{u}, \boldsymbol{x}, \boldsymbol{\gamma}, \boldsymbol{\phi}))$ is Lipschitz continuous. To show that the gradient in question is Lipschitz continuous, we must prove the following inequality is true.

$$\| \frac{\partial}{\partial \phi_r}(L_{sys}(\boldsymbol{u}, \boldsymbol{x}, \boldsymbol{\gamma}, \boldsymbol{\phi})) - \frac{\partial}{\partial \bar{\phi}_r}(L_{sys}(\boldsymbol{u}, \boldsymbol{x}, \boldsymbol{\gamma}, \bar{\boldsymbol{\phi}})) \|_1 \leq K \| \boldsymbol{\phi} - \bar{\boldsymbol{\phi}} \|_1 \tag{C.3}$$

where,

$$\frac{\partial}{\partial \phi_r}(L_{sys}(\boldsymbol{u}, \boldsymbol{x}, \boldsymbol{\gamma}, \boldsymbol{\phi})) = \sum_{i \in \mathcal{N}} x_{ir} - c_r, \quad \forall r \in \mathcal{R}. \tag{C.4}$$

We can write

$$|x_r(\boldsymbol{\phi}) - x_r(\bar{\boldsymbol{\phi}})|_1 \leq \sum_i |x_{ir}(\boldsymbol{\phi}) - x_{ir}(\bar{\boldsymbol{\phi}})|. \tag{C.5}$$

By using the expression x_{ir} presented in eq. (C.1), we can write the following

$$|x_{ir}(\boldsymbol{\phi}) - x_{ir}(\bar{\boldsymbol{\phi}})| = d_{ir} \left| \left(\sum_{r' \in \mathcal{R}} d_{ir'} \phi_{r'}^{\frac{1}{\beta_{r'}}} \right)^{\frac{-1}{\alpha}} - \left(\sum_{r' \in \mathcal{R}} d_{ir'} (\bar{\phi}_{r'})^{\frac{1}{\beta_{r'}}} \right)^{\frac{-1}{\alpha}} \right| \tag{C.6}$$

let

$$y = \sum_{r' \in \mathcal{N}} d_{ir'} \phi_{r'}^{\frac{1}{\beta_{r'}}}$$

Then, consider

$$h(y) = y^{-1/\alpha}$$

Following is the first derivative of the $h(y)$,

$$\frac{\partial h(y)}{\partial y} = -\frac{1}{\alpha} y^{-\frac{1}{\alpha}-1}.$$

We set bounds on the value of $\phi_r \in [\phi_{min}, \phi_{max}]$ such that $\phi_r = \phi_{max}$ when the aggregated resource allocation $\sum_{i \in \mathcal{N}} x_{ir}$ exceeds the available capacity c_r, this prevents the value of ϕ_r reaching ∞, on the contrary, $\phi_r = \phi_{min}$ when $\sum_{i \in \mathcal{N}} x_{ir}$ is significantly lower in comparison to c_r, this prevents the value of ϕ_r from reaching zero.

$$|h(y) - h(\bar{y})| \leq \frac{1}{\alpha} \left(\frac{1}{\phi_{min}} \right)^{-(\frac{1}{\alpha}+1)} |y - \bar{y}| \tag{C.7}$$

Now, let $z = \phi_r$ and $f(z) = z^{\frac{1}{\beta_r}}$, the first derivative of $f(z)$ with respect to z is

$$f'(z) = \frac{1}{\beta_r} z^{\frac{1}{\beta_r}-1}$$

if $\beta_r > 1$, then

$$|f(z) - f(\bar{z})| \leq \frac{1}{\beta_r} (\phi_{min})^{\frac{1}{\beta_r}-1} |z - \bar{z}|, \tag{C.8}$$

if $\beta_r < 1$, then

$$|f(z) - f(\bar{z})| \leq \frac{1}{\beta_r} (\phi_{max})^{\frac{1}{\beta_r}-1} |z - \bar{z}|. \tag{C.9}$$

From eq.(C.8) and (C.9), we can write the following

$$|\sum_{r'\in\mathcal{R}} \phi_{r'}^{\frac{1}{\beta_{r'}}} - \sum_{r'\in\mathcal{R}} (\bar{\phi}_{r'})^{\frac{1}{\beta_{r'}}}| \leq \sum_{r'\in\mathcal{R}} \frac{1}{\beta_{r'}} \underset{r'}{max} \left(d_{ir'}(\phi_{min})^{\frac{1}{\beta_{r'}}-1}, d_{ir'}\phi_{max})^{\frac{1}{\beta_{r'}}-1} \right) |\phi_{r'} - \bar{\phi}_{r'}|. \tag{C.10}$$

Let,

$$K_{r'} = \{\frac{d_{ir'}}{\beta_{r'}}(\phi_{min})^{\frac{1}{\beta_{r'}}-1}, \frac{d_{ir'}}{\beta_{r'}}(\phi_{max})^{\frac{1}{\beta_{r'}}-1}\} \tag{C.11}$$

Now, we can rewrite the eq. (C.10)

$$|\sum_{r'\in\mathcal{R}} \phi_{r'}^{\frac{1}{\beta_{r'}}} - \sum_{r'\in\mathcal{R}} (\bar{\phi}_{r'})^{\frac{1}{\beta_{r'}}}| \leq \underset{r'}{max} K_{r'} \sum_{r'\in\mathcal{R}} |\phi_{r'} - \bar{\phi}_{r'}| \leq \underset{r'}{max} K_{r'} \|\phi_r - \bar{\phi}_r\|_1. \tag{C.12}$$

From eq. (C.5), (C.6), (C.7) and (C.12), we have the following

$$|x_r(\boldsymbol{\phi}) - x_r(\bar{\boldsymbol{\phi}})| \leq d_{ir} \sum_{i\in\mathcal{N}} \frac{1}{\alpha} \left(\frac{1}{\phi_{min}}\right)^{-(\frac{1}{\alpha}+1)} \underset{r'}{max} K_{r'} \|\boldsymbol{\phi} - \bar{\boldsymbol{\phi}}\|_1. \tag{C.13}$$

By taking $\sum_{r\in\mathcal{R}}$ on both sides, we have

$$\|x_r(\boldsymbol{\phi}) - x_r(\bar{\boldsymbol{\phi}})\|_1 \leq R \sum_{i\in\mathcal{N}} \frac{1}{\alpha} \left(\frac{1}{\phi_{min}}\right)^{-(\frac{1}{\alpha}+1)} \underset{r'}{max} K_{r'} \|\boldsymbol{\phi} - \bar{\boldsymbol{\phi}}\|_1. \tag{C.14}$$

Let $\sigma = R \sum_{i\in\mathcal{N}} \frac{1}{\alpha} \left(\frac{1}{\phi_{min}}\right)^{-(\frac{1}{\alpha}+1)} \underset{r'}{max} K_{r'}$, then we can rewrite the above equation as follows

$$\|x_r(\boldsymbol{\phi}) - x_r(\bar{\boldsymbol{\phi}})\|_1 \leq \sigma \|\boldsymbol{\phi} - \bar{\boldsymbol{\phi}}\|_1. \tag{C.15}$$

Therefore, $\frac{\partial}{\partial \phi_r}(L_{sys}(\boldsymbol{u}, \boldsymbol{x}, \boldsymbol{\gamma}, \boldsymbol{\phi}))$ is σ - Lipschitz continuous. Following the proof in [108], we can conclude that if the gradient of $C(\phi)$ is σ - Lipschitz continuous, then given a step size $\delta \in (0, 2/\sigma]$, $\boldsymbol{\phi}$ will converge in $\hat{\boldsymbol{\phi}}$.

□

BIBLIOGRAPHY

[1] 3GPP, *NR; Physical channels and modulation*, Technical Specification (TS) 38.211, Version 17.0.0, january,2022, URL: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3213 (cit. on p. 76).

[2] 3GPP, *System architecture for the 5G System (5GS)*, Technical Specification (TS) 23.501, Version 17.4.0, march,2022, URL: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3144 (cit. on p. 76).

[3] *5G; Management and orchestration; Concepts, use cases and requirements (3GPP TS 128.530 version 15.0.0 Release 15 (2018-10)* (cit. on pp. 29, 30).

[4] Isiaka A. Alimi, Romil K. Patel, Nelson J. Muga, Armando N. Pinto, António L. Teixeira, and Paulo P. Monteiro, "Towards Enhanced Mobile Broadband Communications: A Tutorial on Enabling Technologies, Design Considerations, and Prospects of 5G and beyond Fixed Wireless Access Networks", *in*: *Applied Sciences* 11.*21* (2021), ISSN: 2076-3417, DOI: 10.3390/app112110427, URL: https://www.mdpi.com/2076-3417/11/21/10427 (cit. on pp. 13, 76).

[5] Eitan Altman, Konstantin Avrachenkov, and Andrey Garnaev, "Generalized α-fair resource allocation in wireless networks", *in*: *2008 47th IEEE Conference on Decision and Control*, IEEE, 2008, pp. 2414–2419 (cit. on p. 60).

[6] Eitan Altman, Rachid El-Azouzi, Daniel Sadoc Menasche, and Yuedong Xu, *Forever young: Aging control in dtns*, tech. rep., 2010 (cit. on p. 18).

[7] Eitan Altman, Rachid El-Azouzi, Daniel Sadoc Menasche, and Yuedong Xu, "Forever young: for hybrid networks", *in*: *Proceedings of the Twentieth ACM International Symposium on Mobile Ad Hoc Networking and Computing*, 2019, pp. 91–100 (cit. on pp. 20, 28, 29).

[8] *Amazon EC2 instances*, URL: https://aws.amazon.com/ec2/instance-types/ (cit. on p. 71).

[9] Anthony Barnes Atkinson, "On the Measurement of Inequality', Journal of Economic Theory 2", *in*: *1970. 2: 244* 263 (1970) (cit. on p. 81).

[10] Baran Tan Bacinoglu and Elif Uysal-Biyikoglu, "Scheduling status updates to minimize age of information with an energy harvesting sensor", *in*: *2017 IEEE international symposium on information theory (ISIT)*, IEEE, 2017, pp. 1122–1126 (cit. on p. 19).

[11] S. Bajoudah, C. Dong, and P. Missier, "Toward a Decentralized, Trust-Less Marketplace for Brokered IoT Data Trading Using Blockchain", *in*: *2019 IEEE International Conference on Blockchain (Blockchain)*, 2019, pp. 339–346, DOI: 10.1109/Blockchain.2019.00053 (cit. on p. 25).

[12] Ahmed M Bedewy, Yin Sun, Sastry Kompella, and Ness B Shroff, "Age-optimal sampling and transmission scheduling in multi-source systems", *in*: *Proceedings of the Twentieth ACM International Symposium on Mobile Ad Hoc Networking and Computing*, 2019, pp. 121–130 (cit. on p. 19).

[13] D. Bega, M. Gramaglia, A. Banchs, V. Sciancalepore, K. Samdanis, and X. Costa-Perez, "Optimising 5G infrastructure markets: The business of network slicing", *in*: *Proc. of IEEE INFOCOM*, May 2017, pp. 1–9, DOI: 10.1109/INFOCOM.2017.8057045 (cit. on p. 59).

[14] DP Bertsekas, RG Galleger, and P. Humblet, *Data Networks. Vol. 2*, New Jersey: Prentice-Hall International, 1992, ISBN: 0132009161 (cit. on pp. 77, 85).

[15] Stephen Boyd and Lieven Vandenberghe, *Convex optimization*, Cambridge university press, 2004 (cit. on pp. 22, 85, 87, 106, 108).

[16] William C Brainard and Herbert E Scarf, "How to compute equilibrium prices in 1891", *in*: (2000) (cit. on p. 20).

[17] Simina Brânzei, Yiling Chen, Xiaotie Deng, Aris Filos-Ratsikas, Søren Frederiksen, and Jie Zhang, "The Fisher Market Game: Equilibrium and Welfare", *in*: *Proceedings of the AAAI Conference on Artificial Intelligence* 28.*1* (June 2014) (cit. on pp. 17, 20–22, 60, 70, 81).

[18] Pablo Caballero, Albert Banchs, Gustavo de Veciana, Xavier Costa-Pérez, and Arturo Azcorra, "Network Slicing for Guaranteed Rate Services: Admission Control and Resource Allocation Games", *in*: *IEEE Transactions on Wireless Communications* 17.*10* (2018), pp. 6419–6432, DOI: 10.1109/TWC.2018.2859918 (cit. on p. 60).

[19] Elif Tuğçe Ceran, Deniz Gündüz, and András György, "Average age of information with hybrid ARQ under a resource constraint", *in*: *IEEE Transactions on Wireless Communications* 18.*3* (2019), pp. 1900–1913 (cit. on p. 19).

[20] Aleksandra Checko, Henrik L. Christiansen, Ying Yan, Lara Scolari, Georgios Kardaras, Michael S. Berger, and Lars Dittmann, "Cloud RAN for Mobile Networks—A Technology Overview", *in*: *IEEE Communications Surveys and Tutorials* 17.*1* (2015), pp. 405–426, DOI: 10.1109/COMST.2014.2355255 (cit. on p. 76).

[21] He Chen, Rana Abbas, Peng Cheng, Mahyar Shirvanimoghaddam, Wibowo Hardjawana, Wei Bao, Yonghui Li, and Branka Vucetic, "Ultra-Reliable Low Latency Cellular Networks: Use Cases, Challenges and Approaches", *in*: *IEEE Communications Magazine* 56.*12* (2018), pp. 119–125, DOI: 10.1109/MCOM.2018.1701178 (cit. on pp. 13, 76).

[22] Junghoo Cho and Hector Garcia-Molina, "Effective page refresh policies for web crawlers", *in*: *ACM Transactions on Database Systems (TODS)* 28.*4* (2003), pp. 390–426 (cit. on pp. 18, 19).

[23] *Cisco Annual Internet Report (2018–2023)*, tech. rep., May 2020 (cit. on pp. 25, 29).

[24] D Connors and P Kumar, "Simulated annealing type markov chains and their order balance equations", *in*: *SIAM Journal on Control and Optimization*, 1989, pp. 1440–1461 (cit. on p. 45).

[25] Maice Costa, Marian Codreanu, and Anthony Ephremides, "Age of information with packet management", *in*: *2014 IEEE International Symposium on Information Theory*, IEEE, 2014, pp. 1583–1587 (cit. on p. 19).

[26] Zaher Dawy, Walid Saad, Arunabha Ghosh, Jeffrey G. Andrews, and Elias Yaacoub, "Toward Massive Machine Type Cellular Communications", *in*: *IEEE Wireless Communications* 24.*1* (2017), pp. 120–128, DOI: 10.1109/MWC.2016.1500284WC (cit. on pp. 13, 76).

[27] Peter Decker and Bernhard Walke, "A general packet radio service proposed for GSM", *in*: *GSM in a Future Competitive Environment, Helsinki, Finland* (1993), pp. 1–20 (cit. on p. 12).

[28] Edmund Eisenberg and David Gale, "Consensus of Subjective Probabilities: The Pari-Mutuel Method", *in*: *The Annals of Mathematical Statistics* 30.*1* (1959), pp. 165–168, ISSN: 00034851 (cit. on pp. 22, 60, 66, 77).

[29] Songtao Feng and Jing Yang, "Minimizing age of information for an energy harvesting source with updating failures", *in*: *2018 IEEE International Symposium on Information Theory (ISIT)*, IEEE, 2018, pp. 2431–2435 (cit. on p. 19).

[30] Francesca Fossati, Stéphane Rovedakis, and Stefano Secci, "Distributed Algorithms for Multi-Resource Allocation", *in*: *IEEE Transactions on Parallel and Distributed Systems* 33.*10* (2022), pp. 2524–2539, DOI: 10.1109/TPDS.2022.3144376 (cit. on pp. 78, 86).

[31] Ali Ghodsi, Matei Zaharia, Benjamin Hindman, Andy Konwinski, Scott Shenker, and Ion Stoica, "Dominant resource fairness: Fair allocation of multiple resource types", *in*: *8th USENIX Symposium on Networked Systems Design and Implementation (NSDI 11)*, 2011 (cit. on pp. 60, 77).

[32] Mahdi Ben Ghorbel, David Rodríguez-Duarte, Hakim Ghazzai, Md. Jahangir Hossain, and Hamid Menouar, "Joint Position and Travel Path Optimization for Energy Efficient Wireless Data Gathering Using Unmanned Aerial Vehicles", *in*: *IEEE Transactions on Vehicular Technology* 68.*3* (2019), pp. 2165–2175, DOI: 10.1109/TVT.2019.2893374 (cit. on p. 75).

[33] Lal C Godara, "Application of antenna arrays to mobile communications. II. Beam-forming and direction-of-arrival considerations", *in*: *Proceedings of the IEEE* 85.*8* (1997), pp. 1195–1245 (cit. on p. 14).

[34] Ashish Goel, Reyna Hulett, and Benjamin Plaut, "Markets Beyond Nash Welfare for Leontief Utilities", *in*: *CoRR* abs/1807.05293 (2018), arXiv: 1807.05293, URL: http://arxiv.org/abs/1807.05293 (cit. on pp. 22, 60, 61, 66).

[35] Michael Grant and Stephen Boyd, *CVX: Matlab Software for Disciplined Convex Programming, version 2.1*, http://cvxr.com/cvx, Mar. 2014 (cit. on p. 91).

[36] B Hajek, "Cooling schedules for optimal annealing", *in*: *Mathematics of operations research*, 1988, pp. 311–329 (cit. on pp. 48, 49).

[37] Hassan Halabian, "Distributed Resource Allocation Optimization in 5G Virtualized Networks", *in*: *IEEE Journal on Selected Areas in Communications* 37.*3* (2019), pp. 627–642, DOI: 10.1109/JSAC.2019.2894305 (cit. on pp. 77, 78, 85, 86, 90).

[38] Zhu Han and KJ Ray Liu, *Resource allocation for wireless networks: basics, techniques, and applications*, Cambridge university press, 2008 (cit. on pp. 60, 77).

[39] P. Hansen, M. Labbé, and D. Schindl, "Set covering and packing formulations of graph coloring: Algorithms and first polyhedral results", *in*: *Discrete Optimization* 6.*2* (2009), pp. 135–147, ISSN: 1572-5286, DOI: https://doi.org/10.1016/j.disopt.2008.10.004, URL: https://www.sciencedirect.com/science/article/pii/S1572528608000716 (cit. on p. 54).

[40] W.K. Hastings, "Monte Carlo sampling methods using Markov chains and their applications", *in*: *Biometrika* (1970), pp. 97–109 (cit. on p. 47).

[41] Joao P Hespanha, "Modelling and analysis of stochastic hybrid systems", *in*: *IEE Proceedings-Control Theory and Applications* 153.*5* (2006), pp. 520–535 (cit. on p. 19).

[42] Harri Holma, Antti Toskala, Karri Ranta-aho, and Juho Pirskanen, "High-speed packet access evolution in 3gpp release 7 [topics in radio communications]", *in*: *IEEE Communications Magazine* 45.*12* (2007), pp. 29–35 (cit. on p. 12).

[43] Yu-Pin Hsu, "Age of information: Whittle index for scheduling stochastic arrivals", *in*: *2018 IEEE International Symposium on Information Theory (ISIT)*, IEEE, 2018, pp. 2634–2638 (cit. on p. 19).

[44] Yu-Pin Hsu, Eytan Modiano, and Lingjie Duan, "Scheduling algorithms for minimizing age of information in wireless broadcast networks with random arrivals: The no-buffer case", *in*: *arXiv preprint arXiv:1712.07419* (2017) (cit. on p. 19).

[45] Yih-Chun Hu and David B Johnson, "Ensuring cache freshness in on-demand ad hoc network routing protocols", *in*: *Proceedings of the second ACM international workshop on Principles of mobile computing*, 2002, pp. 25–30 (cit. on pp. 18, 19).

[46] Ibrahim, Tarik Taleb, Konstantinos Samdanis, Adlen Ksentini, and Hannu Flinck, "Network Slicing and Softwarization: A Survey on Principles, Enabling Technologies, and Solutions", *in*: *IEEE Communications Surveys and Tutorials* 20.*3* (2018), pp. 2429–2453, DOI: 10.1109/COMST.2018.2815638 (cit. on pp. 14, 75).

[47] Stefano Iellamo, Janne J. Lehtomaki, and Zaheer Khan, "Placement of 5G Drone Base Stations by Data Field Clustering", *in*: *2017 IEEE 85th Vehicular Technology Conference (VTC Spring)*, 2017, pp. 1–5, DOI: 10.1109/VTCSpring.2017.8108590 (cit. on p. 75).

[48] Stratis Ioannidis, Augustin Chaintreau, and Laurent Massoulié, "Optimal and scalable distribution of content updates over a mobile social network", *in*: *IEEE INFOCOM 2009*, IEEE, 2009, pp. 1422–1430 (cit. on p. 19).

[49] Anura P Jayasumana, Qi Han, and Tissa H Illangasekare, "Virtual sensor networks-a resource efficient approach for concurrent applications", *in*: *Fourth International Conference on Information Technology (ITNG'07)*, IEEE, 2007, pp. 111–115 (cit. on p. 26).

[50] Zhiyuan Jiang, Bhaskar Krishnamachari, Xi Zheng, Sheng Zhou, and Zhisheng Niu, "Timely status update in massive IoT systems: Decentralized scheduling for wireless uplinks", *in*: *arXiv preprint arXiv:1801.03975* (2018) (cit. on p. 19).

[51] Zhiyuan Jiang, Bhaskar Krishnamachari, Sheng Zhou, and Zhisheng Niu, "Can decentralized status update achieve universally near-optimal age-of-information in wireless multiaccess channels?", *in*: *2018 30th International Teletraffic Congress (ITC 30)*, vol. 1, IEEE, 2018, pp. 144–152 (cit. on p. 19).

[52] C. Joe-Wong, S. Sen, T. Lan, and M. Chiang, "Multi-resource allocation: Fairness-efficiency tradeoffs in a unifying framework", *in*: *Proc. of IEEE INFOCOM*, Mar. 2012, pp. 1206–1214, DOI: 10.1109/INFCOM.2012.6195481 (cit. on p. 29).

[53] R. Johari, S. Mannor, and J.N. Tsitsiklis, "Efficiency loss in a network resource allocation game: the case of elastic supply", *in*: *IEEE Transactions on Automatic Control* 50.*11* (2005), pp. 1712–1724, DOI: 10.1109/TAC.2005.858687 (cit. on pp. 23, 77, 78, 86).

[54] David Johnson, Cecilia Aragon, Lyle McGeoch, and Catherine Schevon, "Optimization by Simulated Annealing: An Experimental Evaluation; Part II, Graph Coloring and Number Partitioning", *in*: *Operations Research* 39 (June 1991), pp. 378–406, DOI: 10.1287/opre.39.3.378 (cit. on p. 53).

[55] Igor Kadota, Abhishek Sinha, Elif Uysal-Biyikoglu, Rahul Singh, and Eytan Modiano, "Scheduling policies for minimizing age of information in broadcast wireless networks", *in*: *IEEE/ACM Transactions on Networking* 26.*6* (2018), pp. 2637–2650 (cit. on p. 19).

[56] Mamoru Kaneko and Kenjiro Nakamura, "The Nash social welfare function", *in*: *Econometrica: Journal of the Econometric Society* (1979), pp. 423–435 (cit. on p. 85).

[57] Sanjit K Kaul, Roy D Yates, and Marco Gruteser, "Status updates through queues", *in*: *2012 46th Annual conference on information sciences and systems (CISS)*, IEEE, 2012, pp. 1–6 (cit. on p. 19).

[58] Sanjit Kaul, Marco Gruteser, Vinuth Rai, and John Kenney, "Minimizing age of information in vehicular networks", *in*: *2011 8th Annual IEEE communications society conference on sensor, mesh and ad hoc communications and networks*, IEEE, 2011, pp. 350–358 (cit. on p. 18).

[59] Sanjit Kaul, Roy Yates, and Marco Gruteser, "Real-time status: How often should one update?", *in*: *2012 Proceedings IEEE INFOCOM*, IEEE, 2012, pp. 2731–2735 (cit. on pp. 18, 19).

[60] F P Kelly, A K Maulloo, and D K H Tan, "Rate control for communication networks: shadow prices, proportional fairness and stability", *in*: *Journal of the Operational Research Society* 49.*3* (Mar. 1998), pp. 237–252, ISSN: 1476-9360, DOI: 10.1057/palgrave.jors.2600523 (cit. on pp. 17, 22, 23, 26, 60, 77, 78, 86).

[61] Jalal Khamse-Ashari, Ioannis Lambadaris, George Kesidis, Bhuvan Urgaonkar, and Yiqiang Zhao, "Per-Server Dominant-Share Fairness (PS-DSF): A multi-resource fair allocation mechanism for heterogeneous servers", *in*: *2017 IEEE International Conference on Communications (ICC)*, 2017, pp. 1–7, DOI: 10.1109/ICC.2017.7996727 (cit. on p. 60).

[62] P. Kortoci, L. Zheng, C. Joe-Wong, M. Di Francesco, and M. Chiang, "Fog-based Data Offloading in Urban IoT Scenarios", *in*: *Proc. of IEEE INFOCOM 2019*, 2019, pp. 784–792 (cit. on pp. 20, 28).

[63] Ryong Lee, Rae-young Jang, Minwoo Park, Ga-ye Jeon, Jae-kwang Kim, and Sang-hwan Lee, "Making IoT Data Ready for Smart City Applications", *in*: *2020 IEEE International Conference on Big Data and Smart Computing (BigComp)*, 2020, pp. 605–608, DOI: 10.1109/BigComp48618.2020.00020 (cit. on p. 25).

[64] Jesse Levinson, Jake Askeland, Jan Becker, Jennifer Dolson, David Held, Soeren Kammel, J. Zico Kolter, Dirk Langer, Oliver Pink, Vaughan Pratt, Michael Sokolsky, Ganymed Stanek, David Stavens, Alex Teichman, Moritz Werling, and Sebastian Thrun, "Towards fully autonomous driving: Systems and algorithms", *in*: *2011 IEEE Intelligent Vehicles Symposium (IV)*, 2011, pp. 163–168, DOI: 10.1109/IVS.2011.5940562 (cit. on p. 75).

[65] Rui Li, Qian Ma, Jie Gong, Zhi Zhou, and Xu Chen, "Age of Processing: Age-driven Status Sampling and Processing Offloading for Edge Computing-enabled Real-time IoT Applications", *in*: *arXiv preprint arXiv:2003.10916* (2020) (cit. on p. 29).

[66] Xin Li, Mohammed Samaka, H. Anthony Chan, Deval Bhamare, Lav Gupta, Chengcheng Guo, and Raj Jain, "Network Slicing for 5G: Challenges and Opportunities", *in*: *IEEE Internet Computing* 21.*5* (2017), pp. 20–27, DOI: 10.1109/MIC.2017.3481355 (cit. on p. 76).

[67] Long Liu, Xiaoqi Qin, Zhi Zhang, and Ping Zhang, "Joint Task Offloading and Resource Allocation for Obtaining Fresh Status Updates in Multi-Device MEC Systems", *in*: *IEEE Access* 8 (2020), pp. 38248–38261 (cit. on pp. 20, 28, 29).

[68] Lu Lu, Geoffrey Ye Li, A Lee Swindlehurst, Alexei Ashikhmin, and Rui Zhang, "An overview of massive MIMO: Benefits and challenges", *in*: *IEEE journal of selected topics in signal processing* 8.5 (2014), pp. 742–758 (cit. on p. 14).

[69] Nguyen Cong Luong, Ping Wang, Dusit Niyato, Yonggang Wen, and Zhu Han, "Resource Management in Cloud Networking Using Economic Analysis and Pricing Models: A Survey", *in*: *IEEE Communications Surveys Tutorials* 19.2 (2017), pp. 954–1001, DOI: 10.1109/COMST.2017.2647981 (cit. on pp. 60, 78).

[70] Richard T. B. Ma, Dah Ming Chiu, John C. S. Lui, Vishal Misra, and Dan Rubenstein, *On Resource Management for Cloud Users: A Generalized Kelly Mechanism Approach* (cit. on p. 60).

[71] D. Miorandi, S. Sicari, F. De Pellegrini, and I. Chlamtac, "Internet of things: Vision, applications and research challenges", *in*: *Ad Hoc Networks* 10.7 (2012), pp. 1497–1516, ISSN: 1570-8705, DOI: https://doi.org/10.1016/j.adhoc.2012.02.016 (cit. on p. 25).

[72] K. Mišura and M. Žagar, "Data marketplace for Internet of Things", *in*: *2016 International Conference on Smart Systems and Technologies (SST)*, 2016, pp. 255–260, DOI: 10.1109/SST.2016.7765669 (cit. on p. 25).

[73] J. Mo and J. Walrand, "Fair end-to-end window-based congestion control", *in*: *IEEE/ACM Transactions on Networking* 8.5 (2000), pp. 556–567, DOI: 10.1109/90.879343 (cit. on pp. 60, 65, 78).

[74] J. Mo and J. Walrand, "Fair end-to-end window-based congestion control", *in*: *IEEE/ACM Transactions on Networking* 8.5 (2000), pp. 556–567, DOI: 10.1109/90.879343 (cit. on pp. 77, 84).

[75] Naresh Modina, Rachid El-Azouzi, Francesco De Pellegrini, Daniel Sadoc Menasche, and Rosa Figueiredo, *Joint Traffic Offloading and Aging Control in 5G IoT Networks*, 2022, arXiv: 2201.07615 [cs.NI] (cit. on p. 34).

[76] Eugenio Moro and Ilario Filippini, "Joint Management of Compute and Radio Resources in Mobile Edge Computing: a Market Equilibrium Approach", *in*: *IEEE Transactions on Mobile Computing* (2021), pp. 1–1, DOI: 10.1109/TMC.2021.3091764 (cit. on pp. 60, 77).

[77] Hervé Moulin, *Fair division and collective welfare*, MIT press, 2004, chap. chapter 3 (cit. on pp. 22, 85).

[78] John F Nash Jr, "The bargaining problem", *in*: *Econometrica: Journal of the econometric society* (1950), pp. 155–162 (cit. on p. 85).

[79] Duong Tung Nguyen, Long Bao Le, and Vijay Bhargava, "Price-Based Resource Allocation for Edge Computing: A Market Equilibrium Approach", *in*: *IEEE Transactions on Cloud Computing* 9.*1* (2021), pp. 302–317, DOI: 10.1109/TCC.2018.2844379 (cit. on pp. 21, 60, 77).

[80] Duong Tung Nguyen, Long Bao Le, and Vijay K Bhargava, "A market-based framework for multi-resource allocation in fog computing", *in*: *IEEE/ACM Transactions on Networking* 27.*3* (2019), pp. 1151–1164 (cit. on pp. 60, 77, 85).

[81] Jose Ordonez-Lucena, Pablo Ameigeiras, Diego Lopez, Juan J Ramos-Munoz, Javier Lorca, and Jesus Folgueira, "Network slicing for 5G with SDN/NFV: Concepts, architectures, and challenges", *in*: *IEEE Communications Magazine* 55.*5* (2017), pp. 80–87 (cit. on p. 15).

[82] C. Perera, C. H. Liu, and S. Jayawardena, "The Emerging Internet of Things Marketplace From an Industrial Perspective: A Survey", *in*: *IEEE Transactions on Emerging Topics in Computing* 3.*4* (2015), pp. 585–598, DOI: 10.1109/TETC.2015.2390034 (cit. on p. 25).

[83] M. Pincus, "Monte Carlo method for the approximate solution of certain types of constrained optimization problems", *in*: *Operations Research*, 1970, pp. 1225–1228 (cit. on p. 45).

[84] Petar Popovsk, Volker Brau, Hans-Peter Mayer, Peter Fertl, Zhe Ren, David Gonzales-Serrano, Erik G. Ström, Tommy Svensson, Hidekazu Taoka, Patrick Agyapong, Anass Benjebbour, Gerd Zimmermann, Juha Meinila, Juha Ylitalo, Tommi Jamsa, Pekka Kyosti, Konstantinos D. Dimou, Mikael Fallgren, Yngve Selén, Bogdan Timus, Hugo M. Tullberg, Malte Schellmann, Yuxiang Wu, Martin Schubert, Du Ho Kang, Jan I. Markendahl, Claes Beckman, Mikko A. Uusitalo, Osman N. C. Yilmaz, Carl Wijting, Zexian Li, Patrick Marsch, Krystian Pawlak, Jaakko Vihriala, Alexandre Gouraud, Sebastien Jeux, Mauro Boldi, Gian Michele Dell'Aera, Bruno Melis, Hans D. Schotten, Panagiotis Spapis, Alexandros Kaloxylos, and Konstantinos Chatzikokolakis, "EU FP7 INFSO-ICT-317669 METIS, D1.1 Scenarios, requirements and KPIs for 5G mobile and wireless system", *in*: 2013 (cit. on p. 76).

[85] Martin L Puterman, *Markov decision processes: discrete stochastic dynamic programming*, John Wiley & Sons, 2014 (cit. on pp. 17, 18, 101).

[86] G. Raja, A. Ganapathisubramaniyan, S. Anbalagan, S. B. M. Baskaran, K. Raja, and A. K. Bashir, "Intelligent Reward-Based Data Offloading in Next-Generation Vehicular Networks", *in*: *IEEE Internet of Things Journal* 7.*5* (2020), pp. 3747–3758 (cit. on pp. 20, 28).

[87] Theodore S. Rappaport, Shu Sun, Rimma Mayzus, Hang Zhao, Yaniv Azar, Kevin Wang, George N. Wong, Jocelyn K. Schulz, Mathew Samimi, and Felix Gutierrez, "Millimeter Wave Mobile Communications for 5G Cellular: It Will Work!", *in*: *IEEE Access* 1 (2013), pp. 335–349, DOI: 10.1109/ACCESS.2013.2260813 (cit. on pp. 14, 76).

[88] K. Samdanis, X. Costa-Perez, and V. Sciancalepore, "From network sharing to multi-tenancy: The 5G network slice broker", *in*: *IEEE Communications Magazine* 54.7 (July 2016), pp. 32–39, ISSN: 0163-6804, DOI: 10.1109/MCOM.2016.7514161 (cit. on p. 25).

[89] Fabian Schomm, Florian Stahl, and Gottfried Vossen, "Marketplaces for Data: An Initial Survey", *in*: *SIGMOD Rec.* 42.1 (May 2013), pp. 15–26, ISSN: 0163-5808, DOI: 10.1145/2481528.2481532, URL: https://doi.org/10.1145/2481528.2481532 (cit. on p. 25).

[90] V. Sciancalepore, K. Samdanis, X. Costa-Perez, D. Bega, M. Gramaglia, and A. Banchs, "Mobile traffic forecasting for maximizing 5G network slicing resource utilization", *in*: *Proc. of IEEE INFOCOM*, May 2017, pp. 1–9, DOI: 10.1109/INFOCOM.2017.8057230 (cit. on p. 29).

[91] Amartya Sen, "Welfare inequalities and Rawlsian axiomatics", *in*: *Theory and decision* 7.4 (1976), pp. 243–262 (cit. on p. 85).

[92] Stefania Sesia, Issam Toufik, and Matthew Baker, *LTE-the UMTS long term evolution: from theory to practice*, John Wiley & Sons, 2011 (cit. on p. 12).

[93] Weisong Shi, Jie Cao, Quan Zhang, Youhuizi Li, and Lanyu Xu, "Edge Computing: Vision and Challenges", *in*: *IEEE Internet of Things Journal* 3.5 (2016), pp. 637–646, DOI: 10.1109/JIOT.2016.2579198 (cit. on p. 75).

[94] Sameer Kumar Singh, Rohit Singh, and Brijesh Kumbhani, "The evolution of radio access network towards open-RAN: Challenges and opportunities", *in*: *2020 IEEE Wireless Communications and Networking Conference Workshops (WCNCW)*, IEEE, 2020, pp. 1–6 (cit. on p. 76).

[95] Clint Smith and Daniel Collins, *3G wireless networks*, McGraw-Hill Education, 2002 (cit. on p. 12).

[96] Xianxin Song, Xiaoqi Qin, Yunzheng Tao, Baoling Liu, and Ping Zhang, "Age Based Task Scheduling and Computation Offloading in Mobile-Edge Computing Systems", *in*: *Proc. of IEEE WCNCW*, IEEE, 2019, pp. 1–6 (cit. on pp. 20, 28, 29).

[97] R. Srikant and Lei Ying, *Communication Networks: An Optimization, Control and Stochastic Networks Perspective*, USA: Cambridge University Press, 2014, ISBN: 1107036054 (cit. on p. 60).

[98] R. Srikant and Lei Ying, *Communication Networks: An Optimization, Control and Stochastic Networks Perspective*, USA: Cambridge University Press, 2014, ISBN: 1107036054 (cit. on p. 84).

[99] W. Stallings, F. Agboma, and S. Jelassi, *Foundations of Modern Networking: SDN, NFV, QoE, IoT, and Cloud*, The William Stallings books on computer and data communicaitons technology, Pearson, 2015, ISBN: 9780134175478, URL: https://books.google.fr/books?id=QOUOjwEACAAJ (cit. on p. 75).

[100] Xiao Sun, Tan N. Le, Mosharaf Chowdhury, and Zhenhua Liu, "Fair Allocation of Heterogeneous and InterchangeableResources", *in*: *SIGMETRICS Perform. Eval. Rev.* 46.*2* (Jan. 2019), pp. 21–23, ISSN: 0163-5999, DOI: 10.1145/3305218.3305227, URL: https://doi.org/10.1145/3305218.3305227 (cit. on p. 64).

[101] Yin Sun, Elif Uysal-Biyikoglu, Roy D Yates, C Emre Koksal, and Ness B Shroff, "Update or wait: How to keep your data fresh", *in*: *IEEE Transactions on Information Theory* 63.*11* (2017), pp. 7492–7508 (cit. on p. 19).

[102] William Thomson, "Axiomatic and game-theoretic analysis of bankruptcy and taxation problems: a survey", *in*: *Mathematical social sciences* 45.*3* (2003), pp. 249–297 (cit. on p. 80).

[103] Wenqiang Tian and Kevin Lin, "Chapter 2 - Requirements and scenarios of 5G system", *in*: *5G NR and Enhancements*, ed. by Jia Shen, Zhongda Du, Zhi Zhang, Ning Yang, and Hai Tang, Elsevier, 2022, pp. 41–52, ISBN: 978-0-323-91060-6, DOI: https://doi.org/10.1016/B978-0-323-91060-6.00002-7, URL: https://www.sciencedirect.com/science/article/pii/B9780323910606000027 (cit. on p. 76).

[104] S. Uppoor, O. Trullols-Cruces, M. Fiore, and J. M. Barcelo-Ordinas, "Generation and Analysis of a Large-Scale Urban Vehicular Mobility Dataset", *in*: *IEEE Transactions on Mobile Computing* 13.*5* (2014), pp. 1061–1075 (cit. on p. 40).

[105] Barry D Van Veen and Kevin M Buckley, "Beamforming: A versatile approach to spatial filtering", *in*: *IEEE assp magazine* 5.*2* (1988), pp. 4–24 (cit. on p. 14).

[106] Xianwen Wu, Jing Yang, and Jingxian Wu, "Optimal status update for age of information minimization with an energy harvesting source", *in*: *IEEE Transactions on Green Communications and Networking* 2.*1* (2017), pp. 193–204 (cit. on p. 19).

[107] Jinlai Xu, Balaji Palanisamy, Heiko Ludwig, and Qingyang Wang, "Zenith: Utility-Aware Resource Allocation for Edge Computing", *in*: *2017 IEEE International Conference on Edge Computing (EDGE)*, 2017, pp. 47–54, DOI: 10.1109/IEEE.EDGE.2017.15 (cit. on p. 60).

[108] Haıkel Yaıche, Ravi R Mazumdar, and Catherine Rosenberg, "A game theoretic framework for bandwidth allocation and pricing in broadband networks", *in*: *IEEE/ACM transactions on networking* 8.*5* (2000), pp. 667–678 (cit. on pp. 111, 113).

[109] Roy D Yates and Sanjit K Kaul, "The age of information: Real-time status updating by multiple sources", *in*: *IEEE Transactions on Information Theory* 65.*3* (2018), pp. 1807–1827 (cit. on p. 19).

[110] Roy D Yates, Yin Sun, D Richard Brown, Sanjit K Kaul, Eytan Modiano, and Sennur Ulukus, "Age of information: An introduction and survey", *in*: *IEEE Journal on Selected Areas in Communications* 39.*5* (2021), pp. 1183–1210 (cit. on pp. 18, 19).

[111] Shanhe Yi, Zijiang Hao, Zhengrui Qin, and Qun Li, "Fog Computing: Platform and Applications", *in*: *2015 Third IEEE Workshop on Hot Topics in Web Systems and Technologies (HotWeb)*, 2015, pp. 73–78, DOI: 10.1109/HotWeb.2015.22 (cit. on p. 75).

[112] Chaoqun You, Cheng Ren, and Lemin Li, "Online Multi-Resource Social Welfare Maximization for Non-Preemptive Jobs", *in*: *IEEE Access* 8 (2020), pp. 97920–97934, DOI: 10.1109/ACCESS.2020.2996630 (cit. on p. 60).

[113] A. Zanella, N. Bui, A. Castellani, L. Vangelista, and M. Zorzi, "Internet of Things for Smart Cities", *in*: *IEEE Internet of Things Journal* 1.*1* (2014), pp. 22–32 (cit. on p. 25).

[114] H. Zhang, N. Liu, X. Chu, K. Long, A. Aghvami, and V. C. M. Leung, "Network Slicing Based 5G and Future Mobile Networks: Mobility, Resource Management, and Challenges", *in*: *IEEE Communications Magazine* 55.*8* (Aug. 2017), pp. 138–145, ISSN: 0163-6804 (cit. on p. 25).

[115] Nan Zhang, Ya-Feng Liu, Hamid Farmanbar, Tsung-Hui Chang, Mingyi Hong, and Zhi-Quan Luo, "Network Slicing for Service-Oriented Networks Under Resource Constraints", *in*: *IEEE Journal on Selected Areas in Communications* 35.*11* (2017), pp. 2512–2521, DOI: 10.1109/JSAC.2017.2760147 (cit. on p. 59).

[116] Lei Zhao, Jiadai Wang, Jiajia Liu, and Nei Kato, "Optimal Edge Resource Allocation in IoT-Based Smart Cities", *in*: *IEEE Network* 33.*2* (2019), pp. 30–35, DOI: 10.1109/MNET.2019.1800221 (cit. on p. 60).

[117] Jiaxiao Zheng, Pablo Caballero, Gustavo de Veciana, Seung Jun Baek, and Albert Banchs, "Statistical Multiplexing and Traffic Shaping Games for Network Slicing", *in*: *IEEE/ACM Trans. Netw.* 26.*6* (Dec. 2018), pp. 2528–2541, ISSN: 1063-6692 (cit. on p. 29).

[118] Jiaxiao Zheng and Gustavo de Veciana, "Elastic Multi-resource Network Slicing: Can Protection Lead to Improved Performance?", *in*: *2019 International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOPT)*, 2019, pp. 1–8, DOI: 10.23919/WiOPT47501.2019.9144138 (cit. on p. 60).

[119] Bo Zhou and Walid Saad, "Joint status sampling and updating for minimizing age of information in the Internet of Things", *in*: *IEEE Transactions on Communications* 67.*11* (2019), pp. 7468–7482 (cit. on p. 19).

AVIGNON UNIVERSITÉ

Pricing design and resource allocation for 5G services

by

Naresh Modina

LABORATOIRE INFORMATIQUE D'AVIGNON