UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

MATHEUS CADORI NOGUEIRA

Estratégias de escalonamento OFDMA DL para redes móveis

Dissertação apresentada como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação

Orientador: Prof. Dr. Juergen Rochol

Porto Alegre
2016

CIP – CATALOGAÇÃO NA PUBLICAÇÃO

Nogueira, Matheus Cadori

Estratégias de escalonamento OFDMA DL para redes móveis / Matheus Cadori Nogueira. – Porto Alegre: PPGC da UFRGS, 2016.

88 f.: il.

Dissertação (mestrado) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, Porto Alegre, BR–RS, 2016. Orientador: Juergen Rochol.

1. OFDMA. 2. LTE. 3. QoS. 4. Escalonamento. 5. Parametrizável. I. Rochol, Juergen. II. Título.

AGRADECIMENTOS

Esta dissertação teve muitas colaborações diretas e indiretas e agradeço a todos que se envolveram ou participaram de alguma forma. Obrigado aos meus orientadores, professores Juergen Rochol e Cristiano Both. Obrigado Samuel Marini e Marcelo Caggliani por terem me auxiliado muito no trabalho. Obrigado aos professores do grupo de redes Lisandro Zambenedetti, Luciano Gaspary, Alberto e Marinho.

Alguns colegas do Grupo de Redes que eu tive o prazer de estudar junto e de aprender com vocês: Oscar Caicedo, Goiano, Barata, Ricardo, Angel, Raul, Maurício, Purinho, Nadal, Matheus Lehmman, Julianão, Jaruba, Luquinhas, Tche Pará, Guto, Catarina, Luine, Maicon, Marotta, Leo Faganelo, Jefferson, Raniery, Juan e Jean (Os Peruanos), Lucas Bondan, Batman, Estevão, Jedi, Rodolfo. Obrigado meus amigos queridos Alanzinho Bahea e Renata Neuland.

Obrigado aos meus amigos de Cruz Alta, sempre muito importantes, Flávio Lemos (Joka), Luciano, Modelo, Jorge Henrique. Valeu pela companhia nos momentos de descontração Zanatta, Aurora, Laurinha. Obrigado pessoal da empresa Ivan, Lili, Fábio, Felipe, Douglas, Luciana, Priscila.

Agradeço a Universidade Federal do Rio Grande do Sul pelo excelente nível de ensino, em especial ao instituto de informática que sempre proporcionou todo o suporte as minhas pesquisas.

Agradeço principalmente aos meus pais Fátima Cadori e Dario Nogueira, obrigado por terem me ensinado a buscar o conhecimento sempre, amo vocês. Agradeço aos meus irmãos do coração Michele e Alisson. Obrigado a toda minha família os quais sabem que são fundamentais para mim.

A senhora não foi alfabetizada Dona Margarida Cadori (vózinha), mas dedico esta dissertação a senhora, por ser este exemplo que sempre foi.

"Os livros não mudam o mundo, quem muda o mundo são as pessoas. Os livros só mudam as pessoas."

— MARIO QUINTANA.

ABSTRACT

The huge popularity of mobile devices that provides a ubiquitous Internet broadband access via radio networks and the volume of traffic generated by these devices in the base stations are increasing every year. Furthermore, the frequency which, mobile users are using web-based services, is increasing, requiring high transfer rates such as transmission of interactive videos. These factors have become the main challenges for the scheduling of radio resources. In order to meet these challenges, the Orthogonal Frequency Division Multiple Access (OFDMA), a key technology for multiple access in fourth generation networks, has also been considered for use in next-generation mobile radios. To implement an effective service to users, requirements such as high transfer rates, lower delay tolerance, minimum packet loss and maximum scheduling fairness, should be added to the requirements that emerged after the advent of the popularity of mobile devices. Therefore, new scheduling strategies should be projected. Despite efforts to solve the downlink (DL) scheduling problem on wireless networks, we are not aware of previous attempts that have addressed the above requirements in a single strategy. In this thesis, we took a step further in this direction and still considering the high densities in small cells inherent in modern networks. In additional, we address the radio DL resource scheduling problem for multiple users using LTE networks as a case study. A new optimal scheduler is modeled regarding Quality of Service (QoS) provisioning. In addition, a parameterized heuristic based on user channel quality and service delay is proposed to reach scheduling solutions for overbooked scenarios. Results demonstrate that the proposed scheduling approaches led to a throughput of 7.5% lower than the optimal ones and 25% lower packet losses in overloaded scenarios. Our model also ensures that the resultant scheduling is at least as fair as 0.91 in Jain fairness index. Additionally, the obtained results show a reasonable trade-off between spectral efficiency and QoS metrics.

Keywords: OFDMA, LTE, QoS, scheduling, parameterized.

RESUMO

A grande popularidade dos dispositivos móveis que provêm acesso ubíquo à Internet de banda larga, através de redes de rádio, e o volume de tráfego gerado por estes dispositivos estão aumentando a cada ano. Além disso, vem ampliando consideravelmente a frequência com que usuários de dispositivos móveis estão usando serviços baseados na Web. Alguns destes usuários podem estar acessando serviços que precisam de transmissão contínua como, por exemplo, vídeos interativos, outros podem estar apenas lendo e-mails, o que não exige um fluxo contínuo de dados. Mais do que isso, usuários com altos níveis de sinal podem atingir melhores taxas de transferência do que os com níveis menores. Portanto, encontrar a melhor relação entre os usuários que estão acessando serviços sensíveis ao atraso e aqueles que maximizam a taxa de transferência, e ainda ser justo na transmissão, é um relevante desafio para o escalonamento dos recursos de uma rede sem fio. Embora as pesquisas de escalonamento de recursos em redes sem fio tenham evoluído neste sentido, o recente aumento do volume de tráfego mencionado pode levar a uma sobrecarga no sistema, comprometendo o escalonamento. A fim de enfrentar estes desafios, o *Orthogonal Frequency Division Multiple Access* (OFDMA), tecnologia fundamental para o acesso múltiplo em redes de quarta geração, tem sido considerado também para ser utilizado na próxima geração de rádios móveis. Para implementar um serviço efetivo aos usuários, requisitos, tais como, altas taxas de transferência, tolerância baixa ao atraso, minimização da perda de pacotes e maximização da justiça no escalonamento, devem somar-se à característica, de alta densidade de usuários, que surgiu após o advento da popularização dos dispositivos móveis. Portanto, novas estratégias de escalonamento devem ser idealizadas. Nesta dissertação, deu-se um passo além na proposição de um escalonador para as redes móveis de próxima geração, que busca melhorar a relação entre taxa de transferência e atraso, consequentemente, levando a maiores índices de justiça no escalonamento resultante. O escalonador foi especialmente desenvolvido para lidar com altas densidades de usuários, inerentes às redes modernas, e as redes LTE foram utilizadas como caso de estudo. Desta forma, um novo escalonador ótimo que considera provisão dos requisitos acima mencionados, é modelado. Além disso, uma nova heurística parametrizável, baseada na qualidade do canal do usuário, no atraso permitido por cada serviço e na justiça do escalonamento é proposta, a fim de lidar com cenários sobrecarregados. Resultados demonstram que a abordagem de escalonamento proposta leva a uma taxa de transferência apenas 7,5% menor que os valores ótimos, com 25% a menos de perda de pacotes em cenários sobrecarregados. O modelo também garante que o escalonamento resultante seja pelo menos 0,91 na escala do índice de justiça de Jain. Finalmente, os resultados mostram uma melhor relação entre a eficiência espectral e as métricas de QoS.

Palavras-chave: OFDMA. LTE. QoS. escalonamento. parametrizável.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

3GPP	*3rd Generation Partnership Project*
4G	*Redes Móveis de Quarta Geração*
AMC	*Adaptive Modulation and Coding*
AMPL	*A Mathematical Programming Language*
CQI	*Channel Quality Indicator*
CR	*Code Rate*
DL	*Downlink*
eNB ou eNodeB	*Evolved Node B*
HARQ	*Hybrid Automatic Retransmission Request*
HTTP	*Hypertext Transfer Protocol*
LTE	*Long Term Evolution*
MCS	*Modulation and Coding Scheme*
MIMO	*Multiple Input Multiple Output*
OFDM	*Orthogonal Frequency Division Multiple*
OFDMA	*Orthogonal Frequency Division Multiple Access*
QCI	*QoS Class Identifier*
QoE	*Quality of Experience*
QoS	*Quality of Service*
SB	*Scheduling Block*
SC-FDMA	*Single Carrier Frequency Division Multiple Access*
SNR	*Signal to Noise Ratio*
TTI	*Transmition Time Interval*
TOS	*Type of Service*
UE	*User Equipament*
UL	*Uplink*

SUMÁRIO

1 INTRODUÇÃO

O número de usuários que utilizam sistemas de banda larga sem fio e o volume de tráfego gerado por estes usuários nas estações de transmissão estão aumentando a cada ano. Cada vez mais, usuários de *smartphones* estão usando serviços baseados na Web que exigem altas taxas de transferência, como, por exemplo, vídeos interativos. O tráfego de dados dos usuários é composto por diversos serviços disponibilizados pela internet, e precisam ser transmitidos dos aparelhos dos usuários até os servidores de conteúdo onde os serviços estão armazenados. Nas redes sem fio móveis, a transmissão desses dados é, basicamente, realizada através de rádios e antenas. O atual sistema de rádios móveis, utilizado para realizar a comunicação entre os usuários e os serviços, é o sistema de redes de quarta geração (4G), cuja principal tecnologia usada para implementá-lo é o *Long Term Evolution* (LTE). No LTE, existem rádios com diferentes capacidades de processamento para a transmissão dos dados dos usuários. Os rádios com menor capacidade de processamento são os equipamentos dos usuários, conhecidos como *User Equipament* (UE), e os com maior capacidade são as estações de rádio base conhecidas como *Evolved Node B* (eNodeB). Na comunicação entre os usuários e os serviços, o equipamento do usuário é responsável por enviar e receber dados de uma estação base, e esta estação base encaminha os dados para a próxima etapa do sistema de transmissão. Cada estação de rádio base pode receber os dados de diversos equipamentos de usuários, desde que não exceda o seu limite de capacidade de processamento de dados.

Um dos procedimentos que mais demanda capacidade de processamento nas estações de rádio base é o escalonamento de recursos. Este procedimento é caracterizado como a decisão da ordem de encaminhamento dos dados dos usuários para a próxima etapa de transmissão. Este procedimento é chamado de escalonamento de recursos. Os recursos em LTE são porções de frequência disponíveis para transmitir os dados dos usuários. Estas porções de frequência podem ser consideradas como matrizes bidimensionais, em que o eixo horizontal representa o domínio de tempo e o eixo vertical representa o domínio de frequência. Estas matrizes são subdivididas em blocos e são preenchidas com os dados dos usuários. Cada porção de frequência possui um número limitado de blocos de recursos em um dado tempo, portanto, caso existam mais dados de usuários do que o limite de recursos para um tempo x, os dados excedentes serão enviados somente em um próximo tempo $x + 1$. Este caso, onde não existem recursos suficientes para transmitir os dados dos usuários, é tratado como um cenário sobrecarregado de dados de usuários. Esta sobrecarga de dados no sistema de transmissão leva a um incremento de tempo na comunicação entre o usuário e o serviço sendo acessado, e idealmente, este atraso deve ser evitado. Mais do que isso, alguns serviços precisam de transmissão contínua como, por exemplo, vídeos interativos. Já outros, como por exemplo, e-mail, não exigem um fluxo contínuo de dados.

O escalonador de recursos possui a informação de qual o tipo de serviço o usuário está acessando, portanto, é possível priorizar determinados serviços no escalonamento. Outra infor-

mação importante para o escalonamento é o nível de sinal de cada equipamento de usuário. O usuário que estiver mais próximo da estação base tende a ter melhores níveis de sinal, e os que estiverem mais longe tendem a ter uma pior condição de sinal. Os usuários com altos níveis de sinal podem atingir melhores taxas de transferência que aqueles com níveis menores. Portanto, encontrar a melhor relação entre os serviços sensíveis ao atraso e os usuários com altas taxas de transferência é um desafio para o escalonamento de recursos nas redes sem fio de 4G. Mais uma característica que alguns autores consideram no escalonamento é a justiça, visando não escalonar somente recursos para os usuários com maiores níveis de sinal ou somente para os que estão utilizando os serviços que precisam de um fluxo contínuo de dados.

Embora as pesquisas de escalonamento de recursos em redes sem fio de 4G tenham evoluído no tratamento da relação descrita anteriormente, o recente aumento do volume de tráfego mencionado, inerente às próximas gerações de redes sem fio, pode levar a uma sobrecarga no sistema. Para que o escalonamento de recursos de redes sem fio da próxima geração esteja preparado para grandes volumes de tráfego, é fundamental projetar e desenvolver algoritmos de escalonamento mais otimizados e que considerem características, tais como, serviços sensíveis ao atraso e recursos limitados de rádio. No escalonamento de recursos da próxima geração de redes sem fio, o estado da arte indica utilizar a consciência dos tipos de serviços trafegados. As informações adicionais relacionadas ao conteúdo dos dados, o atraso tolerado por cada tipo de serviço e a justiça na distribuição de recursos são características relevantes para a tomada de decisão no escalonamento, a fim de atender às crescentes expectativas dos usuários (AHMED; JAGANNATHAN; BHASHYAM, 2015).

1.1 Problema e Motivação

As futuras gerações de redes celulares são esperadas por prover acesso a banda larga de forma ubíqua, para um número crescente de usuários móveis (CAPOZZI et al., 2013). A fim de comportar este crescimento, as redes móveis de próxima geração, estão considerando a utilização do *Orthogonal Frequency Division Multiple Access* (OFDMA), como tecnologia de acesso múltiplo (PENG et al., 2014). No OFDMA, a troca de informações entre o equipamento do usuário e a estação base é realizada por meio de dois canais chamados *downlink* (DL) e *uplink* (UL). Estes canais são responsáveis por transportar informações de controle e dados de usuários. Os dados transmitidos no sentido DL utilizam OFDMA, enquanto os dados transmitidos no sentido UL, utilizam uma única portadora *Single Carrier Frequency Division Multiple Access* (SC-FDMA). Portanto, DL herda características do OFDMA, proporcionando que os dados provindos de vários usuários sejam enviados em um único intervalo de transmissão.

Os dados dos usuários são transmitidos pelo escalonador através de recursos de rádio, usando um esquema de modulação e codificação. Este esquema adiciona informações redundantes com base na qualidade do canal dos usuários. A fim de melhorar a eficiência de espectro, algumas estratégias de escalonamento priorizam usuários que tenham uma melhor qualidade de

canal. Embora essa abordagem consiga uma maior taxa de transferência, isto pode fazer com que os usuários com condições de canal ruins fiquem sem transmitir (AVOCANH et al., 2013). Outra abordagem de escalonamento prioriza dados de usuários que estejam a mais tempo na fila para serem transmitidos (TRAN; ELTAWIL, 2012). Apesar da fila de pacotes ser considerada nesta abordagem, o desempenho geral da rede pode diminuir, devido à escolha dos usuários com maior atraso, ao invés de aqueles com melhores condições de canal.

A taxa de transferência máxima, o atraso mínimo dos pacotes e a consideração da justiça são requisitos essenciais para transmissões de recursos em redes sem fio e, portanto, devem ser considerados no processo de escalonamento de redes móveis de próxima geração. Neste sentido, o principal desafio no projeto de um escalonador é como distribuir os recursos de transmissão entre os usuários, visando uma melhor relação entre a taxa de transferência e o atraso, e, ainda assim, considerando a justiça na transmissão. Este desafio deve ser alcançado mediante o cumprimento do maior número de requisitos para tantos usuários quanto possível, ainda assim, sem a perda de eficiência em cenários com altas densidades de usuários, requisito da próxima geração de redes móveis (CAPOZZI et al., 2013). Apesar dos esforços empregados em abordagens anteriores (AVOCANH et al., 2013; TRAN; ELTAWIL, 2012; KWAN; LEUNG; ZHANG, 2008; KWAN; LEUNG; ZHANG, 2009), o advento da popularização dos dispositivos móveis, trouxe à característica de cenários sobrecarregados ao escalonamento e não há conhecimento de trabalhos que abordaram os requisitos acima mencionados, considerando a relação ideal entre eles.

1.2 Objetivos e Principais Contribuições

Esta dissertação está inserida no contexto da próxima geração de redes sem fio, mais especificamente, na busca de um escalonador de recursos de rádio para os usuários que considera os atuais requisitos das redes de 4G, assim como, os requisitos de priorização de serviços em cenários sobrecarregados, que são essências nas redes de próxima geração. O escalonador desenvolvido neste trabalho busca melhorar a relação entre taxa de transferência e atraso, consequentemente, levando a maiores índices de justiça no escalonamento resultante. O escalonador foi especialmente desenvolvido para lidar com altas densidades de usuários, inerentes às redes modernas. Mais especificamente, foi formalizado um modelo DL ótimo que considera requisitos como a taxa de transferência, o atraso e a justiça. Este modelo proposto foi linearizado e avaliado através de uma perspectiva exata, obtendo assim um *baseline*. Mais do que isto, uma métrica de escalonamento, que considera a relação entre a taxa de transferência, a justiça e o atraso, foi proposta e avaliada em relação a este *baseline*. Já para avaliação e validação da métrica proposta, utilizou-se a tecnologia *Long Term Evolution* (LTE) que atualmente implementa as redes de 4G. Ademais, um escalonador de recursos LTE foi construído para facilitar a avaliação e as análises da métrica de escalonamento proposta.

Em suma, as principais contribuições deste trabalho são quatro: (*i*) a formalização de um

modelo de escalonamento DL ótimo que considera diferentes tipos de serviço; (*ii*) o desenvolvimento de uma abordagem heurística, que utiliza a métrica proposta, para trabalhar com cenários de alta densidade de usuários; (*iii*) o desenvolvimento de um simulador para escalonamento LTE; e por fim, (*iv*) a avaliação e discussão dos resultados do impacto do escalonamento de recursos de um grande número de usuários.

1.3 Organização do Documento

O restante deste documento esta organizado como segue. No Capítulo 2, uma breve contextualização da área de escalonamento e de LTE é apresentada, assim como, é descrito o modelo do sistema de escalonamento DL. Alguns trabalhos relacionados são estudados no Capítulo 3, assim como, é explicada a relação destes trabalhos com esta dissertação. No Capítulo 4, é abordado o método linearizado do problema. Ainda neste capítulo, a ideia por trás do método parametrizado heurístico é explicada, assim como, é descrita a forma de escalonamento através do método heurístico. O simulador DL para redes baseadas em OFDMA é sistematicamente explicado no Capítulo 5. Já no Capítulo 6, os resultados das avaliações do modelo linearizado e heurístico são demonstrados e discutidos. E, finalmente no Capítulo 7 há uma breve conclusão desta dissertação destacando os objetivos alcançados e aludindo perspectivas para trabalhos futuros.

2 REFERENCIAL TEÓRICO

Neste capítulo, é apresentada uma profunda análise bibliográfica sobre recursos de redes sem fio e escalonamento. Em seguida, descreve-se o modelo do sistema de escalonamento utilizado. Para tanto, as redes *Long Term Evolution (LTE)*, que atualmente implementam o 4G, são utilizadas neste trabalho para avaliação e validação da abordagem apresentada.

2.1 Recursos DL em Redes LTE

A transmissão de recursos de radiofrequência em LTE é realizada através de rádios e antenas. Além disso, as duas principais tecnologias de transmissão de dados móveis empregadas no LTE são *Time Division Duplex* (TDD) e *Frequency Division Duplex* (FDD). As frequências utilizadas por estes rádios variam de acordo com o país onde a transmissão é licenciada pelo governo. Entretanto, o *3rd Generation Partnership Project* (3GPP) regulamenta as faixas de frequência da tecnologia. Uma lista completa destas frequências pode ser encontrada em (3GPP 33.101 tabela 5.5-1). A seguir, na Tabela 2.1, são sumarizadas as frequências operadas no Brasil.

Tabela 2.1: Frequências utilizadas para o LTE no Brasil.

Banda	Frequência (MHz)	Modo Duplex	Operadoras
3	1800	FDD	Nextel, TIM
7	2600	FDD	Claro, Oi, TIM, Vivo
38	2600	TDD	On Telecom, SKY Brasil

Fonte: pelo autor (2016).

Na operação das redes LTE no Brasil, as frequências acima descritas são utilizadas em diversos rádios e antenas de diferentes marcas e modelos. Além disso, essas frequências são subdivididas em faixas, o que consiste em uma largura de banda utilizada na transmissão, por frequência. Sabendo-se que uma portadora representa um sinal analógico em forma de onda que será modulado (alterado) para representar a informação a ser transmitida, o OFDMA consiste na divisão da banda disponível em subportadoras. As larguras de banda disponíveis no LTE seguem sumarizadas na Tabela 2.2, assim como as respectivas quantidades de recursos e subportadoras disponíveis em cada faixa.

Para transmissão dos recursos de rádio, assume-se que cada transmissor possui domínios de cobertura, os quais são caracterizados como células ou setores. Algumas nomenclaturas comuns neste sentido são *small cells*, *picocells*, *femtocells* e *macrocells*. As *macrocells* geralmente são compostas por mais de uma célula. Além do mais, a combinação entre *macrocells* e *small-cells* é conhecida como redes heterogêneas. O uso de *small cells* está tornando-se importante no cenário das redes de próxima geração, devido à capacidade de disponibilizar mais recursos para o sistema, se comparado com as redes homogêneas de *macrocells*. As *small cells* podem

Tabela 2.2: Larguras de banda em LTE.

Largura de Banda do Canal	Largura de Banda Usável	Total de subportadoras	Total de Recursos
1,4 MHz	1,08 MHz	72	6
3 MHz	2,7 MHz	180	15
5 MHz	4,5 MHz	300	25
10 MHz	9 MHz	600	50
15 MHz	13,5 MHz	900	75
20 MHz	18 MHz	1200	100

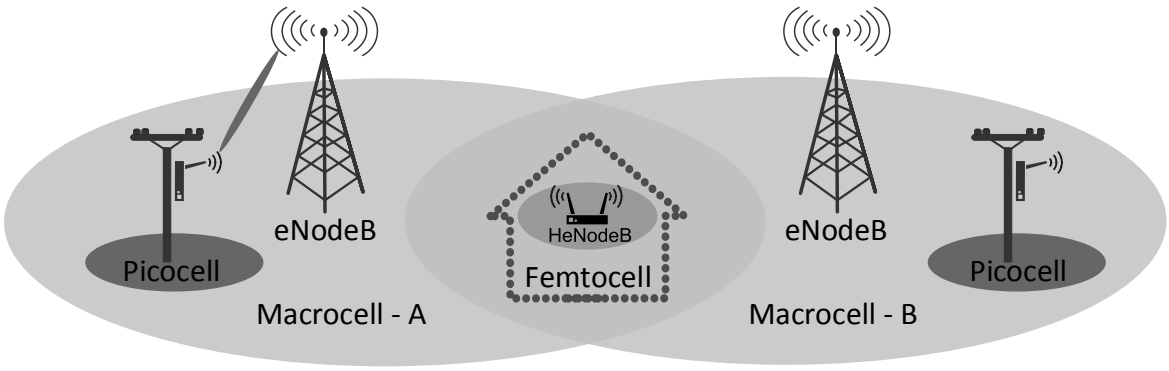
Fonte: pelo autor (2016).

ser caracterizadas tanto por *picocells* (algumas vezes referenciadas por *hot zones*), quanto por *femtocells*. As definições de *picocells* e *femtocells* são um tanto quanto variáveis, mas, normalmente, as principais características de diferenciação podem ser resumidas como se segue (SESIA; TOUFIK; BAKER, 2011):

- As *femtocells* são tipicamente grupos fechados de assinantes. Tais células são acessíveis apenas a um grupo limitado de usuários. Um exemplo disto é a implementação da rede através de estações base residenciais de baixa potência, *e.g.*, (*Home eNodeB*); em contraste, as *picocells* são geralmente abertas a todos os usuários, podendo ocasionalmente oferecer um tratamento preferencial a alguns usuários, por exemplo, para a equipe de um estabelecimento em particular;
- A potência de transmissão das *femtocells* é menor, sendo tipicamente concebida para cobrir uma casa ou um apartamento. Já as *picocells* geralmente operam com uma potência de transmissão superior, suficiente para cobrir uma empresa, um *shopping*, ou simplesmente para estender a cobertura macro celular.

Na Figura 2.1 os conceitos de *femtocells*, *picocells* e *macrocells* são contextualizados:

Figura 2.1: Tipos de células.



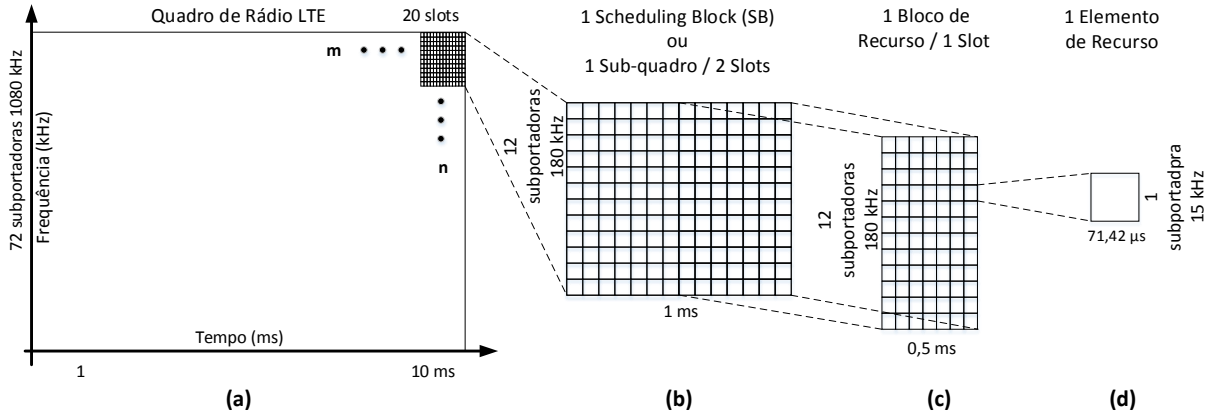
Fonte: adaptada de ((MAROTTA et al., 2015)).

As células descritas são partilhadas em recursos que podem ser de dois tipos, em LTE: *uplink* (UL) e *downlink* (DL). Tais tipos referem-se ao sentido da transmissão. Por um lado, no DL, os dados são transmitidos da estação base para os usuários. Por outro lado, no UL, os dados dos usuários são transmitidos no sentido contrário, *i.e.*, dos usuários para a estação base. Ademais, estes tipos de recursos utilizam diferentes tecnologias para transmissão. Enquanto o DL utiliza o OFDMA, o UL utiliza o SC-FDMA com apenas uma portadora.

Deste modo, um quadro de recurso DL, descrito na Figura 2.2, pode ser considerado uma matriz bidimensional, em que o eixo horizontal representa o domínio de tempo e o eixo vertical representa o domínio de frequência. Em caso de múltiplas antenas (*Multiple Input Multiple Output* (MIMO)) serem utilizadas, uma camada adicional é necessária para representar o domínio correspondente ao espaço. Esta matriz é subdividida de acordo com a especificação LTE e é preenchida com blocos de dados dos usuários. A estrutura de recursos tempo-frequência (veja a Figura 2.2) foi dividida em (a) à (d), para ser melhor representada. Cada parte da figura é apresentada com o nome do recurso e sua respectiva medida no domínio de tempo e frequência. Uma largura de banda de 1,4 MHz foi utilizada na representação.

A primeira parte (a) representa um exemplo de um quadro de rádio de 10 ms por 1,080 kHz, sobre 72 subportadoras. Este quadro é subdividido (b) em 10 sub-quadros de 1 ms por 180 kHz, sobre 12 subportadoras. Cada sub-quadro é ainda subdividido em 2 *slots*, como pode ser observado na parte (c). Um *slot* de 0,5 ms por 180 kHz, sobre 12 subportadoras, é também chamado de Bloco de Recursos (*Resource Bolcks* (RB)). Finalmente, na parte (d), um elemento de recurso ou *Resource Element* (RE) de 71,42 μs por 15 kHz sobre 1 subportadora, é apresentado. Daqui em diante, por convenção, o termo *Scheduling Block* (SB) será usado para se referir a dois blocos de recurso sob um sub-quadro de 1 ms. Na literatura LTE, existem diferentes entendimentos sobre a unidade mínima de transmissão. Entretanto este trabalho baseia-se na norma do 3GPP, onde a menor unidade a ser transmitida no LTE são dois RBs, *i.e.*, um SB.

Figura 2.2: Visão geral do recurso LTE DL nos domínios tempo/frequência.



Fonte: pelo autor (2016).

Na Figura 2.2 a estrutura de recursos LTE demonstrada, representa um quadro DL com

configuração padrão e prefixo cíclico normal. É importante ressaltar que alguns espaços dentro do quadro são reservados para propósito de sinalização da tecnologia, tais como, sinais de sincronização primário e secundário, sinais de referência, indicador de controle de formato e canal físico de *broadcast*, além dos *slots* desperdiçados devido ao modelo de desenvolvimento da tecnologia. Estes símbolos são tratados daqui em diante como símbolos de sinalização (SS). Estes espaços ocasionam uma sobrecarga na transmissão e idealmente deveriam ser levados em consideração pelas estratégias de escalonamento. A fim de melhorar a legibilidade deste trabalho, os SSs não são representados na Figura 2.2. No entanto, duas figuras detalhadas do quadro LTE encontram-se no apêndice A, uma mostrando a estrutura completa sem as sinalizações da tecnologia, e a outra contendo todas as sinalizações SSs. Na próxima subseção, são descritos os principais requisitos de um escalonador LTE.

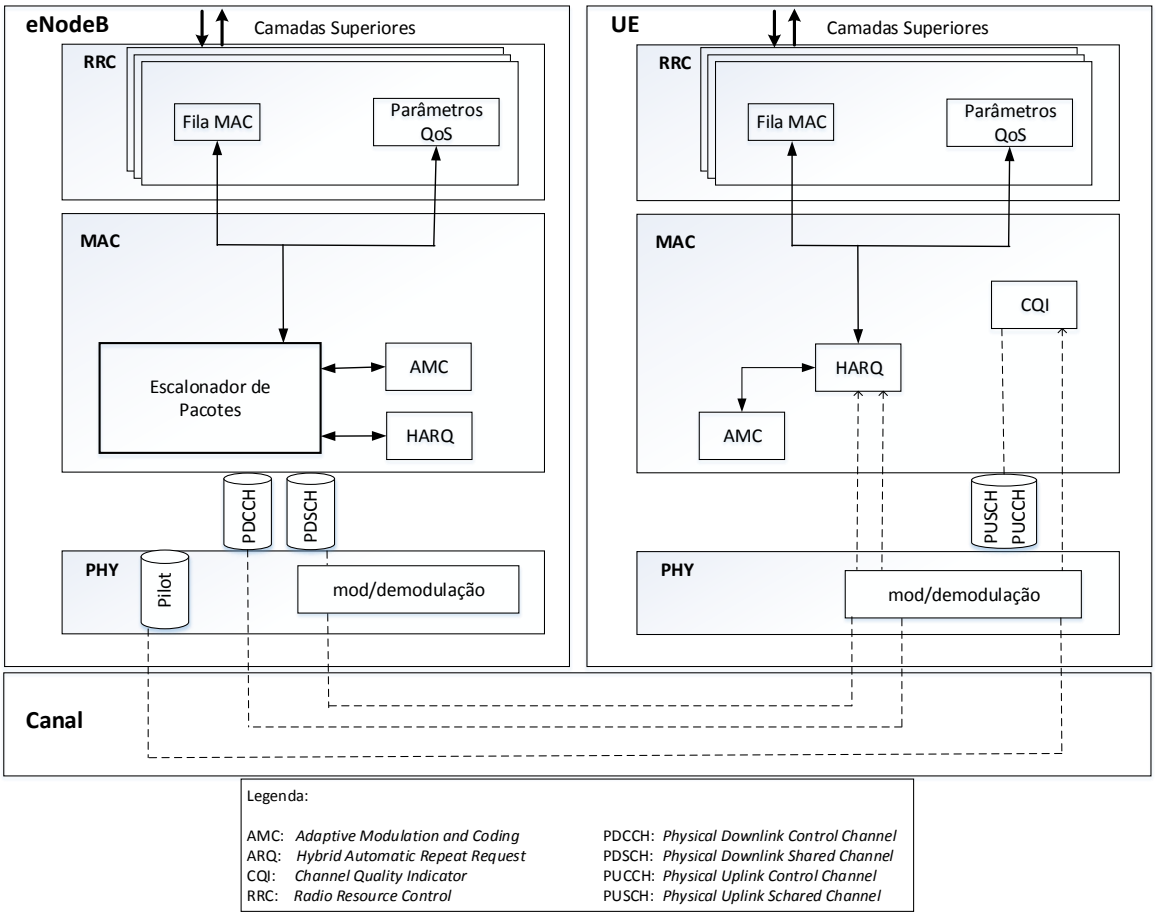
2.2 Principais Requisitos no Escalonamento DL

Em estações base LTE, também chamadas de *Evolved Node B* (eNodeB), uma entidade chamada escalonador lida com a distribuição de recursos entre os usuários. Este escalonador recebe informações de outros componentes da arquitetura LTE, como pode-se observar na Figura 2.3. Além disso, um eNodeB pode servir recursos a múltiplas células ou setores, com estes recursos sendo distribuídos a cada intervalo de tempo, (*Transmission Time Interval* (TTI)) de 1 *ms*. Ademais, um SB é o menor recurso de rádio que pode ser atribuído a um equipamento de usuário, *User Equipament* (UE), para a transmissão de dados (KWAN; LEUNG; ZHANG, 2008). Portanto, com base nas informações recebidas das camadas superiores, o escalonador seleciona um conjunto de dados de usuários para transmitir em cada TTI. Já os dados dos usuários são transmitidos utilizando um *Modulation and Codification Scheme* (MCS), escolhido de acordo com o *Channel Quality Indicator* (CQI) do usuário.

Na Figura 2.3, é ilustrada a arquitetura de escalonamento de recursos em LTE em duas perspectivas, sendo uma da eNodeB e outra do UE. Da mesma forma, são demonstradas as camadas mais importantes ao escalonamento LTE (*Physical Layer* (PHY), *Medium Accesss Control* (MAC), *Radio Resource Control* (RRC)). O LTE utiliza diferentes canais para a transmissão de dados, de usuários e de controle da tecnologia. Os canais PUSCH e PDSCH são utilizados para transmitir os dados dos usuários e os canais PUCCH e PDCCH transmitem as informações de controle da tecnologia.

O *feedback* do CQI é de fundamental importância no escalonamento LTE. Cada CQI é calculado como uma medida quantificada e experimentada do *Signal to Interference plus Noise Ratio* (SINR) do usuário. O principal problema relacionado a métodos de retorno de CQI é encontrar um bom equilíbrio entre uma estimativa precisa de qualidade de canal e uma reduzida quantidade de sinalização utilizada para este fim. O CQI também é diretamente ligado ao módulo *Adaptative Modulation and Coding* (AMC), onde o controle de potência do rádio é regulado. Neste módulo, o MCS é selecionado, tentando maximizar a taxa de *bits* suportada.

Figura 2.3: Visão geral da arquitetura LTE.



Fonte: adaptada de ((CAPOZZI et al., 2013)).

Por isso, usuários com melhores SINRs irão atingir mais altas taxas de transferência, enquanto usuários com baixos níveis de SINR atingirão taxas baixas. É importante destacar que os MCSs são limitados superiormente, logo, a partir de um limiar de qualidade de sinal, a taxa de transferência não aumentará. Algumas abordagens de escalonamento utilizam o AMC para gerenciar o controle de potência, o que não acontece neste trabalho, por motivos detalhados no decorrer da dissertação.

Existem diferentes esquemas de modulação e codificação (MCSs) em LTE, e um atributo que varia entre eles é a quantidade de redundância que cada um deles insere na transmissão. Esta redundância serve para garantir a consistência dos dados de usuários, dada à condição do canal do mesmo. Os MCSs são escolhidos no momento do escalonamento de acordo com o CQI do usuário, que basicamente representa a condição do canal daquele utilizador num determinado momento. Por exemplo, um usuário mais próximo de uma estação base, com boas condições de canal, ou seja, um alto nível de CQI, pode usar um MCS que insere menos redundância, restando assim mais recursos para os dados do usuário. Em contraponto, um usuário longe da

estação base, com más condições de canal, *i.e.*, baixo nível de CQI, deve usar um MCS que insere mais redundância, deixando assim, menos recursos para os dados. Esquemas comuns de modulação em LTE são: QPSK, 16QAM e 64QAM. Os esquemas de MCS em LTE são sumarizados e apresentados na Tabela 2.3 (adaptada da tabela 7.1.7.1-1 do 3GPP 36.213).

Tabela 2.3: Esquemas de MCS.

Índice MCS	Ordem de Modulação	Tipo de Modulação
0-9	2	QPSK
10-16	4	16 QAM
17-28	6	64 QAM

Fonte: pelo autor (2016).

Em LTE, os dados dos usuários são compostos de tráfegos heterogêneos e com diferentes classes de serviços. Aplicações como *Video on Demand* (VoD), *File Transfer Protocol* (FTP) e navegação HTTP convencional possuem requisitos de QoS completamente diferentes. Por exemplo, em vídeo-conferências, o atraso e o *jitter* devem ser levados em conta, enquanto que no FTP, requer-se baixa quantidade de pacotes perdidos. Para considerar múltiplos requisitos de QoS, em LTE, são implementadas diferentes classes de prioridade baseadas em *bearers*, onde cada *bearer* é atribuído a um escalar chamado *QoS Class Identifier* (QCI). A caracterização do QCI é baseada em prioridades, considerando taxa de perda de pacotes, limite de atraso e o tipo do *bearer* que pode ser, *Guaranteed Bit Rate* (GBR) e Non-GBR. Outro fato importante sobre a priorização de serviços em LTE é que esta priorização por QCI é mapeada diretamente para a camada de transporte, e relacionada com os mapeamentos de redes tradicionais, como o *Differentiated Services Code Point* (DSCP), facilitando o manuseio desta diferenciação em cenários heterogêneos. Uma tabela completa dos serviços e das classificações em LTE pode ser encontrada na norma do 3GPP ((3GPP TS 23.203, 2010) tabela 6.1.7). Uma sumarização destes serviços, é apresentada na Tabela 2.4.

Tabela 2.4: Tabela de prioridades de alguns serviços em LTE.

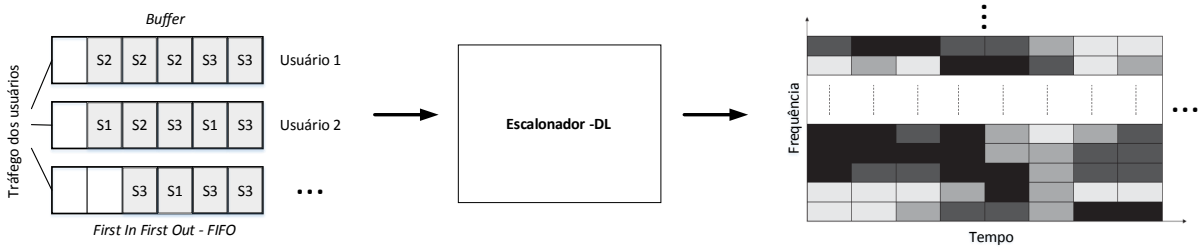
QCI	Tipo do Recurso	Provisão de Atraso	Exemplo de Serviços
1	GBR	100 ms	VoIP (sinalização), Video
4	GBR	50 ms	Jogos Online (Tempo Real), VoIP (chamada)
5	GBR	150 ms	Chat com Vídeo
6	non-GBR	300 ms	HTTP

Fonte: pelo autor (2016).

Além disso, nessas redes, os usuários podem ser móveis e (ou) fixos, de modo que as condições do canal serão naturalmente variáveis. Consequentemente, é necessário o provisionamento de atrasos diferentes para cada tipo de serviço, e, a capacidade de adaptação dos escalonadores

é fundamental. Neste sentido, o escalonador deve atender a estas características, considerando ainda eficientes taxas de transferência e a justiça na transmissão. Uma visão do processo de escalonamento pode ser visto na Figura 2.4.

Figura 2.4: Visão geral do escalonamento.



Fonte: pelo autor (2016).

Primeiramente, como pode ser observado na Figura 2.4, o tráfego dos usuários é armazenado em um *buffer* antes do escalonamento. Em seguida, este tráfego é enviado para o escalonador, usando uma simples abordagem *First In First Out* (FIFO). É possível perceber através da figura que cada usuário possui diversos pacotes de determinados serviços, onde, o serviço 1 é representado por S1, o serviço 2 por S2 e assim sucessivamente. Posteriormente, o tráfego destes usuários chega ao escalonador e precisa ser ordenado e organizado, de forma a preencher a matriz de tempo e frequência que representa os recursos DL disponíveis. Na matriz de recursos, observam-se vários retângulos de diferentes tons de cinza, onde cada usuário é atribuído a um tom. Note que, alguns usuários possuem mais de um bloco de recurso na matriz. As definições de quais blocos são atribuídos para quais usuários é feita seguindo algum critério do escalonador. Na próxima seção, apresentam-se algumas das principais abordagens de escalonamento disponíveis na literatura.

2.3 Tipos de Escalonamento

O escalonamento de múltiplos usuários encontra seus fundamentos na interface entre a teoria da informação e a teoria de filas, sempre considerando a maximização da alocação de recursos. Antes de estabelecer um algoritmo propriamente dito, precisa-se formular uma métrica relacionada com a capacidade que a rede possui, e, em seguida, otimizar todas as soluções possíveis que satisfaçam um conjunto de restrições predeterminadas. Tais restrições podem ser físicas (por exemplo largura de banda e potência total) ou relacionadas com a qualidade de serviço (QoS). A teoria da informação oferece uma gama de possíveis métricas de capacidade que são relevantes nos diferentes cenários de operação do sistema. Dois exemplos proeminentes são explicados na sequência: a capacidade de Shannon e a capacidade de atraso restrita.

2.3.1 Visando a Capacidade de Shannon

A capacidade de Shannon (também conhecida como a capacidade ergódica) é definida como a máxima taxa de dados que pode ser enviada através do canal. Na propagação das ondas de rádio na atmosfera podem ocorrer reflexões causadas por diversos fatores, dentre eles, o solo, as intempéries climáticas e os obstáculos na linha de visada direta entre as antenas. Estas reflexões provocam alterações na amplitude e no caminho percorrido pelas ondas de rádio, ocasionando variações na potência do sinal recebido. Estas variações são chamadas de desvanecimento e são consideradas pela capacidade de Shannon. A métrica de capacidade de Shannon considera a taxa de dados média, no longo prazo, que pode ser entregue a um usuário, quando este não tem quaisquer restrições de latência.

Abordagem *Maximum Sum Rate Scheduling* (MSRS)

Vários trabalhos (KNOPP; HUMBLET, 1995; TSE; HANLY, 2006) mostram que a capacidade de Shannon é alcançada calculando $\sum_{k=1}^{K} R_k$, onde R_k é o *rate* máximo total alocado para o usuário k, considerando uma dada potência de transmissão. Esta abordagem é conhecida na literatura como *Maximum Sum Rate Scheduling* (MSRS). Uma variante desta estratégia de escalonamento, sem o controle da potência de transmissão, é considerada em (KNOPP; HUMBLET, 1995), esta estratégia é chamada, de escalonamento *Maximum-rate Constant-power*, onde somente usuários com os melhores ganhos de canal são escalonados em cada SB, mas sem nenhuma adaptação de potência nos rádios.

Abordagem *Proportional Fair Scheduling* (PFS)

Pelo fato da abordagem ergódica não possuir restrições sobre atraso, os resultados de compartilhamento de recursos do canal acabam tornando-se injustos. Quando o QoS requerido pela aplicação inclui restrições de latência, estratégias de escalonamento que consideram a capacidade ergódica não são aplicáveis, e outras abordagens tem de ser consideradas. Uma destas abordagens é a *Proportional Fair Scheduling*(PFS). Esta conhecida abordagem utiliza a qualidade do canal do usuário, no momento do escalonamento, e compara com a condição média do canal ao longo do tempo. O escalonamento do usuário k_m, no SB m, em qualquer sub-quadro f, é dado por (CAIRE; MÜLLER; KNOPP, 2007):

$$k_m = \arg \max_{k=1,\ldots,K} \frac{R_k(m,f)}{T_k(f)}, \tag{2.1}$$

onde $T_k(f)$ denota a média a longo prazo da taxa de transferência do usuário k, computada no sub-quadro f, e $R_k(m,f) = log(1 + SNR_k(m,f))$ é a taxa de transferência alcançada pelo

usuário k, no SB m, e no sub-quadro f.

2.3.2 Visando a Capacidade de Atraso Restrito

Mesmo que o PFS introduza alguma justiça no sistema, esta forma de justiça pode não ser suficiente para aplicações com latência muito restrita. Para estes casos é necessária uma métrica de capacidade diferente. Um exemplo é a capacidade de atraso restrito.

Abordagem *Guaranteed Delay* (GD)

Esta abordagem também conhecida como capacidade de zero interrupção, é definida como a taxa de transmissão que pode ser assegurada em todos os estados de desvanecimento, sob restrições de energia. Em contraste com a capacidade de Shannon, onde as informações mútuas entre o transmissor e o receptor variam de acordo com o canal, as potências dos rádios nesta abordagem são coordenadas entre os usuários e os SBs com o objetivo de manter constante a informação mútua, independentemente dos estados de desvanecimento. Esta abordagem é relevante para classes de tráfego, onde uma determinada taxa de dados tem de ser garantida ao longo do tempo de conexão, independentemente das quedas de desvanecimento.

Estratégias que utilizam esta abordagem podem ser vistas, *e.g.*, nos trabalhos *Earliest Deadline First* (LIU; LEE, 2005) e *Largest Weighted Delay First* (RAMANAN; STOLYAR, 2001). Um cenário relevante para implantação desta abordagem seria uma comunicação de Voz sobre IP (VoIP), onde estão previstas densidades de várias centenas de usuários com uma restrição de latência, tipicamente exigindo que cada pacote seja entregue com êxito dentro de 50 ms. Na próxima seção, é apresentado o modelo de sistema de escalonamento das redes LTE, o qual, as diferentes estratégias de escalonamento devem considerar.

2.4 Modelo do Sistema

Nesta seção, é descrito o modelo de sistema LTE DL focando o processo de escalonamento. Este modelo de sistema ajuda a definir as restrições de um escalonamento DL. Em um sistema LTE DL, com uma antena e múltiplos usuários, tem-se um conjunto $k = 1, 2, ..., K$ de usuários, onde K representa o número total de usuários. O conjunto $n = 1, 2, ..., N$ representa o número de SBs, onde N é o número total de SBs dentro de um TTI. Mais do que isto, o conjunto de RE dentro de cada SB é representado por $S = 1, 2, ..., N_s$, onde N_s é o número total de RE em um SB e T_s representa a duração de cada RE, sendo $T_s > 0$. Assume-se também que Y é o número total de subportadoras e $Y_y^{(d)}(s)$ é um subconjunto y de subportadoras, as quais, podem ser usadas para carregar sinais de dados d na s^{th} (s-ésima) duração do símbolo OFDM. Este subconjunto considera somente símbolos disponíveis para transmitir dados, excluindo todos os símbolos usados para controle da tecnologia.

Em sistemas DL, os SBs são transmitidos usando um MCS. Os esquemas de modulação comuns em LTE são BPSK, QPSK, 16QAM, e 64QAM. Além do mais, cada MCS tem um *code rate* CR_j e um tamanho de constelação M_j associados, onde CR_j é o *code rate* associado com o MCS j; M_j é o tamanho da constelação MCS j; e $j = 1, 2, ..., J$ é o número total de MCSs suportados na transmissão. Considerando os MCSs comuns previamente mencionados, o valor de J seria 4. Logo, a taxa de bits alcançada por um único SB, considerando um MCS j, é dado pela equação a seguir (KWAN; LEUNG; ZHANG, 2008):

$$r^{(j)} = \frac{CR_j \log_2(M_j)}{T_s N_s} \sum_{s=1}^{N_s} Y_y^{(d)}(s). \tag{2.2}$$

Todo usuário k reporta, via canal de retorno, um vetor de CQIs $CQI_k \in [1, ..., CQI^{max}]$ (em LTE $CQI^{max} = 15$) contendo os valores de CQIs suportados para n SBs individuais (SCHWARZ; MEHLFUHRER; RUPP, 2010). A natureza exata do valor do CQI_k pode variar dependendo do método de retorno adotado (KWAN; LEUNG; ZHANG, 2008). É atribuído um MCS com taxa de bits zero para os SBs que não reportarem os valores de CQI. Os CQIs dentro do vetor correspondem à combinação de esquema de modulação e *code rate* definidos na norma do 3GPP (3GPP TS 36.213, 2008). Mais do que isto, quando um UE é servido em mais de um SB, é necessário encontrar um valor de CQI médio $\overline{CQI_k}$. Neste trabalho, visando simplificar o modelo, utilizou-se números inteiros para valores de CQI. Logo, a média entre os valores de CQIs dos UE é calculada, e, posteriormente, é realizado um arredondamento usando o piso deste valor (SCHWARZ; MEHLFUHRER; RUPP, 2010), como mostrado na fórmula que segue:

$$CQI_k = \lfloor \overline{CQI_k} \rfloor. \tag{2.3}$$

Considerando o modelo do sistema apresentado anteriormente, a taxa de transmissão alcançada por cada usuário pode ser modelada seguindo a equação:

$$r_k = \sum_{n=1}^{N} a_{k,n} \sum_{j=1}^{J} b_{k,j} r^{(j)}, \tag{2.4}$$

onde $a_{k,n}$ é igual a 1 se o usuário k é atribuído ao SB n e 0 caso contrário; e, $b_{k,j}$ é igual a 1 se o usuário k é atribuído ao MCS j e 0 caso contrário.

De acordo com as normas do LTE (3GPP TS 36.213, 2008) e (3GPP TS 36.211, 2008), o problema de escalonamento em LTE é sujeito as seguintes restrições:

$$\sum_{j=1}^{J} b_{k,j} = 1, \tag{2.5}$$

onde a equação (2.5) garante que todos os SBs pertencentes ao mesmo usuário dentro de um

único TTI devem usar o mesmo esquema de MCS:

$$\sum_{k=1}^{K} a_{k,n} = 1, \tag{2.6}$$

onde a equação (2.6) garante que um SB pode ser somente usado por um único UE, em um sistema *Single Input Single Output* (SISO).

Em LTE, existem diferentes classes de serviço, como visto na Tabela 2.4, e o atraso entre essas classes varia definindo as prioridades. Neste trabalho, múltiplos serviços são utilizados, e, portanto, o conjunto $p = 1, 2, ..., P$ é utilizado para representar estes serviços, onde P é o número total de serviços que podem ser usados. Desta forma, a matriz binária provida como entrada do sistema $c_{k,p} = c(k, p)$ define o serviço p associado com cada usuário k. Note que cada usuário pode somente ser associado com um serviço. O tempo de cada pacote do usuário dentro do sistema de escalonamento é denominado *Head Of Line* (HOL) e é calculado com as equações (2.7) e (2.8), descritas como segue:

$$HOL_k = t - t_0, \tag{2.7}$$

onde t é o tempo corrente (hora do sistema) e t_0 é o tempo que o pacote do usuário k, com serviço p, foi marcado quando entrou no sistema $(t > t_0)$. Por isso, o tempo restante para o escalonamento ou atraso (δ) para transmitir os pacotes é apresentado como uma função de HOL para cada usuário k, como:

$$\delta_k = \sigma_p - HOL_k. \tag{2.8}$$

Na equação (2.8), σ é o atraso do pacote para o tipo de serviço p, lembrando que $\sigma > 0$. Segundo as normas do 3GPP, cada tipo de serviço tem um tempo específico para expirar. Desta forma, os tempos de serviços descritos na norma do 3GPP (3GPP TS 23.203, 2010) são considerados.

2.5 Cálculo de Taxa de Transferência

Apesar da equação (2.2) descrever de forma geral o cálculo da taxa de transferência, existem diversas variáveis no sistema LTE que vão influenciar neste cálculo. Por isso, este cálculo é desdobrado de forma a mostrar estas variáveis e suas implicações. O *Code Rate* (CR) é a porção de dados úteis na transmissão e pode ser calculado usando a seguinte equação:

$$CR = \frac{TBS + CRC}{MRE \cdot BPS}, \tag{2.9}$$

onde as siglas representam:

- TBS (*Transport Block Size*) é o valor do tamanho do bloco e pode ser calculado através

da tabela 7.1.7.2.1-1 do *3GPP.36.213*;

- CRC (*Cyclic Redundancy Check*) é o número de bits adicionados para detecção de erros;
- MRE (*Maximum Resource Elements*) é o total de recursos disponíveis para alocação, *i.e.*, desconsiderando os SSs;
- BPS (*Bits Per Simbol*) representa o número de bits que cada símbolo carrega, variando de acordo com o esquema de codificação e modulação escolhido. Valores possíveis são 2,4 e 6 como mostrado na Tabela 2.3.

Segundo o *3GPP.36.213 7.1.7*, o resultado da equação (2.9) deve ser sempre menor que $0,931$. Neste trabalho, assim como em grande parte da literatura, o CR é considerado igual a 1, *i.e.*os limites teóricos da tecnologia. Outro parâmetro de influência na taxa de transferência é a largura de banda utilizada na transmissão. As larguras de banda disponíveis e a quantidade de recursos disponível em cada largura de banda são descritas na Tabela 2.2. Observando-se a tabela de larguras de banda, nota-se que a maior largura de banda no LTE é 20 MHz, possibilitando a utilização de até 100 SBs por *ms* na transmissão de recursos. Na equação (2.2), observa-se os parâmetros $T_s N_s$ que significam o tempo e o número de símbolos dentro de um SB, respectivamente. Como pode ser visto na Figura 2.2 o tempo de um RE é 71,42 μs, e dentro de um SB, com prefixo cíclico normal, são 14 REs. Por fim, a equação (2.2) apresenta uma multiplicação pelos símbolos disponíveis para carregar os dados. Exemplificando o cálculo da taxa de transferência de uma rede LTE, com largura de banda 1,4 MHz, e considerando o melhor *code rate* disponível, obtém-se:

$$r = \frac{1 \cdot \log_2(64)}{71,42 \cdot 14} \cdot 168 = 1,008 Mbits \tag{2.10}$$

Na equação (2.10), foi calculado uma taxa de transferência (*rate*) de $1,008 Mbits$ em um único SB. A largura de banda de 1.4 MHz corresponde a 6 SBs, logo, o *rate* máximo neste caso é igual a $6,048 Mbits$. Note que, caso MIMO estiver sendo considerado, então, esses valores devem ser multiplicados pelo número de antenas presentes.

2.6 Resumo do Capítulo

Neste capítulo, foram descritos os principais fundamentos da tecnologia LTE e do escalonamento de recursos de rádio DL. Além disso, modelou-se matematicamente o sistema de escalonamento de recursos LTE DL, considerando as normas da tecnologia descritas no padrão do 3GPP. A fim de facilitar o entendimento, foi exemplificado o cálculo da taxa de transferência DL, com suas peculiaridades. No próximo capítulo, alguns dos principais trabalhos sobre escalonamento de recursos em LTE serão analisados, assim como, são avaliadas as principais vantagens e desvantagens de cada estratégia.

3 TRABALHOS RELACIONADOS

Neste capítulo, são sumarizados os principais trabalhos da área de escalonamento de recursos em redes celulares. Primeiramente, são destacados os trabalhos que delimitam o escalonamento em redes sem fio que utilizam a tecnologia do OFDMA. Logo após, é realizada uma análise dos trabalhos que formulam matematicamente o problema de escalonamento em LTE. A seguir, é analisada à aplicabilidade de algumas heurísticas e meta-heurísticas disponíveis na literatura. Por fim, são descritas as lacunas existentes nos trabalhos relacionados ao escalonamento, as quais, este trabalho visa cobrir.

3.1 Escalonamento em Redes Sem Fio

O escalonamento em redes sem fio que utilizam OFDMA, é um problema NP-difícil (YANG et al., 2010). Na tentativa de mitigar a dificuldade do problema, algumas abordagens na literatura (POKHARIYAL et al., 2007) tratam o escalonamento em dois níveis. No primeiro, chamado *Time Domain Packet Scheduler* (TDPS), o escalonador seleciona determinados usuários entre aqueles conectados à estação base. E no segundo, chamado *Frequency Domain Packet Scheduler* (FDPS), o escalonador atribui RBs para os usuários escolhidos no nível anterior. A estratégia de divisão em níveis é adotada pelo fato de, a dependência circular entre os dois níveis ser quebrada. No entanto, a complexidade da solução de escalonamento não é reduzida sob as hipóteses padrões (COHEN; KATZIR, 2010).

No trabalho de Cohen *et al.*, os autores propõem uma heurística para resolver o problema de alocação em redes *Worldwide Interoperability for Microwave Access* (WiMAX). A ideia central desta heurística, responsável por acondicionar os pacotes (itens), dentro de um quadro (caixa), é descrita a seguir. Primeiramente, os itens são ordenados e divididos em três conjuntos, de acordo com seus respectivos tamanhos. Desta forma, os itens pequenos são colocados em um conjunto k_1, os itens médios em um conjunto k_2, e em k_3 ficam os itens grandes. Após este particionamento, o quadro também é dividido em proporções, onde então é criada uma partição pequena, uma média e uma grande. Em seguida, com o algoritmo *next fit*, os espaços são preenchidos, até que não se tenha mais espaço suficiente no quadro. Quando os quadros estiverem totalmente preenchidos, novos quadros são criados.

Esta abordagem é eficiente quando os pacotes variam muito de tamanho e existem pacotes dos três tipos definidos (pequenos, médios e grandes). Porém, se os pacotes seguirem uma distribuição normal e o desvio padrão desta distribuição for muito alto, o algoritmo tende a ser ineficiente. Além disso, nesta abordagem, os autores estavam considerando a tecnologia WiMAX, logo, não levaram em consideração as especificações impostas nas normas do LTE.

3.2 Escalonamento em Redes LTE

No trabalho de Assad *et al.*(ASSAAD; MOURAD, 2008), dois escalonadores foram propostos e comparados utilizando diferentes métricas de avaliação. A primeira destas propostas assumia a implementação do AMC e do escalonador separadamente. Já a segunda proposta considerou a união do AMC com o escalonador. Nestes dois escalonadores, o objetivo comum foi equilibrar a relação entre a justiça no escalonamento e a capacidade do sistema. Entretanto, o atraso dos usuários não foi considerado como uma métrica de escalonamento e, além disso, a perda de pacotes desta solução não foi demonstrada. Portanto, torna-se menos provável a utilização destes escalonadores nos atuais cenários das redes sem fio.

Kwan *et al.*(KWAN; LEUNG; ZHANG, 2008), propuseram uma abordagem de escalonamento ótima e sub-ótima considerando múltiplos usuários, com o objetivo de maximizar a taxa de transmissão dos usuários. Neste estudo, os autores formulam matematicamente o problema de escalonamento de recursos nas redes LTE e também descrevem as restrições impostas pela tecnologia. Na abordagem ótima, os autores mostram que a otimização efetuada é a melhor opção em relação a uma estratégia gulosa. E, através de um escalonador sub-ótimo de complexidade reduzida, em relação a abordagem ótima, os autores utilizaram a ideia de escalonamento por estágios. Sendo que, no primeiro estágio, cada SB é atribuído ao usuário que pode suportar a maior taxa de bits e, no segundo estágio, é determinado o melhor MCS para cada usuário. Desta forma, são atribuídos subconjuntos disjuntos de SBs para os usuários. Consequentemente, o problema de otimização com múltiplos usuários é transformado em diversos sub-problemas com usuários unitários, resolvidos de forma paralela.

Os autores demonstram, através de resultados numéricos, que o desempenho do sistema melhora com o aumento da correlação entre as subportadoras OFDMA. Além disso, demonstram que, com limitadas quantidades de informação de *feedback* dos CQIs, é possível atingir bom desempenho no escalonamento. Os autores afirmam ainda que a estratégia sub-ótima é atrativa, quando o número de usuários é grande. No entanto, no escalonamento desenvolvido neste trabalho, os autores levam em conta somente o CQI dos usuários, tornando assim, o escalonamento injusto com os usuários com condições de canal ruins.

Em outro estudo, os autores (KWAN; LEUNG; ZHANG, 2009) modelam o escalonador incluindo justiça no escalonamento. Isso é feito usando o histórico da taxa de transmissão do escalonamento passado. A ideia por trás desta estratégia é de armazenar um histórico da taxa de transferência atingida por um usuário nos últimos escalonamentos. Então, na próxima iteração para decisão de qual usuário será escolhido, a métrica das últimas taxas de transferências atingidas serão relacionadas com a atual taxa de transferência do usuário, e esta relação será atenuada. Um exemplo da relação realizada pode ser apresentado como $\frac{r_i}{\psi_i(t)}$, onde r_i é a taxa de transferência possível do usuário i e $\psi_i(t)$ significa o histórico de taxas de transferência do usuário i no tempo t. Os resultados desta abordagem mostram que o escalonador apresentado

introduz maior justiça no escalonamento, com uma modesta perda de desempenho em relação as taxas de transferências alcançadas, desde que, os SINRs médios dos usuários sejam bastante uniformes.

Embora esta estratégia seja muito interessante, existem muitas limitações que podem ser citadas, como por exemplo, o escalonador ser justo dada uma distribuição de SINRs, porém, se esta distribuição não for uniforme a taxa de transferência dos usuários é fortemente afetada. Outro ponto a ser considerado nesta estratégia é que o tempo dos serviços dos usuários não é considerado, levando a uma grande perda de pacotes dos dados de usuários com serviços prioritários, trafegando na rede.

Outras abordagens encontradas na literatura relacionam o problema de escalonamento em redes sem fio como problema *Multiple Choice Knapsack Problem* (MCKP), que é uma generalização do problema da mochila, tradicional, onde o conjunto de itens é particionado em classes. Porém, a escolha binária de selecionar um item é substituída pela seleção de exatamente um item de cada classe de itens. O MCKP é um dos modelos mais flexíveis dentre os problemas de *knapsack*. Em uma possível formulação para este problema, pode-se considerar m como uma classe de itens $N_1, ..., N_m$ para serem empacotados em uma mochila de capacidade c. Cada item $j \in N_i$ tem um benefício $p_{i,j}$ e um peso $w_{i,j}$. O problema é escolher um item de cada classe, que maximize a soma dos benefícios sem que a soma dos pesos exceda a capacidade c.

Fei *et al.*, em (LIU et al., 2012), os autores consideram que o problema de escalonamento em redes LTE pode ser caracterizado como um MCKP. Neste trabalho, os autores desenvolveram uma relação entre o problema de escalonamento de múltiplos serviços em redes sem fio e o problema da mochila. Os problemas são claramente similares, entretanto, diversas premissas devem ser consideradas caso esta caracterização seja realizada. Uma destas premissas seria a de que pelo menos um pacote de cada tipo deveria ser escolhido a cada momento do escalonamento, algo que não é uma regra no LTE, e portanto, não poderia ser uma premissa. Outra peculiaridade que deveria ser levada em consideração é a expiração dos itens de dentro das mochilas, pois não fica claro se os autores consideram que pacotes expiram no sistema de escalonamento. Portanto, assumindo tais hipóteses, o algoritmo apresenta-se distante da realidade das redes sem fio.

Alguns autores da literatura LTE (CAPOZZI et al., 2013) relatam que uma questão encontrada no estudo da literatura LTE sobre escalonamento é a falta de um cenário de referência comum, que possa ser usado para comparar diferentes soluções. Estes autores relatam que isso é de fundamental importância para uma comparação efetiva das soluções inovadoras com as já existentes. Por esta razão, no trabalho, os autores tentam identificar um cenário de referência para comparar o desempenho de algumas soluções disponíveis na literatura. Na simulação proposta, os autores utilizaram o esquema de escalonamento por justiça como uma estratégia de referência. Graças a esta abordagem, eles foram capazes de demonstrar também a necessidade real de estratégias sofisticadas para enfrentar o problema do provisionamento de QoS.

A literatura sobre escalonamento DL utiliza técnicas tais como *branch-and-bound* (FAN et

al., 2012) escalonamento semi-persistente (GROSS; PARRUCA, 2013), e linearização (KWAN; LEUNG; ZHANG, 2008) para tratar o problema de escalonamento. Entretanto, a otimalidade global para grandes instâncias não pode ser garantida. O escalonamento DL em OFDMA é provado ser NP-difícil, e, por isso não é sensato que se tenha um algoritmo ótimo de tempo polinomial para resolver todas as instâncias do problema (YANG et al., 2010). Como resultado, uma abordagem prática seria projetar algoritmos heurísticos eficientes para enfrentar o problema de escalonamento. Em seguida, são descritas algumas heurísticas e meta-heurísticas de escalonamento, encontradas na literatura LTE.

3.3 Heurísticas e Meta-heurísticas no Escalonamento

Para a resolução de problemas complexos, tal como, escalonamento de recursos, onde o tempo para se encontrar uma solução ótima é não-linear, as alternativas mais naturais são as heurísticas e as meta-heurísticas. No escalonamento LTE, algumas heurísticas são a escolha do melhor CQI e a escolha do pacote mais antigo na fila de transmissão, algumas meta-heurísticas são os Algoritmos Genéticos (AG) e *Simulated Annealing* (SA). Para se designar uma meta heurística que melhor se enquadre ao problema, uma série de características deste problema devem ser levadas em consideração. A principal delas é a relação entre a precisão da solução encontrada com o tempo gasto para encontrar esta solução. Especificamente para o escalonamento de recursos em redes sem fio, apesar da quantidade de dados transmitidos ser fundamental, o atraso dos pacotes deve ocasionar uma grande perda de dados. Portanto, meta-heurísticas que convergem para uma solução viável de forma mais rápida, são mais indicadas, mesmo que a solução esteja distante do valor ótimo.

Neste sentido, os pesquisadores (AYDIN; KWAN; WU, 2012) investigam abordagens heurísticas para o escalonamento em redes LTE, considerando a crescente demanda de usuários nestas redes. Neste estudo, os autores verificam que um escalonamento com abordagens exatas não é o caminho a ser seguido, pois a complexidade do problema torna a utilização inviável. A saída encontrada foi a utilização de métodos heurísticos. Devido as características do problema, as meta-heurísticas, AG e SA, foram escolhidas pelos autores e serão descritas no discorrer deste capítulo. Algumas meta-heurísticas opcionais que poderiam ser empregadas neste problema são, busca tabu, *ant colony* e grasp. Como um resultado das pesquisas deste estudo, os autores compararam os resultados sub-ótimos, ótimos e heurísticos de todas suas pesquisas. Deste modo, conseguiram perceber que as estratégias de escalonamento baseadas em AG e SA aproximam-se de forma satisfatória dos resultados ótimos, com muito menos esforço computacional, viabilizando assim a sua utilização.

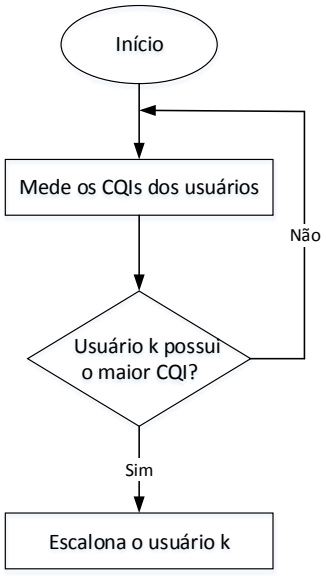
Dentre as duas meta-heurísticas utilizadas, SA obteve melhores resultados na taxa de transferência, atingindo praticamente os mesmos resultados da abordagem ótima, considerando-se o desvio padrão. No entanto, a meta heurística AG obteve melhores resultados refentes a escalabilidade, pois tem uma complexidade menor que o algoritmo baseado em SA. As meta-heurísticas

SA e AG aplicadas para o problema de escalonamento possuem complexidade polinomial e linear, respectivamente. Apesar dos autores evoluírem na escalabilidade da solução de escalonamento, os requisitos de QoS considerando diferentes atrasos não são levados em consideração nestes estudos. A seguir são apresentadas algumas abordagens de escalonamento, heurísticas e meta-heurísticas.

3.3.1 Escalonamento pelo Melhor CQI

Esta heurística utiliza a condição do canal para atribuir os blocos de recursos para os usuários. Para realizar o escalonamento, a estação base envia um sinal de referência para os equipamentos interessados na transmissão, para medir a qualidade do canal até este equipamento. O valor retornado é conhecido como CQI, e quanto maior for este valor, melhores as condições do canal. Esta abordagem pode aumentar a capacidade da célula aos custos de sacrificar a justiça na transmissão. A estratégia de escolher o melhor CQI é um tipo de escalonamento MSRS. A Figura 3.1 apresenta um fluxograma de exemplo do escalonamento por melhor CQI.

Figura 3.1: Fluxograma da abordagem de escalonamento pelo melhor CQI.



Fonte: pelo autor (2016).

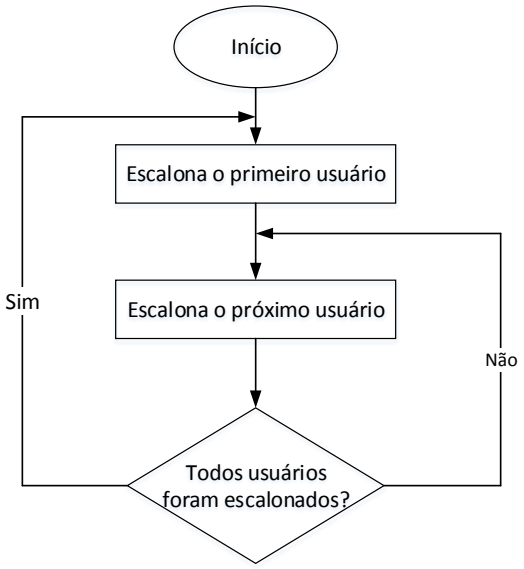
Os usuários mais próximos da estação base idealmente têm melhores condições de canal e consequentemente serão privilegiados. Esta abordagem em cenários de baixa densidade, pode ser uma alternativa sensata. Entretanto, em cenários de alta densidade os usuários provavelmente vão enfrentar uma baixa qualidade na transmissão, tendo em vista que os serviços prioritários possuem baixas taxas de perdas de pacotes. Consequentemente, se um usuário com um serviço prioritário e com um CQI alto perder pacotes na transmissão, no próximo ciclo de escalonamento os dados da retransmissão podem ser privilegiados em relação aos pacotes dos

outros usuários da rede.

3.3.2 Escalonamento *Round Robin*

Nesta estratégia de escalonamento, os usuários são atribuídos aos recursos compartilhados, um após o outro. Assim, cada usuário é igualmente escalonado, sem levar em conta qualquer requisito de QoS. A principal vantagem do escalonamento *Round Robin* é a garantia de equidade para todos os usuários. Além disso, *Round Robin* é uma estratégia de fácil implementação, razão pela qual é geralmente utilizada. Muitos autores baseiam-se nesta estratégia de escalonamento, modificando-a para obter um desempenho mais satisfatório. Na Figura 3.2 é apresentado um fluxograma de exemplo de uma estratégia *Round Robin*.

Figura 3.2: Fluxograma da abordagem de escalonamento *Round Robin*.



Fonte: pelo autor (2016).

Pelo fato de não levar a informação da qualidade de canal em conta, esta abordagem tende a resultar em baixa taxa de transferência de dados dos usuários. Mais do que isto, o atraso do serviço dos usuários não é levado em conta, então, é difícil priorizar qualquer tipo de tráfego utilizando esta estratégia.

3.3.3 Algoritmos Genéticos

O Algoritmo Genético (AG) é uma meta-heurística de busca de soluções aproximadas em problemas de otimização. Mais do que isto, algoritmos genéticos são uma classe particular de algoritmos evolutivos que usam técnicas inspiradas da biologia evolutiva como mutação, recombinação, hereditariedade e seleção natural. A ideia principal é, a partir de uma população de

soluções, realizar a união destas soluções a fim de gerar novos indivíduos nesta população. Em seguida, usando algum critério de desempenho, selecionar as novas soluções que são mais úteis para participar de uma nova geração. Após, continuar com esta evolução enquanto existirem soluções melhores. Caso não exista uma solução melhor em um determinado período, eleger a atual como melhor solução.

Alguns autores (AYDIN; KWAN; WU, 2012), implementam uma abordagem clássica dos algoritmos genéticos a fim de resolver o problema de escalonamento, entretanto, algoritmos genéticos, tem mostrado-se ineficientes para problemas combinatórios de larga escala, considerando tanto o tempo, quanto a qualidade das soluções (JING; LIM; ONG, 2003). Mais do que isso, o AG apresenta complexidade linear $\mathcal{O}(n)$ para este problema.

3.3.4 *Simulated Annealing*

Simulated Annealing (SA) é uma meta-heurística para aproximação global em um grande espaço de busca de soluções. Esta abordagem simula um processo da física que aquece um material a uma temperatura bem alta e resfria aos poucos, dando tempo para o material alcançar seu estado de equilíbrio. Geralmente, esta técnica é utilizada em situações onde encontrar uma solução aceitável para o problema é mais importante do que encontrar a melhor solução global.

Em Kwan *et al.*(AYDIN; KWAN; WU, 2012), os autores mostram que esta meta heurística é melhor que os algoritmos genéticos para o problema de escalonamento. Porém, assim como os AGs, o SA depende de uma randomização para melhorar os resultados da solução, o que geralmente não é o caso em redes sem fio, onde os recursos e os usuários variam pouco em um curto espaço de tempo. Além disso, a complexidade deste algoritmo para o problema de escalonamento é polinomial $\mathcal{O}(n^k)$.

3.4 Resumo do Capítulo

Neste capítulo, foram apresentados diversos trabalhos na área de escalonamento de recursos em redes móveis. Com isso, pôde-se verificar que este tópico foi amplamente estudado nas redes de 4G, indicando fortemente que esta área será extensivamente verificada nas novas redes de rádios. Apesar de diversas estratégias terem sido estudadas, não foi encontrada nenhuma que considere, no escalonamento, as métricas de atraso e de justiça dos diferentes serviços existentes nas redes móveis. Além do mais, não foi identificada alguma estratégia que lide com o escalonamento em redes sem fio de alta densidade. Contudo, alguns trabalhos destacam a importância destas métricas e a necessidade do desenvolvimento de estratégias para manusear o tráfego em redes de alta densidade (PARRUCA; GROSS, 2015), (SANCHEZ et al., 2015). Através dos tipos de escalonamento vistos no capítulo anterior e as estratégias de escalonamento descritas até então, na Tabela 4.1 são sumarizados os requisitos necessários para as redes de próxima geração e quais requisitos são encontrados nas estratégias estudadas. Na tabela, a letra

Tabela 3.1: Tabela de comparação de requisitos das estratégias.

	Atraso	CQI	Justiça	Perda de Pacotes	Escalável
MSRS ou Melhor CQI		X			
DG - Delay Guaranteed	X		X	X	X
Round Robin			X		
AG - Algoritmos Genéticos		X		X	
SA - Simulated Annealing		X		X	X

X representa que o requisito é considerado pela estratégia e o retângulo em branco significa o caso contrário.

Alguns autores também relatam dificuldades relacionadas a avaliação das estratégias propostas, pois não existem cenários predefinidos para avaliar as estratégias de um modo geral. Nesta dissertação, na intenção de não se ter este problema, avaliou-se a métrica proposta para os dois principais cenários existentes, um de baixa densidade e outro de alta densidade.

Finalmente, dados os trabalhos encontrados na literatura, percebe-se que as estratégias de escalonamento existentes nas atuais redes 4G, apesar da quantidade, não estão preparadas para trabalhar com cenários de alta densidade de usuários. Além do mais, requisitos adicionais foram introduzidos pelas redes sem fio de próxima geração, tais como, qualidade de serviço através do atraso e maiores níveis de justiça no escalonamento. Observando que as estratégias atuais não se adequam às necessidades atuais da próxima geração de redes móveis, é apresentada, neste trabalho, uma abordagem de escalonamento de recursos que pode ser utilizada tanto em cenários de redes de 4G quanto para a próxima geração de redes de rádio, sendo idealmente desenvolvida para cenários sobrecarregados.

4 ESCALONAMENTO OFDMA DL

Neste capítulo é descrita a arquitetura de escalonamento OFDMA DL, indicando onde esta dissertação está situada. Em seguida, o escalonamento é abordado em três etapas: (*i*) primeiramente, é proposta uma modelagem do problema de escalonamento OFDMA DL em LTE, assim como, são descritas as restrições deste modelo; (*ii*) O modelo proposto é linearizado para tornar o escalonamento solucionável através de programação linear inteira, alcançando, desta forma, os valores ótimos em algumas instâncias do problema; (*iii*) Por fim, é demonstrada uma abordagem heurística, especialista em cenários sobrecarregados, que considera o QoS através de uma nova métrica parametrizada com valores de CQIs e tempos de serviços.

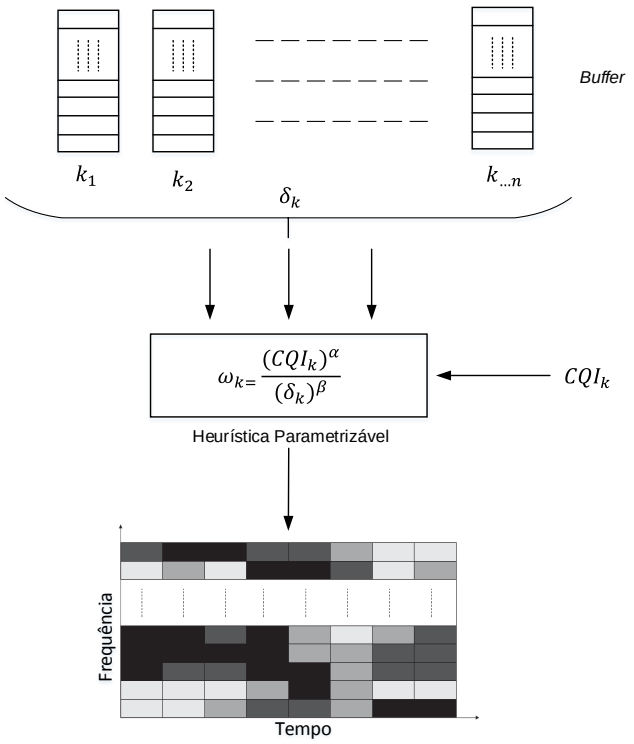
4.1 Arquitetura do Escalonador

Para o melhor entendimento do escalonamento em redes OFDMA DL, é apresentada uma arquitetura que demonstra como a métrica parametrizável, apresentada a seguir, pode ser utilizada para auxiliar o escalonamento em redes sem fio. Na Figura 4.1, os usuários são representados por k. Percebe-se, através da representação, que cada usuário possui um *buffer* de pacotes que estão aguardando para ser transmitidos. Cada um destes pacotes possui um serviço, consequentemente, um tempo limite para ser escalonado. Caso este tempo limite seja excedido, *i.e.*, o pacote fique dentro do sistema um tempo maior do que o seu tempo para expirar, o pacote é descartado. E, através do módulo de retransmissão, é feita uma nova tentativa de envio do pacote para o usuário. Ademais, na Figura 4.1, uma matriz bidimensional representa o escalonamento dos dados dos usuários e cada tom de cinza dentro do quadro representa uma região ocupada por um usuário diferente.

O escalonador deve, idealmente, evitar retransmissões, pois além de causar lentidão para o usuário final, pode gerar uma carga adicional de pacotes para serem retransmitidos. Na Figura 4.1, o tempo para o pacote expirar é representado por δ_k. Esta informação é provida como entrada do escalonador e pode ser calculada através da equação (2.8). Outra entrada do sistema esta representada por CQI_k, que significa o CQI do usuário k, calculado na equação (2.3). A métrica de escalonamento apresentada nesta dissertação esta representada através do retângulo descrito como heurística parametrizável. Note que os parâmetros α e β são ajustáveis e precisam ser atribuídos para definir seus pesos. Para definição destes valores, neste momento, é necessário escolher os valores que melhor adequam-se as necessidades da rede. Desta forma, é possível calcular o valor de ω para todos os pacotes dos k usuários e ordená-los decidindo a prioridade do escalonamento. Os detalhes da métrica, assim como, a ideia por trás da métrica, são apresentados nas próximas seções.

A estratégia de escalonamento através da heurística parametrizável possui flexibilidade para os valores dos parâmetros. Desta maneira, observa-se que, se o valor 0 for atribuído para α e β, o resultado da equação será 1, e portanto, o algoritmo vai ter o mesmo comportamento de uma

Figura 4.1: Arquitetura de escalonamento com a função parametrizada.



Fonte: pelo autor (2016).

estratégia *Round Robin*. Já, se o valor 1 for atribuído para α e o parâmetro β ficar em 0, então o escalonador reproduzirá o comportamento da estratégia de melhor CQI. Entretanto, se o valor 0 for atribuído para o parâmetro α e o valor 1 para β , então, simula-se o comportamento da abordagem DG. Na próxima seção, o problema de escalonamento apresentado na Figura 4.1 é formalizado matematicamente.

4.2 Formulação do Problema

O objetivo principal do escalonador modelado nesta dissertação é maximizar a taxa de transmissão geral dos usuários na rede. Além disso, garantir o escalonamento adequado, respeitando as restrições do modelo de sistema de escalonamento demonstrado no Capítulo 2. O escalonamento de recursos para múltiplos usuários pode ser formalizado como um problema de otimização combinatória, como segue:

$$\max \sum_{k=1}^{K} \sum_{n=1}^{N} \sum_{p=1}^{P} a_{k,n} c_{k,p} \sum_{j=1}^{J} b_{k,j} r^{(j)}, \tag{4.1}$$

sujeito às restrições mostradas nas equações (2.5) e (2.6).

O objetivo na equação (4.1) é maximizar a taxa de transmissão - calculada na equação (2.2) -

dos usuários com serviços prioritários, onde, a variável de decisão binária $a_{k,n}$ será maximizada para alcançar o maior número n de SBs atribuídos para o usuário k. Similarmente, utilizando a variável $c_{k,p}$, a quantidade de usuários k atribuídos para o serviço p será maximizada. Já o $\sum_{j=1}^{J} b_{k,j}$ visa maximizar a escolha do MCS j para o usuário k. Esta formalização do problema não leva em consideração o ajuste da potência de transmissão de rádio, em tempo de escalonamento, ou seja, é assumido que a potência total do sistema é igualmente compartilhada entre os usuários. Como pode-se observar em (CHUNG; GOLDSMITH, 2001), a degradação da taxa de bits resultante desta suposição é pequena quando o *Adaptative Modulation and Coding* (AMC) é usado, como no caso do LTE.

A otimização desenvolvida na equação (4.1) baseia-se nas otimizações apresentadas por Kwan *et al.*, em (AYDIN; KWAN; WU, 2012), e por Wang *et al.*, em (AI et al., 2010). A ideia da união das duas modelagens de escalonamento disponíveis na literatura é adicionar o provisionamento de qualidade de serviço através do atraso inerente a cada serviço. Observe que a variável binária $c_{k,p}$ é inserida no modelo no intuito de maximizar a quantidade de serviços prioritários atendidos, consequentemente, atendendo às necessidades dos usuários das novas redes móveis. Ademais, esta variável permite que, em cenários sobrecarregados, o escalonamento não seja injusto ao escalonar somente usuários com grandes quantidades de CQI.

Observando a equação (4.1) nota-se que os MCSs (j), SBs (n), usuários (k) e serviços (p) são atribuídos todos juntos. Portanto, este modelo ótimo para solução de tal problema pode proporcionar um ganho sobre uma simples abordagem gulosa. Entretanto, a formulação do problema torna-se não-linear, devido à multiplicação entre variáveis binárias. Em geral, problemas de otimização não lineares não podem ser resolvidos de forma ótima em tempo polinomial. Portanto, através da formulação do problema apresentada na equação (4.1), só é possível resolver problemas de escalonamento em cenários de redes LTE com um número limitado de usuários e recursos.

Um cenário de uma rede LTE é representado na abordagem matemática, como uma instância. Uma instância é um conjunto de informações que contém, dentre outros, os seguintes dados, o número de usuários do sistema, a quantidade de bits dos usuários, a quantidade de recursos do sistema e qualidade do canal de cada usuário. As instâncias utilizadas neste trabalho estão disponíveis em <<https://github.com/mcadori/lte_DL_simulator>>. Além disso, no próximo capítulo, é descrito o simulador de escalonamento utilizado nesta dissertação. Através deste simulador, é possível criar instâncias do problema de escalonamento, para posterior resolução com o modelo matemático. Deste modo, na próxima seção, uma linearização do problema é apresentada, visando resolver o problema de escalonamento em cenários de redes LTE complexos.

4.3 Escalonador Linearizado

A linearização é uma técnica conhecida na área da otimização combinatória que consiste da busca de soluções para algumas instâncias de problemas conhecidamente complexos. Este processo torna-se mais importante conforme o aumento da dificuldade de resolução do problema. No contexto do escalonamento de recursos em redes sem fio, é inserida maior complexidade no modelo, devido à inserção do requisito de atraso. Percebe-se, na análise desenvolvida através dos trabalhos relacionados, que o problema de escalonamento OFDMA DL é um problema NP-difícil, então, a menos que P=NP, não pode-se garantir que a decisão de escalonamento possa ser realizada em tempo polinomial, para todas as instâncias. Para tentar mitigar esta dificuldade, nesta seção, o problema é transformado em um problema linear equivalente, através da introdução de uma variável auxiliar.

A linearização apresentada neste trabalho é o resultado de algumas operações matemáticas sobre a equação (4.1). Estas operações são realizadas a fim de tornar este modelo interpretável por um *solver* de programação linear inteira. Através deste procedimento, é possível determinar os resultados ótimos para certas instâncias. É importante destacar que, ao realizar este procedimento de linearização, o espaço de soluções aumenta, e torna a quantidade de restrições criadas para o problema muito maiores. Logo, a utilização da linearização é uma alternativa interessante, apenas para critérios de avaliação.

Baseado na ideia da linearização do escalonamento, obtém-se três abstrações que seguem: (*a*) tornar o modelo de escalonamento exato e solucionável; (*b*) fornecer um *baseline* para comparação de qual a distância das outras abordagens, comparando-se com os valores ótimos; (*c*) conceber um novo modelo de escalonamento linearizado em OFDMA DL, considerando características da próxima geração de redes móveis. Finalmente, considerando a equação (4.1), uma linearização possível é:

$$\max \sum_{k=1}^{K} \sum_{n=1}^{N} \sum_{p=1}^{P} \sum_{j=1}^{J} v_{k,n,p,j} r^{(j)}, \tag{4.2}$$

sujeito às restrições (2.5), (2.6), e:

$$v_{k,n,p,j} \leq b_{k,j}; \tag{4.3}$$

$$v_{k,n,p,j} \leq a_{k,n} \cdot R; \tag{4.4}$$

$$v_{k,n,p,j} \leq (b_{k,j} - (1 - a_{k,n}) \cdot R); \tag{4.5}$$

$$v_{k,n,p,j} \leq (b_{k,j} - (2 - a_{k,n} - c_{k,p}) \cdot R), \tag{4.6}$$

onde, nas equações (4.4), (4.5), e (4.6), o valor R é um valor real grande. O valor de R é usado,

pois, uma abordagem de relaxação conhecida como *big-M* foi utilizada, como em (DESROSIERS; LÜBBECKE, 2005).

A variável $v_{k,n,p,j}$ é o produto das variáveis binárias $a_{k,n}b_{k,j}c_{k,p}$. Mais do que isto, as equações (4.4), (4.5) e (4.6) foram utilizadas para garantir o mesmo comportamento entre $v_{k,n,p,j}$ e $a_{k,n}$, $b_{k,j}$ e $c_{k,p}$. Uma variável auxiliar precisou ser inserida para viabilizar a linearização do problema. Consequentemente, o espaço de soluções possíveis também aumentou. A técnica de inserir uma nova variável ao problema, para linearização, pode ser vista com mais detalhes em (VANDERBEI, 1996).

Após linearizar o problema de acordo com a modelagem apresentada acima, é possível executar algumas instâncias do problema e obter alguns resultados de escalonamento. Mesmo que, ainda não seja possível resolver todas as instâncias do problema em um tempo aceitável, dada a complexidade para a resolução. Instâncias menores, *i.e.*, com poucos usuários, SBs e CQIs, são resolvidas em um tempo aceitável pelo modelo exato linearizado. Todavia, à medida em que a rede começa a aumentar o número de usuários, a resolução do escalonamento poderia levar horas, dias, meses ou até mesmo anos.

Os resultados das instâncias, obtidos através da execução da linearização descrita, serviram como um *baseline* para comparação dos resultados ótimos com as resultados heurísticos, obtidos através da abordagem heurística parametrizável, apresentada na próxima seção. Embora tenham-se auferidos os resultados para algumas instâncias do problema, a garantia de otimalidade para todas as instâncias não é possível de ser alcançada em tempo polinomial. Logo, uma alternativa heurística deve ser desenvolvida para o escalonamento em redes móveis. Portanto, na próxima seção, é proposta uma nova heurística de escalonamento OFDMA DL, que utiliza parâmetros para adaptação ao tipo de tráfego da rede.

4.4 Escalonador Heurístico Parametrizável

Através da abordagem linearizada apresentada anteriormente é possível alcançar os resultados de escalonamento para algumas instâncias do problema. No entanto, o espaço da solução aumenta rapidamente ao utilizar-se esta técnica. Observando essa questão e prevendo a importância de um algoritmo com parâmetros de QoS ajustáveis para as futuras redes móveis, propôs-se uma função peso (ω), denominada de heurística parametrizável, com o objetivo de maximizar $r^{(j)}$ ((2.2)), tendo em conta o atraso e o CQI dos usuários. Esta função prioriza os dados de usuários com os mais altos valores de CQI e também, os pacotes de usuários que estão mais próximos de expirar. Esta métrica foi idealizada intuitivamente e ajustada de forma experimental. Intuitiva, no sentido em que a melhor alternativa para considerar o atraso dos serviços dos usuários e o CQI era o estabelecimento de uma relação entre estas métricas. E experimental, pois os parâmetros da equação foram avaliados em diversos cenários. Explicitamente, pode-se

escrever a função ponderada da seguinte forma:

$$\omega_k = \frac{(CQI_k)^\alpha}{(\delta_k)^\beta}, \tag{4.7}$$

onde, $(\delta_k)^\beta \neq 0$, e α e β são parâmetros que medem a importância do CQI e do atraso, respectivamente. É importante observar que a variável ω depende dos parâmetros α e β, que indicam qual deverá ser a relevância do CQI e do atraso, para cada pacote. Isto é, quanto maior o valor relativo de α, maior é a importância dada ao CQI do pacote, e quanto maior o valor relativo de β, maior é a importância dada ao atraso do pacote. Desta forma, deseja-se evitar que os pacotes atinjam o tempo limite (*deadline*), *e.g.*, aumenta-se o valor relativo do parâmetro β, dando assim, mais importância ao atraso do pacote. De modo análogo, evita-se que alguns usuários entrem no conhecido modo *starvation state*, por falta de recursos, devido a uma momentânea má condição do canal.

Ademais, o escalonador calcula a função peso para cada conjunto de dados e ordena-os por este resultado. Por conseguinte, o conjunto de dados com o maior valor na função peso é escalonado primeiro. Para determinar os parâmetros α e β, que correspondem ao peso correto para cada conjunto de dados dos usuários - e maximizam r -, procede-se com simulações computacionais utilizando o algoritmo descrito na sequência.

O Algoritmo 1 descreve como a métrica proposta foi utilizada no escalonamento. O algoritmo inicia com um conjunto de pacotes ϕ, que corresponde ao conjunto de dados dos usuários que estão esperando escalonamento. A condição na linha (2) garante que, enquanto existirem pacotes de usuários a serem enviados, TTIs continuarão sendo criados para abrigar estes pacotes. Na linha (5), utilizando-se o argumento $technologyLimit$, previne-se que o número limite de SBs por TTI seja excedido. Este argumento pode ser facilmente ajustado, pois depende da tecnologia e da frequência a ser utilizada. Desta maneira, o primeiro conjunto de dados de usuários presente no conjunto ϕ é escalonado. E, finalmente, o TTI é retornado devidamente preenchido. O procedimento acima é repetido até que todos os pacotes tenham sido escalonados.

Note que a abordagem parametrizada foi idealizada para suportar quantos serviços forem necessários. Assim, para utilização em outros cenários, basta ajustar os parâmetros na equação ω (linha 1) para obter a melhor o desempenho de acordo com as necessidades da rede. Além disso, no capítulo de resultados, é apresentada uma avaliação completa dos valores de ω, variando as incógnitas α e β através do Algoritmo 1, para facilitar a identificação de quais parâmetros privilegiam quais ambientes. Tendo em vista a aplicação da estratégia heurística parametrizável, nas futuras redes móveis, é demonstrada uma tabela de comparação de algumas estratégias de escalonamento, auxiliando na decisão de qual abordagem utilizar, dentre as existentes, para melhor preencher as necessidades da rede.

Na Tabela 4.1, cada linha representa uma das estratégias de escalonamento apresentadas até então, nesta dissertação. Já as colunas representam os requisitos inerentes às redes móveis.

Algoritmo 1 Abordagem Parametrizada.

Entradas: Inicia com um conjunto de dados de pacotes dos usuários: ϕ
Entradas: Limite de SBs por TTI: $technologyLimit$

```
1: Calcula ω (φ,α,β)
2: while φ>0 do
3:     OrderBy ω (φ)
4:     CreateNewTTI()
5:     while SBs ≤ technologyLimit do ScheduleTheFirst(φ)
6:     end while
7:
8: end while
9:
```

Saídas: TTI preenchido

Cada requisito apresentado foi extraído da literatura, por representar uma funcionalidade fundamental em pelo menos uma das principais redes móveis. Para tanto, as cinco características selecionadas foram: atraso - significa o tempo que cada pacote tem para expirar (demonstrado na equação (2.8)); CQI - representa a qualidade do canal do usuário (calculado na equação (2.3)); justiça - expressa a justiça do escalonamento dos usuários; perda de pacotes - refere-se à perda de pacotes em cenários sobrecarregados; escalável - representa se a estratégia contempla escalabilidade de usuários ou não; flexível - considera se uma estratégia possui flexibilidade no escalonamento ou não. Por fim, os retângulos marcados com a letra X representam se determinada estratégia de escalonamento, considera o requisito correspondente, caso não considere, o retângulo é deixado em branco.

Tabela 4.1: Tabela de comparação de estratégias.

	Atraso	CQI	Justiça	Perda de Pacotes	Escalável	Flexível
MSRS ou Melhor CQI		X				
DG - *Delay Guaranteed*	X		X	X	X	
Round Robin			X			X
AG - Algoritmos Genéticos		X		X		X
SA - *Simulated Annealing*		X		X	X	X
Heurística Parametrizada	X	X	X	X	X	X

Fonte: pelo autor (2016).

Desta maneira, verifica-se na Tabela 4.1 que a estratégia de escalonamento proposta oferece algumas vantagens, se comparada com as outras abordagens citadas. Embora a abordagem que escalona através do melhor CQI seja uma das opções mais difundidas na literatura, por proporcionar as melhores taxas de transferência, esta abordagem não cobre algumas das mais importantes características das futuras redes móveis, *e.g.*, atraso, justiça e perda de pacotes. Para preencher estes espaços, a estratégia DG toma o atraso em conta. Entretanto, no caso de uma rede com muitos usuários com serviços prioritários e qualidades de canal ruins, o es-

calonador pode baixar a taxa de transferência global da rede. No escalonamento através da estratégia *Round Robin*, devido a imprevisibilidade do tráfego gerado pelos usuários, observa-se que a justiça é inserida no sistema. No entanto, os tempos de serviços e a qualidade do canal dos usuários é desconsiderada. As meta heurísticas SA e AGs são importantes, pois podem ser adaptadas para o problema de escalonamento e, são extremamente flexíveis. Porém, estas heurísticas são voltadas para encontrar uma solução rapidamente, e não para encontrar uma solução de qualidade, considerando diferentes serviços. Na próxima seção, é apresentada uma discussão, com intuito de auxiliar a escolha dos valores da métrica parametrizável.

4.5 Discussão da Métrica de Escalonamento Parametrizável

A escolha dos parâmetros de escalonamento depende da necessidade específica da rede. Por exemplo, caso a rede esteja servindo um ambiente com alta densidade de usuários, o escalonamento, idealmente, deveria considerar o atraso dos serviços e a justiça no escalonamento. Portanto, na métrica parametrizável ω o parâmetro β deve ter um peso maior que o parâmetro α. Entretanto, se o ambiente em que a rede estiver sendo utilizada, possuir baixa densidade de usuários e o objetivo for maximizar a taxa de transferência, o parâmetro α deve possuir um peso maior. Mais do que isso, baseando-se nas simulações realizadas nesta dissertação, pode-se analisar o comportamento da taxa de transferência, da justiça e da perda de pacotes no escalonamento, quando a métrica ω é utilizada. Ademais, diferentes cenários são avaliados, facilitando a escolha dos parâmetros mais adequados. As análises mencionadas podem ser visualizadas no Capítulo 6.

Os parâmetros α e β na equação (4.7) podem ser atribuídos para qualquer valor. Entretanto, dadas às características da equação, os valores mais adequados são entre 0 e 2. Embora, atualmente, os parâmetros de escalonamento dos rádios sejam configurados de forma manual. Os valores de α e β, foram idealizados, para serem definidos de forma dinâmica de acordo com o comportamento da rede. Esta abordagem foi adotada devido à possibilidade da utilização desta métrica, juntamente, com os novos conceitos de rádio que vêm surgindo na literatura, tais como, *Software-defined networking* (SDN) e *Software-defined radio* (SDR).

Um ambiente de simulação é apresentado no próximo capítulo, para auxiliar na tomada de decisão de quais parâmetros satisfazem melhor as necessidades da rede. Através deste ambiente de simulação, é possível implementar diferentes cenários de escalonamento e verificar o comportamento da métrica parametrizável com a combinação de tantos valores quanto necessário.

4.6 Resumo do Capítulo

Neste capítulo, foi apresentado um conjunto de escalonadores OFDMA DL, que consideram requisitos das futuras redes móveis, e que são especialistas em cenários sobrecarregados. Primeiramente, uma arquitetura demonstra e explica a forma que a métrica proposta deve ser

empregada. Complementarmente, modelou-se o problema de escalonamento em redes móveis. Em seguida, o modelo exato foi linearizado para se atingir um *baseline*. Posteriormente, uma heurística parametrizável foi proposta, juntamente com um algoritmo, para utilização da mesma. Além do mais, foi provida uma tabela comparativa entre as principais estratégias de escalonamento da literatura. Por fim, foi apresentada uma discussão da métrica idealizada nesta proposta. No próximo capítulo, é apresentado o ambiente de simulação que auxiliou na avaliação das estratégias implementadas neste trabalho.

5 AMBIENTE DE VALIDAÇÃO DO ESCALONAMENTO OFDMA DL

Neste capítulo, é descrita a implementação do modelo linearizado. Ademais, é apresentado o ambiente de simulação que foi utilizado para validar o algoritmo de escalonamento heurístico proposto para as redes móveis, que utiliza, o OFDMA como tecnologia para acesso múltiplo. Este ambiente de simulação possui algumas funcionalidades básicas de uma rede LTE, focando no escalonamento de recursos.

5.1 Descrição de Implementação do Modelo Linearizado

A ferramenta de software livre chamada GNU MathProg foi utilizada para escrever a modelagem matemática explicada no capítulo anterior. Tal ferramenta trata-se de um ambiente para desenvolvimento de modelos, descrevendo equações matemáticas lineares na linguagem *A Mathematical Programming Language* (AMPL), padrão da ferramenta. Essa linguagem consiste em um conjunto de comandos e estruturas de dados em blocos, construídos pelo usuário, para posterior execução de um *solver*. O GNU MathProg faz parte do GLPK *solver*, sendo necessária a instalação do pacote GLPK e do pacote glpk-utils no ambiente. Apesar de o GLPK possuir seu próprio *solver*, nos testes efetuados neste trabalho ele não foi utilizado. Ao invés dele, utilizou-se o *solver* CPLEX da IBM (versão studio124 x86 64 bits), devido à rapidez e ao tempo de execução para revolver o modelo. A seguir, é demonstrado o principal trecho de código da modelagem desenvolvida.

Figura 5.1: Principal trecho de código da linearização.

```
 maximize Value: sum{i in U, j in SB, e in M, q in S} f[i,j,e,q] *
(log(mcs[i])/log(2)/999.88) * simbols[i];

s.t.  r1{j in SB}: sum{i in U} user_sb[i,j] <= 1;
s.t.  r2{i in U}: sum{e in M} user_mcs[i,e] = 1;
s.t.  r3{i in U}: sum{j in S} user_service[i,j] = 1;

s.t.  r4{i in U, j in SB, e in M, q in S}: f[i,j,e,q] <= user_mcs[i,e];
s.t.  r5{i in U, j in SB, e in M, q in S}: f[i,j,e,q] <= user_sb[i,j]*1000;
s.t.  r6{i in U, j in SB, e in M, q in S}: f[i,j,e,q] >= user_mcs[i,e] -
(1-user_sb[i,j])*1000;
s.t.  r7{i in U, j in SB, e in M, q in S}: f[i,j,e,q] >= user_mcs[i,e] -
(2-user_sb[i,j]-user_service[i,q])*1000;
```

Fonte: pelo autor (2016).

Após o desenvolvimento do modelo acima, é necessário convertê-lo para que seja legível no CPLEX. Este procedimento deve ser realizado sempre, antes de uma execução do modelo no CPLEX. Para automatizar este procedimento, foi escrito, em *shell script*, um programa que: (*a*) efetua a conversão dos modelos desenvolvidos no GLPK em modelos CPLEX *i.e.*, conversão de *.mod* para *.lp*; (*b*) realiza a otimização através do CPLEX; (*c*) filtra e exibe os resultados da

função objetivo.

5.2 Motivação do Desenvolvimento de um Ambiente de Simulação

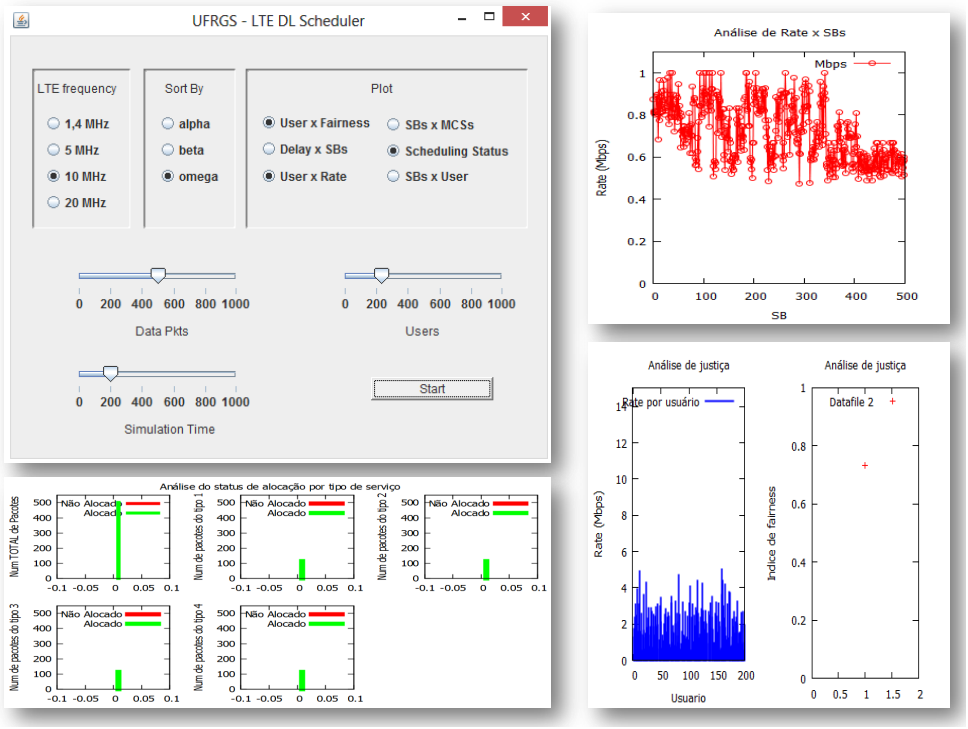
A construção de um ambiente de simulação próprio foi motivada pela escassez de aplicações semelhantes e *open source*, na literatura. Mais do que isso, as poucas ferramentas de simulação disponíveis possuem pouca, ou nenhuma, documentação. Tendo em vista estas dificuldades, desenvolveu-se um ambiente de simulação para avaliação das estratégias de escalonamento propostas. O código fonte do simulador, está livremente disponível em <<https://github.com/mcadori/lte_DL_simulator>>. Ademais, o executável do simulador pode ser acessado no seguinte endereço: <<https://github.com/mcadori/lte_DL_simulator/blob/master/run/scheduler.jar?raw=true>>.

5.3 Principais Funcionalidades

O ambiente de simulação para o escalonamento OFDMA DL em redes móveis foi desenvolvido com uma interface gráfica para facilitar a visualização dos experimentos. Através desta interface, é possível realizar todas as operações envolvidas no processo de escalonamento. Além disso, esta interface serviu de base para a realização dos experimentos efetuados nesta dissertação, possibilitando assim, a análise do comportamento dos gráficos resultantes. A interface do usuário, assim como a demonstração dos gráficos, foi idealizada para ser organizada, fácil e rápida. Observa-se na Figura 5.2, que esta forma de acesso ao ambiente de simulação apresenta uma interface de acesso a escalonamentos mais simples, embora, tenha sido projetado também o acesso à simulações mais complexas. Para isso, concentrou-se a maioria dos parâmetros do simulador em um único arquivo de configuração. Neste arquivo, pode-se definir uma série de informações que especificam qual o algoritmo de escalonamento será utilizado. Além disto, o ambiente de simulação foi integrado a uma biblioteca de gráficos (*gnuplot*), tornando possível gerar gráficos do escalonamento de maneira intuitiva. Na Figura 5.2, pode-se visualizar a interface do simulador assim como alguns gráficos que são gerados após o escalonamento.

De maneira geral, o ambiente de simulação desenvolvido possui um pacote de geração de tráfego (*generateTraffic*), que proporciona a possibilidade de geração de dados de usuários móveis para escalonamento, assim como, efetua a distribuição deste tráfego para os usuários. No auxílio da geração deste tráfego, são disponibilizadas algumas opções de distribuições matemáticas que são utilizadas para geração de cada um dos parâmetros do tráfego dos usuários, *e.g.*, o CQI, os pacotes e os tipos de serviços do tráfego. Em seguida, descrevem-se, de forma detalhada, algumas das distribuições presentes no simulador. Além disso, o simulador possui a opção de interpretar os arquivos de entrada, o que permite que o simulador leia dados de fontes externas para realizar as simulações. Isso é fundamental para a análise de um mesmo conjunto de dados, em diferentes cenários ou parâmetros. Sumarizando as principais funcionalidades do

Figura 5.2: Interface de usuário do simulador.



Fonte: pelo autor (2016).

ambiente de simulação, obtém-se a seguinte lista:

- Geração de tráfego;
- Geração de recursos;
- Escalonamento através de diferentes abordagens;
- Geração de gráficos de escalonamentos através do *gnuplot*;
- Geração de mapas de dados para avaliação de métricas.

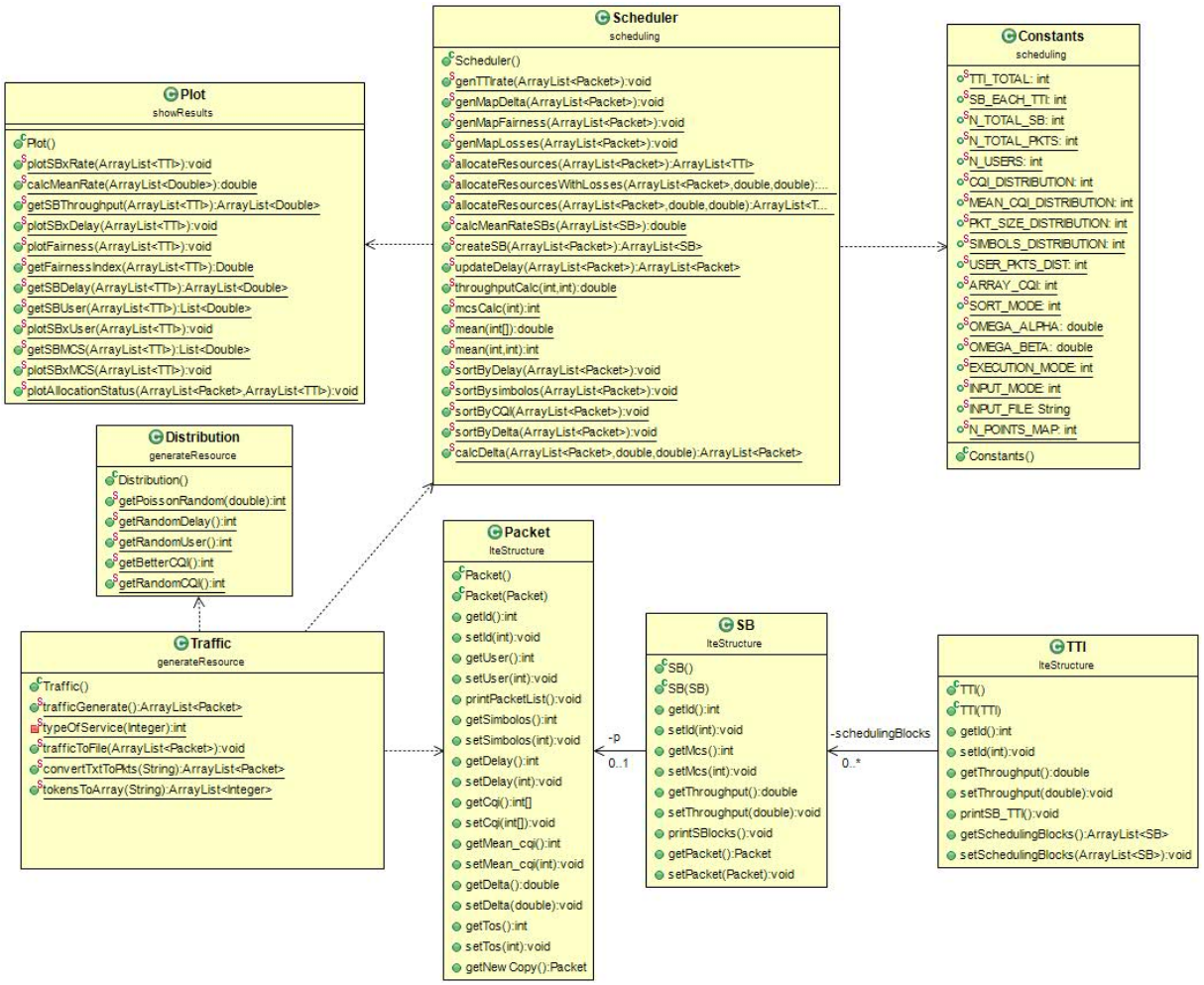
O ambiente de simulação desenvolvido é composto por diversos módulos, o detalhamento de todos os módulos não é apresentado nesta dissertação por motivos de espaço, porém, o diagrama de classes apresentado na Figura 5.3, auxilia no entendimento mais específico do ambiente:

Analisando o diagrama de classes do ambiente é possível perceber as ligações entre as estruturas de dados utilizadas para realização dos experimentos desta dissertação. Na próximas subseções são descritos os principais módulos do ambiente simulação desenvolvido.

5.3.1 Geração de Tráfego

Tendo em vista que os dados dos usuários são os principais componentes no escalonamento em LTE, foi desenvolvido um módulo de geração de tráfego. Este tráfego é composto por um conjunto variado de bits. Este tráfego é mapeado, no ambiente de simulação, para pacotes. Estes pacotes possuem, dentre outras, as seguintes informações: quantidade de símbolos, atraso

Figura 5.3: Diagrama de classes do ambiente de simulação.



Fonte: pelo autor (2016).

e qualidade do canal (CQI) de cada conjunto de dados de usuário, o tipo do serviço daquele tráfego (*Type of Service* - TOS), e, de qual usuário o tráfego foi gerado. O tamanho do pacote é definido pela quantidade de símbolos OFDM. Já o atraso, é definido de acordo com o tipo do serviço. Por fim, os campos usuário e CQI representam, respectivamente, as informações do usuário que gerou o tráfego, juntamente com o CQI do usuário no momento da chegada deste pacote. A seguir, são listadas as distribuições utilizadas na geração dos dados:

- Os símbolos seguem uma distribuição de Poisson com a média centrada em 145, sendo que, em torno de 145 símbolos são disponíveis para a transmissão de dados, desconsiderando os símbolos SSs (utilizados para sinalização da tecnologia). O valor de 145 é um valor médio aproximado. Os valores exatos de símbolos disponíveis em cada SB não são utilizados, pois a especificação destes valores tornaria o simulador menos abrangente, dificultando a adaptação para outras funcionalidades, como por exemplo, o escalonamento UL. Os símbolos para sinalização e para transmissão de dados podem ser observados em detalhes no Apêndice A.

- O tráfego de dados dos usuários foi distribuído para os usuários de forma aleatória, seguindo uma distribuição uniforme com valores entre 1 e o número total de usuários.
- Os valores de CQI foram gerados aleatoriamente, seguindo uma distribuição uniforme com quantias entre 1 e 15.
- O tipo dos serviços foi gerado através de uma distribuição normal padrão (Gaussiana).

5.3.2 Geração de Recursos

Para geração dos recursos LTE DL, foram desenvolvidas, basicamente, três estruturas de dados diferentes para alocar o tráfego gerado pelos usuários. Estas estruturas são TTI, SB e *Simulation*. Onde, um TTI corresponde a um conjunto de SBs e uma simulação (*Simulation*) corresponde a um conjunto de TTIs e SBs, a cada TTI o *throughput* é calculado. O tempo total de um escalonamento, depende do número de TTIs gerados. A estrutura que desempenha o principal papel nesta hierarquia é o SB, que contém, entre outras, as seguintes informações: qual pacote de dados está alocado no SB, qual o *throughput* dos dados alocados neste SB e qual o MCS está sendo utilizado no SB. Os pacotes gerados são alocados em SBs e definem o *throughput* do SB, calculado segundo a equação (2.2). O MCS atribuído a cada SB é definido através do CQI do usuário que gerou o pacote que estará utilizando o SB, no momento do escalonamento. Os valores de CQI e os correspondentes MCSs são definidos pelo 3GPP da seguinte forma, caso o CQI do usuário esteja entre 1 e 6 a modulação QPSK é utilizada, caso esteja entre 7 e 9 a modulação 16QAM é utilizada e caso esteja entre 10 e 15 a modulação 64QAM é utilizada.

Note que o CQI dos usuários é calculado através da equação (2.3), e, mais do que isso, o MCS definido anteriormente deve seguir a restrição descrita na equação (2.5), garantindo que, se mais de um SB for atribuído a um usuário no mesmo TTI, o mesmo MCS deve ser utilizado para estes SBs. Já a restrição descrita na equação (2.6) assegura que um SB seja utilizado por apenas um usuário, esta restrição também é respeitada no simulador.

5.3.3 Escalonamento Através de Diferentes Abordagens

Na literatura LTE, encontram-se diferentes abordagens de escalonamento, como já descrito nos capítulos anteriores. Desta forma, desenvolveu-se um módulo específico para o escalonamento, onde é possível, de forma intuitiva, alternar os algoritmos já implementados, assim como, facilitar a implementação de novas abordagens. A classe de escalonamento (*scheduler*) é utilizada para o manuseio dos diferentes escalonadores. As três abordagens implementadas por padrão neste simulador são: escalonamento por atraso (abordagem utilizada em (LIU; LEE, 2005)), escalonamento por CQI (abordagem utilizada em (KWAN; LEUNG; ZHANG, 2008)) e o escalonamento por ω, que é a métrica proposta nesta dissertação.

5.3.4 Geração de Gráficos Através da API do *gnuplot*

Para facilitar a comparação dos resultados no escalonamento, utilizou-se o pacote *gnuplot*, disponível em <<http://javaplot.panayotis.com/>>. Deste modo, foi possível gerar os resultados gráficos em formato vetorial proporcionando melhor visualização para posterior análise. É importante salientar que a ferramenta *gnuplot*, deve estar previamente instalada no ambiente, para a plotagem dos gráficos do simulador. Ademais, todos os gráficos são acessíveis pela classe de plotagem (*Plot*) e recebem, como parâmetro, uma simulação, *i.e.*, se a necessidade for, construir um gráfico para análise da justiça no escalonamento, é necessário enviar o escalonamento efetuado para o método de plotagem, e o mesmo construirá o gráfico desejado, *e.g.*, (Plot.plotFairness(simulation);). Os gráficos atualmente implementados no simulador são descritos nas subseções seguintes.

5.3.5 Geração de Mapas de Dados

Visando uma melhor análise das métricas fundamentais do escalonamento OFDMA DL é possível gerar, através do ambiente de simulação, gráficos que demonstram mapas de comportamento das variáveis. Estes gráficos mostram o resultado da variação dos parâmetros do sistema de escalonamento. A variação dos pontos é feita, por exemplo da seguinte forma, um parâmetro x é dividido igualmente em 600 pontos entre 0,0 e 2,0, da mesma forma, y é dividido nos mesmos 600 pontos. Em seguida, cada valor de x é avaliado contra todos os valores de y calculando um valor z, logo, cada mapa possuí cerca de 360.000 valores de z. Posteriormente, os pontos de z são plotados com uma escala de tons de cinza em um plano cartesiano, onde os eixos representam valores de x e y.

Em virtude da grande quantidade de dados para geração destes mapas, somente os dados necessários para produzi-los podem ser gerados através deste simulador. Desta forma, após o término da simulação, o arquivo *.dat*, oriundo do simulador, deve ser fornecido diretamente ao *gnuplot* para geração dos mapas. Note que, uma execução para gerar um mapa de dados pode levar várias horas, dependendo do número de pontos do mapa e da capacidade do computador, onde o programa estiver executando.

5.4 Detalhes do Desenvolvimento

Nesta seção, o desenvolvimento do ambiente de simulação é detalhado. Os detalhes do ambiente são apresentados para auxiliar as pesquisas em redes sem fio. O ambiente de simulação proposto foi desenvolvido na linguagem de programação Java, com auxílio das bibliotecas externas *forms* e *gnuplot*.

5.4.1 Entradas do Ambiente de Simulação

Basicamente são suportados dois tipos de entrada de dados no ambiente, um dinamicamente gerado e o outro inserido através de um arquivo de fonte de dados fixo. Embora seja possível uma utilização mesclada entre estas duas formas de entrada. Por exemplo, pode-se gerar os dados dinamicamente em uma simulação e escolher a opção do simulador para armazenar estes dados gerados em uma fonte de dados fixa. Desta forma, é possível a utilização do arquivo de fonte de dados nas simulações posteriores. Estas características são importantes para avaliação das estratégias de forma fidedigna, utilizando o mesmo conjunto de dados. A seguir, pode-se conferir um exemplo de como os dados podem ser formatados para interpretação no simulador.

Figura 5.4: Exemplo dos dados de entrada do simulador.

```
data;
param k :=30000;
param n :=12000;
param m :=3;
param p :=1;

param simbols := 1 132 ...
30000 123 ;
param services := 1 0.16 ...
30000 0.50 ;
param user := 1 908 ...
30000 633 ;
param mcs := 1 16 ...
30000 32 ;
param cqi := 1 5 ...
30000 8 ;
end;
```

Fonte: pelo autor (2016).

O ambiente de simulação reconhece o padrão referido e utiliza estes dados fixos nas simulações. Note que, caso os dados sejam gerados pelo próprio programa, de forma fixa, este também é o formato de saída da fonte de dados. É importante destacar que os dados gerados no simulador podem ser diretamente usados como dados de entrada (uma instância) para um modelo exato construído em *A Mathematical Programming Language* (AMPL), consequentemente, o modelo pode ser resolvido através de um solver, *e.g.*, CPLEX.

5.4.2 Saídas do Ambiente de Simulação

Como um resultado das simulações efetuadas, é possível obter mais de 15 gráficos analisando diferentes perspectivas. Embora seja possível gerar os dados para a construção de todos estes gráficos, alguns deles necessitam ser gerados externamente, pois demandam recursos gráficos não implementados na API utilizada para conexão com o *gnuplot*. Através da visão amigável do ambiente de simulação, são possíveis serem visualizados os seguintes gráficos:

- Análise da justiça dos usuários;
- Análise da distribuição dos SBs para os usuários;
- Análise do atraso de cada pacote;
- Análise da taxa de transferência do usuário para cada SB;
- Análise da taxa de transferência média na transmissão;
- Análise da distribuição dos SBs para os MCSs;
- Análise dos status gerais de alocação, *i.e.*, quantos foram ou não atendidos.

Além dos gráficos listados acima, através da interface avançada do ambiente de simulação, pode-se gerar, pelo menos, os seguintes gráficos:

- Mapa das taxas de transferências da heurística parametrizável;
- Mapa de justiça da heurística parametrizável;
- Mapa de perdas de pacotes da heurística parametrizável;
- Análise da taxa de transferência de cada TTI.

5.4.3 Execução do ambiente

Existem duas formas para a execução do ambiente de simulação, a forma padrão, é através de uma interface gráfica construída na classe *UI*. E, a outra forma de execução, denominada avançada, é realizada através da classe *Scheduling*, sendo esta forma, específica para usuários que já estejam familiarizados com o ambiente de simulação.

Para a execução do escalonamento avançado é necessário que sejam definidos os parâmetros de simulação na classe *Constants*. Nesta classe, é possível definir as informações sobre o tempo de simulação, a quantidade de recursos, usuários e pacotes. Além da definição do tipo de escalonamento a ser realizado. Mais do que isto, nesta classe é possível definir o modo de execução do escalonamento avançado, os modos de execução avançados são: execução de apenas uma simulação, a geração de dados para construção de um gráfico das taxas de transferência por TTIs, a geração do mapa de cores para análise da taxa de transferência variando os parâmetros α e β, a geração do mapa de cores para análise da justiça variando os parâmetros de α e β, a geração do mapa de cores para análise da perda de pacotes variando os parâmetros de α e β.

5.5 Resumo do Capítulo

Neste capítulo, o processo de implementação do escalonamento linearizado foi descrito. Além disso, o ambiente de simulação utilizado para avaliação da métrica heurística parametrizável foi apresentado. Por fim, o ambiente de simulação, aqui descrito, foi utilizado para realização dos experimentos apresentados no próximo capítulo, que apresenta os resultados desta dissertação.

6 AVALIAÇÃO DE DESEMPENHO DO ESCALONAMENTO OFDMA DL

Neste capítulo, são demonstrados dois estudos de caso de escalonamento DL, um com baixa densidade e outro de alta densidade. Em primeiro lugar, é descrito o ambiente simulado em ambas as avaliações. Logo em seguida, os resultados obtidos são avaliados e comparados com o *baseline*, calculado anteriormente, e com outras estratégias. Por fim, apresenta-se uma discussão dos resultados obtidos.

6.1 Ambiente de Simulação

O ambiente de simulação LTE foi construído usando o quadro *downlink*, em modo FDD, com uma periodicidade de 10 ms e com prefixo cíclico normal. Os ajustes de potência dos rádios nas simulações, são considerados constantes. Desta forma, as perdas por desvanecimento ou por interferência não são consideradas. Esta estratégia é conhecida na literatura como, *maximum-rate constant-power scheduling* (SESIA; TOUFIK; BAKER, 2011). Já, as sinalizações de controle do quadro são consideradas conforme descrito, no Capítulo 2. Além disto, em todos os cenários considerados neste trabalho, usou-se os serviços descritos pelos padrões da 3GPP (3GPP TS 23.203, 2010). Assim, o atraso calculado é baseado nos valores apresentados nesta tabela.

A avaliação da abordagem linearizada foi realizada analiticamente e através de simulações computacionais. A fim de avaliar o impacto da métrica parametrizada em relação ao *baseline*. O modelo formalizado anteriormente foi implementado e executado no CPLEX Optimization Studio, versão 12.3. Todos os experimentos foram realizados em uma máquina com quatro processadores AMD Opteron com 6276 núcleos e 64 GB de RAM, usando o Sistema Operacional Ubuntu GNU/Linux versão servidor 11.10 x86 64.

No intuito de demonstrar os resultados da avaliação heurística parametrizada, as métricas de taxa de transferência, justiça e perda de pacotes, são avaliadas. Para cada uma destas métricas, um gráfico mostrando um mapa de comportamento foi construído. Estes gráficos mostram o resultado da variação dos parâmetros α e β de 0,0 até 2,0. A variação dos pontos foi feita da seguinte forma: o parâmetro α foi dividido igualmente em 600 pontos entre 0,0 e 2,0, da mesma forma, β foi dividido nos mesmos 600 pontos. Em seguida, cada valor de α foi avaliado em relação a todos os valores de β, logo, cada mapa possuí ≈ 360.000 valores de ω. Posteriormente, os pontos de ω foram plotados com uma escala de cores em um plano cartesiano, onde os eixos representam valores de α e β. Os limites dos parâmetros α e β, foram escolhidos por reproduzirem o comportamento global do sistema.

Em cada iteração o valor de ω é recalculado e o escalonamento é realizado novamente, considerando o novo valor de ω. Em seguida, as métricas são analisadas novamente e os valores plotados no gráfico. Note que a taxa de bits medida nas avaliações é normalizado *i.e.*, $rate/rate_max$. Além disto, em cada um dos cenários, é inclusa uma tabela com quatro pontos

amostrais coletados. Estes pontos são utilizados para a explicação do comportamento do escalonamento. Esta metodologia foi aplicada para taxa de transmissão, justiça e perda de pacotes.

6.2 Estudo de Caso

No processo de escalonamento OFDMA DL, muitos requisitos precisam ser considerados. Um desses requisitos é o momento para transmitir os dados dos usuários, levando em conta os diferentes tipos de serviços envolvidos. Por exemplo, o atraso médio na transmissão de serviços de vídeo (streaming) é de 100 ms, já o atraso tolerado pelo serviço Web (HTML) é de 300 ms segundo os padrões 3GPP (3GPP TS 36.213, 2008).

Nota-se que o atraso no serviço de transmissão de vídeo é menor que o atraso do serviço de HTTP (isto acontece porque, idealmente vídeos necessitam fluir continuamente, a fim de proporcionar uma melhor qualidade da experiência aos utilizadores). Se o pacote não é entregue no tempo máximo tolerado de atraso, então o pacote é descartado. Para evitar a perda de pacotes em excesso, o algoritmo de escalonamento deve ter ciência do atraso máximo permitido para cada serviço. Além disso, dependendo do CQI entre a estação base e o usuário, a mesma informação deve ser enviada mais do que uma vez. Esta redundância no envio de informações gera sobrecarga e, idealmente, deve ser evitada.

Nas análises apresentadas a seguir os parâmetros α e β, representam respectivamente as métricas de CQI e atraso, quando estes parâmetros assumem valores diferentes significa que um novo valor de ω (que define a prioridade do escalonamento) é obtido, conforme a equação (4.7).

6.2.1 Cenário com Baixa Densidade

No processo de escalonamento DL, a quantidade de dados de usuários é gerida pelo escalonador, que é responsável por atender tantas requisições quanto possível, em um cenário de baixa densidade, *i.e.*, onde os recursos de rádio disponíveis são abundantes para atender as demandas dos usuários, métricas mais simples de escalonamento podem ser utilizadas. Estas métricas demonstram o mesmo efeito de estratégias mais complexas e ainda consumindo menos recursos de processamento. Com a heurística parametrizável proposta neste trabalho, é possível reproduzir o comportamento de outras estratégias da literatura, apenas definindo os parâmetros α e β na equação (4.7). Para demonstrar o comportamento da métrica parametrizável em um cenário de baixa densidade, utilizou-se o seguinte cenário:

6.2.1.1 Análises para Valores de ω

O cenário de baixa densidade, idealmente, não possui perdas de pacotes, pois, existem recursos suficientes para enviar todos os pacotes. Portanto, pacotes são descartados somente se o

Tabela 6.1: Tabela de parâmetros.

Usuários	25
Pacotes	1200
TTIs	200
Largura do Canal	3 MHz
SBs por TTI	15
Total de SBs	3000
α	De 0 até 2
β	De 0 até 2
Serviços	VoIP 50 ms HTTP 300 ms Vídeo (*Buffered*) 150 ms Vídeo (*Streaming*) 100 ms

Fonte: pelo autor (2016).

atraso tolerado pelo serviço for excedido. A Tabela 6.2 apresenta quatro pontos que descrevem o comportamento global do sistema. Na Tabela 6.2 os valores de perda de pacotes são nulos, confirmando a expectativa que todos os pacotes fossem alocados, antes de expirarem. Além disso, os valores de taxa de transferência e de justiça apresentam uma variação, praticamente, irrelevante. Este comportamento, era esperado, pois os recursos são suficientes para alocar todos os usuários. Logo, o tipo de escalonamento, não terá grande impacto nos resultados, devido ao fato que todos os pacotes podem ser alocados.

Tabela 6.2: Pontos de análise - cenário de baixa densidade.

	α	β	$\bar{r}$	f	Perda de Pacotes
P1	0,1	0,1	0,7173	0,9859	0,00
P2	0,1	1,0	0,7167	0,9859	0,00
P3	1,0	1,0	0,7173	0,9855	0,00
P4	1,0	0,1	0,7169	0,9857	0,00

Fonte: pelo autor (2016).

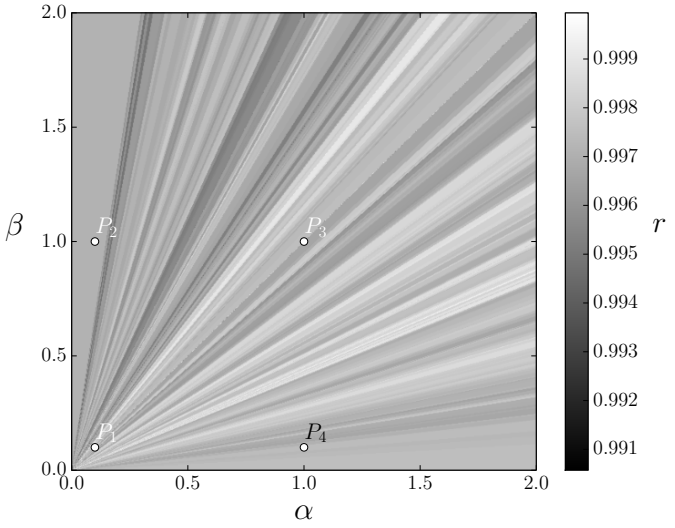
O comportamento das métricas avaliadas na tabela, podem ser observados nos gráficos das Figuras 6.1 e 6.2. Note que, estes gráficos seguem a mesma metodologia apresentada anteriormente.

Análise da Taxa de Transferência

No gráfico apresentado na Figura 6.1, é realizada uma análise, sobre a taxa de transferência em redes sem fio em um cenário de baixa densidade. Neste cenário, 25 usuários estão gerando tráfego de 1200 pacotes, que pertencem a quatro diferentes serviços. O serviço que possui a maior tolerância para expirar na rede é o serviço de HTTP, que expira com 300 ms. Embora

o tempo de simulação seja de 200 ms, em cada 1 ms existem 15 SBs, o que totaliza 3000 SBs para transmissão, viabilizando o escalonamento de todos os pacotes.

Figura 6.1: Análise da taxa de transferência.



Fonte: pelo autor (2016).

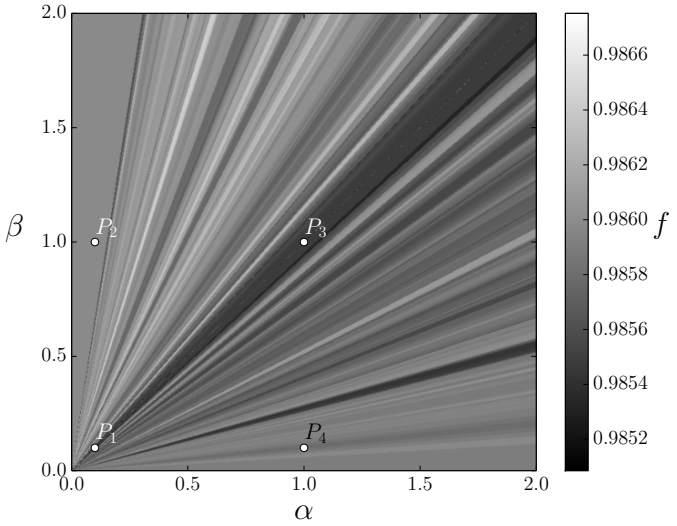
Desta forma, o gráfico demonstra uma baixa variabilidade de cores e uma pequena espessura das linhas, sendo que a maior concentração dos valores para a taxa de bits está entre 0,999 e 0,991. A alta taxa de bits demonstra uma eficiente utilização do canal. A seguir, é demonstrado o comportamento da justiça no escalonamento em redes de baixa densidade.

Análise de Justiça

A Figura 6.2 representa uma análise do índice de justiça no escalonamento, em cenários de baixa densidade. Pode-se observar um comportamento similar ao gráfico apresentado na Figura 6.2. Isto ocorre devido a pouca disputa de recursos existente neste cenário, percebe-se que o índice de justiça se aproxima de um, que representa na escala de Jain, o mais justo possível. Nota-se também, que o gráfico possui características de baixa variabilidade, com valores entre 0,9852 e 0,9866, indicando que independente dos parâmetros atribuídos para a função de peso, os valores tendem a não mudarem significativamente.

A análise de perda de pacotes não foi representada em um mapa neste cenário, pois é uma constate e resultaria em um mapa de cor única. Sumarizando as informações referentes à avaliação da heurística parametrizada nestes dois cenários, é notável a importância desta estratégia especialmente em cenários sobrecarregados, onde a busca por recursos, é fator decisivo para o aumento da eficiência espectral.

Figura 6.2: Análise do índice de justiça.



Fonte: pelo autor (2016).

6.2.2 Cenário com Alta Densidade

Para analisar um caso de estudo com alta densidade de usuários, *i.e.*, sobrecarregado, foram definidos 1000 diferentes usuários que geraram 30000 pacotes uniformemente distribuídos. O tamanho destes pacotes segue uma distribuição de Poisson e precisam ser entregues em 120 TTIs *i.e.*, 120 ms. Mais do que isto, um canal com largura de banda de 20 MHz, o que corresponde a 100 SBs por TTI foi utilizado. Ademais, o CQI entre os usuários foi uniformemente ajustado. Para facilitar, sumarizamos os dados na Tabela 6.3:

Tabela 6.3: Tabela de Parâmetros.

Usuários	1000
Pacotes	30000
TTIs	120
Largura do Canal	20MHz
SBs por TTI	100
Total de SBs	12000
α	De 0 até 2
β	De 0 até 2
Serviços	VoIP 50 ms HTTP 300 ms Vídeo (*Buffered*) 150 ms Vídeo (*Streaming*) 100 ms

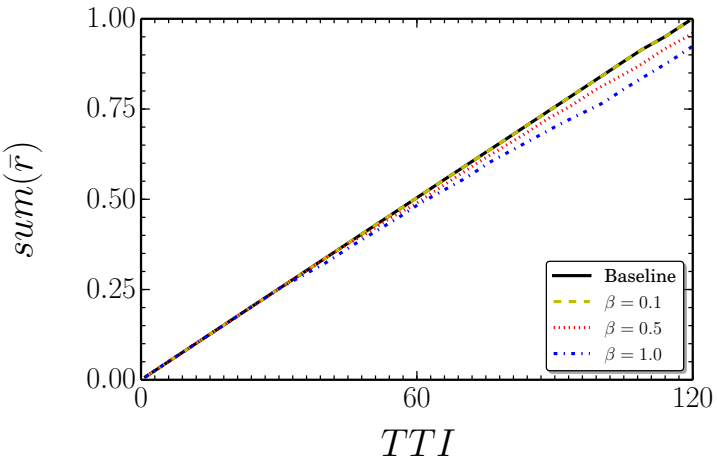
Fonte: pelo autor (2016).

6.2.2.1 Análise da Taxa de Transferência Comparada ao Baseline

A heurística parametrizada proposta nesta dissertação e o *baseline* foram analisados utilizando o cenário sobrecarregado descrito anteriormente. A Figura 6.3 apresenta a soma normalizada $\bar{r}$ dos pacotes escalonados em cada TTI, usando a abordagem heurística e o *baseline* calculado na equação (4.2). O parâmetro β na abordagem heurística foi variado em 0.1, 0.5, e 1, onde, o mais alto valor de β significa que uma maior importância é dada para o atraso.

Na Figura 6.3, pode-se observar que nos resultados do *baseline*, linha sólida, os SBs são atribuídos para os usuários que podem alcançar os mais altos índices de taxa de transferência, *i.e.*, aqueles com os melhores CQIs foram alocados primeiro. Nota-se que o *baseline* tende a alcançar melhores taxas, isto acontece, devido ao fato que os usuários com melhores condições de canal são escalonados ao invés dos que tem condições de canais piores. O *baseline*, representa os resultados ótimos alcançados pela linearização, veja que para otimização da taxa de transferência, o serviço que possibilitava os melhores valores de taxa de bits foi escolhido no escalonamento. Desta forma, a linha sólida considera um único serviço.

Figura 6.3: Análise da heurística parametrizada em relação aos valores exatos (*baseline*).



Fonte: pelo autor (2016).

A abordagem heurística com $\beta = 0.1$, significa que, o algoritmo de escalonamento não considerou os diferentes serviços. Como o esperado, é possível derivar que a linha heurística com $\beta = 0.1$ e o *baseline* possuem a mesma média de $\bar{r}$, isto acontece pois ambos estão considerando os valores mais altos de CQI. A heurística com $\beta = 0.5$, significa que o atraso tem uma relevância ligeiramente maior e como consequência a taxa, é em média, 4% menor que no *baseline*. Já na heurística com $\beta = 1.0$ a taxa de transferência, no geral, é atenuada. Este comportamento acontece devido aos usuários com serviços mais importantes serem escalonados antes daqueles com melhores condições de canal.

O comportamento das linhas demonstra a efetividade da estratégia, pois em cenários sobrecarregados, mesmo considerando os diferentes tipos de serviços, conseguem-se taxas de

transferência de apenas 7.5% menores dos valores de escalonamento ótimos. Além disto, como são apresentados nas próximas seções, estes valores são alcançados sem perda de pacotes significante o que, fatalmente, acontece quando somente o CQI é considerado.

6.2.2.2 Análises para Valores de ω

Note que o valor de ω determina a prioridade de cada pacote do sistema. A Tabela 6.4, apresenta quatro pontos, de um plano cartesiano, coletados das avaliações e seus respectivos valores. Os pontos P1, P2, P3 e P4 foram escolhidos por demonstrarem o comportamento geral em todas as análises.

Tabela 6.4: Pontos de análise - cenário de alta densidade.

	α	β	$\bar{r}$	f	Perda de Pacotes
P1	0,1	0,1	0,9178	0,9158	0,4384
P2	0,1	1,0	0,8755	0,9164	0,4837
P3	1,0	1,0	0,9179	0,9157	0,4486
P4	1,0	0,1	0,9931	0,9224	0,6902

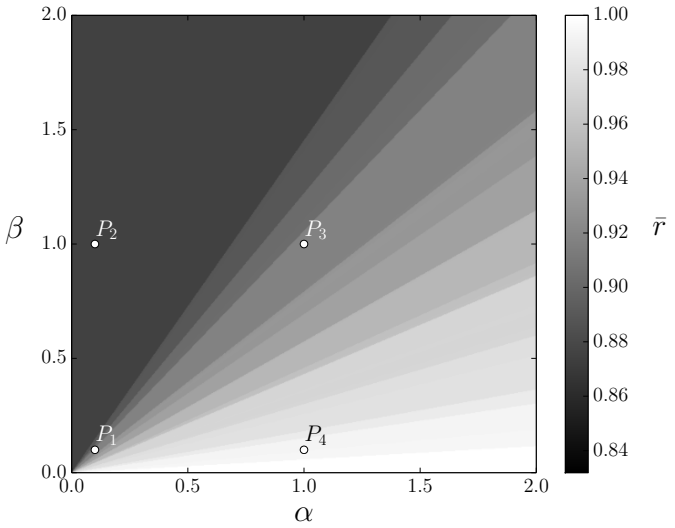
Fonte: pelo autor (2016).

Análise da Taxa de Transferência

Na Figura 6.4 pode-se observar o comportamento do valor médio da taxa de transferência dos usuários $\bar{r}$. Mais especificamente, a figura representa a média de r após o cálculo da equação (4.7) com os parâmetros α e β atribuídos para os valores dos eixos x e y. A região mais clara indica quais parâmetros devem maximizar $\bar{r}$, enquanto a região mais escura indica quais parâmetros devem minimizar $\bar{r}$. Baseado nestes resultados, tons de cinza foram atribuídos para cada ponto, representados pela escala de cinza mostrada a esquerda da figura. Note que, a região ideal é atingida quando a prioridade do atraso é baixa ($0 < \beta < 0.1$).

Baseado na Figura 6.4, é possível inferir que valores baixos para β ($0 < \beta < 0.1$), levam a escolha de peso que efetivamente maximiza a quantidade de informação enviada para a estação base. Este procedimento empírico de ordenar os dados dos usuários é prático e interessante, principalmente, quando cenários sobrecarregados são considerados. Portanto, usando baixo esforço computacional, é possível classificar os pacotes mais relevantes para serem entregues primeiro, desta forma, maximizando a quantidade de informação transmitida. Os resultados obtidos na Figura 6.4 ainda mostram que a desvantagem máxima é de 16% em relação a solução que não considera diferentes serviços.

Figura 6.4: Análise da taxa de transferência.



Fonte: pelo autor (2016).

Análise de Justiça

A justiça no escalonamento de recursos para os usuários foi calculada de acordo com o índice de Jain, proposto por R. Jain *et al.*, em (JAIN; CHIU; HAWE, 1998) e que pode ser calculado como:

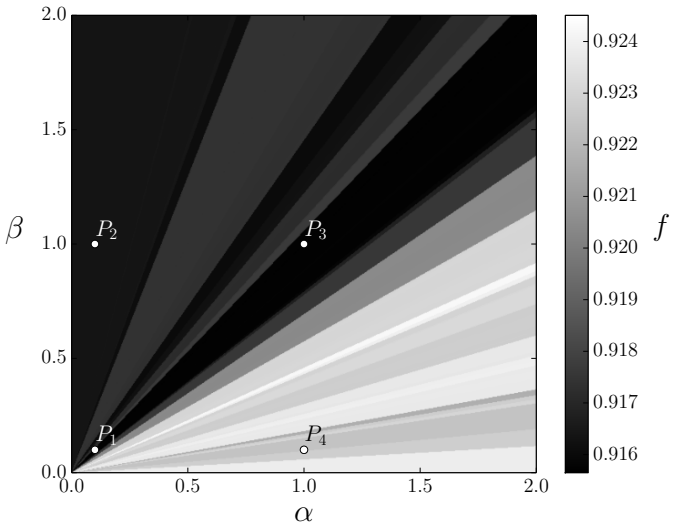
$$f(x) = \frac{(\sum_{k=1}^{K} r_k)^2}{K \sum_{k=1}^{K} r_k^2}, \tag{6.1}$$

onde $\frac{1}{K} \leq f \leq 1$.

O índice de justiça de Jain apresentou $0,9$, na maioria das avaliações (onde o valor máximo é 1), isso pode ser observado na Figura 6.5. Estes resultados ocorrem devido à distribuição usada para representar a quantidade de pacotes de cada usuário na rede. Caso uma distribuição normal seja utilizada nesta representação o índice tenderia para $0,1$.

Na Figura 6.5, a região mais clara indica quais parâmetros devem maximizar a justiça, enquanto a região mais escura indica quais parâmetros devem minimizar o índice de justiça. Observando o fato de que os valores encontrados são todos elevados, pode-se inferir que o índice de justiça é diretamente ligado com a distribuição inicial dos pacotes dos usuários. Desta forma, conclui-se que, ajustando os parâmetros de β para valores menores, o índice de justiça é maximizado.

Figura 6.5: Análise do índice de justiça.

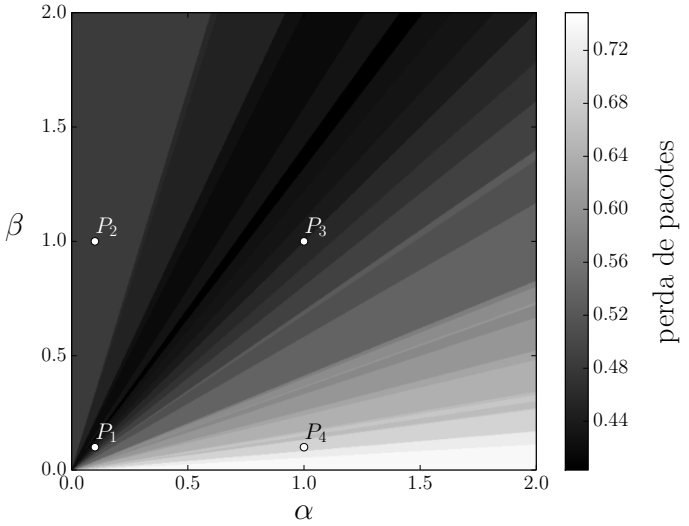


Fonte: pelo autor (2016).

Análise de Perda de Pacotes

Neste cenário, a perda de pacotes acontece, pois não se tem recursos suficientes para acomodar todos os pacotes dos usuários, no entanto, a perda de pacotes pode ser mitigada. Na Figura 6.6 pode se analisar a perda de pacotes no processo de escalonamento. Por um lado, a região mais clara, representa quais parâmetros devem aumentar a perda de pacotes, por outro lado, a região mais escura indica quais parâmetros devem diminuir a perda de pacotes.

Figura 6.6: Análise da perda de pacotes.



Fonte: pelo autor (2016).

Na Figura 6.6, considerando altos valores de CQI (α), obteve-se os mais altos valores de taxa de transferência. Entretanto, o ponto P4 na Figura 6.6 demonstra uma alta perda de pacotes. Já

no ponto P1, quando os parâmetros α e β possuem a mesma importância, percebe-se que a taxa transferência é ligeiramente menor e a perda de pacotes é consideravelmente menor.

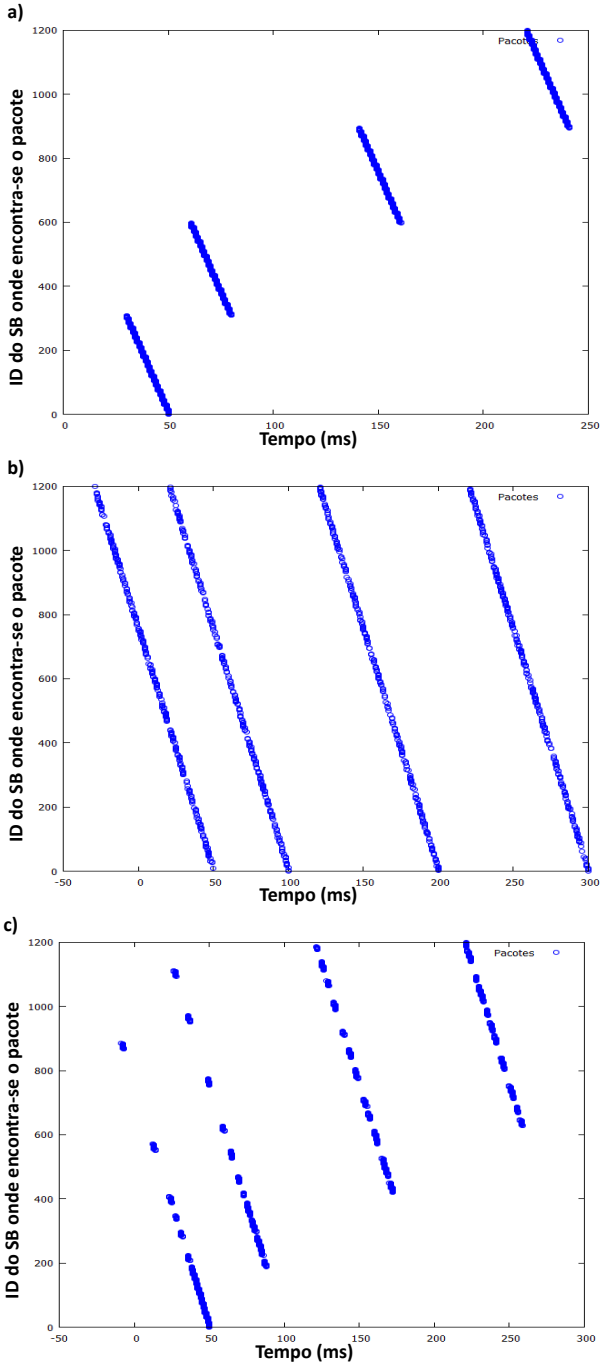
6.3 Comparação das Estratégias

Os gráficos apresentados até então, avaliam o escalonamento utilizando a heurística parametrizável proposta neste trabalho em diferentes cenários. Além disto, é apresentado o comportamento da métrica de escalonamento ω com diferentes valores de α e β. Nesta seção, a métrica de escalonamento é avaliada, frente a algumas outras abordagens disponíveis na literatura. Duas das abordagens de escalonamento mais utilizadas na literatura, *Maximum Sum Rate Scheduling* (MSRS) (ou melhor CQI) e *Guaranteed Delay* (GD), podem ser reproduzidas através da heurística parametrizável, definindo os pesos na equação 4.7. Para reproduzir os resultados da abordagem MSRS, atribui-se $\alpha = 2$ e $\beta = 0$. Já, para reproduzir os resultados de uma abordagem GD, atribui-se $\alpha = 0$ e $\beta = 2$.

Nos resultados apresentados na Figura 6.7 avaliou-se os diferentes comportamentos do *atraso*, usando três abordagens para o escalonamento (GD, MSRS e heurística parametrizável). Nos gráficos apresentados na Figura 6.7 (a), é possível observar o comportamento do escalonamento quando uma abordagem DG é utilizada. Na análise do atraso dos pacotes, percebe-se que, dado o cenário descrito anteriormente, nenhum pacote deixou de ser atendido, ou seja, todos os pacotes foram escalonados no tempo devido, evitando perdas de pacote. Já quando a abordagem MSRS é utilizada, Figura 6.7 (b), percebe-se que os pacotes que estão a esquerda do valor zero no eixo que representa o tempo são descartados. Na Figura 6.7 (c), quando a heurística parametrizada é utilizada, poucos pacotes são perdidos, pois a importância do atraso dos pacotes é levada em consideração assim como o CQI dos usuários.

Na Figura 6.8 (a), quando a estratégia DG é avaliada, nota-se que o gráfico mantém um comportamento padrão, com oscilações de em média 40% para mais e para menos, não há garantia alguma de que a média geral da taxa de bits do escalonamento será a mais alta. Já na Figura 6.8 (b), quando é utilizada uma abordagem MSRS, que leva em conta o CQI do usuário, percebe-se um comportamento que tenta otimizar ao máximo a taxa de transferência final do escalonamento, pois busca os usuários com melhores condições de canal para fazer a transmissão. Ainda analisando a Figura 6.8 (b), verifica-se basicamente 3 padrões da taxa de bits, o gráfico lembra uma escada, onde, os degraus representam os esquemas de modulação utilizados no escalonamento, que são os descritos na Tabela 2.3. Já na Figura 6.8 (c), quando a heurística parametrizada é utilizada, percebe-se o efeito da relação efetuada entre o atraso dos pacotes e a qualidade do canal dos usuários, claramente, o gráfico é intermediário entre as duas estratégias. Embora a abordagem considerando o CQI dos usuários busque a máximo taxa de bits na transmissão, a perda de pacotes desta estratégia pode ser muito elevada, como pode ser visto na Figura 6.7 (b). Tendo em vista que, os pacotes com tempos menores que 0 são descartados, observa-se claramente, que a abordagem que considera CQI é a que mais perde

Figura 6.7: Análise do atraso no escalonamento por: *delay* (a), cqi (b), heurística parametrizada (c).

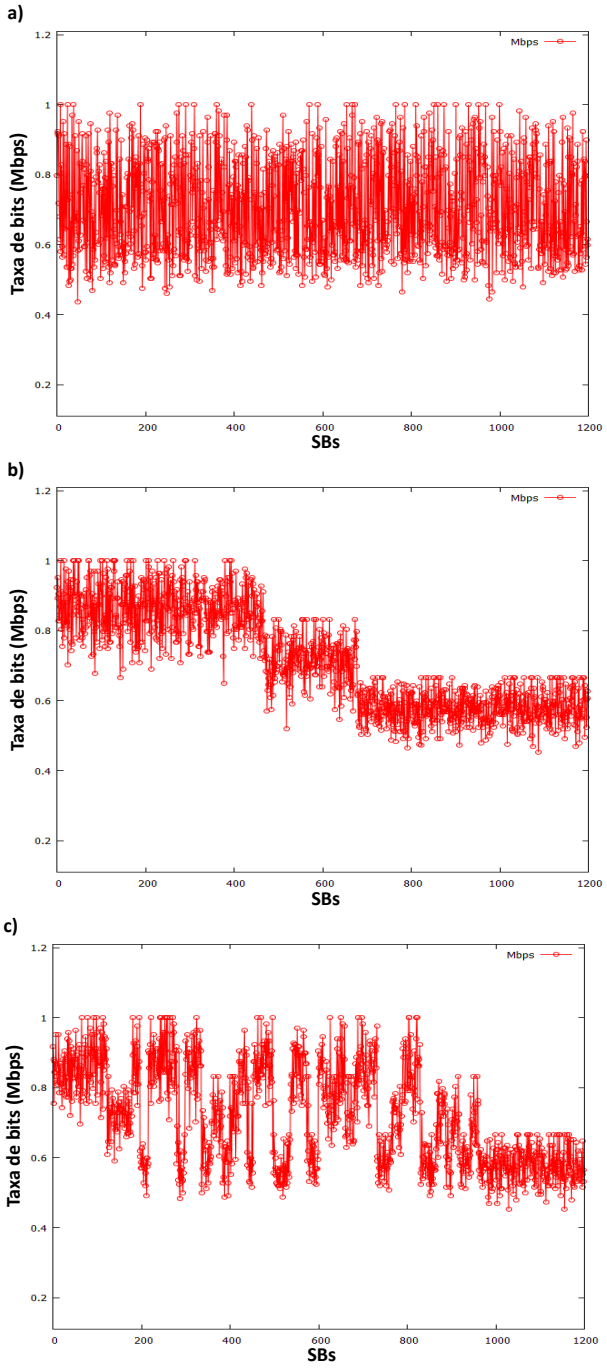


Fonte: pelo autor (2016).

pacotes.

Neste cenário, fica evidente, que uma estratégia considerando o *tradeoff* entre estas abordagens seria conveniente. Deste modo, nos gráficos da Figura 6.7 (c) e na Figura 6.8 (c), os padrões da heurística parametrizada proposta neste trabalho, podem ser percebidos. Na perspectiva do atraso, o número de pacotes perdidos é muito menor que na estratégia considerando

Figura 6.8: Análise da taxa de transferência por: *delay* (a), cqi (b), heurística parametrizada (c).



Fonte: pelo autor (2016).

o CQI. E na perspectiva da taxa de bits, é possível ver que a qualidade do canal dos usuários foi melhor utilizada se comparada com a abordagem que só considerava o atraso, logo, aumentando a taxa de bits média da transmissão.

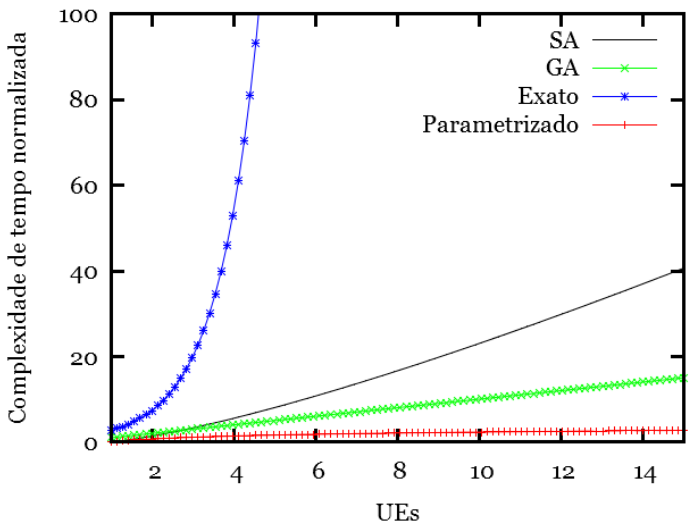
6.3.1 Complexidades

Nas atuais redes móveis é fundamental que a complexidade do escalonamento dos pacotes seja baixa, devido à importância que esta funcionalidade representa ao sistema em geral. Nesta subseção são sumarizadas algumas das complexidades dos algoritmos estudados assim como é descrita a complexidade da estratégia heurística parametrizável proposta nesta dissertação.

No trabalho de (AYDIN; KWAN; WU, 2012), os autores explicam as complexidades dos algoritmos desenvolvidos naquele trabalho. Tradicionalmente, a complexidade da implementação clássica dos algoritmos genéticos (AG) é baseada no laço que define o número de gerações e o tamanho da população, segundo os autores do trabalho estes valores são fixos, e, portanto, a complexidade seria $\mathcal{O}(1)$. A outra complexidade envolvida na abordagem genética é a questão de mutação e do *crossover* resultando em uma complexidade no pior caso $\mathcal{O}(U + U)$ sendo U o número de usuários no sistema. Levando a uma complexidade linear $\mathcal{O}(U)$.

Já o algoritmo *Simulated Annealing* (SA) depende basicamente do laço para controle da temperatura do *Annealing* que os autores afirmam ter fixado em 100. Além deste laço, o SA possui outro laço encadeado que depende de 3 fatores, usuários, MCSs e SBs, logo, a complexidade é $\mathcal{O}(U \times SB \times MCS)$. Na prática os SBs e os MCSs são fixos logo a complexidade se torna $\mathcal{O}(U)$. Entretanto, existe mais a complexidade da função de vizinhos do algoritmo SA que leva em conta a quantidade de usuários, logo, uma complexidade de O(U). Sumarizando as informações desses algoritmos, obtém-se a complexidade total $\mathcal{O}(U) \times \mathcal{O}(U) = \mathcal{O}(U^2)$, que é quadrática no pior caso. Portanto, na complexidade média, obtém-se uma complexidade polinomial.

Figura 6.9: Complexidade dos algoritmos de escalonamento.



Fonte: pelo autor (2016).

Na estratégia heurística parametrizável, utilizou-se uma métrica para o escalonamento baseada no atraso e no CQI do usuário, logo, é necessário, apenas, calcular esta métrica a cada iteração do escalonamento o que corresponde a uma complexidade constate $\mathcal{O}(1)$. Após este cálculo ser realizado para todos os usuários necessita-se ordenar os usuários em função desta métrica. Este ordenamento pode ser realizado através de uma busca binária com uma complexidade $\mathcal{O}(\log n)$ o que leva a uma complexidade média $\mathcal{O}(\log n)$.

6.4 Discussão dos Resultados

Considerando o cenário de alta densidade, todos os pontos discutidos são da Tabela 6.4. Pode-se observar que o ponto P4, aonde $\alpha = 1.0$ e $\beta = 0.1$, os mesmos valores de taxa de transmissão do *baseline* são atingidos, como mostrado na Figura 6.4. Portanto, é possível afirmar que P1 tem uma taxa de transmissão de apenas 7.5% menor que os valores ótimos. No entanto, a perda de pacote em P4 é 25% maior do que em P1. Além disso, observa-se que o índice de justiça entre o *baseline* e a estratégia heurística parametrizável é insignificante, menor de 1%. Com esta análise, é possível concluir também que mesmo considerando uma métrica crucial para cenários sobrecarregados, como o atraso, os resultados de taxa de transmissão são muito próximos do *baseline*, ainda entregando mais pacotes para os usuários.

Kwan *et al.*, (KWAN; LEUNG; ZHANG, 2008) (KWAN; LEUNG; ZHANG, 2009) tentavam maximizar a taxa de transmissão baseados no CQI de usuários. Estas estratégias foram reproduzidas com os dados encontrados em seus artigos e podem ser vistas ajustando nossas métricas parametrizadas para $\alpha = 1.0$ e $\beta = 0.1$. O ponto P4 na Tabela 6.4 representa essas estratégias e pode ser usado para fins de comparação.

Outra constatação, observada nesta análise dá conta que a métrica de justiça, proposta por R. Jain *et al.*, em (JAIN; CHIU; HAWE, 1998) não deve ser diretamente usada para a próxima geração de redes móveis, porque leva em conta apenas a taxa de transferência de dados dos usuários. Por fim, destaca-se que todos os dados deste trabalho são normalizados para permitir a comparação, por isso, a diferença entre os pontos pode ser calculada de forma direta.

6.5 Resumo do Capítulo

Neste capítulo, analisou-se a métrica proposta em relação ao *baseline* provido pela estratégia linearizada. Além disto, a métrica parametrizável foi avaliada em dois cenários de redes LTE, sendo um de baixa densidade e outro de alta densidade de usuários. A complexidade da estratégia proposta foi comparada com outras estratégias avaliadas. Finalmente, foi apresentada uma discussão dos resultados obtidos através dos experimentos realizados. No próximo capítulo, apresentam-se as conclusões atingidas nesta dissertação, assim como, algumas sugestões para trabalhos futuros na área de escalonamento de recursos OFDMA DL.

7 CONCLUSÕES

No início deste trabalho buscavam-se alternativas de solução para o escalonamento nas mais recentes redes de rádios móveis. Através das pesquisas realizadas, percebeu-se que o LTE é a tecnologia que atualmente implementa o 4G. Na literatura, existem diversas pesquisas sobre escalonamento de recursos em LTE. Grande parte destes trabalhos, efetuam o escalonamento considerando a qualidade do canal do usuário. Desta forma, os usuários com condições de rádio momentaneamente favoráveis são privilegiados no escalonamento, levando os usuários com condições desfavoráveis a uma baixa taxa de transferência. Alguns autores tentam mitigar este impacto, incluindo requisitos como a justiça e o atraso tolerado pelos serviços da rede. Cada um dos requisitos mencionados é tido como importante para entregar um serviço efetivo ao usuário final. Todavia, nenhum dos trabalhos pesquisados propôs uma solução considerando todos estes requisitos. Além disso, com a possibilidade de adoção do OFDMA pela próxima geração de redes móveis, torna-se fundamental que as estratégias de escalonamento estejam preparadas para lidar com um número cada vez maior de usuários em cenários sobrecarregados.

Desta forma, nesta dissertação foi idealizada uma nova métrica de escalonamento, eficiente para cenários com altas densidades de usuários, característica das futuras redes de rádios móveis. Esta métrica leva em consideração o atraso que cada serviço pode suportar e a qualidade do canal de cada usuário. Devido aos demais escalonadores estudados durante esta dissertação considerarem apenas a utilização de uma métrica no escalonamento. Constata-se que a nova métrica proposta, provêm maior flexibilidade no escalonamento de dados dos usuários. E, através dos ambientes de simulação apresentados nos resultados deste documento, observa-se a importância da métrica parametrizável em cenários com altas densidades de usuários.

7.1 Contribuições Sumarizadas

No intuito de prover um escalonamento possível para as futuras redes móveis, foi formalizado um escalonador ótimo para DL OFDMA, o qual considera diferentes tipos de serviço através do atraso, sem perdas significantes na taxa de transmissão. Mais do que isso, o modelo exato do escalonador foi linearizado para obter os valores ótimos do escalonamento, provendo um *baseline* de comparação para validar a heurística proposta. Adicionalmente, foi proposta uma nova métrica de escalonamento projetada, especialmente, para cenários sobrecarregados, que considera o atraso dos serviços e a taxa máxima de transmissão, levando a altos índices de justiça no escalonamento, mesmo em cenários com alta densidade de usuários. Para avaliação desta métrica foi desenvolvido um ambiente de simulação de escalonamento de recursos de rádio baseado nas redes LTE, seguindo como base as regulamentações descritas pelo 3GPP. Através deste simulador, foi possível implementar e estudar as principais estratégias de escalonamento presentes, atualmente, na literatura LTE.

A nova métrica de escalonamento foi utilizada na heurística parametrizada proposta e atin-

giu taxas de transmissão apenas 7,5% mais baixas que os valores ótimos, com perdas de pacotes 25% menores que as perdas do *baseline*, e alcançando um índice de justiça (*Jain fairness index*) de 0,91 na escala de Jain em cenários sobrecarregados.

Como um resultado desta pesquisa, foi observado que as mais novas redes de rádio móvel, podem considerar a heurística proposta neste trabalho para implementar as decisões de escalonamento nas redes de rádio móveis de próxima geração. Além disso, no decorrer do trabalho, percebeu-se que a métrica para cálculo da justiça, proposta por R. Jain *et al.*(JAIN; CHIU; HAWE, 1998), não deve ser diretamente utilizada nas redes sem fio da próxima geração, pois esta métrica leva em conta somente a taxa de transmissão dos usuários, deixando de lado métricas importantes, como por exemplo, o atraso.

7.2 Trabalhos Futuros

A seguir, elenca-se alguns tópicos de pesquisa que podem ser abordados em trabalhos relacionados a esta dissertação:

- Evoluir a métrica de justiça para as redes de próxima geração, incluindo características como atraso e perda de pacotes, neste índice. Isso decorre da evolução das redes sem fio, e, assim, as métricas de avaliação destas redes devem ser reconsideradas;
- Inserir a questão da mobilidade dos UEs e a opção de vários eNBs para escalonamento no simulador proposto. Em detrimento da inserção destas variáveis, o desvanecimento e da perda de pacotes pela interferência dos canais adjacentes deverão ser considerados. Como um resultado, isso tornará o simulador mais próximo ao ambiente de rádio real;
- Desenvolver um índice de *Quality of Experience* (QoE), especialmente aplicado no contexto de escalonamento de recursos em redes sem fio de próxima geração. Uma definição formal de um índice para avaliação dos algoritmos de escalonamento pode trazer, para os níveis mais básicos da arquitetura das redes sem fio, as percepções do usuário. Com isso, as abordagens conseguiriam identificar pontos que devem ser melhorados, através do *feedback* imediato dos usuários destas redes;
- Considerar o consumo energético no escalonamento;
- Implementar uma meta-heurística busca tabu para tratar com o escalonamento em regiões de alta densidade;
- Avaliar a heurística proposta, considerando mais características das novas redes sem fio.

REFERÊNCIAS

3GPP TS 23.203. *Evolved Universal Terrestrial Radio Access (E-UTRA); Policy and charging control architecture*. [S.l.], 2010.

3GPP TS 36.211. *Evolved Universal Terrestrial Radio Access (E-UTRA); Physical channels and modulation*. [S.l.], 2008.

3GPP TS 36.213. *Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures*. [S.l.], 2008.

AHMED, H.; JAGANNATHAN, K.; BHASHYAM, S. Queue-aware optimal resource allocation for the lte downlink with best m subband feedback. **IEEE Transactions on Wireless Communications**, v. 14, n. 9, p. 4923–4933, Sept 2015. ISSN 1536-1276.

AI, Q. et al. Qos-guaranteed cross-layer resource allocation algorithm for multiclass services in downlink lte system. In: **Wireless Communications and Signal Processing (WCSP), 2010 International Conference on**. [S.l.: s.n.], 2010. p. 1–4.

ASSAAD, M.; MOURAD, A. New Frequency-Time Scheduling Algorithms for 3GPP/LTE-like OFDMA Air Interface in the Downlink. In: **Vehicular Technology, IEEE Conference**. [S.l.: s.n.], 2008. p. 1964–1969.

AVOCANH, S. et al. A new two-level scheduling algorithm for the downlink of LTE networks. **IEEE Globecom Workshops (GC Wkshps)**, p. 4519–4523, 2013.

AYDIN, M. E.; KWAN, R.; WU, J. Multiuser scheduling on the LTE downlink with meta-heuristic approaches. **Physical Communication**, feb. 2012. ISSN 18744907.

CAIRE, G.; MÜLLER, R.; KNOPP, R. Hard fairness versus proportional fairness in wireless communications: the single-cell case. **IEEE Transactions on information theory, Volume 53, Issue 4, April 2007**, 04 2007. Available from Internet: <http://www.eurecom.fr/publication/2214>.

CAPOZZI, F. et al. Downlink Packet Scheduling in LTE Cellular Networks: Key Design Issues and a Survey. **IEEE Communications Surveys Tutorials**, v. 15, n. 2, p. 678–700, 2013. ISSN 1553-877X.

CHUNG, S. T.; GOLDSMITH, A. Degrees of freedom in adaptive modulation: a unified view. **IEEE Transactions on Communications**, v. 49, n. 9, p. 1561–1571, Sep 2001. ISSN 0090-6778.

COHEN, R.; KATZIR, L. Computational analysis and efficient algorithms for micro and macro OFDMA downlink scheduling. **IEEE/ACM Trans. Netw.**, v. 18, n. 1, p. 15–26, 2010.

DESROSIERS, J.; LÜBBECKE, M. E. **A primer in column generation**. [S.l.]: Springer, 2005.

FAN, J. et al. Joint User Pairing and Resource Allocation for LTE Uplink Transmission. **IEEE Transactions on Wireless Communications**, v. 11, n. 8, p. 2838–2847, August 2012. ISSN 1536-1276.

GROSS, J.; PARRUCA, D. Rate selection analysis under semi-persistent scheduling in LTE networks. **ICNC - International Conference on Computing, Networking and Communications**, IEEE Computer Society, Los Alamitos, CA, USA, v. 0, p. 1184–1190, 2013.

JAIN, R.; CHIU, D.-M.; HAWE, W. A Quantitative Measure Of Fairness And Discrimination For Resource Allocation In Shared Computer Systems. **CoRR**, cs.NI/9809099, 1998.

JING, T.; LIM, M. H.; ONG, Y. S. A parallel hybrid ga for combinatorial optimization using grid technology. In: **Evolutionary Computation, 2003. CEC '03. The 2003 Congress on**. [S.l.: s.n.], 2003. v. 3, p. 1895–1902 Vol.3.

KNOPP, R.; HUMBLET, P. Multiple-accessing over frequency-selective fading channels. In: **Personal, Indoor and Mobile Radio Communications, 1995. PIMRC'95. Wireless: Merging onto the Information Superhighway., Sixth IEEE International Symposium on**. [S.l.: s.n.], 1995. v. 3, p. 1326–.

KWAN, R.; LEUNG, C.; ZHANG, J. Multiuser Scheduling on the Downlink of an LTE Cellular System. **Rec. Lett. Commun.**, Hindawi Publishing Corp., New York, NY, United States, p. 3:1–3:4, jan. 2008. ISSN 1687-6741.

KWAN, R.; LEUNG, C.; ZHANG, J. Proportional Fair Multiuser Scheduling in LTE. **IEEE Signal Processing Letters**, v. 16, n. 6, p. 461–464, 2009. ISSN 1070-9908.

LIU, D.; LEE, Y.-H. An efficient scheduling discipline for packet switching networks using earliest deadline first round robin. **Telecommunication Systems**, v. 28, n. 3-4, p. 453–474, 2005. Available from Internet: <http://dblp.uni-trier.de/db/journals/telsys/telsys28.html#LiuL05>.

LIU, F. et al. A novel qoe-based carrier scheduling scheme in lte-advanced networks with multi-service. In: **Vehicular Technology Conference (VTC Fall), 2012 IEEE**. [S.l.: s.n.], 2012. p. 1–5. ISSN 1090-3038.

MAROTTA, M. et al. Resource sharing in heterogeneous cloud radio access networks. **Wireless Communications, IEEE**, v. 22, n. 3, p. 74–82, June 2015. ISSN 1536-1284.

PARRUCA, D.; GROSS, J. Throughput analysis of proportional fair scheduling for sparse and ultra-dense interference-limited OFDMA/LTE networks. **CoRR**, abs/1510.06530, 2015. Available from Internet: <http://arxiv.org/abs/1510.06530>.

PENG, M. et al. Energy-efficient resource assignment and power allocation in heterogeneous cloud radio access networks. **CoRR**, abs/1412.3788, 2014.

POKHARIYAL, A. et al. HARQ Aware Frequency Domain Packet Scheduler with Different Degrees of Fairness for the UTRAN Long Term Evolution. In: **IEEE 65th Vehicular Technology Conference.** [S.l.: s.n.], 2007. p. 2761–2765. ISSN 1550-2252.

RAMANAN, K.; STOLYAR, A. L. Largest Weighted Delay First Scheduling: Large Deviations and Optimality. v. 11, n. 1, p. 1–48, 2001. Available from Internet: <http://ProjectEuclid.org/getRecord?id=euclid.aoap/998926986>.

SANCHEZ, M. et al. Tackling the increased density of 5g networks: The crowd approach. In: **Vehicular Technology Conference (VTC Spring), 2015 IEEE 81st**. [S.l.: s.n.], 2015. p. 1–5.

SCHWARZ, S.; MEHLFUHRER, C.; RUPP, M. Low complexity approximate maximum throughput scheduling for LTE. In: **Signals, Systems and Computers (ASILOMAR), 2010 Conference Record of the Forty Fourth Asilomar**. [S.l.: s.n.], 2010. p. 1563–1569. ISSN 1058-6393.

SESIA, S.; TOUFIK, I.; BAKER, M. **LTE - The UMTS Long Term Evolution: From Theory to Practice**. 2. ed. [S.l.]: Wiley, 2011. Hardcover. ISBN 0470660252.

TRAN, S.; ELTAWIL, A. Optimized scheduling algorithm for LTE downlink system. In: **IEEE Wireless Communications and Networking Conference (WCNC)**. [S.l.: s.n.], 2012. p. 1462–1466. ISSN 1525-3511.

TSE, D. N.; HANLY, S. V. Multiaccess fading channels. i. polymatroid structure, optimal resource allocation and throughput capacities. **IEEE Trans. Inf. Theor.**, IEEE Press, Piscataway, NJ, USA, v. 44, n. 7, p. 2796–2815, sep. 2006. ISSN 0018-9448. Available from Internet: <http://dx.doi.org/10.1109/18.737513>.

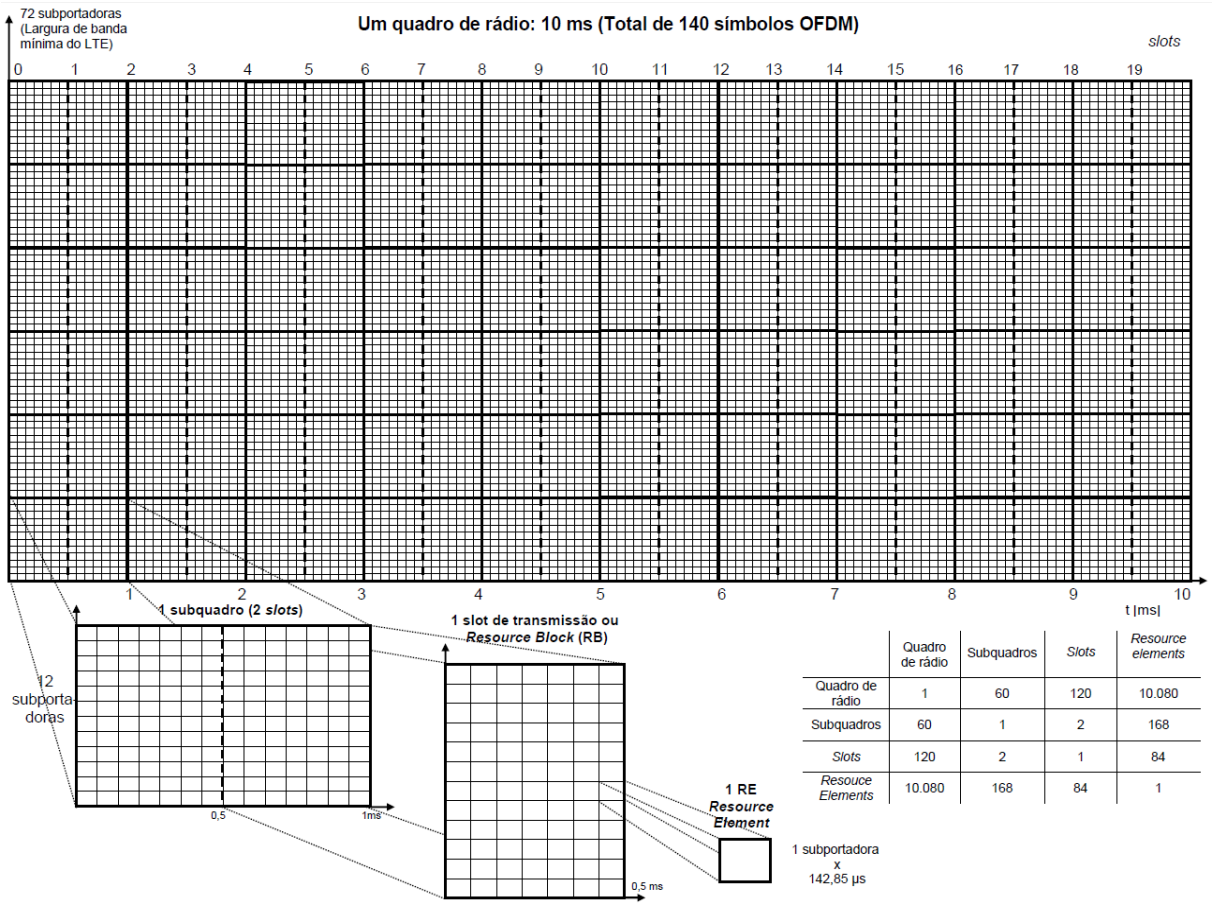
VANDERBEI, R. **Linear Programming: Foundations and Extensions**. [S.l.]: Springer, 1996. (International Series in Operations Research & Management Science). ISBN 9780792398042.

YANG, H. et al. Frequency-Domain Packet Scheduling for 3GPP LTE Uplink. In: **IEEE INFOCOM**. [S.l.: s.n.], 2010. p. 2597–2605. ISBN 978-1-4244-5838-7.

ANEXO A REPRESENTAÇÃO DOS RECURSOS DE RÁDIO LTE DL

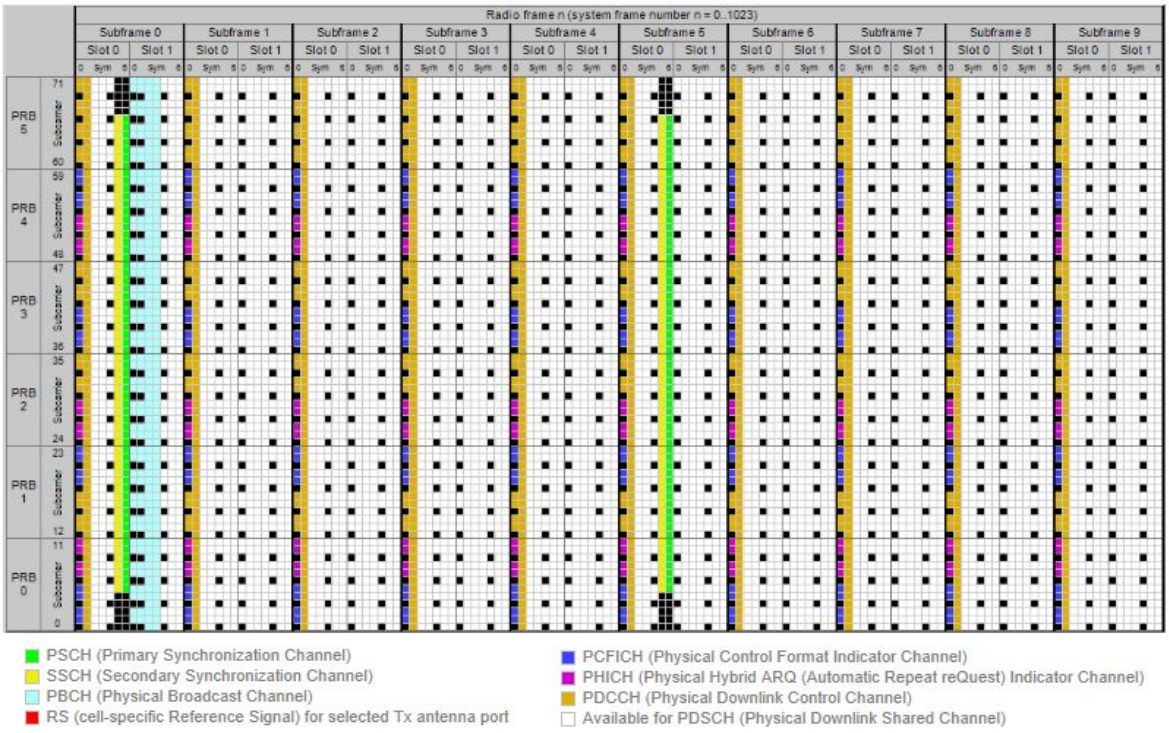
A seguir são demonstradas duas figuras que representam, detalhadamente, os recursos OFDMA DL das redes LTE. Na primeira representação, os sinais de referência da tecnologia são omitidos, já na segunda, estes sinais são apresentados.

Figura A.1: Quadro de rádio LTE DL nos domínios tempo/frequência



Fonte: pelo autor (2015).

Figura A.2: Quadro LTE DL nos domínios tempo/frequência com todas as sinalizações da tecnologia



Fonte: Online. <<http://dhagle.in/LTE>>

ANEXO B ARTIGO SUBMETIDO – SBRC 2016

Long Term Evolution is an emerging as a promising technology that provides a ubiquitous Internet broadband access. However, several requirements must be considered to implement an effective service to users, such as, throughput, delay, and fairness. Despite efforts to solve the downlink (DL) scheduling problem on LTE networks, we are not aware of previous attempts that have addressed the above requirements together in overbooked scenarios. In this paper, we address the radio DL resource scheduling problem for multiple users. A new optimal scheduler is modeled regarding Quality of Service (QoS) provisioning through delay. Moreover, a parameterized heuristic based on user channel quality and service delay is proposed to reach scheduling solutions for overbooked scenarios. Our scheduling shows results of throughput with at most 15% of loses from the optimal value, always considering fairness among users. The proposed approach take into account a good trade-off between spectral efficiency and QoS provisioning. Furthermore, the execution time of our scheduler is insignificant when compared with exact strategies.

- **Título –**
  QoS Aware Schedulers for Multi-users on OFDMA Downlink: Optimal and Heuristic
- **Conferência –**
  XXXIV Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC-2016)
- **Tipo –**
  Trilha Principal (paper-completo)
- **Qualis –**
  B1

QoS Aware Schedulers for Multi-users on OFDMA Downlink: Optimal and Heuristic

Matheus Cadori[1], Samuel Marini[1*], Marcelo Caggiani[1], Cristiano C. Both[2], Juergen Rochol[1]

[1]Instituto de Informática, [1*]Instituto de Física
Universidade Federal do Rio Grande do Sul (UFRGS)
Av. Bento Gonçalves, 9500 - Porto Alegre, RS - Brazil

[2]Instituto de Informática
Universidade Federal de Ciências da Saúde de Porto Alegre (UFCSPA)
Rua Sarmento Leite, 245 - Porto Alegre - RS - Brazil

{mcadori, mluizelli, cbboth, juergen}@inf.ufrgs.br, marini@ufrgs.br

Abstract. *Long Term Evolution is an emerging as a promising technology that provides a ubiquitous Internet broadband access. The growing popularity of mobile devices has brought new scheduling challenges to this technology. Therefore, several requirements must be considered to implement an effective service to users, such as, throughput, delay, packet loss, and fairness. Despite efforts to solve the downlink (DL) scheduling problem on LTE networks, we are not aware of previous attempts that have addressed the above requirements together in overloaded scenarios. In this paper, we address the radio DL resource scheduling problem for multiple users. A new optimal scheduler is modeled regarding Quality of Service (QoS) provisioning through delay. In addition, a parameterized heuristic based on user channel quality and service delay is proposed to reach scheduling solutions for overloaded scenarios. Results demonstrate that the proposed scheduling approaches led to a throughput of 7.5% lower than the optimal ones and 25% lower packet losses in overloaded scenarios. Our model also ensures that the resultant scheduling be as fair as 0.91 (according to Jain fairness index). Additionally, the obtained results show a reasonable trade-off between spectral efficiency and QoS metrics.*

1. Introduction

Future generation cellular networks are expected to provide ubiquitous broadband access to a continuously growing number of mobile users [Capozzi et al. 2013]. In order to support this growth, the Long Term Evolution (LTE), the main technology that has been used to implement fourth generation networks (4G), uses Orthogonal Frequency Division Multiple Access (OFDMA) as a multiple access technology [Peng et al. 2014]. In OFDMA, the exchange of information between user equipment and base station is performed via two channels, namely downlink (DL) and uplink (UL). In a broad sense, these channels are responsible for carrying control information and user data. DL sends data from the base station to the users. In addition, the data provided from multiple users can be carried in a single transmission interval.

The users data are transmitted and scheduled through radio resources using a modulation and codification scheme, so as to provide redundant information based on user

channel quality. To improve the spectral efficiency, the scheduler prioritizes users who have better channel quality. Although this approach achieves a higher throughput, it can also lead users, who have bad channel condition, to a starve state [Avocanh et al. 2013]. Another scheduling approach prioritizes data users with greater delay [Tran and Eltawil 2012]. In spite of the Quality of Service (QoS) be considered in this approach, the overall network performance may decline, due to the choice by users with greater delay instead of those with better channel conditions.

Requirements such the throughput, delay, and fairness are essential for OFDMA transmission and are expected to be considered in the network scheduling process as well. In this regard, the main scheduler challenge consists in how to allocate transmission resources to users, aiming the optimal trade-off between throughput and delay. Additionally, the allocation procedure has to ensure the QoS requirements for as many users as possible [Capozzi et al. 2013]. Despite efforts to solve the DL scheduling problem on Long Term Evolution (LTE), [Avocanh et al. 2013, Tran and Eltawil 2012, Kwan et al. 2008, Kwan, R. and Leung, C. and Jie Zhang 2009], we are not aware of previous attempts that have addressed all above requirements.

In this paper, we take a step further in proposing a scheduler for overloaded LTE networks. Our scheduler leverages the allocation trade-off between throughput and delay and, as a consequence, lead to higher fairness index in resultant solutions. In particular, we formalize an optimal DL scheduler model that consider requirements such as delay and type of service. The proposed model was evaluated through exact and heuristic perspectives. The main contributions of this paper are threefold: (*i*) the formalization of an optimal OFDMA DL scheduler model, which considers different type of services; (*ii*) the design of a heuristic approach to cope with a larger number of users, and (*iii*) the evaluation and discussion of the impact of many priority services being attended on resource scheduling.

The remainder of this paper is organized as follow. In Section 2, a brief background on scheduling process in LTE is presented. In Section 3, we explain the DL system model. The model formulation and linearization of the scheduler are presented in Sections 4.1 and 4.2. The parameterized algorithm is presented in Section 4.3. In Section 5, we present and discuss the results of our evaluation. Finally, in Section 6, we conclude this paper with final remarks and perspectives for future work.

2. Background and Related Work

In this section, we first introduce a basic background related to the resources and the scheduling process in LTE DL. Then, we highlight the main strategies related to the DL scheduling problem that is addressed in this paper.

2.1. LTE DL Resources

A DL-frame resource can be considered as a two dimensional matrix, where the horizontal axis represents time domain and a vertical axis frequency domain. This matrix is subdivided according to LTE specification to be filled with user data blocks. The time-frequency frame resource structure (see Figure 1) is divided in (a) to (d) parts. Each one of these parts is presented with the name of the resource and its respective measure in time and frequency domain.

The Fig. 1 part (a) represents an example of one radio frame of 10 ms/1080 kHz over 72 subcarriers. This subframe is subdivided (b) into 10 subframes of 1 ms/180 kHz over 12 subcarriers. Each subframe is subdivided into 2 slots as can be observed in part (c), where one slot of 0,5 ms/180 kHz over 12 subcarriers is called one Resource Block (RB). Finally, in part (d) one Resource Element (RE) of 71,42 μs/15 kHz over 1 subcarrier is presented. From now on, by convention, the Scheduling Block (SB) term is used to refer two RB over one subframe of 1 ms.

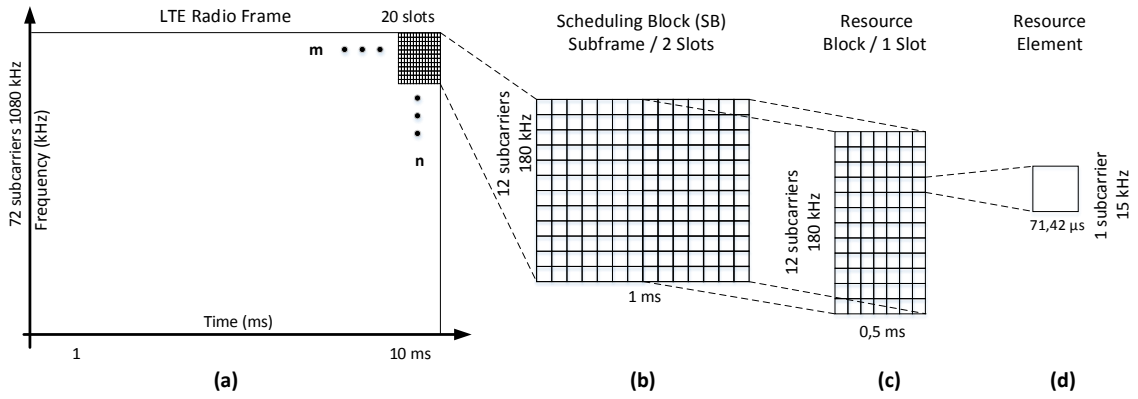


Figure 1. LTE Resource Structure Time/Frequency Overview

The default configuration and normal cyclic prefix is considered in the above representation of the LTE resource structure. The scheduling requirements are detailed in the next subsection.

2.2. Main Requirements of LTE DL Scheduling

Based on, information as an *e.g.* type of service, and quantity of traffic the scheduler selects a user data set to transmit in Transmission Time Intervals (TTIs) *ms*. In addition, the users data is transmitted using a Modulation and Coding Scheme (MCS), chosen according to the user Channel Quality Indicator (CQI).

There are different MCSs in LTE and the amount of redundancy used in transmissions vary among different MCSs. This redundancy serves to ensure the consistency of users data, based on the user channel condition. MCSs are chosen according to the CQI that represents the channel condition of the user at the instant of time. For instance, a user close to some base station and with good channel condition, *i.e.*, high CQI, may use an MCS that inserts less redundancy (and, therefore, leaving more resources for the data itself). Furthermore, a user far away from the base station and with bad channel conditions, *i.e.*, low CQI, may have to use an MCS that inserts more redundancy (thereby leaving fewer resources to user data).

2.3. Related Work

The DL scheduling in LTE is an NP-hard problem [Yang et al. 2010]. Trying to mitigate the hardness of the scheduling problem, the LTE literature [Pokhariyal et al. 2007] treats scheduling on two levels. The first one, Time Domain Packet Scheduler (TDPS), selects certain users among those connected to the base station. The second, Frequency Domain Packet Scheduler (FDPS), assigns RBs to these users. This strategy is adopted because in

doing that, the circular dependency between the two levels is broken. Nevertheless, finding an optimal scheduling algorithm is not possible under standard assumptions [Cohen and Katzir 2010].

Kwan et al. [Kwan et al. 2008] propose an optimal and suboptimal multiuser schedulers, aiming to maximize the users rate in FDPS scheduling. In another study, the authors model the DL scheduler including fairness [Kwan, R. and Leung, C. and Jie Zhang 2009], using the throughput history to make scheduling decisions. Moreover, the importance of heuristic approaches has been investigated by the same author in [Aydin et al. 2012]. This investigation shows the increasingly importance of new suboptimal and heuristic approaches that fulfill as many users requirements as possible. However, QoS requirements considering different delays are not taken into account in these studies.

The literature about DL scheduling uses techniques such as branch-and-bound [Fan et al. 2012], semi-persistent scheduling [Gross and Parruca 2013], and linearization [Kwan et al. 2008] to solve the scheduling problem. However, it can not guarantee global optimality for large instances. The optimal scheduling in LTE networks is proven to be NP-hard, so it is not reasonable to have a polynomial time algorithm to solve all instances of them [Yang et al. 2010]. As a result, a practical approach would be designing efficient heuristic algorithms to deal with the scheduling problem. In this way, the LTE DL scheduling process is mathematically formulated in the Section 3 and a parameterized approach is presented in the Section 4.

3. System Model

In this section, we describe the LTE DL system model, focusing on the scheduling process. This system model, helps to define the constraints of a DL scheduling.

In a LTE DL single antenna multiuser system, we have a set $k = 1, 2..., K$ of users, where K represents the total number of users. Also, the set $n = 1, 2..., N$ represents the number of SBs, where N is the total number of SBs inside one TTI. Furthermore, the set of OFDM symbols inside each SB is represented by $S = 1, 2..., N_s$, where N_s is the total number of symbols in one SB and T_s represents the time of each symbol ($T_s > 0$). Let Y be the total number of subcarriers and $Y_y^{(d)}(s)$ be a subset y of subcarriers, which can be used to carry data signals d in the s^th OFDM symbol duration. This subset considers only symbols available to transmit data, excluding simbols used to technology control.

In DL systems, SBs are transmitted using an MCS. The common modulation schemes in LTE are: BPSK, QPSK, 16QAM, and 64QAM. In addition, each MCS has one code rate CR_j and one constellation size M_j; where CR_j is the code rate associated with the MCS j; M_j is the constellation size of MCS j, and $j = 1, 2..., J$ is the total number of MCS supported in the transmission. Considering commons MCS previously mentioned, the value of J would be 4. The bit rate reached by one single SB considering an MCS j is given [Kwan et al. 2008] by the following equation:

$$r^{(j)} = \frac{CR_j \log_2(M_j)}{T_s N_s} \sum_{s=1}^{N_s} Y_y^{(d)}(s). \tag{1}$$

Every user k reports, via a feedback channel, a CQI vector $CQI_k \in [1..., CQI^{max}]$

containing the supported CQI values for n individuals SBs (in LTE $CQI^{max} = 15$) [Schwarz et al. 2010]. The exact nature of CQI_k vector varies depending on the feedback method adopted [Kwan et al. 2008]. It is assigned an MCS with zero rate to SBs which that have not reported the CQI values. Moreover, CQIs inside a vector corresponds to the supported modulation order and code rate combinations defined in 3GPP [3GPP TS 36.213 2008]. If a UE is served on several SBs, it is necessary to find a CQI value average $\overline{CQI_k}$. In this paper, aiming to simplify the model, we use only integer numbers. Then, it is necessary to round the CQI values [Schwarz et al. 2010] as follow:

$$CQI_k = \lfloor \overline{CQI_k} \rfloor, \tag{2}$$

considering the system model presented earlier, the data rate achieved by each user can be modeled as:

$$r_k = \sum_{n=1}^{N} a_{k,n} \sum_{j=1}^{J} b_{k,j} r^{(j)}, \tag{3}$$

where, $a_{k,n}$ is equal to 1 if the user k is attributed to the SB n and 0 otherwise; and, $b_{k,j}$ is equal to 1 if the user k is attributed to the MCS j and 0 otherwise.

Based on, LTE DL scheduling constraints presented in [Erpek et al. 2015], the scheduling problem is subjected to the following constraints:

$$\sum_{j=1}^{J} b_{k,j} = 1, \tag{4}$$

where, the equation (4) assures that all SBs belonging to the same user within one single TTI must use a unique MCS scheme:

$$\sum_{k=1}^{K} a_{k,n} = 1, \tag{5}$$

where, the equation (5) ensures that a SB can only be used by a single UE in a Single Input Single Output (SISO) system. Note that, SISO, was employed for practical reasons. However, the use of multiple antennas, MIMO, would not have influence in contributions of this work.

In LTE, there are different classes of service, the delay among them varies defining service priorities. In this paper multi-services are used, then the set $p = 1, 2..., P$ represents the services, where P is the total number of services that can be used. Moreover, the binary matrix provided as input, $c_{k,p} = c(k, p)$ defines the service p associated with each user k. Note that, one user can only be associated with one service. The time inside of the system Head Of Line (HOL), of each user data packet, is calculated with the equations (6) and (7):

$$HOL_k = t - t_0, \tag{6}$$

where, t is the current time (clock time) and t_0 is the time that the packet of the user k and service p was timestamped, $(t > t_0)$. Therefore, the remaining scheduling time or delay (δ) to transmit the packets is presented as a function of HOL to each user k as:

$$\delta_k = \sigma_p - HOL_k. \tag{7}$$

In equation (7), σ is the packet delay for the type of service p, also $\sigma > 0$. According to the 3GPP standard, each type of service has a specific time to expire. In this paper we have considered the services given in 3GPP [3GPP TS 23.203 2010].

4. QoS Aware Scheduling

In this section, we formalize the problem and solve it by linearized approach, achieving the optimal values for some instances. Following, we propose a heuristic that provides QoS through of a new parameterized metric considering CQI and delay.

4.1. Problem Formulation

The main goal of our scheduler is to maximize the overall throughput of users in the network. At the same time, our scheduler also ensures a correct and suitable resource allocation that respect all the system model constraints. Considering the presented system model, the scheduling of resource blocks for multiple users, can be formulated as a combinatorial optimization problem as follows:

$$\text{maximize} \sum_{k=1}^{K} \sum_{n=1}^{N} \sum_{p=1}^{P} a_{k,n} c_{k,p} \sum_{j=1}^{J} b_{k,j} r^{(j)}, \tag{8}$$

subject to the constraints showed in equations (4) and (5).

The objective in equation (8) is to maximize the throughput of users, calculated at equation (1), with priority services. Observing the equation (8) we can note that the MCSs, SBs, users, and services are assigned together. Therefore, the problem formulation is non-linear due to the multiplication among binary variables. In this way, we propose a linearization of the problem in order to solve some instances of the problem and to provide the basis to idealize a heuristic approach to deal with DL scheduling.

4.2. Linearized Scheduler

The idea behind linearization is threefold: (*a*) make it solvable by integer linear programming; (*b*) provide a baseline to compare how far the heuristic approach is from the optimal values; (*c*) provide a new linearized model for an OFDMA scheduler considering 5G wireless networks features. Finally, considering the equation (8), one possible linearization is:

$$\text{maximize} \sum_{k=1}^{K} \sum_{n=1}^{N} \sum_{p=1}^{P} \sum_{j=1}^{J} v_{k,n,p,j} r^{(j)}, \tag{9}$$

subject to the constraints (4), (5), and:

$$v_{k,n,p,j} \leq b_{k,j}; \tag{10}$$

$$v_{k,n,p,j} \leq a_{k,n} * R; \tag{11}$$

$$v_{k,n,p,j} \leq b_{k,j} - (1 - a_{k,n}) * R; \tag{12}$$

$$v_{k,n,p,j} \leq b_{k,j} - (2 - a_{k,n} - c_{k,p}) * R, \tag{13}$$

in equations (11), (12), and (13) R is a big real number. The R value is used, due to the big-M [Desrosiers and Lübbecke 2005] relaxation approach be used.

The variable $v_{k,n,p,j}$ is the product of binary variables $a_{k,n}b_{k,j}c_{k,p}$. Furthermore, the equations (11), (12), and (13) were used to ensure the same behavior between $v_{k,n,p,j}$ and the $a_{k,n}$,$b_{k,j}$ and $c_{k,p}$. An auxiliary variable needed to be inserted, hence the solution space increases. For this reason, the time to solve this problem even considering small instances may be too long. To perform the above procedure, we obtain a baseline for later use in section 5.

4.3. Parameterized Heuristic Scheduler

In section 2.3 we realized that the OFDMA DL scheduling problem is a NP-hard problem, so unless P = NP we can not guarantee that the scheduling decision can be performed in polynomial time for all instances. To avoid this difficulty, in section 4.2, the problem was transformed into an equivalent linear problem by introducing an auxiliary variable, however, the solution space increase rapidly using this technique.

Observing the above issues and foreseeing the importance of an algorithm with adjustable QoS parameters to future wireless networks, we propose an weight function (ω) aiming to maximize $r^{(j)}$ taking into account the delay and the CQI of the users. This function prioritizes the data from users with both the highest CQI and the delay nearest to expire. Explicitly, we can write the weighted function as follows:

$$\omega_k = \frac{(CQI_k)^\alpha}{(\delta_k)^\beta}, \tag{14}$$

where $(\delta_k)^\beta \neq 0$, and α, and β, are parameters that measure the importance of the CQI and delay, respectively. In addition, the scheduler calculates the weight function for each set of data and sorts them by weight. In such way, the set of data with larger weights is scheduled first. To determine parameters α and β that assign the correct weight to each data set (maximize r), we proceed with computational simulations.

The Algorithm 1 describes how to use the proposed metric, in simulations. The algorithm starts with a set of packets ϕ, that represents the users data. The condition in line (1) ensures that while the packets from users are not finished, TTIs will continue being created. At line (1), using the argument $technologyLimit$, we prevent the SBs per TTIs limit from being exceeded, depending on the technology. In this way, the first user data of the set ϕ is scheduled. Finally, we return the properly filled TTI.

Note that the parameterized approach was idealized to support as many services as necessary, so only parameters in equation ω (line 1) must be adjusted to obtain the performance that best fit the network needs. In the next section, the experimental results of the Algorithm 1 varying α and β is demonstrated.

Input: Start with a data set of packets from users: ϕ
Input: Limit of SBs per TTI: $technologyLimit$
Output: Filled TTI
Calculate ω (ϕ,α,β)
while $\phi > 0$ **do**
  OrderBy ω (ϕ)
  CreateNewTTI()
  while $SBs \leq technologyLimit$ **do**
    ScheduleTheFirst(ϕ)
  end
end

Algorithm 1: Parameterized Approach

5. Evaluation

In this section, we show a case study example of DL scheduling scenario. First, we describe the case study and the simulated environment. Then, we evaluate and compare the obtained results with the proposed baseline.

5.1. Case Study

In the DL scheduling process, the amount of data from users is handled by the scheduler (which is responsible for attending as many users request as possible). In this process, many requirements must be considered. One of these requirements is the time to transmit the user data, regarding the different type of services involved. For instance, the mean delay in transmitting video (streaming) services is 100 ms and to HTML is 300 ms following the 3GPP standards [3GPP TS 36.213 2008]. Note that in all scenarios considered in this paper, we are using the services provided by 3GPP standards [3GPP TS 36.213 2008]. Thus, equation 7 is based on these delay values to each service.

Observing the case study above, the delay observed in the video transmission service is smaller than the delay of the HTML service (it happens because ideally videos need to flow continuously in order to provide a better quality of experience to users). If the packet is not being handed out in a timely way, it is dropped. To avoid the excessive packet loss, the scheduler must take into account the maximum delay allowed for each service. Moreover, depending on the CQI between the base station and the user, the same information needs to be sent more than once. This redundancy in sending information generates overhead and ideally needs to be avoided.

5.2. Scenario and Environment

The LTE radio environment was built using the channel bandwidth of 1.4 MHz in FDD mode, which corresponds to 6 SBs per TTI. A downlink frame with a 10 ms scheduling request periodicity and cyclic prefix type of 1/4 was considered. Moreover, the temporal variation model of the fading channel was characterized following the Rician fading distribution.

The scenario used in the evaluation represents the analysis of an overloaded LTE scenario with packets from 1000 different users, generating 30000 packets uniformly distributed. These packets needed be delivered over 120 TTIs *i.e.*, 120 ms. The size of those

packets follows a Poisson distribution. Moreover, the CQI among users was uniformly adjusted.

The evaluation of the linearized approach was performed analytically and through computer simulations. In order to evaluate the impact of the parameterized metric against the baseline. The model formalized in the previous section was implemented and executed in CPLEX Optimization Studio, version 12.3. All experiments were performed on a machine with four AMD Opteron 6276 processors and 64 GB of RAM, using the Operational System Ubuntu GNU/Linux Server 11.10 x86 64.

5.3. Results

In order to demonstrate the parameterized heuristic results, in this section three metrics are evaluated: rate, fairness, and packet loss.

5.3.1. Rate Analysis in an overloaded Scenario

The parameterized heuristic and the baseline were analyzed using the overloaded scenario, and the case study, previously described. The graphic in Figure 2 presents the normalized sum $\bar{r}$ of the packets being scheduled in each TTI, using the heuristic approach and the baseline calculated with the equation (9). The parameter β in heuristic approach was varied in 0.1, 0.5, and 1, intuitively, the higher the value of β greater significance is given to the delay.

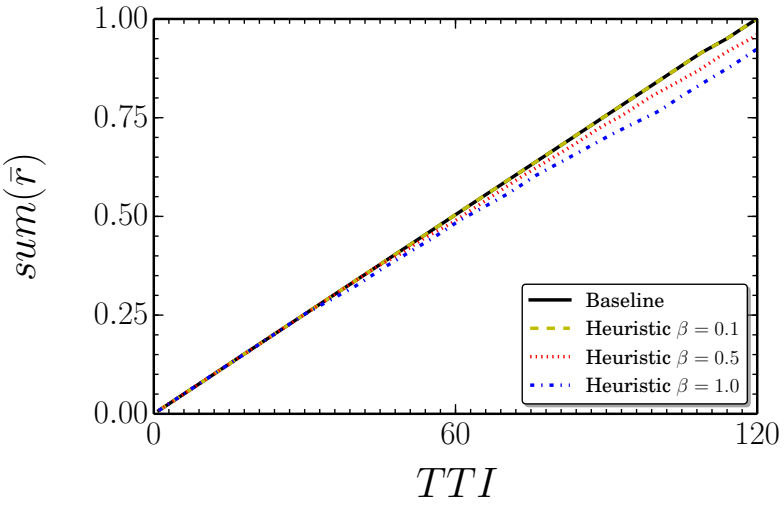


Figure 2. Simulation of an overloaded LTE scheduling scenario

In Fig. 2, we can observe that in the baseline results, solid line, the SBs are assigned to users who achieve the highest rates, *i.e.*, those with best CQI values. We can observe that baseline tends to reach better rate results because the users that have higher rates are allocated rather than those users with worse channel conditions. However, baseline does not take into account service classes.

The heuristic approach with $\beta = 0.1$, dashed line, means that the scheduler does not take into account service classes. Therefore, it is possible to derive that heuristic

and the baseline keep the same average of $\bar{r}$, it happens because both are considering highest CQI values and not considering service classes. The heuristic with $\beta = 0.5$, dotted line, means that the delay has a slightly higher relevance and hence the throughput is, on average, 4% lower than the baseline. In heuristic with $\beta = 1.0$, dot/dashed line, the throughput of users, in general, is attenuated. This behavior happens due to the fact that users with more important services are scheduled before than those who have better channel conditions.

In conclusion, the lines behavior demonstrates the effectiveness of our strategy in relation to baseline in overloaded scenarios, even considering the delay we can reach the throughput only 7.5% lower than baseline, in addition, our strategy do not lose many packets in this process, as we can see next.

5.3.2. Rate Analysis

The next three charts show the results of variation of parameters α and β from 0.0 to 2.0. The parameters α and β were chosen because it reproduces the global system behavior. At each iteration, we recalculated the ω, retracing the scheduling and evaluating a determined metric to build the charts. We applied this methodology for rate, fairness, and packet loss.

In order to provide a better readability of our results, we present the following Table 1, showing 4 points and their values, collected from the next 3 graphics. The points P1, P2, P3, and P4 were chosen for demonstrating the behavior followed in the analysis.

Table 1. Analysis Points

	α	β	$\bar{r}$	f	packet loss
P1	0.1	0.1	0.9178	0.9158	0.4384
P2	0.1	1.0	0.8755	0.9164	0.4837
P3	1.0	1.0	0.9179	0.9157	0.4486
P4	1.0	0.1	0.9931	0.9224	0.6902

The Figure 3 (a) shows the behavior of $\bar{r}$ value for users rate. In addition, the figure represents the mean of r obtained through the sorting of packets by weight, calculated by equation (14). The lighter region indicates which parameters might maximize $\bar{r}$ while the darker region indicates which parameters might minimize $\bar{r}$. Based on the results, a color was assigned to each point such that the highest values of $\bar{r}$ was assigned to darker color and lowest $\bar{r}$ was represented by lighter colors, according to the color scale shown on the left of the figure. Note that, the ideal region is achieved when low priority is given to the delay ($0 < \beta < 0.1$).

Based on Fig. 3 (a), it is possible to infer that low values to β ($0 < \beta < 0.1$), lead to the choice of weight that effectively maximizes the amount of information sent by the base station. This empirical procedure of sorting the users data is practical and interesting to consider in overloaded systems. Therefore, using a low computational cost, we can classify the most relevant packets to be delivered to users and thereby maximizing the amount of information to be transmitted. Results reached through this procedure, present a maximum disadvantage of 15% of the exact approach.

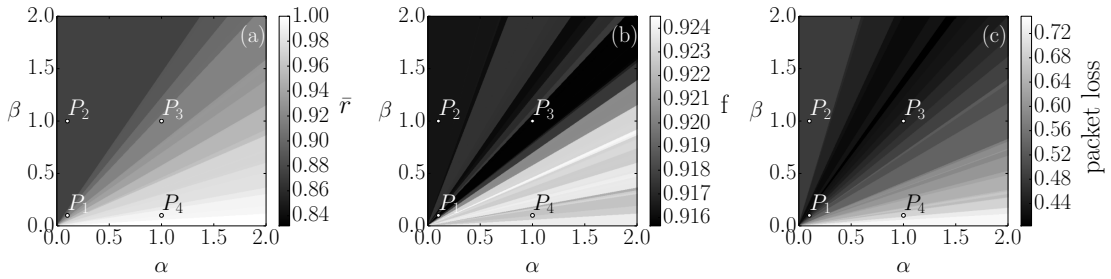


Figure 3. Parameterized map using a color scale index, analyzing the rate (a), analyzing the fairness (b), and analyzing the packet loss (c).

5.3.3. Fairness Analysis

The fairness among users were computed via Jain fairness index proposed by R. Jain *et al.* in [Jain et al. 1998] which can be stated as:

$$f(x) = \frac{(\sum_{k=1}^{K} r_k)^2}{K \sum_{k=1}^{K} r_k^2}, \tag{15}$$

where $\frac{1}{K} \leq f \leq 1$.

In our results we reach a Jain fairness index around 0.9, in most of the evaluations (where the most fair is 1), it can be observed in Figure 3 (b). These results are due to the uniform distribution used to distribute packets to users. In the case of the packets users following a normal distribution, we would have a justice index around 0.1 for the heuristic approach.

In Fig. 3 (b), the lighter region indicates which parameters might maximize the fairness (calculated in equation (15)), while the darker region indicates which parameters might minimize the fairness index. Observing this data we can infer that the fairness index is directly related to the distribution of packets among users. Moreover, we can conclude that adjusting the parameter β of our heuristic led to best fairness.

5.3.4. Packet Loss Analysis

Fig. 3 (c) shows the packet loss in the scheduling process. In addition, the lighter region indicates which parameters might increase the packet losses, while the darker region indicates which parameters might decrease the packet loss.

As can be observed at the point P4 in Fig. 3 (c), considering a high value CQI (α) we obtained a high rate of value. However, as the value of β is low, the packet loss is the highest observed among the four points. The point P1 shows that when the parameters α and β have the same importance, we obtained a slightly lower rate, but the loss of considerably smaller packages.

5.4. Discussion of the Results

Considering P4, where $\alpha = 1.0$ and $\beta = 0.1$, we reach the same rate values of our baseline, as shown in Fig. 3. Therefore, we can affirm that P1 has a rate 7.5% worse than

the baseline. However, the loss of packet P4 is 25% higher than in P1. Besides, we figure out that the fairness between baseline and our strategy is negligible, less than 1%. In this analysis, we can conclude that even considering a crucial metric to overloaded scenarios, as the delay, the rate results are very close to the baseline, yet delivering more packets to users.

Several authors, [Kwan et al. 2008], [Kwan, R. and Leung, C. and Jie Zhang 2009], [Ahmed et al. 2015], try to maximize the rate based on the CQI of users. These strategies have been reproduced with the data found in your articles and can be seen adjusting our parameterized metrics to $\alpha = 1.0$ and $\beta = 0.1$. P4 in Table 1 point represents these strategies and can be used to a comparison purposes. Other approaches, [Abdel-Hadi and Clancy 2014], try to maximize the fairness in data transmission, *i.e.*, $\alpha = 0.1$ and $\beta = 1.0$. However, as observed in Figure 3 and in Table 1, this approach wastes a lot of packtes, and consequently, attenuates the transmission rate.

6. Conclusion

In this work, we formalize an optimal OFDMA DL scheduler, which considers the type of services through the delay, without significant throughput losses. In addition, we linearized it to reach optimal values for some instances providing a throughput baseline. Additionally, we developed a parameterized heuristic considering different types of services, fairness and packet loss. The scheduling using our parameterized heuristic, achieves throughput only 7.5% lower than the optimal values, with 25% less packet loss, and reaching a Jain fairness index of 0.91 in overloaded scenarios.

As a result of this research, we note that the newest radio networks can consider the parameterized heuristic proposed in this paper to implement the scheduling decisions, ideally, in overloaded scenarios. In addition, we observe that the fairness metric proposed by R. Jain *et al.* [Jain et al. 1998], just takes into account the throughput of the users. Therefore, important metrics such as delay and packet loss are not considered.

As a future work, we suggest to include the delay and packet loss metrics in the Jain fairness index, enabling it to be used in future networks. Furthermore, we envision developing a meta-heuristic tabu search algorithm to compare with our parameterized heuristic. Finally, we can highlight another possibility for future work, to consider the impact of different packet distributions representing sundry scenarios, in the fairness index.

References

3GPP TS 23.203 (2010). Policy and charging control architecture . Technical report, 3rd Generation Partnership Project .

3GPP TS 36.213 (2008). Evolved Universal Terrestrial Radio Access; Physical layer procedures. Technical report, 3rd Generation Partnership Project .

Abdel-Hadi, A. and Clancy, C. (2014). A utility proportional fairness approach for resource allocation in 4g-lte. In *Computing, Networking and Communications (ICNC), 2014 International Conference on*, pages 1034–1040.

Ahmed, H., Jagannathan, K., and Bhashyam, S. (2015). Queue-aware optimal resource allocation for the lte downlink with best m subband feedback. *Wireless Communications, IEEE Transactions on*, 14(9):4923–4933.

Avocanh, S., Thierry, J., Abdennebi, M., and Ben-Othman, J. (2013). A new two-level scheduling algorithm for the downlink of LTE networks. *IEEE Globecom Workshops (GC Wkshps)*, pages 4519–4523.

Aydin, M. E., Kwan, R., and Wu, J. (2012). Multiuser scheduling on the LTE downlink with meta-heuristic approaches. *Physical Communication*.

Capozzi, F., Piro, G., Grieco, L., Boggia, G., and Camarda, P. (2013). Downlink Packet Scheduling in LTE Cellular Networks: Key Design Issues and a Survey. *IEEE Communications Surveys Tutorials*, 15(2):678–700.

Cohen, R. and Katzir, L. (2010). Computational analysis and efficient algorithms for micro and macro OFDMA downlink scheduling. *IEEE/ACM Trans. Netw.*, 18(1):15–26.

Desrosiers, J. and Lübbecke, M. E. (2005). *A primer in column generation*. Springer.

Erpek, T., Abdelhadi, A., and Clancy, T. (2015). An optimal application-aware resource block scheduling in lte. In *Computing, Networking and Communications (ICNC), 2015 International Conference on*, pages 275–279.

Fan, J., Li, G., Yin, Q., Peng, B., and Zhu, X. (2012). Joint User Pairing and Resource Allocation for LTE Uplink Transmission. *IEEE Transactions on Wireless Communications*, 11(8):2838–2847.

Gross, J. and Parruca, D. (2013). Rate selection analysis under semi-persistent scheduling in LTE networks. *ICNC - International Conference on Computing, Networking and Communications*, 0:1184–1190.

Jain, R., Chiu, D.-M., and Hawe, W. (1998). A Quantitative Measure Of Fairness And Discrimination For Resource Allocation In Shared Computer Systems. *CoRR*, cs.NI/9809099.

Kwan, R., Leung, C., and Zhang, J. (2008). Multiuser Scheduling on the Downlink of an LTE Cellular System. *Rec. Lett. Commun.*, pages 3:1–3:4.

Kwan, R. and Leung, C. and Jie Zhang (2009). Proportional Fair Multiuser Scheduling in LTE. *IEEE Signal Processing Letters*, 16(6):461–464.

Peng, M., Zhang, K., Jiang, J., Wang, J., and Wang, W. (2014). Energy-efficient resource assignment and power allocation in heterogeneous cloud radio access networks. *CoRR*, abs/1412.3788.

Pokhariyal, A., Pedersen, K., Monghal, G., Kovacs, I. Z., Rosa, C., Kolding, T., and Mogensen, P. (2007). HARQ Aware Frequency Domain Packet Scheduler with Different Degrees of Fairness for the UTRAN Long Term Evolution. In *IEEE 65th Vehicular Technology Conference.*, pages 2761–2765.

Schwarz, S., Mehlfuhrer, C., and Rupp, M. (2010). Low complexity approximate maximum throughput scheduling for LTE. In *Signals, Systems and Computers (ASILOMAR), 2010 Conference Record of the Forty Fourth Asilomar*, pages 1563–1569.

Tran, S. and Eltawil, A. (2012). Optimized scheduling algorithm for LTE downlink system. In *IEEE Wireless Communications and Networking Conference (WCNC)*, pages 1462–1466.

Yang, H., Ren, F., Lin, C., and Zhang, J. (2010). Frequency-Domain Packet Scheduling for 3GPP LTE Uplink. In *IEEE INFOCOM*, pages 2597–2605.