Utility-based Packet Scheduling and Resource Allocation Algorithms with Heterogeneous Traffic for Wireless OFDMA Networks

by

Alireza Sharifian

A thesis submitted to the Faculty of Graduate and Postdoctoral Affairs
in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

in

System and Computer Engineering

Carleton University
Ottawa, Ontario, Canada

September 2014

Abstract

The focus of this thesis is on studying various packet scheduling and resource block (RB) allocation algorithms, for different flow requirements and architecture. In the second chapter, we investigate packet scheduling and resource allocation algorithms for realtime (RT) and non-realtime (NRT) packet-switched flows, in an orthogonal frequency division multiple access (OFDMA) wireless networks. We start by specifying different quality-of-service (QoS) requirements for RT and NRT flows and present different packet scheduling and RB allocation algorithms based on the bit-rate utilities and the delay utilities, in a harmonized manner. For RT flows, we proposed and analyzed a novel delay fairness notion and proved two related propositions. We, then, summarize the machinery for attaining different fairness and QoS requirements in a classification table and show a novel intuitive decomposition of the structure of each packet scheduling and RB allocation core based its properties.

When considering RT and NRT flows together, the commonly-used approach is the one based on sequential scheduling of RT and NRT flows. This approach cannot exploit the potent existent multiuser diversity, in wireless OFDMA networks. In the third chapter, we propose a novel unified disutility minimization, in a common pool of RBs. We use mean bit-rate, mean queue length, and head-of-the-line (HOL) delay information, in addition to channel information embedded in gradient of dis-utilities, to match the demand and supply. Since the packet scheduling and RB allocation algorithms are taken place for RT and NRT flows from a common pool of RBs (without static priority separation), the overall spectral efficiency is increased. The novel formulations are used to devise computationally-efficient packet schedulers that surpass baseline schedulers in terms of output bit-rates and delay performances. Our approach can be extended for broader QoS re-

quirements and for the utilities of the future applications. We also develop a novel general model for input-output bit-rate behaviour in resource allocation of the heterogenous traffic. This model sheds light on identifying different input load regions and understanding of the system in a simple intuitive manner.

When it comes to providing very high bit-rate coverage, wireless networks require cost-effective radio access network (RAN) devices, such as multiuser enabled amplify-and-forward (AF) relays, with proper fair packet scheduling and RB allocation algorithm. These relays are cost-effective, simpler to implement, and introduce less delay in comparison to other relay based routers. In the fourth chapter, we develop novel fair packet scheduling and RB allocation algorithm in this kind of OFDMA based AF relays.

Finally, we will discuss a number of interesting candidate research topics, as future direction, in the last chapter.

Acknowledgements

First and foremost, I am grateful to God, the Gracious and the Merciful. I am very thankful to my lovely parents for their endless support and the encouragement through my entire life. I would like to acknowledge the encouragement, patient guidance, and the support I received from Prof. Hallim Yanıkömeroğlu, as well as his tremendous efforts to build and to continuously expand a dynamic large research group with members from the industry and from the academia. Many thanks go to all the members of our research group, especially, my project managers, industry collaborators, and friends, Dr. Petar Djukic from Ciena-Ottawa, Dr. Rainer Schoenen from Carleton university, Dr. Gamini Senarath from Huawei-Ottawa, Dr. Tony Zhang from Huawei-Shenzhen, and Dr. Ho Ting Cheng from BLiNQ networks for their comments on my research and for being instrumental in the patent filing processes.

This work was supported by Huawei-Shenzhen, Huawei-Ottawa, Ontario Ministry of Economic Development and Innovations ORF-RE (Ontario Research Fund - Research Excellence) program, and Carleton university.

Contents

List of Figures

List of Tables

List of Acronyms

The acronyms, used in the thesis, are summarized alphabetically in the following table. The acronyms are defined, throughout the text, independently for each chapter and independently for the abstract of each chapter.

Acronym	**Definition**
AC	Admission controller
AF	Amplify-and-forward
AFGPF	Algorithm generalized proportional fair for amplify-and-forward
AFGPF-EFF	Efficient implementation of the AFGPF
AFMM	Algorithm maxmin bit-rate fair for amplify-and-forward
AMC	Adaptive modulation and coding
ARQ	Automatic repeat request
BE	Best-effort
BER	Bit error rate
BS	Base station
CDF	Cumulative distribution function
CP	Cyclic prefix
CDMA	Code division multiple access
CSFB	Circuit-switched fall back
CSI	Channel state information

DCH	Dedicated channel
DF	Decode-and-forward
EDF	Earliest deadline first
EJT	Efficiency-Jain tradeoff
EXP	Exponential-rule
FEC	Forward error correction
FDD	Frequency division multiple access
FFT	Fast Fourier transform
GPF	Generalized proportional fairness
GNRT	Generalized-non-realtime
HOL	Head-of-the-line
HDR	High data rate legacy network
ICIC	Intercell Interference Coordination
IoT	Internet of things
LMS	Least mean-square
LTE	Long-term evolution networks
M2M	Machine-to-machine communication
MAC layer	Medium access control layer
MDP	Markov decision process
MDU	Maximum delay utility
MLWDF	Modified largest weighted delay first
NRT	Non-realtime
OSP	Ordered sub-channel pairing
OFDMA	Orthogonal frequency division multiple access.
PF	Proportional fair
PHY layer	Physical layer
PLR	Packet loss ratio

QoE	Quality of experience
QoS	Quality-of-service
RAN	Radio access network
RB	Resource block
RRM	Radio resource management
RS	Relay station
RT	Realtime
RTOFT	RT-only output for the first time
sat.	Saturation
sat. I	Saturation region I
sat. II	Saturation region II
st. d.	Standard deviation
SFT	Saturated for the first time
SINR	Signal to noise and interference ratio
SVLTE	Simultaneous voice and LTE
TDMA	Time division multiple access
TDD	Time division duplexing
TCP	Transmission control protocol
TUF	Time-utility function
WT	Wireless terminal
VoIP	Voice-over-IP
VoLTE	Voice-over-LTE
WGPF	Weighted generalized proportional fairness
4G	4-th generation cellular networks
5G	5-th generation cellular networks

List of Symbols

Lists of symbols are separated for each chapter, after the introduction of each chapter, in order to be accessed more comfortably.

Chapter 1

Introduction

1.1 Motivation

New Services, Heterogeneity, Importance of QoS, Over-provisioning, and Capacity Crunch

The main requirement for the next generation of wireless network is that it should cost-effectively provide guaranteed quality-of-service (QoS), especially in terms of delay and bit-rate requirement, with ubiquitous high bit-rate coverage, when and where required [2]. The challenge is the potentially high level of traffic heterogeneity over WTs with different channel state information (CSI) and over time, which did not previously exist in the context of voice and text traffic. This heterogeneity is becoming more pronounced with multimedia data traffic. Wireless networks are part of a highly complex heterogeneous interactive system, where consumers share limited radio resources for a broad range of services such as Voice-over-IP (VoIP), tele-medicine, online games, industrial & home automation, wearable connected devices, Hulu, Netflix, and Chrome OS. The flows for these vastly different services require highly different QoS. In addition, Traditionally, QoS in cellular communications has been implemented with *over-provisioning*, or through costly higher layer mechanisms and overheads. Over-provisioning results in a network design for its peak load which makes the system highly inefficient. In this setup, when the network becomes congested (load approaching the capacity), conventional rate limiters or bandwidth throttling is used which causes

user dissatisfaction [3, 4]. In addition, A WT may initiate a high-demanding application, such as high definition video-streaming, but then may simply cancel the request as its attention moves to some other item during web browsing. The combined actions of many WTs create a highly difficult situation to address randomness in traffic fluctuation. In fact, the way this randomness is addressed in the current wireline networks is by gross over provisioning of resources. Tomorrow's networks with more frequent congestion problems will not have the luxury of over-provisioning or using various forms of excessive overhead. Advanced access technologies, such as the long term evolution (LTE) network, are purely scheduled system based on orthogonal frequency division multiple access (OFDMA) which creates the opportunity to dynamically and efficiently exploit various types of diversity and to schedule for diverse requirements, instead of over-provisioning. The main question, then, is how to perform the resource allocation and packet scheduling to treat different flows with different requirements and different wireless links.

Need for Reengineering the Architecture of Data and Voice

Because of the pre-existent voice services and the gradual emersion of data traffic in the cellular networks, voice services are designed separately from packet-switched data connections in the pre-4G cellular networks, such as GSM, UMTS and CDMA2000 [5]. In these networks, only non-realtime (NRT) flows are subject to scheduling over the shared channel and voice services are often served in circuit-switching mode over the dedicated channel (DCH) [5]. This static separation sacrifices multiuser diversity. Today's wireless networks, such as LTE and the upcoming LTE-A, are moving towards packet-switching and IP-flat architecture to serve a broad range of applications with many diverse requirements [6]. Designing the flexible resource allocation and packet scheduling framework which can manage heterogeneity of the traffic across time and among WTs is crucial. In fact, providing properly engineered differentiable data flows is more cost-effective than providing voice and data in separation. In the flat architecture, voice will be one of the differentiated packet-switched data flows that the packet scheduling and RB allocation algorithm is responsible to guarantee its QoS requirements. With voice as a realtime (RT) differentiated data flow, networks will have reduced access delay, shorter wake-from-idle time and be able to offer not only

regular voice calls but also differentiated higher quality (audio or video) calls. There are three main ways of implementation for the new IP-flat architecture, in LTE, in regards to voice services, namely, voice-over-LTE (VoLTE), circuit-switched fall back (CSFB), and simultaneous voice and LTE (SVLTE) [7]. Among these three, VoLTE is the only one that really allows the delivery of voice as a data flow, within the LTE data bearer. CSFB and SVLTE are still dependent on the old pre-4G architecture and fall back to the legacy 2G or 3G circuit-switching in voice calls. CSFB and SVLTE have several inefficiencies in regards to VoLTE, such as longer call access delay, more expensive handsets & access points, and high power consumption on handset [7]. In regards to VoLTE, the joint resource allocation of RT and NRT flows is becoming more important in improving the efficiency of the cellular networks. Note that, although there are some partial solutions for the integration of RT and NRT flows in the application layer such as WebRTC [8], the packet and resource block (RB) scheduler in the centre of medium access control (MAC) layer is the main component to be designed efficiently for guaranteeing differentiated QoS.

Diverse New Applications: Need for Harmonization

Future wireless communication is a part of highly complex and differentiated services, such as digital distribution platforms (for audio, video, books, games, magazines, e.g. Google Play store) or emerging machine to machine (M2M) communication (e.g. auto sync always connected services). In contrast to voice-only circuit switched communication, the new evolved packet switched communication requires to consider several QoS elements, in addition to bit-rate information. In addition, efficient utilizing of radio resource, for making the most revenue, is one of the centre of interests in 5-th generation (5G) cellular networks, rather than a new physical (PHY) layer modification. Therefore, advanced schedulers should takes the operator's aspects (such as evolved fairness concepts, traffic aspects, user priority, and dynamic adjustment of system parameters through feedbacks in radio resource management (RRM) decision making (See Figure 2.1 in the next chapter). The novel framework should care about managing a broad range of QoS elements directing into the objective as the first thoughts, not afterthoughts, to reduce the stress on transmission control protocol (TCP). Therefore, a harmonized structure to position legacy packet scheduling and resource

scheduling concepts in perspective is necessary.

Need for Cost-effective RAN Architectures

Current state-of-the-art wireless standardization activities are leading towards high bit-rates in the order of one gigabit per second in the downlink with a fair coverage. While it is still early for the standardization bodies to consider much higher bit-rates, this is clearly a timely and important research topic due to the exponential growth of WT traffic on existing networks. Since wireless channel impairments and transmit power limitations prevent high spectral efficiency even for moderately long links, it is necessary to consider advanced cost-effective RANs, such as relay networks, empowered with fair efficient RRM techniques, which effectively collect and distribute wireless signals. Relay deployment opens the potential space to enhance the cell edge performance. To achieve the full potential of the advanced RANs with fairness, efficient packet scheduling and RB allocation techniques are also necessary to match the demand with the limited wireless resources.

Amplify-and-forward Relays as a Cost-effective RAN Element

OFDMA-based AF relays buffer quantized samples of the symbols until they are amplified and transmitted at a later time. These relays are cost-effective, simpler to implement, and introduce less delay in comparison to the decode-and-forward (DF) relay based routers. As evident in today's networks, implementing hop-by-hop routing is challenging at high bit-rates due to the hardware complexities of fast packet header inspection. AF relaying eliminates these issues from the very high bit-rate wireless networks. AF relays forward data without examining network layer headers, and is possible due to the synchronicity of OFDMA systems. In addition, since the AF relays do not decode the packets, channel decoder delays are eliminated, reducing its impact on higher layers. Therefore, AF relays are good candidates for enhancing the coverage in the next generation of wireless network. Providing a fair RRM framework for this kind of relay is crucial. Next generation of wireless networks aim at providing ubiquitous very high bit-rate coverage. Traditional throughput maximization fails to provide fairness because it results in scheduling starvation. Therefore, packet scheduling and RB allocation algorithm to exchange the fairness and total throughput needs to be

developed for OFDMA-based AF relays.

General Client & Server Model

It is worth mentioning that some of the concepts in this thesis can be adopted in a general clients & server scenarios, in addition to communication servers, including social and economics models, such as the ones in [9, 10]. Concepts can be used for example in cloud file storage and cloud processing, where basically the queue & server model is the same, but the meaning of RB and packet should be changed.

1.2 Organization and Contributions

- **Chapter 2: Packet Scheduling and Resource Block Allocation in Wireless OFDMA networks: A Block Utility-based Classification**

  In **Chapter 2**, we make a novel classification and intuitive decomposition of the packet scheduling and RB allocation algorithms for NRT flows and RT flows.

  The contributions, in this chapter, are categorized into following main parts:

  1. A classification of the packet scheduling and RB allocation algorithm, by means of a harmonized core.
  2. Novel delay fairness through weighted generalized proportional fairness (WGPF) for OFDMA networks: This novel delay fairness framework extends the conventional fairness notions in order to accommodate heterogeneity of traffic in time and among flows which is an important emerging problem. Our framework with a special disutility function is adjustable between two extreme objectives: minimizing the sum mean delay and minimizing the maximum mean delay. Delay fairness is an example of the evolved QoS-fairness notions. QoS-fairness is a generalized notion of fairness in comparison to having fairness on only bit-rate quantities.
  3. Decomposition of the structure of the core of various packet scheduling and RB allocation algorithms, based on their properties. The decomposition and classification viewpoint

enable to introduce several novel core for packet scheduling and RB allocation.

4. Propositions about the structure of the bit-rate proportional fairness (GPF), bit-rate maxmin, and delay fairness through WGPF.

The publications based on this chapter include [11–13].

- **Chapter 3: Joint Realtime and Non-realtime Flows Packet Scheduling and Resource Block Allocation in Wireless OFDMA networks**

  In **Chapter 3**, our novel joint RT and NRT packet scheduling and RB allocation will be proposed.

  The main contributions, in this chapter, can be categorized into two main items:

  1. We propose a novel joint resource allocation for RT and NRT flows based on head-of-the-line (HOL) delay, queue length information, and bit-rate information besides the embedded channel information in the dis-utilities. The proposed approach responds to heterogenous delay requirements for RT flows and manages NRT flows effectively within a common pool of RBs, rather than the sequential resource allocation of RT and NRT flows. Furthermore, the developed framework enables putting different algorithms in the literature in perspective and in a unified manner. Our approach is also a joint optimization in terms of both packet scheduling and RB allocation in a single shot.
  2. We developed a novel model for input-output bit-rate behaviour in packet scheduling and resource allocation of the mixture of RT and NRT flows. This model sheds light to identifying different load regions, explaining them, and understanding of the system in a simple and intuitive manner.

  The publication and the patent based on this chapter include [14, 15].

- **Chapter 4: Fair Scheduling in Multiuser Amplify-and-forward Relays**

  In **Chapter 4**, we design novel fair packet scheduling and RB allocation for OFDMA based AF relays.

  This chapter presents three main contributions to the multiuser AF relay scheduling:

1. We devise a novel scheduling framework for AF relay. Flows, from different WTs, over two-hops, are assigned bit-rate utility functions. According to the value of the embedded parameter, the utilities are able to gradually change resource allocation from throughput optimal, to proportionally fair, and to maxmin fair.

2. Since finding schedules and allocations are computationally hard, we propose two novel algorithms to quickly find schedules in each frame. The first algorithm is based on the gradient of the utility functions, so it is similar to the proportionally fair [16] scheduling algorithm, which was proposed for conventional cellular time division multiple access (TDMA), and OFDMA networks. However, unlike [16], which finds *long-term* fair bit-rates, our algorithm finds *short-term* fair bit-rates in each frame. The other algorithm is based on our observation that as the embedded parameter becomes large, the steepest gradient corresponds to the flow with the minimum bit-rate. In our simulations, we have observed that by changing the value of α, the algorithms achieve something similar to cell-breathing. In effect, instead of using power control to change the size of the cell, our RRM technique achieves it by combining adaptive modulation and coding (AMC) with time and channel allocation.

3. Third, we develop the efficient implementation of the above-mentioned algorithms by exploiting the super-modularity structure of the AMC table, in a AF relay system.

The publications and the patent based on this chapter include [17–19].

- **Chapter 5: Future Directions**

  Finally, in **Chapter 5**, we discuss a number of interesting items as future works, categorized in five dimensions.

Chapter 2

Packet Scheduling and Resource Allocation in Wireless OFDMA Networks: A Utility-based Classification

Abstract

This chapter presents a classification of a broad range of packet scheduling and resource block (RB) allocation algorithms, based on the main objectives in scheduling and allocation algorithms (of packet-switched wireless networks), namely, transmission efficiency, fairness, and quality-of-service (QoS)-requirements of bit-rate and delay. The chapter surveys the conceptual evolution of different packet scheduling and RB allocation regimes, with a comprehensive list of references, based on bit-rate utility functions and delay disutility functions appropriate for non-realtime (NRT) and realtime (RT) flows, respectively. The chapter is written with the motivation to serve as a through guide on how to design packet scheduling and RB allocation algorithms for the next-generation (beyond-4G) wireless orthogonal frequency division multiple access (OFDMA) networks. The chapter starts with a general discussion on the conceptual evolution of packet scheduling and

RB allocation algorithms, with respect to the gradient-based assignment which describes the core of the algorithm that determines iteratively which flow should be served with which RB.

The chapter consists of five main parts. In the first part, the motivation, the conceptual evolution of packet scheduling and RB allocation, and the general utility-based scheduling and allocation approach are discussed. In the second part, the system model is presented. In the third part, packet scheduling and RB allocation algorithms for NRT flows are discussed, namely, the generalized proportional fairness (GPF), bit-rate maxmin fairness, the Jain's index maximization of bit-rate, and the variants of minimum bit-rate guarantee. In the forth part, packet scheduling and RB allocation for RT flows are studied, namely, the weighted GPF for delay fairness, delay minmax fairness, and the variants of maximum delay guarantee. In the last part, the chapter is concluded with a pair of streamlined classification tables, based on the core of the algorithms. In addition, the chapter provides insights into the basic components of the core of the packet scheduling and RB allocation structures and highlights some common performance trends of various types of algorithms. It is our hope that this chapter, especially the discussion on the classification and decomposition of various algorithms, instigates research in a number of novel directions in packet scheduling and RB allocation in wireless networks.

2.1 Introduction: The Conceptual Evolution of the Core of the Packet Scheduling and RB Allocation Algorithms

MAC versus PHY

The research focus in wireless communications used to be on the enhancement of the physical (PHY) layer metrics, such as bit error rate (BER), in point-to-point links based on techniques mitigating, the noise, the interference, and the channel impairments. However, today's challenges lie in reliable and efficient end-to-end radio resource management (RRM) designs with quality-of-service (QoS) provisioning for heterogeneous services, especially over wireless networks. With the emergence of diverse flow classes and QoS-requirements, the emphasis in design has been shifting from PHY layer to the MAC and network layers. In other words, the decades long advancements

in the PHY layer have moved the bottleneck to the higher layers in wireless networks.

Wireline versus Wireless

Scheduling and allocation designs for mobile wireless channels are fundamentally different than the ones for less hostile mediums, such as wireline channels. Limited precious bandwidth, channel fluctuations (large-scale and small-scale signal variations), potent multiuser diversity, and burst errors (which depend on the WT locations and interference levels) make the packet scheduling and RB allocation design of the wireless environment more challenging than that of wireline networks. Channel fluctuation of the wireless networks makes the compensation of WT in poor signal location or high velocity, with granting more resource block (RB), inevitable. One conventional approach is the algorithm design based on the layered network architecture and simply to model the wireless channel as being a binary on and off. However, this restrictive view is inefficient in comparison to a fully channel-aware design. The scheduler design for the wireless channels not only should address the heterogeneity of the demands (QoS requirements and fairness), but also should address the heterogeneity of the supply (channel capacities). In addition, a cross-layer paradigm shift is beginning to take place in designing the packet scheduling and RB allocation algorithm, based on embedding the QoS and fairness requirements, right in the algorithm design. This is in contrast to having a rather naïve algorithm and handling those requirements through more costly higher layer mechanisms. This paradigm shift makes wireless networks more efficient, with the matching of the supply (RBs capacities) with demand (QoS-requirements) in the packet-switched networks.

Need for Channel-awareness

Primitive packet scheduling and RB allocation algorithms were designed with neither channel-awareness nor queue-awareness, such as round-robin. Then, channel-aware algorithms were proposed to exploit the multiuser diversity in wireless networks. In fact, multiuser diversity can turn the challenges of multi-path fading into an opportunity. The opportunity comes from the fact that the average capacity of the fading channels monotonically increase and exceed that of the deterministic one, as the number of the WT increases [20]. Intuitively, as the number of WT increases, the

probability of all of the WTs being in deep fade decreases sharply. In other words, the probability of having one or more WTs near their peak adaptive modulation and coding (AMC) modes gets close to one, as the number of WTs increase. From the scheduling point of view, this multiuser diversity opportunity (independent fading process over different WTs) can be exploited by channel-aware designs.

From another angle, the initial design of wireless networks was based on the legacy dedicated circuit-switched designs especially for voice. The circuit-switched mode is inefficient not only for data traffic with bursty nature, but also for voice traffic. The round-robin mechanism can be considered as the simplest packet scheduling algorithm, in transition from circuit-switching to packet-switching, where flows are periodically allocated RBs irrespective of the their backlog and their channel conditions. Channel-aware designs can avoid bursty errors, with deferring the transmission of packets with bad channel condition in the worst case scenario, at the link layer instead of relying on the higher layers for costly error recovery. For a simple example on the inefficiency of a non-channel-aware design, see [21] where it is shown that the efficiency of even a very simple channel-aware design can be as high as twice of that of the round-robin type, for only three WTs. As the number of WTs increases, the gain of the multiuser diversity increases further.

Need for QoS-requirements

The next important consideration in scheduling and allocation design is the queue awareness, or more generally QoS aware designs. Scheduling and allocation algorithms, such as GPF, are ignorant to the queue length related information, therefore, they could not guarantee queue stability of RT packet-switched flows. The more advanced algorithms exploit the fluctuations in both channels and queue lengths, in order to maximize the efficiency. More generally, advance designs should guarantee specific QoS such as minimum bit-rate and maximum delay.

Sub-optimality of the Independent Packet Scheduling and RB allocation

Many of the existing packet scheduling and resource allocation designs, either perform the RB scheduling without considering QoS-requirements (ignoring the QoS attributes of the bit-pipe they

are building) or perform packet scheduling independent of the RB allocation. Combining the independently designed RB allocation and packet scheduling algorithm results in a suboptimal system design. To avoid this sub-optimality, a joint approach for packet scheduling and RB allocation (a single-shot decision making) should be adopted.

General System Block Diagram

To better present the general components of a system where a packet scheduling and RB allocation algorithm works within, Figure 2.1 is provided to depict an input and output system view. The

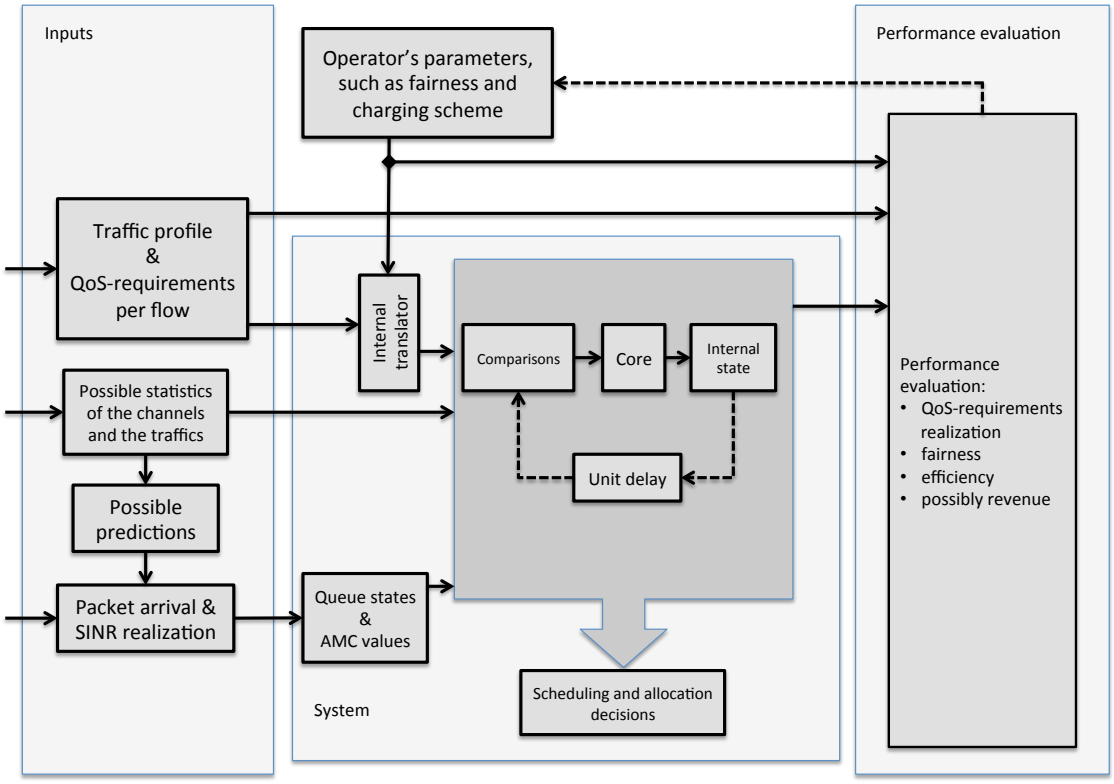

Figure 2.1: General system view block diagram of the packet scheduling and RB allocation algorithm.

leftmost blocks define the input to the system. These blocks consider the heterogeneity coming

from different QoS demands and different channel conditions, such as varying different arrival rates and wireless channel impairment effects. The system, which is depicted by the middle box, works based on system state which includes the queue length information (which are controllable) and the CSI (which are observable). The parameter translator, in the system, translates parameters from outside (operator's parameters in the top of the middle box) to the algorithm-specific parameter set such as utility curves and the comparison functions. The rightmost blocks evaluate the results of the RRM algorithm based on the input demands and algorithm outputs.

Top-down Utility-based Optimization

Utility-based packet scheduling and RB allocation algorithm has been inspired by an economic concept. The general methodology is based on a top-down optimize-able packet scheduling and RB allocation algorithm design. This methodology provides a way to design protocols, which support a wide range of RRM algorithms, ensuring that RRM is not restricted by radio access network (RAN) protocols. Protocol design starts with a global utility maximization problem. The objective function of the optimization is chosen so that, at the optimum (equilibrium) point, the QoS elements satisfy some criterion, specified by the network operator. Combining the flow and WT centric RRM with the top-down optimize-able RAN protocol design approach provides promising approach for advanced RRM in beyond-4G networks.

Without any loss of generality, we suppose only one flow per WT. However, to accommodate several flows per WT, scheduling can be implemented either on a per flow basis or on a per WT basis. The first option can be interpreted with multiple flows as multiple virtual WT. The second option can be interpreted when WTs request aggregate services from the base station (BS) and then distribute the aggregate resource assignments among its own multiple flows. In this case, the aggregate flows of a WT plays as a virtual flow in our framework.

Harmonized Packet Scheduling and Resource Block Allocation Core

Utility-based scheduling algorithms have a good feature of possibility to present different algorithms based on a defining core. We are motivated to simplify the presentation of different scheduling

algorithms based on a harmonized defining core. In general for different utility-based algorithms, at the core of the algorithms, there is a calculation of a priority for different traffic flow based on finding maximum element in a list as

$$\Big(\phi^*[k], j^*[k]\Big) = \arg\max_{(\phi,j)} \frac{\partial S_{\text{net}}^{\text{QoS}}(.)}{\partial x_{\phi}^{(j)}[k]}, \tag{2.1}$$

where $S_{\text{net}}^{\text{QoS}}(.)$ is the summation of individual utility function, $S_{\phi}^{\text{QoS}}(.)$, the optimization variable $x_{\phi}^{(j)}[k]$ is the scheduling variable (describes how many RB in sub-channel j is allocated to flow ϕ in frame k), and $\Big(\phi^*[k], j^*[k]\Big)$ is the selected flow $\phi^*[k]$ to be served by a RB on the sub-channel $j^*[k]$. The argument of the utility function depends on the design. It can be as simple as the flow bit-rate or in general a sophisticated version of QoS elements. We elaborate extensively on the argument of the utility function, in this chapter. We use the superscript of the utility function to show different special version of its argument. In the most general form of the utility function, we use the QoS as its superscript (which indicate its general argument). We use the subscript of the utility function to show whether it is referring to individual flow utility or the total sum utility.

The scheduling core specifies which flow and which RB should be selected for the allocation. However, it should be noted that when more than one RB is available in a frame (such as the case in OFDMA), the core will be executed in an iterative manner after the allocation of each RB to a flow during a *single frame*. The fine granularity of OFDMA makes possible these updates in a single frame. In frame updates makes the utility-based scheduling algorithms to work closer to the optimum point despite their low complexity.

General Scheduling and Allocation Core Block Diagram

Figure 2.2 shows more details around the scheduling core of the system, where the packets from certain flows are assigned to RB with appropriate AMC. The decision maker uses the state of system (queue length information & channel information), QoS-requirements, and other information to control the OFDMA server.

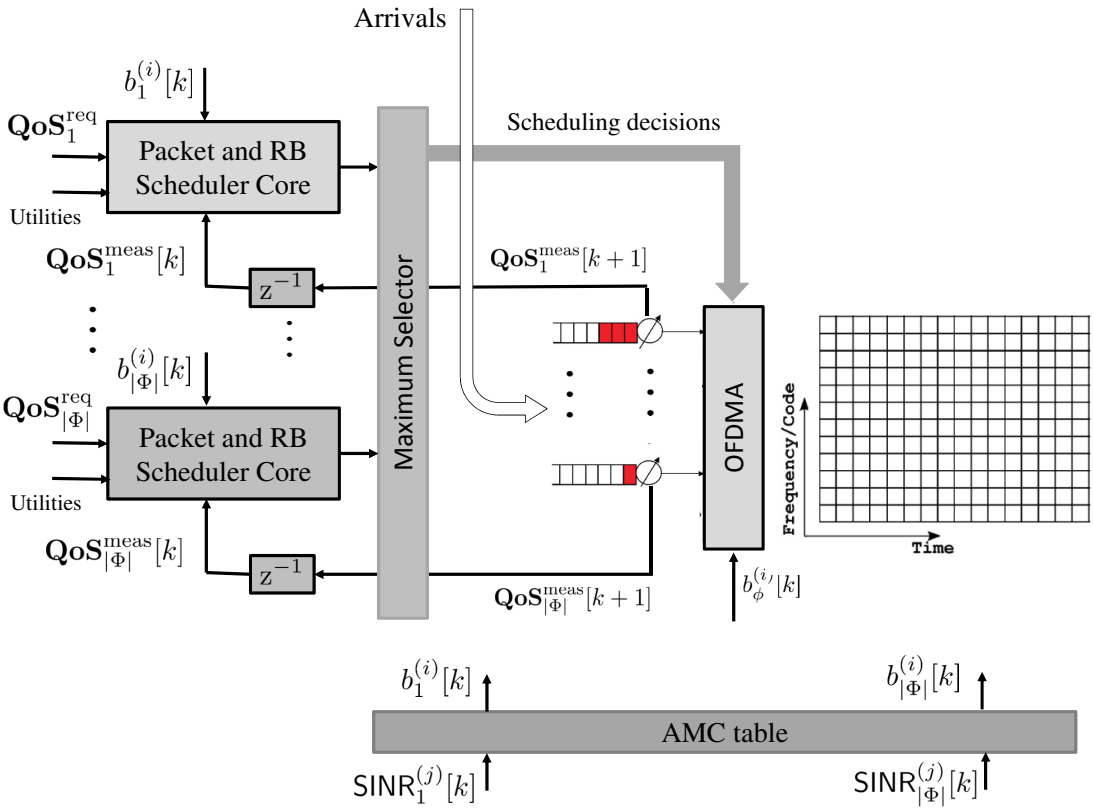

Figure 2.2: Block diagram of the data plane of the packet scheduling and RB allocation algorithm.

Basic Model Assumption for the Inputs

Different assumption can be made for the two main input to the system, channels and traffic. Mainly, four possible models for the arrival model and channel model can be assumed in formulating the scheduling problem, namely, finite-backlog with adversary channel model, infinite-backlog with adversary channel model, infinite-backlog with stationary channel, and finite-backlog with stationary channels [22]. When there is no information about dynamics of the system input, including CSI and traffic, robust approach [23, 24] and online learning [25–27] are recommended. The on-line learning algorithm learns the statistics of channel and arrivals, the two random natures of the problem. The on-line learning algorithm can be used in predictive (See the third leftmost block

in Figure 2.1) packet scheduling and RB allocation. The predictive makes packet scheduling and RB allocation decisions based on information from time horizon a head of the current frame index. When multi-hops is considered an algorithm called back-pressure comes into picture [23, 28]. In practice, however, adversary channel models are generally too complicated and can be avoided.

The infinite-backlog assumption can be used for simplification and only for NRT flows. In the infinite-backlog model, it is assumed that each flow always has data to transmit (full queue). Nevertheless, for RT flows, the finite-backlog should be considered. We use the finite-backlog and stationary channels in order to consider both RT and NRT flows.

Rate-revenue Decoupling and Differentiated Services

One of the main current challenge and opportunity facing the wireless industry is the expected rapid growth tied with high diverse QoS expectation in the cellular sector. However, because of rate-revenue decoupling, business model for wireless industry should be re-engineered. It is anticipated that real differentiated services in the MAC layer opens up opportunities for innovative timely business model to address diverse QoS for new data plans, Internet-of-things (IoT), and M2M communications. As an upcoming example Amazon prime free data plan is expected to be bundled with the phone, suggesting that customers will have discounted or even free data deals to access to Amazon store. The novel business models to address rate-revenue decoupling cannot be implemented efficiently without a harmonized QoS modelling.

Chapter Contributions

The contributions in this chapter are categorized into following main parts:

1. A classification of the packet scheduling and RB allocation algorithm, through a harmonized core (the core specifies which flow should be served with which RB), based on the utility of the algorithm.

2. Novel delay fairness through weighted generalized proportional fairness (WGPF) for OFDMA networks: This novel delay fairness framework extends the conventional fairness notions in order to accommodate heterogeneity of traffic in time and among flows which is an important

emerging problem. Our framework with a special disutility function is adjustable between two extreme objectives: minimizing the sum mean delay and minimizing the maximum mean delay. Delay fairness is an example of the evolved QoS-fairness notions. QoS-fairness is a generalized notion of fairness in comparison to having fairness on only bit-rate quantities.

3. Decomposition of the structure of the core of different packet scheduling and RB allocation algorithms, based on their properties. The decomposition and classification viewpoint enable to introduce several novel core for packet scheduling and RB allocation.

4. Propositions about the structure of the bit-rate generalized proportional fairness (GPF), bit-rate maxmin, and delay fairness through WGPF.

Chapter Organization

This chapter consists of five main parts as in the following list.

1. **Introduction** (Section 2.1), including motivation, general evolution of the packet scheduling and RB allocation, organization, and symbol list.

2. **System model** (Section 2.2), including the OFDMA frame, basic bit-rate quantities, QoS modelling, and the required CSI feedback and scheduling announcement feedforward assumptions.

3. **Scheduling and allocation for NRT flows** (Section 2.3), including bit-rate GPF, bit-rate maxmin, the Jain's index maximization of bit-rate, the Jain's index maximization with general utility of bit-rate, multi-objective approach for efficiency and fairness, minimum bit-rate guarantee with virtual token, minimum bit-rate guarantee with the Lagrangian multipliers, and minimum bit-rate guarantee with the modified largest weighted delay first (MLWDF).

4. **Scheduling and allocation for RT flows** (Section 2.4), including mean-delay fairness, mean-delay minmax fairness, the Jain's index maximization for delay, maximum mean-delay guarantee with the Lagrangian multipliers, discussion on the mean-delay versus head-of-the-line (HOL) delay, Lagrangian multiplier assisted admission controller, earliest deadline first,

modified largest weighted delay first, and exponential-rule.

5. **Discussions** (Section 2.5), including classification, structural decomposition, notes on performance comparison of scheduling cores, and the concluding remarks.

List of Symbols

In this section, we summarize the notations used throughout this chapter with a short definition of them in Table 2.1

Table 2.1: List of symbols used in the classification.

Symbol	**Definition**
$\left(\phi^*[k], j^*[k]\right) = \arg\max_{(\phi,j)} \frac{\partial S_{\text{net}}^{\text{QoS}}(.)}{\partial x_\phi^{(j)}[k]}$	General utility based scheduling core
$\phi^*[k]$ & $j^*[k]$	Selected flow $\phi^*[k]$ to be served by an RB on sub-channel $j^*[k]$
Φ_{RT}	Set of RT flows
Φ_{NRT}	Set of NRT flows
N	Number of frequency sub-channels
T	Number of time slots per frequency sub-channel
T_b	Time span of each RB in second
W_b	Frequency span of each RB in Hertz
ϕ	Flow index
j	Sub-channel index
k	Frame index
$\mathfrak{f}\left(\mathsf{SINR}_\phi^{(j)}[k]\right)$	Function describing the AMC table
$b_\phi^{(j)}[k]$	AMC value of the an RB on sub-channel j, for flow ϕ, in frame k
$\mathsf{SINR}_\phi^{(j)}[k]$	SINR on sub-channel j, for flow ϕ, in frame k

$x_{\phi}^{(j)}[k]$	Scheduling variable for sub-channel j, flow ϕ, in frame k
$\mathsf{BER}\left(\mathsf{SINR}_{\phi}^{(j)}[k], b_{\phi}^{(j)}[k]\right)$	BER as a function of the SINR and AMC value (selected PHY mode)
$c_{\mathrm{MOD}_{\phi}^{(\mathrm{j})}}$ & $c_{\mathrm{FEC}_{\phi}^{(\mathrm{j})}}$	modulation FEC constants, for flow ϕ & sub-channel j
$\hbar[k]$	Forgetting factor
$\mathbf{QoS}_{\phi}^{\mathrm{req}}$	QoS-requirement vector, for flow ϕ
$\mathbf{QoS}_{\phi}^{\mathrm{meas}}[k]$	QoS-measurements vector, for flow ϕ, in frame k
$\mathbf{QoS}_{\phi}^{\mathrm{class}}$	QoS-requirement class vector, for flow ϕ
$S_{\phi}^{\mathrm{QoS}}(.)$	General utility function for flow ϕ
$S_{\mathrm{net}}^{\mathrm{QoS}}(.)$	General network utility function
$d_{\phi}^{\mathrm{HOL}^{\max}}$	Delay deadline of HOL-delay, for flow ϕ
$\bar{d}_{\phi}^{\max}$	Maximum mean-delay, for flow ϕ
$r_{\phi}^{\min}$	Minimum frame bit-rate, for flow ϕ
$\bar{r}_{\phi}^{\min}$	Minimum mean bit-rate, for flow ϕ
$\epsilon_{\phi}^{\max}$	Maximum PLR, for flow ϕ
$\bar{d}_{\phi}[k]$	Mean-delay, until frame k
$d_{\phi}^{\mathrm{HOL}}[k]$	Mean-delay, until frame k
$\bar{r}_{\phi}[k]$	Mean bit-rate of flow ϕ, until frame k
$r_{\phi}[k]$	Frame bit-rate, for flow ϕ, in frame k
$\epsilon_{\phi}[k]$	Measured PLR, for flow ϕ, in frame k
$\theta_{\phi} \in \Theta$	Flow type of flow ϕ in the ensemble of flow types Θ
χ_{ϕ}	Colour level of flow ϕ
$\Phi_{\theta_{\phi}}$	Class ensemble of flow ϕ, such as Φ_{RT} and Φ_{NRT}
$\bar{\mathbf{r}}[k]$ & $\bar{\mathbf{d}}[k]$	mean bit-rate vector and mean delay vector
amc_m	AMC mode for WT m

M	Number of WT
$S_{\phi}^{\bar{\mathrm{r}}}\left(\bar{r}_{\phi}[k]\right)$	Mean bit-rate utility function
$S_{\mathrm{net}}^{\bar{\mathrm{r}}}\left(\bar{\mathbf{r}}[k]\right)$	Network mean delay utility function
α & $w_{\phi}^{\mathrm{GPF}}[k]$	GPF parameter and WGPF weights
$\mathcal{C}^{\mathrm{PHY}}$	PHY layer constraints
$\mathbf{w}^{\mathrm{off}}[k]$	Weights vector corresponding to the offline version of PF
$v_{\phi}[k]$	The general quantity to be distributed in Jain's index maximization
$\mathbf{v}_{\phi}[k]$	The vector version of the quantity to be distributed in Jain's index maximization
$\eta(\mathbf{v}[k])$	The efficiency in Jain's index maximization
$J(\mathbf{v}[k])$	Jain's index of the vector $\mathbf{v}[k]$
$\mathbf{1}_{\lvert\Phi\rvert}$	Vector of ones
β	Fairness parameter in Jain's index maximization
$S_{\mathrm{net}}^{\bar{\mathrm{r}}}\left(\bar{\mathbf{r}}[k]\right)$ & $\Pi_{\mathrm{net}}^{\bar{\mathrm{r}}}\left(\bar{\mathbf{r}}[k]\right)$	Network efficiency and network discrimination, in multi-objective approach
$\mathcal{C}^{\bar{\mathrm{r}}^{\min}}, \mathcal{C}^{\bar{\mathrm{r}}^{\max}}$	Constraints enforcing the minimum bit-rates and maximum bit-rates requirements.
$\lambda_{\phi}^{\mathrm{token}}[k]$	Token counter in minimum bit-rate guarantee with virtual tokens, for flow ϕ
$\lambda_{\phi}^{\bar{\mathrm{r}}^{\min}}[k]$	Lagrangian multiplier associated with minimum mean bit-rate guarantee
$\bar{q}_{\phi}[k]$	Mean queue length of flow ϕ, until frame k
$q_{\phi}[k]$	Queue length of flow ϕ, in frame k
$S_{\mathrm{net}}^{(\bar{\mathbf{q}},\bar{\mathbf{r}})}\left(\bar{\mathbf{q}}[k],\bar{\mathbf{r}}[k]\right)$	Utility function, based mean bit-rate and mean queue length, used in mean delay fairness

$\nu^{\bar{\mathrm{d}}}$	Mean delay fairness parameter
$\Lambda_\phi[k]$	Actual arrival bit-rate of flow ϕ
$\hat{\Lambda}_\phi[k]$	Estimate of arrival bit-rate of flow ϕ, until frame k
$A_\phi[k]$	Arrival bits in frame k in bits
$D_\phi^{\bar{\mathrm{d}}}(\bar{d}_\phi[k])$	Mean delay disutility, for flow ϕ
$D_{\mathrm{net}}^{\bar{\mathrm{d}}}\left(\bar{\mathbf{d}}[k]\right)$	Mean delay network disutility
$\tilde{d}_\phi[k]$	Mean delay estimate in frame k
$\zeta_\phi^{(1)}[k], \zeta_\phi^{(2)}[k], \zeta_\phi^{(3)}[k], \& \ \zeta_\phi^{(4)}[k]$	Parameters in the estimation of the of $\tilde{d}_\phi[k]$
$t_\phi^{\mathrm{HOL}}[k]$	Arrival time stamp of the HOL packet, for flow ϕ, in frame k
$t_\phi^{(l)}[k]$	Timestamp of the l-th bits ($l \in [1, q_\phi[k]]$) in queue of flow ϕ
$d_\phi^{(l)}[k]$	Delay of the he l-bits in the queue for flows ϕ
$D_\phi^{\mathrm{d^{HOL}}}\left(d_\phi^{\mathrm{HOL}}[k]\right)$	HOL-delay disutility
$\nu^{\mathrm{d^{HOL}}}$	HOL-delay fairness parameter
$\beta^{\bar{\mathrm{d}}}$	Mean delay fairness parameter, through Jain' index maximization
$\beta^{\mathrm{D}^{\bar{\mathrm{d}}}}$	Mean delay fairness with general disutility, through Jain' index maximization
$\mathcal{C}^{\bar{\mathrm{d}}^{\max}}$	Maximum mean delay constraints
$\lambda_\phi^{\bar{\mathrm{d}}^{\max}}[k]$	Lagrangian multiplier associated with maximum mean delay constraints
$\partial D_\phi^{(.)}(.)/\partial x_\phi^{(j)}[k]$	General disutility gradient, for flow ϕ, with respect to its scheduling variable on sub-channel j and frame k
μ_ϕ	EXP and MLWDF parameter, for flow ϕ
δ_ϕ	Maximum probability of exceeding the HOL-delay in MLWDF, for flow ϕ

τ_ϕ	Maximum delay threshold in MLWDF, for flow ϕ, interpretable as $d_\phi^{\mathrm{HOL}^{\max}}$
$\overline{d^{\mathrm{HOL}}}[k]$	Average of HOL-delay over flows, in frame k
η	EXP parameter controlling the delay fairness
Φ_{BE}	Set of BE flows, as an NRT example
Φ_{VoIP}	Set of VoIP flows, as an RT example
$\nu^{\bar{\mathrm{d}}}$	Mean-delay fairness parameter, controlling the mean-delay equalization
$\mathfrak{O}^{\mathrm{d-fair}}(\cdot)$	Conceptual delay fairness component in the scheduling core
$\mathfrak{O}^{\mathrm{r-fair}}(\cdot)$	Conceptual bit-rate fairness component in the scheduling core
$\mathfrak{O}^{\mathrm{PLR}}(\cdot)$	Conceptual PLR control component in the scheduling core
$\mathfrak{O}^{\mathrm{eff}}(\cdot)$	Conceptual transmission efficiency component in the scheduling core
$\mathfrak{O}^{\mathrm{r-min}}(\cdot)$	Conceptual min bit-rate guarantee component in the scheduling core
$\mathfrak{O}^{\mathrm{d-max}}(\cdot)$	Conceptual max delay guarantee component in the scheduling core
ϱ	Multi-objective scalar-ization parameter

In the following sections, after the explanation of system model, scheduling algorithms for NRT flows, as well as RT flows will be discussed.

2.2 System Model

A downlink scenario with an OFDMA air interface, which serves RT flows in set Φ_{RT} and NRT flows in set Φ_{NRT} in a single cell, is considered here. The framework can be generalized to consist higher number of classes, in addition to RT and NRT classes. A flow, ϕ, is a packet-switched connection from layer-three between the BS and a WT. Following a handoff, packets for the flows queued at the first BS will be transferred to the possibly reserved resources in the second BS, with adjusted time-stamps of the transferred packets. A packet is a fixed size of bits. A WT may have several flows. The WTs use a variety of RT and NRT applications which are delay intolerant and delay tolerant, respectively. We assume that radio resources (frequency and power) are allocated to base-stations a priori, in a way that coordinates the inter cell interference. The framework can be extended to the uplink if a central resource allocation for uplink is possible.

In this chapter, we study work-conserving schedulers. A work-conserving scheduler may not be idle when there is backlogged packets in the queues. In contrast, a non-work-conserving scheduler can be idle, even when there is backlogged packets in the queues. Non-work-conserving schedulers are designed based on the reservation of a certain portion of the frame resources for the higher-priority packet, no matter they exist or not. Non-work-conserving schedulers may be relevant for applications where time jitter is the most important factor, otherwise they are not common in the scheduler design.

2.2.1 OFDMA

Why OFDMA

OFDMA has several technical advantages, namely, subcarrier adaptive bit loading capability, ISI reduction, multi-path fading mitigation, and efficient implementation through fast Fourier transform (FFT) [20,29]. In addition, in the MAC layer it offers a fine granularity which can be exploited for scheduling. This feature gives the opportunity to adapt the transmission across sub-channels, among WTs, and dynamically between frames [13]. To exploit this feature in packet scheduling, resource scheduling, and AMC control, we select the OFDMA interface as the access technology in

our study.

OFDMA Frame

The total bandwidth is divided into N sub-channels consisting of several OFDMA sub-carriers. Each sub-channel is further divided in time into T time-slots. In this way, the time-frequency plane, for each frame, is divided into NT RBs, each of which spans T_b seconds in time and W_b Hertz in frequency. For example, in the long term evolution (LTE) systems $T_b = 1$ ms and $W_b = 180$ kHz. Since we use OFDMA, the framework applies to frequency-selective fading, in addition to flat fading. It is worth mentioning that we extend the OFDMA plane framework to have more than one single RB on an specific sub-channel, within a frame, over time. This generalization gives the flexibility of including future technologies, where time-division within a frame is possible. When this flexibility is not possible, $T = 1$ reduces the model to the conventional OFDMA plane of LTE. Time-division within a frame, if possible, results in higher granularity and increases the efficiency in resource allocation algorithm. We note that in contrast to the code division multiple access (CDMA) in which the residual power is an efficiency problem or in the time division multiple access (TDMA) where only one WT can be served in a frame, different AMC modes (even with different BER requirements) can be used in OFDMA for simultaneously several WTs, over different sub-channels.

AMC Values

The transmission frames are indexed by notation k, sequentially. We use the frame to refer to frame index. In frame k, the highest available spectral efficiency and corresponding AMC level, for a single RB on sub-channel j for flow ϕ is

$$b_{\phi}^{(j)}[k] = \mathfrak{f}\left(\mathsf{SINR}_{\phi}^{(j)}[k]\right), \text{ in bits/Hz./sec.}, \tag{2.2}$$

where $\mathsf{SINR}_{\phi}^{(j)}[k]$ is the signal to noise and interference ratio (SINR) of RBs associated with flow ϕ on sub-channel j in frame k and $\mathfrak{f}(\cdot)$ represents the AMC table which depends on BER. The BER,

is a function of AMC transmission bit-rates and SINR and approximated as

$$\mathsf{BER}\left(\mathsf{SINR}_{\phi}^{(j)}[k], b_{\phi}^{(j)}[k]\right) = c_{\mathrm{MOD}_{\phi}^{(\mathrm{j})}} \exp\left(-\frac{c_{\mathrm{FEC}_{\phi}^{(\mathrm{j})}} \mathsf{SINR}_{\phi}^{(j)}[k]}{2^{b_{\phi}^{(j)}[k]} - 1}\right), \tag{2.3}$$

where $c_{\mathrm{MOD}_{\phi}^{(\mathrm{j})}}$ and $c_{\mathrm{FEC}_{\phi}^{(\mathrm{j})}}$ are modulation and forward error correction (FEC) constants, respectively, for flow ϕ & sub-channel j (See [30] for the details on adaptive modulation and coding and its BER estimation). Given a requested BER bound, the AMC modes are a function of different SINR levels which can be found by solving

$$b_{\phi}^{(j)}[k] = \max_{b_{\phi}^{(j)}[k]} \left\{ b_{\phi}^{(j)}[k] \middle| \mathsf{BER}\left(\mathsf{SINR}_{\phi}^{(j)}[k], b_{\phi}^{(j)}[k]\right) \leq \mathsf{BER}^{\mathrm{req}} \right\}, \tag{2.4}$$

where it equivalently finds the AMC table. Each WT transmits using AMC mode, consist of a pair of a modulation index and an FEC index, based on its SINR level.

2.2.2 Frame Bit-rate and Mean Bit-rate

Radio resources are assigned to the flows in terms of RBs; each RB carries data of only one flow at a time. The bit-rate of a flow is determined from the number of RBs it is allocated in the frame and the AMC level used in each RB as

$$r_{\phi}[k] = W_b \sum_{j=1}^{N} b_{\phi}^{(j)}[k]\, x_{\phi}^{(j)}[k], \tag{2.5}$$

in bits/sec, where $b_{\phi}^{(j)}[k]$ is the spectral efficiency of RBs on sub-channel j for flow ϕ in frame k, W_b is the frequency span of RB, and $x_{\phi}^{(j)}[k]$ is the number of RBs allocated to flow ϕ on sub-channel j in frame k, as defined earlier.

Likewise, the mean bit-rate will be defined based on

$$\bar{r}_{\phi}[k] = \Big(1 - \hbar[k]\Big)\bar{r}_{\phi}[k-1] + \hbar[k] r_{\phi}[k], \tag{2.6}$$

where $\hbar[k] \in (0, 1)$ is the forgetting factor.

Bold-face and regular-face fonts will be used to denote vectors and scalars, respectively. The set of length-n vectors with non-negative real entries will be denoted by $\mathbb{R}^n$ and the length-n all-one and all-zero vectors will be denoted by $\mathbf{1}_n$ and $\mathbf{0}_n$, respectively. The Euclidean norm will be denoted by L_p. It is worth mentioning that although our framework is OFDMA, the concepts can be carried over to any orthogonal channelization scheme.

2.2.3 QoS-requirements, -measurement, -class Vectors, and Utility Functions

QoS-requirements

A flow is identified by its flow index, $\phi \in \Phi$. The QoS-requirements of a flow are described by its QoS elements. We use five QoS elements which form the QoS-requirement vector as

$$\mathbf{QoS}_\phi^{\text{req}} \triangleq \left(d_\phi^{\text{HOL}^{\max}}, \bar{d}_\phi^{\max}, \epsilon_\phi^{\max}, r_\phi^{\min}, \bar{r}_\phi^{\min}\right), \tag{2.7}$$

where the following list defines the set points.

- The set point $d_\phi^{\text{HOL}^{\max}}$ is the maximum tolerable HOL-delay per flow in seconds.
- The set point $\bar{d}_\phi^{\max}$ is the maximum tolerable mean delay per flow in seconds. This item for an NRT flow can be limited.
- The set point $\epsilon_\phi^{\max}$ is equal to $\max \Pr\left(d_\phi^{\text{HOL}}[k] > d_\phi^{\text{HOL}^{\max}}\right)$, the maximum probability of packet loss (PL).
- The set point $r_\phi^{\min}$ is the frame required minimum bit-rate in bps.
- The set point $\bar{r}_\phi^{\min}$ is the required minimum mean bit-rate in bps.

Since we are considering RT and NRT flows together, the framework should be able to capture both mean-delay (suitable for NRT flows) and more HOL-delay (suitable for RT flows). The QoS-elements are chosen with considering services that needs guaranteed bit-rate within a certain time window, guaranteed hard delay thresholds, such as voice-over-IP (VoIP), and delay non sensitive services for which the experience could be enhanced if the mean-delay is shorten, such as FTP. This

design results in QoS-aware which is the useful version of the queue-awareness. Queue-awareness is not sufficient in the practice, as the decision maker should know the QoS-requirements and the utility characteristics to calculate the perceived QoS.

QoS-measurements

Similar to the QoS-requirement vector, the measurement counterparts of the QoS-requirement makes the $\mathbf{QoS}_{\phi}^{\text{meas}}[k]$ as

$$\mathbf{QoS}_{\phi}^{\text{meas}}[k] \triangleq \Big(d_{\phi}^{\text{HOL}}[k], \bar{d}_{\phi}[k], \epsilon_{\phi}[k], r_{\phi}[k], \bar{r}_{\phi}[k]\Big), \tag{2.8}$$

where the elements are the measurement counterparts of the (2.7). Vector $\mathbf{QoS}_{\phi}^{\text{req}}$ determines the set points while the $\mathbf{QoS}_{\phi}^{\text{meas}}$ determines the actual values. In other words, $d_{\phi}^{\text{HOL}}[k]$ is the measured (or estimated) delay of flow ϕ in frame k, $\bar{d}_{\phi}[k]$ is the measured mean-delay of flow ϕ in frame k, $\epsilon_{\phi}[k]$ is the measured PL in a reasonable window until frame k, $r_{\phi}[k]$ is the measured frame bit-rate, and $\bar{r}_{\phi}[k]$ is the measured mean bit-rate.

QoS-class

Along with QoS-requirements vector, the class of the flow ϕ makes the QoS class, $\mathbf{QoS}_{\phi}^{\text{class}}$, as

$$\mathbf{QoS}_{\phi}^{\text{class}} \triangleq (\theta_{\phi}, \chi_{\phi}), \tag{2.9}$$

where θ_{ϕ} is a unit-less quantity that describes the flow type, such as the two main categories of RT and NRT flows as $\theta_{\phi} \in \Theta = \{\text{RT}, \text{NRT}\}$. Set Θ is the ensemble of flow types. In general, sets Φ_{θ}s, for $\theta \in \Theta$, form a partition of the set of all flows, Φ, over the flow types. For example, for two classes of RT and NRT flows, $\Phi = \Phi_{\text{NRT}} \bigcup \Phi_{\text{RT}}$. The optional parameter χ_{ϕ} is also a unit-less quantity (named colour property) that describes the priority level of a flow among a set of same flow type and same QoS-requirements. As an example, $\chi_{\phi} \in \{\text{bronze}, \text{silver}, \text{gold}\}$. The colour property, χ_{ϕ}, of a flow adds priority by giving a larger QoS provisioning margin to the corresponding flows in order to ensure higher reliability. Two flows can have same flow type and QoS-requirements but

the one which pays more will be placed in higher level of colour priority level.

Utility Function

The optimization objective is defined as the combined function on the QoS-measurement vector and QoS-requirement vector as

$$S_{\phi}^{\mathbf{QoS}}\left(\mathbf{QoS}_{\phi}^{\mathrm{req}}, \mathbf{QoS}_{\phi}^{\mathrm{meas}}[k], \mathbf{QoS}_{\phi}^{\mathrm{class}}\right). \tag{2.10}$$

In general, the combination of QoS-requirement and QoS-measurements elements such as their difference or their division play role as argument. The summation of the individual utility makes the network utility as

$$S_{\mathrm{net}}^{\mathbf{QoS}}\left(\begin{bmatrix}\mathbf{QoS}_{\phi}^{\mathrm{req}}\\ \vdots \\ \mathbf{QoS}_{|\Phi|}^{\mathrm{req}}\end{bmatrix}, \begin{bmatrix}\mathbf{QoS}_{\phi}^{\mathrm{meas}}\\ \vdots \\ \mathbf{QoS}_{|\Phi|}^{\mathrm{meas}}\end{bmatrix}, \begin{bmatrix}\mathbf{QoS}_{\phi}^{\mathrm{class}}\\ \vdots \\ \mathbf{QoS}_{|\Phi|}^{\mathrm{class}}\end{bmatrix}\right) = \sum_{\phi=1}^{|\Phi|} S_{\phi}^{\mathbf{QoS}}\left(\mathbf{QoS}_{\phi}^{\mathrm{req}}, \mathbf{QoS}_{\phi}^{\mathrm{meas}}[k], \mathbf{QoS}_{\phi}^{\mathrm{class}}\right). \tag{2.11}$$

The relative effectiveness of each QoS elements can be controlled through parametrization of utility, where the parameters can be modified by WT behaviour, WT subscription class, and other operator dynamic requirements [15]. In general, harmonized QoS modelling enables the framework to define future (presently undefined) QoS classes and accommodate broader soft differentiated services including a continuum between pure RT and pure NRT connections. For example, a class of NRT flows practically can have a delay deadline on the mean delay. The general harmonized QoS modelling establishes a framework for selling differentiated QoS to a wider range of emerging services, such as different services required by IoT and M2M communications.

2.2.4 CSI Feedback and Scheduling Announcement

It is assumed that CSIs, or their quantized version, are available at (or fed back to) the BS for making the scheduling decisions. In this section, the CSI feedback and the scheduling announcements will be described. In a time division sduplexing (TDD) system, CSIs are available for both

downlink and uplink at the BS, due to reciprocity. In a frequency division duplexing (FDD) system, however, in downlink, feedback channels are needed for reporting CSI. Since the CSI is used for calculating AMC values, the infinite precision values of CSI are not necessary. In fact, AMC values are sufficient to make the RRM decisions at the BS. We assume that the number of AMC modes for WT m is equal to amc_m.

In the uplink, CSIs are known at the BS and the AMC values are calculated at the BS, where the decision will be made. Therefore, no need for CSI reporting in uplink. The BS, then, announces the RRM decisions to the WTs. The decision announcement can be described with $\left\lceil \log_2 \left(1 + \sum_m^M \mathsf{amc}_m\right) \right\rceil$ number of bits, where integer numbers between 0 and $\sum_m^M \mathsf{amc}_m$ are used to indicate which WT with which AMC mode should transmit. Here, $m \in \{1, \ldots, M\}$ corresponds to M WTs.

In the downlink, however, AMC values are calculated at the WT and must be reported to the BS. The WTs calculate the AMC values and report them to BS, for decision making. Therefore, each WT needs to report its calculated AMC which requires a $\lceil \log_2 (1 + \mathsf{amc}_m) \rceil$ number of bits per WT. Then, BS decides the scheduling and announce it with $\lceil \log_2(M) \rceil$ number of bits, in order to indicate which WT should transmit.

2.3 Packet Scheduling and Resource Block Allocation for NRT flows

In this part, different scheduling algorithms for NRT flows will be surveyed. Since the NRT flows are delay insensitive, the utility functions will be based on bit-rate information.

2.3.1 Bit-rate GPF

From a networking perspective the NRT flows bit-rates should satisfy bit-rate fairness, otherwise the network operator may have too many unhappy customers who are starved out by the WTs with high spectral efficiency. The utility based scheduling is taken from networking research [16, 31–37], where each WT is assigned a utility function, such that when the network utility is maximum for

a given set of bit-rates, over all other bit-rates, those bit-rates are fair with respect to the utilities. Utility functions have originated from economics to quantify the degree of satisfaction a user enjoys in using a certain resource. Different fairness goals are achieved with different utility functions. In general, utility functions are assumed to be differentiable, concave, non-decreasing on some interval.

Bit-rate GPF utility functions, which result in α-fair, are defined with

$$S_{\phi}^{\bar{\mathrm{r}}}\Big(\bar{r}_{\phi}[k]\Big) = \begin{cases} \frac{w_{\phi}^{\mathrm{GPF}}[k]}{1-\alpha}\big(\bar{r}_{\phi}[k]\big)^{1-\alpha}, & \text{if} \quad 0 \leq \alpha < 1, \\ w_{\phi}^{\mathrm{GPF}}[k] \ \log\big(\bar{r}_{\phi}[k]\big), & \text{if} \quad \alpha = 1, \end{cases} \tag{2.12}$$

where α is the parameter influencing the kind of bit-rate fairness we expect, $\bar{r}_{\phi}[k]$ is the mean bit-rates received by flow ϕ until frame k as defined in (2.6), and $\hbar[k] \in (0,1)$ is the forgetting factor. The forgetting factor can be selected to be either asymptotically vanishing as $\hbar[k] = 1/k$ or can be selected as a small constant. Similar to the least mean-square (LMS) algorithm, a constant forgetting factor is robust to fluctuation in channels, while vanishing step size ensures convergence of the gradient algorithm in terms of converging mean bit-rates when the channel processes are stationary [38]. Each flow has a mean bit-rate utility function denoted by $S_{\phi}^{\bar{\mathrm{r}}}(\bar{r}_{\phi}[k])$. Note that even though the exponential in the utility function is undefined for $\alpha = 1$, its derivative approaches the derivative of log when $\alpha \to 1$.

The sum utility is the network utility as

$$S_{\mathrm{net}}^{\bar{\mathrm{r}}}\Big(\bar{\mathbf{r}}[k]\Big) \triangleq \sum_{\phi=1}^{|\Phi_{\mathrm{NRT}}|} S_{\mathrm{net}}^{\bar{\mathrm{r}}}\Big(\bar{r}_{\phi}[k]\Big), \tag{2.13}$$

where $\mathbf{r}[k]$ is the vector of flow bit-rate in frame k. Bit-rates, which maximize the sum utility for a specific α are said to be α-fair.

The GPF optimization, which maximizes network utility is

$$\max_{x_{\phi}^{(j)}[k] \, \in \, \mathcal{C}^{\mathrm{PHY}}} \quad \sum_{\phi=1}^{|\Phi_{\mathrm{NRT}}|} \frac{w_{\phi}^{\mathrm{GPF}}[k]}{1-\alpha}\Big(\bar{r}_{\phi}[k]\Big)^{1-\alpha}, \tag{2.14}$$

where $\mathcal{C}^{\text{PHY}}$ is the feasible set defined by PHY layer constraints as

$$\mathcal{C}^{\text{PHY}} = \Big\{ x_{\phi}^{(j)}[k] \Big| \quad \forall\, j: \sum_{\phi=1}^{|\Phi_{\text{NRT}}|} x_{\phi}^{(j)}[k] \leq T, \quad \forall\, \phi, j: \; x_{\phi}^{(j)}[k] \in \{0, \ldots, T\} \Big\}, \tag{2.15}$$

where $|\Phi_{\text{NRT}}|$ is the number of NRT flows, N is the number of sub-channels, T_b is the time duration of the RB, α is the parameter which sets the type of bit-rate fairness, $b_{\phi}^{(j)}[k]$ is the number of bits that can be transmitted to flow ϕ on sub-channel j, and $x_{\phi}^{(j)}[k]$ is the number of RBs assigned to flow ϕ on sub-channel j. The constraints ensure that the total number of allocated blocks does not exceed what is available in the frame. In the sequel, we use the notation $\mathcal{C}^{\text{PHY}}$ to describe the feasible set of described in the above optimization, induced by physical limitation. Depending on the parameter α, the optimization results in different types of the bit-rate fairness, as explained.

Flexible Bit-rate Fairness based on GPF

Different types of fairness [31] can be achieved by changing the parameter α, where $\alpha \to 1$ corresponds to weighted PF [39], $w_{\phi}^{\text{GPF}}[k] = 1$ and $\alpha = 0$ corresponds to *maximum throughput* and $w_{\phi}^{\text{GPF}}[k] = 1$ and $\alpha \to \infty$ corresponds to *maxmin fairness* [31].

For $\alpha \to 0$ the network utility corresponds to throughput as defined by

$$S_{\text{net}}^{\bar{\mathbf{r}}}\Big(\bar{\mathbf{r}}[k]\Big) = \sum_{\phi=1}^{|\Phi_{\text{NRT}}|} \bar{r}_{\phi}[k], \qquad \text{if} \quad \alpha = 0. \tag{2.16}$$

For $\alpha \to 1$, the bit-rates maximizing the network utility are PF [39] with objective defined as

$$S_{\text{net}}^{\bar{\mathbf{r}}}\Big(\bar{\mathbf{r}}[k]\Big) = \sum_{\phi=1}^{|\Phi_{\text{NRT}}|} \log\Big(\bar{r}_{\phi}[k]\Big), \qquad \text{if} \quad \alpha \to 1. \tag{2.17}$$

Early adoptions of PF, in the legacy Qualcomm high data rate (HDR) network, are reported in [29, 40, 41]. In environments with little scattering or slow fading the multiuser diversity is low, therefore, [29] proposes the use of multiple transmit antennas to induce fast channel fluctuations so that multiuser diversity can still be exploited by a PF. The PF is characterized with its interpreted

fairness notion that if we use another scheduling algorithm to increase the throughput of a specific flow by x % over what that flow receives under the PF scheduling, the summation of all the percentage decreases suffered by the throughput of other flows will be more than x %. Scheduling with PF results in serving the flows with equal probability regardless of their possibly different average channel quality [42]. Since the packets, scheduled with PF, are from flows near its possible peak bit-rate, transmission time will be reduced. This results in power efficiency due to the fact that the WT can remain longer in energy-conserving mode. However, GPF are relevant only for NRT flows and cannot control potential traffic delay violations resulting from delaying transmission until a WT link reaches its peak AMC bit-rate [29, 43].

For $\alpha = 2$, the bit-rates maximizing the GPF utility minimize the *potential delay* [44] as

$$S_{\text{net}}^{\bar{\mathbf{r}}}\Big(\bar{\mathbf{r}}[k]\Big) = -\sum_{\phi=1}^{|\Phi_{\text{NRT}}|} \frac{1}{\bar{r}_{\phi}[k]}, \qquad \text{if} \quad \alpha = 2, \tag{2.18}$$

where each term in the summation is the potential delay which is inversely proportional to flow bit-rate. This potential delay is valid only when we have same backlog among queues.

The Gradient for Bit-rate GPF

Although the relaxed version of GPF optimization is convex, solving with convex programming suffers from several deficiencies. First, the optimal solution consists of real-numbers, which should somehow be converted to integers. Second, the size of the optimization can quickly get out of control. Therefore, integer-based solution for the network utility maximization, with relatively simple complexity is of interest. The main solution idea is based on the fact that the maximum change in the objective function, that can be obtained from increasing one time-allocation by one, is obtained by adding time allocation in the direction of the steepest gradient of the objective function. To make the best change in the objective by increasing only one scheduling variable, the variable with the steepest gradient should be chosen. Suppose we want to add one slot to flow ϕ

on sub-channel j. By Taylor's expansion, the network utility can be approximated by

$$S^{\bar{\mathrm{r}}}_{\mathrm{net}}(\cdots, x^{(j)}_{\phi}[k]+1, \cdots) \approx S^{\bar{\mathrm{r}}}_{\mathrm{net}}(\cdots, x^{(j)}_{\phi}[k], \cdots) + \frac{\partial}{\partial x^{(j)}_{\phi}[k]} S^{\bar{\mathrm{r}}}_{\mathrm{net}}(\cdots, x^{(j)}_{\phi}[k], \cdots). \tag{2.19}$$

This is the general idea for the gradient-based algorithm. Therefore, the bit-rate utility, $S^{\bar{\mathrm{r}}}_{\mathrm{net}}(\cdots)$, can be replaced with the general utility, $S^{\mathrm{QoS}}_{\mathrm{net}}(\cdots)$. For specific GPF utilities, the gradient is

$$\frac{\partial}{\partial x^{(j)}_{\phi}[k]} S^{\bar{\mathrm{r}}}_{\mathrm{net}}(\ldots, x^{(j)}_{\phi}[k], \ldots) = \frac{W_b \; b^{(j)}_{\phi}[k]}{\Big(W_b \sum_{j=1}^{N} b^{(j)}_{\phi}[k] x^{(j)}_{\phi}[k]\Big)^{\alpha}}. \tag{2.20}$$

Accordingly, if we are given a choice of increasing any one of $x^{(j)}_{\phi}[k]$, we should increase the time allocation of the sub-channel j for flow ϕ with the highest partial derivative, to maximize the incremental change in the objective function. The gradient algorithms are greedy in the sense that winner flow takes the entire RB.

The obtained sequence, indexed by frame index, converges in probability to the optimal allocation under mild conditions, regardless of the initialization. For asymptotic optimality proof of gradient algorithm see [16, 45]. Using the previous observation about the objective function, we devised an iterative greedy heuristic algorithm for amplify-and-forward (AF) relay context in [13, 18, 19]. It has been reported that the Taylor approximation works closer to optimal when $NT/|\Phi| >> 1$, especially for short-term fairness.

Online Algorithm versus Offline Algorithm

It is worth mentioning that PF, and in general GPF, can also be implemented based on linear weighted sum bit-rate maximization, as

$$\max_{x^{(j)}_{\phi}[k] \in \mathcal{C}^{\mathrm{PHY}}} \bar{\mathbf{r}}[k] \;\; \Big(\mathbf{w}^{\mathrm{off}}[k]\Big)^{\mathrm{T}}, \tag{2.21}$$

where $\mathbf{w}^{\mathrm{off}}[k]$ is the weights vector corresponding to the bit-rate vector $\bar{\mathbf{r}}[k]$. When the weights are selected as proportional or equal to the mean of the SINR, the above maximization is equivalent

to PF. This approach is an off-line approach, as the weights depends on the knowledge of the statistics of channels. The gradient based PF, on the other hand, is an online approach as it does not require a priori knowledge of the channels. Online scheduling algorithms learns dynamically the appropriate weights to adapt to the intended channel statistics and converge to the optimal allocation, as the number of frames becomes large enough.

TCP Reverse Engineering Example

Without getting into the mathematical details of utility functions, one can still mention that the transmission control protocol (TCP) was reverse engineered as a utility optimization, which makes the network proportionally fair [42]. In other words, one can find utility functions to model different variants of TCP [46] as an optimization.

2.3.2 Bit-rate Maxmin Fairness

After introducing the GPF, in this section we describe the bit-rate maxmin as an asymptotic case of GPF. The maxmin fairness is interesting by the fact that it is able to provide the most ubiquitous bit-rate coverage. The GPF formulation is asymptotically equivalent to the maxmin bit-rate fairness as the parameter of the GPF, α, goes to infinity. We will prove that as parameter α goes to infinity the previously GPF gradient defined in (2.44) can be simplified as selecting the minimum bit-rate flow at each step. As discussed in Section 2.3.1, in GPF scheduling, α parameter gradually changes schedules from sum bit-rate optimal to PF. As $\alpha \to \infty$, it asymptotically achieve maxmin fairness among the bit-rates which is proved based on **Proposition 1** and **Proposition 2**.

As $\alpha \to \infty$, the network utility leads to a maxmin fair allocation of bit-rates [31]. This can be shown by noting an interesting relationship between the network utility maximization and the L_p norm minimization of the inverse of bit-rates values.

Proposition 1. *For $\alpha \to \infty$, the network utility maximization leads to bit-rate maxmin objective.*

Proposition 2. *For α sufficiently large, assigning RB to the flow with the minimum current bit-rate on its best sub-channel is equivalent to assigning resources to the flow with largest gradient in*

(2.1).

Using the proposition, we see that finding the largest derivative is equivalent to assigning RB to the flow with the minimum current bit-rate to its best sub-channel.

Similar to GPF bit-rate fairness, maxmin bit-rate fairness is suitable for NRT flows, where the traffic is delay insensitive. Maxmin bit-rate fairness provides the highest bit-rate fairness.

2.3.3 Bit-rate Fairness through Jain's Index Maximization

Generally, scheduling encounters by conflicting goals. Fair utility-based resource allocation trade-offs the efficiency and fairness. For instance, favouring a high spectral efficient WT can increase the system efficiency, but would result in the dissatisfaction of other WTs. It is shown in [9] that α-type fairness provides optimum tradeoff between efficiency and a *variable fairness measure that changes with the extrinsic parameter* α. However, since varying α changes the fairness measure itself, a question that arises is whether the α-fair policy achieves the optimal efficiency-fairness tradeoff in practical resource allocation scenarios wherein the fairness measure does not depend on extrinsic parameters such as α. The answer is that α-type fair scheduler does not guarantee optimum tradeoff between efficiency and a fixed fairness measure (that does not change with α), such as the Jain's fairness index [47].

As an accepted intuitive fairness measure, the Jain's index is defined as

$$J\Big(\mathbf{v}[k]\Big) = \left(\sum_{\phi=1}^{|\Phi_{\mathrm{NRT}}|} v_\phi[k]\right)^2 \Bigg/ |\Phi_{\mathrm{NRT}}| \sum_{\phi=1}^{|\Phi_{\mathrm{NRT}}|} (v_\phi[k])^2, \tag{2.22}$$

to measure the fairness of the allocated resources in the vector $\mathbf{v}[k]$, or equivalently to measures how similar are the allocated quantities, in frame k. For $J\left(\mathbf{v}[k]\right)$ close to one the quantities, in frame k, are the most similar, so the system is in extreme fair case in terms this quantity in frame k. For $J\left(\mathbf{v}[k]\right)$ close to $\frac{1}{|\Phi_{\mathrm{NRT}}|}$, the quantities are the least similar, so the system is in a unfair case. The Jain's index can be interpreted as the squared first sample moment of bit-rates over second sample moment of bit-rates. The Jain's index implicitly determines the fairness improvement (decreasing in variance) with considering the cost (decreasing in mean) of this improvement. Among many

fairness notions [9], the Jain's index is an standard interpretable fairness measure. For instance, the Jain's index of $\frac{p}{100}$ can be regarded as the fairness index of an equivalent resource allocation in which p % of the flows receive equal non-zero benefits and the remaining $(100-p)$ % flows receive zero benefits [48].

It is shown that α-type fairness cannot achieve the Pareto front of efficiency and the Jain's index tradeoff (EJT), except for the case of $|\Phi_{\rm NRT}| = 2$ flows. References [47, 49] derive sufficient conditions for optimizing the efficiency while preserving the fairness as

$$\max_{J_0 \leq J(\mathbf{v}[k])} \eta\big(\mathbf{v}[k]\big), \tag{2.23}$$

where the efficiency, $\eta(\mathbf{v}[k])$, and the constraint, $J_0 \leq J(\mathbf{v}[k])$, are defined as

$$\eta(\mathbf{v}[k]) = \sum_{\phi=1}^{|\Phi_{\rm NRT}|} v_\phi[k] \quad \text{and} \quad J_0 \leq \frac{\Big(\sum_{\phi=1}^{|\Phi_{\rm NRT}|} v_\phi[k]\Big)^2}{|\Phi_{\rm NRT}| \sum_{\phi=1}^{|\Phi_{\rm NRT}|} \big(v_\phi[k]\big)^2}. \tag{2.24}$$

In other words, the optimization finds the Pareto optimal [50] front of η and J. Pareto optimal points are the ones at which efficiency cannot be increased without decreasing the Jain's index, and likewise, the Jain's index cannot be increased without decreasing efficiency [50].

Monotonic Tradeoff Property for Jain's Index Maximization

The solution to (2.23), in general, is not trackable, however, a property named *monotonic tradeoff property* is proved in [47] to simplify the problem. A set $\mathcal{C}^{\rm PHY}$ is said to have the monotonic tradeoff property if the Jain's index J^*_σ is strictly decreasing in σ, for $\sigma > \sigma^*$, and constant otherwise. The yields to the fact that a decrease in efficiency results in a strict increase in the Jain's index, until σ^* is reached. Decreasing efficiency beyond σ^* maintains the Jain's index at its maximum. It is shown that if the set $\mathcal{C}^{\rm PHY}$ has monotonic tradeoff property [47] then

$$\mathbf{v}^*_\sigma[k] = \left\{ \mathbf{v}[k] \middle| \mathbf{v}[k] = \arg \max_{\sigma \leq \eta(\mathbf{v}[k]),\, \mathbf{v}[k] \in \mathcal{C}^{\rm PHY}} J(\mathbf{v}[k]) \right\} \tag{2.25}$$

is equivalent to the case when the inequality is replaced by equality. Further, since the efficiency is constant (note that the Jain's index numerator is the square of efficiency) we have the equivalent minimization as

$$\min_{\eta(\mathbf{v}[k])=\sigma,\ \mathbf{v}[k]\in\mathcal{C}^{\mathrm{PHY}}} \left\|\mathbf{v}[k]\right\|^2. \tag{2.26}$$

In contrast to the previous equivalent form (2.23), the objective is strictly convex, which implies that when $\mathcal{C}^{\mathrm{PHY}}$ is also convex, the solution to the above optimization problem is trackable and unique [50]. In summary, if the *monotonic tradeoff property* is satisfied, $\mathbf{v}^*_\sigma[k]$ can be found by solving second norm minimization in (2.26), which is significantly easier to solve than the optimization problems in (2.23) for an arbitrary $\mathcal{C}^{\mathrm{PHY}}$.

It is proved [47] that a set $\mathcal{C}^{\mathrm{PHY}}$ possesses the monotonic tradeoff property if the set $\mathcal{C}^{\mathrm{PHY}}$ is convex, vector $v_{\min}\mathbf{1}_{|\Phi_{\mathrm{NRT}}|} \in \mathcal{C}^{\mathrm{PHY}}$, and for all $\mathbf{v}[k] \in \mathcal{C}^{\mathrm{PHY}}$ satisfies $\mathbf{v}[k] \geq v_{\min}\mathbf{1}_{|\Phi_{\mathrm{NRT}}|}$.

In other words, a benefit vector $\mathbf{v}^*_\sigma[k]$ achieves the optimal EJT if there is no feasible benefit vector $\mathbf{u} \neq \mathbf{v}^*_\sigma[k]$ such that $\eta(\mathbf{u})$ is closer to the fairest solution $\sigma\mathbf{1}_{|\Phi_{\mathrm{NRT}}|}/|\Phi_{\mathrm{NRT}}|$. The solution is the projection of fairest potentially unfeasible solution to the feasible set as

$$\min_{\eta(\mathbf{v}[k])=\sigma,\ \mathbf{v}[k]\in\mathcal{C}^{\mathrm{PHY}}} \left\|\mathbf{v}[k] - \beta\mathbf{1}_{|\Phi_{\mathrm{NRT}}|}\right\|. \tag{2.27}$$

Based on the monotonic tradeoff property, the solution of second norm minimization in (2.26) remains unchanged if the equality constraint is replaced by the inequality $\eta(\mathbf{v}[k]) \geq \sigma$. Hence, 2β is the non-negative Lagrange multiplier corresponding to this constraint.

Gradient based Core

Although for static channels, time sharing and solving using a standard convex optimization problem is possible, for time-varying channels, time sharing and standard convex optimization is implausible. Therefore, the scheduler should be simplified, using the gradient-based approach. For optimizing the long-term (or steady-state) mean bit-rates, the benefit vector is defined as

$$\mathbf{v}[k] = \lim_{k\to\infty} \bar{\mathbf{r}}[k]. \tag{2.28}$$

Given the exponentially-weighted mean bit-rates at frame index $k-1$ and the instantaneous bit-rates, $b_\phi^{(j)}[k]$, the task of the scheduler is to determine the instantaneous scheduling variables, $x_\phi^{(j)}[k]$, in such a way that maximizes a given system utility in (2.27). The corresponding gradient scheduling algorithm uses the first order Taylor's series expansion of (2.27) in frame index k. Accordingly, the problem will be simplified to selecting the one that has the largest distance to β as

$$\Big(\phi^*[k], j^*[k]\Big) = \arg\max_{(\phi,j)} b_\phi^{(j)}[k]\Big(\beta - \bar{r}_\phi[k-1]\Big). \tag{2.29}$$

This gradient scheduling algorithm provides the Pareto optimal tradeoff, by changing β, between the Jain's index and sum bit-rate [47].

2.3.4 General Utility Fairness for Bit-rate through Jain's Index Maximization

As we discussed in the previous section, maximizing the Jain's index provides efficiency and fairness trade-offs. However, optimizing the Jain's index of the raw bit-rate does not capture the heterogeneity among different flows which have different utilities of the bit-rate. As an example, one can consider a cellular system where the WTs in the cell edge have very tough utility with respect to bit-rates. Formulating the problem based on raw bit-rate cannot capture this heterogeneity. Instead, the Jain's index of utilities should be maximized. This makes the MAC layer aware of the real utility of the resources, when maximizing the Jain's index. Based on this idea, the fairness is generalized in equalizing the utilities rather than the raw bit-rate. Then, the formulation is

$$\max_{J_0 \le J\left(\bar{r}_\phi[k]\right),\ x_\phi^{(j)}[k] \in \mathcal{C}^{\mathrm{PHY}}} \ \sum_{\phi=1}^{|\Phi_{\mathrm{NRT}}|} S_\phi^{\bar{\mathrm{r}}}\big(\bar{r}_\phi[k]\big). \tag{2.30}$$

Since different applications have different utility perception, *fairness* in this heterogeneous context means distributing resources with respect to different utilities. In other words, utility fairness is the notion which equalizes the utilities among flows, instead of equalizing the raw QoS elements among flows.

Gradient based Core

The corresponding scheduling core is suggested as

$$\Big(\phi^*[k], j^*[k]\Big) = \max_{(\phi,j)} b_\phi^{(j)}[k]\Big(\beta^{\mathrm{S}^{\bar{\mathrm{r}}}} - S_\phi^{\bar{\mathrm{r}}}(\bar{r}_\phi[k-1])\Big), \tag{2.31}$$

where $\beta^{\mathrm{S}^{\bar{\mathrm{r}}}}$ is the parameter which changes the tradeoff between bit-rate fairness and efficiency.

2.3.5 Efficiency and Fairness tradeoff through Multi-objective Optimization

As we already discussed, naïve optimization of the throughput will result in starvation for some of the WTs, decrease the WT loyalty and end up being very far from the optimal network efficiency. Here, we are proposing to maximize the efficiency and at the same time minimize the discrimination. Having discrimination minimization results in *service-ubiquity* in the cell, generalizing the bit-rate ubiquity.

We model the efficiency of the network by $S_{\mathrm{net}}^{\bar{\mathrm{r}}}\Big(\bar{\mathbf{r}}[k]\Big)$ and the discrimination by $\Pi_{\mathrm{net}}^{\bar{\mathrm{r}}}\Big(\bar{\mathbf{r}}[k]\Big)$, with respect to the mean bit-rate, as

$$S_{\mathrm{net}}^{\bar{\mathrm{r}}}\Big(\bar{\mathbf{r}}[k]\Big) = \sum_{\phi=1}^{|\Phi_{\mathrm{NRT}}|} S_\phi^{\bar{\mathrm{r}}}(\bar{r}_\phi[k]) \quad \text{and} \quad \Pi_{\mathrm{net}}^{\bar{\mathrm{r}}}\Big(\bar{\mathbf{r}}[k]\Big) = \sum_{\phi=1}^{|\Phi_{\mathrm{NRT}}|} \left(S_\phi^{\bar{\mathrm{r}}}\big(\bar{r}_\phi[k]\big) - \frac{S_{\mathrm{net}}^{\bar{\mathrm{r}}}\Big(\bar{\mathbf{r}}[k-1]\Big)}{|\Phi_{\mathrm{NRT}}|} \right)^2. \tag{2.32}$$

The *discrimination* is indeed the *efficiency risk*, as a variance, that the operator will endure with regards to its efficiency.

Formulation and the Gradient based Core

We now propose to use a multi-objective optimization (with respect to the first Euclidean quadrant cone, $\mathbb{R}_+^2$), as maximizing the total utility while minimizing the discrimination as

$$\max_{\bar{\mathbf{r}}[k]\in\mathcal{C}^{\mathrm{PHY}}} \Big[S_{\mathrm{net}}^{\bar{\mathrm{r}}}\Big(\bar{\mathbf{r}}[k]\Big), -\Pi_{\mathrm{net}}^{\bar{\mathrm{r}}}\Big(\bar{\mathbf{r}}[k]\Big)\Big]_{\mathbb{R}_+^2} \text{ which is equivalent to } \max_{\bar{\mathbf{r}}[k]\in\mathcal{C}^{\mathrm{PHY}}} S_{\mathrm{net}}^{\bar{\mathrm{r}}}\Big(\bar{\mathbf{r}}[k]\Big) - \varrho\Pi_{\mathrm{net}}^{\bar{\mathrm{r}}}\Big(\bar{\mathbf{r}}[k]\Big). \tag{2.33}$$

The equivalent formulation is based on the fact that for a multi-objective optimization, a feasible allocation is *Pareto* optimal, if it is optimal for the scalar version of the problem [50]. Parameter ϱ tradeoffs different Pareto optimal allocation. The operator will have this choice to force the algorithm to control the discrimination minimization by increasing or decreasing parameter ϱ, achieving a predetermined tolerable discrimination. Applying the gradient scheduling algorithm, we get

$$\Big(\phi^*[k], j^*[k]\Big) = \arg\max_{(\phi,j)} \frac{\partial S^{\bar{\mathrm{r}}}_{\phi}\big(\bar{r}_{\phi}[k]\big)}{\partial x^{(j)}_{\phi}[k]} \Big[\frac{1}{2\varrho} - S^{\bar{\mathrm{r}}}_{\phi}\big(\bar{r}_{\phi}[k]\big) + \frac{S^{\bar{\mathrm{r}}}_{\mathrm{net}}\big(\mathbf{r}[k-1]\big)}{|\Phi_{\mathrm{NRT}}|}\Big]. \tag{2.34}$$

Changing parameter ϱ from 0 to ∞ increases the fairness from sum bit-rate maximization to scheduling resembling the Jain's index maximization, similar to efficiency and Jain's index maximization.

Intra-class Fairness as an Application of the Multi-objective Optimization

Interestingly, the multi-objective approach can be used to implement intra-class fairness. In fact, the flows with the same QoS-class should experience a same level of QoS, in average. This can be implemented by intra-class fairness to equalize the utilities within a class. Based on the concept developed in Section 2.3.5, here we modify the core and outline the intra-class fairness. The utility fairness term is to penalize the objective if $S^{\mathrm{QoS}}_{\mathrm{net}}$ is far from the average of it in the same subscription class based on the following core as

$$\Big(\phi^*[k], j^*[k]\Big) = \arg\max_{(\phi,j)} \frac{\partial S^{\bar{\mathrm{r}}}_{\phi}\big(\bar{r}_{\phi}[k]\big)}{\partial x^{(j)}_{\phi}[k]} \, \Im_{\phi} \left(\frac{1}{2\varrho} - S^{\bar{\mathrm{r}}}_{\phi}\big(\bar{r}_{\phi}[k]\big) + \frac{1}{|\Phi_{\theta_{\phi}}|} \sum_{\phi' \in \Phi_{\theta_{\phi}}} S^{\bar{\mathrm{r}}}_{\phi}(\bar{r}_{\phi}[k-1]) \right), \tag{2.35}$$

where $\Im_{\phi}(\cdot)$ is the generalization of the perceived fairness on the utility, as a general function ($\Im_{\phi}(\cdot)$) on the distance between flow utility and the intra-class average utility in the previous frame, and $\Phi_{\theta_{\phi}}$ is the ensemble of flows in the same QoS class as the flow ϕ.

In some scenarios, the network is interested in proving different notion of fairness among different subset of flows. For example, we can have a very tight fairness among a certain subset of flows while we have a relaxed fairness among another set of flows. Nevertheless, when we mix the flows, it is in contrast to the fairness notion definition to have more than one fairness notions in a same

pool of RBs.

2.3.6 Minimum Bit-rate QoS Guarantee through Virtual Token

References [51,52] studied the general utility maximization, including GPF, subject to the minimum and maximum bit-rate constraints in infinite-backlog queues in single channel case, based on token counter. The optimization maximizes network utility subject to the minimum and maximum bit-rate requirement as

$$\max_{x_\phi^{(j)}[k] \in \mathcal{C}^{\mathrm{PHY}} \bigcap \mathcal{C}^{\bar{\mathrm{r}}^{\min}} \bigcap \mathcal{C}^{\bar{\mathrm{r}}^{\max}}} \quad \sum_{\phi=1}^{|\Phi_{\mathrm{NRT}}|} S_\phi^{\bar{\mathrm{r}}}\big(\bar{r}_\phi[k]\big), \tag{2.36}$$

where $\mathcal{C}^{\bar{\mathrm{r}}^{\min}}$ and $\mathcal{C}^{\bar{\mathrm{r}}^{\max}}$ describe the minimum bit-rates and maximum bit-rates requirements as

$$\mathcal{C}^{\bar{\mathrm{r}}^{\min}} = \Big\{ x_\phi^{(j)}[k] \Big| \quad \forall\, \phi :\ \bar{r}_\phi^{\min} \le \bar{r}_\phi[k] \Big\} \quad \text{and} \quad \mathcal{C}^{\bar{\mathrm{r}}^{\max}} = \Big\{ x_\phi^{(j)}[k] \Big| \quad \forall\, \phi :\ \bar{r}_\phi[k] \le \bar{r}_\phi^{\max} \Big\}. \tag{2.37}$$

The virtual token mechanism is a modification on the gradient algorithm solving the corresponding unconstrained problem. The modification produces a gradient algorithm solving the problem with maximum and minimum bit-rate constraints. This approach selects the flow $\phi^*[k]$ to be transmitted on RB $j^*[k]$ based on

$$\Big(\phi^*[k], j^*[k]\Big) = \arg\max_{(\phi,j)} \ \mathrm{e}^{\lambda_\phi^{\mathrm{token}}[k]} \ \frac{\partial S_\phi^{\bar{\mathrm{r}}}\left(\bar{r}_\phi[k]\right)}{\partial \bar{r}_\phi[k]} \ \frac{\partial \bar{r}_\phi[k]}{\partial x_\phi^{(j)}[k]}, \tag{2.38}$$

where $\lambda_\phi^{\mathrm{token}}[k]$ is the token counter for flow ϕ. If flow ϕ receives services less than $\bar{r}_\phi^{\min}$, then its token counter has a positive drift, increasing the priority level of corresponding flow. If flow ϕ receives more than $\bar{r}_\phi^{\max}$, then the its token counter has a negative drift, decreasing the priority level of the of the corresponding flow and less likely to be served. The update rule for token counter is

$$\lambda_\phi^{\mathrm{token}}[k] = \begin{cases} \lambda_\phi^{\mathrm{token}}[k-1] + \bar{r}_\phi^{\min} - r_\phi[k], & \text{if} \quad \lambda_\phi^{\mathrm{token}}[k-1] \ge 0, \\ \lambda_\phi^{\mathrm{token}}[k-1] - \bar{r}_\phi^{\max} - r_\phi[k], & \text{if} \quad \lambda_\phi^{\mathrm{token}}[k-1] < 0. \end{cases} \tag{2.39}$$

Intuitive Explanation of the Virtual Tokens

The intuition behind update of the $\lambda_\phi^{\text{token}}[k]$ can be interpreted as follows. There is a virtual token queue (which can take either positive or negative values) corresponding to each flow. The tokens arrive in the token queue (token counter is incremented or decremented) at the rate $\bar{r}_\phi^{\min}$ or $\bar{r}_\phi^{\max}$ per iteration, depending on whether the token counter is positive or negative, respectively. If flow ϕ is served, the tokens are removed from the token queue equal to received service. Intuitively, if in an iteration, the mean bit-rate of flow ϕ is less than its minimum requirements, the token counter has positive drift, therefore, the priority of serving flow ϕ gradually increases. On the other hand, if the mean bit-rate of flow ϕ is higher than its maximum, then its token counter has negative drift, thus gradually decreasing the priority to be served. Associated with guaranteeing minimum bit-rate, the length of the averaging window, in which the requirement is satisfied, can be interpreted as the target delay of the minimum bit-rate requirement. Long delay windows allow longer periods of not being scheduled, whereas a short delay window will help schedule flows more often. For other similar approaches on incorporating minimum bit-rates with GPF see [33].

Guaranteeing the minimum bit-rate is important for some application, such as the streaming applications where a minimum bit-rate for a multi-description coding is needed to transmit the basic layer. In addition, it can smooth out the wireless channel and make the packet-based flows emulating a circuit-based flows. Differentiating the minimum bit-rate is a tool for selling different kind of service with different prices. On the other hand, the maximum bit-rate gives the incentive to the WTs to upgrade their service.

Ineffectiveness of Applying QoS-requirement Constraints through Barrier Functions

It is worth mentioning that another natural approach for implementing bit-rate constraints is an unconstrained optimization that deals with the bit-rate constraints by modifying the utility function, with barrier functions, in a way that penalizes bit-rate constraint violations. However, it is demonstrated in [51] that such an approach does not work well. The reason is that an algorithm with a modified utility function typically overreacts to temporary bit-rate constraint violations which significantly degrades the achieved value of the utility function.

2.3.7 Minimum Bit-rate QoS Guarantee through Lagrangian Multipliers

Using the method of the Lagrange multipliers, an alternative approach for implementing the bit-rate constraints in $\mathcal{C}^{\mathrm{r}^{\min}}$ can be derived. The Lagrangian of the constrained optimization, with $\mathcal{C}^{\mathrm{r}^{\min}}$, is

$$L^{\bar{\mathrm{r}}^{\min}}\left(\boldsymbol{\lambda}^{\bar{\mathrm{r}}^{\min}}[k], \bar{\mathbf{r}}[k]\right) = \sum_{\phi}^{|\Phi_{\mathrm{NRT}}|} \lambda_{\phi}^{\bar{\mathrm{r}}^{\min}}[k]\left(\bar{r}_{\phi}[k] - \bar{r}_{\phi}^{\min}\right) + \sum_{\phi}^{|\Phi_{\mathrm{NRT}}|} S_{\phi}^{\bar{\mathrm{r}}}\left(\bar{r}_{\phi}[k]\right), \tag{2.40}$$

where the Lagrangian multiplier vector, associated with constraints $\mathcal{C}^{\mathrm{r}^{\min}}$, is denoted by vector $\boldsymbol{\lambda}^{\mathrm{r}^{\min}}[k]$. This form transforms the constrained optimization into an unconstrained optimization of maximizing the Lagrangian over the primal and dual variables. The steepest gradient theory [50], should be performed on primal variables, $x_{\phi}^{(j)}[k]$, as well as the dual variables, $\lambda_{\phi}^{\bar{\mathrm{r}}^{\min}}[k]$. Accordingly, the primal variables are updated based on the gradient with respect to primal variables as

$$\frac{\partial L^{\bar{\mathrm{r}}^{\min}}\left(\boldsymbol{\lambda}^{\bar{\mathrm{r}}^{\min}}[k], \bar{\mathbf{r}}[k]\right)}{\partial x_{\phi}^{(j)}[k]} = \lambda_{\phi}^{\bar{\mathrm{r}}^{\min}}[k]\,\frac{\partial \bar{r}_{\phi}[k]}{\partial x_{\phi}^{(j)}[k]} + \frac{\partial S_{\phi}^{\bar{\mathrm{r}}}\left(\bar{r}_{\phi}[k]\right)}{\partial \bar{r}_{\phi}[k]}\,\frac{\partial \bar{r}_{\phi}[k]}{\partial x_{\phi}^{(j)}[k]}, \tag{2.41}$$

which makes the scheduling core as

$$\left(\phi^{*}[k], j^{*}[k]\right) = \arg\max_{(\phi,j)}\left(\lambda_{\phi}^{\bar{\mathrm{r}}^{\min}}[k] + \frac{\partial S_{\phi}^{\bar{\mathrm{r}}}\left(\bar{r}_{\phi}[k]\right)}{\partial \bar{r}_{\phi}[k]}\right)\frac{\partial \bar{r}_{\phi}[k]}{\partial x_{\phi}^{(j)}[k]}. \tag{2.42}$$

Intuitively, the flows with low spectral efficiency values (or AMC) but high minimum bit-rate are assisted through the addition of the of the Lagrangian multiplier to their derivative.

The Lagrange multiplier update is determined by the gradient of the Lagrangian with respect to the dual variables (Lagrange multipliers) equal to

$$\frac{\partial L^{\bar{\mathrm{r}}^{\min}}\left(\boldsymbol{\lambda}^{\bar{\mathrm{r}}^{\min}}[k], \bar{\mathbf{r}}[k]\right)}{\partial \lambda_{\phi}^{\bar{\mathrm{r}}^{\min}}[k]} = \bar{r}_{\phi}[k] - \bar{r}_{\phi}^{\min} \tag{2.43}$$

and the projection of the gradient into positive numbers, similar to the LMS update, as

$$\lambda_{\phi}^{\bar{\mathrm{r}}^{\min}}[k+1] = \max\left[0, \lambda_{\phi}^{\bar{\mathrm{r}}^{\min}}[k] - \hbar[k]\left(\bar{r}_{\phi}[k] - \bar{r}_{\phi}^{\min}\right)\right]. \tag{2.44}$$

This is an online algorithm which learns the channel statistics, along the scheduling satisfying the mean bit-rate constraints. The iterations are terminated when $\lambda_\phi[k+1] - \lambda_\phi[k] < \mathsf{err}$, for all ϕ. If the minimum bit-rate constraint is eliminated (equivalently when $\bar{r}_\phi^{\min} = 0$ for a certain ϕ) the $\lambda_\phi[k]$ will fall to 0 and the aforementioned update rule for the Lagrangian multipliers will reduce to gradient algorithm without bit-rate constraints.

Forgetting Factor Effect on the Convergence and the Robustness

The forgetting factor step size for minimum bit-rate brings mean bit-rates to a neighbourhood with a size of order of step size, $\mathcal{O}(\hbar[k])$, in number of iteration inversely proportional to step size $\mathcal{O}(1/\hbar[k])$. Similar to GPF gradient algorithm, convergence is ensured with an asymptotically vanishing step size, with the cost of lowering the robustness to channel non fluctuation. Smaller step size results in faster convergence but larger variability. Interestingly, the robustness and convergence tradeoff is similar to tracking and optimality tradeoff encountered with the LMS algorithm [38]. The minimum bit-rate implemented by the token mechanism if converge, converge to optimal. On the other hand, the Lagrangian approach is proved to converge to suboptimal solution. For detailed discussion on minimum bit-rate guarantee based on the Lagrangian approach see [53, 54].

2.3.8 Minimum Bit-rate QoS Guarantee through MLWDF

Another alternative for guaranteeing minimum bit-rare is using MLWDF scheduling in conjunction of virtual token buckets. Each flow will be associated with a virtual token queue. Tokens in flow ϕ arrive at the constant rate $\bar{r}_\phi^{\min}$, where $\bar{r}_\phi^{\min}$ is the required minimum bit-rate. Then, in each iteration, the priority decisions of RBs to flow mapping are made according to the MLWDF rule, when $d_\phi^{\mathrm{HOL}}[k]$ is replaced with HOL-delay of the longest waiting token in token bucket ϕ, instead of being the real HOL-delay of flow ϕ. Subsequently, the number of tokens in the corresponding bucket is reduced by the actual amount of data served. The HOL-delay of token can be implemented just as a counter. This is due to the fact that the tokens arrive at a constant rate and $d_\phi^{\mathrm{HOL}}[k]$, in this case, is equal to $d_\phi^{\mathrm{HOL}}[k] = q_\phi[k]/\bar{r}_\phi^{\min}$.

Since MLWDF is a throughput optimal scheduling rule [55] and token queues are stable, the

actual throughput of each flow ϕ is at least $\bar{r}_{\phi}^{\min}$. In this case, parameters μ_{ϕ} also control the time scale on which throughput guarantees is provided. The greater the μ_{ϕ} for a certain flow, the desired minimum bit-rate is provided on a finer time scale. In other words, the mean bit-rate is satisfied over shorter time scale [55,56]. The MLWDF also is used for maximum HOL-delay guarantee which will be discussed later in scheduling for RT flows section.

2.4 Packet Scheduling and Resource Block Allocation for RT flows

2.4.1 Mean-delay Fairness through WGPF

Majority of results reported in the literature on fairness are confined to bit-rate fairness, often without adequate attention to delay fairness. In this section we design delay fairness objectives, in which delay measures are equalized among flows. This section is based on our previous work on the delay fairness through scheduling [12]. We show a design methodology for mean-delay fairness, as an example of QoS-fairness. In general, QoS-fairness can be defined as applying fairness on a general subset of QoS elements, instead of a single QoS element.

For a finite-backlog queue model, any bit-rate fairness such as bit-rate GPF or bit-rate maxmin cannot guarantee to what extent the queueing delay quantities are equalized. In other words, fairness in terms of bit-rates, such as GPF, cannot guarantee the fairness on delay because the objective is blind to any queue information [12]. Delay fairness is generally the effective fairness answer to the finite-backlog scenario in order to equalize the delay among flows.

We start with non-traffic-aware weighed generalized proportional fairness (WGPF) objective and modify it to make an algorithm for delay fairness. We show how the WGPF objectives are connected and are equivalent to the delay-fair algorithm, derived from the Little's law [36]. We, then, show how the fair framework can be interpreted as minimizing the total delay disutility. We use convex increasing functions to describe the *delay disutility* of flows as the delay counterpart of bit-rate utility functions. This interpretation explains how the utility functions fit in the traffic-aware schedulers. It will be shown how the developed framework for a special disutility function, asymptotically (with respect to the associated parameter $\nu_{\bar{\mathrm{d}}}$) lead to the minmax mean-delays

fairness. The developed framework, in this special case, can be adjusted between two extreme objectives: minimizing total mean delay and minimizing the maximum mean delay. Finally, we prove that the gradient scheduling algorithm is equivalent asymptotically to serving the flow with the largest mean-delay, at each iteration.

Mean-delay Fairness Formulation

To guarantee the delay fairness, the scheduler should be both queue-aware and channel-aware. The general idea is that serving the flows with the longer mean queue lengths results in greater satisfaction than serving the flows with smaller mean queue lengths. We set the weights of WGPF as $w_{\phi}^{\mathrm{GPF}}[k] = (\bar{q}_{\phi}[k])^{\bar{\nu^{\mathrm{d}}}}$ which results in network utility as

$$S_{\mathrm{net}}^{(\bar{\mathbf{q}},\bar{\mathbf{r}})}\Big(\bar{\mathbf{q}}[k],\bar{\mathbf{r}}[k]\Big) = \sum_{\phi=1}^{|\Phi_{\mathrm{RT}}|} \frac{\Big(\bar{q}_{\phi}[k]\Big)^{\bar{\nu^{\mathrm{d}}}}}{1-\alpha}\Big(\bar{r}_{\phi}[k]\Big)^{1-\alpha}, \tag{2.45}$$

where $\bar{q}_{\phi}[k]$ is the mean queue length of flow ϕ until frame k as

$$\bar{q}_{\phi}[k] = \left(1-\frac{1}{k}\right)\bar{q}_{\phi}[k-1] + \frac{1}{k}q_{\phi}[k] \tag{2.46}$$

and $q_{\phi}[k]$ is the frame queue length at the end of frame k, for flow ϕ. Parameter $\bar{\nu^{\mathrm{d}}}$ is a constant determining the importance level of the traffic-aware characteristic. For the constant mean queue lengths (infinite-backlog) the objective reduces to conventional GPF.

Now, we show the interesting connection between this design and delay minimization. For $2 \leq \alpha$ and $\bar{\nu^{\mathrm{d}}} = \alpha - 1$, the network utility becomes

$$S_{\mathrm{net}}^{(\bar{\mathbf{q}},\bar{\mathbf{r}})}\Big(\bar{\mathbf{q}}[k],\bar{\mathbf{r}}[k]\Big) = -\frac{1}{\bar{\nu^{\mathrm{d}}}}\sum_{\phi=1}^{|\Phi_{\mathrm{RT}}|}\left(\frac{\bar{q}_{\phi}[k]}{\bar{r}_{\phi}[k]}\right)^{\bar{\nu^{\mathrm{d}}}}, \quad \text{where} \quad \frac{\bar{q}_{\phi}[k]}{\bar{r}_{\phi}[k]} \quad \text{is equal to} \quad \bar{d}_{\phi}[k]. \tag{2.47}$$

Note that for $\alpha \geq 2$, the constant factor, $1/(1-\alpha)$, makes the maximization, a minimization. This positive constant is irrelevant to the optimization.

Since the mean incoming bit-rate is equal to mean outgoing bit-rate for each flow, mean bit-rate

$\bar{r}_\phi[k]$ is proportional to the estimate ($\hat{\Lambda}_\phi[k]$) of actual arrival bit-rate (Λ_ϕ) as

$$\bar{r}_\phi[k] \triangleq \frac{\sum_{k'=1}^{k} r_\phi[k']}{k} \approx \frac{1}{T_b}\hat{\Lambda}_\phi[k], \quad \text{where } \hat{\Lambda}_\phi[k] \text{ is an estimate of } \quad \Lambda_\phi = \lim_{k\to\infty} \frac{\sum_{k'=1}^{k} A_\phi[k']}{k}, \tag{2.48}$$

in frame k, where $r_\phi[k]$ is the frame bit-rate for flow ϕ in frame k, $A_\phi[k]$ is the arrival bits in frame k, and $\bar{r}_\phi[k]$ is the mean bit-rate received by flow ϕ until frame k.

With the ergodicity assumption for queue lengths, the mean length of queue, ϕ, $\bar{q}_\phi[k]$, is the multiplication of the estimated arrival bit-rate, $\hat{\Lambda}_\phi[k]$, by the mean latencies in that queue, $\bar{l}_\phi[k]$. This can be explained by the Little's law as

$$\bar{q}_\phi[k] = \bar{l}_\phi[k]\ \hat{\Lambda}_\phi[k] \quad \text{or} \quad \bar{q}_\phi[k] = \bar{l}_\phi[k]\ \bar{r}_\phi[k]. \tag{2.49}$$

Therefore, in fact $\bar{d}_\phi[k]$ is the mean delay, equal to $\bar{l}_\phi[k]$.

Having designed a WGPF with daly consideration, we make a more general formulation based on delay disutility. Each flow is associated with a disutility as $D_\phi^{\bar{\mathrm{d}}}(\bar{d}_\phi[k])$ in frame k, where the disutility is a convex increasing function and $\bar{d}_\phi[k]$ is an estimation of the mean-delay at the end of frame k. The network objective is, then, to minimize the total disutility, $D_{\mathrm{net}}^{\bar{\mathrm{d}}}\left(\bar{\mathbf{d}}[k]\right)$, as

$$\min_{x_\phi^{(j)}[k]\in\mathcal{C}^{\mathrm{PHY}}} D_{\mathrm{net}}^{\bar{\mathrm{d}}}\Big(\bar{\mathbf{d}}[k]\Big), \quad \text{or} \quad \min_{x_\phi^{(j)}[k]\in\mathcal{C}^{\mathrm{PHY}}} \sum_{\phi=1}^{|\Phi_{\mathrm{RT}}|} D_\phi^{\bar{\mathrm{d}}}\Big(\bar{d}_\phi[k]\Big), \tag{2.50}$$

where the constraints are induced by PHY limitation of RBs in a frame, as before. Minimizing the above objective, with

$$D_\phi^{\bar{\mathrm{d}}}\left(\bar{d}_\phi[k]\right) = \Big(\bar{d}_\phi[k]\Big)^{\nu^{\bar{\mathrm{d}}}} \Big/ \nu^{\bar{\mathrm{d}}} \tag{2.51}$$

is equivalent to WGPF maximization. This connection explains the relationship between bit-rate utility and mean-delay disutility. We note that our delay minimization approach is also equivalent to minimization of delay violation probability in heavy traffic situation in [57].

Mean Delay Estimate as a Function of Frame Bit-rate

In the next section, we derive an estimate for mean delay, needed for studying the daly fairness. Since queue lengths at the end of each frame are unknown, we need an estimate for queue lengths. Across the frames the queue lengths are flactuating based on the arrival bits, $A_\phi[k]$, and the received service, $r_\phi[k]$, as

$$q_\phi[k] = \max\Big[q_\phi[k-1] - T_b\ r_\phi[k] + A_\phi[k], 0\Big]. \tag{2.52}$$

An estimate for mean queue lengths at the end of frame k can be achieved with considering the effect of allocation on the expected value of queue length at the end of frame k with respect to arrivals. The arrival bit-rates, in frame k, is estimated with mean bit-rate at the end of frame $k-1$. Then, the estimate for $\bar{d}_\phi[k]$, which is denoted by $\tilde{d}_\phi[k]$, is

$$\tilde{d}_\phi[k] = \frac{\big(1-\frac{1}{k}\big)\,\bar{q}_\phi[k-1] + \frac{1}{k}\Big(q_\phi[k-1] - r_\phi[k] + \mathbb{E}(A_\phi[k])\Big)}{\big(1-\frac{1}{k}\big)\,\bar{r}_\phi[k-1] + \frac{1}{k} r_\phi[k]}. \tag{2.53}$$

Separating the constant factors and the optimization variable factors in frame k, we get

$$\tilde{d}_\phi[k] = \frac{\big(\frac{k-2}{k}\big)\,\bar{q}_\phi[k-2] + \frac{2}{k} q_\phi[k-1] - \frac{k-1}{k^2}\Big(r_\phi[k] - \bar{r}_\phi[k-1]\Big)}{\big(1-\frac{1}{k}\big)\,\bar{r}_\phi[k-1] + \frac{1}{k} r_\phi[k]} = \frac{\zeta_\phi^{(1)}[k] - \zeta_\phi^{(2)}[k] r_\phi[k]}{\zeta_\phi^{(3)}[k] + \zeta_\phi^{(4)}[k] r_\phi[k]}, \tag{2.54}$$

where $\zeta_\phi^{(1)}[k], \zeta_\phi^{(2)}[k], \zeta_\phi^{(3)}[k],\ \&\ \zeta_\phi^{(4)}[k]$ are constants, independent of current frame allocation. The constants $\zeta_\phi^{(1)}[k]$ & $\zeta_\phi^{(3)}[k]$ are based on the past allocations and $\zeta_\phi^{(2)}[k]$ & $\zeta_\phi^{(4)}[k]$ are functions of k. This estimate, of the mean delay, is a decreasing convex bi-linear function of the frame bit-rate. Since the mean delay estimates are decreasing convex functions and disutility functions are increasing convex functions, the summation of the substitution compositions in (2.50), with real variable relaxation, is a convex decreasing function [50].

Gradient based Core for the Mean-delay Fairness

Having formulated the delay disutility minimization, now, we discuss the gradient based algorithm for it. The maximum change in the objective can be achieved in the direction of gradient. Based

on the chain rule the gradient can be decomposed as

$$\frac{\partial D_{\text{net}}^{\bar{\text{d}}}\left(\bar{\mathbf{d}}[k]\right)}{\partial x_{\phi}^{(j)}[k]} = \frac{\partial D_{\text{net}}^{\bar{\text{d}}}(\bar{\mathbf{d}}[k])}{\partial \bar{d}_{\phi}[k]} \frac{\partial \bar{d}_{\phi}[k]}{\partial r_{\phi}[k]} \frac{\partial r_{\phi}[k]}{\partial x_{\phi}^{(j)}[k]}, \tag{2.55}$$

where each component is further specified by

$$\frac{\partial D_{\text{net}}^{\bar{\text{d}}}\left(\bar{\mathbf{d}}[k]\right)}{\partial \bar{d}_{\phi}[k]} = \left(\bar{d}_{\phi}[k]\right)^{\nu^{\bar{\text{d}}}-1}, \quad \frac{\partial \bar{d}_{\phi}[k]}{\partial r_{\phi}[k]} = -\frac{\zeta_{\phi}^{(1)}[k]\zeta_{\phi}^{(4)}[k] + \zeta_{\phi}^{(2)}[k]\zeta_{\phi}^{(3)}[k]}{\left(\zeta_{\phi}^{(3)}[k] + \zeta_{\phi}^{(4)}[k] r_{\phi}[k]\right)^2}, \quad \text{and} \quad \frac{\partial r_{\phi}[k]}{\partial x_{\phi}^{(j)}[k]} = b_{\phi}^{(j)}[k]. \tag{2.56}$$

Since the overall objective is a decreasing function in terms of scheduling variables, the derivatives are negative. Therefore, for finding the smallest gradient, it is sufficient to find the largest absolute value of the gradient, which finds the flow and the associated RB based on the largest absolute value of the gradient with respect to the optimization variable as

$$\left(j^*[k], \phi^*[k]\right) = \arg\max_{(j,\phi)} \left| \frac{\partial D_{\text{net}}^{\bar{\text{d}}}\left(\bar{\mathbf{d}}[k]\right)}{\partial x_{\phi}^{(j)}[k]} \right|. \tag{2.57}$$

2.4.2 Mean-delay Minmax Fairness

Similar to maxmin bit-rate fairness as an asymptotic version of GPF, a minmax mean delay counterpart is the asymptotic version of the aforementioned WGPF for delay fairness. For (2.51) and $1 \leq \nu^{\bar{\text{d}}}$, the scheduler can be adjusted between sum mean delay minimizer and minmax delay objective. The term minmax mean-delay refers to the case when any change in the allocation cannot result in reducing the maximum mean-delay among flows. The equivalent explanation is that no mean-delay $\bar{d}_{\phi_0}$ can be reduced without increasing the other mean delay less than $\bar{d}_{\phi_0}$.

For $\nu^{\bar{\text{d}}} = 1$, the sum mean delay minimization is the straight result of substitution. The following propositions prove the delay minmax case. The **Proposition 3** shows that in the asymptotic case the network objective (2.50) corresponds to the *minmax* mean-delay objective.

Proposition 3. *For (2.51), as $\nu^{\bar{\text{d}}} \to \infty$, the network disutility leads to a minmax delay-fair allocation.*

The next proposition, **Proposition 4**, shows how the scheduler can be simplified in asymptotic

case for sufficiently large $\nu^{\bar{\mathrm{d}}}$.

Proposition 4. *For a sufficiently large $\nu^{\bar{\mathrm{d}}}$, assigning a RB to the flow with the maximum current mean-delay, in the iterations in the frame k, on its best sub-channel is equivalent to assigning RB to the flow with the smallest gradient.*

We proved that increasing parameter $\nu^{\bar{\mathrm{d}}}$ in the proposed adjustable scheduler leads to minmax objective for mean delays. This proof rationalizes a simple intuitive mechanism of serving the worst flow, in terms of mean delay, and connects it to the gradient-based algorithms. The idea of selecting the flow with the maximum mean delay is the traffic-aware counterpart of the selecting the flow with the minimum bit-rate in each iteration which was used in [17, 34, 58].

2.4.3 HOL-delay

Formally, the HOL-delay, in each frame, is defined as the difference of the current frame index, k, and the arrival time-stamp frame index of HOL packet, $t_{\phi}^{\mathrm{HOL}}[k]$, as

$$d_{\phi}^{\mathrm{HOL}}[k] = k - t_{\phi}^{\mathrm{HOL}}[k]. \tag{2.58}$$

Similar to mean delay, HOL delays are also updated frame-by-frame as the result of the frame allocation, the arrivals and their previous states. The $d_{\phi}^{\mathrm{HOL}}[k]$ is time the HOL packets are delayed until frame k, in the scale of the frame time span. Accordingly, the HOL packets are delayed equal to $T_b\ d_{\phi}^{\mathrm{HOL}}[k]$ in seconds, until frame k. In frame k, the timestamp of the arrivals (if there is arrival in frame k or $A_{\phi}[k] \neq 0$) is

$$t_{\phi}^{(l)}[k] = k, \quad \text{for} \quad l \quad \text{such that} \quad q_{\phi}[k-1] - T_b\, r_{\phi}[k] \ < \ l \ \leq \ q_{\phi}[k-1] - T_b\, r_{\phi}[k] + A_{\phi}[k], \tag{2.59}$$

where $t_{\phi}^{(l)}[k]$ denotes the timestamp of the l-th bits ($l \in [1, q_{\phi}[k]]$) and the condition is targeting the new arrival in the queue. In addition, $r_{\phi}[k]$ is the output bits from flow ϕ in frame k (results of decisions in frame k) and $A_{\phi}[k]$ is the arrival in bits to the flow ϕ in frame k.

Note that $t_{\phi}^{(l)}[k]$ is equal to the former used notation as $t_{\phi}^{\mathrm{HOL}}[k]$, for $1 \leq l \leq$. Since the HOL-delay plays an special rule, we used an special notation for it. Similarly, $d_{\phi}^{(l)}[k]$ is defined as the

l-bits delay for flows ϕ, where $d_\phi^{\mathrm{HOL}}[k]$ is defined as $d_\phi^{\left(q_\phi[k]\right)}[k]$. Accordingly,

$$d_\phi^{(l)}[k] = \begin{cases} d_\phi^{\left(l+r_\phi[k]\right)}[k-1]+1, & \text{if} \quad 0 \ \leq \ l \ \leq \ q_\phi[k-1] - T_b \, r_\phi[k], \\ 0, & \text{if} \quad q_\phi[k-1] - T_b \, r_\phi[k] \ < \ l \ \leq \ q_\phi[k-1] - T_b \, r_\phi[k] + A_\phi[k], \end{cases} \tag{2.60}$$

In other words, the delay of the new arrivals will be equal to 0 and the delay of the not transmitted packets will be increased by 1 (in frame index scale). Arrival is considered in the beginning of each frame leading to possible zero delay.

2.4.4 HOL-delay Fairness through WGPF

Similar to mean-delay fairness, with the selection of disutility over HOL-delay as

$$D_\phi^{\mathrm{d^{HOL}}}\left(d_\phi^{\mathrm{HOL}}[k]\right) = \left(d_\phi^{\mathrm{HOL}}[k]\right)^{\nu^{\mathrm{d^{HOL}}}} \Big/ \bar{r}_\phi[k], \text{ for } \phi \in \Phi_{\mathrm{RT}}, \tag{2.61}$$

our framework will be reduced to HOL-delay fairness proposed in [59], similar to mean-delay fairness in [12]. The same observation in [12] for trade-off between delay fairness and throughput (or equivalently resource efficiency) can be observed for HOL-delay fairness, by controlling the parameter $\nu^{\mathrm{d^{HOL}}}$ ($1 \leq \nu^{\mathrm{d^{HOL}}}$) [60]. Reference [59] used the HOL-delay fairness and GPF mechanism sequentially (or with static separation) for RT and NRT packet scheduling and resource allocation.

2.4.5 Delay Fairness through Jain's Index Maximization for Delay

In previous sections based on [12], we presented the parametric utility delay fairness. This approach is to allocate the resources in a way that maximizes a parametric utility to control the tradeoff of the efficiency and fairness, in terms of delay. However, increasing parameter $\nu^{\bar{\mathrm{d}}}$ results in allocation which has high fairness in a sense that it does not necessarily conform to the Jain's index of delay profile of flows, as it will be shown hereinafter.

Another approach for achieving delay fairness is suggested as maximizing the Jain's index of mean-delay extending [47]. Similar to delay fairness based on WGFP in the previous sections, maximizing the Jain's index of delay is a fairness notion for finite-backlog scenario. This is in

contrast to bit-rate fairness which is for infinite backlog scenario. Based on this approach the following core is suggested as

$$\Big(\phi^*[k], j^*[k]\Big) = \max_{(\phi,j)} b_\phi^{(j)}[k]\Big(\bar{d}_\phi[k-1] - \beta^{\bar{\mathrm{d}}}\Big), \tag{2.62}$$

where $\beta^{\bar{\mathrm{d}}}$ is the parameter which changes the tradeoff between delay fairness and efficiency.

Delay Fairness through Jain's Index Maximization for Delay with General Disutility

Similar to Section 2.3.4, for maximizing the Jain's index on the general disutility of the delay the following scheduling core is suggested as

$$\Big(\phi^*[k], j^*[k]\Big) = \max_{(\phi,j)} b_\phi^{(j)}[k]\Big(D_\phi^{\bar{\mathrm{d}}}\left(\bar{d}_\phi[k-1]\right) - \beta^{\mathrm{D}^{\bar{\mathrm{d}}}}\Big), \tag{2.63}$$

where $\beta^{\mathrm{D}(\bar{\mathrm{d}})}$ is the parameter which changes the tradeoff between delay fairness and efficiency, in this case.

2.4.6 Maximum Mean-delay QoS Guarantee through Lagrangian Multipliers

Similar to guaranteeing minimum bit-rate, Lagrangian method can be used for guaranteeing the maximum of the mean delay. Reference [54] considers mean delay based utility maximization subject to the mean delay constraints defined as

$$\mathcal{C}^{\bar{\mathrm{d}}^{\max}} = \Big\{ x_\phi^{(j)}[k] \Big| \quad \forall\, \phi:\ \bar{d}_\phi[k] \le \bar{d}_\phi^{\max} \Big\}. \tag{2.64}$$

Then, the OFDMA version of the gradient scheduling algorithm in [54] can be described as

$$\Big(\phi^*[k], j^*[k]\Big) = \arg\max_{(\phi,j)} \left(\lambda_\phi^{\bar{\mathrm{d}}^{\max}}[k] + \frac{\partial D_\phi^{\bar{\mathrm{d}}}\left(\bar{d}_\phi[k]\right)}{\partial r_\phi[k]} \right) \frac{\partial r_\phi[k]}{\partial x_\phi^{(j)}[k]}. \tag{2.65}$$

Similar to guaranteeing the minimum bit-rate in Section 2.3.7, intuitively, the flows with low AMC values, but low mean delay requirement, are assisted through the addition of the Lagrangian multiplier. Likewise, the corresponding Lagrangian multipliers are determined by approximation

of the sub-gradient projection as

$$\lambda_{\phi}^{\bar{\mathrm{d}}^{\max}}[k+1] = \max\left[0, \lambda_{\phi}^{\bar{\mathrm{d}}^{\max}}[k] - \hbar[k]\left(\bar{d}_{\phi}^{\max} - \bar{d}_{\phi}[k+1]\right)\right]. \tag{2.66}$$

It is worth mentioning that when only part of the RB is required to serve all the data in selected flow queue, the remaining capacity of the RB will be assigned to the flow with the second highest argument in the arg max. This process continue until the all the RBs are assigned to flows or all the all queues become empty. Similar to its bit-rate counterpart, minimum bit-rate guarantee in Section 2.3.7, when the maximum mean delay requirement is high, the Lagrangian multiplier becomes zero, not assisting the corresponding flow.

Mean-delay versus HOL-delay

Interestingly, guaranteeing maximum mean-delay makes a bound on the HOL-delay outage. This can be explained in terms of the Markov inequality [61, sec. 4.6]. Based on the Markov inequality, the mean of a random variable makes an upper bound on its complementary-CDF. Therefore, guaranteeing the maximum mean-delay makes an upper bound on the outage of HOL-delay. Applying the Markov inequality implies that

$$\mathbf{Pr}\left(d_{\phi}^{\mathrm{HOL}}[k] > d_{\phi}^{\mathrm{HOL}^{\max}}\right) \leq \frac{\mathbf{E}\left(d_{\phi}^{\mathrm{HOL}}[k]\right)}{d_{\phi}^{\mathrm{HOL}^{\max}}}. \tag{2.67}$$

Now, if we estimate $\bar{d}_{\phi}[k]$ with $\mathbf{E}\left(d_{\phi}^{\mathrm{HOL}}[k]\right)$ and assume to have a bound on the mean-delay as $\bar{d}_{\phi}[k] < \bar{d}_{\phi}^{\max}$, then

$$\mathbf{Pr}\left(d_{\phi}^{\mathrm{HOL}}[k] > d_{\phi}^{\mathrm{HOL}^{\max}}\right) \leq \frac{\mathbf{E}\left(d_{\phi}^{\mathrm{HOL}}[k]\right)}{d_{\phi}^{\mathrm{HOL}^{\max}}} \leq \frac{\bar{d}_{\phi}^{\max}}{d_{\phi}^{\mathrm{HOL}^{\max}}}. \tag{2.68}$$

This suggests that to control maximum HOL-delay, we can control the maximum mean-delay, indicated by $\bar{d}_{\phi}^{\max} = d_{\phi}^{\mathrm{HOL}^{\max}}\delta_{\phi}$. However, it should be noted that this approach is an indirect approach for HOL-delay control which its resulting bound on the HOL delay is not tight.

Lagrange Multiplier Assisted Admission Controller

The Lagrangian multipliers, associated with the minimum bit-rate constraint and maximum delay constraint formulation, are also useful in admission controller (AC) decision making, whether to admit additional flows, or drop flows in accordance to their priorities as resources are consumed. Intuitively, the Lagrange multiplier values show the slackness of their corresponding constraints [50]. The Lagrange multipliers are relatively small, if resources are more than enough, signalling feasibility of admitting additional flows. In other hands, the Lagrange multipliers start to become relatively large, if the resources are not enough to render QoS-requirements, signalling that low priority flows should be dropped, in order to maintain higher priority QoS-requirements feasible.

As a conceptual guide, in [62], the AC is proposed to have different threshold on Lagrangian multipliers for each class of traffic. However, we note that the dropping can be implemented automatically in the RRM design, without separate AC mechanism, similar to [14].

The Lagrange multipliers are also suggested to be used as a guideline for dynamic pricing of the resources. When initiating a flow, the BS negotiates with the corresponding WT to agree on a price, for utilizing the network resources, based on the Lagrange multiplier. When a Lagrange multiplier grows, the BS increase the corresponding price [62].

2.4.7 Maximum HOL QoS Control through Earliest Deadline First

The simplest mechanism for controlling HOL delay is the earliest deadline first (EDF) [63]. The EDF approach is based on serving the flow with the earliest deadline. The EDF is described by

$$\Big(\phi^*[k], j^*[k]\Big) = d_\phi^{\mathrm{HOL}}[k] - d_\phi^{\mathrm{HOL}^{\max}}, \tag{2.69}$$

which is a non channel aware scheduler. A channel aware version can be considered as

$$\Big(\phi^*[k], j^*[k]\Big) = b_\phi^{(j)}[k]\Big(d_\phi^{\mathrm{HOL}}[k] - d_\phi^{\mathrm{HOL}^{\max}}\Big), \tag{2.70}$$

where it prioritize the flow with higher transmission efficiencies.

2.4.8 Maximum HOL QoS Guarantee through Modified Largest Weighted Delay First

A more advanced mechanism, in comparison to EDF, for controlling HOL delay for RT flows, is MLWDF [56,64]. In each frame k, a flow $\phi^*[k]$ is selected to be transmitted on an RB on sub-channel $j^*[k]$ according to

$$\Big(\phi^*[k], j^*[k]\Big) = \frac{b_\phi^{(j)}[k]}{\bar{r}_\phi[k]}\, \mu_\phi\, d_\phi^{\mathrm{HOL}}[k], \quad \phi \in \Phi_{\mathrm{RT}}, \tag{2.71}$$

where μ_ϕ is suggested to be selected as

$$\mu_\phi = -\frac{\log \delta_\phi}{\tau_\phi}, \tag{2.72}$$

as a result of large deviation optimality in [65], $\bar{r}_\phi[k]$ is the mean bit-rate in frame k, τ_ϕ is the maximum allowable delay threshold, and δ_ϕ is a maximum probability of exceeding the delay threshold [66] as

$$\Pr\left(d_\phi^{\mathrm{HOL}}[k] > \tau_\phi\right) < \delta_\phi. \tag{2.73}$$

It is worth mentioning that by fixing a close-to-one percentile for $1-\delta_\phi$ (or equivalently small δ_ϕ), we can interpret τ_ϕ as the delay deadline which we denoted earlier by $d_\phi^{\mathrm{HOL}^{\max}}$. As an example, for $1-\delta_\phi$ equals to the 99^{th} percentile, we have $\mu_\phi \approx 2/d_\phi^{\mathrm{HOL}^{\max}}$. By this interpretation, it can be noticed that the HOL-delays are divided by their deadlines in the structure of MLWDF (see (2.71) and (2.72)). It is in contrast to EDF, where the difference of HOL-delays and their deadlines forms the structure.

Intuitive Interpretation of the Parameter μ_ϕ

Parameter μ_ϕ embodies the different HOL-delay requirements among flows. For example, if flows 1 and 2 have the same desired delay thresholds $\tau_1 = \tau_2$, but the desired maximum HOL delay violation probability for flow 2, δ_2, is four times less than that of flows 1, δ_1, then $\mu_2 = 2\mu_1$, and therefore, flow $\phi = 2$ is treated with higher priority over flow $\phi = 1$. The greater the flow HOL-delay, or the higher the channel quality relative to its average level, or the higher the HOL-delay requirement,

prioritize the scheduling of the associated flow. The MWLDF rule approximately balances different probabilities of the deadline violation, relative to their maximum allowed violation values. It is shown in [67] that the policy which minimizes the total packets in the system (with Bernoulli i.i.d. arrival and Bernoulli on & off channel model), is the one that serves the flow whose channel is on and has the longest queue length which resembles MWLDF.

Packet Loss Ratio

In packet-switching networks, packet loss is inevitable. In addition to bit-rate and delay, packet loss ratio (PLR) is another QoS measure. Among the scheduling cores which are discussed in this chapter, MWLDF and exponential-rule (EXP) cores are able to control the PLR, besides other elements. In fact, in the structure of MLWDF and EXP, the maximum PLR is embedded in μ_ϕ as δ_ϕ, which we denoted earlier by $\epsilon_\phi^{\max}$. A packet loss may happen due to the several reasons including link error or passed delay deadline. Generally, when a packet arrives, it is timestamped and placed in the packet queue, in addition to a unique associated virtual token with the same timestamp. This token, with the same timestamp, will be preserved until either the acknowledge message for receiving the packet is received, or the HOL-delay (for RT flows) of the packet is passed. Other origins of the packet loss may include overflow occurrence, or the retransmission timeout of automatic repeat request (ARQ). Basically, the delay exceeding probability is kept around 10^{-2} by the scheduler. Then, ARQ reduces it to order value of 10^{-6}. Finally, TCP retransmits if a packet is lost or delayed beyond its expectation.

More generally, there can be a packet loss counter in the system which counts the occurrences of losses, until frame k. The packet loss counters are updated frame-by-frame. Then, based on the packet loss counter, the empirical PLR is calculated and is passed through the PLR-disutility function, which is an *increasing function* of the empirical PLR and acts as a weight emphasizing or de-emphasizing on serving a flow. If PLR for certain flow increases beyond its requirements, the PLR-disutility will also be increased and motivate the core to serve this flow with higher priority until its empirical PLR is corrected to conform the PLR requirements [15].

2.4.9 Maximum HOL QoS Guarantee through Exponential-rule

Another mechanism called exponential rule or EXP [66, 68] has been also proposed for resource allocation of RT flows with a similar structure in MLWDF, but with different dependency on HOL-delay. An OFDMA version of EXP can be represented as

$$\Big(\phi^*[k], j^*[k]\Big) = \arg\max_{(\phi,j)} \ \frac{b_\phi^{(j)}[k]}{\bar{r}_\phi[k]} \mathrm{e}^{\left(\frac{\mu_\phi d_\phi^{\mathrm{HOL}}[k] - \overline{d^{\mathrm{HOL}}}[k]}{1+\left(\overline{d^{\mathrm{HOL}}}[k]\right)^\eta}\right)}, \tag{2.74}$$

where in each frame k, a flow $\phi^*[k]$ is selected to be transmitted on an RB on sub-channel $j^*[k]$, the exponent $0 < \eta < 1$ is making the delay equalization, μ_ϕ is defined the same as in (2.72) (inversely proportional to delay deadline), and

$$\overline{d^{\mathrm{HOL}}}[k] = \frac{1}{|\Phi_{\mathrm{RT}}|} \sum_{\phi=1}^{|\Phi_{\mathrm{RT}}|} d_\phi^{\mathrm{HOL}}[k] \tag{2.75}$$

is the average of HOL-delays.

EXP equalizes the HOL-delays. If a flow HOL-delay would have a larger delay more than its deadline $d_\phi^{\mathrm{HOL}^{\max}}$, then the exponent term becomes very large and override the channel awareness (as long as it is not zero), hence leading to that flow getting priority. Note that the term $\overline{d^{\mathrm{HOL}}}[k]$ in the exponent numerator can be dropped without changing the rule. However, the presence of this term makes the explanation of the structure more intuitive. The constant term 1 in the denominator is to prevent the overall quantity to be unbounded.

Note that the expression inside the arg max in (2.71) and (2.74) can be interpreted as gradient of the disutility. Therefore, MLWDF can be considered as a gradient scheduling, with quadratic utility, $\mu_\phi(\bar{d}_\phi[k])^2/2$. Likewise, EXP can be casted as a gradient scheduling. However, its actual corresponding optimization is more sophisticated. The EXP is an online algorithm that can stabilize the queues without explicit knowledge of the arrival bit-rates or channel statistics [68]. As discussed earlier, MLWDF and EXP, in conjunction with virtual token queues (with constant deterministic arrival rate), can be used to implement minimum bit-rate in resource allocation and

packet scheduling [51, 69].

Queue Length Driven Versions of MWLDF and EXP

The HOL-delay term, $d_\phi^{\mathrm{HOL}}[k]$, in (2.71) and (2.74) can be replaced with queue-length term as $q_\phi[k]$ to obtain the queue-length-driven versions of aforementioned schedulers. However, there is often significant difference between queue-length optimality and delay optimality. In fact, there exist arrival patterns such that algorithms with small queue backlog can still lead to large delay. When queue lengths of the flows are equal or close to each other (or equivalently the delay differences are small; see [70] for its formal definition), EXP and MLWDF reduce to PF scheduler [70]. It is worth mentioning that EXP is suitable for the cases where the delay equalization is preferable. However, there has been analyses [71] showing that this has a cost. In fact, EXP sacrifices the asymptotic system throughput when the queues grow asymptotically. This is the cost of the delay equalization. The structures of EXP and MLWDF are based on division of HOL-delay by its deadline, while the structure of EDF is based on the difference of HOL-delay and its deadline.

Bit-rate Timescale

The mean bit-rate (See (2.6)) in the packet scheduling and RB allocation core can be replaced by a mixed timescale of mean bit-rate [13, 72]. In other words, the exact definition share may depends on the application, as different applications may require averaging over different time duration or timescales. The mixed timescale introduces refined definition of fairness that take into account the time horizon (duration) over which the mean bit-rates are calculated. The mixed timescale of bit-rates is denoted by $\hat{r}_\phi[k]$ in this chapter.

Intercell Interference Assumptions

It is worth highlighting that the scheduling and allocation cores, presented throughout this chapter, have been designed based on the static interference assumption. In fact, the intercell interference coordination (ICIC) [73] works in a longer timescale than the scheduling and allocation algorithm to specify which RB should be muted or de-muted for each cell or sector. Since the ICIC schemes

often do not aim for sophisticated QoS-requirements or sophisticated fairness notions, integrating ICIC right into the scheduling and allocation core, through a systematic design, can be suggested as a promising future direction.

2.5 Discussions and Conclusion

2.5.1 Classification of the Scheduling and Allocation Cores based on the QoS-requirements and the Embedded Fairness Notion

In this section, we make a summary of the discussed scheduling cores and a classification based on the traffic type and the requirements, the cores have been designed for. We are motivated to find an efficient way for classifying the relevant packet scheduling and RB allocation algorithms which is also suitable for a text book. The packet scheduling and RB allocation algorithms for RT and NRT flows are summarized and classified in Table 2.3 and Table 2.2, respectively. This classification offers an efficient way to present algorithm for both educational and research purposes and reveals the interrelations among many previous works. Based on this classification, the common and distinguishing structural patterns in the cores can be decomposed conceptually. Inspired by the classification, in the next section, we decompose conceptually the structure of the scheduling core and connect different decomposed components to their properties.

A typical set of order values of the QoS-requirements, for a number of packet-switched flow types, is summarized in Table 2.4 based on 3GPP specifications in [1, p. 37].

Table 2.2: Classification of the discussed packet scheduling and RB allocation cores for NRT flows.

Flow type	**Requirements**	**Description**	**Core**
NRT flows	Bit-rate GPF	Sum bit-rate max [29, 34]	$\max\limits_{(\phi,j)} b_\phi^{(j)}[k]$
		PF [39, 40, 42]	$\max\limits_{(\phi,j)} \frac{b_\phi^{(j)}[k]}{\bar{r}_\phi[k]}$
		Bit-rate maxmin [17, 31]	$\phi^* \leftarrow \arg\max\limits_{\phi} \frac{1}{\bar{r}_\phi[k]}, \quad \max\limits_{j} b_{\phi^*}^{(j)}[k]$
		GPF with generalized time-scale [13, 18]	$\max\limits_{(\phi,j)} \frac{b_\phi^{(j)}[k]}{\left(\hat{r}_\phi[k]\right)^\alpha}$
		Utility fairness with GPF [Novel]	$\max\limits_{(\phi,j)} \frac{b_\phi^{(j)}[k]}{S_\phi^{\bar{\mathrm{r}}}\left(\bar{r}_\phi[k]\right)} \frac{\partial S_\phi^{\bar{\mathrm{r}}}\left(\bar{r}_\phi[k]\right)}{\partial \bar{r}_\phi[k]}$
	Jain's index for bit-rate	EJT [47]	$\max\limits_{(\phi,j)} b_\phi^{(j)}[k]\left(\beta^{\bar{\mathrm{r}}} - \bar{r}_\phi[k-1]\right)$
		EJT with generalized time-scale [Novel]	$\max\limits_{(\phi,j)} b_\phi^{(j)}[k]\left(\beta^{\hat{\mathrm{r}}} - \hat{r}_\phi[k-1]\right)$
		Utility fairness with EJT [Novel]	$\max\limits_{(\phi,j)} b_\phi^{(j)}[k]\left(\beta^{\mathrm{S}^{\bar{\mathrm{r}}}} - S_\phi^{\bar{\mathrm{r}}}(\bar{r}_\phi[k-1])\right)$
	Min bit-rate guarantee	Min bit-rate with EXP [55, 56]	$\max\limits_{(\phi,j)} \frac{b_\phi^{(j)}[k]}{\bar{r}_\phi[k]} \mathrm{e}^{\left(\frac{\mu_\phi d_\phi^{\mathrm{HOL}}[k]}{1+\left(\overline{d^{\mathrm{HOL}}}[k]\right)^\eta}\right)}$
		Min bit-rate with MWLDF [55, 56]	$\max\limits_{(\phi,j)} \frac{b_\phi^{(j)}[k]}{\bar{r}_\phi[k]} \mu_\phi d_\phi^{\mathrm{HOL}}[k]$
		Min & max bit-rate with virtual token [51, 52]	$\max\limits_{(\phi,j)} \mathrm{e}^{\lambda_\phi^{\mathrm{token}}[k] \frac{\partial S_\phi^{\bar{\mathrm{r}}}\left(\bar{r}_\phi[k]\right)}{\partial \bar{r}_\phi[k]}} \frac{\partial \bar{r}_\phi[k]}{\partial x_\phi^{(j)}[k]}$
		Min bit-rate with the Lagrangian multiplier [54]	$\max\limits_{(\phi,j)} \left(\lambda_\phi^{\bar{\mathrm{r}}^{\min}}[k] + \frac{\partial S_\phi^{\bar{\mathrm{r}}}\left(\bar{r}_\phi[k]\right)}{\partial \bar{r}_\phi[k]}\right) \frac{\partial \bar{r}_\phi[k]}{\partial x_\phi^{(j)}[k]}$

Table 2.4: Example of min bit-rate, max HOL-delay, and max PLR requirements [1, p. 37].

Row	Description	$\mathbf{QoS}_\phi^{\mathrm{class}}$	$d_\phi^{\mathrm{HOL}^{\max}}$	$\bar{r}_\phi^{\min}$	$\epsilon_\phi^{\max}$
1	Conversational voice	RT	100 ms	32-128 kbps	10^{-2}
2	Non-conv. voice	RT	300 ms - 1 s	48 kbps	10^{-3}
3	Conversational video	RT	150 ms	128 kbps - 2 Mbps	10^{-3}
4	Non-conv. video	RT	300 ms - 1 s	128 kbps - 10 Mbps	10^{-6}
5	Interactive browsing	RT	1 s	n/a	10^{-8}
6	FTP	NRT	n/a	depends	10^{-8}
7	Email & Passive browsing	NRT /BE	n/a	n/a	10^{-8}

Table 2.3: Classification of packet scheduling and RB allocation cores for RT flows.

Flow type	Requirements	Description	Core
RT flows	Delay aware WGPF	Mean-delay aware WGPF [12]	$\max_{(\phi,j)} b_{\phi}^{(j)}[k]\left(\bar{d}_{\phi}[k]\right)^{\nu^{\bar{\mathrm{d}}}}$
		HOL-delay aware WGPF [60]	$\max_{(\phi,j)} b_{\phi}^{(j)}[k]\left(d_{\phi}^{\mathrm{HOL}}[k]\right)^{\nu^{\mathrm{d}^{\mathrm{HOL}}}}$
		Delay utility fairness with WGPF [Novel]	$\max_{(\phi,j)} b_{\phi}^{(j)}[k] D_{\phi}^{\bar{\mathrm{d}}}\left(\bar{d}_{\phi}[k]\right) \frac{\partial D_{\phi}^{\bar{\mathrm{d}}}\left(\bar{d}_{\phi}[k]\right)}{\partial \bar{d}_{\phi}[k]}$
	Jain index for delay	EJT for mean-delay [Novel]	$\max_{(\phi,j)} b_{\phi}^{(j)}[k]\left(\bar{d}_{\phi}[k-1]-\beta^{\bar{\mathrm{d}}}\right)$
		EJT for HOL-delay [Novel]	$\max_{(\phi,j)} b_{\phi}^{(j)}[k]\left(d_{\phi}^{\mathrm{HOL}}[k-1]-\beta^{\mathrm{d}^{\mathrm{HOL}}}\right)$
		Utility fairness with EJT for delay [Novel]	$\max_{(\phi,j)} b_{\phi}^{(j)}[k]\left(D_{\phi}^{\bar{\mathrm{d}}}\left(\bar{d}_{\phi}[k-1]\right)-\beta^{\mathrm{D}^{\mathrm{d}}}\right)$
	Max delay guarantee	Max delay guarantee with EDF [63]	$\max_{(\phi,j)} d_{\phi}^{\mathrm{HOL}}[k]-d_{\phi}^{\mathrm{HOL}^{\max}}$
		Max HOL-delay with EXP [56, 68]	$\max_{(\phi,j)} \frac{b_{\phi}^{(j)}[k]}{\bar{r}_{\phi}[k]} \mathrm{e}^{\left(\frac{\mu_{\phi} d_{\phi}^{\mathrm{HOL}}[k]}{1+\left(\overline{d^{\mathrm{HOL}}}[k]\right)^{\eta}}\right)}$
		Max HOL-delay guarantee with MWLDF [56, 64, 66]	$\max_{(\phi,j)} \frac{b_{\phi}^{(j)}[k]}{\bar{r}_{\phi}[k]} \mu_{\phi} d_{\phi}^{\mathrm{HOL}}[k]$
		Max mean-delay with the Lagrangian [53, 54]	$\max_{(\phi,j)} \left(\lambda_{\phi}^{\bar{\mathrm{d}}^{\max}}[k]+\frac{\partial D_{\phi}^{\bar{\mathrm{d}}}\left(\bar{d}_{\phi}[k]\right)}{\partial \bar{r}_{\phi}[k]}\right) \frac{\partial \bar{r}_{\phi}[k]}{\partial x_{\phi}^{(j)}[k]}$

2.5.2 Structural Decomposition: Connecting Structural Patterns of different Scheduling and Allocation Cores to their Properties

In this section, we decompose the structure of the discussed scheduling algorithm and connect each element of the core to the main aforementioned discussed objectives, namely, transmission efficiency, QoS-requirements, and fairness type. We conceptually show different distinguishable component in the core structures by $\mathfrak{O}^{\mathrm{eff}}$, $\mathfrak{O}^{\mathrm{r-fair}}$, $\mathfrak{O}^{\mathrm{d-fair}}$, $\mathfrak{O}^{\mathrm{r-min}}$, $\mathfrak{O}^{\mathrm{d-max}}$, and $\mathfrak{O}^{\mathrm{PLR}}$ corresponding to transmission efficiency, bit-rate fairness, delay fairness, minimum bi-rate, maximum delay, and maximum PLR. Components $\mathfrak{O}$ show that the associated structure has aforementioned terms. This decomposition sheds light into how the different terms in the structure of scheduling core, such as MLWDF, EXP, and GPF, reflect on the bit-rate fairness, delay fairness, total output bit-rate, and delay performance. The transmission efficiency (channel-awareness) is based on terms as $\mathfrak{O}^{\mathrm{eff}}\left(b_{\phi}^{(j)}\right)$. The bit-rate fairness in EXP, MLWDF, and GPF is originated by terms as $\mathfrak{O}^{\mathrm{r-fair}}\left(1/(\bar{r}_{\phi}[k])^{\alpha}\right)$. The delay fairness in EXP is originated by terms as $\mathfrak{O}^{\mathrm{d-fair}}\left(1/1+\left(\overline{\mathrm{d}^{\mathrm{HOL}}}[k]\right)^{\eta}\right)$. The PLR requirements, in MLWDF and EXP, δ_{ϕ} is embedded in μ_{ϕ} by term as $\mathfrak{O}^{\mathrm{PLR}}\left(-\log(\delta_{\phi})\right)$, where the above-mentioned term increase very rapidly when the PLR, δ_{ϕ}, approaches zero. The maximum

bit-rate, minimum bit-rate and maximum delay guarantee are embedded based on the Lagrangian multiplier of $\mathfrak{O}^{\mathrm{r-min}}\left(\lambda_{\phi}^{\bar{\mathrm{r}}^{\min}}[k]\right)$, $\mathfrak{O}^{\mathrm{r-max}}\left(\lambda_{\phi}^{\bar{\mathrm{r}}^{\max}}[k]\right)$, and $\mathfrak{O}^{\mathrm{d-max}}\left(\lambda_{\phi}^{\bar{\mathrm{d}}^{\max}}[k]\right)$.

Individual Decomposition of the Cores

In the following we decompose the structure of the cores, intuitively, based on the discussed properties, namely, channel-awareness, bit-rate fairness, delay fairness, max & min bit-rate guarantee, max delay guarantee, and PLR requirements.

GPF

$$\max_{(\phi,j)} \quad \underbrace{b_{\phi}^{(j)}[k]}_{\text{channel-awareness}} \quad \overbrace{\frac{1}{(\hat{r}_{\phi}[k])^{\alpha}}}^{\text{bit-rate fairness}}. \tag{2.76}$$

General bit-rate utility fairness through GPF

$$\max_{(\phi,j)} \quad \underbrace{b_{\phi}^{(j)}[k]}_{\text{channel- awareness}} \quad \overbrace{\frac{1}{S_{\phi}^{\bar{\mathrm{r}}}(\hat{r}_{\phi}[k])}\frac{\partial S_{\phi}^{\bar{\mathrm{r}}}\left(\bar{r}_{\phi}[k]\right)}{\partial \bar{r}_{\phi}[k]}}^{\text{bit-rate utility fairness}}. \tag{2.77}$$

General bit-rate utility fairness through Jain's Index Maximization

$$\max_{(\phi,j)} \quad \underbrace{b_{\phi}^{(j)}[k]}_{\text{channel-awareness}} \quad \overbrace{\left(\beta^{\mathrm{S}^{\hat{\mathrm{r}}}} - S_{\phi}^{\hat{\mathrm{r}}}(\hat{r}_{\phi}[k])\right)}^{\text{Jain's utility fairness}}. \tag{2.78}$$

Max & min bit-rate QoS guarantee through virtual token and general utility

$$\max_{(\phi,j)} \quad \underbrace{\mathrm{e}^{\lambda_{\phi}^{\mathrm{token}}[k]}}_{\text{max \& min bi-rate}} \quad \overbrace{\frac{\partial S_{\phi}^{\bar{\mathrm{r}}}\left(\bar{r}_{\phi}[k]\right)}{\partial \bar{r}_{\phi}[k]}}^{\text{bit-rate fairness}} \quad \underbrace{\frac{\partial \bar{r}_{\phi}[k]}{\partial x_{\phi}^{(j)}[k]}}_{\text{channel-awareness}}. \tag{2.79}$$

Min bit-rate QoS guarantee through Lagrangian with general utility

$$\max_{(\phi,j)} \left(\underbrace{\lambda_{\phi}^{\bar{\mathrm{r}}^{\min}}[k]}_{\text{min bit-rate}} + \overbrace{\frac{\partial S_{\phi}^{\bar{\mathrm{r}}}\left(\bar{r}_{\phi}[k]\right)}{\partial \bar{r}_{\phi}[k]}}^{\text{bit-rate fairness}} \right) \underbrace{\frac{\partial \bar{r}_{\phi}[k]}{\partial x_{\phi}^{(j)}[k]}}_{\text{channel-awareness}}. \tag{2.80}$$

Mean-delay fairness through WGPF

$$\max_{(\phi,j)} \underbrace{b_{\phi}^{(j)}[k]}_{\text{channel- awareness}} \overbrace{\left(\bar{d}_{\phi}[k]\right)^{\nu^{\bar{\mathrm{d}}}}}^{\text{WGPF delay fairness}}. \tag{2.81}$$

Mean-delay fairness with general utility

$$\max_{(\phi,j)} \underbrace{b_{\phi}^{(j)}[k]}_{\text{channel-awareness}} \overbrace{D_{\phi}^{\bar{\mathrm{d}}}\left(\bar{d}_{\phi}[k]\right) \frac{\partial D_{\phi}^{\bar{\mathrm{d}}}\left(\bar{d}_{\phi}[k]\right)}{\partial \bar{d}_{\phi}[k]}}^{\text{general utility delay fairness}}. \tag{2.82}$$

HOL-delay fairness through the Jain's index maximization

$$\max_{(\phi,j)} \underbrace{b_{\phi}^{(j)}[k]}_{\text{channel-awareness}} \overbrace{\left(d_{\phi}^{\mathrm{HOL}}[k] - \beta^{\mathrm{d}^{\mathrm{HOL}}}\right)}^{\text{Jain's fairness of HOL-delay}}. \tag{2.83}$$

HOL-delay fairness through the Jain's index maximization with general disutility

$$\max_{(\phi,j)} \underbrace{b_{\phi}^{(j)}[k]}_{\text{channel-awareness}} \overbrace{\left(D_{\phi}^{\bar{\mathrm{d}}}\left(\bar{d}_{\phi}[k]\right) - \beta^{\mathrm{D}^{\bar{\mathrm{d}}}}\right)}^{\text{Jain's utility fairness of mean-delay}}. \tag{2.84}$$

Max HOL-delay QoS guarantee through EDF

$$\max_{(\phi,j)} \underbrace{d_{\phi}^{\mathrm{HOL}}[k] - d_{\phi}^{\mathrm{HOL}^{\max}}}_{\text{max HOL-delay guarantee}}. \tag{2.85}$$

Max HOL-delay QoS guarantee through EXP

$$\max_{(\phi,j)} \underbrace{b_{\phi}^{(j)}[k]}_{\text{channel- awareness}} \overbrace{\frac{1}{\bar{r}_{\phi}[k]}}^{\text{PF bit-rate fairness}} \mathrm{e}^{\left(\frac{\overbrace{\mu_{\phi} d_{\phi}^{\mathrm{HOL}}[k]}^{\text{max HOL-delay guarantee through division}}}{\underbrace{1 + \left(\overline{d^{\mathrm{HOL}}}[k] \right)^{\eta}}_{\text{HOL-delay fairness}}} \right)}. \tag{2.86}$$

Max HOL-delay QoS guarantee through MWLDF

$$\max_{(\phi,j)} \underbrace{b_{\phi}^{(j)}[k]}_{\text{channel- awareness}} \overbrace{\frac{1}{\bar{r}_{\phi}[k]}}^{\text{PF bit-rate fairness}} \underbrace{\mu_{\phi} d_{\phi}^{\mathrm{HOL}}[k]}_{\text{max HOL-delay guarantee}}. \tag{2.87}$$

Max mean-delay QoS guarantee through Lagrangian with general bit-rate utility

$$\max_{(\phi,j)} \left(\underbrace{\lambda_{\phi}^{\bar{\mathrm{d}}^{\max}}[k]}_{\text{max mean-delay guarantee}} + \overbrace{\frac{\partial D_{\phi}^{\bar{\mathrm{d}}}\left(\bar{d}_{\phi}[k]\right)}{\partial \bar{r}_{\phi}[k]}}^{\text{mean-delay fairness}} \right) \underbrace{\frac{\partial \bar{r}_{\phi}[k]}{\partial x_{\phi}^{(j)}[k]}}_{\text{channel-awareness}}. \tag{2.88}$$

The novel structures which proposed in Sections 2.3.4, 2.3.5, and 2.4.5 have been inspired based on the structural decomposition and the classification which have been presented as the novel scheduling core structure in Table 2.3 and Table 2.2 with shaded background in the table. These novel cores make an example of a methodology for designing novel scheduling cores based on separating the math analyses and engineering aspects.

2.5.3 Notes on the Performance Evaluation and Comparison of Scheduling and Allocation Cores

Having elaborated on different packet scheduling and RB allocation algorithms, a few important notes regarding the performance evaluation of the scheduling algorithms is worth highlighting. Some of the notes may seems trivial, however, there are many examples in the literature which violate them. First, in comparing different algorithms, the algorithms designed for a single traffic type should be compared. Second, different algorithms can outperform in different range of the parameters, especially the input load. As an example, in addition to the SINR regime in PHY layer, in MAC layer the load regime is another input parameter where for different load regime algorithms behave differently. Especially, different algorithms should be designed for different input load situation. A possibility is to design a comprehensive algorithm which can change its gear, including its fairness, based on the flows requirements and input load to adapt to the system. For example, in under-load, the largest supported cell size (which can be realized through bit-rate maxmin) and in over-load the sum bit-rate maximization which has the smallest cell size should be used for NRT traffic. Third, the scheduling design problem is not a single commodity problem. In other words, for a certain design, one of the output performance metrics (such as sum bit-rate, delay performances, fairness, to name a few) can be improved while some degradation in the other metrics may be inevitable. Nevertheless, the overall trade-offs on efficiency (capacity) for different traffic type, fairness, and the QoS determines the quality of a scheduling core. In fact, the achievable total bit-rate, an OFDMA frames can support, depends on three main components, namely, the OFDMA raw AMC values, the QoS-requirements, and the core of the scheduling algorithm. Two scheduling cores can handle a same QoS-requirements on an identical OFDMA realization but with different capacity.

An alternative for performance evaluation of the scheduling algorithm is sub-optimality gap. If a tight bound on the objective of the mathematical optimization is possible, the performance evaluation can be based on a sub-optimality gap, defined by the absolute difference of the value of the objective and the bound. As an example, an upper bound on a discrete maximization is possible to make by relaxing the discrete optimization variables. When the relaxed version has

a convex structure, the upper bound can be found by the available off-the-shelf standard convex solvers.

2.5.4 Concluding Remarks

In this chapter, a harmonized survey of the packet scheduling and RB allocation algorithms, for wireless OFDMA networks, based on the main design objectives, namely, efficiency, fairness, and QoS-requirements, was provided in details with extensive elaboration on the interrelation of different cores and on the design methodology. The schedulers were classified based on their core and decomposed based on their main properties. We connected the properties of the schedulers to their core structure. These intuitive connections make the study and research of the schedulers easier and streamlined.

It can be observed that further research is still needed in design and performance evaluation of scheduling algorithms for the different types of traffic, in wireless networks. Data traffic on the next generation of wireless networks is expected to be very diverse with various requirements, including different maximum and minimum bounds on bit-rate, delay, and jitter, in addition to fairness requirements. Several aspects of the packet scheduling and RB allocation algorithm, such as optimality criteria, performance metrics, and algorithms are still in a rather non-mature stage, especially in heterogeneous QoS provisioning. The challenge becomes more pronounced, especially, because the demand for enabling emerging new applications in the mobile wireless networks (such as IoT) has grown exponentially and has overtaken the study of scheduling algorithms.

We observed that the research in this area is rather fragmented in different disciplines. This formidable challenge makes harmonization, unification, and axiomatization [9] crucial. In this chapter, we made an effort to present different packet scheduling and RB allocation algorithms by reducing their presentation to their fundamental properties in a non-redundant self-consistent manner, towards unifying known notions, proving basic properties, and towards finding novel notions. This presentation approach paves the way for systematic design.

Appendix

2.A Proof of Bit-rate Maxmin Case of GPF, for NRT flows

Proposition 1. *For α sufficiently large, assigning a time slot to the flow with the minimum mean bit-rate on its best sub-channel is equivalent to assigning resources to the flow with largest gradient in (2.1).*

Proof. Define the best sub-channel AMC value for flow ϕ with $b_\phi^{(j_{\max}(\phi))}[k]$, the mean bit-rate of flow $\phi_{\min}$ with the lowest bit-rate among all flows with $\bar{r}_{\phi_{\min}}[k]$ as

$$b_\phi^{(j_{\max}(\phi))}[k] \triangleq \max_{1\le j\le N}\left\{b_\phi^{(j)}[k]\right\},\ \ \bar{r}_{\phi_{\min}}[k] \triangleq \min_{\phi\in\Phi_{\mathrm{NRT}}}\left\{\Big(1-\hbar[k]\Big)\bar{r}_\phi[k-1]+\hbar[k]W_b\sum_{j=1}^{N} b_\phi^{(j)}[k]\ x_\phi^{(j)}[k]\right\}, \tag{2.89}$$

and a threshold on α as

$$\alpha_0 \triangleq \max_{\phi\in\Phi_{\mathrm{NRT}}}\left\{\log\left(\frac{b_\phi^{(j_{\max}(\phi))}[k]}{b_{\phi_{\min}}^{(j_{\max}(\phi_{\min}))}[k]}\right)\Big/\log\left(\frac{\bar{r}_\phi[k]}{\bar{r}_{\phi_{\min}}[k]}\right)\right\}. \tag{2.90}$$

Since for all $\phi\neq\phi_{\min}$, we have $0<\log\Big(\bar{r}_\phi[k]\Big/\bar{r}_{\phi_{\min}}[k]\Big)$, the following is true, for all $\alpha\ge\alpha_0$, $\phi\neq\phi_{\min}$.

$$\alpha\log\left(\frac{\bar{r}_\phi[k]}{\bar{r}_{\phi_{\min}}[k]}\right)\ge\log\left(\frac{b_\phi^{(j_{\max}(\phi))}[k]}{b_{\phi_{\min}}^{(j_{\max}(\phi_{\min}))}[k]}\right), \tag{2.91}$$

or

$$\frac{\Big(\bar{r}_\phi[k]\Big)^\alpha}{\Big(\bar{r}_{\phi_{\min}}[k]\Big)^\alpha}\ge\frac{b_\phi^{(j_{\max}(\phi))}[k]}{b_{\phi_{\min}}^{(j_{\max}(\phi_{\min}))}[k]}\quad\leftrightarrow\quad\frac{b_\phi^{(j_{\max}(\phi))}[k]}{\Big(\bar{r}_\phi[k]\Big)^\alpha}\le\frac{b_{\phi_{\min}}^{(j_{\max}(\phi_{\min}))}[k]}{\Big(\bar{r}_{\phi_{\min}}[k]\Big)^\alpha}, \tag{2.92}$$

by taking the exponent of both sides of the first inequality. Finally, by definition of $b_{\phi}^{(j_{\max}(\phi))}[k]$ the followings are true

$$\frac{b_{\phi}^{(j)}[k]}{(\bar{r}_{\phi}[k])^{\alpha}} \leq \frac{b_{\phi}^{(j_{\max}(\phi))}[k]}{(\bar{r}_{\phi}[k])^{\alpha}} \leq \frac{b_{\phi_{\min}}^{(j_{\max}(\phi_{\min}))}[k]}{(\bar{r}_{\phi_{\min}}[k])^{\alpha}} \quad \Rightarrow \quad \max_{\phi\in\Phi_{\mathrm{NRT}}} \max_{1\leq j\leq N} \frac{b_{\phi}^{(j)}[k]}{(\bar{r}_{\phi}[k])^{\alpha}} \leq \frac{b_{\phi_{\min}}^{(j_{\max}(\phi_{\min}))}[k]}{(\bar{r}_{\phi_{\min}}[k])^{\alpha}}, \tag{2.93}$$

for all $\alpha \geq \alpha_0$, proving the proposition. □

Proposition 2. *For $\alpha \to \infty$, the network utility maximization leads to maxmin objective.*

Proof. For $\alpha > 1$, the network utility maximization becomes a minimization problem as

$$\min_{x_{\phi}^{(j)}\in\mathcal{C}^{\mathrm{PHY}}} \quad \frac{1}{\alpha-1} \sum_{\phi=1}^{|\Phi_{\mathrm{NRT}}|} \frac{1}{(\bar{r}_{\phi}[k])^{\alpha-1}}. \tag{2.94}$$

Because of the a monotonically increasing property of the root function $(.)^{\frac{1}{\alpha-1}}$ for $1 \leq \alpha$, the above optimization is equivalent to

$$\min_{x_{\phi}^{(j)}\in\mathcal{C}^{\mathrm{PHY}}} \sqrt[\alpha-1]{\sum_{\phi=1}^{|\Phi_{\mathrm{NRT}}|} \frac{1}{(\bar{r}_{\phi}[k])^{\alpha-1}}} = \min_{x_{\phi}^{(j)}\in\mathcal{C}^{\mathrm{PHY}}} L_{\alpha-1}\left(\frac{1}{\bar{r}_1[k]}, \ldots, \frac{1}{\bar{r}_{|\Phi_{\mathrm{NRT}}|}[k]}\right), \tag{2.95}$$

where

$$L_p(\mathbf{v}) \triangleq \sqrt[p]{\sum_{\phi=1}^{|\Phi_{\mathrm{NRT}}|} (v_{\phi})^p} \tag{2.96}$$

is the L_p norm of vector $(v_1, \ldots, v_{|\Phi_{\mathrm{NRT}}|})$. As $p \to \infty$, the L_p norm becomes the L_∞ norm

$$L_p(\mathbf{v}) \underset{p\to\infty}{=} L_{\infty}(\mathbf{v}) \triangleq \max_{\phi}\{v_{\phi}\}. \tag{2.97}$$

Therefore, as $\alpha \to \infty$ we have

$$\min_{x_{\phi}^{(j)}[k]} \quad \max_{\phi} \quad \frac{1}{\bar{r}_{\phi}[k]} = \min_{x_{\phi}^{(j)}[k]} \quad \frac{1}{\min_{\phi} \bar{r}_{\phi}[k]}, \tag{2.98}$$

which is equivalent to the following optimization due to the monotonically decreasing property of

function $f(u) = \frac{1}{u}$.

$$\max_{x_\phi^{(j)}[k]} \min_{\phi} \bar{r}_\phi[k]. \tag{2.99}$$

□

2.B Proof of Delay Minmax Case of WGPF, for RT flows

Proposition 3. *For (2.51), as $\nu^{\bar{\mathrm{d}}} \to \infty$, the network disutility leads to a minmax delay-fair allocation.*

Proof. This can be shown by noting an interesting relationship between the network disutility minimization and the L_p norm minimization. Consider the network disutility minimization as

$$\min_{x_\phi^{(j)} \in \mathcal{C}^{\mathrm{PHY}}} \quad \sum_{\phi=1}^{|\Phi_{\mathrm{RT}}|} \left(\bar{d}_\phi[k]\right)^{\nu^{\bar{\mathrm{d}}}}. \tag{2.100}$$

Because of the monotonically increasing property of the root function $(.)^{\frac{1}{\nu^{\bar{\mathrm{d}}}}}$, for $1 \leq \nu^{\bar{\mathrm{d}}}$, the above optimization is equivalent to

$$\min_{x_\phi^{(j)} \in \mathcal{C}^{\mathrm{PHY}}} \sqrt[\nu^{\bar{\mathrm{d}}}]{\sum_{\phi=1}^{|\Phi_{\mathrm{RT}}|} \left(\bar{d}_\phi[k]\right)^{\nu^{\bar{\mathrm{d}}}}} = \min_{x_\phi^{(j)} \in \mathcal{C}^{\mathrm{PHY}}} L_{\nu^{\bar{\mathrm{d}}}} \left(\bar{d}_1[k], \ldots, \bar{d}_{|\Phi_{\mathrm{RT}}|}[k]\right). \tag{2.101}$$

As $p \to \infty$, the L_p norm becomes the L_∞ norm

$$L_p \left(v_1, \ldots, v_{|\Phi_{\mathrm{RT}}|}\right) \underset{p\to\infty}{=} L_\infty \left(v_1, \ldots, v_{|\Phi_{\mathrm{RT}}|}\right) \triangleq \max_{1\leq\phi\leq|\Phi_{\mathrm{RT}}|} \{v_\phi\}. \tag{2.102}$$

Therefore, as $\nu^{\bar{\mathrm{d}}} \to \infty$, the main optimization becomes equivalent to

$$\min_{x_\phi^{(j)} \in \mathcal{C}^{\mathrm{PHY}}} \quad \max_{\phi} \quad \bar{d}_\phi[k]. \tag{2.103}$$

This objective guarantees that the maximum mean-delay among flows is minimized in frame k. □

Proposition 4. *For a sufficiently large $\nu^{\bar{\mathrm{d}}}$, assigning a resource to the flow with the maximum*

current mean-delay in the iterations in the frame k on its best sub-channel is equivalent to assigning resource to the flow with the smallest gradient, or the largest absolute value gradient.

Proof. Define the best sub-channel bit-rate for flow ϕ denoted by $b_{\phi}^{(j_{\max}(\phi))}[k]$ and the mean-delay of flow $\phi_{\max}$ with the largest measure among all flows, in frame k with $\bar{d}_{\phi_{\max}}[k]$, as

$$b_{\phi}^{(j_{\max}(\phi))}[k] \triangleq \max_{1\leq j\leq N}\left\{b_{\phi}^{(j)}[k]\right\}, \qquad \bar{d}_{\phi_{\max}}[k] \triangleq \max_{1\leq \phi\leq |\Phi_{\mathrm{RT}}|}\left\{\bar{d}_{\phi}[k]\right\}, \tag{2.104}$$

and an axillary variable $\Upsilon_{\phi}^{(j)}[k]$ as

$$\Upsilon_{\phi}^{(j)}[k] \triangleq \frac{\partial \bar{d}_{\phi}[k]}{\partial x_{\phi}^{(j)}[k]} \Bigg/ \frac{\partial \bar{d}_{\phi_{\max}}[k]}{\partial x_{\phi_{\max}}^{(j_{\max}(\phi_{\max}))}[k]}, \tag{2.105}$$

which is finite and is equal to

$$\Upsilon_{\phi}^{(j)}[k] = \frac{b_{\phi}^{(j)}[k]}{b_{\phi_{\max}}^{(j_{\max}(\phi_{\max}))}[k]} \; \frac{\zeta_{\phi}^{(1)}[k]\zeta_{\phi}^{(4)}[k] + \zeta_{\phi}^{(2)}[k]\zeta_{\phi}^{(3)}[k]}{\zeta_{\phi_{\max}}^{(1)}[k]\zeta_{\phi_{\max}}^{(4)}[k] + \zeta_{\phi_{\max}}^{(2)}[k]\zeta_{\phi_{\max}}^{(3)}[k]} \; \frac{\left(\zeta_{\phi_{\max}}^{(3)}[k] + \zeta_{\phi_{\max}}^{(4)}[k] r_{\phi_{\max}}[k]\right)^2}{\left(\zeta_{\phi}^{(3)}[k] + \zeta_{\phi}^{(4)}[k] r_{\phi}[k]\right)^2}, \tag{2.106}$$

and a threshold on $\nu^{\bar{\mathrm{d}}}$ as

$$\nu_0^{\bar{\mathrm{d}}} \triangleq \max_{\phi\neq\phi_{\max}, 1\leq j\leq N}\left\{\log\left(\Upsilon_{\phi}^{(j)}[k]\right) \Big/ \log\left(\frac{\bar{d}_{\phi_{\max}}[k]}{\bar{d}_{\phi}[k]}\right)\right\} + 1. \tag{2.107}$$

Since by definition $\bar{d}_{\phi_{\max}}[k]$ is the largest delay in frame k, $\forall\ \phi \neq \phi_{\max}$, we have $0 < \log\left(\frac{\bar{d}_{\phi_{\max}}[k]}{\bar{d}_{\phi}[k]}\right)$, and the following is true as

$$\forall\ \nu^{\bar{\mathrm{d}}} \geq \nu_0^{\bar{\mathrm{d}}},\ \forall\ \phi \neq \phi_{\max},\ \text{and}\ \forall\ j: \quad (\nu^{\bar{\mathrm{d}}} - 1)\log\left(\frac{\bar{d}_{\phi_{\max}}[k]}{\bar{d}_{\phi}[k]}\right) \geq \log\left(\Upsilon_{\phi}^{(j)}[k]\right). \tag{2.108}$$

Taking the exponent of both sides of the above inequality we get

$$\frac{\left(\bar{d}_{\phi_{\max}}[k]\right)^{(\nu^{\bar{\mathrm{d}}}-1)}}{\left(\bar{d}_{\phi}[k]\right)^{(\nu^{\bar{\mathrm{d}}}-1)}} \geq \Upsilon_{\phi}^{(j)}[k], \tag{2.109}$$

or

$$b_{\phi}^{(j)}[k]\ \frac{\zeta_{\phi}^{(1)}[k]\zeta_{\phi}^{(4)}[k]+\zeta_{\phi}^{(2)}[k]\zeta_{\phi}^{(3)}[k]}{\left(\zeta_{\phi}^{(3)}[k]+\zeta_{\phi}^{(4)}[k]r_{\phi}[k]\right)^{2}}\ \left(\bar{d}_{\phi}[k]\right)^{\nu^{\bar{\mathrm{d}}}-1} \leq \\ b_{\phi_{\max}}^{(j_{\max}(\phi_{\max}))}[k]\ \frac{\zeta_{\phi_{\max}}^{(1)}[k]\zeta_{\phi_{\max}}^{(4)}[k]+\zeta_{\phi_{\max}}^{(2)}[k]\zeta_{\phi_{\max}}^{(3)}[k]}{\left(\zeta_{\phi_{\max}}^{(3)}[k]+\zeta_{\phi_{\max}}^{(4)}[k]r_{\phi_{\max}}[k]\right)^{2}}\ \left(\bar{d}_{\phi_{\max}}[k]\right)^{\nu^{\bar{\mathrm{d}}}-1}. \tag{2.110}$$

Note that $\frac{\partial r_{\phi}[k]}{\partial x_{\phi}^{(j)}[k]} = b_{\phi}^{(j)}[k]$, and the other two components, in each side of the above expression, are $\frac{\partial D_{\mathrm{net}}^{\bar{\mathrm{d}}}\left(\bar{\mathbf{d}}[k]\right)}{\partial \bar{d}_{\phi}[k]}$ & $\frac{\partial \bar{d}_{\phi}[k]}{\partial r_{\phi}[k]}$ in (2.55).

Therefore, for $\forall\ \phi \neq \phi_{\max}$ and $\forall\ j$, we have

$$\left|\frac{\partial D_{\mathrm{net}}^{\bar{\mathrm{d}}}\left(\bar{\mathbf{d}}[k]\right)}{\partial x_{\phi}^{(j)}[k]}\right| \leq \left|\frac{\partial D_{\mathrm{net}}^{\bar{\mathrm{d}}}\left(\bar{\mathbf{d}}[k]\right)}{\partial x_{\phi_{\max}}^{(j_{\max}(\phi_{\max}))}[k]}\right|. \tag{2.111}$$

This ensures that the flow with the highest mean delay has the smallest gradient (the largest absolute value of the gradient), proving the proposition. □

Chapter 3

Joint Realtime and Non-Realtime Flows Packet Scheduling and Resource Block Allocation in Wireless OFDMA Networks

Abstract

In this chapter, we consider the resource allocation and packet scheduling for realtime (RT) and non-realtime (NRT) packet-switched flows in orthogonal frequency division multiple access (OFDMA) wireless networks. Radio resource blocks (RB)s in OFDMA plane are to be distributed among RT and NRT flows. In the conventional approach, resource allocation for RT and NRT flows are executed sequentially. This sequential approach is inefficient, because an RT flow may presumably have enough time until its deadline while its channel is in deep fade. In this situation, the transmission of NRT with high efficient transmission opportunities can be performed. Intuitively speaking, the conventional sequential approach is too conservative approach that should be reengineered. We propose a novel joint unified utility based packet scheduling and RB allocation, in a common pool of RBs. The proposed joint approach, or using a common pool of RBs, enlarges the effective capacity

of the associated wireless system, when compared to the separated pool of RBs. We use mean bit-rate, mean queue length, and instantaneous queueing delay information, in addition to channel information embedded in gradient of dis-utilities, to match the demand and supply. Exploiting the inherent diversity, including time, frequency, spatial, and multiuser diversity in a wireless system is a key to improve its performance. Joint RT and NRT flows packet scheduling and RB allocation exploits multiuser diversity, better than conventional approach. It is worth mentioning that the increase in the number of WTs and the increase in the heterogeneity of QoS requirements result in the increase of the potent multiuser diversity. The increase in the multiuser diversity is exploitable through the proposed joint approach. Furthermore, we develop a novel model for input-output bit-rate behaviour in resource allocation of the mixture of RT and NRT flows. This model sheds light to identifying different load regions, explaining them, and understanding of the system in a simple and intuitive manner. Our approach and methodology can be extended for broader quality-of-service requirements and for the utilities of the future applications. Extensive simulation results show that the proposed framework can unify the RT and NRT and achieves higher admissible bit-rate when handling mixed RT and NRT flows compared to baselines.

3.1 Introduction

New Services, Importance of QoS, Over-provisioning, and Capacity Crunch

The main requirement for next generation of wireless network is that it should cost-effectively provide guaranteed QoS, especially in terms of delay and bit-rate requirement, with ubiquitous high bit-rate coverage, when and where required [2]. Wireless networks are part of a highly complex heterogeneous interactive system, where consumers share limited radio resources for a broad range of services such as Voice-over-IP (VoIP), tele-medicine, online games, industrial/home automation, wearable connected devices, Hulu, Netflix, Chrome OS. The flows for these vastly different services require highly different quality-of-service (QoS). Traditionally, QoS in cellular communications has been implemented with *over-provisioning*, or through costly higher layer mechanisms and overheads. Over-provisioning results in a network design for its peak load which makes the system highly

inefficient. In this setup, when the network becomes congested (load approaching the capacity), conventional rate limiters or bandwidth throttling is used which causes user dissatisfaction [3, 4]. Tomorrow's networks with more frequent congestion problems will not have the luxury of over-provisioning or using various forms of excessive overhead. Advanced access technologies, such as the long term evolution (LTE), are purely scheduled system based on orthogonal frequency division multiple access (OFDMA) which creates the opportunity to dynamically and efficiently exploit various types of diversity and to schedule for diverse requirements, instead of over-provisioning. The main question, then, is how to do the resource allocation to treat different flows with different demands and different wireless links.

Conventional Sequential Approach

The packet-switched connections can generally be divided into realtime (RT) and non-realtime (NRT) flows. Conventionally the resource allocation for RT flows is designed based on the earliest deadline first (EDF) [63]. On the other hand, some version of proportional fairness (PF) [29,34,39], or generalized proportional fairness (GPF) [18, 19, 31], scheduler is used for NRT flows, where the PF algorithm serves a flow whose instantaneous bit-rate divided by its mean bit-rate is the highest in each frame. Note that PF scheduler is queue-blind so it cannot be used properly for RT flows and it is not stable with respect to queues [66].

Inefficiency of the Sequential Approach

In the conventional approach, resource allocation for RT and NRT flows are executed sequentially [74]. In other words, RT flows are served first, and if any resources are still available, NRT flows are served subsequently. This static priority separation, or sequential approach, is inefficient. The reason for this inefficiency is that an RT flow may presumably have enough time until its deadline while its channel is in deep fade. In this situation the transmission of other flows, and possibly an NRT flow with good channel condition, can take place. Moreover, NRT flows are not completely insensitive to delay, especially to the mean-delay. Therefore, a joint resource allocation of RT and NRT flows which achieves higher level of multiuser diversity, and can also consider the QoS of NRT

flows, becomes important. The joint resource allocation of RT and NRT flows exploits both intra- and inter-class opportunism across flows. Heterogeneous QoS requirements in time and among flows, different load conditions, limited capacity in comparison to demands, and more frequent congestions make the joint resource allocation of RT and NRT flows more significant.

Need for Reengineering the Architecture of Data and Voice Services

Because of the pre-existent voice services and the gradual emersion of data traffic in the cellular networks, voice services are designed separately from packet-switched data in the pre-4G cellular networks, such as GSM, UMTS and CDMA2000 [5]. In these networks only NRT flows are subject to scheduling over the shared channel and voice services are served in circuit-switched mode over the dedicated channel (DCH) [5]. This static separation sacrifices multiuser diversity. Today's wireless networks such as LTE and the upcoming LTE-A are moving towards packet-switching and IP-flat architecture to serve a broad range of applications with many diverse requirements [6]. Designing the flexible resource allocation framework which considers the heterogeneity is crucial. In fact, providing properly engineered differentiable data flows is more cost-effective than providing voice and data in separation. In the flat architecture, voice will be one of the differentiated packet-switched data flows that the scheduler is responsible to guarantee its QoS requirements. With voice as a RT differentiated data flow, networks will have reduced access delay, shorter wake-from-idle time and be able to offer not only regular voice calls but also differentiated higher quality (audio or video) calls. There are three main ways of implementation for the new IP-flat architecture in regards to voice services, namely, voice-over-LTE (VoLTE), circuit-switched fall back (CSFB), and simultaneous voice and LTE (SVLTE) [7]. Among these three, VoLTE is the only one that really allows the delivery of voice as a data flow, within the LTE data bearer. CSFB and SVLTE are still dependent on the old pre-4G architecture and fall back to the legacy 2G or 3G circuit-switching in voice calls. CSFB and SVLTE have several inefficiencies in regards to VoLTE, such as longer call access delay, more expensive handsets and access points, and high power consumption on handset [7]. In regards to VoLTE, the joint resource allocation of RT and NRT flows is becoming more important in improving the efficiency of the cellular networks. Note that, although there are

some partial solutions for the integration of RT and NRT flows in the application layer such as WebRTC [8], the packet and resource block (RB) scheduler in the centre of medium access control (MAC) layer is the main component to be designed efficiently for guaranteeing differentiated QoS.

Bit-rate Utilities and Queue-awareness

There has been much prior works on scheduling and resource allocation in wireless networks based on the channel information and the queue information; see, for example, [21, 22, 70, 75, 76] and the references therein. Many resource allocation algorithms based on bit-rate utilities [16, 33, 35, 39, 47, 49, 69, 77–86] have been developed and studied in the literature. Bit-rate-based-only utilities are relevant to commonly called infinite-backlog scenario, where the queues are assumed to be always full, independent of service. Nevertheless, bit-rate-based utilities are blind to the requirements of flows, especially to the RT ones, where the requirements are more important. Even for NRT flows, the queue-blindness of bit-rate-based-only utilities makes the system potentially unstable [66, 87]. Therefore, queue-awareness should be a key feature in designing the joint resource allocation of RT and NRT flows.

Related Works

Prior studies on resource allocation of RT and NRT flows mainly focus on a single QoS element; such as head-of-the-line (HOL)-delay [51, 66, 88], or the mean delay [36, 37]. Two heuristics called exponential-rule (EXP), and modified largest weighted delay first (MLWDF) have been also studied in [52,64,68] based on the HOL delay. Another study is a joint channel- and queue-aware scheduling maximum delay utility (MDU) [36, 37], where the Little's delay is utilized. Since the averaging is a low pass filter, RT flows suffer when the Little's delay (which is essentially a mean-delay) is used. In other words, the decisions based on the mean-delay are insensitive to HOL-delay (or equivalently, instantaneous delay) requirements. Another approach based on time-utility function (TUF) has been proposed in [89–91], where the idea is to force the scheduler to transmit at near the deadlines. This approach has the shortcoming of not fully exploiting the multiuser diversity as well as increasing the number of dropped packets due to the passed deadlines. A recent study is

a utility-based adaptive approach with traffic prioritization [92, 93], where at the expense of 15 % loss in throughput it gains a decreased delay variance (or equivalently improved delay fairness) among WTs. The utility functions in [92] are assumed as the summation of the separate utilities on its allocated RBs, rather than a single utility on top of its delay, or bit-rate. Previously, we used the mean-delay dis-utilities in [12] to introduce and analyze the minmax mean-delay fairness mathematical notion. Here in this chapter, we use a framework inspired from [12] but incorporates HOL-delay, queue-length information, and bit-rate information jointly in the dis-utilities, and design the dis-utilities to advance a unified framework for joint resource allocation of RT and NRT flows.

Chapter Contributions

The main contributions in this chapter can be summarized into two main items:

1. We propose a novel *joint RT and NRT flows packet scheduling and RB allocation based on HOL-delay, queue length information, and bit-rate information, besides the embedded channel information in the dis-utilities*. The proposed approach responds to heterogenous delay requirements for RT flows and manages NRT flows effectively within a *common pool of RBs*, rather than the sequential resource allocation of RT and NRT flows. Although the framework is designed for wireless networks, whenever there is heterogeneity among resources, including (but not limited to) multiuser diversity or any large-scale or small-scale signal variation, the proposed joint RT and NRT scheduling and allocation can offer higher performance, in comparison to the the sequential approach. Furthermore, the developed framework enables putting different algorithms in the literature in perspective and in a unified manner (see Section 3.4). Our approach is also a joint optimization in terms of both packet scheduling and resource allocation in one shot.

2. We developed a novel model for input-output bit-rate behaviour in resource allocation of the mixture of RT and NRT flows. The model elaborates on different capacity definitions (necessary to describe the system of packet scheduling and RB allocation with the heterogeneous traffic), and their dependence on the input load. In addition, this model sheds light on identifying under-load region, over-load regions, the general trends of output bit-rate of RT

& NRT flows, and understanding of the system in a simple and intuitive manner.

Chapter Organization

This chapter is organized into seven main parts:

1. Introduction, background, and the motivation are explained in Section 3.1.
2. The system model and definitions, for the joint RT and NRT packet scheduling and RB allocation, will be given in Section 3.2.
3. The formulation of the novel proposed joint RT and NRT flows packet scheduling and RB allocation will be described in Section 3.3, where the sequential approach in Section 3.3.1, will be compared with the proposed joint approach in Section 3.3.2.
4. The special cases of the proposed framework will be discussed in Section 3.4.
5. The proposed novel algorithm will be discussed in Section 3.5.
6. The high-level input-output system behaviour study of the system will be discussed in Section 3.6, as a benchmark and explanation for the simulations, in the next section.
7. Finally, simulation experiments, through two experiments, will be presented in Section 3.7.

List of Symbols

In this section, we summarize the symbols used throughout this chapter, with a short definition of them in Table 3.1.

Table 3.1: List of symbols used in the joint RT and NRT packet scheduling and RB allocation.

Symbol	**Definition**
Φ_{RT}	Set of RT flows
Φ_{NRT}	Set of NRT flows
Φ	Set of all flows
N	Number of frequency sub-channels

T	Number of time slots per frequency sub-channel
T_b	Time span of each RB in second
W_b	Frequency span of each RB in Hertz
ϕ	Flow index
j	Sub-channel index
k	Frame index
$\mathfrak{f}\left(\mathsf{SINR}_{\phi}^{(j)}[k]\right)$	Function describing the AMC table ,from SINR values
$b_{\phi}^{(j)}[k]$	AMC value of the an RB on sub-channel j, for flow ϕ, in frame k
$\mathsf{SINR}_{\phi}^{(j)}[k]$	SINR on sub-channel j, for flow ϕ, in frame k
$x_{\phi}^{(j)}[k]$	Scheduling variable for sub-channel j, flow ϕ, in frame k
$r_{\phi}[k]$	Frame bit-rate for flow ϕ, in frame k
$q_{\phi}[k]$	Frame queue length for flow ϕ, in frame k
$D_{\phi}^{\text{joint}}\left(d_{\phi}^{\text{HOL}}[k], \bar{q}_{\phi}[k], \bar{r}_{\phi}[k]\right)$	Disutility function for flow ϕ
$\mathcal{C}^{\text{PHY}}$	Feasible set for the scheduling variable
$D_{\text{net}}^{\text{joint}}\left(\mathbf{d}^{\text{HOL}}[k], \bar{\mathbf{q}}[k], \bar{\mathbf{r}}[k]\right)$	Network disutility
$\mathbf{d}^{\text{HOL}}[k], \bar{\mathbf{q}}[k],$ & $\bar{\mathbf{r}}[k]$	HOL-delay vector, mean queue length vector, mean bit-rate
$d_{\phi}^{\text{HOL}}[k]$	HOL-delay of flow ϕ, in frame k
$\bar{q}_{\phi}[k]$	Mean queue lenght of flow ϕ, until frame k
$\bar{r}_{\phi}[k]$	Mean bit-rate of flow ϕ, until frame k
$t_{\phi}^{\text{HOL}}[k]$	Time stamp of arrival of the HOL packet, for flow ϕ, in frame k
$d_{\phi}^{\text{HOL}^{\max}}$	Delay deadline of HOL-delay, for flow ϕ
$\Delta d_{\phi}^{\text{HOL}}[k]$	Difference of HOL-delay and its deadline, for flow ϕ

$\frac{\partial D_{\phi}^{\text{joint}}\left(d_{\phi}^{\text{HOL}}[k],\bar{q}_{\phi}[k],\bar{r}_{\phi}[k]\right)}{\partial x_{\phi}^{(j)}[k]}$	Disutility gradient, for flow ϕ, with respect to its scheduling variable on sub-channel j and frame k
κ	Channel awareness factor of RT flows
$F_{\phi}^{\bar{\text{r}}}(\bar{r}_{\phi}[k])$	Bit-rate importance function, for flow ϕ
$F_{\phi}^{\text{d}^{\text{HOL}}}(d_{\phi}^{\text{HOL}}[k])$	HOL-delay importance function, for flow ϕ
ϑ	Scale factor
ξ	NRT disutility gradient maximum
$F_{\phi}^{\bar{\text{q}}}(\bar{q}_{\phi}[k])$	Mean queue size importance function, for flow ϕ
$F_{\phi}^{\bar{\text{r}}}(\bar{r}_{\phi}[k])$	Mean bit-rate importance function, for flow ϕ
π	Interpretable design parameter for ξ
α	GPF parameter
μ_{ϕ}	EXP and MLWDF parameter, for flow ϕ
δ_{ϕ}	Probability of exceeding the HOL-delay in MLWDF, for flow ϕ
τ_{ϕ}	Maximum delay threshold in MLWDF, for flow ϕ, same as $d_{\phi}^{\text{HOL}^{\max}}$
$\overline{d^{\text{HOL}}}[k]$	Average of HOL-delay over flows, in frame k
$\lvert\Phi_{\text{RT}}\rvert$ & $\lvert\Phi_{\text{NRT}}\rvert$	Number of RT flows, number of NRT flows
$j^{*}[k]$	Algorithm internal variable for selected sub-channel, in frame k
$\phi^{*}[k]$	Algorithm internal variable for selected flow, in frame k
η	EXP parameter controlling the delay fairness
$\bar{d}_{\phi}[k]$	Mean-delay, until frame k
$\bar{d}_{\phi}^{\max}$	Maximum mean-delay, for flow ϕ
Φ_{BE}	Set of BE flows, as an NRT example
Φ_{VoIP}	Set of VoIP flows, as an RT example

$\nu^{\bar{\mathrm{d}}}$	Mean-delay fairness parameter, controlling the mean-delay equalization
$\nu^{\mathrm{d^{HOL}}}$	HOL-delay fairness parameter, controlling the HOL-delay equalization
ι_{RT}	RT flows transmission interval in TUF
ι_{NRT}	NRT flows transmission interval in TUF
l_{ϕ}	Length of interval for RT flows transmissions in TUF, for flow ϕ
$T^{(j)}$	Algorithm internal variable for available slots per sub-channel j
$\tilde{b}_{\phi}^{(j)}$	Algorithm internal variable for AMC on sub-channel j and flow ϕ
Ω_{RT}	Capacity of system when output consists of RT-only flows
$\Gamma^{(j)}_{\phi\in\Phi_{\mathrm{RT}}}[k]$	The selecting criterion for RT flows
$\Gamma^{(j)}_{\phi\in\Phi_{\mathrm{NRT}}}[k]$	The selecting criterion for NRT flows
Ω_{NRT}	Capacity of system when output consists of NRT-only flows
Ω_{SFT}	Capacity of system when saturated for the first time
$\Lambda_{\Sigma}^{\mathrm{in}}$	Total input bit-rate to the system
$\Lambda_{\mathrm{RT}}^{\mathrm{in}}$	Total RT flows input bit-rate
$\Lambda_{\mathrm{NRT}}^{\mathrm{in}}$	Total NRT flows input bit-rate
f_{RT}	Fraction of RT input, to the total input
$\Lambda_{\mathrm{RT}}^{\mathrm{out}}$	Total RT flows output bit-rate
$\Lambda_{\mathrm{NRT}}^{\mathrm{out}}$	Total NRT flows output bit-rate
$\Lambda_{\Sigma}^{\mathrm{out}}$	Total output bit-rate
ρ	Normalized load coefficient

ρ_{RTOFT}	Normalized lowest load coefficient, when the output consists of RT-only
Ω_ρ	Capacity of system for the normalized load coefficient ρ
w_ϕ	Fraction of flow ϕ input, to the total input
$\mathfrak{U}^{\text{non-frag}}$	Non-fragmented utilization
$\mathfrak{U}^{\text{frag}}$	Fragmented utilization
N_{used}	Total number of used RBs
N_{total}	Total number RBs
B_n^{tx}	Number of transmitted bits on n^{th} RB
B_n^{cap}	Capacity of n^{th} RB
$J\left(w_1\Lambda_1^{\text{out}}, \ldots, w_{\lvert\Phi_{\text{NRT}}\rvert}\Lambda_{\lvert\Phi_{\text{NRT}}\rvert}^{\text{out}}\right)$	Weighted Jain's index of the flow-by-flow output bit-rates

3.2 System Model

A downlink scenario with an OFDMA air interface, which serves RT flows in set Φ_{RT}, and NRT flows in set Φ_{NRT} in a single cell, is considered here. The union of the flows is denoted by Φ as

$$\Phi = \Phi_{\text{RT}} \bigcup \Phi_{\text{NRT}}. \tag{3.1}$$

3.2.1 OFDMA Frame

The total bandwidth is divided into N sub-channels consisting of several OFDMA sub-carriers. Each sub-channel is further divided in time into T time-slots. In this way, the time-frequency plane, for each frame, is divided into NT RBs, each of which spans T_b seconds in time and W_b Hertz in frequency. It is worth mentioning that we extend the OFDMA plane framework to have more than one single RB on specific sub-channel, within a frame, over time. This generalization gives the flexibility of including future technologies, where time-division within a frame is possible.

When this flexibility is not possible, $T = 1$ reduces the model to the conventional OFDMA plane of LTE. Time-division within a frame, if possible, results in higher granularity, and increases the efficiency in resource allocation algorithm.

The transmission frames are indexed by notation k, sequentially. We use the frame to refer to frame index throughout this paper. In frame k, the highest available spectral efficiency and corresponding adaptive modulation and coding (AMC) level, for a single RB on sub-channel j for flow ϕ is

$$b_{\phi}^{(j)}[k] = \mathfrak{f}\left(\mathsf{SINR}_{\phi}^{(j)}[k]\right), \tag{3.2}$$

in bits/sec/Hz, where $\mathsf{SINR}_{\phi}^{(j)}[k]$ is the signal to the interference and noise ratio (SINR) of RBs associated with flow ϕ on sub-channel j in frame k, and $\mathfrak{f}(\cdot)$ represents the AMC table which depends on bit error rate (BER), as well. In Section 3.7, we will use arrays of modulation levels, coding rates, and SINR thresholds which will define an specific $\mathfrak{f}(\cdot)$.

3.2.2 Frame Bit-rate

Radio resources are assigned to the flows in terms of RBs; each RB carries data of only one flow at a time. The bit-rate of a flow is determined from the number of RBs it is allocated in the frame and the AMC level used in each RB. The bit-rate of the flow ϕ, in frame k, is

$$r_{\phi}[k] = W_b \sum_{j=1}^{N} b_{\phi}^{(j)}[k] x_{\phi}^{(j)}[k], \tag{3.3}$$

in bits/sec, where $b_{\phi}^{(j)}[k]$ is the spectral efficiency of RBs on sub-channel j for flow ϕ in frame k, W_b is the frequency span of RB as defined earlier, and $x_{\phi}^{(j)}[k]$ is the number of RBs allocated to flow ϕ on sub-channel j in frame k.

3.3 Joint RT and NRT Flows Scheduling and Allocation Formulation

In this section, we formulate the main joint resource allocation of RT and NRT flows based on disutility functions.

Disutility Functions

Bit-rate utility functions, have been proposed first in [39] which inspired many other works; see, for example, [18, 29, 33, 42, 81, 82]. However, there exists other QoS measures, such as delay, which are independent of bit-rate. It is for the same underlying reason that in order to meet the packet delay deadlines of RT flows, it is not sufficient to only guarantee a minimum mean bit-rate to those flows [70]. Therefore, recently delay has been taken into account as an input to the disutilities [12, 15, 36]. In this study, we use disutility functions with respect to the HOL-delay, queue length information, and bit-rate information for joint resource allocation of RT and NRT flows. RT flows have sensitivities based on HOL-delay while NRT flows sense the mean-delay mainly. The concept of using HOL-delay, beside other information inside disutility functions creates the opportunity to use the framework for QoS classes that may emerge in the future. This concept will be further elaborated in Section 3.4.7.

Main Formulation

The network objective is to minimize the total disutility, $D_{\text{net}}^{\text{joint}}\left(\mathbf{d}^{\text{HOL}}[k], \bar{\mathbf{q}}[k], \bar{\mathbf{r}}[k]\right)$, which depends on the bit-rate vector $\bar{\mathbf{r}}[k]$, queue length vector $\bar{\mathbf{q}}[k]$, and HOL-delay vector $\mathbf{d}^{\text{HOL}}[k]$. The corresponding optimization problem can be casted as

$$\min_{x_\phi^{(j)}[k] \in \mathcal{C}^{\text{PHY}}} D_{\text{net}}^{\text{joint}}\left(\mathbf{d}^{\text{HOL}}[k], \bar{\mathbf{q}}[k], \bar{\mathbf{r}}[k]\right), \quad \text{or} \quad \min_{x_\phi^{(j)}[k] \in \mathcal{C}^{\text{PHY}}} \sum_{\phi=1}^{|\Phi|} D_\phi^{\text{joint}}\left(d_\phi^{\text{HOL}}[k], \bar{q}_\phi[k], \bar{r}_\phi[k]\right), \tag{3.4}$$

where $D_\phi^{\text{joint}}\left(d_\phi^{\text{HOL}}[k], \bar{q}_\phi[k], \bar{r}_\phi[k]\right)$ describes the combined disutility with respect to the HOL-delay (denoted by $d_\phi^{\text{HOL}}[k]$), mean queue length (denoted by $\bar{q}_\phi[k]$), and mean bit-rate (denoted by $\bar{r}_\phi[k]$).

HOL-delay, for flow ϕ, is defined as the delay experienced by the packet at the HOL of the associated queue. Formally, this is the difference of the current frame index, k, and the arrival time-stamp frame index of HOL packet, $t_\phi^{\mathrm{HOL}}[k]$, as

$$d_\phi^{\mathrm{HOL}}[k] = k - t_\phi^{\mathrm{HOL}}[k]. \tag{3.5}$$

It is worth noticing that $d_\phi^{\mathrm{HOL}}[k]$ is different from the mean-delay. We elaborate on mean-delay when describing the MDU algorithm, later in this section. Mean queue length $\bar{q}_\phi[k]$ is defined, based on frame queue length, $q_\phi[k]$, recursively as

$$\bar{q}_\phi[k] = \left(1 - \frac{1}{k}\right) \bar{q}_\phi[k-1] + \frac{1}{k} q_\phi[k]. \tag{3.6}$$

Likewise, mean bit-rate $\bar{r}_\phi[k]$ is defined, based on frame bit-rate, $r_\phi[k]$, recursively as

$$\bar{r}_\phi[k] = \left(1 - \frac{1}{k}\right) \bar{r}_\phi[k-1] + \frac{1}{k} r_\phi[k]. \tag{3.7}$$

As defined earlier, number of frequency sub-channels is denoted by N and number of time slots per frequency sub-channel is denoted by T. The total number of flows is denoted by $|\Phi|$. The constraints are induced by the physical (PHY) layer limitation of RBs in a frame and the fact that scheduler does not map an RB to more than one flow by the feasible set for scheduling variable as

$$\mathcal{C}^{\mathrm{PHY}} = \left\{ x_\phi^{(j)}[k] \middle| \quad \forall\, j:\ \sum_{\phi=1}^{|\Phi|} x_\phi^{(j)}[k] \le T, \quad \forall\, \phi, j:\ x_\phi^{(j)}[k] \in \{0, \ldots, T\} \right\}, \tag{3.8}$$

Note that the scheduling variable $x_\phi^{(j)}[k]$ indicates how many slots in sub-channel j are assigned to flow ϕ, in frame k, which is an integer number in $[0, T]$. In addition, since each frequency sub-channel has T time-slots, the total assignment to any frequency sub-channel should be less than T. The channel information is embedded in the optimization. The dependence of the optimization to the channel information will show itself when we use the gradient of the network disutility

function to develop the resource allocation algorithm in the next section. The disutility functions are non-decreasing in their delay, and queue length arguments and are non-increasing in their bit-rate argument. To the best of our knowledge, the formulation (3.4) is novel in the sense that it incorporates the HOL-delay, mean queue length, and mean bit-rate information.

In the following section, we demonstrate how the framework is the generalization of the static separation, present the ways of choosing disutility functions for RT and NRT flows in the proposed joint approach, and show extensively the perspective with respect to the relevant literature. Section 3.4 will list candidates of resource allocation for RT and NRT from the literature, their main properties, and the underlying reason of their properties in their structure. Note that since we use gradient scheduling, we directly design the gradient of the disutility functions in the next section.

3.3.1 An Example of the Sequential Approach

Conventionally, resource allocation of RT flows ($\phi \in \Phi_{\mathrm{RT}}$) and NRT flows ($\phi \in \Phi_{\mathrm{NRT}}$) is executed based on two *sequential* algorithms, where Φ_{RT} and Φ_{NRT} are set of RT flows and NRT flows, respectively. The sequential approaches result in complete separation of RBs into two sets for RT and NRT flows. In other words, RBs are assigned to RT flows based on an RT scheduler, and if there is any RBs remains, the NRT flows are served. Traditionally, EDF is used for RT flows based on their HOL-delay, and their deadlines. EDF works based on the HOL-delay margin (denoted by $\Delta d_{\phi}^{\mathrm{HOL}}[k]$) which is the difference of the flow's current HOL-delay and its maximum threshold as

$$\Delta d_{\phi}^{\mathrm{HOL}}[k] \triangleq d_{\phi}^{\mathrm{HOL}}[k] - d_{\phi}^{\mathrm{HOL}^{\max}}, \tag{3.9}$$

where $d_{\phi}^{\mathrm{HOL}^{\max}}$ is the HOL-delay deadline of flow ϕ. A flow ϕ is in the safe region when $\Delta d_{\phi}^{\mathrm{HOL}}[k] < 0$. Accordingly, the gradient of dis-utilities of RT flows, in EDF, can be interpreted as

$$\frac{\partial D_{\phi}^{\mathrm{joint}}\Big(d_{\phi}^{\mathrm{HOL}}[k], \bar{r}_{\phi}[k], \bar{q}_{\phi}[k]\Big)}{\partial x_{\phi}^{(j)}[k]} = \Delta d_{\phi}^{\mathrm{HOL}}[k], \quad \text{if} \quad \phi \in \Phi_{\mathrm{RT}}. \tag{3.10}$$

It is worth mentioning that EDF packet scheduling is channel-blind. This is can be noticed from the fact that left hand side of (3.10) is dependent on both sub-channel index j and flow index

ϕ, while the right hand side is not dependent on j. In other words, a flow is scheduled based on EDF irrespective of its sub-channel condition, even if their sub-channels are in deep fade. A channel-aware version of EDF can be defined by its disutility gradient as

$$\frac{\partial D_{\phi}^{\text{joint}}\Big(d_{\phi}^{\text{HOL}}[k], \bar{r}_{\phi}[k], \bar{q}_{\phi}[k]\Big)}{\partial x_{\phi}^{(j)}[k]} = b_{\phi}^{(j)}[k]\ \Delta d_{\phi}^{\text{HOL}}[k], \quad \text{if} \quad \phi \in \Phi_{\text{RT}}. \tag{3.11}$$

Both side of the (3.11) are dependent on ϕ and j for any k. The gradient of the disutility, for any frame index k, forms a two dimensional array where its maximum value plays an important role in decision making. We elaborate on this matter in Section 3.5, and particularly in (3.39).

Having finished RT flows resource allocation, if any RBs are still available, NRT flows are scheduled based on PF scheduler which can be determined as

$$\frac{\partial D_{\phi}^{\text{joint}}\Big(d_{\phi}^{\text{HOL}}[k], \bar{r}_{\phi}[k], \bar{q}_{\phi}[k]\Big)}{\partial x_{\phi}^{(j)}[k]} = \frac{b_{\phi}^{(j)}[k]}{\bar{r}_{\phi}[k]}, \quad \text{if} \quad \phi \in \Phi_{\text{NRT}}, \tag{3.12}$$

where $\partial D_{\phi}^{\text{joint}}\Big(\bar{r}_{\phi}[k], \bar{q}_{\phi}[k], d_{\phi}^{\text{HOL}}[k]\Big)/\partial x_{\phi}^{(j)}[k]$ is the gradient of the corresponding disutility function, in this special case. Note that we use the notation $D_{\phi}^{\text{joint}}\Big(\bar{r}_{\phi}[k], \bar{q}_{\phi}[k], d_{\phi}^{\text{HOL}}[k]\Big)$, as the disutility, for both the sequential approach and for the joint approach. The design of the disutility, or its gradient, with respect to the QoS measurements and QoS requirements, distinguishes the joint from the sequential approach. This unified disutility definition paves the way toward a unified theory for the packet scheduling and RB allocation for addressing heterogeneous traffic.

3.3.2 Proposed Joint Approach

As discussed earlier, the complete separation RBs sets for RT and NRT flows results in system inefficiency. RT flows can presumably have sufficient time with respect to their deadlines while their channel are in deep fade. In this situation NRT flows with good channels can be transmitted. Therefore, joint resource allocation of RT and NRT flows achieves higher multiuser diversity. We propose a generalized framework for disutility functions of RT and NRT flows, towards the joint resource allocation in a common pool of RBs. This framework not only enables the joint resource

allocation of RT and NRT flows, but also paves the way for future unified designs in RT and NRT resource allocation.

In this section, we propose the general structure of the gradient of the disutility for RT and NRT flows. Different information are relevant to RT and NRT resource allocation. HOL-delay, as an instantaneous quantity, plays the important role in RT flows resource allocation. While HOL-delay is the most relevant information in decision making for RT flows, long-term information such as mean bit-rate are relevant for NRT flows.

Design of the Gradient of the Disutility for the RT Flows

For RT flows we define the gradient of disutility function as

$$\frac{\partial D_{\phi}^{\text{joint}}\Big(d_{\phi}^{\text{HOL}}[k],\bar{r}_{\phi}[k],\bar{q}_{\phi}[k]\Big)}{\partial x_{\phi}^{(j)}[k]}=\frac{\Big(b_{\phi}^{(j)}[k]\Big)^{\kappa}}{\digamma_{\phi}^{\bar{\text{r}}}\Big(\bar{r}_{\phi}[k]\Big)}\,\digamma_{\phi}^{\text{d}^{\text{HOL}}}\left(d_{\phi}^{\text{HOL}}[k]\right),\quad \text{if}\quad \phi\in\Phi_{\text{RT}}, \tag{3.13}$$

where κ is the channel-awareness exponent of RT flows, and $\digamma_{\phi}^{\bar{\text{r}}}\left(\bar{r}_{\phi}[k]\right)$ & $\digamma_{\phi}^{\text{d}^{\text{HOL}}}\Big(d_{\phi}^{\text{HOL}}[k]\Big)$ are non-decreasing function, $\mathbb{R}\longrightarrow\mathbb{R}^{+}$, which represents the component of the gradient of the disutility with respect to the mean bit-rate and HOL-delay, respectively. Particularly, we use

$$\digamma_{\phi}^{\text{d}^{\text{HOL}}}\Big(d_{\phi}^{\text{HOL}}[k]\Big)=\vartheta\ \text{e}^{\Delta d_{\phi}^{\text{HOL}}[k]}, \tag{3.14}$$

where ϑ is a scale factor. The channel-awareness exponent of RT flows, κ, can be used to tradeoff the RT output bit-rates with fulfilling different HOL-delay requirement. Parameter κ can also be adjusted for preferring the RT flows in cell edge, instead of the NRT flows's with high channel quality. We use a simple choice of $\kappa=1$, in the simulation section. The image of function $\digamma_{\phi}^{\text{d}^{\text{HOL}}}\Big(d_{\phi}^{\text{HOL}}[k]\Big)$ should be positive ($\mathbb{R}^{+}$) because it is multiplied with channel-awareness factor. Function $\digamma_{\phi}^{\bar{\text{r}}}\left(\bar{r}_{\phi}[k]\right)$ enforces the bit-rate fairness of RT flows. Note that when $\Delta d_{\phi}^{\text{HOL}}[k]$ becomes positive, the deadline has been passed and the corresponding RT flows' packets will be discarded. Therefore,

$$\Delta d_{\phi}^{\text{HOL}}[k]\in\Big(-d_{\phi}^{\text{HOL}^{\max}},0\Big), \tag{3.15}$$

when the packet deletion, due to passed deadline, is on. Moreover, note that the HOL-delay in RT gradient should be amplified so that it can be compared with NRT gradient. With choice (3.14), RT gradient is in the interval of

$$\digamma_{\phi}^{\mathrm{d}^{\mathrm{HOL}}}\left(d_{\phi}^{\mathrm{HOL}}[k]\right) \in \left[\vartheta\ \mathrm{e}^{-d_{\phi}^{\mathrm{HOL}^{\max}}}, \vartheta\right]. \tag{3.16}$$

Design of the Gradient of the Disutility for the NRT Flows

For NRT flows resource allocation, long-term information, namely, mean queue size and mean bit-rate, are relevant. For NRT flows, we form the general structure of the gradient of the disutility function as

$$\frac{\partial D_{\phi}^{\mathrm{joint}}\left(d_{\phi}^{\mathrm{HOL}}[k], \bar{r}_{\phi}[k], \bar{q}_{\phi}[k]\right)}{\partial x_{\phi}^{(j)}[k]} = \min\left[\xi, b_{\phi}^{(j)}[k]\frac{\digamma_{\phi}^{\bar{\mathrm{q}}}\left(\bar{q}_{\phi}[k]\right)}{\digamma_{\phi}^{\bar{\mathrm{r}}}\left(\bar{r}_{\phi}[k]\right)}\right],\ \text{if}\ \phi \in \Phi_{\mathrm{NRT}}, \tag{3.17}$$

where ξ is the instrument for sliding between the complete RT and NRT flow separation and the common pool of RBs, $\bar{q}_{\phi}[k]$ is the mean queue-length of flow ϕ until frame k, $\bar{r}_{\phi}[k]$ is the mean bit-rate of flow ϕ until frame k, $\digamma_{\phi}^{\bar{\mathrm{q}}}\left(\bar{q}_{\phi}[k]\right)$ is mean queue-length importance function, and $\digamma_{\phi}^{\bar{\mathrm{r}}}\left(\bar{r}_{\phi}[k]\right)$ is mean bit-rate importance function. It is worth mentioning that the framework is channel-aware to have spectrally efficient transmissions. Based on this fact, the cell-edge effect is compensated in the proposed framework for their poor links.

The general structures in (3.17) and (3.13) are inspired by keeping the desired pattern in the scheduling design and generalizing the structure in order to exploit existent degrees of freedom. The proposed approach is evolved further in [15] to incorporate the operators interest as well right into the scheduling frameworks.

Flexible Delay Fairness and Bit-rate Fairness

The proposed disutility, in (3.17), with using GPF functions for the NRT flows can be adjusted between sum bit-rate maximization to the maxmim bit-rate fairness for NRT flows. As the counterpart to NRT flows, the proposed disutility, in (3.13), with using WGPF functions for the RT

flows can be adjusted between sum delay minimization to the minmax delay fairness for RT flows (See Chapter 2 or [11] for the proofs).

Design of the Parameter ξ

One of the crucial fact in designing the joint resource allocation of RT and NRT flows is that the NRT flows' disutility gradient have to be bounded. Otherwise, when the RT packets pass their deadlines and are discarded, RT flows cannot compete with NRT flows, where the packets are kept in the queues for much longer time. Therefore, the structure in (3.17) essentially should be bounded. This is done through clipping by ξ. In formulation (3.17), we bound NRT flows disutility gradient by ξ. On the one hand, because of the discarding of RT packets after their deadline, the RT flows disutility gradient value will decrease, after the passed deadlines of RT packets. On the other hand, NRT packets have been kept in their queues, contribute highly to their queue length attributes, and result in increase in their gradient. We design parameter ξ based on the fact that when the RT packets reach π fraction ($0 < \pi < 1$) of their deadline (equivalently pass $1-\pi$ fraction of their deadline), the RT disutility gradient, divided by its channel-awareness term, should be strictly larger than their NRT counterparts as

$$\forall\, k: \quad \max_{\phi\in\Phi_{\mathrm{NRT}}} \frac{\partial D_{\phi}^{\mathrm{joint}}\Big(d_{\phi}^{\mathrm{HOL}}[k], \bar{r}_{\phi}[k], \bar{q}_{\phi}[k]\Big)}{\partial x_{\phi}^{(j)}[k]} < \min_{\phi\in\Phi_{\mathrm{RT}}} \frac{\partial D_{\phi}^{\mathrm{joint}}\Big(d_{\phi}^{\mathrm{HOL}}[k], \bar{r}_{\phi}[k], \bar{q}_{\phi}[k]\Big)}{\partial x_{\phi}^{(j)}[k]}\Bigg|_{k=\pi d_{\phi}^{\mathrm{HOL}^{\max}}}. \tag{3.18}$$

This ensures that, in this fraction, RT flows will be exclusively transmitted. Accordingly, and based on (3.14), the following rule is derived for designing ξ as

$$\xi = \vartheta\, \mathrm{e}^{-\pi \max_{\phi}\left(d_{\phi}^{\mathrm{HOL}^{\max}}\right)}. \tag{3.19}$$

Increasing π makes the scheduler to prioritize RT flows over NRT flows with the cost of lower multiuser diversity. In other words, we divide adaptively the delay margin of RT flows into two regions: Region one where NRT flows can compete with RT flows, based on their utility, and region two where RT flows are given strick higher priority. Nevertheless, note that these two regions are

not necessary separated in time and any combination of RT and NRT flows transmissions can take place, unless when RT flows load does not allow. This enables the joint approach and exploits the hidden multiuser diversity in the conventional designs. Note that parameter π is a design parameter for ξ. Relationship between parameters π and ξ is one-to-one. However, π is interpretable in terms of when RT flows will be given strict priority.

Figure 3.1 depicts the block diagram of the proposed joint approach.

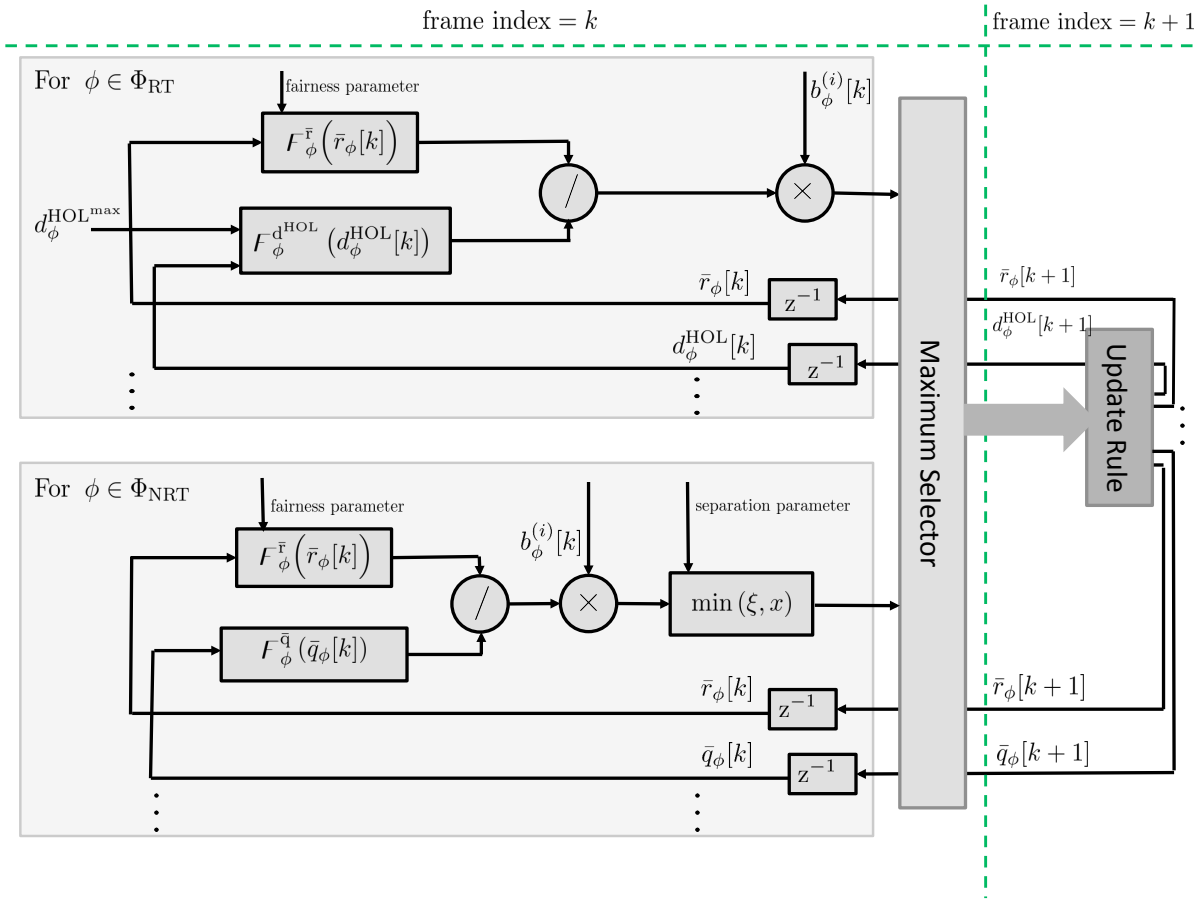

Figure 3.1: Block diagram of the joint RT and NRT packet scheduling and RB allocation.

Note on the Interference

It is worth highlighting that the proposed joint approach, in this chapter, is designed based on the static interference assumption. In fact, the intercell interference coordination (ICIC) [73] works in a much longer timescale, in comparison to the scheduling and allocation algorithm, to specify which RB should be muted or de-muted for each cell or sector. Note that the ICIC schemes often do not aim for the QoS-requirements or sophisticated fairness notions. Therefore, integrating ICIC (See [73] and references within) right into the scheduling and allocation core, through a systematic design, can be suggested as a promising future direction.

3.4 Special Cases

3.4.1 Sequential EDF and GPF

The framework falls back into complete separation of RT, and NRT, served by EDF, and GPF [18, 19, 31, 84], sequentially with the following special choices:

$$\frac{\partial D_{\phi}^{\text{joint}}\Big(d_{\phi}^{\text{HOL}}[k], \bar{r}_{\phi}[k], \bar{q}_{\phi}[k]\Big)}{\partial x_{\phi}^{(j)}[k]} = \Delta d_{\phi}^{\text{HOL}}[k], \quad \text{if} \quad \phi \in \Phi_{\text{RT}} \tag{3.20}$$

and

$$F_{\phi}^{\bar{\text{r}}}\Big(\bar{r}_{\phi}[k]\Big) = \Big(\bar{r}_{\phi}[k]\Big)^{\alpha}, \quad F_{\phi}^{\bar{\text{q}}}\Big(\bar{q}_{\phi}[k]\Big) = 1, \quad \text{for} \quad \phi \in \Phi_{\text{NRT}}, \tag{3.21}$$

or equivalently,

$$\frac{\partial D_{\phi}^{\text{joint}}\Big(d_{\phi}^{\text{HOL}}[k], \bar{r}_{\phi}[k], \bar{q}_{\phi}[k]\Big)}{\partial x_{\phi}^{(j)}[k]} = \min\left[\xi, \frac{b_{\phi}^{(j)}[k]}{\big(\bar{r}_{\phi}[k]\big)^{\alpha}}\right], \quad \text{if} \quad \phi \in \Phi_{\text{NRT}}, \tag{3.22}$$

where $0 \leq \alpha$ is the parameter of GPF that influences the bit-rate fairness. Different types of bit-rate fairness can be achieved by changing α: Cases $\alpha \to 0$, $\alpha = 1$, $\alpha \to \infty$ correspond to sum bit-rate maximization (or max-SINR; first introduced in [94]), PF, and maxmin bit-rate fairness, respectively [31, 46]. The parameter α slides the allocation from no bit-rate fairness to the highest bit-rate fairness in maxmin bit-rate fairness, where it maximizes the minimum mean bit-rate. In

other words, increasing α increases the lower percentiles (such as the 5^{th} percentile) bit-rate at the cost of decrease in higher percentile (such as the 95^{th} percentile) of bit-rate [17, 19]. Parameter α in GPF can be used in a closed loop control system to achieve a target level of fairness [60].

It is worth mentioning that the sequential scheduling (complete static separation) of RT and NRT flows can be achieved within the joint approach when the NRT disutility gradient is clipped by the minimum disutility in RT flows set. This results in selecting ξ as

$$\xi = \min_{\phi \in \Phi_{\text{RT}},\ 1 \le j \le N} \left[\frac{\partial D_{\phi}^{\text{joint}}\Big(d_{\phi}^{\text{HOL}}[k], \bar{r}_{\phi}[k], \bar{q}_{\phi}[k]\Big)}{\partial x_{\phi}^{(j)}[k]} \right], \tag{3.23}$$

which makes the RT and NRT flows disutility gradient disjoint, in their value.

Similar to Section 3.4.1, by substituting EDF with EXP, or MLWDF, the static priority (sequential resource allocation) version of EXP-GPF, or MLWDF-GPF, is constructed. We describe MLWDF-GPF and EXP-GPF in the two next sections.

3.4.2 Sequential MLWDF and GPF

A popular approach for resource allocation of RT flows is the MLWDF [56, 64]. MLWDF can be considered as an advance algorithm in comparison to EDF. In each frame k, a flow $\phi^*[k]$ is selected to be transmitted on an RB on sub-channel $j^*[k]$ according to

$$\Big(\phi^*[k], j^*[k]\Big) = \arg \max_{\substack{\phi \in \Phi_{\text{RT}}, \\ 1 \le j \le N}} \ \mu_{\phi} \frac{b_{\phi}^{(j)}[k]}{\bar{r}_{\phi}[k]} d_{\phi}^{\text{HOL}}[k], \quad \phi \in \Phi_{\text{RT}}, \tag{3.24}$$

where μ_{ϕ} is suggested to be selected as

$$\mu_{\phi} = -\frac{\log \delta_{\phi}}{\tau_{\phi}}, \tag{3.25}$$

by large deviation optimality results of [65], $\bar{r}_{\phi}[k]$ is the mean-bit-rate in frame k, τ_{ϕ} is the maximum allowable delay threshold, and δ_{ϕ} is a maximum probability of exceeding the delay threshold [66] as

$$\Pr\Big(d_{\phi}^{\text{HOL}}[k] > \tau_{\phi}\Big) < \delta_{\phi}. \tag{3.26}$$

It is worth mentioning that by fixing a close-to-one percentile for $1-\delta_\phi$ (or equivalently small δ_ϕ), we can interpret τ_ϕ as the delay deadline which we denoted earlier by $d_\phi^{\mathrm{HOL}^{\max}}$. As an example: For $1-\delta_\phi$ equal to the 99^{th} percentile, we have

$$\mu_\phi \approx \frac{2}{d_\phi^{\mathrm{HOL}^{\max}}}. \tag{3.27}$$

By this interpretation, it can be noticed that the HOL-delays are divided by their deadlines in the structure of MLWDF (see (3.24) and (3.27)). It is in contrast to EDF, where the difference of HOL-delays and their deadlines form the structure.

We note that by selecting the disutility gradient components as

$$\frac{F_\phi^{\mathrm{HOL}}\big(d_\phi^{\mathrm{HOL}}[k]\big)}{F_\phi^{\bar{\mathrm{r}}}\big(\bar{r}_\phi[k]\big)} = \frac{\mu_\phi}{\bar{r}_\phi[k]} d_\phi^{\mathrm{HOL}}[k], \tag{3.28}$$

our framework for RT flows falls back to MLWDF. For a mathematical analysis of MLWDF see [55].

Similar to previous section on EDF and GPF, the sequential MLWDF and GPF is constructed by serving the NRT flows, after RT flows are served, by GPF as described in (3.21).

3.4.3 Sequential EXP and GPF

Another mechanism called EXP [68] has been also proposed for resource allocation of RT flows with a similar structure in MLWDF, but with different dependence on HOL-delay. An OFDMA version of EXP can be represented as

$$\Big(\phi^*[k], j^*[k]\Big) = \arg \max_{\substack{\phi\in\Phi_{\mathrm{RT}},\\ 1\le j\le N}} \frac{b_\phi^{(j)}[k]}{\bar{r}_\phi[k]} \mathrm{e}^{\left(\frac{\mu_\phi d_\phi^{\mathrm{HOL}}[k]}{1+\left(\overline{d^{\mathrm{HOL}}}[k]\right)^\eta}\right)}, \tag{3.29}$$

where in each frame k, a flow $\phi^*[k]$ is selected to be transmitted on an RB on sub-channel $j^*[k]$, $0<\eta<1$, μ_ϕ is defined the same as in (3.27) (inversely proportional to delay deadline), and

$$\overline{d^{\mathrm{HOL}}}[k] = \frac{1}{|\Phi_{\mathrm{RT}}|} \sum_{\phi=1}^{|\Phi_{\mathrm{RT}}|} d_\phi^{\mathrm{HOL}}[k] \tag{3.30}$$

is the average of HOL-delays over RT flows.

The HOL-delay term, $d_\phi^{\text{HOL}}[k]$, in (3.24) and (3.29) can be replaced with queue-length term as $q_\phi[k]$ to obtain the queue-length-driven versions of aforementioned schedulers. Similar to EDF, both MLWDF and EXP are heuristics designed for delay sensitive flows. However, EXP in conjunction with virtual token queues (with constant deterministic arrival rate) can be used to guarantee a minimum bit-rate in resource allocation [51,69]. When queue length of flows are equal or close (see [70] for its formal definition), EXP and MLWDF reduce to PF scheduler [70]. EXP is suitable for the cases where the delay equalization is preferable. However, there has been analysis [71] showing that EXP sacrifices the asymptotic system throughput when the queues grow asymptotically as the cost of emphasis on delay equalization. As discussed earlier, the structure of EXP and MLWDF is based on division of HOL-delay by its deadline, while the structure of EDF is based on the difference of HOL-delay and its deadline.

The sequential EXP and GPF is constructed by serving the NRT flows by GPF as described in (3.21), after RT flows are served.

Note that the gradient of the disutility can be interpreted as the expression inside the arg max in (3.24) and (3.29).

3.4.4 MDU

An algorithm, called MDU, based on two different functions on the Little's delay for RT and NRT flows has been used in [37, 95] for resource allocation of RT and NRT flows. As an example of RT, [37, 95] used

$$\frac{\partial D_\phi^{\bar{\text{d}}}(\bar{d}_\phi[k])}{\partial x_\phi^{(j)}[k]} = b_\phi^{(j)}[k] * \begin{cases} \bar{d}_\phi[k], & \text{if } \bar{d}_\phi[k] \le \frac{\bar{d}_\phi^{\max}}{4}, \quad \phi \in \Phi_{\text{VoIP}}, \\ \left(\bar{d}_\phi[k]\right)^{1.5} - \left(\frac{\bar{d}_\phi^{\max}}{4}\right)^{1.5} + \frac{\bar{d}_\phi^{\max}}{4}, & \text{if } \frac{\bar{d}_\phi^{\max}}{4} \le \bar{d}_\phi[k], \quad \phi \in \Phi_{\text{VoIP}}, \end{cases} \tag{3.31}$$

for the VoIP disutility gradient where $\bar{d}_\phi[k]$ is the Little's delay for flow ϕ in frame k, and $\bar{d}_\phi^{\max}$ is the maximum tolerable Little's delay for flow ϕ. Note that we generalized their utility for general $\bar{d}_\phi^{\max}$ as the mean-delay threshold. Nevertheless, they used $\bar{d}_\phi^{\max} = 100$ ms in simulation experiments.

In parallel, as an example of NRT, [37, 95] used

$$\frac{\partial D_{\phi}^{\bar{\mathrm{d}}}(\bar{d}_{\phi}[k])}{\partial x_{\phi}^{(j)}[k]} = b_{\phi}^{(j)}[k]. \begin{cases} \left(\bar{d}_{\phi}[k]\right)^{0.5}, & \text{if } \bar{d}_{\phi}[k] \leq 100, \phi \in \Phi_{\mathrm{BE}}, \\ 100^{0.5}, & \text{if } 100 \leq \bar{d}_{\phi}[k], \phi \in \Phi_{\mathrm{BE}}, \end{cases} \tag{3.32}$$

for the gradient of the disutility of the best effort (BE) traffic. Notations Φ_{VoIP} and Φ_{BE} denote the VoIP and BE flows set, respectively. Accordingly, the flow $\phi^*[k]$ is selected to be transmitted on an RB in sub-channel $j^*[k]$ based on

$$\Big(\phi^*[k], j^*[k]\Big) = \arg \max_{\substack{1 \leq \phi \leq |\Phi|, \\ 1 \leq j \leq N}} \frac{\partial D_{\phi}^{\bar{\mathrm{d}}}\left(\bar{d}_{\phi}[k]\right)}{\partial x_{\phi}^{(j)}[k]}. \tag{3.33}$$

Note that the Little's delay ($\bar{d}_{\phi}[k]$) is in fact can be approximated [12, 37] by

$$\bar{d}_{\phi}[k] \approx \bar{q}_{\phi}[k] \Big/ \bar{r}_{\phi}[k]. \tag{3.34}$$

Therefore, MDU can be considered as special case of the proposed approach. Nevertheless, since the averaging is a low pass filter and the Little's delay is essentially an averaging mechanism, RT flows suffer when the Little's delay is only used, as an argument, for their disutility functions. In reality however, RT flows sense the HOL-delay, rather than the Little's delay or mean-delay. For handling heterogeneous HOL-delay deadlines, the design needs to incorporate the HOL-delay, which are the relevant delay measures for RT flows. We also note that, since MDU design is merely based on mean-delay, it suffers from bit-rate fairness point of view, in high loads.

3.4.5 Delay Fairness through WGPF

Similar to bit-rate fairness, delay fairness is a concept referring to mechanisms that equalize delay in the system. A resource allocation framework for achieving mean-delay fairness has been studied in [12]. We note that with selection of

$$F_{\phi}^{\bar{\mathrm{r}}}\Big(\bar{r}_{\phi}[k]\Big) = \Big(\bar{r}_{\phi}[k]\Big)^{\nu^{\bar{\mathrm{d}}}}, \quad F_{\phi}^{\bar{\mathrm{q}}}\Big(\bar{q}_{\phi}[k]\Big) = \Big(\bar{q}_{\phi}[k]\Big)^{\nu^{\bar{\mathrm{d}}}}, \text{ for } \phi \in \Phi_{\mathrm{NRT}} \tag{3.35}$$

and $\xi = \infty$, the proposed gradient in (3.17) reduces to the Little's delay disutility function in [12]. Indeed Parameter, $1 \leq \nu^{\bar{\mathrm{d}}}$, is controlling the trade-off between mean-delay fairness and throughput (or resource efficiency). It has been proved in [12] that sufficiently large $\nu^{\bar{\mathrm{d}}}$ corresponds to the *minmax* mean-delay fairness.

On the other hand with selection of disutility over HOL-delay as

$$\frac{F_{\phi}^{\mathrm{HOL}}\big(d_{\phi}^{\mathrm{HOL}}[k]\big)}{F_{\phi}^{\bar{\mathrm{r}}}\big(\bar{r}_{\phi}[k]\big)} = \frac{\big(d_{\phi}^{\mathrm{HOL}}[k]\big)^{\nu^{\mathrm{d}^{\mathrm{HOL}}}}}{\bar{r}_{\phi}[k]}, \quad \text{for } \phi \in \Phi_{\mathrm{RT}}, \tag{3.36}$$

our framework will be reduced to HOL-delay fairness proposed in [59], similar to mean-delay fairness in [12]. The same observation in [12] for trade-off between delay fairness and throughput (or equivalently resource efficiency) can be seen for HOL-delay fairness, by controlling the parameter $\nu^{\mathrm{d}^{\mathrm{HOL}}}$ ($1 \leq \nu^{\mathrm{d}^{\mathrm{HOL}}}$) [60]. Reference [59] used the HOL-delay fairness and GPF mechanism sequentially (or with static separation) for RT and NRT resource allocation.

3.4.6 TUF

An idea based on TUF for joint resource allocation of RT and NRT flows has been proposed in [89–91]. In general, RT scheduler should transmit RT packets any time within their deadline for satisfying the delay requirement, not necessarily with EDF. References [89,91] used the z-shape TUF adopted from [96] as the urgency criterion for RT resource allocation and an efficiency criterion for NRT resource allocation. Based on TUF, RT flows are transmitted, just before their deadline, within a short interval, ι_{RT}, defined as

$$\iota_{\mathrm{RT}} = \Big[d_{\phi}^{\mathrm{HOL}^{\max}} - l_{\phi}, d_{\phi}^{\mathrm{HOL}^{\max}}\Big], \tag{3.37}$$

where $d_{\phi}^{\mathrm{HOL}^{\max}}$ is the flow ϕ deadline and l_{ϕ} determines the length of interval for RT flows transmissions. NRT packets are transmitted during the remaining time interval, ι_{NRT}, defined as

$$\iota_{\mathrm{NRT}} = \Big[0, d_{\phi}^{\mathrm{HOL}^{\max}} - l_{\phi}\Big). \tag{3.38}$$

The TUF approach transmits RT packets near the deadline without channel information. Therefore, it has the shortcoming of not exploiting the multiuser diversity, and it increases the number of dropped packets due to channel-blindness attribute in TUF for RT flows. Reference [97] also used a similar approach in [89] where [97] introduces a transmission guard interval which gives high priority (overriding the NRT packets) to RT packets approaching the delay deadline. Note that in our approach, generally, any combination of RT and NRT packets transmissions can take place in time and RT and NRT packets are not necessary separated in time, in contrast to TUF approach.

3.4.7 Others

References [79, 80] considered the unit-step utility functions for non-BE flows and concave non-decreasing utility functions for BE flows. Mainly, they proved three theorems for bounds on the optimality of their proposed algorithms, based on the inverse of the utility functions. However, [79, 80] did not consider the queue information, or delay information, in their framework. Accordingly, they used bit-rate-based utilities, and sequential resource allocation in mixture of traffic.

Generalized Flow Concept

The joint resource allocation of RT and NRT flows through disutility functions not only increase the efficiency of the system but also has a futuristic application. It is common to have the HOL-delay deadline requirement for RT flows. However most NRT flows are not completely insensitive to delay. Based on our approach one can introduce different levels of delay sensitivity for NRT flows as well. We use the term generalized-non-realtime (GNRT) for those NRT flows which have a mean-delay deadline, $\bar{d}_{\phi}^{\max}$. The concept GNRT is a good model for NRT flows QoS measures, such as the file transfer time. We further elaborate on GNRT concept in [15]. The joint resource allocation of RT and NRT flows enables to define future QoS classes and to accommodate differentiated services between pure RT and pure NRT flows.

3.5 Proposed Joint RT and NRT Scheduling Algorithm

Given the optimization of RT and NRT flows in Section 3.3, we develop the gradient-based algorithm in this section. To make the best change in the objective by increasing only one scheduling variable, the variable with the steepest gradient should be chosen. Note that advanced technologies such as LTE offers high enough granularity that the loss in optimality, due to this step size, is negligible. Since the overall objective is a decreasing function in terms of scheduling variables, $x_{\phi}^{(j)}[k]$, the overall gradient is negative. Therefore, finding the steepest gradient is equivalent to finding the largest absolute value of the gradient:

$$\Big(\phi^*[k], j^*[k]\Big) = \arg \max_{\substack{1\leq\phi\leq|\Phi|,\\ 1\leq j\leq N}} \left| \frac{\partial D_{\text{net}}^{\text{joint}}\Big(\mathbf{d}^{\text{HOL}}[k], \bar{\mathbf{r}}[k], \bar{\mathbf{q}}[k]\Big)}{\partial x_{\phi}^{(j)}[k]} \right|, \tag{3.39}$$

where $D_{\text{net}}^{\text{joint}}\left(\bar{\mathbf{r}}[k], \bar{\mathbf{q}}[k], \mathbf{d}^{\text{HOL}}[k]\right)$ is the objective (the summation of the disutility functions over flows), defined in (3.4), $\bar{\mathbf{r}}[k]$ is the vector of bit-rates, $\bar{\mathbf{q}}[k]$ is the vector of queue lengths, and $\mathbf{d}^{\text{HOL}}[k]$ is the vector of HOL-delays. This equation determines the flow $\phi^*[k]$ to be transmitted on an RB on the sub-channel $j^*[k]$, in frame k. The corresponding values of the (3.39), or equivalently the (3.13) and (3.17), for RT and NRT flows will be denoted by $\Gamma_{\phi\in\Phi_{\text{RT}}}^{(j)}[k]$ and $\Gamma_{\phi\in\Phi_{\text{RT}}}^{(j)}[k]$, respectively, in the proposed algorithm in the sequel.

In this part, we propose an algorithm that we refer to it as **Algorithm** Joint RT-and-NRT Packet Scheduling and RB Allocation. The algorithm is based on the gradient of the objective, which schedules RT flows and NRT flows packets to the OFDMA plane RBs, in an iterative manner. Step 1 makes a copy of AMC bits in frame k and the number of available slots in each sub-channel. Step 2 implements a loop until all RBs are assigned in frame k. Step 3 decides which flow, denoted by $\phi^*[k]$, should be emptied on which sub-channel, denoted by $j^*[k]$, in frame k, based on the largest element in $\Gamma_{\phi}^{(j)}[k]$. Variable $\Gamma_{\phi}^{(j)}[k]$ is equal to the gradient for RT flows and NRT flows for $\phi \in \Phi_{\text{RT}}$ and $\phi \in \Phi_{\text{NRT}}$, respectively. The ties are broken with a uniform random variable. Step 4 implements the decided schedule. The number of unassigned RBs is updated in Step 5. Step 6 updates the delay quantities based on the last decision. In other words, the scheduler

makes decisions one RB at a time, and updates queues and other quantities, such as HOL-delay, after each assignment, and before finding the next flow for next RB. Steps $7-9$ make the AMC copy of the fully occupied sub-channels to zero so that RBs on the corresponding sub-channels are not selected again.

Algorithm JOINT RT-AND-NRT PACKET SCHEDULING AND RB ALLOCATION $\left(b_\phi^{(j)}[k], q_\phi[k], d_\phi^{\mathrm{HOL}}[k], \Phi_{\mathrm{RT}}, \Phi_{\mathrm{NRT}}, N, T, F_\phi^{\bar{\mathrm{q}}}\left(\bar{q}_\phi[k]\right), F_\phi^{\bar{\mathrm{r}}}\left(\bar{r}_\phi[k]\right), F_\phi^{\mathrm{d^{HOL}}}\left(d_\phi^{\mathrm{HOL}}[k]\right), \xi\right)$.

1: $\forall\, j, \phi:\ \tilde{b}_\phi^{(j)} \leftarrow b_\phi^{(j)}[k], \quad \forall\, j : T^{(j)} = T.$
2: **while** $\exists\, T^{(j)} > 0$ **do**
3: $\quad \forall\, \phi \in \Phi_{\mathrm{NRT}}: \quad \Gamma_\phi^{(j)} \leftarrow \min\left[\xi, b_\phi^{(j)}[k]\dfrac{F_\phi^{\bar{\mathrm{q}}}\left(\bar{q}_\phi[k]\right)}{F_\phi^{\bar{\mathrm{r}}}\left(\bar{r}_\phi[k]\right)}\right].$
4: $\quad \forall\, \phi \in \Phi_{\mathrm{RT}}: \quad \Gamma_\phi^{(j)} \leftarrow \left(b_\phi^{(j)}[k]\right)^\kappa F_\phi^{\mathrm{d^{HOL}}}\left(d_\phi^{\mathrm{HOL}}[k]\right) \Big/ F_\phi^{\bar{\mathrm{r}}}\left(\bar{r}_\phi[k]\right).$
5: $\quad \left(\phi^*[k], j^*[k]\right) \leftarrow \arg\max_{\phi, j}\ \Gamma_\phi^{(j)}.$
6: $\quad$ Empty flow, $\phi^*[k]$, to an empty RB on sub-channel, $j^*[k]$.
$\quad x_{\phi^*[k]}^{(j^*[k])}[k] \leftarrow x_{\phi^*[k]}^{(j^*[k])}[k] + 1.$
7: $\quad T^{(j^*[k])} \leftarrow T^{(j^*[k])} - 1.$
8: $\quad$ Update $d_\phi^{\mathrm{HOL}}[k], \bar{q}_\phi[k]$, and $\bar{r}_\phi[k]$ based on the assigned RB.
9: $\quad$ **if** $T^{(j^*[k])} = 0$ **then**
10: $\quad\quad \forall\, \phi \in \Phi : \tilde{b}_\phi^{(j^*[k])} \leftarrow 0.$
11: $\quad$ **end if**
12: **end while**

3.6 Behavioural Study of Input-output Bit-rates

In this section, the behaviour of the RT and NRT output bit-rates versus total input bit-rate will be analyzed. We start by a discussion on how the capacity of the system depends on input as well as the structure of the scheduler. Later, the input model, RT output, and NRT output will be analysed. Finally, the two regions of under-load, moderate saturation, and severe saturation will be identified and explained.

This study sheds light in understanding the system input-output bit-rates dynamics, in identifying different load regions as well as in explaining the simulation results in the next section.

3.6.1 Capacity Definitions and its Dependence on the Load and the Algorithm

The capacity of the system depends on the structure of resource allocation algorithm, as well as its input traffic mixture. In other words, the capacity depends on the way scheduler allocates resources to the RT and NRT mixture in the input, and shapes the mixture in the output. We denote the capacity when the system only allocate resources to RT flows by Ω_{RT}. This case happens when either there is no resource remaining for NRT flows (RT flows occupy the system capacity and the system is in over-load), or the input only consists of RT-only flows. Similarly, the capacity of the system in NRT-only traffic is denoted by Ω_{NRT}. This capacity is achievable when the input only consists of NRT-only flows. Besides Ω_{RT} and Ω_{NRT}, when the system is *saturated for the first time (SFT)*, the capacity is denoted by Ω_{SFT}. The system is at SFT, when the input to the system reaches the point that the system is at the edge of the over-load, and the under-load. This is when the server is full with the lowest input bit-rate. From this point on the system cannot serve the total arrivals. Generally,

$$\Omega_{\mathrm{RT}} \quad \leq \quad \Omega_{\mathrm{SFT}} \quad \leq \quad \Omega_{\mathrm{NRT}}. \tag{3.40}$$

The underlying reason for the first inequality in (3.40) is twofold: First when the system capacity allows to have both RT and NRT flows in the output, the multiuser diversity level is higher than that when RT-only flows are in the output. Second, serving RT-only flows when the system capacity is reached, reduces the opportunity in the scheduler to wait for RBs with better links SINR due to RT flows' delay deadlines. The second inequality in (3.40) is due to the similar higher level of multiuser diversity when we have NRT flows (no deadline) in comparison to the case when we have both RT and NRT flows. In other words, HOL-delay deadlines in RT flows override the opportunistic transmissions and reduce multiuser diversity. Inequality (3.40) will be further explained in Section 3.6.3, after preliminary discussion in Section 3.6.2.

3.6.2 Input Model

Assume that the total input bit-rate, $\Lambda_{\Sigma}^{\text{in}}$, is composed of RT bit-rate, $\Lambda_{\text{RT}}^{\text{in}}$, and NRT bit-rate, $\Lambda_{\text{NRT}}^{\text{in}}$, as

$$\Lambda_{\Sigma}^{\text{in}} = \Lambda_{\text{RT}}^{\text{in}} + \Lambda_{\text{NRT}}^{\text{in}}, \tag{3.41}$$

with

$$\Lambda_{\text{RT}}^{\text{in}} = f_{\text{RT}} \Lambda_{\Sigma}^{\text{in}}, \tag{3.42}$$

and

$$\Lambda_{\text{NRT}}^{\text{in}} = (1 - f_{\text{RT}}) \Lambda_{\Sigma}^{\text{in}}, \tag{3.43}$$

all in bps, where $0 \leq f_{\text{RT}} \leq 1$, similar to input model in [74].

3.6.3 RT Output Bit-rate

Given the aforementioned definitions in the input model, RT output bit-rate, $\Lambda_{\text{RT}}^{\text{out}}$, is limited to its input, $\Lambda_{\text{RT}}^{\text{in}}$, and naturally to the system capacity for RT-only traffic, Ω_{RT}, as

$$\Lambda_{\text{RT}}^{\text{out}} = \min\left(\Lambda_{\text{RT}}^{\text{in}}, \Omega_{\text{RT}}\right). \tag{3.44}$$

We normalize the total input bit-rate to the SFT capacity, when the saturation happens first, in order to define the load coefficient as

$$\rho = \frac{\Lambda_{\Sigma}^{\text{in}}}{\Omega_{\text{SFT}}}. \tag{3.45}$$

Two specific ρ values, namely,

$$\rho_{\text{SFT}} = 1, \tag{3.46}$$

and

$$\rho_{\text{RTOFT}} = \frac{\Omega_{\text{RT}}}{\Omega_{\text{SFT}} f_{\text{RT}}} \tag{3.47}$$

are important in determining the different load regions and capacity of the system. Point, $\rho_{\text{SFT}} = 1$,

is where the saturation happens first. Although the system is saturated for load coefficient larger than 1, the system can still serves a portion of NRT flows until load coefficient ρ_{RTOFT}, where the system output consist of RT-only output for the first time (RTOFT). $\rho_{\mathrm{RTOFT}} = \frac{\Omega_{\mathrm{RT}}}{\Omega_{\mathrm{SFT}} f_{\mathrm{RT}}}$, corresponds to the point where the RT output reaches RT-only capacity. This happens when total input bit-rate is high enough that its RT flows portion ($f_{\mathrm{RT}}\Lambda_{\Sigma}^{\mathrm{in}}$) is equal to RT-only capacity (Ω_{RT}). See (3.42) and (3.44).

Dependence of the Capacity on Input Load

For finding the NRT output bit-rate, we need knowledge on the capacity of the system after the first saturation point. We show the capacity when load coefficient equals specific ρ, from the first saturation point ($\rho_{\mathrm{SFT}} \leq \rho$) until when the RT-only flows occupy the system ($\rho \leq \rho_{\mathrm{RTOFT}}$), by Ω_{ρ}. This capacity is showing the dependence of capacity on input load. The capacity Ω_{ρ} is a non-increasing function over ρ due to the decrease in multiuser diversity and the pressure of RT flows' deadlines after the first saturation. The tighter the RT flows' requirements, the more degradation in capacity when the system serves RT-only flows, in comparison to when it serves NRT flows. In fact, the RT flows' requirements are casted as $d_{\phi}^{\mathrm{HOL}^{\max}}$. The tighter the values of $d_{\phi}^{\mathrm{HOL}^{\max}}$, the lower the capacity of system. A same observation of the degradation in capacity due to RT requirements has been reported in [98]. The discussion on the dependence of the capacity on the input bit-rate of RT and NRT flows mixture is beyond this study. In fact, for many combined resource allocation of RT and NRT flows, the capacity has a complex behaviour versus total input bit-rate (we will observe this in simulation experiments in the Section 3.7). Moreover, this section does not consider the effects of packet deletion due to passed deadline and/or finite-buffer assumptions. However, we show a linear model, with respect to ρ for the general behaviour of the capacity, after SFT until the RTOFT. We will see in simulation section that this model can show the general behaviour of the system.

As discussed earlier, capacity, Ω_{ρ}, is equal to Ω_{SFT} at first saturation load $\rho = \rho_{\mathrm{SFT}}$, and is equal to Ω_{RT} when RT-only output occupies the system for the first time $\rho = \rho_{\mathrm{RTOFT}}$. Therefore,

the general behaviour of the capacity can be represented as

$$\Omega_\rho = \begin{cases} \Omega_{\mathrm{SFT}}, & \rho \leq 1, \\ \frac{f_{\mathrm{RT}}\Omega_{\mathrm{SFT}}(\Omega_{\mathrm{RT}}-\Omega_{\mathrm{SFT}})}{\Omega_{\mathrm{RT}}-f_{\mathrm{RT}}\Omega_{\mathrm{SFT}}}\big(\rho-1\big)+\Omega_{\mathrm{SFT}}, & 1 \leq \rho \leq \frac{\Omega_{\mathrm{RT}}}{\Omega_{\mathrm{SFT}}f_{\mathrm{RT}}}, \\ \Omega_{\mathrm{RT}}, & \frac{\Omega_{\mathrm{RT}}}{\Omega_{\mathrm{SFT}}f_{\mathrm{RT}}} \leq \rho. \end{cases} \tag{3.48}$$

This model shows that the system capacity decreases versus input bit-rate when $1 \leq \rho \leq \frac{\Omega_{\mathrm{RT}}}{\Omega_{\mathrm{SFT}}f_{\mathrm{RT}}}$. Note that we use this model only to show the general system behaviour. The simulation experiments, in the next section, is independent of capacity model in current section. The capacity model from this section neither is used nor is necessary in simulation experiments.

3.6.4 NRT Output Bit-rate

NRT output bit-rate, $\Lambda^{\mathrm{out}}_{\mathrm{NRT}}$, is also limited to its input, $\Lambda^{\mathrm{in}}_{\mathrm{NRT}}$, capacity of the system in NRT-only traffic, Ω_{NRT}, and the remaining capacity after RT flows scheduling (equal to $\max(0, \Omega_\rho - \Lambda^{\mathrm{in}}_{\mathrm{RT}})$), as

$$\Lambda^{\mathrm{out}}_{\mathrm{NRT}} = \min\Big(\Lambda^{\mathrm{in}}_{\mathrm{NRT}}, \Omega_{\mathrm{NRT}}, \max(0, \Omega_\rho - \Lambda^{\mathrm{in}}_{\mathrm{RT}})\Big), \tag{3.49}$$

or by substituting (3.42) and (3.43) in (3.49), as

$$\Lambda^{\mathrm{out}}_{\mathrm{NRT}} = \min\Big((1-f_{\mathrm{RT}})\Lambda^{\mathrm{in}}_{\Sigma}, \Omega_{\mathrm{NRT}}, \max(0, \Omega_\rho - f_{\mathrm{RT}}\Lambda^{\mathrm{in}}_{\Sigma})\Big). \tag{3.50}$$

We note that, since the input bit-rate ($\Lambda^{\mathrm{in}}_{\Sigma}$) is unbounded, the term max in $\max(0, \Omega_\rho - f_{\mathrm{RT}}\Lambda^{\mathrm{in}}_{\Sigma})$ is necessary to keep the remaining capacity non-negative.

3.6.5 Discussion on the RT and NRT Outputs Behaviour

Understanding the behaviour of RT and NRT output bit-rate versus input bit-rate is important. Equations (3.44) and (3.50) describe this behaviour for RT and NRT, respectively. Generally, it is convenient to depict the behaviour of normalized RT and NRT output based on normalized input bite-rate, defined in (3.45), as

$$\frac{\Lambda^{\mathrm{out}}_{\mathrm{RT}}}{\Omega_{\mathrm{SFT}}} = \min\left(f_{\mathrm{RT}}\rho, \frac{\Omega_{\mathrm{RT}}}{\Omega_{\mathrm{SFT}}}\right), \tag{3.51}$$

and

$$\frac{\Lambda_{\mathrm{NRT}}^{\mathrm{out}}}{\Omega_{\mathrm{SFT}}} = \min\Big((1-f_{\mathrm{RT}})\rho, \frac{\Omega_{\mathrm{NRT}}}{\Omega_{\mathrm{SFT}}}, \max(0, 1-f_{\mathrm{RT}}\rho)\Big). \tag{3.52}$$

Normalization helps to focus on the general behaviour and less assumption for absolute values of the capacities.

Figures 3.2 and 3.3 show RT, NRT, and total output bit-rates versus total input bit-rate, all normalized to Ω_{SFT}, when RT input bit-rate is dominant ($f_{\mathrm{RT}} > 0.5$) and vice versa ($f_{\mathrm{RT}} < 0.5$), respectively. The value of $\frac{\Omega_{\mathrm{RT}}}{\Omega_{\mathrm{SFT}}}$ is assumed equal to 0.85 in Figures 3.2 and 3.3 for showing a typical behaviour. The region, from the first saturation point (or the load at SFT) until the output of system consist of RT-only (or the load at RTOFT), is highlighted in Figures 3.2 and 3.3. This region is labeled by `sat. I`.

Note that the absolute value of the system capacities depend on the RT flows's requirements. As an example, for tighter values of $d_{\phi}^{\mathrm{HOL}^{\max}}$, the RT-only capacity degrades more when passing RTOFT point. In other words, for tighter values of $d_{\phi}^{\mathrm{HOL}^{\max}}$, the value of $\frac{\Omega_{\mathrm{RT}}}{\Omega_{\mathrm{SFT}}}$ decreases. As we discussed earlier in this chapter, decreasing the parameter ξ makes the scheduler to go toward the sequential scheduling. Therefore, decreasing ξ decreases Ω_{SFT}. It is worth mentioning that, in this section, we intentionally made general assumption, not confining as much as possible to construct a simple yet effective model to explain the general trends in the input-output bit-rates of the system.

The following Sections 3.6.6 and 3.6.7 elaborate more on RT and NRT output bit-rates versus ρ for both Figures 3.2 and 3.3, in different load coefficient regions.

3.6.6 Under-load Region

As shown in Figures 3.2 & 3.3, RT output bit-rate, $\Lambda_{\mathrm{RT}}^{\mathrm{out}}$ is equal to its input (goes up with slop equal to f_{RT}) until the input reaches RT-only system capacity, Ω_{RT}, and becomes constant at $\Lambda_{\Sigma}^{\mathrm{in}} = \frac{\Omega_{\mathrm{RT}}}{f_{\mathrm{RT}}}$, or equivalently at $\rho = \frac{\Omega_{\mathrm{RT}}}{\Omega_{\mathrm{SFT}} f_{\mathrm{RT}}}$.

NRT output is equal to its input until the input reaches the system capacity for the first time, $\Lambda_{\Sigma}^{\mathrm{in}} = \Omega_{\mathrm{SFT}}$, or equivalently $\rho = 1$. At this point, $\rho = 1$, the NRT output reaches $\Lambda_{\mathrm{NRT}}^{\mathrm{out}} = (1-f_{\mathrm{RT}})\Omega_{\mathrm{SFT}}$, and the RT output reaches $\Lambda_{\mathrm{RT}}^{\mathrm{out}} = f_{\mathrm{RT}}\Omega_{\mathrm{SFT}}$. The system becomes saturated which means that all the resources in OFDMA server are occupied. Note that although the system is

saturated, the output of the system still consist of both RT and NRT flows.

This explanation summarizes the first part of the outputs, where RT output $\frac{\Lambda_{\text{RT}}^{\text{out}}}{\Omega_{\text{SFT}}} \in (0, f_{\text{RT}})$, and NRT output $\frac{\Lambda_{\text{NRT}}^{\text{out}}}{\Omega_{\text{SFT}}} \in (0, 1 - f_{\text{RT}})$. See Figures 3.2 and 3.3 in the region where $\rho < 1$. This region is labeled as `under-load`.

3.6.7 Over-load Regions: Saturation I and Saturation II Regions

For $\Lambda_{\Sigma}^{\text{in}} > \Omega_{\text{SFT}}$, or equivalently $\rho > 1$, the priority of RT flows makes the NRT output bit-rate to break and to go down until $\Lambda_{\Sigma}^{\text{in}} = \frac{\Omega_{\text{RT}}}{f_{\text{RT}}}$, or equivalently $\rho = \frac{\Omega_{\text{RT}}}{\Omega_{\text{SFT}} f_{\text{RT}}}$, despite the increase in its input. In saturation region (or equivalently $\rho > 1$ which is labeled by `saturation`) two subregions can be identified, further labeled by `sat. I` and `sat. II`. For the region where $\frac{\Omega_{\text{RT}}}{\Omega_{\text{SFT}} f_{\text{RT}}} > \rho > 1$ the system is saturated but still is able to pass a portion of NRT flows. This region is depicted in shaded color. In fact, networks are designed to operate in moderate saturation region of sat. II, where the operator benefits from its investment efficiently, while clients can have both RT and NRT flows. For region, $\frac{\Omega_{\text{RT}}}{\Omega_{\text{SFT}} f_{\text{RT}}} \leq \rho$, RT flows merely fill the system capacity and the system will not even be able to pass all RT flows, yet the NRT flows. This region is labeled by `sat. II`.

This explanation outlines the second part of the outputs, where $\frac{\Lambda_{\text{RT}}^{\text{out}}}{\Omega_{\text{SFT}}} \in (f_{\text{RT}}, \frac{\Omega_{\text{RT}}}{\Omega_{\text{SFT}}})$, and $\frac{\Lambda_{\text{NRT}}^{\text{out}}}{\Omega_{\text{SFT}}}$ goes down from $1 - f_{\text{RT}}$ to 0. See Figures 3.2 and 3.3 in the regions where $1 \leq \rho \leq \frac{\Omega_{\text{RT}}}{\Omega_{\text{SFT}} f_{\text{RT}}}$ (shaded color), and $\rho \leq \frac{\Omega_{\text{RT}}}{\Omega_{\text{SFT}} f_{\text{RT}}}$.

Note that the input $\Lambda_{\Sigma}^{\text{in}}$, or its normalized value ρ, is unbounded but f_{RT} is between 0 and 1. System capacity, Ω_{ρ}, after the first saturation ($\rho > 1$) is equal to total output bit-rate, $\Lambda_{\text{RT}}^{\text{out}} + \Lambda_{\text{NRT}}^{\text{out}}$.

In the case, when RT and NRT flows share the equal input bit-rate ($f_{\text{RT}} = 0.5$), RT output, $\Lambda_{\text{RT}}^{\text{out}}$, and NRT output, $\Lambda_{\text{NRT}}^{\text{out}}$ go up with same slop equal to $f_{\text{RT}} = 0.5$ until $\frac{\Lambda_{\Sigma}^{\text{in}}}{\Omega_{\text{SFT}}} = 1$. Then $\frac{\Lambda_{\text{RT}}^{\text{out}}}{\Omega_{\text{SFT}}}$ continues increasing but $\frac{\Lambda_{\text{NRT}}^{\text{out}}}{\Omega_{\text{SFT}}}$ decreases until $\frac{\Lambda_{\Sigma}^{\text{in}}}{\Omega_{\text{SFT}}} = 2$, when RT flows saturate the system and there is no remaining capacity for NRT flows. Due to space limitation, we did not include the graphical representation of this case.

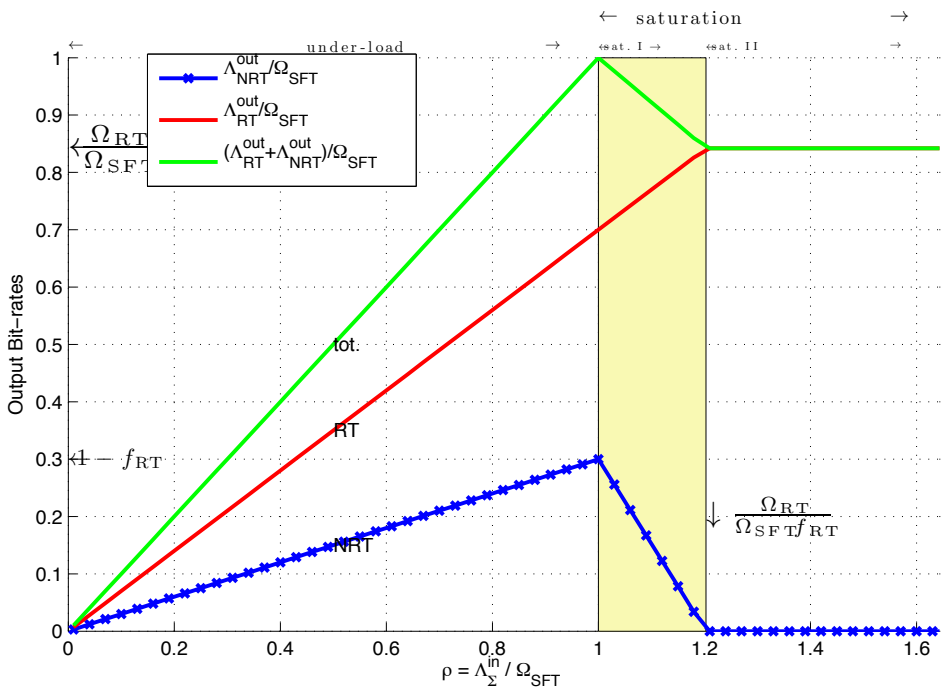

Figure 3.2: Normalized output bit-rates vs. normalized input bit-rate for $f_{\mathrm{RT}} = 0.7$, for behavioural study.

3.7 Simulation

We developed a comprehensive simulation platform for packet delay simulation for a single cell. The platform incorporates correlated fading in time and frequency with Rayleigh fading, shadowing, and path-loss based on [99]. The total OFDMA bandwidth is 10 MHz divided into 20 sub-channels, each sub-channel consist of 20 sub-carriers, each with 25 KHz span in frequency. We used 14 AMC level (including zero) for AMC table (represented by ƒ in (3.2) in the system model) which is the result of QPSK, 16-QAM, and 64-QAM in conjunction with 14 code rates from 0.105 up to 0.801 [100].

3.7.1 SINR Distribution

First, we test the algorithm based on equal average SINR for all flows. Second, we use two-level SINR to test the algorithm in heterogeneous SINR situation. In two-level SINR, we assume that the number of flows is an even number. For each type of traffic, half of the flows have the same average SINR of 15 dB (good flows), and half of them have the same average SINR of 8 dB (bad flows).

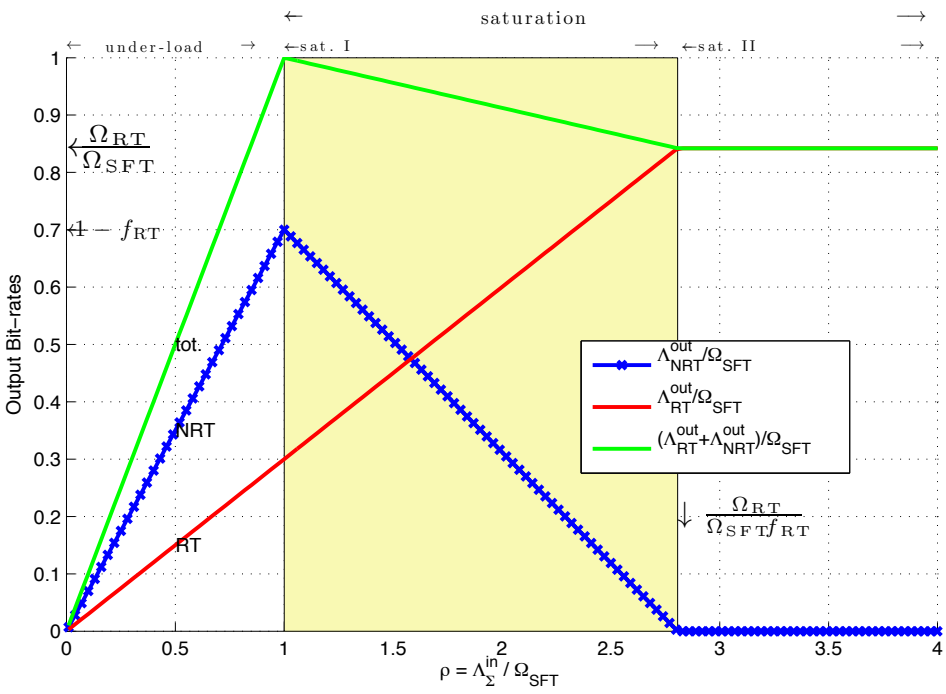

Figure 3.3: Normalized output bit-rates vs. normalized input bit-rate for $f_{\mathrm{RT}} = 0.3$, for behavioural study.

For saving time on wireless channel simulation, we used a three-steps method: First we find a high resolution SINR distribution, resulting from path-loss and shadowing (large-scale fading). Second, one cell is simulated by finding the SINRs from the aforementioned SINR distribution. Third we generate small-scale time and frequency Rayleigh fading, independently for each flow, and add the corresponding average SINRs (result of the large-scale fading, drawn in the second step) in order to get the instantaneous SINR [100]. Simulation parameters are summarized in Table 3.2.

3.7.2 Simulation Assumptions

We tested the proposed joint resource allocation of RT and NRT flows with three arrival scenarios, four RT flows, four NRT flows, and a mixture of two RT flows and two NRT flows. The assumption for delay requirements are 50 ms (or 50 frames) on RT flows. The arrival bit-rates are proportional to $[1, 2, 3, 4]$ according to the total load in the system as $\Lambda_1^{\mathrm{in}} = \frac{1}{10}\Lambda_\Sigma^{\mathrm{in}}, \Lambda_2^{\mathrm{in}} = \frac{2}{10}\Lambda_\Sigma^{\mathrm{in}}, \Lambda_3^{\mathrm{in}} = \frac{3}{10}\Lambda_\Sigma^{\mathrm{in}}$, and $\Lambda_4^{\mathrm{in}} = \frac{4}{10}\Lambda_\Sigma^{\mathrm{in}}$, where $\Lambda_\Sigma^{\mathrm{in}}$ is the total input bit-rate. We investigated output bit-rate and the 99^{th} percentile of the HOL-delay cumulative distribution function (CDF) versus total input bit-

Table 3.2: Simulation parameters for the joint RT and NRT scheduling.

Parameter	**Value**
Fading	Rayleigh [99]
Shadowing	Log-normal, s.d. 5 dB
Doppler shift	37 Hz
Path loss	$38.4 + 2.35 \log_{10}(d)$ dB
Total bandwidth	10 MHz
Number of sub-channels	$N = 20$
Sub-carrier bandwidth	25 kHz
Sub-carriers per sub-channel	20
Slots per frame	$T = 1$
Frame duration	1 ms
Number of flows	$\|\Phi\| = 4$
Cell radius	1000 m
Close-in minimum distance	35 m
Transmit power	30 dBm BS
Antenna gain	5 dBi BS, 0 dBi WTs
Noise figure	2 dB BS, 2 dB WTs

rate. The results of complete separation of RT and NRT flows with EDF-PF are also produced for comparison. Figure 3.4 shows the input scenario to queues (input bit-rates and delay thresholds) for homogeneous average SINR test, as a point for each flow, vs. average SINR (due to shadowing and pathloss) for an specific load. Higher (lower) loads have the same input pattern, but with higher (lower) total input bit-rate.

To find the load range that covers interesting load regions, we estimate the capacity of our OFDMA system based on the 0.3 portion of the highest AMC level on RBs. Based on this estimate we found a load coefficient range by multiplying this capacity from 0.05 to 5. Note that the interval $[0.05, 5]$ is large enough to cover all the interesting load situations.

We assume that the RT packets will be discarded, if their deadlines are passed. We also use a finite-buffer assumption for both RT and NRT flows, equal to 40 Mbits. However, the probability of overflow for RT flows is very low due to their deadline timescale.

3.7.3 RT-only and NRT-only Traffic

For the proof-of-concept, we first test the algorithm with either solely RT or solely NRT flows. We expect the proposed algorithm to reduce to the channel-aware version of EDF (see (3.11)) for RT-

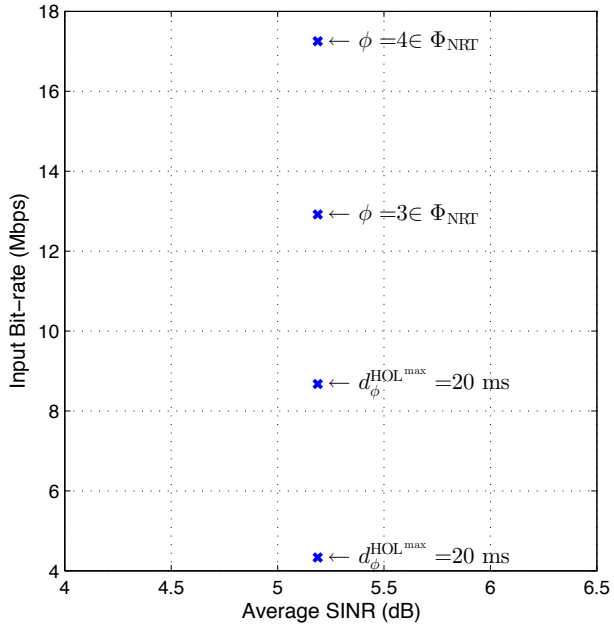

Figure 3.4: Input scenario: Input bit-rates and delay thresholds vs. average SINR for the joint RT and NRT scheduling.

only flows, and PF for NRT-only flows. Note that in RT-only flows, finding the highest value from equation (3.13) with $\kappa = 1$ is equivalent to finding the highest from (3.11). We observed that the proposed algorithm outperforms the baseline (which is EDF for RT flows) algorithm and releases the potential increase in admissible bit-rates without any compromise in the 99^{th} percentile of the HOL-delay. Moreover, we also observed that irrespective of the load situation and link qualities, the output bit-rates per flow are in the same order of the input bit-rates which is showing that the system works as expected. We also tested the proposed approach versus PF for four NRT-only flows. In NRT-only experiment, we observed that in under-load situation the output bit-rates are proportional to input bit-rates which is the result of imposed bit-rate fairness. However, in over-load situation the flows with better wireless links get higher number of RBs in compliance with the expected resource-fairness. Figures for RT-only experiment and NRT-only experiment are not included, due to their less importance in comparison to mixed traffic.

3.7.4 Mixed Traffic Experiment-one

We tested the algorithm in the mixed scenario of two RT flows and two NRT flows. Flows 1 & 2 are RT, and flows 3 & 4 are NRT in this experiment. We tested the proposed approach versus separate EDF and PF in mixed traffic. The output bit-rate per flow, total output bit-rate of RT and NRT flows, and the 99^{th} percentile of the HOL-delay are shown in Figures 3.5, 3.6, and 3.7, respectively.

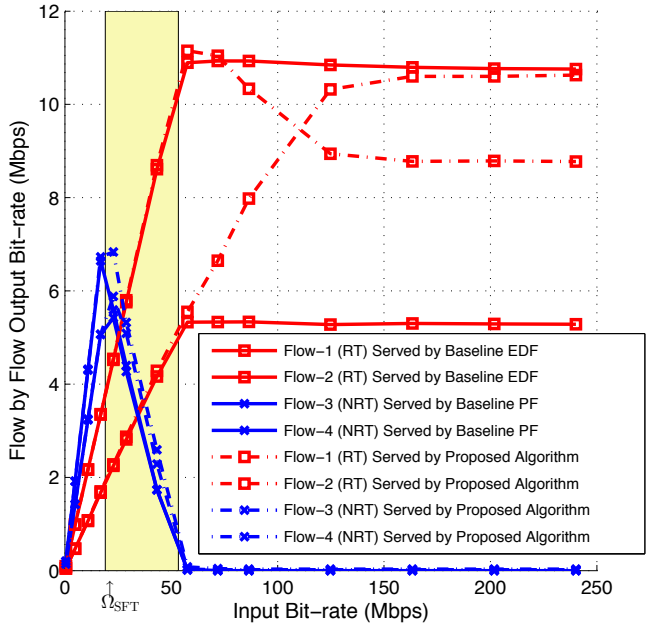

Figure 3.5: Flow by flow output bit-rates vs. input bit-rate in mixed scenario.

Output Bit-rate Performance

Based on the analysis in Section 3.6 and since RT flows occupy

$$f_{\text{RT}} = \sum_{\phi \in \Phi_{\text{RT}}} \Lambda_{\phi}^{\text{in}} \Big/ \sum_{\phi \in \Phi} \Lambda_{\phi}^{\text{in}} \tag{3.53}$$

equal to $\frac{1+2}{1+2+3+4}$, fraction of the total input bit-rate, the saturation of system with RT flows (when there is only RT flows in the output) happens around $\frac{1}{f_{\text{RT}}} = \frac{10}{3}$ in normalized value, for the baseline

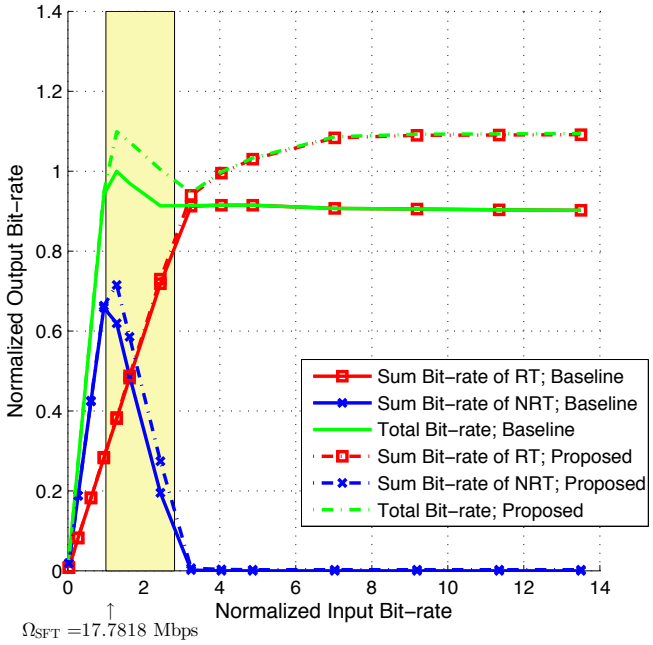

Figure 3.6: Sum RT and NRT output bit-rates vs. input bit-rate in mixed scenario.

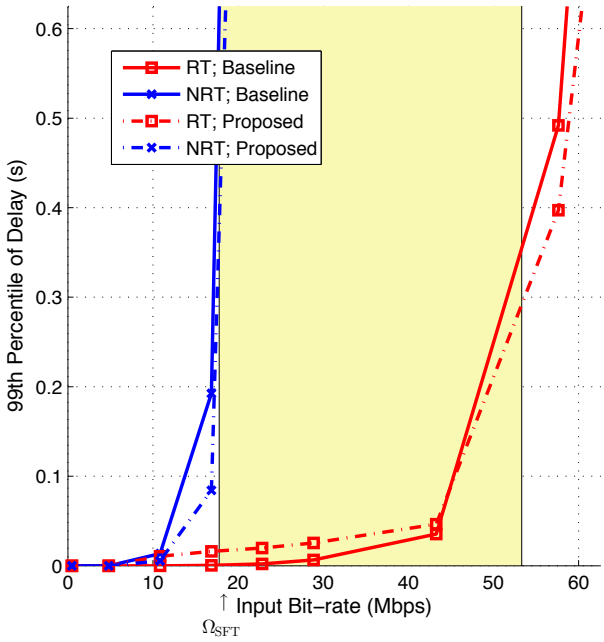

Figure 3.7: The 99th percentile of delay CDF vs. input bit-rate in mixed scenario.

(separate EDF-PF). In addition, NRT output bit-rate at its highest point is around $1 - f_{\mathrm{RT}} = 0.7$ with respect to its normalized value. Figure 3.6 verifies the expected results from Section 3.6. Due to the fact that figures are crowded with plots, we inevitably used coloured plots; Please see the soft copy. As we anticipated, the framework allows for potential increase in admissible bit-rates. The red lines indicate the total output bit-rate of the RT flows. The blue lines indicate the total output bit-rate of the NRT flows. For both red and blue lines, the solid line is the baseline algorithm and dotted line is the proposed algorithm. The green lines are the sum of the red and blue lines which is equal to server capacity, when the system is saturated. In over-load situation, the NRT output bit-rate decreases until RT flows saturates the OFDMA server. This is when the NRT outputs will be zero. When the system goes to saturation II region (or in other words, when the NRT flows output is zero), the RT output (red line) is the same as the green lines (sum of RT and NRT). The reason for the gap, in saturation II region, between solid green (which is equal to solid red because NRT is zero) and dotted green (which is equal to dotted red because NRT is zero) is two fold. In the moderate saturation region (sat. I region), the gain is due to better exploitation of multiuser diversity by the joint RT and NRT scheduling in a common pool of RBs, instead of sequential scheduling. In severe saturation region (sat. II)., the gain is due to channel awareness of the design. As discussed in Section 3.6 networks are designed to operate in moderate saturation region of sat. II, where the operator benefits from its investment efficiently, while clients can have both RT and NRT flows. Especially, the joint approach enable to increase the capacity of moderate saturation region (sat. I).

Delay Performance

Figure 3.7 shows that the average of the delay requirements of RT are not compromised, in a significant portion of the load situation. It is worth mentioning that, based on the range of the input load to the system, one algorithm potentially can have better delay (or output bit-rate) performance with respect to others, even in a single class of RT flows, or NRT flows. In the case of proposed approach, the delay performance is better in a significant portion of the load range. This means that the increase in system capacity is achieved with negligible violation in the delay

requirements. The increase in system capacity is the result of better exploitation of inter- and intra classes multiuser diversity, as well as traffic heterogeneity in time and among flows, in the joint approach. Note that the stationary and reliable delays are valid up to a fraction of the simulation time. Since we run the simulation for 10s, the first 2500 frames in delay are reliable (Figure 3.7). Beyond that, the system is in over-load situation and delays are not stationary and not reliable.

Capacity Dependence on the Load and the Algorithm

We observe that, for static separated EDF-PF when all the links have a same average SINR, the capacity of the system in the first saturation $\Omega_{\mathrm{SFT}} = 17.74$ Mbps is larger than saturation II region $\Omega_{\mathrm{RT}} = 16.14$ Mbps. When the system capacity allows to have both RT and NRT in the output, the multiuser diversity level is higher. Serving RT-only flows with input bit-rate equal to system capacity, reduces the opportunity in the scheduler to wait for better RBs, due to RT flows's delay deadlines. We also observed that when delay deadlines of RT flows becomes tighter, the increase in system capacity due to joint resource allocation of RT and NRT flows vanishes. In other words, the tighter the delay deadlines, the lesser the chances to wait for a good RB thus reducing the multiuser diversity. Note that for the proposed algorithm, some of sophisticated effects in output bit-rate cannot be predicted by the model in Section 3.6: For example, the non-linearity, and the undershoot in Figure 3.6, for the total output bit-rate versus total input bit-rate. This is due to the unavoidable simplification in the Section 3.6, namely, not considering the packet deletion due to passed deadline, static separation assumption, non-linearity of AMC table versus SINR. Having evaluated various resource allocation rules, novel joint RT and NRT resource allocation algorithm that can change their core structure based on the load situation is recommended for future works.

3.7.5 Mixed Traffic Experiment-two

To further test our proposed approach, we reproduced the algorithm MLWDF-PF & EXP-PF (described in Section 3.4) and compared the proposed algorithm against them, in addition to Section 3.7.4. Figures 3.8, 3.9, and 3.10 show the flow-by-flow output bit rates, total output bit-rates, and delay performance, respectively, for the proposed approach against MLWDF-PF. In parallel,

Figures 3.11, 3.12, and 3.13 show the flow-by-flow output bit rates, total output bit-rates, and delay performance, respectively, for the proposed approach against EXP-PF. The same observation in Section 3.7.4 are valid in comparing the proposed joint approach against MLWDF-PF and EXP-PF. The claimed gain in joint approach is also observable for the proposed approach, in sat. I region, in comparison to both MLWDF-PF and EXP-PF, with a comparable delay performances. Nevertheless, we anticipate that in a real network with several flows with many heterogenous delay deadline, higher gain is possible, due to higher level of potent multiuser diversity.

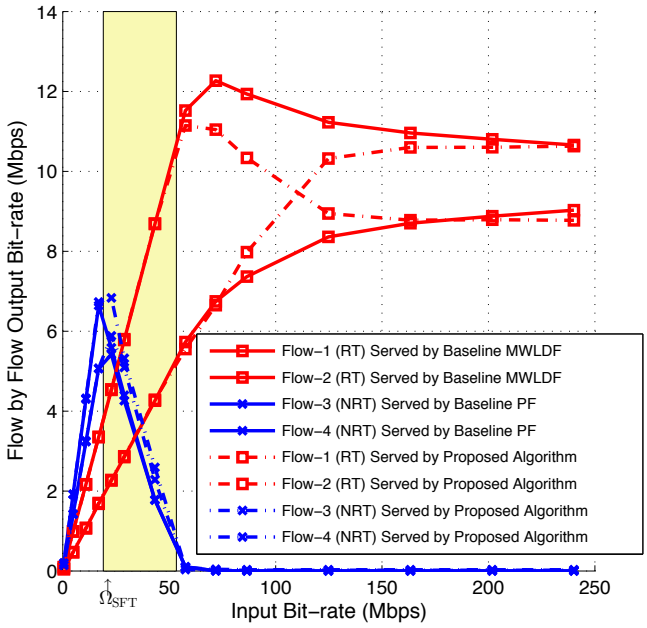

Figure 3.8: Flow by flow output bit-rates vs. input bit-rate in mixed scenario, in comparing MLWDF-PF, with the proposed.

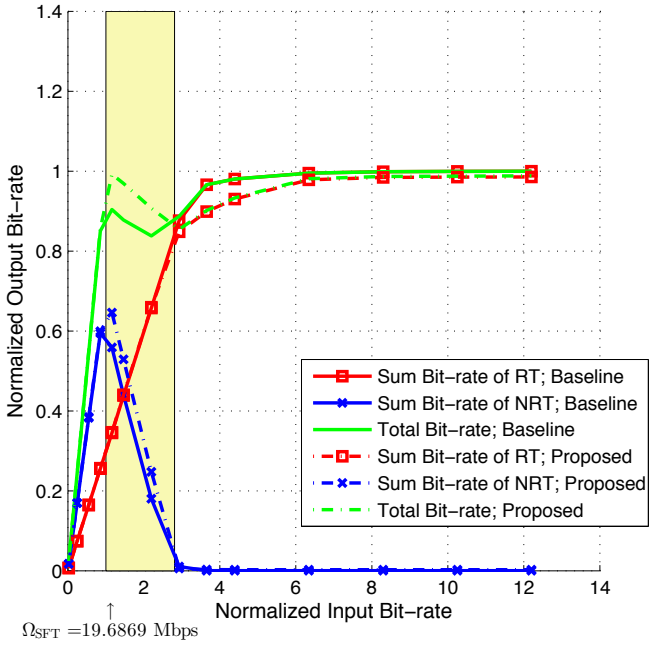

Figure 3.9: Sum RT and NRT output bit-rates vs. input bit-rate in mixed scenario, in comparing MLWDF-PF, with the proposed.

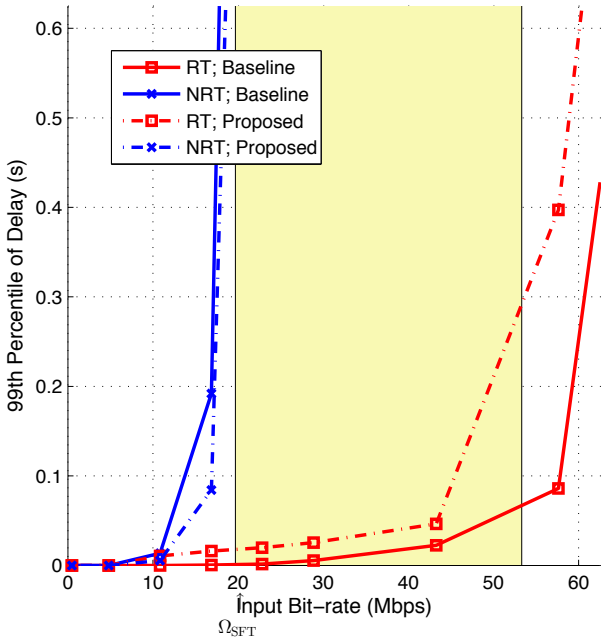

Figure 3.10: The 99th percentile of delay CDF vs. input bit-rate in mixed scenario, in comparing MLWDF-PF, with the proposed.

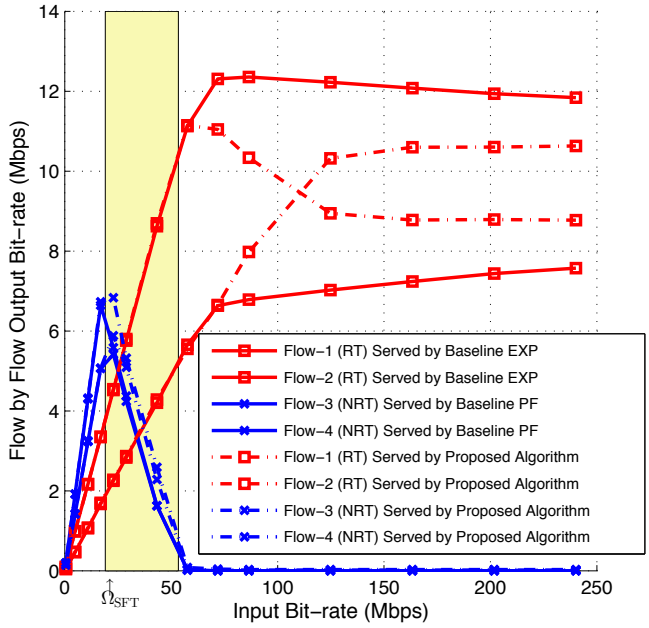

Figure 3.11: Flow by flow output bit-rates vs. input bit-rate in mixed scenario, comparing EXP-PF, with proposed.

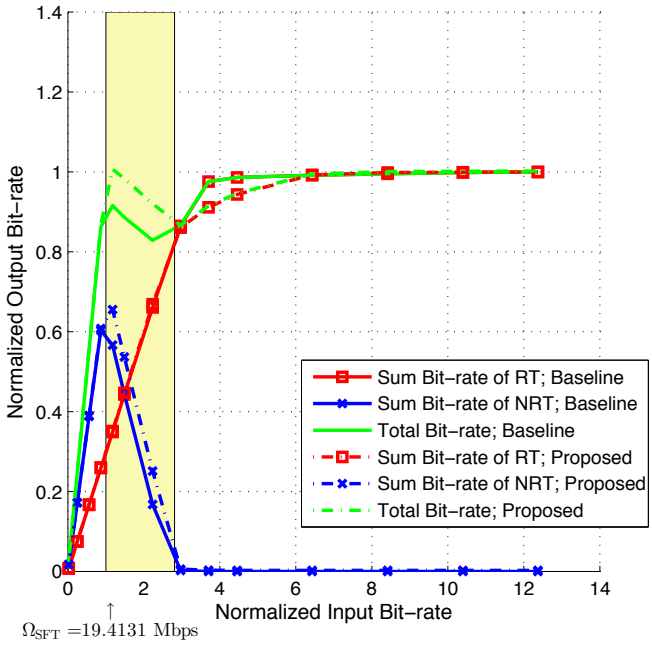

Figure 3.12: Sum RT and NRT output bit-rates vs. input bit-rate in mixed scenario, comparing EXP-PF, with proposed.

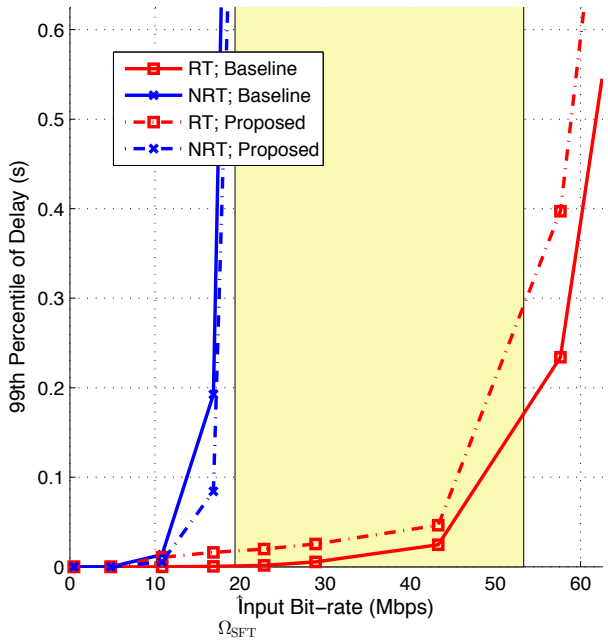

Figure 3.13: The 99^{th} percentile of delay CDF vs. input bit-rate in mixed scenario, comparing EXP-PF, with proposed.

3.7.6 RB Utilization

Generally, channel RB utilization is the ratio of the consumed RBs over the total available RBs in the OFDMA plane. In resource allocation and assigning packets to the RBs, when the number of bits that are assigned to a certain RB is less than its capacity (or AMC available bits that it can carry), it is possible that RB is used (filled) partially. Therefore, we used two measures for RB utilization, one which is just considering the number of RBs, and one that account for the portion that the RBs are filled. As an example when a flow with only a few bits, in its queue, is assigned to an RB, it is possible that the RB is filled partially.

Non-fragmented RB Utilization Measure

The first one, denoted by $\mathfrak{U}^{\text{non-frag}}$ (called non-fragmented utilization), is based on the number of consumed RBs over the total available RBs in the OFDMA plane as

$$\mathfrak{U}^{\text{non-frag}} = N_{\text{used}} \Big/ N_{\text{total}}, \tag{3.54}$$

where N_{used} is the number of used RBs, and N_{total} is the total number of available RBs in a frame.

Fragmented RB Utilization Measure

Since some of the RBs can be used partially, we adjusted a novel RB utilization measure, denoted by $\mathfrak{U}^{\text{frag}}$ (called fragmented utilization), as the average value of the number of the transmitted bits on each RB over the capacity of each RB defined by

$$\mathfrak{U}^{\text{frag}} = \left(\sum_{n=1}^{N_{\text{total}}} B_n^{\text{tx}} / B_n^{\text{cap}} \right) \Big/ N_{\text{total}}, \tag{3.55}$$

where N_{total} is the total number RBs, B_n^{tx} is the number of transmitted bits on n^{th} RB, and B_n^{cap} is the capacity of n^{th} RB. RB utilization depends on the input bit-rate or load.

Figure 3.14 shows the RB utilization versus the input bit-rate. We use the terms non-fragmented utilization, and fragmented utilization for definitions (3.55), and (3.54), respectively, in Figure 3.14. The non-fragmented utilization in the proposed approach is lower than the baseline, although by small margin. This observation has been also reported in [70], where they used EXP for resource allocation of RT and NRT flows. Interestingly, it can be observed that fragmented utilization in the proposed algorithm can use the whole capacity of the available RBs while baseline algorithm is not able to achieve that. This is due to the fact that the proposed algorithm uses the RBs with a higher level of multiuser diversity, in comparison to the baseline algorithm.

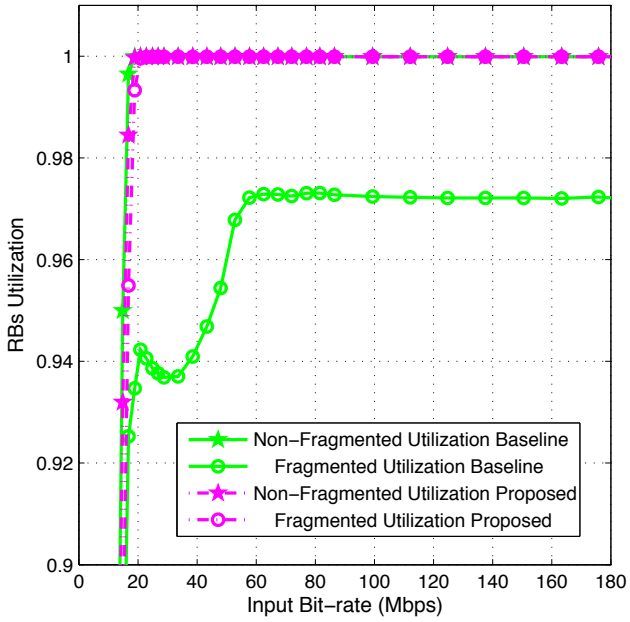

Figure 3.14: Channel RB utilization vs. input bit-rate.

3.7.7 Fairness Evaluation

We also evaluated the Jain's index of bit-rates and the Jain's index of the 99^{th} percentile of the HOL-delay. We observe that the Jain's index of the weighted NRT flows' output bit-rates before the first saturation, the Jain's index of NRT flows' output bit-rates after the saturation, and the Jain's index of the 99^{th} percentile for RT flows in the proposed approach is in the range of 0.93 to 1, depending on the load situation. This proves that NRT flows are treated bit-rate-fair, and RT flows are treated delay-fair. Note that before the first saturation since the there is idle capacity in the system, and importantly since the arrival bit-rates are intentionally heterogenous, the system serves the NRT flows based on their input bit-rates. Therefore, we used a weighted Jain's index [101] for measuring the NRT output bit-rate's fairness, before the first saturation point, as

$$J\Big(w_1\Lambda_1^{\text{out}},\ldots,w_{|\Phi|}\Lambda_{|\Phi|}^{\text{out}}\Big)=\left(\sum_{\phi=1}^{|\Phi|}w_\phi\Lambda_\phi^{\text{out}}\right)^2\Bigg/|\Phi|\sum_{\phi=1}^{|\Phi|}\big(w_\phi\Lambda_\phi^{\text{out}}\big)^2, \tag{3.56}$$

which compensates the input bit-rate heterogeneity. The weights are inversely proportional to the input arrival proportions, $w_\phi = \frac{\Lambda^{\text{in}}_\Sigma}{\Lambda^{\text{in}}_\phi}$. However, after the first saturation point, the system should ignore the input bit-rate heterogeneity and provide the service based on bit-rate fairness. The bit-rate fairness of the NRT flows, after the first saturation, is due to the structure in (3.17), where the term $\digamma^{\bar{r}}_\phi(\bar{r}_\phi[k])$ equalizes the output NRT flow's bit-rates. We used pure Jain's index, or equivalently $w_\phi = 1$, in this load situation to assess the bit-rate fairness.

Interestingly, the bit-rate fairness is observable from output bit-rates in Figure 3.5. In fact, before the saturation, any arrival to the system (either RT or NRT) will be served. In other words, the slope of each red or blacklines is equal to their proportion in the arrival (that is $\frac{1}{10}, \frac{2}{10}, \frac{3}{10}, \frac{4}{10}$). Therefore, the fairness is not interesting or challenging in this under-load situation. When the system goes to saturation II region (when NRT output becomes zero), the fairness enforcement kicks in. Now since the system is full, fairness governs the resource allocation. This is the underlying reason that the output bit-rate of RT flows show convergent property (see Figure 3.5). In fact, for $\phi = 2$ (the one that has the $\frac{2}{10}$ of arrival proportion) the output decreases until it reaches the output of $\phi = 1$ (the one that has the $\frac{2}{10}$ of arrival proportion). The convergence of RTs is because of fairness feature, designed into objective.

3.8 Conclusion

We developed a channel-aware and delay-aware framework, and designed appropriate dis-utilities, for the joint RT and NRT flows packet scheduling and RB allocation. The shape of the disutility function reflects how delay sensitive are the applications with respect to both the HOL- and the mean-delay. The choice of disutility functions results in sorting and making the prioritization within a global set of RT and NRT flows. The effectiveness of the proposed approach is validated by extensive simulations. The simulations shown that the proposed algorithm is able to utilize the potent intra- and inter-class multiuser diversity, and the heterogeneity of traffic in time and among flows when the RT and NRT flows are jointly scheduled, in a common pool of RBs. This increases the effective system capacity without significant compromise in delay performance. It is worth highlighting that whenever there is heterogeneity in the resources, including (but not

limited to) multiuser diversity or any large-scale or small-scale signal variation, the proposed joint RT and NRT scheduling and allocation can offer higher performance, in comparison to the the sequential approach. Therefore, the proposed joint approach is able to offer gains in other medium in addition to mobile wireless networks. We anticipate that whenever increasing the heterogeneity in the requirements, when theres is heterogeneity in the resources, increases the the gain of the proposed joint approach. We also observed that one fixed algorithm rule cannot outperform in output bit-rate and delay performances over all the loads. Accordingly, design of scheduling rule that can change its structure depending on the load situation is suggested as future works. We also developed a novel model for input-output bit-rate behaviour in packet scheduling and resource allocation of the mixture of RT and NRT flows. This model sheds light to the understanding of the system in a simple and intuitive manner, based on non restrictive assumptions. The framework can also be extended to incorporate other QoS vector elements, such as other statistics of the delay, in the objective function. The developed framework, in this chapter, enables putting several algorithms for packet scheduling and resource allocation of RT and NRT flows, in the literatures as well as the proposed approach, in perspective.

Chapter 4

Fair Scheduling with Sub-channel Pairing for Multiuser Amplify-and-Forward Relays in Wireless OFDMA Networks

Abstract

Providing ubiquitous very high data rate coverage in the next generation wireless networks requires cost-effective radio access networks (RAN) devices, such as multiuser enabled amplify-and-forward (AF) relays with proper fair packet scheduling and resource block (RB) allocation. These relays are cost-effective, simpler to implement, and introduce less delay in comparison to other relay based routers. In this chapter, we investigate the packet scheduling and RB allocation algorithms for the orthogonal frequency division multiple access (OFDMA) based AF relays. In the single-user case, the problem reduces to the well known assignment problem, which maximizes the WT bit-rate. For the multiuser case, we devise a generalized proportional fairness (GPF) scheduling framework, which its objective function gradually changes schedules from sum bit-rate maximization to proportionally fair and asymptotically to maxmin fair. Since the relaxed version of the optimization is a convex

problem, we are able to devise a near-optimal gradient-based algorithm to solve it quickly for the non-asymptotic cases. For the asymptotic case, we show that the gradient of the objective function can be simplified and devise a second near-optimal gradient-based algorithm to produce maxmin fair schedules. In addition, we develop the efficient implementation of the above-mentioned two algorithms, by exploiting the super-modularity structure of the general adaptive modulation and coding (AMC) table in AF relay systems. Simulations show that due to their gradient origins both algorithms achieve results very close to the optimum solutions and can tradeoff fairness and efficiency as expected.

4.1 Introduction

Need for the Cost-effective RAN Architectures

Current state-of-the-art wireless standardization activities are leading towards high bit-rates in the order of one gigabit per second in the downlink with a fair coverage. While it is still early for the standardization bodies to consider much higher bit-rates, this is clearly a timely and important research topic due to the exponential growth of the WT traffic on the existing networks. Since wireless channel impairments and transmit power limitations prevent high spectral efficiency even for moderately long links, it is necessary to consider advanced cost-effective radio access network (RAN), such as relay networks, empowered with fair efficient radio resource management (RRM) techniques, which effectively collect and distribute wireless signals. Relay deployment opens the potential space to enhance the cell edge performance. To achieve the full potential of the advanced RANs with fairness, efficient packet and resource block (RB) scheduling techniques are also necessary to match the demand with the limited wireless resources.

Amplify-and-forward Relays as a Cost-effective RAN Element

We consider fair allocation algorithms for RANs with OFDMA-based amplify-and-forward (AF) relays, which multiplex WT data. OFDMA-based AF relays buffer quantized samples of the symbols until they are amplified and transmitted at a later time. These relays are cost-effective, simpler

to implement, and introduce less delay in comparison to the decode-and-forward (DF) relay based routers. As evident in today's networks, implementing hop-by-hop routing is challenging at high bit-rates due to the hardware complexities of fast packet header inspection (See [102] for an example on hop-by-hop routing through software patch for WiFi to take advantage of 802.16 MCF and [103] for method that finds conflict-free TDMA schedules with minimum scheduling delay) AF relaying eliminates these issues from the very high bit-rate wireless networks. AF relays forward data without examining network layer headers, and is possible due to the synchronicity of OFDMA systems. In addition, since the AF relays do not decode the packets, channel decoder delays are eliminated, reducing its impact on higher layers. Therefore, AF relays are good candidates for enhancing the coverage in the next generation of wireless network. Providing a fair RRM framework for this kind of relay is of an great importance. In this chapter, we propose a general framework for fair packet scheduling and RB allocation for AF relays.

Role of the Scheduling

Next generation of wireless networks aim at providing ubiquitous very high bit-rate coverage. Traditional throughput maximization fails to provide fairness and result in scheduling starvation. Therefore, packet scheduling and RB allocation algorithms to exchange the fairness and total throughput needs to be developed for OFDMA-based AF relays.

Related Works

Previous research show that scheduling for AF relay networks holds great promise in the single-user setup [104–110]. With a single-user, the scheduling problem becomes matching the input sub-carriers to the output sub-carriers to maximize the sum bit-rate capacity, a process called sub-channel pairing. Since this problem is equivalent to the graph-theoretic assignment problem, it can be solved by the Hungarian algorithm [106, 109]. However, due to the special structure of the problem, a solution can also be obtained by matching input and output sub-carriers which were first sorted according to their spectral efficiency [104, 105]. This technique is known as ordered sub-channel pairing (OSP). A sufficient condition on the objective function of the related optimization

for optimality of the OSP is found in [110]. This condition also includes the minimization of bit error rate (BER) in a high signal to the noise ratio (SNR) regime [111], in addition to the sum channel capacity objective. Extensions taking interference information into account are also possible [104], as well as extensions that include power allocation [112–118]. However, since the equal power allocation achieves similar performance of adaptive power allocation in OFDMA systems [119], we use equal power allocation to reduce the implementation complexity. Relay selection in multiple relay scenario with single-user is considered in [120, 121]. Throughout this chapter, we assume a single relay.

Resource allocation through OSP for the DF relaying is considered in [122, 123] where adaptive modulation and coding (AMC) levels are kept unchanged. Approaches with changing the AMC levels are presented in [124, 125] where fairness is also considered. Distributed scheduling and power allocation for uplink OFDMA relaying is examined in [126] with game theoretic approaches. Proportional fair resource allocation for OFDMA-based DF relay networks through the objective function is studied in [127, 128]. Long-term proportional fair allocation for AF relays, based on the separate sub-channel assignment to the WTs and sub-channel permutation, is considered in [129].

Approaches based on incorporating fairness in the constraints is studied in [130–133] where throughput maximization, subject to fairness constraint with minimum predefined bit-rate, is adopted. Note that adding fairness through constraints has a disadvantage of losing the meaning of achieved fairness. Sub-carrier allocation both in AF and DF with graph theoretical approach is considered in [134], using network flow programming. The authors focus on maximizing the network throughput subject to load balancing among relays, rather than satisfying WT bit-rates, where it assumes the paired source and relay transmit on the same subcarrier. Extension of [134] considering fairness among WTs, in addition to load balancing among relays, is addressed in [135]. Dual approach is adopted in [133, 136], where the same formulation in [137] is considered.

Nevertheless, most of the discussed works, including [130–137], consider the same sub-channel on the second-hop which limits the capacity. In this chapter, we consider the most general case where any pairing between the first-hop and the second-hop are allowed, in a multiuser setup. AF scheduling in the multiuser setup is not as simple as scheduling in the single-user setup. Compli-

cations arise from the need to provide end-to-end fairness among WTs. Traditional throughput maximization fails to provide fairness and result in scheduling starvation. Therefore, algorithm to exchange the system throughput with fairness needs to be developed. In addition, unlike other works, our upper bound on the performance of the RRM algorithms, allows us to examine how far the algorithms are from the optimally fair bit-rates, which cannot be addressed if algorithms are developed in an ad-hoc manner.

Chapter Contributions

This chapter presents three main contributions to the multiuser AF relay scheduling.

1. We devise a generalized proportional fairness (GPF) scheduling and allocation framework for AF relay scheduling. Flows, from different WTs, are assigned utility functions, which take the bit-rate and a parameter α as inputs. According to the value of the α parameter, the utilities are able to gradually change resource allocation from throughput optimal, to proportionally fair, and to maxmin fair. To the best of our knowledge, our work is the first to consider this type of flexible scheduling and allocation framework in the context of AF relaying.

2. Since finding GPF schedules are computationally hard, we propose two algorithms to quickly find schedules in each frame. The first algorithm is based on the gradient of the α-fair utility functions, so it is similar to the proportionally fair scheduling algorithm [16], which was proposed for conventional cellular time division multiple access (TDMA), and OFDMA networks. However, unlike [16], which finds *long-term* fair bit-rates, our algorithm finds *short-term* fair bit-rates in each frame. The other algorithm is based on our observation that as α becomes large, the steepest gradient corresponds to the flow with minimum bit-rate. We evaluate the performance of our algorithm with extensive simulations and show that the algorithm works close to the optimal solution. In our simulations, we have observed that the two algorithms have a greater usefulness than simply a way to allocate the bit-rates. In effect, by changing the value of α, the algorithms achieve something similar to cell-breathing. Instead of using power control to change the size of the cell, our RRM technique achieves it by combining AMC with time and channel allocation.

3. We develop the efficient implementation of the above-mentioned algorithms by exploiting the super-modularity structure of the AMC table in AF relay systems.

Chapter Organization

This chapter organized based on five main parts:

1. Introduction, motivation, and the related works are explained in Section 4.1.
2. The specific system model and definitions, for AF relay systems, will be given in Section 4.2.
3. The formulation of the fair packet scheduling and RB allocation will be described in Section 4.3.
4. The proposed algorithms for the fair packet scheduling and RB allocation will be described in Section 4.4, where the GPF algorithm for AF relay is discussed in Section 4.4.1, the maxmin algorithm for AF relay is developed in Section 4.4.2, and the efficient implementation of the algorithms is investigated in Section 4.4.3.
5. Finally, simulation experiments will be presented in Section 4.5.

List of Symbols

In this section, we summarize the symbols used throughout this chapter, with a short definition of them in Table 4.1

Table 4.1: List of symbols used in the multiuser AF relay fair scheduling.

Symbol	**Definition**
Φ	Set of all flows
N	Number of frequency sub-channels
T	Number of time slots per frequency sub-channel
T_b	Time span of each RB in second
W_b	Frequency span of each RB in Hertz

ϕ	Flow index
i, j	Sub-channel index
k	Frame index
$B_{\phi}^{(i,j)}[k]$	Shannon capacity in sub-channel pairing of (i, j)
$b_{\phi}^{(i,j)}[k]$	Combined AMC value over two hops
$x_{\phi}^{(i,j)}[k]$	Number of allocated RB to flow ϕ on sub-channel pairing (i, j)
$S_{\text{net}}^{\text{r}}\left(\ldots, x_{\phi}^{(i,j)}[k], \ldots, \alpha\right)$	Sum of the frame bit-rate utility, showing dependence to the scheduling variable
$S_{\text{net}}^{\text{r}}\left(\mathbf{r}[k]\right)$	Sum of the frame bit-rate utility
$\mathsf{SNR}_{\phi}^{(i,j)}[k]$	Combined SNR in sub-channel pairing of (i, j) in AF relay
$\mathsf{SNR}_{\text{R}}^{(i)}[k]$	First-hop SNR at RS on sub-channel i for AF relay
$\mathsf{SNR}_{\phi}^{(j)}[k]$	Second-hop SNR at flow ϕ, for AF relay
$\mathfrak{f}(\cdot)$	Function describing the AMC table, from capacity to AMC values
$r_{\phi}[k]$	Frame bit-rate, for flow ϕ, in frame k
$S_{\phi}^{\text{r}}\left(r_{\phi}[k]\right)$	Frame bit-rate utility, for flow ϕ
$\mathcal{C}^{\text{PHY}-\text{AF}}$	Feasible set for AF relay scheduling
α	GPF parameter
$J\big(r_1[k] \ldots, r_{\lvert\Phi\rvert}[k]\big)$	Jain's fairness index
$\Big(i^*[k], j^*[k], \phi^*[k]\Big)$	Selected first-hop subcarrier, second-hop sub-carrier, and flow in frame k, for AF relay scheduling
$T_{\text{BS}}^{(i)}, T_{\text{RS}}^{(j)}$	Algorithm internal variables, for AF relay
$\tilde{b}_{\phi}^{(i,j)}$	Algorithm internal variables, for AF relay
$\hat{b}_{\phi}^{(i,j)}[k]$	Auxiliary variable used in the maxmin proposition
$\delta_{\phi}[k], \delta_{\text{H}}[k]$	Similarities test, for AF relay
$\Delta_{\text{H}}[k]$	Sub-optimality gap, for AF relay
$h\left(\begin{bmatrix} F^{(1)} \\ S^{(1)} \end{bmatrix}\right)$	Super-modular example function

$\text{sort}^{\downarrow}$	Vector sorting notation
$\mathfrak{P}_{\text{R}}(l)$	Permutations for sorting first-hop SNR vectors, for AF relay
$\mathfrak{P}_{\phi}(l)$	Permutations for sorting first-hop SNR vectors, for AF relay
$\widehat{\mathsf{SNR}}_{\phi}^{(l)}$	Sorted SNR in the first-hop, for AF relay
$\widehat{\mathsf{SNR}}_{\text{R}}^{(l)}$	Sorted SNR in the second-hop, for AF relay
$\mathfrak{P}_{\phi}^{-1}(l)$	Inverse permutation for sorting first-hop SNR vector, for AF relay

4.2 AF System Model

OFDMA-based AF Relay

The orthogonal sub-carriers are grouped in time and frequency into RBs, with duration of T_b seconds and a frequency span of W_b Hertz. There are N available sub-channels, each with T time slots. The relay station (RS) receives the signal from the base station (BS), samples it, performs the fast-Fourier transform to get the received modulation symbols on each sub-carrier, and stores them in its buffer. After receiving the signal for T_b seconds, the RS has one RB in its buffer for each sub-channel, so it may re-map the RBs to different sub-channels, before performing the inverse fast Fourier transform (FFT) to obtain the output signal, similar to single-user chunk-based sub-channel pairing [106]. Because of the buffering, the RS has NT RBs before re-transmitting them to WT, allowing for scheduling of multiple WTs in the same frame and on the same sub-channel (see Figure 4.1). We assume a network of Φ flows, connected to the BS through a predetermined RS at any given time. A higher layer process determines which Φ flows are connected to the BS through this RS. Each WT is assumed to have a single flow, in this chapter.

Combined Double-hop AMC and SNR Values and Sub-channel Pairing

The number of bits that can be carried in an RB depends on the AMC used in the combined transmission over the two hops. We denote the number of bits transmitted in an RB, allocated to flow ϕ on sub-channel pairing (i, j), with $b_{\phi}^{(i,j)}[k]$. The sub-channel pairing (i, j) refers to the

transmission from BS to the RS on sub-channel i coupled with the transmission from the RS to the WT, with flows, on sub-channel j. We assume that the time coherence of the channels are less than frame duration. Therefore, $b_{\phi}^{(i,j)}[k]$ remains the same for all RBs in sub-channel pairing (i,j) for flow ϕ. AMC values $b_{\phi}^{(i,j)}[k]$s are obtained from a table which maps combined double-hop signal-to-noise ratios (SNRs) to the appropriate AMC. We only consider scheduling on the downlink; uplink scheduling is identical.

The highest AMC value available on the combined link from the BS to the RS and from RS to the WTs, in an RB, depends on the combined SNR in its time interval and frequency span. The combined end-end SNR of the double-hop transmission is

$$\mathsf{SNR}_{\phi}^{(i,j)}[k] = \frac{\mathsf{SNR}_{\mathrm{R}}^{(i)}[k]\ \mathsf{SNR}_{\phi}^{(j)}[k]}{\mathsf{SNR}_{\mathrm{R}}^{(i)}[k] + \mathsf{SNR}_{\phi}^{(j)}[k] + 1}, \tag{4.1}$$

where $\mathsf{SNR}_{\mathrm{R}}^{(i)}[k]$ is the SNR from the base-station to the relay on channel i and $\mathsf{SNR}_{\phi}^{(j)}[k]$ is the SNR from the relay to the flow ϕ on channel j (See [138] for further physical (PHY) layer analysis of the AF relay). The highest available AMC bit-rate for a single RB, in each frame, on a specific sub-channel pairing is determined as

$$B_{\phi}^{(i,j)}[k] = \log\left(1 + \mathsf{SNR}_{\phi}^{(i,j)}[k]\right), \tag{4.2}$$

in bits/Hz/sec. Without any loss of generality, we assume that the bit-rate achieved by transmitting a single RB in each frame on a specific sub-channel pairing is a function of the highest available bit-rate as

$$b_{\phi}^{(i,j)}[k] = \mathfrak{f}\left(B_{\phi}^{(i,j)}[k]\right), \tag{4.3}$$

in bits/Hz/sec, where $\mathfrak{f}(\cdot)$ represents the AMC mapping from the Shannon capacity to the AMC value. This definition of the AMC mapping is helpful in this chapter in order to exploit the structure. The term $b_{\phi}^{(i,j)}[k]$ is the number of bits that can be transmitted to flow ϕ on sub-channel pairing (i,j) in the frame k.

Frame Bit-rate

Radio resources are assigned to the flows in terms of RBs. Each RB carries data of only one flow at a time. The bit-rate of a flow ϕ is determined from the number of RBs it is allocated in the frame and the AMC used in each RB as

$$r_\phi[k] = W_b \sum_{i=1}^{N} \sum_{j=1}^{N} b_\phi^{(i,j)}[k]\; x_\phi^{(i,j)}[k], \tag{4.4}$$

in bits/sec, where $x_\phi^{(i,j)}[k]$ is the number of RBs assigned to flow ϕ on sub-channel pairing (i, j), and T_b is the time span of a RB (see Figure 4.1). We use $x_\phi^{(i,j)}[k]$ to indicate that the slot allocations are the unknowns the algorithm is searching for. If we limit each sub-channel in a frame to be allocated to only one WT, $x_\phi^{(i,j)}[k]$ is equal to one if the sub-channel pairing (i, j) is allocated to flow ϕ in frame k, otherwise it is zero (equivalently $T/2$ and 0).

A Simple AF Schedule Example

To better reflecting the mechanism of the OFDMA-based AF relay and explaining the definition of the $x_\phi^{(i,j)}[k]$, in this chapter, we use an example from Figure 4.1. Figure 4.1 shows that 7 RBs in the sub-channel 1 are coupled with the sub-channel 4 in the second-hop conveying the data of flow $\phi = 2$, defining $x_2^{(1,4)}[k] = 7$. The rest of the sub-channel 1 in the first-hop, which consists of 3 RBs, is coupled with the sub-channel 5 in the second-hop conveying the data of flow $\phi = 3$, resulting to $x_3^{(1,5)}[k]$. The main question, we answer in the rest of the work, is how to allocate the RB to the flows for each sub-channel pairing, $x_\phi^{(i,j)}[k]$.

4.3 Proposed Formulation

4.3.1 Flexible Fairness Criteria

From the networking perspective the bit-rates should satisfy some type of fairness, otherwise the network operator may have too many unhappy flows who are starved out by the WTs with high AMC values. Instead of requiring strict proportional fairness [42], we use a more general fairness,

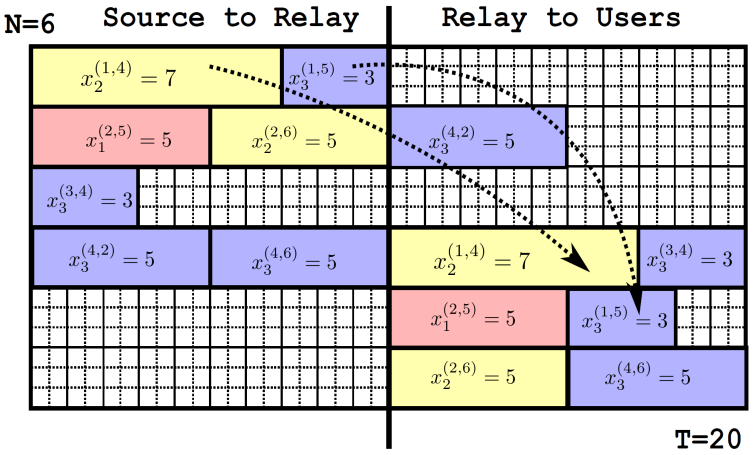

Figure 4.1: An example for AF schedule.

where the network operator has the flexibility to modify the scheduler to exchange fairness for throughput. To quantify the flow utility with the resources it is given, each flow is assigned a utility function. When the network utility is maximized over all possible bit-rates, the bit-rates are called fair with respect to the utilities [31].

An important family of utility functions, which results in a range of fairness notions, is defined as

$$S_\phi^{\mathrm{r}}\left(\ldots,x_\phi^{(i,j)}[k],\ldots,\alpha\right)=\begin{cases}\frac{1}{1-\alpha}\left(W_b\sum_{i=1}^{N}\sum_{j=1}^{N}b_\phi^{(i,j)}[k]x_\phi^{(i,j)}[k]\right)^{1-\alpha}, & \text{if } \alpha\neq 1\\ \log\left(W_b\sum_{i=1}^{N}\sum_{j=1}^{N}b_\phi^{(i,j)}[k]x_\phi^{(i,j)}[k]\right), & \text{if } \alpha=1,\end{cases} \tag{4.5}$$

where $\alpha\geq 0$ is the parameter influencing the kind of fairness and the term in the brackets is the flows bit-rates as given in (4.4). The constant factor is necessary to make the utility concave for $\alpha>1$. The sum utility makes the network utility, as

$$S_{\mathrm{net}}^{\mathrm{r}}\left(\ldots,x_\phi^{(i,j)}[k],\ldots,\alpha\right)\triangleq\sum_{\phi=1}^{|\Phi|}S_\phi^{\mathrm{r}}\left(\ldots,x_\phi^{(i,j)}[k],\ldots,\alpha\right). \tag{4.6}$$

Bit-rates, which maximize the sum utility for a specific α are said to be α-fair [31].

Different types of fairness can be achieved by changing the parameter α to trade the fairness and the throughput [31]. For $\alpha = 0$ the utilities become an identity function in terms of bit-rates, i.e., $S_{\phi}^{\mathrm{r}}(r_{\phi}[k]) = r_{\phi}[k]$, and the network utility corresponds to *throughput* given as

$$S_{\mathrm{net}}^{\mathrm{r}}\left(\ldots, x_{\phi}^{(i,j)}[k], \ldots, \alpha\right) \underset{\alpha \to 0}{=} W_b \sum_{\phi=1}^{|\Phi|} \sum_{i=1}^{N} \sum_{j=1}^{N} b_{\phi}^{(i,j)}[k]\; x_{\phi}^{(i,j)}[k]. \tag{4.7}$$

For $\alpha \to 1$, the bit-rates maximizing the network utility are *proportionally fair* (PF) [31] which make the network utility as

$$S_{\mathrm{net}}^{\mathrm{r}}\left(\ldots, x_{\phi}^{(i,j)}[k], \ldots, \alpha\right) \underset{\alpha \to 1}{=} \sum_{\phi=1}^{|\Phi|} \log\left(W_b \sum_{i=1}^{N} \sum_{j=1}^{N} b_{\phi}^{(i,j)}[k] x_{\phi}^{(i,j)}[k]\right). \tag{4.8}$$

Finally as $\alpha \to \infty$, the network utility leads to a maxmin fair allocation of bit-rates. This can be shown also by noting an interesting relationship between the network utility maximization and the L_p norm minimization of the inverse of the bit-rates..

4.3.2 Proposed Formulation

We now formulate the optimization problem that finds time allocations for the AF relay resulting in α-fair bit-rates. We call a set of bit-rates α-fair, if for a given α they maximize the network utility (4.6) over all possible bit-rates (sub-channel pairing). The optimization, which maximizes network utility over all feasible bit-rates to find the α-fair rates is

$$\max_{x_{\phi}^{(i,j)}[k] \in \mathcal{C}^{\mathrm{PHY-AF}}} \quad \sum_{\phi=1}^{|\Phi|} \frac{1}{1-\alpha} \left(W_b \sum_{i=1}^{N} \sum_{j=1}^{N} b_{\phi}^{(i,j)}[k] x_{\phi}^{(i,j)}[k]\right)^{1-\alpha}. \tag{4.9}$$

The constraints is imposed by the bit-rate feasible set, based on AF relay architecture, described as

$$\mathcal{C}^{\mathrm{PHY-AF}} = \left\{ x_{\phi}^{(i,j)}[k] \middle| \forall\, i: \sum_{\phi=1}^{|\Phi|} \sum_{j=1}^{N} x_{\phi}^{(i,j)}[k] \leq \frac{T}{2},\ \forall\, j: \sum_{\phi=1}^{|\Phi|} \sum_{i=1}^{N} x_{\phi}^{(i,j)}[k] \leq \frac{T}{2},\ \forall\, i,j,\phi:\ x_{\phi}^{(i,j)}[k] \in \left\{0, 1, \ldots, \frac{T}{2}\right\} \right\}. \tag{4.10}$$

The constraints (4.10) ensure that the total number of allocated blocks does not exceed what is available in the frame and ensure that the scheduling variables are integers.

The discrete nature of constraint (4.10) makes the problem computationally hard. However, if the integrality of time allocations is relaxed, the optimization becomes a convex problem, due to the fact that the constraints are affine, and the objective is a summation of concave functions which can be solved with an off-the-shelf convex optimization package. A very useful feature of the relaxed optimization is that it is an upper bound on the integer solution of (4.9). We use this fact later to verify the performance of the proposed algorithm.

Since the utilities are strictly concave there is a unique optimizer for the problems. However, the allocation which lead to optimum rates may be not unique. Furthermore, since the objective is strictly increasing in terms of rates, the optimum point must be on the boundary of the feasible region (the constraints are active)

4.4 Proposed Algorithms

In this section the proposed algorithm for finite and infinite value of α will be discussed. Later, at the last part of this section, the efficient implementation of the algorithms will be designed and be presented.

Finding the time-allocations with convex programming suffers from several deficiencies. First, the optimal solution consists of the real-numbers, which should somehow be converted to integers. Second, the size of the optimization can quickly get out of control. The optimization has MN^2 variables and $2N$ constraints. For a 30 WTs network, each with one flow, with 50 sub-channels, there are 75000 variables in the optimization, which challenges even the best solvers. Therefore, we develop sub-optimal algorithms with relatively simple complexity for the network utility maximization. First we devise an algorithm for $\alpha < \infty$, then we devise an algorithm for $\alpha \to \infty$.

Our gradient-based algorithms are based on the fact that the maximum change in the objective function, that can be obtained from increasing one $x_{\phi}^{(i,j)}[k]$ by 1, is obtained by adding time to $x_{\phi}^{(i,j)}[k]$s in the direction of the steepest gradient of the objective function. This can be justified by

the Taylor's expansion of the network utility as

$$S^{\mathrm{r}}_{\mathrm{net}}\left(\cdots, x^{(i,j)}_{\phi}[k]+1, \cdots\right) \approx S^{\mathrm{r}}_{\mathrm{net}}\left(\cdots, x^{(i,j)}_{\phi}[k], \cdots\right) + \frac{\partial}{\partial x^{(i,j)}_{\phi}[k]} S^{\mathrm{r}}_{\mathrm{net}}\left(\cdots, x^{(i,j)}_{\phi}[k], \cdots\right). \quad (4.11)$$

Therefore, if we are given a choice of increasing any one $x^{(i,j)}_{\phi}[k]$, we should increase the one with the highest partial derivative, to maximize the incremental change in the objective function.

4.4.1 Proposed Algorithm for GPF for $\alpha < \infty$: AFGPF

Using the above observation about the gradient of the objective function, we now devise an iterative greedy algorithm to solve the optimization for $\alpha < \infty$ (**Algorithm** AFGPF). The algorithm starts by making a copy of $b^{(i,j)}_{\phi}[k]$s (Step 2), which are used later in the algorithm. It works in iterations, where in each iteration, the flow with the highest partial derivative is allocated an RB on its highest available sub-channel pairing based on

$$\left(i^*[k], j^*[k], \phi^*[k]\right) \leftarrow \underset{\phi,i,j}{\arg\max} \quad b^{(i,j)}_{\phi}[k] \Bigg/ \left(W_b \sum_{i=1}^{N}\sum_{j=1}^{N} b^{(i,j)}_{\phi}[k]\; x^{(i,j)}_{\phi}[k]\right)^{\alpha}. \quad (4.12)$$

The flow selection and the sub-channel pairing selection is performed in Steps 3. Variables $T^{(i)}_{\mathrm{BS}}$ and $T^{(j)}_{\mathrm{RS}}$ keep track of the available slots on each channel in the first and second parts of the frame, corresponding to the constraints in (4.10). After each iteration $T^{(i)}_{\mathrm{BS}}$ and $T^{(j)}_{\mathrm{RS}}$ are updated if any slots are allocated on their channels (Steps 5-6). The copy of bits per slot values, $\tilde{b}^{(i,j)}_{\phi}$s, are also updated (set to zero) according to the availability of RBs. This ensures that allocated slots are not considered in the next iteration (Steps 7-12). Note that since $b^{(i,j)}_{\phi}[k]$s are used to find the bit-rates at each iteration, they are fixed on a frame, throughout of the algorithm. On the other hand, $\tilde{b}^{(i,j)}_{\phi}$s change as the algorithm runs and are used to find the best pairing for a selected flow, in each iteration.

The complexity of the **Algorithm** AFGPF depends on the implementation of the search in Step 3. We do not get into the specifics of the algorithm's implementation, in this section. However, we note that the search in Step 3 can be implemented with multiple sorted lists holding SNRs,

namely, $|\Phi|$ lists for second-hop SNR measured at WTs, and one list for first-hop SNR measured at RS. It takes $N\log(N)$ steps to sort each list. With the sorted lists, in Step 3, the best available pairing, for each flow, is at the top of list and can be obtained in a single shot. Therefore, we can find the flow with the best partial derivative in $|\Phi|$ steps. Taking into account that there are $N\frac{T}{2}$ RBs in each hop, the algorithm goes through $N\frac{T}{2}$ iterations, for each user. The complexity of the algorithm is $\mathcal{O}\left((|\Phi|+1)N\log(N)+MN\frac{T}{2}\right)$. We further elaborate on the mechanism of sorted lists in Section 4.4.3.

It is worth highlighting that the scheduling and allocation core in (4.12), have been designed based on the static interference assumption. In fact, the intercell interference coordination (ICIC) (See for example [73]) works in a longer timescale than the scheduling and allocation algorithm to specify which RB is allowed to be used for each cell or sector. The proposed framework, in this chapter, can be extended to incorporate joint single-hop & double-hop scheduling & allocation and multiple relays with polynomial complexity.

We note the relationship between the algorithm and one of the procedures proposed for single channel, single-hop, networks [16]. In contrast, our algorithm is for double-hop AF networks. The connection is not unexpected given the fact that both our approach and [16] use the same utility functions to achieve fairness. The difference is that our utility function takes the instantaneous frame bit-rate, while in [16] the utility function takes in the mean bit-rates. Our optimization is performed in every frame for *short-term* bit-rate fairness, whereas the optimization in [16] is performed in every frame to obtain *long-term* bit-rate fairness.

4.4.2 Proposed Algorithm for GPF for $\alpha \to \infty$: AFMM

The **Algorithm** AFGPF is valid for $\alpha < \infty$. However, for $\alpha \to \infty$ the gradient becomes very small and the algorithm exhibits odd behaviours. To solve that problem, we investigate the problem for $\alpha \to \infty$ and devise another gradient-based algorithm to solve it in the asymptotic case. This special case of α corresponds to the maxmin bit-rate allocation. The algorithm is based on the results of the following proposition:

Proposition 5. *For α sufficiently large, assigning a time slot to the flow with the minimum current*

Algorithm AFGPF $\left(b_\phi^{(i,j)}[k], |\Phi|, N, T, \alpha\right)$.

Initialize: $\forall\, i,j : T_{\rm BS}^{(i)} \leftarrow T/2,\, T_{\rm RS}^{(j)} \leftarrow T/2.$

1: $\forall\, i,j,\phi: \quad \tilde{b}_\phi^{(i,j)} \leftarrow b_\phi^{(i,j)}[k].$

2: **while** $\exists\, T_{\rm BS}^{(i)} > 0$ **and** $\exists\, T_{\rm RS}^{(j)} > 0$ **do**

3: $\quad (i^*[k], j^*[k], \phi^*[k]) \leftarrow \underset{\substack{1\le\phi\le|\Phi|,\\ 1\le i,j\le N}}{\arg\max}\, \tilde{b}_\phi^{(i,j)} \Big/ \left(W_b \sum_{i=1}^N \sum_{j=1}^N b_\phi^{(i,j)}[k] x_\phi^{(i,j)}\right)^\alpha.$

4: $\quad x_{(\phi^*[k])}^{(i^*[k],j^*[k])} \leftarrow x_{(\phi^*[k])}^{(i^*[k],j^*[k])} + 1.$

5: $\quad T_{\rm BS}^{(i^*[k])} \leftarrow T_{\rm BS}^{(i^*[k])} - 1.$

6: $\quad T_{\rm RS}^{(j^*[k])} \leftarrow T_{\rm RS}^{(j^*[k])} - 1.$

7: $\quad$ **if** $T_{\rm BS}^{(i^*[k])} = 0$ **then**

8: $\quad\quad \tilde{b}_\phi^{(i^*[k],j)} \leftarrow 0, \quad 1 \le \phi \le |\Phi|, \quad 1 \le j \le N.$

9: $\quad$ **end if**

10: $\quad$ **if** $T_{\rm RS}^{(j^*[k])} = 0$ **then**

11: $\quad\quad \tilde{b}_\phi^{(i,j^*[k])} \leftarrow 0, \quad 1 \le \phi \le |\Phi|, \quad 1 \le i \le N.$

12: $\quad$ **end if**

13: **end while**

bit-rate on its best sub-channel pairing is equivalent to assigning resources to the flow with the largest gradient as given in (4.12).

Proof. Define the best sub-channel pairing for flow ϕ and the bit-rate of flow with the lowest bit-rate among all flows, denoted by $\underline{\phi}$, with

$$\hat{b}_\phi^{(i,j)}[k] \triangleq \max_{1\le i,j\le N} \left\{ b_\phi^{(i,j)}[k] \right\}, \quad r_{\underline{\phi}}[k] \triangleq \min_{1\le\phi\le|\Phi|} \left\{ W_b \sum_{i=1}^N \sum_{j=1}^N b_\phi^{(i,j)}[k] x_\phi^{(i,j)}[k] \right\}, \tag{4.13}$$

and a threshold on α as

$$\alpha_0 \triangleq \max_{1\le\phi\le|\Phi|} \left\{ \log\left(\frac{\hat{b}_\phi^{(i,j)}[k]}{\hat{b}_{\underline{\phi}}^{(i,j)}} \right) \Big/ \log\left(\frac{r_\phi[k]}{r_{\underline{\phi}}[k]} \right) \right\}. \tag{4.14}$$

Since $\forall\ \phi \neq \underline{\phi}$, we have $0 < \log\left(\frac{r_\phi[k]}{r_{\underline{\phi}}[k]}\right)$, the following is true as

$$\alpha \log\left(\frac{r_\phi[k]}{r_{\underline{\phi}}[k]}\right) \geq \log\left(\frac{\hat{b}_\phi^{(i,j)}[k]}{\hat{b}_{\underline{\phi}}^{(i,j)}}\right), \quad \Rightarrow \quad \frac{(r_\phi[k])^\alpha}{(r_{\underline{\phi}}[k])^\alpha} \geq \frac{\hat{b}_\phi^{(i,j)}[k]}{\hat{b}_{\underline{\phi}}^{(i,j)}} \quad \Rightarrow \quad \frac{\hat{b}_\phi^{(i,j)}[k]}{(r_\phi[k])^\alpha} \leq \frac{\hat{b}_{\underline{\phi}}^{(i,j)}}{(r_{\underline{\phi}}[k])^\alpha}, \quad (4.15)$$

for $\forall\ \alpha \geq \alpha_0, \forall\ \phi \neq \underline{\phi}$.

Finally since by definition of $\hat{b}_\phi^{(i,j)}[k]$,

$$\frac{b_\phi^{(i,j)}[k]}{(r_\phi[k])^\alpha} \leq \frac{\hat{b}_\phi^{(i,j)}[k]}{(r_\phi[k])^\alpha} \leq \frac{\hat{b}_{(\underline{\phi})}^{(i,j)}}{(r_{\underline{\phi}}[k])^\alpha}, \quad \text{we have} \quad \max_{1\leq\phi\leq|\Phi|} \max_{1\leq i,j\leq N} \frac{b_\phi^{(i,j)}[k]}{(r_\phi[k])^\alpha} \leq \frac{\hat{b}_{(\underline{\phi})}^{(i,j)}}{(r_{\underline{\phi}}[k])^\alpha}, \quad (4.16)$$

for $\forall\ \alpha \geq \alpha_0$, proving the proposition. □

Using the proposition, we see that as $\alpha \to \infty$, finding the largest derivative is equivalent to assigning time to the flow with the minimum current bit-rate to its best sub-channel pairing. Based on this fact, we devise a gradient-based algorithm to find the maxmin fair allocation of bit-rates by substituting the step 3 in the **Algorithm** AFGPF with

$$\phi^*[k] \leftarrow \arg\min_{1\leq\phi\leq|\Phi|} \left\{ \sum_{i=1}^{N} \sum_{j=1}^{N} b_\phi^{(i,j)}[k] x_\phi^{(i,j)}[k] \right\}, \quad \text{and then} \quad \left(i^*[k], j^*[k]\right) \leftarrow \arg\max_{1\leq i,j\leq N} \tilde{b}_{\phi^*}^{(i,j)}. \quad (4.17)$$

The substituted Steps 3 perform the search according to the proposition: first the minimum bit-rate flow is found, then its best sub-channel pairing is found. The rest of the algorithm corresponds to **Algorithm** AFGPF. In the sequel, we refer to this version of algorithm as **Algorithm** AFMM. It is worth noting that approaching the problem with the linear maxmin objective does not produce an obviously good heuristic.

A similar algorithm is also used in the context conventional cellular networks without relays [34], where in each iteration the WT with the minimum bit-rate is allocated resources on its best sub-channel. However, unlike [34] which is for conventional cellular networks, our algorithm is for OFDMA-based AF relay networks. Since we derive our algorithm from the convex utility-based maxmin fair resource allocation problem, we also have an explanation of why the allocations derived

by this kind of algorithm are so close to optimum.

In our simulations, we compare the performance of the **Algorithm** AFGPF and **Algorithm** AFMM with the exact solution of the relaxed optimization, which finds the upper bound. We used the `CVX` solver [139]. Even though the relaxed version of the optimization can find the upper bound for the maxmin bit-rate allocation when $\alpha \to \infty$, the convex solver exhibits poor convergence for the large values of α, so we use an alternative approach to find the upper bound for maxmin bit-rate allocation with the maxmin objective function as

$$\max_{x_{\phi}^{(i,j)}[k]} \quad \min_{\phi} \quad W_b \sum_{i=1}^{N} \sum_{j=1}^{N} b_{\phi}^{(i,j)}[k]\; x_{\phi}^{(i,j)}[k], \tag{4.18}$$

which makes (4.9) as a linear optimization.

4.4.3 Efficient Implementation based on the Super-modularity of the AMC Table: AFGPF-EFF

We did not go into implementation of the step 3 (the argmax in step 3), not to make the algorithm confusing. In this part, we will exploit the special characteristic of the mapping between the hop's SNRs, combined SNR, and AMC values to develop an efficient implementation of step 3 of the **Algorithm** AFGPF. Note that the reduction in the complexity of such a realtime decision maker is of a great importance.

The idea is based on an observation, from sub-channel pairing, for single-user case [104,106,110, 111]. To maximize the throughput for a single-user scenario, in AF relay, it is sufficient to couple the first-hop sub-carriers with second-hop sub-carriers in a same order of their SNRs. This is called OSP and is based on the super-modular property [140] of the functions of mapping the combined SNR to AMC values. Suppose that for a certain WT, or its flow, a pair of first-hop sub-carriers are coupled with a pair second-hop sub-carriers unordered, the super-modularity is ensuring that the pairing can be reversed which increases the frame bit-rate of that flow, with no effect on the other flow frame bit-rate. Therefore, the sub-carriers of different WT must be coupled based on the ordered list.

Super-modular Function

First, we further elaborate on the super-modular functions. We use an special indexing for the arguments of a super-modular function, in order to make it easier to connect it to our context. A scalar function on two variables $h\left(A,B\right):\mathbb{R}^2\rightarrow\mathbb{R}$ is called super-modular [140] if and only if

$$h\left(\begin{bmatrix}F^{(1)}\\S^{(1)}\end{bmatrix}\right)+h\left(\begin{bmatrix}F^{(2)}\\S^{(2)}\end{bmatrix}\right)\leq h\left(\begin{bmatrix}\max(F^{(1)},F^{(2)})\\\max(S^{(1)},S^{(2)})\end{bmatrix}\right)+h\left(\begin{bmatrix}\min(F^{(1)},F^{(2)})\\\min(S^{(1)},S^{(2)})\end{bmatrix}\right). \tag{4.19}$$

A super modular function can be identified if its cross derivative is positive [140] as

$$0\leq\partial^2 h\left(\begin{bmatrix}F^{(1)}\\S^{(1)}\end{bmatrix}\right)\Big/\partial F^{(1)}\partial S^{(1)}. \tag{4.20}$$

As an example, the capacity of the AF relay is a super-modular function. In other words, if $F^{(2)}\leq F^{(1)}$ and $S^{(1)}\leq S^{(2)}$ correspond to the hop's SNRs, the following is true as

$$\log\left(1+\frac{F^{(1)}S^{(1)}}{F^{(1)}+S^{(1)}+1}\right)+\log\left(1+\frac{F^{(2)}S^{(2)}}{F^{(2)}+S^{(2)}+1}\right)\leq\log\left(1+\frac{F^{(1)}S^{(2)}}{F^{(1)}+S^{(2)}+1}\right)+\log\left(1+\frac{F^{(2)}S^{(1)}}{F^{(2)}+S^{(1)}+1}\right). \tag{4.21}$$

Super-modularity of the AMC table function

Now we show that the AMC values of an AF relay is a super-modular function. Based on the chain rule, the cross derivative of the AMC values, as a function of hop's SNR,(4.3), is

$$\frac{\partial^2\mathfrak{f}\left(B_{\phi}^{(i,j)}[k]\left(\mathsf{SNR}_{\mathrm{R}}^{(i)},\mathsf{SNR}_{\phi}^{(j)}[k]\right)\right)}{\partial\mathsf{SNR}_{\mathrm{R}}^{(i)}\partial\mathsf{SNR}_{\phi}^{(j)}[k]}=\frac{\partial^2 B_{\phi}^{(i,j)}[k]}{\partial\mathsf{SNR}_{\mathrm{R}}^{(i)}\partial\mathsf{SNR}_{\phi}^{(j)}[k]}\mathfrak{f}^{'}\left(B_{\phi}^{(i,j)}[k]\right)+\frac{\partial B_{\phi}^{(i,j)}[k]}{\partial\mathsf{SNR}_{\mathrm{R}}^{(i)}}\frac{\partial B_{\phi}^{(i,j)}[k]}{\partial\mathsf{SNR}_{\phi}^{(j)}[k]}\mathfrak{f}^{''}\left(B_{\phi}^{(i,j)}[k]\right). \tag{4.22}$$

We note that the above two terms are both non negative. The cross derivative of $B_{\phi}^{(i,j)}[k]$ is non negative as

$$0\ \leq\ \frac{\partial^2 B_{\phi}^{(i,j)}[k]}{\partial\mathsf{SNR}_{\mathrm{R}}^{(i)}\partial\mathsf{SNR}_{\phi}^{(j)}[k]}=1\Big/\left(1+\mathsf{SNR}_{\mathrm{R}}^{(i)}+\mathsf{SNR}_{\phi}^{(j)}[k]\right)^2. \tag{4.23}$$

The nondecreasing property of the Shannon capacity, in terms of both hop's SNRs ($B_{\phi}^{(i,j)}[k]$), ensures the non negativity of the second term. Therefore, a sufficient condition for super-modularity of the AMC mapping function, in terms of hop's SNR, is piecewise linearity non decreasing condition of the AMC table. Interestingly, although it is a sufficient condition, all AMC mapping table satisfy this.

Now, we can exploit super-modularity property and reduce the complexity of the algorithm. As explained earlier, instead of the finding the largest $b_\phi^{(i,j)}[k]/(r_\phi[k])^\alpha$ among all i, j, ϕ, we will use a set of sorted SNR lists to simplify the search in step 3. Assume that an RB is to be given to a flow in an iteration, based on super-modularity, we must give the best RB remaining in the first-hop coupled with the best second-hop remaining RB. Toward this purpose, we sort the first-hop SNR vector and second-hop SNR vectors of different flows, before algorithm starts. Then at each iteration, we only compute a one dimensional vector of a fraction of best $b_\phi^{(i,j)}[k]$ over denominator for different flows, in order to select the best flow. The selected flow gets the best available sub-channel pairing, based on sorted SNR lists.

Explanation of the Efficient Implementation

The **Algorithm** AFGPF-EFF starts in Step 1 and Step 2 with sorting first-hop SNR vector and second-hop SNR vectors of different flows. We denote the sorted list as $\widehat{\mathsf{SNR}}_{\mathrm{R}}^{(i)}$ and $\widehat{\mathsf{SNR}}_{\phi}^{(j)}[k]$ for the first-hop and second-hop, respectively. The corresponding permutations will be denoted as $\mathfrak{P}_{\mathrm{R}}(l)$ and $\mathfrak{P}_\phi(l)$. In other words,

$$\forall\, l:\ \widehat{\mathsf{SNR}}_{\mathrm{R}}^{(l)} = \mathsf{SNR}_{\mathrm{R}}^{(\mathfrak{P}_{\mathrm{R}}(l))} \quad \text{and} \quad \forall\, l, \phi:\ \widehat{\mathsf{SNR}}_{\phi}^{(l)}[k] = \mathsf{SNR}_{\phi}^{(\mathfrak{P}_\phi(l))}[k]. \tag{4.24}$$

For the second-hop SNRs, we also need the sorted version of the permutation (Step 4) in order to find the position of each sub-carrier in the sorted list. This can be found without computation. Suppose that we sort the second-hop permutation, which is $\mathfrak{P}_\phi^{(l)}$, in ascending order, therefore

$$\left(\widehat{\mathfrak{P}}_\phi(l), \mathfrak{P}_\phi^{-1}(l)\right) \leftarrow \mathrm{sort}^{\downarrow}\left(\mathfrak{P}_\phi^{(l)}\right), \qquad \text{or equivalently} \qquad \widehat{\mathfrak{P}}_\phi(l) = \mathfrak{P}_\phi\left(\mathfrak{P}_\phi^{-1}(l)\right). \tag{4.25}$$

Now, note that

$$\forall\, \phi:\ \widehat{\mathfrak{P}}_\phi(l) = l, \tag{4.26}$$

which makes the inverse permutation, trivial. We can now locate a certain sub-carrier of the original list, in the sorted list, as

$$\widehat{\mathsf{SNR}}_{\phi}^{\left(\mathfrak{P}_\phi^{-1}(l)\right)}[k] = \mathsf{SNR}_{\phi}^{(l)}[k]. \tag{4.27}$$

Having explained the sorted lists, Step 5 searches in a one dimensional vector, instead of previously searching through a matrix. Then, the algorithm gives the current top SNRs in the sorted list of first-hop and second-hop to the selected flow, in step 6 and step 7. Steps 8, 9 will update the availability of RB. If one of the first-hop sub-carriers is exhausted, it will be deleted from the first-hop sorted SNR list in step 11. A certain sub-carrier in the second-hop may also become exhausted because of its assignment in previous iterations. In this case, the corresponding second-hop SNR will set to zero (or deleted), from the sorted SNR lists of all of the flows, in step 14. Note that these deletion will not violate the order of the list. The permutation vectors will be updated accordingly, in Step 11 and Step 14.

Algorithm AFGPF-EFF $\left(b_{\phi}^{(i,j)}[k], \mathrm{SNR}_{\mathrm{R}}^{(i)}, \mathrm{SNR}_{\phi}^{(j)}, |\Phi|, N, T, \alpha\right)$.

Initialize: $1 \le i, j \le N : T_{\mathrm{BS}}^{(i)} = T/2,\ T_{\mathrm{RS}}^{(j)} = T/2.$

1: $(\widehat{\mathsf{SNR}}_{\mathrm{R}}^{(l)}, \mathfrak{P}_{\mathrm{R}}(l)) \leftarrow \mathrm{sort}^{\uparrow}\left(\mathsf{SNR}_{\mathrm{R}}^{(i)}\right)$ equivalently $\widehat{\mathsf{SNR}}_{\mathrm{R}}^{(l)} = \mathsf{SNR}_{\mathrm{R}}^{(\mathfrak{P}_{\mathrm{R}}(l))}$.

2: $(\widehat{\mathsf{SNR}}_{\phi}^{(l)}, \mathfrak{P}_{\phi}(l)) \leftarrow \mathrm{sort}^{\uparrow}\left(\mathsf{SNR}_{\phi}^{(j)}\right)$ equivalently $\widehat{\mathsf{SNR}}_{\phi}^{(l)} = \mathsf{SNR}_{\phi}^{(\mathfrak{P}_{\phi}(l))}$.

3: $(\widehat{\mathfrak{P}}_{\phi}(l), \mathfrak{P}_{\phi}^{-1}(l)) \leftarrow \mathrm{sort}^{\downarrow}\left(\mathfrak{P}_{\phi}(l)\right)$ equivalently $\widehat{\mathfrak{P}}_{\phi}(l) = \mathfrak{P}_{\phi}(\mathfrak{P}_{\phi}^{-1}(l)) = l$.

4: **while** $\exists\ T_{\mathrm{BS}}^{(i)} > 0$ **and** $\exists\ T_{\mathrm{RS}}^{(j)} > 0$ **do**

5: $\phi^* \leftarrow \arg\max_{\phi} W_b\, \mathfrak{f}\left(\log\left(1 + \frac{\widehat{\mathsf{SNR}}_{\mathrm{R}}^{(1)}[k]\widehat{\mathsf{SNR}}_{\phi}^{(1)}[k]}{\widehat{\mathsf{SNR}}_{\mathrm{R}}^{(1)}[k] + \widehat{\mathsf{SNR}}_{\phi}^{(1)}[k] + 1}\right)\right) \Big/ (r_{\phi}[k])^{\alpha}$.

6: $i^* \leftarrow \mathfrak{P}_{\mathrm{R}}(1), \quad j^* \leftarrow \mathfrak{P}_{\phi^*}(1)$.

7: $x_{\phi^*}^{(i^*,j^*)} \leftarrow x_{\phi^*}^{(i^*,j^*)} + 1$.

8: $T_{\mathrm{BS}}^{(i^*)} \leftarrow T_{\mathrm{BS}}^{(i^*)} - 1$.

9: $T_{\mathrm{RS}}^{(j^*)} \leftarrow T_{\mathrm{RS}}^{(j^*)} - 1$.

10: **if** $T_{\mathrm{BS}}^{(i^*)} = 0$ **then**

11: update $\widehat{\mathsf{SNR}}_{\mathrm{R}}^{(l)}$ by deleting its first element.
update $\mathfrak{P}_{\phi}(l)$ by shifting its vector accordingly
and make its last non-zero element to zero.

12: **end if**

13: **if** $T_{\mathrm{RS}}^{(j^*)} = 0$ **then**

14: update $\widehat{\mathsf{SNR}}_{\phi}^{(l)}$ by deleting its $\mathfrak{P}_{\phi}(1)$-th element.
update $\mathfrak{P}_{\phi}(l)$ & $\mathfrak{P}_{\phi}^{-1}(l)$ by shifting its vector accordingly
and make its last non-zero element to zero.

15: **end if**

16: $r_{\phi^*}[k] \leftarrow W_b \sum_{i=1}^{N} \sum_{j=1}^{N} b_{\phi^*}^{(i,j)}[k] x_{\phi^*}^{(i,j)}$

17: **end while**

4.5 Simulation

We ran two sets of Monte-Carlo simulations to evaluate the proposed algorithm. In both simulations, we consider a network of $|\Phi| = 30$ flows connected to the BS through a predetermined RS. In each iteration of the simulation, we randomly drop the WTs with a uniform density in the area around the relay. Each WT is assumed to have a single flow. From WT locations, we calculate each users' channel coefficients to the RS and use a detailed channel model to find the number of bits carried in an RB on each sub-channel. Details of the simulation parameters are shown in Table 4.2.

Table 4.2: Simulation parameters for the fair scheduling in AF relay networks.

Parameter	**Value**
BS to RS channel	Rician, K=10 dB [99]
BS to RS shadowing	Log-normal, variance 3 dB
BS to RS doppler shift	4 Hz
RS to WTs channel	Rayleigh [99]
RS to WTs shadowing	Log-normal, variance 5 dB
RS to WTs doppler shift	37 Hz
Path loss	$38.4 + 2.35 \log_{10}(d)$ dB
Sub-carrier bandwidth	10.9375 kHz
Sub-carriers per sub-channel	18
Number of WTs	$\|\Phi\| = 30$
Number of sub-channels	$N = 50$
Slots per frame	$T = 20$
Cell radius	1000 m
BS to RS distance	500 m
Transmit power	40 dBm BS, 30 dBm RS
Antenna gain	10 dB BS, 5 dB RS, 0 dB WTs
Noise figure	2 dB RS, 2 dB WTs

Sub-optimality and Optimizer Similarity Evaluation

In the first set of simulations, we measure the sub-optimality of the algorithm. There are a total of 40 drops for a total of 800 distinct inputs to the optimization. Even this modest number of drops took more than 20 hours to run due to the time it takes to find the upper bound values. Due to the slow convergence of the convex solver for large values of α, we use linear programming to find the upper bound for the maxmin bit-rate allocation.

For each drop, we calculate time allocations using the relaxed optimization, denoted by $\hat{x}_{\phi}^{(i,j)}[k]$, and time allocations with the proposed algorithm, denoted by $\tilde{x}_{\phi}^{(i,j)}[k]$. Then, the sub-optimality gap is bounded by

$$\Delta_{\mathrm{H}}[k] \triangleq \left| S_{\mathrm{net}}^{\mathrm{r}}\left(\cdots, \tilde{x}_{\phi}^{(i,j)}[k], \cdots\right) - S_{\mathrm{net}}^{\mathrm{r}}\left(\cdots, \hat{x}_{(\phi)}^{(i,j)}, \cdots\right) \right| \Bigg/ S_{\mathrm{net}}^{\mathrm{r}}\left(\cdots, \hat{x}_{\phi}^{(i,j)}[k], \cdots\right), \tag{4.28}$$

where $|\cdot|$ is the absolute value of its operand. As the granularity of the OFMDA, in comparison to the number of flows, increases ($NT/|\Phi|$ increases) the upper bound approaches the actual optimization. Nevertheless, since the relaxed optimization only *strictly* upper bounds the value of the integer optimization, this is larger than actual gap between the optimal integer solution and the proposed algorithm. In other words, the upper bound is not tight and the aforementioned gap is a *pessimistic* performance evaluation for the algorithms.

For a more detailed comparison, we also find the difference between the individual bit-rates of the relaxed optimization and the proposed algorithms as

$$\delta_{\phi}[k] = \left| W_b \sum_{i=1}^{N} \sum_{j=1}^{N} b_{\phi}^{(i,j)}[k] \left(\hat{x}_{\phi}^{(i,j)}[k] - \tilde{x}_{\phi}^{(i,j)}[k] \right) \right|. \tag{4.29}$$

Based on $\delta_{\phi}[k]$, we can find the similarity between the way the proposed algorithm and the optimal solution allocate bit-rates as

$$\delta_{\mathrm{H}}[k] = \left(\sum_{\phi=1}^{|\Phi|} \delta_{\phi}[k] \right)^2 \Bigg/ |\Phi| \sum_{\phi=1}^{|\Phi|} \left(\delta_{\phi}[k]\right)^2. \tag{4.30}$$

Table 4.3 shows the mean and standard deviation (st. d.) of the sub-optimality gap, $\Delta_{\mathrm{H}}[k]$, and the similarity of the allocated bit-rates, $\delta_{\mathrm{H}}[k]$. We see that the sub-optimality gap is relatively small. In fact, for $\alpha = 0$, there is no gap within this precision because the proposed algorithm finds the optimal bit-rates maximizing the system throughput in this case. The similarity between the algorithm and the optimal solution is relatively small in this case, indicating the presence of a large amount of multiuser diversity in the network which causes the algorithm and the upper bound to have the same objective value with different allocations. For $\alpha = 0.2$, $\alpha = 0.45$, $\alpha = 1.0$,

Table 4.3: Optimality performance for fair scheduling in AF relay networks.

	$\alpha = 0.00$	$\alpha = 0.20$	$\alpha = 0.45$	$\alpha = 1.00$	MM
Δ_{H} (mean)	0.00 %	3.43 %	2.92 %	0.45 %	7.90 %
Δ_{H} (st. d.)	0.00 %	0.68 %	0.43 %	0.06 %	1.59 %
δ_{H} (mean)	11.10 %	34.70 %	49.08 %	58.63 %	93.50 %
δ_{H} (st. d.)	5.95 %	9.01 %	9.60 %	7.53 %	3.08 %

and $\alpha \to \infty$, the sub-optimality gap (Δ_{H}) is still smaller than 8 % on average. In these cases, the similarity (δ_{ϕ}) between the bit-rates increases since the bit-rates are allocated fairly despite the available multiuser diversity.

Fairness and Throughput Evaluation

In the second set of simulations, we measure the fairness performance as well as throughput performance of the algorithm with a total of 8000 distinct drops.

Figure 4.2 shows the total system bit-rate allocated to WTs as a function of distance from the relay. We observe that as α decreases, more system resources are assigned to WTs close to the RS, increasing the system throughput. We observe that as we move from the maxmin allocation, to the PF allocation, to the maximum throughput allocation, more resources are assigned to WTs closer to the RS, increasing the throughput. For a large enough α, such as $\alpha = 2$, the distribution of the resources is similar to the maxmin fair distribution of resources. The distribution of resources results in the most ubiquitous coverage with the **Algorithm** AFMM. It is worth mentioning that changing α achieves an effect similar to cell-breathing.

Figure 4.3 shows the cumulative distribution function (CDF) of the bit-rates for different value of α. We see that as α increases, the CDF approaches to a step function. The step function CDF corresponds to the most uniform distribution of resources in the network. In other words, as parameter α increases, the bit-rates become more uniform in the network with the maximum uniformity achieved for **Algorithm** AFMM. Note that 67 % of the WTs get no allocation of resources for maximum throughput allocation, in which $\alpha = 0$. Figure 4.3 also confirms that the **Algorithm** AFGPF behaviour converges to the**Algorithm** AFMM for sufficiently large α.

Figure 4.4 and Figure 4.5 show the performance of the lowest 5^{th} percentile of the WTs bit-rate

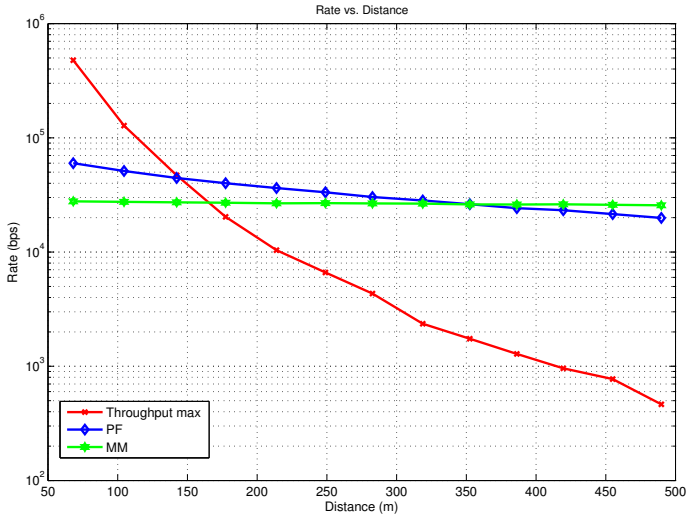

Figure 4.2: WT bit-rates vs. distance for fair scheduling in AF relay networks.

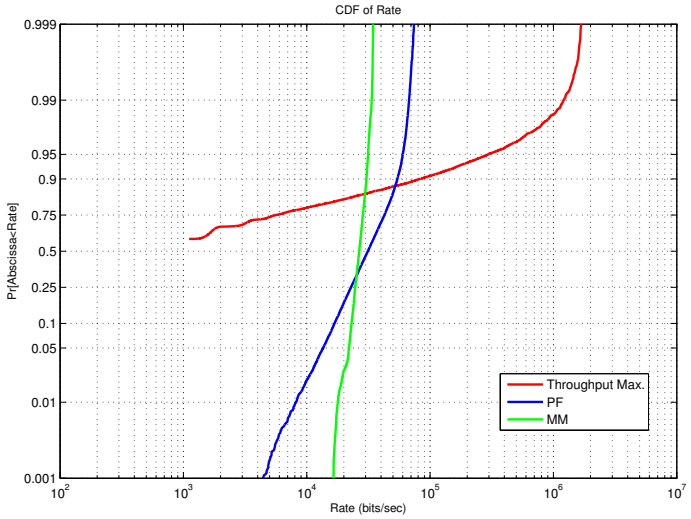

Figure 4.3: Cumulative distribution function of bit-rate for fair scheduling in AF relay networks.

(the cell-edge) and the highest 5^{th} percentile of the WTs bit-rate, for different α values. The lowest 5^{th} percentile of the bit-rates resembles the cell edge performance and the highest 5^{th} percentile of the bit-rates resembles the centre cell WTs performance. We observe that increasing α increases the cell-edge bit-rate, at the expense of WTs with good channels, where the system is fairer and transferring resources to the weaker WTs. As expected, the bit-rates of the lowest 5^{th} percentile and the highest 5^{th} percentile converge to the maxmin fair bit-rates as α increases (see Figure 4.5).

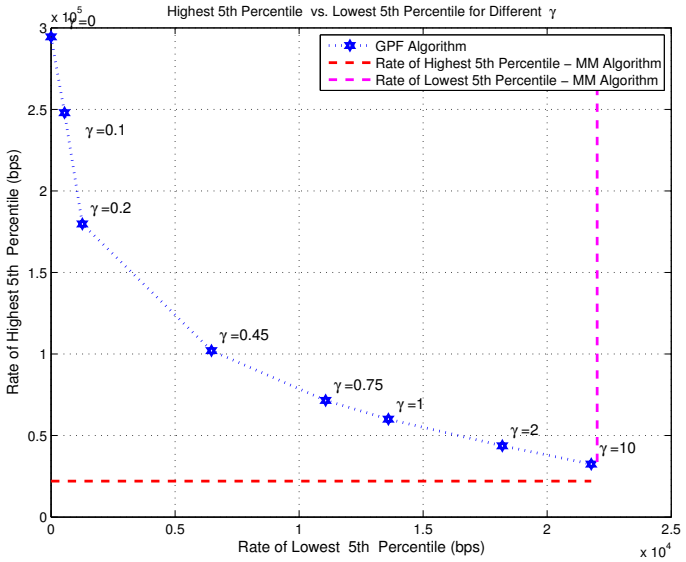

Figure 4.4: The 95^{th} percentile bit-rate vs. the 5^{th} percentile bit-rate ($\gamma = \alpha$).

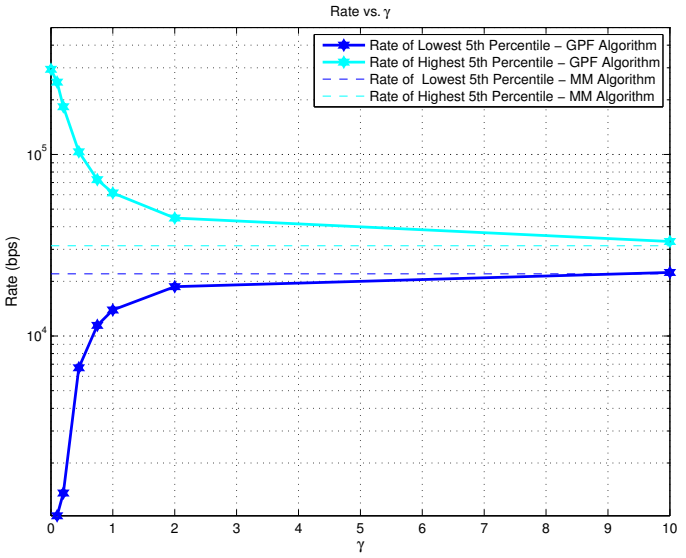

Figure 4.5: The lowest 5^{th} percentile and the highest 5^{th} percentile bit-rates vs. $\gamma = \alpha$ for fair scheduling in AF relay networks.

To further compare the impact of allocations on the fairness, we use the Jain's fairness index as

$$J\Big(r_1[k],\ldots,r_\Phi[k]\Big) = \left(\sum_{\phi=1}^{|\Phi|} r_\phi[k]\right)^2 \Big/ |\Phi| \sum_{\phi=1}^{|\Phi|} (r_\phi[k])^2, \tag{4.31}$$

which measures how similar bit-rates are. For Jain's index close to one the bit-rates are the most similar, so the system is in extreme fair case, and for the Jain's index close to $\frac{1}{|\Phi|}$, the rates are the

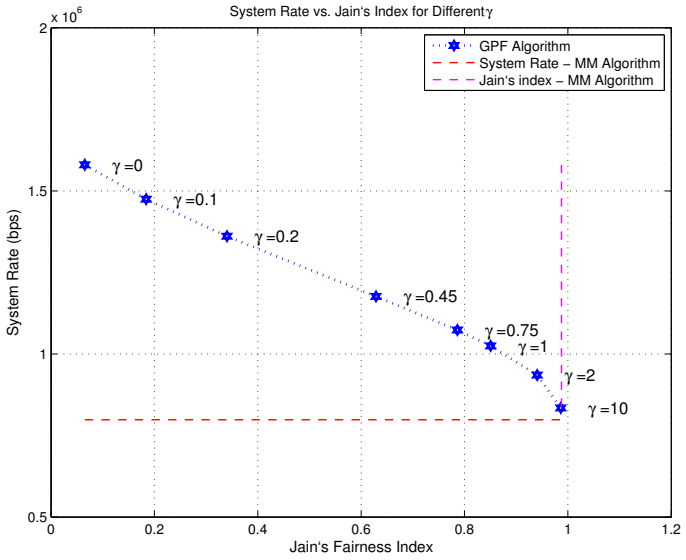

Figure 4.6: System bit-rate vs. the Jain's index for fair scheduling in AF relay networks ($\gamma = \alpha$).

least similar so the system is in extreme unfair case. Figure 4.6 shows that as α increases, the Jain index is improved. As expected, for the **Algorithm** AFMM, where the system is more fair and transfers resources to the weaker WTs, the system throughput has the lowest value while the Jain's index is the highest. The trade-off between system bit-rate (as a measure of system satisfaction) and fairness (as a measure of WTs satisfaction) can be seen clearly.

Alternatively, Table 4.4 summarizes the bit-rate of the lowest 5^{th} percentile of the bit-rate, the highest 5^{th} percentile of the bit-rate, the Jain's index, and the total throughput of the system for α equal to 0, 1, and ∞.

Table 4.4: The lowest 5^{th} percentile of the bit-rate, the highest 5^{th} percentile of the bit-rate, the Jain's index, and the total throughput for fair scheduling in AF relay networks.

	Throughput max.	PF	MM
5^{th} percentile (Kbps)	0.000	13.600	22.010
95^{th} percentile (Mbps)	0.2945	0.0600	0.0220
Throughput (Mbps)	1.580	1.003	0.800
Jain's index	0.06531	0.84920	0.98770

4.6 Conclusion

We investigate OFDMA-based AF relays and devise near-optimal algorithms for packet scheduling and RB allocation for these relays. We devise a GPF scheduling framework, considering the possibility of the sub-channel pairing. We devise two sub-optimal gradient-based algorithms to find the bit-rates close to the optimum solution of the GPF scheduler and the maxmin scheduler. Our simulations show that the network operator is able to adjust the parameter of the fairness to move between sum bit-rate maximization, PF, and maxmin fairness. For the asymptotic case of GPF, we show that the gradient of the objective function can be simplified in order to produce maxmin fair schedules. Simulations show that the maxmin allocation is more fair than the allocations by PF and achieves the most ubiquitous coverage. Both algorithms achieve results very close to the optimum solutions, due to their gradient origin. Our simulation also show that this RRM technique achieves similar results to cell-breathing, without the need to dynamically adjust the transmit power.

Chapter 5

Future Directions

The future directions on packet scheduling and resource allocation are suggested in four main dimensions:

1. The objective: The objective of packet scheduling and resource allocation can be as simple as sum bit-rate maximization, and be more advanced, such as maxmin bit-rate or minmax delay. Extension are suggested on advancement of incorporating long-term operator's interests based on user-satisfaction to minimize the clients incentive in leaving the operator. Extension on the revenue-awareness to incorporates general pricing mechanism, based on the dissatisfaction, as well as determining and adjusting the optimum charging policies are also recommended. The framework can be used in self-optimizing networks, where the contracts and billing are updated based on solid revenue maximization, along the other self-updating mechanisms.

2. Other radio resources: The joint optimization of packet scheduling, RB scheduling, route scheduling, and power allocation can be suggested in this dimension. The admission controller (AC) design can also be jointly considered in formulation, by modelling the cost of not admitting a flow with an extra disutility function.

3. RAN Architecture: In the RAN architecture dimension, developing QoS guarantee algorithms in multi-cell architectures, ICIC, and CoMP are suggested as a future work. Mechanisms such as ICIC, cell switch-offs, CoMP, relay networks, and MIMO conventionally are designed based

on maximizing sum bit-rate, or utmost a primitive fairness mechanism. In other words, rarely QoS or advanced fairness notions is incorporated in the design of such systems. Especially, in cell switch-off, QoS-requirements and -measurements can be adopted to improve the decisions on whether or not the network is under-utilized enough in order to switch-off cells. Likewise, many of scheduling and allocation algorithms in the relay-based system lacks sophistication of QoS requirements which can be suggested as a future work.

4. Distributed vesions: To coordinate individual WT and network optimality at a same time with distributed approach, the problem can be solved partly in network controller and partly in WTs. In this setup, WTs solve their own optimization based on a set of the Lagrange multipliers which are issued by the network controller and the network controller updates the Lagrange multipliers based on WTs feedbacks and current Lagrange multipliers. Based on this approach, the developing the distributed version of the algorithm is suggested. Two important topics to be investigated are how to deal with possibly unfeasible bit-rate requests, and the second is the convergence issues.

5. Improving the QoS machinery: In this dimension, further study on the minimum bit-rate and maximum delay guarantee by incorporating concepts from proportional integral derivative (PID) controller from control theory is suggested. This idea is anticipated to make improvement by reducing the overshoot effects when controlling the measured quantities with respect to their set points. In fact, in the two common approach of virtual token and Lagrangian approach, updating of the virtual token counter or the Lagrangian multiplier can be done not only based on previous frame mean values but also based on the last two frame difference values.

List of References

[1] 3GPP, "Technical specification group services and system aspects; policy and charging control architecture Release 11," 3rd Generation Partnership Project (3GPP), TS 23.203, December 2011. [Online]. Available: {http://www.3gpp.org/DynaReport/23203.htm}

[2] H. Yanikomeroglu and J. Zhang, "Beyond-4G cellular networks: Advanced radio access network (RAN) architectures, advanced radio resource management (RRM) techniques, and other enabling technologies," in *Proc. of WWRF Meeting*, October 2008, pp. 13–15.

[3] Canadian Radio and Telecommunication Commision, "Internet traffic management practices guidelines for responding to complaints and enforcing framework compliance by internet service providers," http://www.crtc.gc.ca/eng/archive/2011/2011-609.htm, September 2011.

[4] Chief Judge Tatel, U.S. Court of Appeal, "Comcast vs. FCC," https://www.eff.org/files/Comcast, April 2010.

[5] H. Holma and A. Toskala, *WCDMA for UMTS*. West Sussex, England: John Wiley & Sons, 2002.

[6] G. Fettweis and S. Alamouti, "5G: Personal mobile internet beyond what cellular did to telephony," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 140–145, February 2014.

[7] ROHDE & SCHWARZ, "Voice and SMS in LTE technology," *ROHDE & SCHWARZ White Paper. Available at rohde-schwarz.com*, 2011.

[8] A. Bergkvist, D. Burnett, C. Jennings, and A. Narayanan, "WebRTC 1.0: Real-time communication between browsers," http://dev.w3.org/2011/webrtc/editor/webrtc.html, September 2012, World Wide Web Consortium.

[9] T. Lan, D. Kao, M. Chiang, and A. Sabharwal, "An axiomatic theory of fairness in network resource allocation," in *Proc. of IEEE Annual Joint Conference of the Computer and Communications Societies (INFOCOM)*, March 2010, pp. 1–9.

[10] A. Odlyzko, "Network neutrality, search neutrality, and the never-ending conflict between efficiency and fairness in markets," *Review of Network Economics*, vol. 8, no. 1, pp. 1–21, March 2009.

[11] A. Sharifian and H. Yanikomeroglu, "Packet scheduling and resource block allocation in wireless OFDMA networks: A utility-based classification," *submitted to IEEE Communications Surveys and Tutorials*, August 2014.

[12] ——, "On the delay-fairness through scheduling for wireless OFDMA networks," in *Proc. of IEEE 73rd Vehicular Technology Conference (VTC2011-Spring)*, May 2011, pp. 1–5.

[13] A. Sharifian, P. Djukic, H. Yanikomeroglu, and J. Zhang, "Mixed time-scale generalized fair scheduling for amplify-and-forward relay networks," in *Proc. of IEEE Global Communication Conference (GLOBECOM)*, December 2010, pp. 1–5.

[14] A. Sharifian, R. Schoenen, and H. Yanikomeroglu, "Joint realtime and non-realtime flows packet scheduling and resource block allocation in wireless OFDMA networks," *submitted to IEEE Transactions on Vehicular Technology*, April 2014.

[15] A. Sharifian, R. Schoenen, H. Yanikomeroglu, G. Senarath, H. T. Cheng, and P. Djukic, "System and Method for Network Resource Allocation Considering Users Experience, Satisfaction, and Operators Interest," US patent application no: 14/181,160, February 2014, PCT/US2014/016575, July 14, 2014, Filed by Huawei-Ottawa, Canada, including:
System and Method for User Satisfaction Modelling for Radio Resource Management in Wireless Communications, US provisional application no: 61/764,903, 14 February 2013, Filed by

Huawei-Ottawa, Canada,

System and Method for Joint Packet and Resource Scheduling, US provisional application no: 61/764,895, 14 February 2013, Filed by Huawei-Ottawa, Canada.

[16] H. Kushner and P. Whiting, "Convergence of proportional-fair sharing algorithms under general conditions," *IEEE Journal on Wireless Communication*, vol. 3, no. 4, pp. 1250–1259, July 2004.

[17] A. Sharifian, P. Djukic, H. Yanikomeroglu, and J. Zhang, "Max-min fair resource allocation for multi-user amplify-and-forward relay networks," in *Proc. of IEEE 72nd Vehicular Technology Conference (VTC2010-Fall)*, September 2010, pp. 1–5.

[18] ——, "Generalized proportionally fair scheduling for multi-user amplify-and-forward relay networks," in *Proc. of IEEE 71st Vehicular Technology Conference (VTC2010-Spring)*, May 2010, pp. 1–5.

[19] ——, "Resource Allocation Methods and Devices for Amplify-and-Forward Relay Network," US Patent 20,130,028,171, Granted on July 2013, CN201010175925.1, PCT/CN2011/071416, Huawei-Shenzhen, China.

[20] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. New York, NY, USA: Cambridge University Press, 2005.

[21] S. Shakkottai, T. Rappaport, and P. Karlsson, "Cross-layer design for wireless networks," *IEEE Communications Magazine*, vol. 41, no. 10, pp. 74–80, October 2003.

[22] M. Andrews, *A Survey of Scheduling Theory in Wireless Data Networks*, ser. The IMA Volumes in Mathematics and its Applications, Wireless Communications. New York, NY, USA: Springer, 2007.

[23] L. Tassiulas and A. Ephremides, "Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks," *IEEE Transactions on Automatic Control*, vol. 37, no. 12, pp. 1936–1948, December 1992.

[24] A. L. Stolyar, "Maximizing queueing network utility subject to stability: Greedy primal-dual algorithm," *Queueing System Theory and Application*, vol. 50, no. 4, pp. 401–457, August 2005.

[25] N. Salodkar, A. Bhorkar, A. Karandikar, and V. Borkar, "An on-line learning algorithm for energy efficient delay constrained scheduling over a fading channel," *IEEE Journal on Selected Areas in Communications*, vol. 26, no. 4, pp. 732–742, May 2008.

[26] D. P. Bertsekas, *Dynamic programming and optimal control*. Massachusetts, MA, USA: Athena Scientific, 2005.

[27] R. S. Sutton and A. G. Barto, *Reinforcement learning: an introduction*. Cambridge, MA, USA: MIT press, 1998.

[28] L. Bui, R. Srikant, and A. L. Stolyar, "Novel architectures and algorithms for delay reduction in back-pressure scheduling and routing," in *Proc. of IEEE Annual Joint Conference of the Computer and Communications Societies (INFOCOM)*, April 2009, pp. 2936–2940.

[29] P. Viswanath, D. Tse, and R. Laroia, "Opportunistic beamforming using dumb antennas," *IEEE Transactions on Information Theory*, vol. 48, no. 6, pp. 1277–1294, June 2002.

[30] A. Goldsmith and S.-G. Chua, "Variable-rate variable-power MQAM for fading channels," *IEEE Transactions on Communications*, vol. 45, no. 10, pp. 1218–1230, October 1997.

[31] J. Mo and J. Walrand, "Fair end-to-end window-based congestion control," *IEEE/ACM Transactions on Networking*, vol. 8, no. 5, pp. 556–567, October 2000.

[32] P. Svedman, S. Wilson, and B. Ottersten, "A QoS-aware proportional fair scheduler for opportunistic OFDM," in *Proc. of IEEE 60th Vehicular Technology Conference (VTC2004-Fall)*, September 2004, pp. 558–562.

[33] P. Svedman, S. K. Wilson, L. J. Cimini, and B. Ottersten, "Opportunistic beamforming and scheduling for OFDMA systems," *IEEE Transactions on Communications*, vol. 55, no. 5, pp. 941–952, May 2007.

[34] W. Rhee and J. M. Cioffi, "Increase in capacity of multi-user OFDM system using dynamic sub-channel allocation," in *Proc. of IEEE 51st Vehicular Technology Conference (VTC2000-Spring)*, May 2000, pp. 1085–1089.

[35] G. Song and Y. Li, "Utility-based resource allocation and scheduling in OFDM-based wireless broadband networks," *IEEE Communications Magazine*, vol. 43, no. 12, pp. 127–134, December 2005.

[36] G. Song, Y. Li, and L. J. Cimini, "Joint channel- and queue-aware scheduling for multiuser diversity in wireless OFDMA networks," *IEEE Transactions on Communications*, vol. 57, no. 7, pp. 2109–2121, July 2009.

[37] G. Song, Y. Li, J. Cimini, L. J., and H. Zheng, "Joint channel-aware and queue-aware data scheduling in multiple shared wireless channels," in *Proc. of IEEE Wireless Communications and Networking Conference (WCNC)*, March 2004, pp. 1939–1944.

[38] S. S. Haykin, *Adaptive filter theory*. Pearson Education, India, 2005.

[39] F. Kelly, A. Maulloo, and D. Tan, "Rate control in communication networks: Shadow prices, proportional fairness and stability," *Journal of the Operational Research Society*, vol. 49, no. 3, pp. 237–252, March 1998.

[40] A. Jalali, R. Padovani, and R. Pankaj, "Data throughput of CDMA-HDR a high efficiency-high data rate personal communication wireless system," in *Proc. of IEEE 51st Vehicular Technology Conference (VTC 2000-Spring)*, May 2000, pp. 1854–1858.

[41] G. Barriac and J. Holtzman, "Introducing delay sensitivity into the proportional fair algorithm for CDMA downlink scheduling," in *Proc. of IEEE International Symposium on Spread Spectrum Techniques and Applications*, September 2002, pp. 652–656.

[42] F. Kelly, "Fairness and stability of end-to-end congestion control," *European Journal of Control*, vol. 9, no. 2, pp. 159–176, April 2003.

[43] A. Gyasi-Agyei and S.-L. Kim, "Cross-layer multiservice opportunistic scheduling for wireless networks," *IEEE Communications Magazine*, vol. 44, no. 6, pp. 50–57, June 2006.

[44] L. Massoulie and J. Roberts, "Bandwidth sharing: objectives and algorithms," *IEEE/ACM Transactions on Networking*, vol. 10, no. 3, pp. 320–328, June 2002.

[45] A. L. Stolyar, "On the asymptotic optimality of the gradient scheduling algorithm for multiuser throughput allocation," *Operations Research*, vol. 53, no. 1, pp. 12–25, 2005.

[46] S. Low and D. Lapsley, "Optimization flow control. i. basic algorithm and convergence," *IEEE/ACM Transactions on Networking*, vol. 7, no. 6, pp. 861–874, December 1999.

[47] A. Sediq, R. Gohary, R. Schoenen, and H. Yanikomeroglu, "Optimal tradeoff between sum-rate efficiency and jain's fairness index in resource allocation," *IEEE Transactions on Wireless Communications*, vol. 12, no. 7, pp. 3496–3509, June 2013.

[48] R. Jain, D. Chiu, and W. Hawe, "A quantitative measure of fairness and discrimination for resource allocation in shared computer system," in *DEC Technical Report 301*, 1984.

[49] A. B. Sediq, R. H. Gohary, and H. Yanikomeroglu, "Optimal tradeoff between efficiency and jain's fairness index in resource allocation," in *Proc. of IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, September 2012, pp. 577–583.

[50] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York, NY, USA: Cambridge University Press, 2004.

[51] M. Andrews, L. Qian, and A. Stolyar, "Optimal utility based multi-user throughput allocation subject to throughput constraints," in *Proc. of IEEE Annual Joint Conference of the Computer and Communications Societies (INFOCOM)*, March 2005, pp. 2415–2424.

[52] S. Shakkottai and A. Stolyar, "Optimal utility based multi-user throughput allocation subject to throughput constraints," in *Proc. of International Teletraffic Congress*, September 2001, pp. 793–804.

[53] X. Wang and G. Giannakis, "Stochastic primal-dual scheduling subject to rate constraints," in *Proc. of IEEE Wireless Communications and Networking Conference (WCNC)*, March 2007, pp. 1527–1531.

[54] ——, "Resource allocation for wireless multiuser OFDM networks," *IEEE Transactions on Information Theory*, vol. 57, no. 7, pp. 4359–4372, July 2011.

[55] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, R. Vijayakumar, and P. Whiting, *"CDMA data QoS scheduling on the forward link with variable channel conditions"*. Massachusetts, MA, USA: Bell Laboratories, Lucent Technologies, 2000.

[56] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, P. Whiting, and R. Vijayakumar, "Providing quality of service over a shared wireless link," *IEEE Communications Magazine*, vol. 39, no. 2, pp. 150–154, February 2001.

[57] S. Asmussen, *Applied Probabability and Queues*. New York, NY, USA: Springer, 2000.

[58] H. Rasouli and A. Anpalagan, "An asymptotically fair subcarrier allocation algorithm in OFDM systems," in *Proc. of IEEE 69th Vehicular Technology Conference (VTC2009-Spring)*, April 2009, pp. 1–5.

[59] E. Rodrigues and F. Casadevall, "Control of the trade-off between resource efficiency and user fairness in wireless networks using utility-based adaptive resource allocation," *IEEE Communications Magazine*, vol. 49, no. 9, pp. 90–98, September 2011.

[60] E. B. Rodrigues, F. R. Lima, F. Casadevall, and F. R. P. Cavalcanti, "Capacity, fairness, and QoS trade-offs in wireless networks with applications to LTE," in *Resource Allocation and MIMO for 4G and Beyond*. Springer, 2014, pp. 157–211.

[61] A. Leon-Garcia, *Random processes, Probability, and Statistics for electrical engineering*. Upper Saddle River, NJ, USA: Prentice Hall, 2008.

[62] X. Wang, G. Giannakis, and A. Marques, "A unified approach to QoS-guaranteed scheduling for channel-adaptive wireless networks," *Proceedings of the IEEE*, vol. 95, no. 12, pp. 2410–2431, December 2007.

[63] P. P. Bhattacharya and A. Ephremides, "Optimal scheduling with strict deadlines," *IEEE Transactions on Automatic Control*, vol. 34, no. 7, pp. 721–728, July 1989.

[64] M. Andrews, S. C. Borst, F. Dominique, P. R. Jelenkovic, K. Kumaran, K. Ramakrishnan, and P. A. Whiting, "Dynamic bandwidth allocation algorithms for high-speed data wireless networks," *Bell Labs Technical Journal*, vol. 3, no. 3, pp. 30–49, July 1998.

[65] A. L. Stolyar and K. Ramanan, "Largest weighted delay first scheduling: Large deviations and optimality," *JSTOR Annals of Applied Probability*, vol. 11, no. 1, pp. 1–48, February 2001.

[66] S. Shakkottai and A. L. Stolyar, "Scheduling algorithms for a mixture of real-time and non-real-time data in HDR," in *Proc. of International Teletraffic Congress*, December 2001, pp. 793–804.

[67] L. Tassiulas and A. Ephremides, "Dynamic server allocation to parallel queues with randomly varying connectivity," *IEEE Transactions on Information Theory,*, vol. 39, no. 2, pp. 466–478, March 1993.

[68] S. Shakkottai and A. Stolyar, "Scheduling for multiple flows sharing a time-varying channel: The exponential rule," *American Mathematical Society Translations*, vol. 207, pp. 185–202, December 2002.

[69] R. Agarwal, V. Majjigi, R. Vannithamby, and J. Cioffi, "Efficient scheduling for heterogeneous services in OFDMA downlink," in *Proc. of IEEE Global Telecommunications Conference (GLOBECOM)*, November 2007, pp. 3235–3239.

[70] S. Bilal, M. Ritesh, and S. Ashwin, "Downlink scheduling for multiclass traffic in LTE," *EURASIP Journal on Wireless Communications and Networking*, vol. 2009, pp. 1:9–18:9, November 2009.

[71] B. Sadiq, S. J. Baek, and G. De Veciana, "Delay-optimal opportunistic scheduling and approximations: The log rule," *IEEE/ACM Transactions on Networking*, vol. 19, no. 2, pp. 405–418, August 2011.

[72] E. Altman, K. Avrachenkov, and S. Ramanath, "Multiscale fairness and its application to resource allocation in wireless networks," *Elsevier Computer Communications*, vol. 35, no. 7, pp. 820–828, February 2012.

[73] M. Rahman and H. Yanikomeroglu, "Enhancing cell-edge performance: a downlink dynamic interference avoidance scheme with inter-cell coordination," *IEEE Transactions on Wireless Communications,*, vol. 9, no. 4, pp. 1414–1425, April 2010.

[74] R. Schoenen and A. Otyakmaz, "QoS and flow management for future multi-hop mobile radio networks," in *Proc. of IEEE 72nd Vehicular Technology Conference (VTC2010-Fall)*, September 2010, pp. 1–5.

[75] S. Shakkottai and T. Rappaport, "Research challenges in wireless networks: a technical overview," in *Proc. of International Symposium on Wireless Personal Multimedia Communications*, October 2002, pp. 12–18.

[76] M. Salem, A. Adinoyi, M. Rahman, H. Yanikomeroglu, D. Falconer, Y.-D. Kim, E. Kim, and Y.-C. Cheong, "An overview of radio resource management in relay-enhanced OFDMA-based networks," *IEEE Communications Surveys and Tutorials*, vol. 12, no. 3, pp. 422–438, Third quarter 2010.

[77] N. Enderle and X. Lagrange, "User satisfaction models and scheduling algorithms for packet-switched services in UMTS," in *Proc. of IEEE 57th Vehicular Technology Conference (VTC2003-Spring)*, May 2003, pp. 1704–1709.

[78] L. Chen, B. Wang, X. Chen, X. Zhang, and D. Yang, "Utility-based resource allocation for mixed traffic in wireless networks," in *Proc. of IEEE Annual Joint Conference of the Computer and Communications Societies (INFOCOM) Workshop*, April 2011, pp. 91–96.

[79] W.-H. Kuo and W. Liao, "Utility-based optimal resource allocation in wireless networks," in *Proc. of IEEE Global Telecommunications Conference (GLOBECOM)*, December 2005, pp. 3508–3512.

[80] ——, "Utility-based resource allocation in wireless networks," *IEEE Transactions on Wireless Communications*, vol. 6, no. 10, pp. 3600–3606, October 2007.

[81] G. Song and Y. Li, "Cross-layer optimization for OFDM wireless networks-part I: Theoretical framework," *IEEE Transactions on Wireless Communications*, vol. 4, no. 2, pp. 614–624, March 2005.

[82] ——, "Cross-layer optimization for OFDM wireless networks-part II: Algorithm development," *IEEE Transactions on Wireless Communications*, vol. 4, no. 2, pp. 625–634, March 2005.

[83] R. Madan, S. Boyd, and S. Lall, "Fast algorithms for resource allocation in wireless cellular networks," *IEEE/ACM Transactions on Networking*, vol. 18, no. 3, pp. 973–984, June 2010.

[84] E. Rodrigues and F. Casadevall, "Adaptive radio resource allocation framework for multi-user OFDM," in *Proc. of IEEE 69th Vehicular Technology Conference (VTC2009-Spring)*, April 2009, pp. 1–6.

[85] E. B. Rodrigues and F. Casadevall, "Rate adaptive resource allocation with fairness control for OFDMA networks," in *Proc. of European Wireless Conference*, April 2012, pp. 18–20.

[86] A. Pantelidou and A. Ephremides, "What is optimal scheduling in wireless networks?" in *Proc. of Annual International Conference on Wireless Internet*, November 2008, pp. 1–8.

[87] M. Andrews, "Instability of the proportional fair scheduling algorithm for HDR," *IEEE Transactions on Wireless Communications*, vol. 3, no. 5, pp. 1422–1426, September 2004.

[88] K. Navaie, S. Valaee, A.-R. Sharafat, and E. Sousa, "Optimum model-based non-real-time downlink data transmission in heterogeneous DS-CDMA cellular networks," *IEEE Transactions on Wireless Communications,*, vol. 6, no. 6, pp. 2357–2367, June 2007.

[89] S. Ryu, B. Ryu, H. Seo, and M. Shin, "Urgency and efficiency based packet scheduling algorithm for OFDMA wireless system," in *Proc. of IEEE International Conference on Communications (ICC)*, May 2005, pp. 2779–2785.

[90] S. Ryu, B. Ryu, H. Seo, and M. Shi, "Urgency and efficiency based wireless downlink packet scheduling algorithm in OFDMA system," in *Proc. of IEEE 61st Vehicular Technology Conference (VTC2005-Spring)*, June 2005, pp. 1456–1462.

[91] S. Ryu, B. Ryu, and H. Seo, "Adaptive and QoS downlink multimedia packet scheduling for broadband wireless systems," in *Advances in Multimedia Information Processing*. Springer-Verlag, 2005, pp. 417–428.

[92] M. Katoozian, K. Navaie, and H. Yanikomeroglu, "Utility-based adaptive radio resource allocation in OFDM wireless networks with traffic prioritization," *IEEE Transactions on Wireless Communications*, vol. 8, no. 1, pp. 66–71, January 2009.

[93] ——, "Optimal utility-based resource allocation for OFDM networks with multiple types of traffic," in *Proc. of IEEE 67th Vehicular Technology Conference, (VTC2008-Spring)*, May 2008, pp. 2223–2227.

[94] R. Knopp and P. Humblet, "Information capacity and power control in single-cell multiuser communications," in *Proc. of IEEE International Conference on Communications (ICC)*, June 1995, pp. 331–335.

[95] G. Song, "Cross-layer resource allocation and scheduling in wireless multicarrier networks," Ph.D. dissertation, Georgia Institute of Technology, 2005.

[96] T. Schwarzfischer, "Quality and utility-towards a generalization of deadline and anytime scheduling," in *Proc. of International Conference on Automated Planning and Scheduling*, June 2003, pp. 277–286.

[97] H. Lei, M. Yu, A. Zhao, Y. Chang, and D. Yang, "Adaptive connection admission control algorithm for LTE systems," in *Proc. of IEEE 67th Vehicular Technology Conference (VTC2008-Spring)*, May 2008, pp. 2336–2340.

[98] S. Patil and G. de Veciana, "Managing resources and quality of service in heterogeneous wireless systems exploiting opportunism," *IEEE/ACM Transactions on Networking*, vol. 15, no. 5, pp. 1046–1058, October 2007.

[99] L. Hentila, P. Kyasti, M. Koske, M. Narandzic, and M. Alatossava, "Matlab implementation of the WINNER phase II channel model ver1.1," http://projects.celtic-initiative.org/winner+/phase_2_model.html, December 2007.

[100] R. Schoenen, A. B. Sediq, H. Yanikomeroglu, G. Senarath, Z. Chao, and H. T. Cheng, "Spectral efficiency and fairness tradeoffs in cellular networks with realtime+nonrealtime traffic mix using stochastic Petri nets," in *Proc. of IEEE 76th Vehicular Technology Conference (VTC2012-Fall)*, September 2012, pp. 1–5.

[101] H. T. Cheng and W. Zhuang, "An optimization framework for balancing throughput and fairness in wireless networks with QoS support," *IEEE Transactions on Wireless Communications*, vol. 7, no. 2, pp. 584–593, February 2008.

[102] P. Djukic and S. Valaee, "Towards guaranteed QoS in mesh networks: Emulating WiMAX mesh over WiFi hardware," in *Proc. of International Conference on Distributed Computing Systems Workshops*, June 2007, pp. 8–15.

[103] ——, "Delay aware link scheduling for multi-hop tdma wireless networks," *IEEE/ACM Transactions on Networking,*, vol. 17, no. 3, pp. 870–883, June 2009.

[104] A. Hottinen and T. Heikkinen, "Optimal subchannel assignment in a two-hop OFDM relay," in *Proc. of IEEE Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, June 2007, pp. 1–5.

[105] ——, "Subchannel assignment in OFDM relay nodes," in *Proc. of Conference on Information Sciences and Systems (CISS)*, March 2006, pp. 1314–1317.

[106] M. Herdin, "A chunk based OFDM amplify-and-forward relaying scheme for 4G mobile radio systems," in *Proc. of IEEE International Conference on Communications (ICC)*, June 2006, pp. 4507–4512.

[107] T. Riihonen, R. Wichman, and A. Hottinen, "Analysis of subcarrier pairing in a cellular OFDMA relay link," in *Proc. of International ITG Workshop on Smart Antennas*, February 2008, pp. 104–111.

[108] A. Hottinen, "Apparatus, method and computer program product providing sub-channel assignment for relay node," November 2006, US patent publication no. US2007/0098102 A1.

[109] M. Herdin, "Method for relaying information received via a first channel to a second channel and relay apparatus," August 2006, US patent publication no. US 2007/0036071 A1.

[110] A. Pandharipande and C. K. Ho, "Spectrum pool reassignment for a cognitive OFDM-based relay system," in *Proc. of International Conference on Cognitive Radio Oriented Wireless Networks and Communications, (CrownCom)*, August 2007, pp. 90–94.

[111] C. K. Ho and A. Pandharipande, "BER minimization in relay-assisted OFDM systems by subcarrier permutation," in *Proc. of IEEE 67th Vehicular Technology Conference (VTC2008-Spring)*, May 2008, pp. 1489–1493.

[112] Y. Guan-Ding, Z. Zhao-Yang, C. Yan, C. Shi, and Q. Pei-Liang, "Power allocation for non-regenerative OFDM relaying channels," in *Proc. of International Conference on Wireless Communications, Networking and Mobile Computing*, September 2005, pp. 185–188.

[113] I. Hammerstrom and A. Wittneben, "Joint power allocation for nonregenerative MIMO-OFDM relay links," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2006, pp. 49–50.

[114] ——, "On the optimal power allocation for nonregenerative OFDM relay links," in *Proc. of IEEE International Conference on Communications (ICC)*, June 2006, pp. 4463–4468.

[115] ——, "Power allocation schemes for amplify-and-forward MIMO-OFDM relay links," *IEEE Transactions on Wireless Communications*, vol. 6, no. 8, pp. 2798–2802, August 2007.

[116] L. Lihua, Z. Mingyu, Z. Xiaoxia, W. Haifeng, and Z. Ping, "Adaptive bit, power allocation and sub-carrier coupling for AF-OFDM relaying systems," in *Proc. of IEEE 67th Vehicular Technology Conference (VTC2008-Spring)*, May 2008, pp. 1494–1498.

[117] Y. Li, W. Wang, J. Kong, and M. Peng, "Subcarrier pairing for amplify-and-forward and decode-and-forward OFDM relay links," *IEEE Communications Letters*, vol. 13, no. 4, pp. 209–211, April 2009.

[118] Y. Li, W. Wang, J. Kong, W. Hong, X. Zhang, and M. Peng, "Power allocation and subcarrier pairing in OFDM-based relaying networks," in *Proc. of IEEE International Conference on Communications (ICC)*, May 2008, pp. 2602–2606.

[119] J. Jang and K. B. Lee, "Transmit power adaptation for multiuser OFDM systems," *IEEE Journal on Selected Areas in Communications*, vol. 21, no. 2, pp. 171–178, February 2003.

[120] E. Jeon, J. Yang, and D. K. Kim, "A low complexity subcarrier pairing scheme for OFDM based multiple AF relay systems," in *Proc. of International Conference on Convergence and Hybrid Information Technology (ICCIT)*, November 2008, pp. 675–678.

[121] Z. Mingyu, L. Lihua, W. Haifeng, Z. Ping, and T. Xiaofeng, "Sub-carrier coupling for OFDM based AF multi-relay systems," in *Proc. of IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, September 2007, pp. 1–5.

[122] W. Ying, Q. Xin-Chun, W. Tong, and L. Bao-Ling, "Power allocation and subcarrier pairing algorithm for regenerative OFDM relay system," in *Proc. of IEEE 65th Vehicular Technology Conference (VTC2007-Spring)*, April 2007, pp. 2727–2731.

[123] H. Li, H. Luo, X. Wang, and C. Li, "Throughput maximization for OFDMA cooperative relaying networks with fair subchannel allocation," in *Proc. of IEEE Wireless Communications and Networking Conference (WCNC)*, April 2009, pp. 1–6.

[124] W. Nam, W. Chang, S.-Y. Chung, and Y. Lee, "Transmit optimization for relay-based cellular OFDMA systems," in *Proc. of IEEE International Conference on Communications (ICC)*, June 2007, pp. 5714–5719.

[125] Z. Tang and G. Wei, "Resource allocation with fairness consideration in OFDMA-based relay networks," in *Proc. of IEEE Wireless Communications and Networking Conference (WCNC)*, April 2009, pp. 1–5.

[126] A. Hottinen and T. Heikkinen, "Distributed subchannel assignment in an OFDMA relay," in *Proc. of Conference on Performance Evaluation Methodologies and Tools*, October 2007, pp. 1–6.

[127] W.-G. Ahn and H.-M. Kim, "Proportional fair scheduling in relay enhanced cellular OFDMA systems," in *Proc. of IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, September 2008, pp. 1–4.

[128] L. Xiao and L. Cuthbert, "A two-hop proportional fairness scheduling algorithm for relay based OFDMA systems," in *Proc. of Conference on Wireless Communications, Networking and Mobile Computing (WiCOM)*, October 2008, pp. 1–4.

[129] B. Fan, W. Wang, Y. Lin, L. Huang, and K. Zheng, "Subcarrier allocation for OFDMA relay networks with proportional fair constraint," in *Proc. of IEEE International Conference on Communications (ICC)*, June 2009, pp. 1–5.

[130] M. Awad and X. Shen, "OFDMA-based two-hop cooperative relay network resources allocation," in *Proc. of IEEE International Conference on Communications (ICC)*, May 2008, pp. 4414–4418.

[131] X. Zhang, S. Chen, and W. Wang, "Multiuser radio resource allocation for multiservice transmission in OFDMA-based cooperative relay networks," *EURASIP Journal on Wireless Communications and Networking*, vol. 2009, pp. 1–13, January 2009.

[132] H. Jeong, J. H. Lee, and H. Seo, "Resource allocation for uplink multiuser OFDM relay networks with fairness constraints," in *Proc. of IEEE 69th Vehicular Technology Conference (VTC2009-Spring)*, May 2009, pp. 1–5.

[133] D. Zhang, Y. Wang, and J. Lu, "On QoS-guaranteed downlink cooperative OFDMA systems with amplify-and-forward relays: Optimal schedule and resource allocation," in *Proc. of IEEE Wireless Communications and Networking Conference (WCNC)*, April 2009, pp. 1–5.

[134] G. Li and H. Liu, "Resource allocation for OFDMA relay networks with fairness constraints," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 11, pp. 2061–2069, November 2006.

[135] J. Wang, Y. Zhao, and T. Korhonen, "Cross layer optimization with complete fairness constraints in OFDMA relay networks," in *Proc. of IEEE Global Telecommunications Conference (GLOBECOM)*, December 2008, pp. 1–5.

[136] L. You, M. Song, J. Song, Q. Miao, and Y. Zhang, "Adaptive resource allocation in OFDMA relay-aided cooperative cellular networks," in *Proc. of IEEE 67th Vehicular Technology Conference, (VTC2008-Spring)*, May 2008, pp. 1925–1929.

[137] D. Zhang, Y. Wang, and J. Lu, "Resource allocation in OFDMA based cooperative relay networks," in *Proc. of IEEE Global Telecommunications Conference (GLOBECOM)*, December 2008, pp. 1–5.

[138] J. N. Laneman, D. N. Tse, and G. W. Wornell, "Cooperative diversity in wireless networks: Efficient protocols and outage behaviour," *IEEE Transactions on Information Theory*, vol. 50, no. 12, pp. 3062–3080, December 2004.

[139] M. Grant and S. Boyd, *CVX: Matlab software for disciplined convex programming (web page and software). http://stanford.edu/ boyd/cvx.* Stanford University, 2009.

[140] D. M. Topkis, *Supermodularity and Complementarity.* USA, Princeton, New Jersey: Princeton University Press, 1998.