UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

Compressão de CSI para MIMO Distribuído com Processamento Centralizado

Marcos Davi Lima da Silva

DM: 15/2024

UFPA / ITEC / PPGEE
Campus Universitário do Guamá
Belém-Pará-Brasil
2024

UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

Marcos Davi Lima da Silva

Compressão de CSI para MIMO Distribuído com Processamento Centralizado

DM: 15/2024

UFPA / ITEC / PPGEE
Campus Universitário do Guamá
Belém-Pará-Brasil
2024

UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

Marcos Davi Lima da Silva

Compressão de CSI para MIMO Distribuído com Processamento Centralizado

Submetido à banca examinadora do departamento de pós-graduação em Engenharia Elétrica da Universidade Federal do Pará em cumprimento parcial dos requisitos para o curso de Mestrado em Engenharia Elétrica com ênfase em Telecomunicações.

UFPA / ITEC / PPGEE
Campus Universitário do Guamá
Belém-Pará-Brasil
2024

Dados Internacionais de Catalogação na Publicação (CIP) de acordo com ISBD
Sistema de Bibliotecas da Universidade Federal do Pará
Gerada automaticamente pelo módulo Ficat, mediante os dados fornecidos pelo(a) autor(a)

S586c Silva, Marcos Davi Lima da.
Compressão de CSI para MIMO Distribuído com Processamento Centralizado / Marcos Davi Lima da Silva. — 2024.
100 f. : il. color.

Orientador(a): Prof. Dr. Leonardo Lira Ramalho
Dissertação (Mestrado) - Universidade Federal do Pará, Instituto de Tecnologia, Programa de Pós-Graduação em Engenharia Elétrica, Belém, 2024.

1. D-MIMO. 2. Cell-Free. 3. Fronthaul. 4. Compressão. I. Título.

CDD 621.3822

UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

Compressão de CSI para MIMO Distribuído com Processamento Centralizado

AUTOR: **MARCOS DAVI LIMA DA SILVA**

DISSERTAÇÃO DE MESTRADO SUBMETIDA À BANCA EXAMINADORA APROVADA PELO COLEGIADO DO PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA, SENDO JULGADA ADEQUADA PARA A OBTENÇÃO DO GRAU DE MESTRE EM ENGENHARIA ELÉTRICA NA ÁREA DE TELECOMUNICAÇÕES.

Aprovada em: 18/06/2024

BANCA EXAMINADORA:

Prof. Dr. Leonardo Lira Ramalho
(Orientador – PPGEE/UFPA)

Prof. Dr. Aldebaro Barreto da Rocha Klautau Júnior
(Avaliador Interno – PPGEE/UFPA)

Prof. Dr. Diego de Azevedo Gomes
(Avaliador Externo – FACEEL/UNIFESSPA)

VISTO:

Prof. Dr. Diego Lisboa Cardoso
(Coordenador do PPGEE/ITEC/UFPA)

Agradecimentos

Agradeço a Deus por me conceder forças para conseguir vencer todas as etapas da minha vida e por me mostrar o caminho certo a seguir, por me proporcionar a cada dia o sustento e saúde para continuar seguindo, e por sempre estar ao meu lado, independente do momento e da circunstância.

Agradeço a minha família (meus pais, Odir e Denise, e meu irmão Daniel) por sempre estar disposta a me ajudar e por sempre acreditarem na minha capacidade. Por me acolherem com amor e carinho em cada momento da minha vida e por me proporcionarem a base pela qual pude chegar até aqui.

Agradeço a minha namorada e companheira Fernanda por me conceder muito amor, carinho e paz, e por me apoiar e me dar forças em todas as áreas da minha vida, sendo essencial para eu conseguir cumprir esta etapa tão importante e celebrar comigo todas as minhas conquistas.

Agradeço a todos os meus outros familiares que sempre estiveram torcendo por mim e que em todo momento esperavam ansiosamente por mais essa conquista em minha vida. Pela minha avó Maria que sempre se preocupou comigo e que presenciou cada conquista alcançada por mim.

Agradeço a todos os meus amigos que me incentivaram a continuar me esforçando em busca dos meus sonhos, proporcionando momentos de alegria, paz e harmonia, os quais contribuíram para que eu pudesse alcançar essa conquista tão importante.

Agradeço aos meus orientadores Aldebaro Klautau e Leonardo Ramalho por prestarem todo o suporte necessário para a minha vida acadêmica, o qual fez com que eu pudesse alcançar todos os resultados presentes nesse trabalho. Por contribuírem com o aperfeiçoamento do meu leque de conhecimentos adquiridos durante toda esta trajetória. Por sempre estarem dispostos a me ajudar e a me ensinar de maneira mais compreensível possível. E por fazerem do ambiente de trabalho também um lugar de alegria e harmonia.

Agradeço a Ericsson, a FADESP e a CAPES por contribuírem financeiramente com a pesquisa e por me proporcionarem suporte técnico e científico para que eu pudesse alcançar resultados melhores e mais satisfatórios.

Marcos Davi Lima da Silva

Junho 2024

Pois desde a criação do mundo os atributos invisíveis de Deus, seu eterno poder e sua natureza divina, têm sido vistos claramente, sendo compreendidos por meio das coisas criadas, de forma que tais homens são indesculpáveis;
Romanos 1:20

Apóstolo Paulo

Lista de Acrônimos

3GPP *3rd Generation Partnership Project*

AP *Antenna Point*

BBU *Baseband Unit*

BFP *Block Floating Point*

BLER *Block Error Rate*

BS *Block Scaling*

CB *Conjugate Beamforming*

CDF *Cumulative Distribution Function*

CF-mMIMO *Cell-Free Massive Multiple-Input Multiple-Output*

CP *Cyclic Prefix*

CPU *Central Processing Unit*

CRC *Cyclic Redundancy Check*

CSI *Channel State Information*

CSI-RS *Channel State Information Reference Signal*

CU *Central Unit*

DCC *Dynamic Cooperation Clustering*

DFT *Discrete Fourier Transform*

D-MIMO *Distributed MIMO*

DM-RS *Demodulation Reference Signal*

DU *Distributed Unit*

FDD *Frequency-Division Duplex*

FFT *Fast Fourier Transform*

gNB *Next Generation NodeB*

IFFT *Inverse Fast Fourier Transform*

IoT *Internet of Things*

IP-MMSE *Improved Partial Minimum Mean Square Error*

IQ *in-phase and quadrature*

IQR *Interquantile Range*

IR *Incremental Redundancy*

LDPC *Low-Density Parity Check Code*

LoS *line-of-sight*

LSB *least significant bit*

LTE *Long Term Evolution*

LUT *lookup table*

MCS *Modulation and Coding Scheme*

MIMO *multiple-input multiple-output*

mMIMO *massive MIMO*

MMSE *Minimum Mean Squared Error*

MRT *Maximum Ratio Transmission*

MU-MIMO *Multiple User - Multiple Input Multiple Output*

NLoS *non-line-of-sight*

NMSE *Normalised Mean Square Error*

NR *New Radio*

O-DU *O-RAN Distributed Unit*

OFDM *Orthogonal Frequency-Division Multiplexing*

OLLA *Outer Loop Link Adaptation*

O-RAN *Open Radio Access Network*

O-RU *O-RAN Radio Unit*

PBCH *Physical Broadcast Channel*

PDSCH *Physical Downlink Shared Channel*

PRACH *Physical Random Access Channel*

PRB *Physical Resource Block*

PSS *Primary Synchronization Signal*

PUSCH *Physical Uplink Shared Channel*

QAM *Quadrature Amplitude Modulation*

QPSK *Quadrature Phase Shift Keying*

RAN *Radio Access Network*

RAR *Random Access Response*

RE *Resource Element*

RF *Radio Frequency*

RRH *Remote Radio Head*

RRU *Remote Radio Unit*

SE *Spectral Efficiency*

SFN *System Frame Number*

SINR *Signal-to-Interference-plus-Noise Ratio*

SNR *Signal-to-Noise Ratio*

SRS *Sounding Reference Signal*

SS *Synchronization Signal*

SSS *Secondary Synchronization Signal*

SVD *Singular Value Decomposition*

TDD *Time-Division Duplex*

TTI *Transmission Time Interval*

UE *User Equipment*

VQ *Vector Quantization*

vRAN *Virtualized Radio Access Network*

ZF *Zero-Forcing*

Lista de Símbolos

β_{km} *Large scale fading*

γ_{km} Variância do erro de estimação do canal

Δf Espaçamento de subportadoras

ρ_d Potência máxima transmitida em cada AP

ρ_u Potência do *uplink* transmitida

σ_{est} Desvio padrão do ruído em cada AP receptor

σ_{shadow} Desvio padrão de *shadow*

σ_w Desvio padrão de ruído

τ Comprimento da sequência do piloto enviada

A' Pré-codificador linear

a' Elementos do pré-codificador

A Matriz de pré-codificação de limite de potência

b_{CSI} Número de bits para representar cada amostra *Channel State Information* (CSI)

B_{c} Constante de Boltzmann

b_{IQ} Número de bits para representar cada amostra *in-phase and quadrature* (IQ)

$b_{CSI,BFP}$ Número efetivo de bits por amostra CSI com *Block Floating Point* (BFP)

$b_{CSI,BS}$ Número efetivo de bits por amostra CSI com *Block Scaling* (BS)

B_{CSI} Número de bits para representar as amostras CSI com compressão

$BLER_T$ Taxa de erro de bloco alvo

C_{BW} Número de subportadoras na largura de banda de coerência

d_{km} Distância entre a antena *m* e o UE *k* em km

$\bar{G}$ Distorção da compressão entre os *Antenna Points* (APs) e *User Equipments* (UEs)

$\dot{G}$ Matriz de canal estimada entre os APs e UEs sem compressão

$\hat{G}$ Matriz de canal estimada entre os APs e UEs com compressão

$\tilde{G}$ Erro de estimação de canal entre os APs e UEs

G Matriz de canal entre os APs e os UEs

h_{km} *Small scale fading*

I Matriz identidade

K Número de UEs

M Número de APs

N_{PRB} Número máximo de *Physical Resource Blocks* (PRBs) por largura de banda

N_L Número de camadas

N_r Antenas receptoras

N_t Antenas transmissoras

N_{SC} Número de subportadoras do sinal multiportadora

NF *Noise Figure*

P Matriz diagonal com elementos de potência associados a cada símbolo dos UEs

p_i Elementos de potência associados a cada símbolo dos UEs

R Taxa de dados no *fronthaul* para transmitir IQ em cada *O-RAN Radio Unit* (O-RU)

R_{CSI} Taxa de transmissão do CSI para a *Central Processing Unit* (CPU)

R_a Taxa de dados no *fronthaul* agregada

R_k	*Taxa alcançável de downlink para o k-ésimo UE*
R_P	Taxa de pico no *fronthaul* no *uplink*
R_{IQ}^{DL}	Tráfego *fronthaul* necessário para transmitir no *downlink* amostras IQ para cada AP
R_{IQ}^{UL}	Tráfego *fronthaul* necessário para transmitir no *uplink* amostras IQ para cada AP
s	Símbolos *Orthogonal Frequency-Division Multiplexing* (OFDM) transmitidos
T_{sym}	Período de um único símbolo OFDM
T_{slot}	Período de um *slot*
W	Matriz de canal entre as antenas transmissoras e as *layers*
w	Vetor de ruído
X_{km}	Matriz de *shadow*
x	Símbolos transmitidos
y	Sinais de recepção

Lista de Figuras

2.1 Representação da tecnologia MIMO. 8

2.2 Eficiência espectral para o cenário de rede celular e para o cenário de D-MIMO [1]. 10

2.3 Sistema celular tradicional e MIMO distribuído. 11

2.4 Topologia de rede tradicional (superior) e topologia de rede *Distributed MIMO* (D-MIMO) (inferior) [2]. 12

2.5 Representação dos canais no cenário com *multiple-input multiple-output* (MIMO) distribuído. 15

2.6 Diagrama de representação do método *Zero-Forcing* [3]. 18

2.7 Representação esquemas duplex [4]. 19

2.8 A pilha de protocolos LTE/NR com camadas e subcamadas. 21

2.9 Estrutura do Bloco SS [4]. 22

2.10 Representação dos quadros, subquadros e slots em NR [4]. 26

2.11 Estrutura do CSI-RS [4]. 29

2.12 Representação do CSI-RS aplicado ao filtro espacial (F) antes de serem mapeados para antenas físicas [4]. 29

2.13 Representação da estrutura do SRS [4]. 30

2.14 Representação do SRS aplicado ao filtro espacial (F) antes de serem mapeados para antenas físicas [4]. 30

2.15 Representação do processamento da canal de transporte. 31

2.16 Mapeamento de transmissão multiantena [4]. 37

2.17 Tráfego *fronthaul uplink* para dados IQ (R_{IQ}^{UL}), dados CSI (R_{CSI}) e taxa agregada de pico (R_P) para cada *link* de *fronthaul* que conecta o AP a CU para diferentes números de UEs e com $N_{SC} = 1200$ subportadoras. 40

2.18 Tráfego *fronthaul uplink* para dados IQ (R_{IQ}^{UL}), dados CSI (R_{CSI}) e taxa agregada de pico (R_P) para cada *link* de *fronthaul* que conecta o AP a CU para diferentes números de UEs e com $N_{SC} = 600$ subportadoras. 40

2.19 Tráfego *fronthaul uplink* para dados IQ (R_{IQ}^{UL}), dados CSI (R_{CSI}) e taxa agregada de pico (R_P) para cada *link* de *fronthaul* que conecta o AP a CU para diferentes números de UEs e com $N_{SC} = 180$ subportadoras. 41

2.20 Tráfego *fronthaul uplink* para dados IQ (R_{IQ}^{UL}), dados CSI (R_{CSI}) e taxa agregada de pico (R_P) de todos os APs somadas para cada *link* de *fronthaul* que conecta o AP a CU para diferentes números de APs e com $N_{SC} = 1200$ subportadoras, considerando $K = 15$ UEs. 41

2.21 Tráfego *fronthaul uplink* para dados IQ (R_{IQ}^{UL}), dados CSI (R_{CSI}) e taxa agregada de pico (R_P) de todos os APs somadas para cada *link* de *fronthaul* que conecta o AP a CU para diferentes números de APs e com $N_{SC} = 600$ subportadoras, considerando $K = 15$ UEs. 42

2.22 Tráfego *fronthaul uplink* para dados IQ (R_{IQ}^{UL}), dados CSI (R_{CSI}) e taxa agregada de pico (R_P) de todos os APs somadas para cada *link* de *fronthaul* que conecta o AP a CU para diferentes números de APs e com $N_{SC} = 180$ subportadoras, considerando $K = 15$ UEs. 42

2.23 Arquitetura de compartilhamento de dados pelo *link ethernet*. 43

3.1 Representação do método de compressão BFP. 47

3.2 Representação do método de compressão BS. 51

3.3 Análise do NMSE de compressão dos métodos de compressão *Block Floating Point* (BFP) e *Block Scaling* (BS) para diferentes números de bits de compressão por amostra *Channel State Information* (CSI). 52

3.4 CDF da taxa de usuário alcançável para cada método de compressão e para $B_{CSI} = 16, 12$ e 8 bits considerando $M = 128$ APs. 53

3.5 CDF da taxa de usuário alcançável para cada método de compressão e para $B_{CSI} = 16, 12$ e 8 bits considerando $M = 96$ APs. 53

3.6 CDF da taxa de usuário alcançável para cada método de compressão e para $B_{CSI} = 16, 12$ e 8 bits considerando $M = 64$ APs. 54

3.7 CDF da taxa de usuário alcançável para cada método de compressão e para $B_{CSI} = 16, 12$ e 8 bits considerando $M = 32$ APs. 54

3.8 Taxa de usuário alcançável em cenários com diferentes números de UEs (superior) e taxa de usuário agregada alcançável em cenários com diferentes números de APs (inferior) sem compressão e com diferentes configurações de compressão. 56
3.9 Tráfego no *fronthaul* alcançável por *Antenna Point* (AP) em cenários com diferentes números de UEs (superior) e tráfego no *fronthaul* alcançável agregado em cenários com diferentes números de APs (inferior) sem compressão e com diferentes configurações de compressão. 57

4.1 Exemplo de Slot TDD NR composto por 14 símbolos OFDM, onde 9 símbolos são alocados para *downlink* e 4 para *uplink*. 61
4.2 Diagrama de sequência ilustrando as tarefas e sinalização na rede. 61
4.3 Estatísticas de taxa de dados individuais para diferentes números de *User Equipment* (UEs), considerando o cenário com MIMO distribuído sem compressão e com compressão. 67
4.4 Estatísticas de taxa de dados individuais para diferentes números de UEs, considerando o cenário com MIMO distribuído sem compressão e com compressão com b_{CSI} = 24 e 16 bits. 68

Lista de Tabelas

2.1 Termos do NR intercambiáveis. 14
2.2 Numerologia de Bloco SS e Faixas de Frequência [4]. 22
2.3 Mapeamento de Camadas por *Codeword* [5]. 34

4.1 Tráfego no *Fronthaul* por *O-RAN Radio Unit* (O-RU) para Várias Configurações de Bandwidth com b_{IQ} = 20 bits, b_{CSI} = 64 bits e K = 16 UEs. 64
4.2 Resumo dos parâmetros da simulação e taxa de dados no *fronthaul*. 66

Conteúdo

Agradecimentos **vi**

Lista de Acrônimos **ix**

Lista de Símbolos **xv**

Lista de Figuras **xvi**

Lista de Tabelas **xix**

Conteúdo **xx**

1 Introdução **1**

1.1 Motivações e Relevância 1

1.2 Objetivos 2

1.3 Trabalhos Relacionados 3

1.4 Contribuições 5

1.5 Publicação 6

1.6 Esboço da Dissertação 6

2 Fundamentos Teóricos **7**

2.1 Tecnologia MIMO 7

2.2 MIMO Distribuído 9

2.3 Pré-codificação Linear 13

2.3.1 Pré-codificador Zero-Forcing 16

2.4 Estrutura Geral de Transmissão do NR 17

2.4.1 Esquemas Duplex 18

2.4.2 Functional Split 19

2.4.3 Identificação da Rede 20
2.4.4 Acesso Aleatório 22
2.4.5 Transmissão de Tempo-Frequência 23
2.4.6 Estrutura no Domínio do Tempo 24
2.4.7 Portas de Antena 25
2.4.8 Sinais de Referência 26
2.4.9 Processamento do Canal de Transporte 30
2.4.10 Transmissão Multiantena 35
2.5 Modelo do Sistema e Tráfego no Fronthaul 37
2.5.1 Descrição do Sistema 37
2.5.2 Tráfego no Fronthaul 39

3 Esquemas de Compressão do O-RAN para D-MIMO 45
3.1 Aliança O-RAN 45
3.2 Métodos de Compressão em Blocos 46
3.2.1 Compressão Block Floating Point 46
3.2.2 Compressão Block Scaling 49
3.3 Resultados da Simulação com os Métodos de Compressão 49

4 Desempenho do D-MIMO em Sistemas 5G/NR 58
4.1 Modelo do Sistema 58
4.2 Compressão em Sistemas NR 64
4.2.1 Resultados Numéricos 65

5 Conclusão e Trabalhos Futuros 69
5.1 Conclusão 69
5.2 Trabalhos Futuros 70

Referências Bibliográficas 71

Abstract

In Distributed-MIMO (D-MIMO), a large number of distributed Antenna Points (APs) are coordinated by a Central Unit (CU) to serve a limited number of users with the same time/frequency resources, which brings improvements in spectral efficiency. The success of D-MIMO depends on precoding and power allocation, which can be performed completely centrally on the CU or distributed across APs. The centralized approach has greater spectral efficiency than the distributed implementation, but requires a significant spike in fronthaul traffic due to the exchange of Channel State Information (CSI) between APs and CU. In this work, CSI compression schemes are proposed to enable practical and centralized implementation of D-MIMO. It is shown that depending on the compression configuration, the spectral efficiency can be as good as in the case without compression. Furthermore, this work explores the implementation of multiple-input multiple-output (MIMO) within the framework of the New Radio (NR) architecture. The study evaluates a distributed MIMO deployment using NR signals with compression and evaluates its performance compared to the uncompressed scenario. Through simulations using the NR physical layer, the results also show that the spectral efficiency can be as good as in the uncompressed case depending on the compression configuration. Finally, the simulations with NR signals highlight important practical aspects and the feasibility of implementing D-MIMO in the 5G architecture and beyond 5G.

***Keywords* —** D-MIMO, cell-free, fronthaul, compression

Resumo

Em *Distributed-MIMO* (D-MIMO), um grande número de *Antenna Points* (APs) distribuídos são coordenados por uma *Central Unit* (CU) para atender um número limitado de usuários com os mesmos recursos de tempo/frequência, que traz melhorias na eficiência espectral. O sucesso do D-MIMO depende da pré-codificação e alocação de potência, onde podem ser realizadas de forma totalmente centralizada na CU ou distribuídas nos APs. A abordagem centralizada apresenta maior eficiência espectral do que a implementação distribuída, mas requer um pico significativo de tráfego *fronthaul* devido à troca de *Channel State Information* (CSI) entre APs e CU. Neste trabalho, são propostos esquemas de compressão CSI para permitir a implementação prática e centralizada de D-MIMO. É mostrado que dependendo da configuração de compressão, a eficiência espectral pode ser tão boa quanto no caso sem compressão. Além disso, este trabalho explora a implementação de *multiple-input multiple-output* (MIMO) dentro da estrutura da arquitetura *New Radio* (NR). O estudo avalia uma implantação MIMO distribuída usando sinais NR com compressão e avalia seu desempenho em comparação ao cenário sem compressão. Através de simulações utilizando a camada física NR, os resultados também mostram que a eficiência espectral pode ser tão boa quanto no caso sem compressão dependendo da configuração de compressão. Por fim, as simulações com sinais NR mostram aspectos práticos importante e a viabilidade de se implementar D-MIMO na arquitetura 5G e além do 5G.

Palavras-chave — D-MIMO, cell-free, fronthaul, compressão

Capítulo 1

Introdução

Este capítulo apresenta a pesquisa desenvolvida para esta dissertação. Primeiramente, descreve-se as motivações e a relevância deste trabalho (Seção 1.1), em seguida são destacados os objetivos deste trabalho (Seção 1.2). Num segundo momento são apresentados os trabalhos relacionados ao tema que será abordado (Seção 1.3). Por fim, são mostradas as contribuições da pesquisa (Seção 1.4), assim como publicação (Seção 1.5) e um resumo da dissertação (Seção 1.6).

1.1 Motivações e Relevância

Os sistemas 4G/5G/6G visam promover avanços significativos em telecomunicações, tais como: alta *performance* em termos de capacidade e links com ultra-alta taxa de dados. Além disso, essas tecnologias buscam prover eficiência energética, eficiência de transporte de dados, robustez, baixa latência, mobilidade, seguridade e flexibilidade de implantação. Para isso, as redes D-MIMO podem melhorar essas métricas oferecendo alto nível de macro-diversidade e serviços consistentes sobre a área de cobertura por meio da coordenação de *Remote Radio Units* (RRUs) [6].

Um sistema sem fio abrangente deve fornecer um serviço uniformemente bom em toda uma área designada. Esse é um problema enfrentado pelas redes de telecomunicações, pois alguns sistemas não conseguem cobrir toda a área designada, oferecendo menor qualidade de serviço aos usuários de borda. Para esse fim, o *massive MIMO* (mMIMO) atraiu considerável atenção como candidato para a tecnologia de camada física de quinta geração [7].

Os sistemas D-MIMO têm sido amplamente investigados nas últimas duas décadas e, re-

centemente, têm sido associados ao conceito de redes *cell-free*. Em ambos os casos, vários APs distribuídos atendem simultaneamente um número menor de UEs usando os mesmos recursos de tempo/frequência. Os APs são coordenados por uma *Central Unit* (CU) através de um link *fronthaul* e, ao realizar a pré-codificação e alocação de potência, a rede D-MIMO alcança maior eficiência espectral, confiabilidade e justiça entre os usuários [7]. As redes celulares típicas têm a desvantagem de aumentar a interferência entre células, particularmente quando um UE está localizado próximo aos limites das células [8]. Ao mesmo tempo, o D-MIMO permite uma melhor utilização de recursos através da implementação de pré-codificação e alocação de potência distribuída nos APs [7, 8, 9] ou centralizada na CU [10, 11, 12].

1.2 Objetivos

Há uma grande variedade de estratégias para realizar alocação de potência e cálculo de pré-codificação em D-MIMO [7, 8, 9, 10, 11, 12, 13]. Neste trabalho será avaliada a abordagem *Zero-Forcing* (ZF), que centraliza o cálculo da pré-codificação e alocação ótima de potência na CU através de um procedimento que requer CSI de curto prazo. A vantagem desta abordagem centralizada é a melhoria na eficiência espectral quando comparada às abordagens distribuídas [7]. Entretanto, o pré-codificador *Zero-Forcing* exige que os APs enviem a estimação do CSI para a CU a cada intervalo de coerência, o que impõe requisitos mais robustos ao tráfego *fronthaul* do uplink [12].

Neste trabalho, foram propostos métodos práticos de compressão do CSI com baixa complexidade que poderiam ser implementados em APs de baixo custo equipados com uma única antena. Além disso, o método proposto não depende de uma etapa de treinamento ou de suposições sobre o modelo do canal. Mais especificamente, os métodos propostos são versões modificadas dos métodos de compressão de blocos propostos na especificação *Open Radio Access Network* (O-RAN) para amostras *in-phase and quadrature* (IQ) [14]: BFP e BS. Esses métodos foram escolhidos com base na sua baixa complexidade e simplicidade de implementação. O desempenho dos novos métodos é comparado ao caso sem compressão, onde métricas de taxa de usuário alcançável e tráfego *fronthaul* são calculadas para cada caso. Eles não são comparados com algoritmos de compressão CSI anteriores para D-MIMO devido às diferenças no número de antenas.

É importante notar a diferença entre o CSI em sistemas MIMO convencionais com ante-

nas co-localizadas e o D-MIMO. No primeiro caso, o CSI normalmente exibe alta correlação espacial devido à distância física limitada entre as antenas no transmissor ou receptor. Esta propriedade pode ser explorada para atingir altos fatores de compressão usando várias técnicas [15]. Porém, em sistemas D-MIMO, todas as antenas de transmissão e recepção estão localizadas em posições geográficas distintas, dificultando a exploração de recursos de correlação ou métodos convencionais de compressão CSI, especialmente com algoritmos de baixa complexidade.

O presente trabalho também investiga uma implementação potencial de um sistema com antenas distribuídas centrado em células dentro da especificação *New Radio* (NR) existente. O estudo investiga resultados de simulação no contexto da implementação de 5G/NR com antenas distribuídas, abrangendo taxas de dados no *fronthaul* e oferecendo uma análise comparativa entre D-MIMO centrado em células sem compressão e com compressão em relação à taxa de transferência de rádio. A abordagem *cell-centric* foi escolhida neste trabalho pois apresenta benefícios em relação a outras tecnologias, como: melhoria na eficiência espectral e na qualidade de serviço, otimização da capacidade dos recursos disponíveis, escalabilidade e simples implementação, uma vez que não é necessária a cooperação entre CUs de células diferentes [1].

Pesquisas anteriores abordam desafios, como processamento de sinal para estimativa de canal e transmissão/recepção de dados, sinalização no *fronthaul* para compartilhamento de dados e CSI e otimização de controle de potência [1, 13]. Essa abordagem envolve o tratamento de problemas de processamento de sinal na camada física do NR, mantendo a independência da sinalização no *fronthaul* dos números de UE e compartilhamento do CSI, empregando um método de otimização de baixa complexidade para controle de potência.

Este trabalho avalia a implantação do D-MIMO utilizando uma simulação compatível com NR. A taxa de transferência de *downlink* de MIMO distribuído sem compressão e com compressão é avaliada – a taxa de transferência de *uplink* está fora do escopo. Até onde se sabe, a maioria dos trabalhos que avaliam D-MIMO ou *Cell-Free Massive Multiple-Input Multiple-Output* (CF-mMIMO) não implementam simulações com sinais NR.

1.3 Trabalhos Relacionados

Embora vários artigos na literatura discutam a alta sobrecarga causada pela estratégia *Zero-Forcing*, até onde sabemos, apenas alguns trabalhos propõem estratégias para diminuir o tráfego CSI no *fronthaul* do D-MIMO. Por exemplo, os autores de [16] utilizam *Vector Quanti-*

zation (VQ) para comprimir o CSI de um AP aproveitando a correlação entre suas antenas. No entanto, o método VQ necessita de treinamento *on-line* para funcionar bem quando o grau de correlação entre as antenas muda [16].

Em [12], os autores descrevem um método para decompor uma matriz de canal em componentes de *line-of-sight* (LoS) e *non-line-of-sight* (NLoS) usando um modelo de desvanecimento *Rician* entre usuários e AP com múltiplas antenas. O componente LoS é representado com um método baseado em modelo, enquanto o componente NLoS é compactado usando *Singular Value Decomposition* (SVD). O estudo constatou que este método tem baixo impacto na eficiência espectral de rádio, mas seu desempenho depende da validade da suposição do modelo Rician para o canal.

Em [17], o artigo discute um método proposto para comprimir CSI em AP com múltiplas antenas usando quantização uniforme simples com otimização de tamanho de passo. O artigo avalia o desempenho da eficiência espectral do *uplink* comparando o envio da versão quantizada do canal estimado com o sinal quantizado disponível na CU versus o envio apenas do sinal quantizado ponderado disponível na CU. Ambas as abordagens são semelhantes, e os autores propõem melhorias na otimização da taxa max-min considerando a restrição do tráfego *fronthaul*.

Os trabalhos mencionados acima consideram cenários com múltiplas antenas por AP e contam com uma etapa de treinamento [16, 17] ou suposições sobre o modelo de canal [12]. No atual trabalho, o desempenho do sistema não depende da etapa de treinamento e de suposição do canal. Além disso, vários trabalhos abordam questões importantes de escalabilidade no CF-mMIMO [1, 13, 18, 19, 20], mas raramente estabelecem uma conexão clara com sistemas práticos, como o padrão NR.

O trabalho de [1] está entre os primeiros esforços para resolver problemas de escalabilidade no CF-mMIMO, que consiste na capacidade do sistema expandir-se ou adaptar-se de maneira eficiente para suportar um número crescente de usuários, dispositivos e demandas de tráfego sem comprometer significativamente o desempenho ou a qualidade de serviço do sistema. Eles identificam desafios de escalonamento na formulação canônica do *cell-free*, onde cada AP atende a cada UE e uma única CU controla todos os APs. Para enfrentar esses desafios, eles propõem uma estrutura abrangente que abrange processamento de dados, topologia de rede e controle de potência, utilizando conceitos de *cluster* centrados em célula e centrados no usuário.

Um estudo subsequente [13] investiga a implementação escalável de CF-mMIMO à medida que o número de usuários se aproxima do infinito. Sua estrutura centrada no usuário garante a implementação de tarefas essenciais, como processamento de sinal para estimação de canal, recepção e transmissão de dados, sinalização *fronthaul* para compartilhamento de dados e CSI e otimização de controle de potência, mesmo em cenários com um número infinitamente crescente de usuários.

Com base nas conclusões de [1] e [13], os autores de [18] abordam um cenário de acesso massivo, onde o número de UEs é comparável ao número de APs. Sua proposta incorpora estratégias para acesso inicial, decodificação de dados, atribuição de piloto e controle de potência, levando em consideração fatores como CSI imperfeito, *Spectral Efficiency* (SE), densidade de usuários e justiça entre UEs.

A limitação de usar uma única CU para controlar todos os APs é um ponto focal em [19], onde um sistema multi-CU é proposto usando os conceitos de "*clusters* reais" e "*clusters* virtuais". Além disso, diferentes graus de cooperação entre essas múltiplas CUs são definidos e analisados.

Entre as contribuições mais recentes, [20] propõe um novo esquema de combinação/pré-codificação. Entre as contribuições mais recentes, propõe um novo esquema de combinação/pré-codificação, denominado *Improved Partial Minimum Mean Square Error* (IP-MMSE), para permitir avaliação de desempenho escalável de sistemas CF-mMIMO.

1.4 Contribuições

As contribuições mais relevantes desta pesquisa podem ser resumidos como:

- Investigação da redução de tráfego no *fronthaul* do *uplink* por meio da compressão de CSI para a CU.
- Implementação de métodos de compressão do CSI com baixo custo computacional e com bom desempenho para o sistema adotado.
- Comparação de desempenho em sistemas D-MIMO em cenários sem compressão e com compressão de CSI.
- Investigação de implementação dos sistemas D-MIMO em sistemas 5G/NR.

- Comparação de desempenho em sistemas 5G/NR com antenas distribuídas sem compressão e com compressão de CSI.

1.5 Publicação

- M. Silva, L. Ramalho, I. Almeida, E. Medeiros and A. Klautau, "CSI Compression for Distributed-MIMO With Centralized Precoding and Power Allocation,"in IEEE Communications Letters, vol. 27, no. 6, pp. 1535-1539, June 2023, doi: 10.1109/LCOMM.2023.3264944.

1.6 Esboço da Dissertação

O restante deste trabalho está organizado como descrito a seguir:

- **Capítulo 1:** Contém a introdução do trabalho, assim como motivação, objetivos trabalhos relacionados e contribuições.
- **Capítulo 2:** Aborda sobre os conceitos básicos usados no desenvolvimento da pesquisa.
- **Capítulo 3:** Destaca os esquemas de compressão do CSI adotados e os resultados obtidos com os esquemas de compressão.
- **Capítulo 4:** Explana sobre os resultados obtidos com a implementação dos sistemas D-MIMO em 5G/NR com a presença de compressão e sem compressão de CSI.
- **Capítulo 5:** Completa esta dissertação fornecendo conclusões gerais e sugestões para trabalhos futuros.

Capítulo 2

Fundamentos Teóricos

Este capítulo apresenta os principais conceitos básicos abordados nesta pesquisa. Primeiramente, o capítulo descreve aspectos sobre tecnologia MIMO (Seção 2.1) e características sobre o sistema D-MIMO (Seção 2.2). Além disso, são destacados alguns aspectos gerais da transmissão NR (Seção 2.4.5). Por fim, são mostradas algumas análises feitas correspondentes ao modelo do sistema adotado e tráfego no *fronthaul* (Seção 2.5).

2.1 Tecnologia MIMO

A tecnologia MIMO é um sistema avançado empregado em comunicações sem fio. Neste sistema, diversas antenas são agrupadas para transmitir e receber dados, visando minimizar erros, aumentar a velocidade de transmissão e aprimorar a capacidade das comunicações. Uma das vantagens primordiais do MIMO reside na sua habilidade de gerar várias versões do mesmo sinal. Isso implica em mais oportunidades para os dados circularem e alcançarem as antenas receptoras sem serem prejudicados por problemas como o desvanecimento do sinal. A Figura 2.1 ilustra um cenário MIMO com antenas co-localizadas.

O desvanecimento representa uma condição em que a intensidade e a qualidade de um sinal de rádio variam ao longo do tempo e da distância. Diversos fatores, como propagação de múltiplos caminhos, condições atmosféricas e movimentação de objetos no percurso de transmissão, podem desencadear esse fenômeno. A tecnologia MIMO oferece uma solução para superar os desafios do desvanecimento, melhorando significativamente a *Signal-to-Noise Ratio* (SNR). Isso culmina em uma comunicação mais estável, segura e eficiente, beneficiando tanto os usuários quanto as aplicações dependentes de uma conexão sem fio robusta e de alta

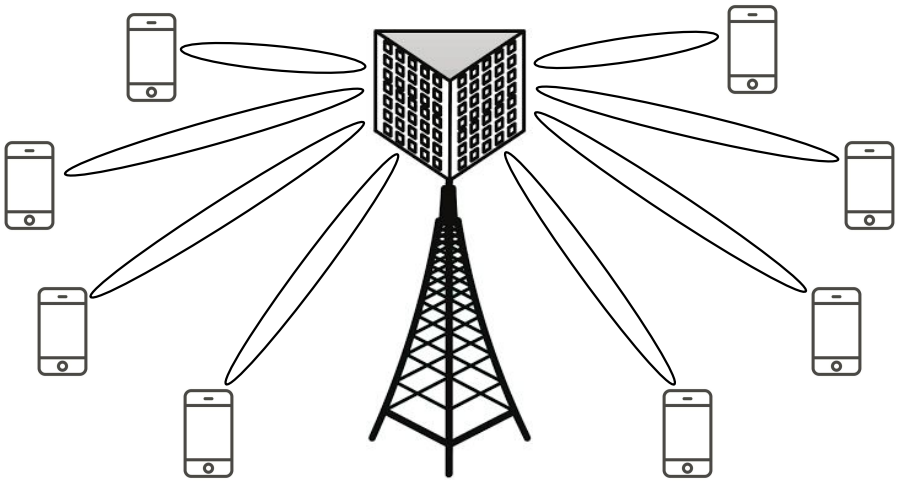


Figura 2.1: Representação da tecnologia MIMO.

qualidade [21].

As redes de telefonia móvel atuais operam por meio da criação de células autônomas, mas essa abordagem tem limitações devido à cobertura desigual que não atendem muitos usuários. Em um mundo cada vez mais dependente das comunicações sem fio, é imperativo ter uma rede celular com vários pontos de acesso que seja altamente eficiente em termos de eficiência espectral e potência. Para atender a essas necessidades, surge a tecnologia mMIMO. Esta inovação não apenas é adequada para as demandas atuais das comunicações sem fio, mas também tem o potencial de solucionar os problemas de interferência que são prevalentes no cenário atual [22].

Essa inovação tecnológica abriu novas possibilidades para o desenvolvimento e implementação de recursos espaciais em sistemas de comunicação móvel. Ao introduzir a utilização de recursos tridimensionais, que envolvem espaço, tempo e frequência, o design dos sistemas pode agora superar as limitações dos recursos bidimensionais de tempo e frequência. Isso implica que os sistemas não estão mais restritos apenas às dimensões tradicionais de tempo e frequência, mas também podem explorar o espaço, resultando em uma melhoria significativa na capacidade e eficiência das comunicações sem fio multiusuário. Esse avanço representa um salto importante no campo das comunicações móveis, permitindo uma experiência de usuário mais eficaz e confiável.

Essas características tridimensionais não apenas ampliam as capacidades técnicas, mas também estimulam a criação de novas teorias e abordagens na área das comunicações sem fio. Ao integrar esses componentes espaciais, os sistemas conseguem melhorar significativamente

a eficiência espectral e aumentar de forma notável a capacidade total do sistema [23]. Esse avanço representa um marco importante no campo das comunicações móveis, proporcionando um desempenho mais sólido e uma experiência aprimorada para os usuários finais.

2.2 MIMO Distribuído

Os sistemas MIMO desempenham um papel crucial no progresso das redes móveis, possibilitando um aumento considerável na velocidade de transmissão. Isso ocorre ao empregar várias antenas tanto no transmissor quanto no receptor. No entanto, canais que operam nessas frequências mais altas enfrentam desafios como atenuação e desvanecimento do sinal, o que precisa ser gerenciado para garantir uma comunicação eficaz.

Para superar esses desafios, uma solução inovadora foi criada: a conexão simultânea de um UE a várias antenas espalhadas geograficamente. Essa estratégia proporciona ao UE diversas chances de se comunicar com o sistema, mesmo em locais sujeitos a interferências e variações de sinal. Esse sistema, onde um UE pode se ligar a várias antenas simultaneamente, é denominado D-MIMO.

Esta abordagem não só aprimora a estabilidade e confiabilidade da conexão, mas também aumenta consideravelmente a capacidade e eficiência da rede móvel. Ao possibilitar que os dispositivos aproveitem vários caminhos de transmissão, o D-MIMO revoluciona as redes móveis, proporcionando uma experiência de usuário mais estável e eficaz, mesmo em ambientes de transmissão adversos.

O D-MIMO, também conhecido como CF-mMIMO na literatura, traz diversas melhorias para a rede. Ele impacta positivamente a taxa de usuário alcançável, eficiência espectral, eficiência energética, confiabilidade, escalabilidade e redução da interferência [7, 9, 24]. Nesse cenário, um grande número de APs atende a um determinado grupo de UEs, possibilitando maior multiplexação espacial, diversidade de canais, escalabilidade, propagação favorável e uma melhoria geral na qualidade do canal [1, 9]. Além disso, a propagação favorável contribui para mitigar interferências entre usuários, tornando eficazes esquemas de pré-codificação linear como *Zero-Forcing*, *Conjugate Beamforming* (CB) e *Minimum Mean Squared Error* (MMSE).

A ilustração na Figura 2.2 destaca esses efeitos mostrando a eficiência espectral em uma área coberta por 9 APs, revelando uma variação significativa no desempenho entre as redes celulares convencionais e as redes D-MIMO, ambas configuradas com a mesma distribuição

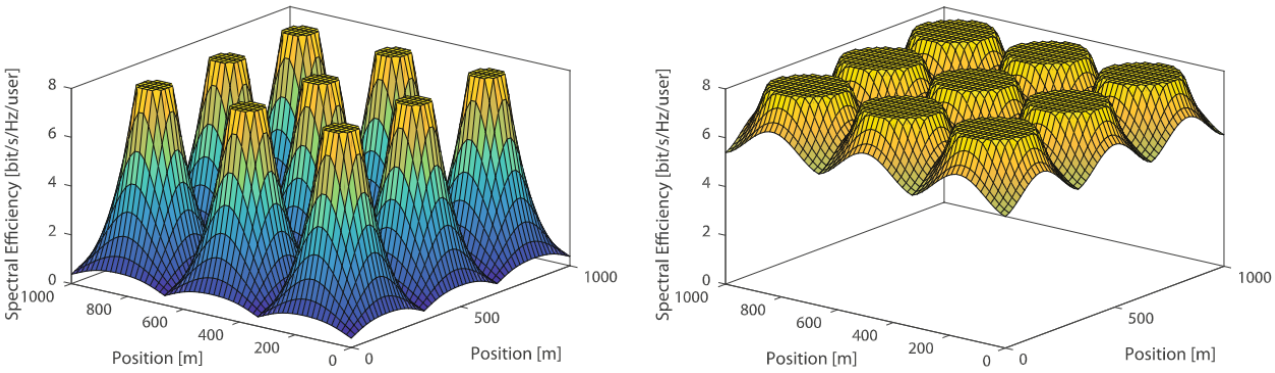


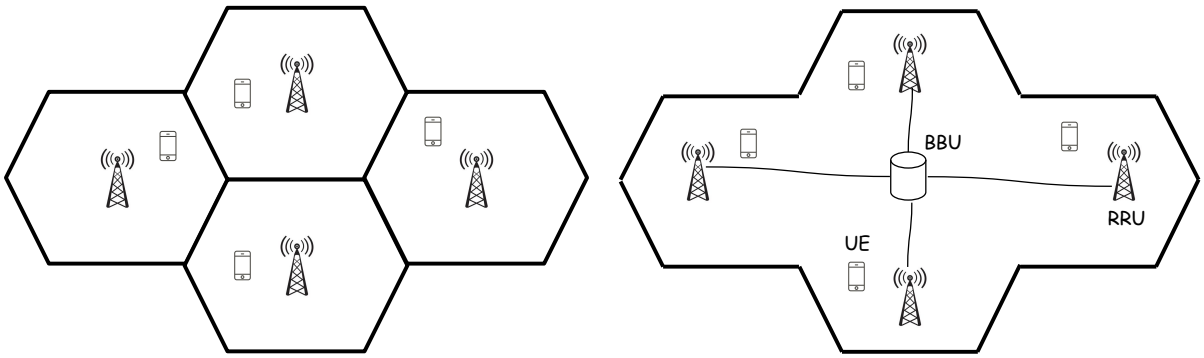
Figura 2.2: Eficiência espectral para o cenário de rede celular e para o cenário de D-MIMO [1].

de APs. À esquerda, a rede celular exibe um desempenho inferior nas bordas das células devido à interferência intensa entre elas. Em contraste, à direita, a rede D-MIMO demonstra um desempenho superior, já que consegue mitigar consideravelmente esse problema utilizando pré-codificação linear.

As unidades de antena em sistemas D-MIMO estão distribuídas em várias localidades geográficas e podem ser empregadas tanto em sistemas de estação base unicelular quanto em sistemas multicelulares, criando assim um sistema de comunicação móvel sem restrições de células. Esse sistema oferece serviços para todos os usuários utilizando os mesmos recursos de tempo e frequência, sem a necessidade de planejamento de frequência entre as células. Os recursos do sistema podem ser dinamicamente alocados em todas as dimensões, aumentando significativamente a flexibilidade na distribuição de recursos do sistema e otimizando a utilização dos recursos sem fio [23].

A Figura 2.3 ilustra a comparação entre um sistema de comunicação móvel tradicional e um sistema de comunicação distribuído. No cenário tradicional, diferentes recursos de espectro são atribuídos a um grupo de células próximas para prevenir interferências e melhorar a capacidade do sistema. Esse grupo, conhecido como *cluster*, é formado por células vizinhas que compartilham entre si o espectro disponível. Já no sistema de comunicação distribuído, a abordagem é diferente, permitindo uma alocação mais dinâmica e flexível dos recursos de espectro para otimizar o desempenho da rede.

Nas redes celulares tradicionais, cada AP é designado para cobrir uma área específica, e os UEs são conectados ao AP que abrange essa área. Em contraste, na tecnologia D-MIMO, toda a área de cobertura pode ser tratada como uma única célula, com vários APs distribuídos geograficamente e capazes de atender simultaneamente todos os UEs [7, 8, 9]. As diferenças



(a) Sistema Celular Tradicional **(b)** MIMO distribuído

Figura 2.3: Sistema celular tradicional e MIMO distribuído.

entre as topologias das redes tradicionais e do D-MIMO são detalhadas na Figura 2.4. A Figura superior representa a topologia do cenário de rede tradicional, enquanto a Figura inferior representa a topologia do cenário D-MIMO. Na abordagem D-MIMO, os APs são conectados à CU através do *link* denominado *fronthaul* e as CUs são conectadas entre si através do *link backhaul*.

Outra abordagem das redes D-MIMO consiste no cenário *user centric* que define explicitamente um *cluster* de serviço para cada usuário, limitando assim o número de transmissores de serviço. Diferentemente da abordagem destacada no parágrafo anterior na qual o usuário pode ser servido conjuntamente por todos os transmissores da rede [25].

Uma distinção entre sistemas MIMO *cell free* centrados no usuário e sistemas CF-mMIMO é que o primeiro define explicitamente um *cluster* de serviço para cada usuário, limitando assim o número de transmissores de serviço, enquanto o último assume que, teoricamente, o usuário pode ser servido conjuntamente por todos os transmissores da rede [1]. No entanto, na prática, mesmo este último define implicitamente um *cluster* de serviço devido à perda de percurso. Outra distinção é a suposição da aplicabilidade das propriedades de canais MIMO massivos, especificamente, a aplicabilidade teórica de propagação favorável e endurecimento de canal, que será discutida detalhadamente posteriormente. Portanto, não está claro se o termo "massivo" deve ser usado/descartado quando se refere ao esquema CF-mMIMO centrado no usuário com um *cluster* de serviço limitado.

A arquitetura dos sistemas celulares tradicionais é transformada pelo MIMO distribuído. Nesse novo modelo, não há uma forma padrão para se implementar essa tecnologia. Alguns trabalhos na literatura mostram diferentes formas de implementação visando melhorias de escala-

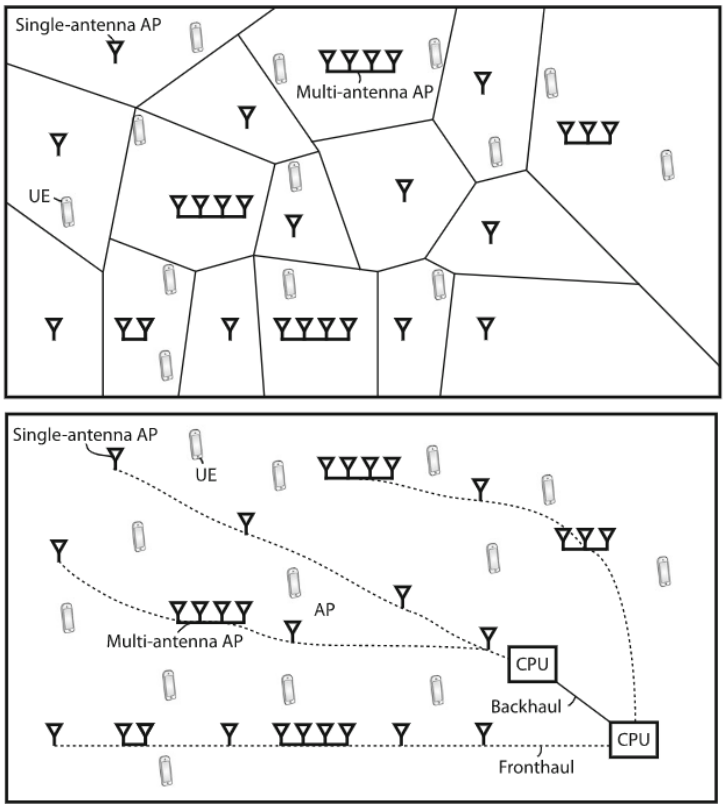


Figura 2.4: Topologia de rede tradicional (superior) e topologia de rede D-MIMO (inferior) [2].

bilidade, baixa complexidade computacional e diminuição de requisitos no *fronthaul* [1, 26, 27]. De modo geral, o AP permanece separado do sistema de comunicação móvel e é conhecido como RRU. As RRUs são conectadas à *Baseband Unit* (BBU) (CU) por meio de fibras ópticas ou ethernet, permitindo o processamento conjunto de sinais na BBU. Esse arranjo possibilita o suporte a múltiplas antenas distribuídas e, consequentemente, o atendimento de múltiplos usuários nos mesmos recursos de tempo e frequência, aumentando assim a eficiência espectral do sistema. Além disso, essa abordagem melhora significativamente a cobertura do sistema e o desempenho na borda da célula [28].

O método tradicional de distribuição de recursos de frequência é conhecido por ser estático, uma vez que a maior parte dos recursos é gerenciada diretamente na estação base. Apesar de sua simplicidade, essa abordagem estática limita a flexibilidade na alocação de recursos em sistemas de comunicação, como mencionado em [23].

O sistema D-MIMO não segue o princípio convencional de cobertura celular. Uma das estratégias em D-MIMO adota uma abordagem focada no usuário por meio do *Dynamic Cooperation Clustering* (DCC) de APs. Esses agrupamentos de APs proporcionam aos usuários

um desempenho estável, garantindo diversidade geográfica em distâncias curtas. Além disso, as interferências indesejadas podem ser controladas por meio da técnica conhecida como pré-codificação linear, que envolve a colaboração entre os APs [28].

Nesse trabalho, os termos CF-mMIMO e D-MIMO serão intercambiáveis, contudo o termo D-MIMO será priorizado. O termo "*cell-free*" no âmbito da comunicação sem fio não implica a ausência de células ou estações base. Pelo contrário, ele denota uma abordagem inovadora na organização e gestão de redes sem fio. Em um sistema massivo MIMO *cell-free*, todos os pontos de acesso colaboram para atender simultaneamente todos os usuários, rompendo com o conceito de fronteiras fixas de células.

Embora o termo "*cell-free*" seja utilizado para descrever este conceito, pode ser enganador porque ainda existem células (no sentido de áreas de cobertura) e estações base (na forma de pontos de acesso). A utilização do termo "*cell-free*" pode criar confusão ou mal-entendidos sobre o funcionamento fundamental da tecnologia. Dessa maneira, o termo D-MIMO foi escolhido por padrão neste trabalho.

Além disso, outros termos também serão intercambiáveis, como *Antenna Point* em relação à AP e CU em relação à *Central Processing Unit* (CPU). O motivo de considerarmos nesse trabalho *Antenna Point* é devido à generalização no modelo de comunicação estabelecido pelo *fronthaul*. *Access Point* pode ser considerado como um dispositivo que se comunica via *wireless*, contudo, no sistema D-MIMO pode haver outras formas de comunicação. O termo CU é considerado nesse trabalho de forma mais relevante, pois esse termo referencia o componente de processamento geral do sistema, enquanto o termo CPU pode fazer referência a termos mais específicos como componente primário do computador que realiza processo que pode estar relacionado com o *hardware* e o processamento do computador. Os termos mencionados no texto estão destacados na Tabela 2.1.

2.3 Pré-codificação Linear

A pré-codificação linear é uma estratégia eficaz para mitigar interferências em sistemas MIMO. Nesse método, o sinal transmitido é processado de maneira a maximizar a qualidade do sinal recebido por receptores e antenas específicos, ao mesmo tempo em que minimiza a interferência para os demais receptores e antenas. No contexto de D-MIMO, há duas abordagens bem conhecidas na literatura, a centralizada e a distribuída. A primeira faz uso das informações de

Tabela 2.1: Termos do NR intercambiáveis.

Literatura	Adaptado
Cell-Free (CF-mMIMO)	MIMO distribuído (D-MIMO)
Access Point	Antenna Point
CPU	CU

CSI no transmissor para otimizar o desempenho e aumentar a eficiência espectral. No entanto, para implementar a pré-codificação por meio de algumas estratégias, como o *zero-forcing*, é necessário que todos os APs compartilhem suas informações de CSI locais com um processador central através de links de *fronthaul* com capacidade limitada, o que consome consideráveis recursos de largura de banda [12].

Por outro lado, no sistema D-MIMO com processamento distribuído, uma CU está presente, mas a comunicação entre os APs e a CU está restrita às informações essenciais, como dados de carga útil e coeficientes de controle de potência, que mudam lentamente. Não há compartilhamento instantâneo de informações sobre o CSI entre os APs ou com a unidade central. A estimação de todos os canais é realizada nos APs por meio de pilotos de *uplink*. Essas estimações de canal são então utilizadas para pré-codificar os dados enviados no *downlink* e para detectar os dados recebidos no *uplink*. A técnica CB, também conhecida como *Maximum Ratio Transmission* (MRT), é um exemplo de pré-codificação usualmente implementada presente na literatura que utiliza a abordagem distribuída [8].

A Figura 2.5 apresenta um modelo da tecnologia MIMO. Nesse contexto, quando as N_t antenas do transmissor estão localizadas em diferentes posições geográficas, elas coletivamente formam o que é chamado de MIMO distribuído. O CSI de um sistema MIMO pode ser representado por uma matriz $N_r \times N_t$. Se essa matriz for conhecida e for invertível, as interferências entre os canais sem fio podem ser completamente eliminadas, resolvendo assim as equações lineares associadas. De modo geral, os símbolos recebidos podem ser calculados como

$$\mathbf{y} = \mathbf{G}\mathbf{x} + \mathbf{w}, \tag{2.1}$$

onde os símbolos transmitidos são denotados por $\mathbf{x} \in \mathbb{C}^{N_t \times 1}$ e podem ser recuperados utilizando os sinais de recepção $\mathbf{y} \in \mathbb{C}^{N_r \times 1}$. A matriz de canal é denotada por $\mathbf{G} \in \mathbb{C}^{N_r \times N_t}$ e $\mathbf{w} \in \mathbb{C}^{N_r \times 1}$ é o vetor de ruído, onde $\mathbf{w} \sim \mathcal{N}_{\mathbb{C}}(0, \mathbf{I}_{N_r}\sigma_w^2)$. Essa tecnologia MIMO introduz a multiplexação

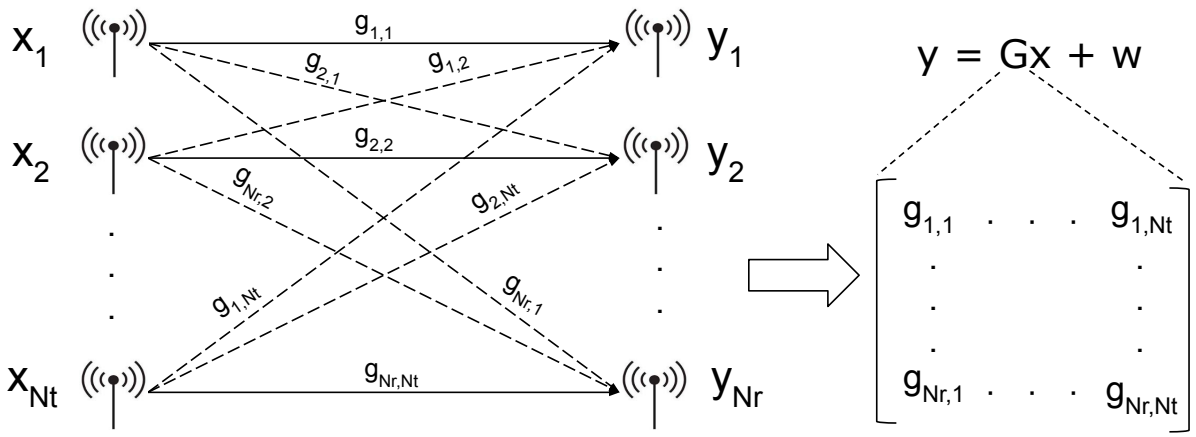


Figura 2.5: Representação dos canais no cenário com MIMO distribuído.

espacial da comunicação sem fio. Se a matriz do canal sem fio for invertível, então a taxa de transmissão e a capacidade do sistema aumentarão linearmente com o número de antenas, melhorando dessa forma a eficiência espectral do sistema de comunicação móvel [28].

O D-MIMO aproveita o poder do processamento conjunto para suprimir interferências entre antenas vizinhas usando equações lineares. Em uma área de cobertura do usuário, cada AP pode usar a mesma frequência, eliminando interferências por meio do processamento conjunto entre todas as antenas. Em contrapartida, o sistema celular convencional só consegue eliminar interferências entre antenas alocando recursos de frequência diferentes, já que as estações base não conseguem processar em conjunto.

A pré-codificação é uma etapa importante no processamento de sinais no sistema MIMO, onde os sinais são combinados de maneira predefinida e entregues na proporção adequada aos diversos elementos de antena. Em sistemas mMIMO, esses algoritmos desempenham um papel fundamental. O mMIMO com pré-codificação oferece a capacidade de transmitir diferentes sinais por meio de cada elemento da antena, com ganho e ajuste de fase adequados. Para alcançar isso, os pré-codificadores digitais são projetados de forma a criar múltiplos feixes principais, permitindo assim o foco simultâneo em vários usuários [29].

A pré-codificação em mMIMO simplifica a transmissão de dados para vários usuários, ajustando cuidadosamente os sinais de cada antena. O objetivo é calcular um vetor de peso que maximize o desempenho do receptor. A complexidade do pré-codificador aumenta com o nú-

mero de antenas, então é crucial equilibrar complexidade e desempenho ao escolher algoritmos para sistemas MIMO massivos [29].

2.3.1 Pré-codificador Zero-Forcing

Um dos métodos de pré-codificação linear utilizado na literatura é o *Zero-Forcing*, que tem recebido bastante atenção devido às dificuldades computacionais inerentes às estratégias de pré-codificação linear. O *Zero-Forcing* é uma abordagem simples que decompõe o canal multiusuário em subcanais independentes, tratando o problema como uma alocação de potência. Ele se destaca em situações de alta SNR ou quando há um grande número de usuários, proporcionando graus completos de liberdade. No contexto do ZF, dois critérios de projeto comuns são a maximização da justiça e do desempenho. Além disso, devido à sua simplicidade, o método *Zero-Forcing* também se tornou uma escolha atrativa para transmissões em sistemas MIMO [30].

No sistema D-MIMO, um AP é controlado por uma CU por meio de uma conexão chamada *fronthaul*. Esta CU é encarregada de calcular a pré-codificação linear e determinar a distribuição ideal de potência com base no conhecimento do CSI. A abordagem *Zero-Forcing* concentra as responsabilidades de pré-codificação e alocação de potência na CU. Esse processo exige informações de CSI de curto prazo, o que coloca exigências mais rigorosas no tráfego de *uplink* transmitido através da conexão *fronthaul*.

Em termos gerais, o projeto do transmissor é concebido considerando uma limitação total de potência. Na prática, há um interesse crescente em lidar com restrições individuais de potência por antena. Para isso, métodos de alocação de potência são usados com o intuito de otimizar o desempenho da rede e maximizar a eficiência espectral. A alocação eficiente de potência é crucial para garantir que todos os usuários sejam atendidos com a qualidade de serviço adequada, evitando interferências desnecessárias e maximizando a utilização dos recursos disponíveis. Isso pode ser alcançado ajustando dinamicamente a potência de transmissão em cada antena para atender às demandas de cada usuário, minimizando a interferência entre eles e otimizando o uso do espectro de frequência disponível. Além disso, os métodos de alocação de potência também são importantes para lidar com as características dinâmicas do ambiente de comunicação sem fio, como variações na demanda de tráfego, mudanças na topologia da rede e interferências externas. Esses métodos permitem que o sistema se adapte rapidamente às mudanças nas condições da rede, garantindo um desempenho consistente e confiável. Nestes

sistemas, cada transmissor estará sujeito às suas próprias limitações de potência [30].

O projeto de pré-codificação *Zero-Forcing* está estreitamente ligado ao conceito de inversos generalizados na área da álgebra linear. Essa relação é compreensível, pois o pré-codificador *Zero-Forcing* essencialmente reverte o canal multiusuário. Trabalhos anteriores que lidaram com limitações de potência total e restrições individuais de potência por antena partiram da premissa de que o pré-codificador segue uma forma específica de inverso generalizado conhecida como pseudo-inverso. O pré-codificador baseado em pseudo-inverso é considerado ideal entre os inversos generalizados para maximizar qualquer métrica de desempenho sob uma restrição de potência total. No entanto, quando estão envolvidas as restrições de potência por antena, essa abordagem não é mais ótima e outros tipos de inversos podem superá-la. Encontrar a matriz ideal torna-se uma tarefa não trivial e depende do critério de desempenho específico em questão [30].

Com a abordagem *Zero-Forcing*, tipicamente, no início de cada intervalo de coerência, os K UEs transmitem sinais ortogonais aos APs para estimar as condições do canal. Posteriormente, cada AP comunica as informações do canal estimadas para a CU. A CU, por sua vez, calcula os parâmetros de pré-codificação e a alocação de potência, realiza a pré-codificação dos símbolos e, em seguida, transmite os símbolos pré-codificados de volta para cada AP. Finalmente, os APs transmitem os símbolos pré-codificados para os UEs. Esse processo de pré-codificação de símbolos, modulação de frequência e transmissão sem fio é repetido até o próximo intervalo de coerência [3]. O diagrama que ilustra esse método está representado na Figura 2.6.

Uma das principais desvantagens de um sistema *Zero-Forcing* completamente centralizado é a necessidade de uma grande quantidade de sinalização entre os APs e a CU. No entanto, nenhuma sinalização é exigida entre os próprios APs.

2.4 Estrutura Geral de Transmissão do NR

As seções anteriores deste capítulo detalharam diversos aspectos da estrutura do D-MIMO descritos na literatura, sendo estes fundamentais para a implementação da rede, tanto nas tecnologias *Long Term Evolution* (LTE) quanto NR. Agora, dedicaremos esta seção a abordar elementos significativos da estrutura NR, que ampliarão a compreensão sobre como a implementação do D-MIMO pode ser efetuada na tecnologia NR.

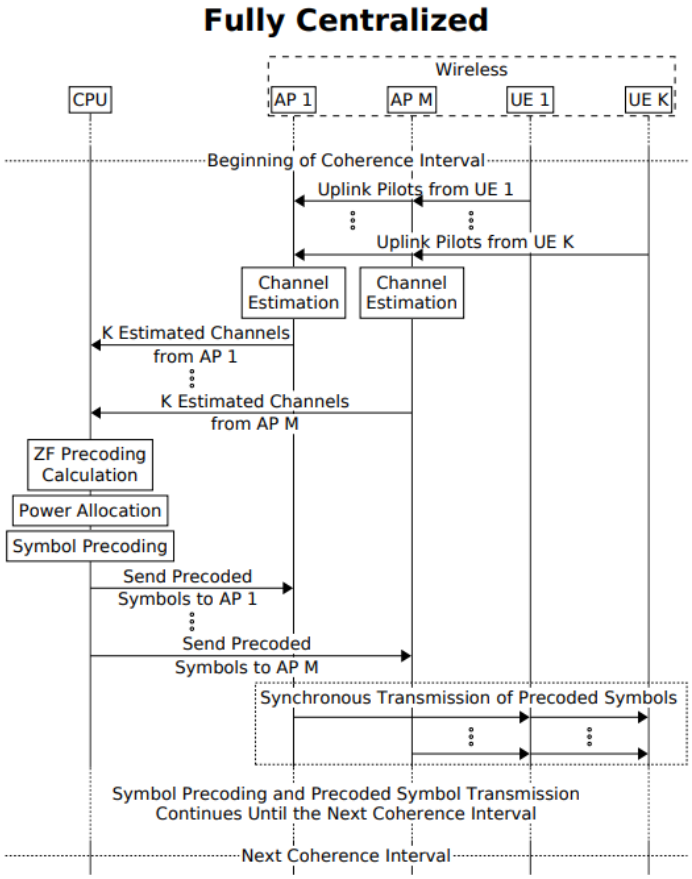


Figura 2.6: Diagrama de representação do método *Zero-Forcing* [3].

2.4.1 Esquemas Duplex

A estrutura básica do NR oferece suporte para a separação de *uplink* e *downlink* em termos de tempo e/ou frequência, podendo operar em modos *half-duplex* ou *full-duplex*, utilizando uma estrutura de quadro única. Existem três esquemas de duplex: *Time-Division Duplex* (TDD), *Frequency-Division Duplex* (FDD) e FDD *half-duplex* [4].

No modo TDD, as transmissões de *uplink* e *downlink* compartilham a mesma frequência portadora, sendo separadas apenas no tempo. Por outro lado, no modo FDD, as transmissões de *uplink* e *downlink* utilizam frequências diferentes, permitindo que ocorram simultaneamente. Além disso, existe o FDD *half-duplex*, onde as transmissões de *uplink* e *downlink* são separadas tanto em frequência quanto em tempo. Esse modo é especialmente adequado para dispositivos mais simples que operam em espectros emparelhados.

A Figura 2.7 oferece uma representação mais detalhada dos diferentes modelos esquemáticos duplex [4, 31]. Essa flexibilidade nos esquemas de *duplex* permite a adaptação dos

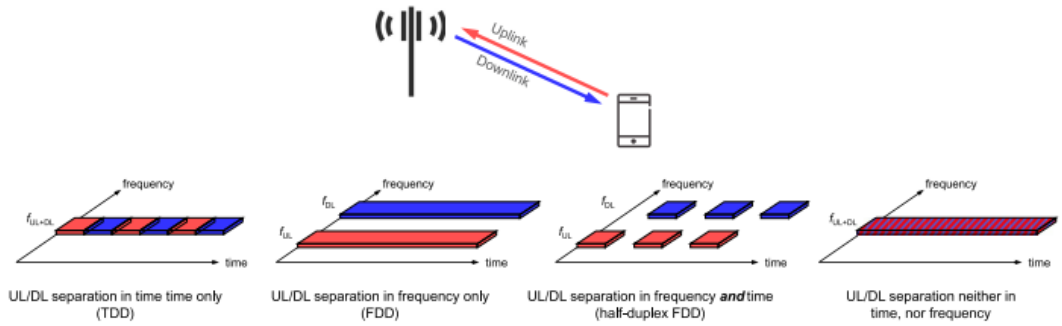


Figura 2.7: Representação esquemas duplex [4].

sistemas NR de acordo com as necessidades específicas de comunicação, proporcionando eficiência e confiabilidade nas transmissões de dados [4].

Nas redes D-MIMO é amplamente adotada a tecnologia TDD devido à sua flexibilidade e características promissoras. Nas redes TDD, a estação base pode usar a reciprocidade do canal e a partir da estimação do canal do *uplink* consegue obter informação sobre o canal de *downlink*, assim, construir o *precoder* sem precisar da recepção do canal do próprio UE. Uma característica chave do TDD é a capacidade de acomodar vários padrões de tráfego *uplink* e *downlink* em redes celulares (o que não é tão fácil em um sistema FDD) [32]. Ou seja, essa tecnologia possui maior flexibilidade no uso do espectro de frequências, pois o TDD permite uma maior flexibilidade na alocação de recursos de tempo para transmissão e recepção. Isso pode ser benéfico em ambientes dinâmicos, nos quais as condições de rede mudam de forma mais abrupta. Outro aspecto consiste no menor *overhead* de controle quando comparado ao FDD, pois a alocação de tempo pode ser gerenciada de forma mais eficiente sem a necessidade de faixas de frequências separadas.

2.4.2 Functional Split

Nas redes móveis 3G, foi introduzida a ideia de separar a estação rádio base em duas unidades, o *Remote Radio Head* (RRH) e a BBU. O RRH continha apenas as funções de rádio e estava localizado próximo à antena na torre da célula e a BBU continha todas as funções de processamento de banda base. A conexão entre o RRH e a BBU ocorre por meio do link denominado *fronthaul*. No NR, o tráfego está crescendo continuamente, por esse motivo, pesquisadores estão trabalhando em novas formas para reduzir as taxas de bits no link *fronthaul*. Uma possibilidade consiste em incluir mais funções localmente nos sites e processar mais os

sinais antes de eles serem transmitidos pelo *fronthaul*. O *functional split* determina a quantidade de funções realizadas no RRH e na BBU [33].

O *3rd Generation Partnership Project* (3GPP) propôs oito opções de *functional split*, incluindo várias subopções. A pilha de protocolos do 3GPP é dividida em *Distributed Unit* (DU) e CU. As funções presente na DU estão mais próximas do UE, pois estarão próximas das antenas e as funções localizadas na CU se beneficiarão da centralização do processamento. Quanto mais funções estiverem presentes na DU, mais processamento já foi feito antes que os dados sejam transmitidos na rede *fronthaul* e menor será a taxa de bits na rede *fronthaul*. No O-RAN, a DU é dividida em duas unidades: *O-RAN Distributed Unit* (O-DU) e O-RU, sendo que a O-RU tem funções mais próximas do rádio [33].

A descrição das divisões funcionais segue a pilha de protocolos LTE e NR conhecida das BSs tradicionais. A parte inferior desta pilha de protocolos inclui três camadas; mais baixa é a camada física, depois segue a camada de enlace de dados e no topo está a camada de rede. As funções implementadas nas diferentes camadas estão ilustradas na Figura 2.8.

Atualmente na literatura, a opção 7-2 tem sido bastante utilizada como ponto de divisão entre a CU e DU. No contexto do O-RAN, a DU ainda pode ser dividida entre O-DU e O-RU, sendo o *functional split* 7.2 usado no O-RAN para dividir as tarefas entre O-DU e O-RU. A divisão Low PHY/High PHY é a abordagem mais aceitável, pois é menos complexa e suporta vários requisitos de *fronthaul* e, o mais importante, tem altos benefícios de virtualização. Esta divisão foi otimizada ainda mais pela O-RAN Alliance em duas variantes: divisão 7.2a e divisão 7.2b. O Split 7.2x vem com técnicas de compressão *fronthaul*, como compressão e descompressão de IQ BFP, para reduzir ainda mais a largura de banda de transporte [33].

2.4.3 Identificação da Rede

O UE precisa realizar a identificação da rede na qual ele vai se conectar. A pesquisa de células é realizada quando um dispositivo entra inicialmente na área de cobertura de um sistema. Para permitir que os dispositivos encontrem uma célula ao entrar no sistema, bem como encontrar novas células ao se moverem dentro do sistema, um sinal de sincronização composto por duas partes, o *Primary Synchronization Signal* (PSS) e o *Secondary Synchronization Signal* (SSS), deve ser utilizado, os quais são transmitidos periodicamente no *downlink* de cada célula NR. Esses sinais, juntamente com o *Physical Broadcast Channel* (PBCH), são referidos como um bloco de *Synchronization Signal* (SS) [4].

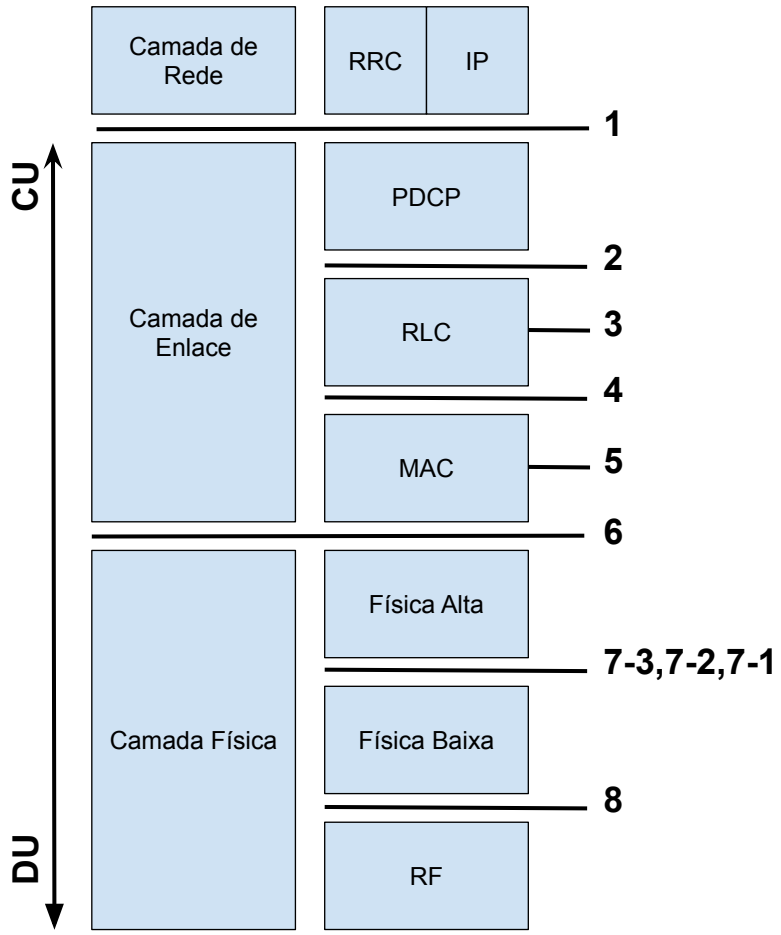


Figura 2.8: A pilha de protocolos LTE/NR com camadas e subcamadas.

Há algumas características que diferenciam os sinais de sincronização do LTE para o do NR. Uma dessas características consiste na redução da quantidade de sinais "sempre ligados" e a possibilidade de formação de feixe durante o acesso inicial. As transmissões do bloco SS é baseada em *Orthogonal Frequency-Division Multiplexing* (OFDM). A estrutura do bloco é mostrada na Figura 2.9. O bloco SS abrange quatro símbolos OFDM no domínio do tempo e 240 subportadoras no domínio da frequência [4].

Há diferentes numerologias que podem ser usadas para transmissão de blocos SS. Para limitar a necessidade de dispositivos procurarem simultaneamente blocos SS de diferentes numerologias, na maioria dos casos existe apenas uma única numerologia de bloco SS definida para uma determinada banda de frequência. A Tabela 2.2 mostra as diferentes numerologias aplicáveis à transmissão do bloco SS, juntamente com a largura de banda do bloco SS e a duração do tempo correspondentes. A vantagem de ser usada numerologia mais alta consiste na

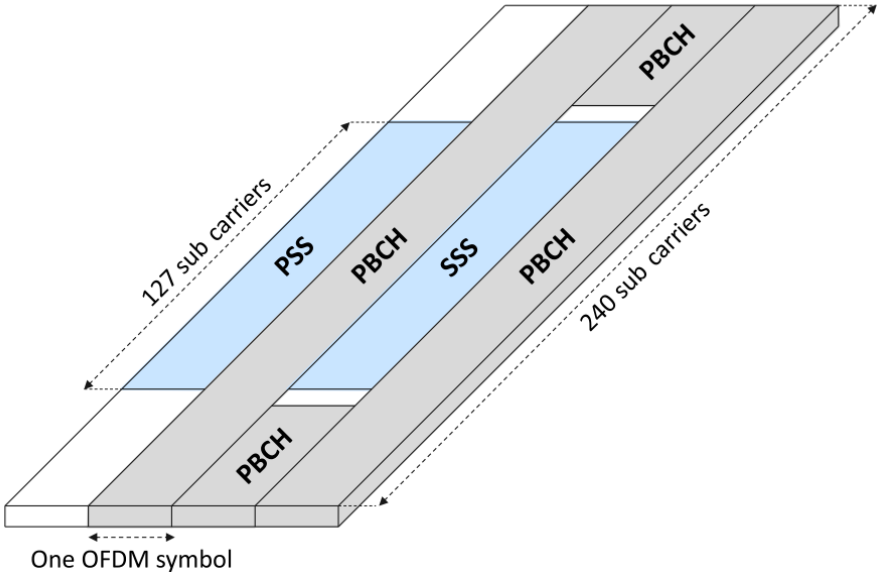


Figura 2.9: Estrutura do Bloco SS [4].

Tabela 2.2: Numerologia de Bloco SS e Faixas de Frequência [4].

Numerologia(kHz)	Largura de Banda(MHz)	Duração(μs)
15	3.6	≈ 285
30	7.2	≈ 143
120	28.8	≈ 36
240	57.6	≈ 18

duração curta de cada bloco que pode ser vantajoso para casos de varredura de feixes com um grande número correspondente de blocos SS multiplexados no tempo [4].

2.4.4 Acesso Aleatório

Na maioria dos casos, uma transmissão *uplink* NR ocorre usando um recurso dedicado que é atribuído pela rede ou célula para essa transmissão específica. Assim, não há risco de colisão com transmissões de outros dispositivos dentro da célula. Os procedimentos para acesso aleatório ocorrem por meio de 3 etapas principais: transmissão de um preâmbulo pelo dispositivo conhecido como *Physical Random Access Channel* (PRACH), transmissão de uma *Random Access Response* (RAR) e troca de mensagens entre dispositivos e rede [4, 34].

Diferentes tempos de guarda podem ser configurados para a recepção do PRACH para

corresponder às incertezas no tempo de recepção do preâmbulo, por exemplo, devido a diferentes tamanhos de células. Depois que um dispositivo transmite um preâmbulo de acesso aleatório, ele aguarda uma resposta ao acesso aleatório informando que o preâmbulo foi recebido corretamente. A resposta de acesso aleatório inclui: o índice do preâmbulo de acesso aleatório que a rede detectou, uma correção de tempo calculada pela rede com base no tempo de recepção do preâmbulo (o dispositivo deve atualizar o tempo de transmissão de *uplink* de acordo com a correção antes de novas transmissões do *uplink*), uma concessão de agendamento indicando qual recurso deverá ser utilizado para transmissão da mensagem subsequente, e por fim, uma identidade temporária deverá ser usada para comunicação adicional entre o dispositivo e a rede [4, 34, 35].

Após haver a recepção da resposta ao acesso aleatório pelo dispositivo, ele sincroniza o *uplink* no tempo. Cada dispositivo recebe uma identidade própria antes de realizar a transmissão de dados. A identidade é importante, pois será usada como parte do mecanismo de resolução de contenção na próxima etapa. Após isso, o UE começa a transmitir as mensagens necessárias à estação base. Por fim, a última etapa do procedimento consiste em uma mensagem de *downlink* para resolução de contenção que é usada para garantir que um dispositivo não utilize incorretamente a identidade de outro dispositivo [4, 34].

2.4.5 Transmissão de Tempo-Frequência

Nos sistemas NR, o uso do OFDM foi adotado como uma forma de onda ideal devido à sua robustez contra a dispersão do tempo e à capacidade de explorar eficientemente os domínios de tempo e frequência. Essa característica torna o OFDM uma escolha adequada ao definir a estrutura para diferentes canais e sinais nas direções de transmissão de *downlink* e *uplink*.

Um aspecto crucial do OFDM é a seleção da numerologia, especialmente o espaçamento entre as subportadoras e o comprimento do prefixo cíclico. Um espaçamento considerável entre as subportadoras traz benefícios significativos em termos de erro de frequência, reduzindo o impacto dos erros de frequência e do ruído de fase. Além disso, possibilita implementações que abrangem uma largura de banda considerável com um tamanho relativamente modesto de *Fast Fourier Transform* (FFT). Entretanto, para um determinado comprimento de prefixo cíclico em microssegundos, a sobrecarga relativa aumenta à medida que o espaçamento das subportadoras aumenta. Nessa perspectiva, um espaçamento menor das subportadoras seria mais preferível [4, 36].

No LTE, é empregado um espaçamento de subportadora equivalente a 15 kHz e um prefixo cíclico de aproximadamente 4.7 μs. Esses parâmetros foram cuidadosamente escolhidos para alcançar um equilíbrio ideal entre várias restrições inerentes aos cenários para os quais o LTE foi originalmente desenvolvido [37]. Esse cuidadoso ajuste permite que o LTE atenda de forma eficiente às demandas de comunicação em diversos ambientes, garantindo desempenho e confiabilidade consistentes em diferentes situações de uso.

No contexto do NR, a tecnologia foi estrategicamente desenvolvida para lidar com uma ampla variedade de cenários de implementação, que vão desde células grandes operando em frequências de portadora inferiores a 1 GHz até implantações de ondas milimétricas com alocações de espectro muito amplas. Nesse contexto diversificado, não faz sentido ter uma única numerologia que sirva para todos os cenários. Em vez disso, o NR adota uma abordagem flexível, reconhecendo a necessidade de uma numerologia adaptável [4].

2.4.6 Estrutura no Domínio do Tempo

No domínio do tempo, as transmissões no NR são organizadas em quadros com duração de 10 milissegundos, cada um desses quadros é subdividido em 10 partes iguais, cada uma com 1 milissegundo de comprimento. Cada uma dessas partes, conhecida como subquadro, é dividida em *slots*, sendo que cada *slot* contém 14 símbolos OFDM [5].

Em um nível mais alto de organização, cada quadro é identificado por um *System Frame Number* (SFN). O SFN desempenha um papel crucial na definição de diferentes ciclos de transmissão, incluindo ciclos de paginação em modo de economia de energia (modo sleep). É importante notar que o período do SFN é fixado em 1024 quadros, equivalendo a 10.24 segundos. Assim, o SFN se repete após a transmissão de 1024 quadros, garantindo uma sincronização precisa e confiável no sistema NR [4].

Para um espaçamento de subportadora de 15 kHz, um *slot* NR segue a mesma estrutura de um subquadro LTE com prefixo cíclico normal. Essa escolha de 15 kHz como espaçamento básico de subportadora é vantajosa para a coexistência harmoniosa entre NR e LTE. No entanto, essa decisão também implica que o prefixo cíclico para o primeiro e o oitavo símbolos em um *slot* de 15 kHz é ligeiramente maior em comparação com os outros símbolos [4]. No LTE, o tamanho padrão da subportadora é de 15 kHz. Isso significa que cada subportadora na banda de frequência é espaçada por 15 kHz. No NR, há uma maior flexibilidade em relação ao tamanho da subportadora. O NR suporta diferentes tamanhos de subportadora, incluindo: 15, 30 e 60

kHz.

Em detalhes, um símbolo OFDM é dividido em dois símbolos OFDM da próxima numerologia superior, e 14 símbolos consecutivos formam um *slot*. A Figura 2.10 fornece uma representação detalhada das divisões dos quadros, subquadros e *slots* correspondentes ao esquema OFDM [37]. Esse arranjo estrutural contribui para uma integração eficiente entre as tecnologias NR e LTE, assegurando uma transmissão de dados estável e eficaz no ambiente de comunicação sem fio [4, 5].

A definição de um bloco de recursos no contexto da tecnologia NR difere da definição usada no LTE. Em NR, um bloco de recursos é uma medida unidimensional que engloba apenas o domínio da frequência. Em contraste, na LTE, são utilizados blocos de recursos bidimensionais compostos por doze subportadoras no domínio da frequência e um *slot* LTE no domínio do tempo [5].

Uma razão fundamental para adotar a definição de blocos de recursos unidimensionais no domínio da frequência no NR é a flexibilidade na duração do tempo para diferentes transmissões. Isso significa que as transmissões NR podem variar em sua extensão temporal, proporcionando uma adaptabilidade para diferentes cenários de comunicação. Em contrapartida, no LTE, especialmente na versão original, as transmissões eram rigidamente alocadas em um *slot* completo, limitando a capacidade de ajustar a duração do tempo conforme necessário [4].

Essa abordagem inovadora no NR não apenas melhora a eficiência espectral, permitindo uma utilização mais eficaz do espectro de frequência disponível, mas também aumenta a flexibilidade operacional, tornando a tecnologia NR mais adaptável às demandas variáveis das redes de comunicação modernas. Essa flexibilidade é crucial para oferecer uma experiência de comunicação mais eficiente e robusta em ambientes diversificados e desafiadores.

2.4.7 Portas de Antena

A transmissão de múltiplas antenas no *downlink* é uma tecnologia essencial no contexto do NR. Em sistemas com várias antenas, os sinais transmitidos de diferentes antenas percorrem canais de rádio distintos, mesmo que todas as antenas estejam fisicamente localizadas no mesmo ponto. Cada antena é designada como uma "porta de antena", permitindo que se saiba através de qual canal um símbolo específico é transmitido a partir dessa porta [5].

Em outras palavras, cada transmissão individual no *downlink* é realizada por uma porta de antena específica, cuja identidade é conhecida pelo dispositivo receptor. Isso implica que o

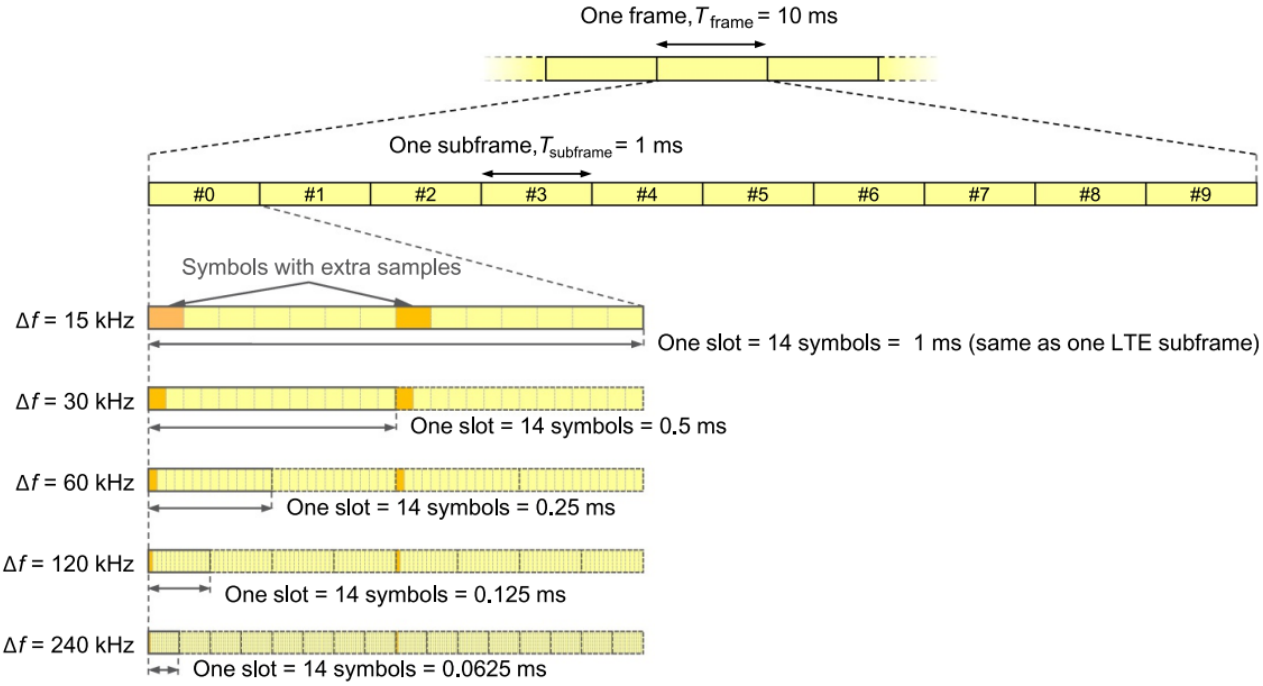


Figura 2.10: Representação dos quadros, subquadros e slots em NR [4].

dispositivo pode deduzir o canal pelo qual um símbolo é transmitido com base na porta de antena utilizada. Além disso, o dispositivo pode assumir que dois sinais transmitidos compartilham o mesmo canal de rádio somente se forem emitidos pela mesma porta de antena [4, 5].

Na prática, cada porta de antena pode ser considerada como correspondente a um sinal de referência específico, especialmente no *downlink*. Um dispositivo pode utilizar esse sinal de referência para estimar o canal associado a essa porta de antena específica. Além disso, os sinais de referência oferecem ao dispositivo a oportunidade de obter informações detalhadas sobre o estado do canal em relação a essa porta de antena [4, 5].

É fundamental salientar que o termo "porta de antena" representa um conceito abstrato que não está vinculado necessariamente a uma antena física específica. Por exemplo, dois sinais distintos podem ser transmitidos de maneira idêntica através de diversas antenas físicas. Quando um dispositivo receptor capta esses sinais, ele os percebe como se estivessem se propagando por um único canal, resultado da combinação dos canais das diferentes antenas.

2.4.8 Sinais de Referência

No contexto das modernas tecnologias de acesso de rádio, é essencial possuir um conhecimento detalhado sobre as diversas características do canal de rádio através do qual os sinais

são transmitidos. Isso inclui a compreensão da perda de percurso aproximada do canal de rádio para ajuste de potência de transmissão, bem como informações detalhadas sobre a amplitude e fase do canal ao longo do tempo, frequência e/ou domínio espacial.

Existem várias maneiras de adquirir esse conhecimento sobre o canal. Pode ser obtido por meio de medições realizadas tanto no transmissor quanto no receptor de um enlace de rádio. As informações coletadas durante as transmissões de *downlink* ou *uplink* podem ser relatadas à rede, permitindo a definição de diferentes parâmetros para transmissões subsequentes.

Em casos onde o canal é considerado recíproco, a rede pode adquirir conhecimento sobre as características do canal de *downlink* estimando as mesmas características na direção de *uplink*. Para facilitar esse processo, são utilizados sinais de referência específicos. No *downlink*, esses sinais são chamados de *Channel State Information Reference Signal* (CSI-RS), enquanto no *uplink* são referidos como *Sounding Reference Signal* (SRS) [4]. Esses sinais desempenham um papel fundamental ao fornecer informações cruciais para otimizar a qualidade e confiabilidade das transmissões em ambientes de rádio, garantindo uma comunicação eficaz e estável [4, 34].

2.4.8.1 Canal de Downlink - CSI-RS

Um CSI-RS desempenha um papel importante ao fornecer informações sobre as características do canal de transmissão. Além disso, ele é utilizado para estimar o nível de interferência, realizando a subtração entre o sinal recebido esperado e o que é efetivamente recebido no recurso CSI-RS. Cada CSI-RS configurado pode corresponder a até 32 portas de antenas diferentes, cada uma representando um canal específico a ser emitido [4].

No contexto do NR, é importante destacar que um CSI-RS é sempre configurado por dispositivo. No entanto, essa configuração não implica que um CSI-RS transmitido esteja restrito a ser utilizado apenas por um único dispositivo. Na prática, CSI-RS idênticos, utilizando o mesmo conjunto de elementos de recursos, podem ser configurados separadamente para vários dispositivos. Isso significa que um único CSI-RS pode ser compartilhado entre diferentes dispositivos, possibilitando uma utilização mais eficiente dos recursos disponíveis [4].

Um CSI-RS de porta única ocupa apenas um elemento de recurso dentro de um bloco correspondente a um bloco de recursos no domínio da frequência e um *slot* no domínio do tempo. Em teoria, o CSI-RS pode ser configurado para ocorrer em qualquer posição dentro deste bloco, mas na prática existem algumas restrições para evitar colisões com outros canais

e sinais físicos de *downlink*. Por outro lado, um CSI-RS multiporta pode ser visualizado como vários CSI-RS transmitidos de forma ortogonal por portas de antena diferentes, compartilhando o mesmo conjunto geral de elementos de recursos atribuídos ao CSI-RS multiporta configurado. A estrutura do CSI-RS é ilustrada na Figura 2.11 [4].

O CSI-RS multiportas compreende um conjunto de portas de antena, o qual pode ser utilizado para sondar os canais correspondentes às portas de antena. Frequentemente, uma porta CSI-RS não está diretamente associada a uma antena física específica, o que significa que o canal transmitido com base em um CSI-RS pode não ser o canal de rádio físico real. Portanto, é possível aplicar qualquer tipo de transformação ou filtragem espacial ao CSI-RS antes de mapeá-lo para antenas físicas. Consequentemente, o número de antenas físicas pode ser maior do que o número de portas CSI-RS. Quando um dispositivo realiza a sondagem de canal com base no CSI-RS, nem o filtro nem as antenas físicas são explicitamente visíveis; em vez disso, são os canais correspondentes às portas de antena que são observados. A Figura 2.12 ilustra uma representação da filtragem dos sinais CSI-RS antes de serem mapeados para antenas físicas [4].

2.4.8.2 Canal de Uplink - SRS

Em diversos aspectos, o SRS pode ser considerado equivalente ao CSI-RS, embora com direções de transmissão diferentes. No entanto, ao analisarmos em detalhes, a estrutura do SRS difere significativamente do CSI-RS. Além disso, o SRS é restrito a um máximo de quatro portas de antena, enquanto o CSI-RS é capaz de suportar até 32 portas de antena [4].

A Figura 2.13 ilustra a estrutura fundamental de tempo e frequência de um SRS. Em geral, um SRS pode consistir em um, dois ou quatro símbolos OFDM consecutivos, e é posicionado dentro dos últimos seis símbolos de um *slot*. No domínio da frequência, um SRS segue uma estrutura em pente, o que significa que é transmitido em cada conjunto de subportadoras, com N assumindo os valores dois ou quatro (conhecidos como "comb-2"e "comb-4", respectivamente) [4].

As transmissões SRS de diversos dispositivos podem ser multiplexadas em frequência dentro da mesma faixa de frequência, sendo-lhes atribuídos diferentes *slots* correspondentes a diferentes deslocamentos de frequência. No esquema comb-2, ou seja, quando o SRS é transmitido a cada segunda subportadora, dois SRSs podem ser multiplexados em frequência. No caso do comb-4, é possível multiplexar até quatro SRSs na mesma faixa de frequência [4].

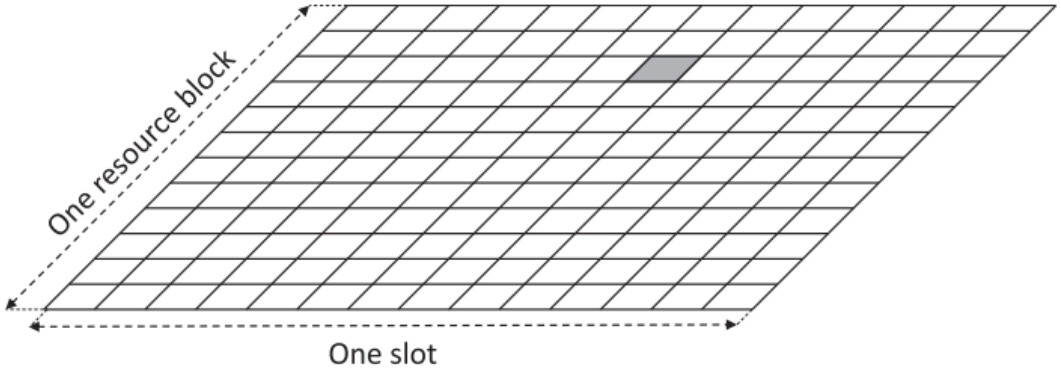


Figura 2.11: Estrutura do CSI-RS [4].

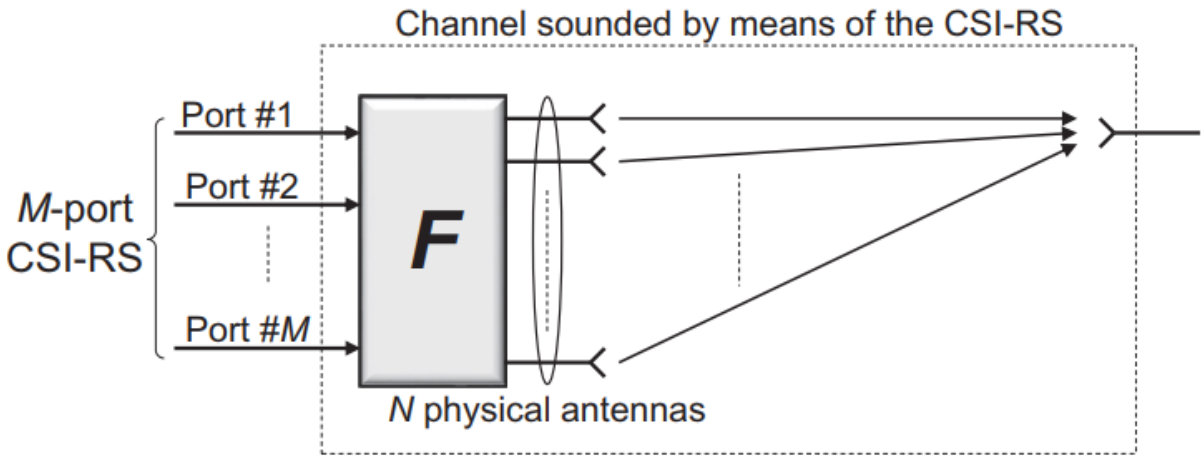


Figura 2.12: Representação do CSI-RS aplicado ao filtro espacial (F) antes de serem mapeados para antenas físicas [4].

Em sistemas SRS que suportam múltiplas portas de antena, essas portas compartilham o mesmo conjunto de elementos de recursos e a mesma sequência básica de SRS. Para diferenciá-las, são aplicadas diferentes rotações de fase. Esta técnica é realizada no domínio da frequência, mas pode ser entendida como um deslocamento cíclico no domínio do tempo. Na especificação NR, essa operação é formalmente denominada "mudança cíclica", embora seja matematicamente descrita como uma mudança de fase no domínio da frequência.

À semelhança do CSI-RS, as portas SRS geralmente não são diretamente mapeadas para as antenas físicas do dispositivo. Em vez disso, são mapeadas através de um filtro espacial, como ilustrado na Figura 2.14. A rede pode sondar o canal com base em um sinal enviado pelo dispositivo (SRS) e, em seguida, decidir sobre uma matriz pré-codificadora que o dispositivo deve usar para a transmissão posterior no enlace. Presume-se que o dispositivo utilize essa ma-

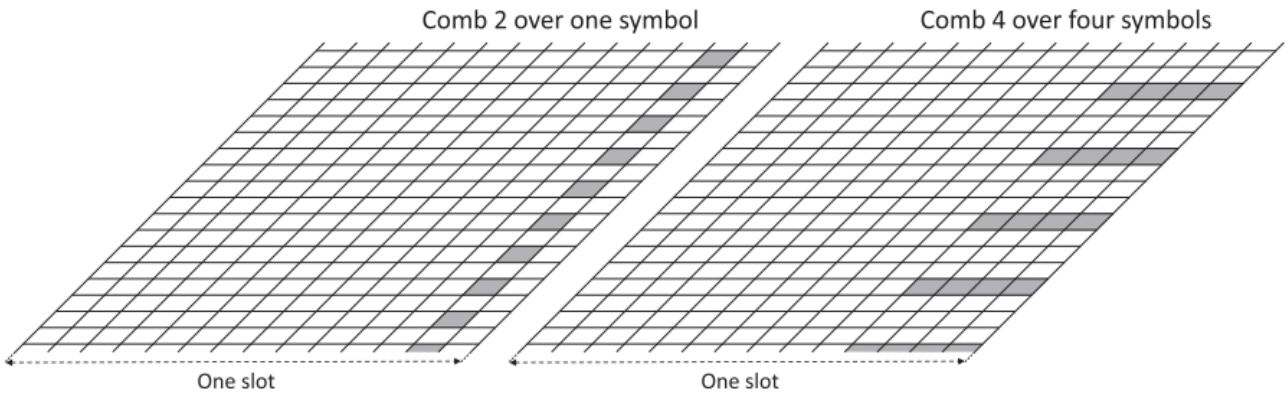


Figura 2.13: Representação da estrutura do SRS [4].

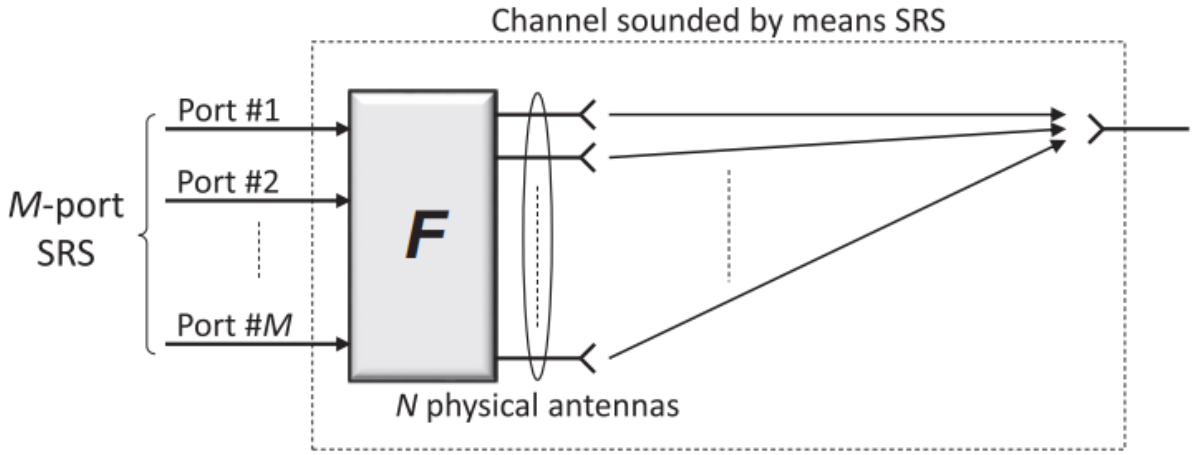


Figura 2.14: Representação do SRS aplicado ao filtro espacial (F) antes de serem mapeados para antenas físicas [4].

triz pré-codificadora em conjunto com o filtro espacial F aplicado ao SRS. Um dispositivo pode ser explicitamente programado para a transmissão de dados usando portas de antena definidas por um determinado SRS. Na prática, isso implica que o dispositivo transmite usando o mesmo filtro espacial F que foi aplicado ao SRS [4].

2.4.9 Processamento do Canal de Transporte

O processamento geral do canal de transporte é semelhante tanto para *uplink* quanto para *downlink*, como detalhado na Figura 2.15. Em cada *Transmission Time Interval* (TTI), até dois blocos de transporte de tamanho dinâmico são entregues à camada física e transmitidos pela interface de rádio para cada portadora componente. A utilização de dois blocos de transporte

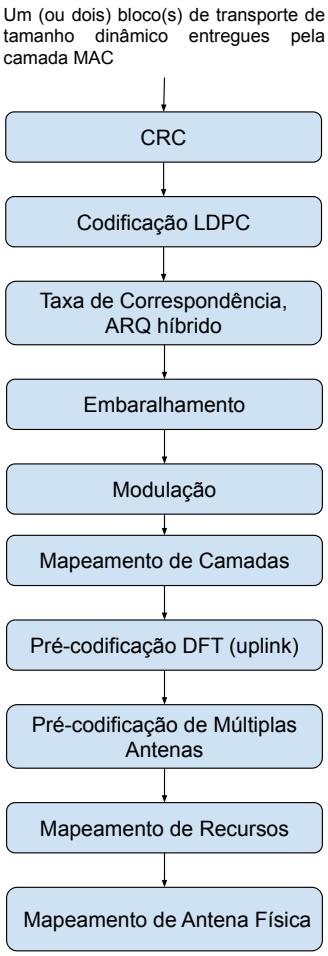


Figura 2.15: Representação do processamento da canal de transporte.

ocorre somente em casos de multiplexação espacial com mais de quatro camadas, sendo suportada apenas no *downlink*. Esse cenário é especialmente útil em ambientes com SNR muito alta [4].

2.4.9.1 Codificação CRC

O processo é iniciado com a codificação dos canais, na qual é anexado um *Cyclic Redundancy Check* (CRC) ao bloco de transporte para facilitar a detecção de erros. Em seguida, o bloco de código é segmentado. O CRC possibilita a detecção de erros no lado do receptor, após a decodificação do bloco de transporte, podendo ser utilizado como sinal para solicitar retransmissões quando necessário [4, 38].

2.4.9.2 Codificação LDPC

A segmentação de blocos de código é uma técnica utilizada para lidar com blocos grandes, onde os divide em tamanhos iguais. O codificador *Low-Density Parity Check Code* (LDPC) em NR é projetado para operar até um tamanho específico de bloco de código [4, 34].

2.4.9.3 ARQ híbrido

A funcionalidade ARQ híbrido se refere ao esquema de retransmissão automática utilizada para garantir a integridade e a confiabilidade das transmissões de dados em redes móveis. O ARQ híbrido no NR combina técnicas de ARQ tradicional com técnicas de *Incremental Redundancy* (IR), proporcionando melhorias significativas na eficiência e na robustez da transmissão de dados [4].

2.4.9.4 Embaralhamento

A etapa de embaralhamento é uma parte fundamental do processo de transmissão de dados para garantir a segurança e a integridade das informações transmitidas. Os bits após a codificação do canal são embaralhados com uma sequência pseudo-aleatória de bits. A saída do codificador de canal pode, teoricamente, ser igual a outro sinal interferente, o que pode levar à falha de decodificação na extremidade do receptor. Este módulo de embaralhamento adiciona aleatoriedade extra aos bits codificados, o que fornece ganho de codificação ao suprimir a interferência. A sequência pseudo-aleatória é diferente e única para cada canal [4].

2.4.9.5 Modulação

A etapa de modulação é crucial no processo de transmissão de dados, pois transforma o bloco de bits embaralhados em um conjunto correspondente de símbolos de modulação complexos. Esta transformação é essencial para a comunicação eficaz, e os esquemas de modulação suportados nesta etapa incluem QPSK, 16QAM, 64QAM e 256QAM, tanto no uplink quanto no downlink [4, 34].

2.4.9.6 Mapeamento de Camadas

Na etapa de mapeamento de camadas, o objetivo é distribuir os símbolos de modulação entre as diversas camadas de transmissão. Esse processo é semelhante ao que ocorre no LTE:

cada símbolo é mapeado para todas as camadas. Um bloco de transporte codificado pode ser distribuído em até quatro camadas. No caso de cinco a oito camadas, suportadas exclusivamente no *downlink*, um segundo bloco de transporte é mapeado para as camadas cinco a oito, seguindo o mesmo princípio do primeiro bloco de transporte. A transmissão multicamadas é possível apenas em conjunto com a técnica OFDM, que é a forma de onda de linha de base em NR [4].

O mapeamento de camadas é realizado baseado nos *codewords* que consistem em sequências de bits que são transmitidos no canal de comunicação, Os *codewords* são utilizados para representar informações que são enviadas entre a estação base e o dispositivo celular. De acordo com [5], cada *codeword* é composta por até 4 camadas, sendo que no sistema de comunicação podem ser transmitidos no máximo 2 *codewords*. A Tabela 2.3 mostra o mapeamento das camadas por *codeword*.

2.4.9.7 Pré-Codificação DFT de Uplink

No caso de a pré-codificação *Discrete Fourier Transform* (DFT) ser aplicada no uplink, blocos de símbolos são alimentados por uma DFT de tamanho M, onde M corresponde ao número de subportadoras atribuídas para a transmissão. A razão para a pré-codificação DFT é permitir maior eficiência do amplificador de potência [4].

2.4.9.8 Pré-Codificação de Múltiplas Antenas

Esta etapa tem como objetivo mapear as diversas camadas de transmissão para um conjunto de portas de antena utilizando uma matriz pré-codificadora. No contexto do 5G (NR), as operações de pré-codificação e múltiplas antenas variam entre os canais de *downlink* e *uplink*.

No processo de *downlink*, o *Demodulation Reference Signal* (DM-RS), utilizado para a estimação do canal, é submetido à mesma pré-codificação que o *Physical Downlink Shared Channel* (PDSCH). Como resultado, essa pré-codificação não é diretamente perceptível para o receptor; em vez disso, ela é integrada ao canal global. Essa abordagem é comparável à filtragem espacial transparente ao receptor discutida anteriormente no contexto de CSI-RS e SRS [4].

Os sinais de referência de demodulação são transmitidos nos blocos de recursos programados. A partir desses sinais de referência, o dispositivo pode estimar o canal, incluindo qualquer pré-codificação e filtragem espacial solicitadas pelo PDSCH. Em princípio, o conhecimento sobre a correlação entre as transmissões do sinal de referência é valioso, tanto em termos

Tabela 2.3: Mapeamento de Camadas por *Codeword* [5].

Número de Camadas	Número de *Codewords*
1	1
2	1
3	1
4	1
5	2
6	2
7	2
8	2

da correlação introduzida pelo canal de rádio quanto da correlação no uso do pré-codificador. Esse conhecimento pode ser explorado pelo dispositivo para aprimorar a precisão da estimação do canal [4].

Assim como no *downlink*, os sinais de referência para demodulação de *uplink*, estão sujeitos à mesma pré-codificação aplicada ao *Physical Uplink Shared Channel* (PUSCH) de *uplink*. Portanto, no contexto do *uplink*, a pré-codificação não é diretamente perceptível pelo receptor; em vez disso, ela é considerada como parte integrante do canal global [35].

2.4.9.9 Mapeamento de Recursos

O mapeamento de blocos de recursos envolve a alocação dos símbolos de modulação a serem transmitidos em cada porta de antena, associando-os aos elementos de recursos disponíveis em um conjunto de blocos de recursos atribuídos pelo escalonador MAC para a transmissão. Cada bloco de recursos compreende 12 subportadoras de largura, e geralmente, múltiplos blocos de recursos e símbolos OFDM são empregados durante a transmissão. Entretanto, é importante notar que alguns ou todos os elementos de recursos dentro dos blocos programados podem não estar disponíveis para a transmissão no canal de transporte. Isso ocorre porque esses elementos podem ser reservados para sinais de referência de demodulação, como CSI-RS e SRS, sinalização de controle *downlink*, sinais de sincronização e informações do sistema, entre outras funções essenciais [4].

Para a transmissão, os recursos de tempo-frequência são identificados como um conjunto

de blocos de recursos virtuais e um conjunto de símbolos OFDM. Esses blocos de recursos virtuais, que contêm os símbolos de modulação, são mapeados para blocos de recursos físicos na parte da largura de banda designada para a transmissão [4].

2.4.9.10 Sinais de Referência

Sinais de referência são padrões predefinidos que ocupam posições específicas na grade de tempo-frequência do *downlink*. A especificação NR inclui uma variedade de sinais de referência transmitidos de diferentes maneiras, cada um projetado para servir a propósitos distintos para um dispositivo receptor. Ao contrário do LTE, que depende fortemente de sinais de referência sempre presentes e específicos da célula no *downlink* para demodulação coerente, estimação de qualidade do canal para relatórios de CSI e rastreamento geral de tempo-frequência, o NR utiliza sinais de referência de *downlink* variados para diferentes finalidades [4].

Em cada instância de DM-RS, é possível criar diversos sinais de referência ortogonais. Esses sinais são distintos nos domínios da frequência e do código, e no caso de um DM-RS de símbolo duplo, também no domínio do tempo. Existem dois tipos de sinais de referência de demodulação, denominados Tipo 1 e Tipo 2, que se diferenciam no mapeamento no domínio da frequência e no número máximo de sinais de referência ortogonais. O Tipo 1 pode fornecer até quatro sinais ortogonais utilizando um DM-RS de símbolo único e até oito sinais ortogonais usando um DM-RS de símbolo duplo. Para o Tipo 2, esses números são seis e doze, respectivamente. É possível combinar diferentes tipos de mapeamento com diversos sinais de referência. A geração da sequência de sinal de referência em todos os blocos de recursos garante que a mesma sequência subjacente seja utilizada para vários dispositivos programados em recursos de tempo-frequência sobrepostos, especialmente em cenários de MIMO multiusuário [4].

2.4.10 Transmissão Multiantena

A utilização de múltiplas antenas no lado do transmissor e/ou receptor oferece uma solução eficaz para mitigar os efeitos do desvanecimento. Isso ocorre porque os canais observados por diferentes antenas podem ser, em parte, independentes entre si. Essa independência pode ser resultado da distância significativa entre as antenas ou da diferença de polarização entre elas. Além disso, ao ajustar cuidadosamente a fase e, em alguns casos, a amplitude de cada elemento da antena, as múltiplas antenas no lado do transmissor podem ser utilizadas para criar um efeito direcional, conhecido como *beamforming*. Em outras palavras, é possível concentrar

a potência transmitida em uma direção específica ou em locais específicos no espaço. Esse direcionamento não apenas aumenta as taxas de dados e o alcance alcançável, devido ao aumento da potência que atinge o receptor desejado, mas também reduz a interferência em outras conexões, melhorando, assim, a eficiência global do espectro [4].

Da mesma forma, múltiplas antenas no lado do receptor podem ser empregadas para criar direcionalidade na recepção, permitindo focalizar a captação na direção de um sinal desejado, enquanto suprime interferências provenientes de outras direções. Além disso, a presença de múltiplas antenas tanto no lado do transmissor quanto no lado do receptor possibilita a realização de multiplexação espacial. Isso significa que é possível transmitir várias camadas de dados em paralelo, utilizando os mesmos recursos de tempo e frequência. Dessa forma, a utilização estratégica de múltiplas antenas não apenas melhora a qualidade da comunicação sem fio, mas também maximiza a eficiência espectral, proporcionando uma transmissão mais robusta e confiável.

Em um sistema de comunicações móveis, onde os dispositivos estão localizados em diversas direções em relação à estação base e possuem uma rotação essencialmente aleatória, o uso de antenas fixas altamente direcionais não é prático. No entanto, é possível alcançar um efeito semelhante, ou seja, estender a área total da antena de recepção para permitir uma transmissão mais direcional, por meio de um painel de antena composto por vários pequenos elementos.

Nesse cenário, cada elemento da antena e a distância entre eles são proporcionais ao comprimento de onda da frequência utilizada. À medida que a frequência aumenta, o tamanho dos elementos e suas distâncias mútuas são reduzidos. Para compensar essa redução mantendo o tamanho global da antena constante, pode-se aumentar o número de elementos da antena. A vantagem desse painel de antena com muitos pequenos elementos, em comparação com uma única antena grande, é que a direção do feixe transmissor pode ser ajustada de forma independente, alterando a fase dos sinais aplicados a cada elemento de antena, tanto no transmissor quanto no receptor.

De maneira geral, qualquer sistema de transmissão linear que envolva múltiplas antenas pode ser representado conforme ilustrado na Figura 2.16, com N_L camadas, as quais são capturadas pelo vetor $\bar{x}$ e posteriormente mapeadas para N_T antenas de transmissão, representadas pelo vetor $\bar{y}$. Esse mapeamento é realizado por meio de uma multiplicação com uma matriz $\mathbf{W}$ de dimensões $N_T \times N_L$ [4].

Um aspecto importante da pré-codificação em sistemas com múltiplas antenas diz res-

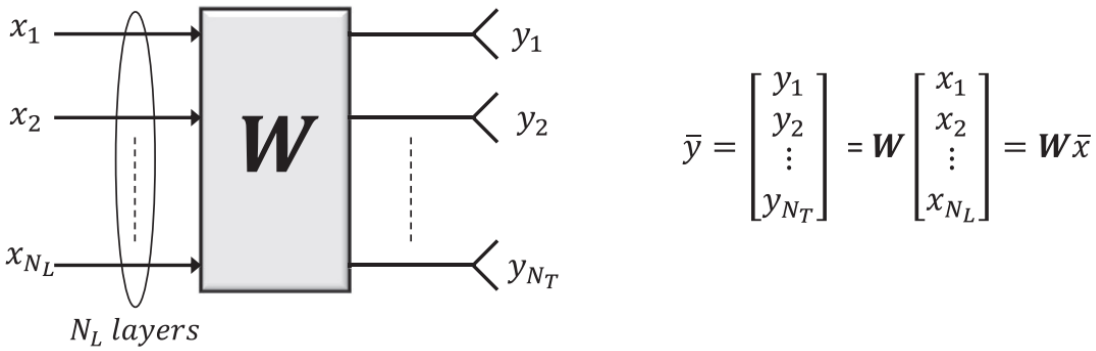


Figura 2.16: Mapeamento de transmissão multiantena [4].

peito à aplicação da técnica aos DM-RS utilizados para possibilitar a demodulação coerente do sinal pré-codificado. Quando o DM-RS não é pré-codificado, é necessário informar ao receptor qual pré-codificador está sendo usado no transmissor para permitir a demodulação adequada dos dados pré-codificados. Em contrapartida, se os sinais de referência são pré-codificados juntamente com os dados, a pré-codificação é considerada parte do canal multidimensional global, visto pelo receptor como uma matriz de dimensões $N_R \times N_T$, em vez da concatenação do canal $\mathbf{H}$ e da matriz de pré-codificação $\mathbf{W}$ no transmissor. Nesse contexto, a pré-codificação se torna transparente para o receptor, simplificando o processo de demodulação coerente [4].

2.5 Modelo do Sistema e Tráfego no Fronthaul

2.5.1 Descrição do Sistema

Este trabalho considera um cenário de D-MIMO, no qual uma CU controla M APs distribuídos que possuem uma única antena para comunicar-se coerentemente com K UEs que possuem uma única antena, onde $K \ll M$. O número de UEs atendidos refere-se a usuários que usam os mesmos recursos de tempo e frequência com multiplexação espacial, mas o número de usuários na rede pode ser maior, desde que eles usem outros recursos de tempo ou subportadoras. No *downlink*, os símbolos recebidos em um determinado *Resource Element* (RE) podem ser representados como mostra a Equação 2.1.

Supondo que existam K símbolos por RE a serem transmitidos aos usuários (um símbolo para cada usuário por RE), é possível usar um pré-codificador *Zero-Forcing* $\mathbf{A}' \in \mathbb{C}^{M \times K}$ que pode mitigar a interferência entre os símbolos *downlink*, ou seja, $\mathbf{GA}' \approx \mathbf{I}_K$. Na prática, a

matriz do pré-codificador *Zero-Forcing* pode ser calculada como o pseudo-inverso da matriz do canal estimada:

$$\mathbf{A}' = \hat{\mathbf{G}}^H(\hat{\mathbf{G}}\hat{\mathbf{G}}^H)^{-1}, \tag{2.2}$$

onde $\hat{\mathbf{G}} = \mathbf{G} - \tilde{\mathbf{G}} - \bar{\mathbf{G}}$ é a matriz do canal estimada na CU que representa o CSI, $\mathbf{G}$ é o canal verdadeiro, $\tilde{\mathbf{G}}$ é o erro de estimação do canal no APs, e $\bar{\mathbf{G}}$ é a distorção da compressão. Cada elemento de $\mathbf{G}$ é modelado aqui como uma combinação do *large scale fading* β_{km} e *small scale fading* h_{km}, entre o usuário k e o AP m: $g_{km} = \sqrt{\beta_{km}}h_{km}$, onde $h_{km} \sim \mathcal{N}_{\mathbb{C}}(0,1)$ [7, 11].

Assumimos uma rede operando no modo duplex por TDD com reciprocidade de canal perfeita. Assim, no início do intervalo de coerência, os usuários podem enviar pilotos ortogonais aos APs, que estimam os CSIs e os enviam para a CU. Os K CSIs estimados no m-ésimo AP pode ser representado pelo vetor $\dot{\mathbf{g}}_m = [\dot{g}_{1m}, \dot{g}_{2m}, \cdots, \dot{g}_{Km}]$, e se considerarmos todos os APs: $\dot{\mathbf{G}} = \mathbf{G} - \tilde{\mathbf{G}}$, onde $\dot{\mathbf{G}} = \left[\dot{\mathbf{g}}_1^T \dot{\mathbf{g}}_2^T \cdots \dot{\mathbf{g}}_M^T\right]$. O canal estimado entre o m-ésimo AP e o k-ésimo UE pode ser representado por $\dot{g}_{km} = g_{km} - \tilde{g}_{km}$.

O erro de estimação é modelado como $\tilde{\mathbf{G}}$, cujos elementos são variáveis aleatórias gaussianas com média zero e variância γ_{km}, ou seja, $\tilde{g}_{km} \sim \mathcal{N}_{\mathbb{C}}(0, \gamma_{km})$. No caso TDD e assumindo um estimador de canal com MMSE, a variância de $\tilde{g}_{km}$ é dada por $\gamma_{km} = \beta_{km} - \frac{\rho_u \tau \beta_{km}^2}{\sigma_{est}^2 + \rho_u \tau \beta_{km}}$ [11, 7], onde ρ_u é a potência do *uplink* transmitida, τ é o comprimento da sequência do piloto enviada e σ_{est}^2 é a variância do ruído em cada AP receptor.

A matriz de distorção de compressão $\bar{\mathbf{G}}$ é criada pela representação de bits finitos da estimação do canal, por exemplo, devido à quantização ou compressão do CSI. Cada AP estima K *taps* de canal de K UEs ativos e quantiza as estimações ou executa algoritmos de compressão neles.

Quando a estimação do canal $\hat{\mathbf{G}}$ está pronta na CU, a matriz de pré-codificação *Zero-Forcing* é calculada de acordo com (2.2), mas uma segunda matriz é necessário para limitar a potência de transmissão em cada antena. Assim, a matriz de pré-codificação é definida como $\mathbf{A} = \mathbf{A}'\mathbf{P}$, onde a matriz $\mathbf{P} \in \mathbb{R}_+^{K \times K}$ é uma matriz diagonal com elementos $\sqrt{p_i}$, $i = 1, \ldots, K$, e p_i é a potência associada ao i-ésimo símbolo do usuário após executar pré-codificação. Então, os símbolos transmitidos em cada RE são $\mathbf{x} = \mathbf{A}\mathbf{s}$, onde os elementos de $\mathbf{s}$, representados no vetor coluna $\mathbf{s} \in \mathbb{C}^{K \times 1}$, tem potência unitária média ($\mathbf{E}\left[|s_k|^2\right] = 1, k = 1, \ldots, K$).

A matriz diagonal de alocação de potência $\mathbf{P}$ pode ser encontrada com uma variedade de algoritmos, como o uso de métodos de otimização max-min [7, 8, 11, 10], ou abordagens

baseadas em heurísticas. Neste trabalho, é utilizada a solução heurística de baixa complexidade baseada na alocação igual de potência [7]:

$$p_{11} = p_{22} = \ldots = p_{KK} = \eta = \frac{\rho_d}{\max_m \sum_{i=1}^{K} |a'_{mi}|^2}, \tag{2.3}$$

onde $\sqrt{p_{11}}, \ldots, \sqrt{p_{KK}}$ são os elementos diagonais de $\mathbf{P}$, ρ_d é a restrição de potência do downlink e a' são os elementos do pré-codificador *Zero-Forcing* $\mathbf{A}'$.

2.5.2 Tráfego no Fronthaul

Considerando um cenário com sinais OFDM e o D-MIMO com pré-codificação e alocação de potência totalmente centralizadas, no início de cada intervalo de coerência, os K UEs enviam pilotos ortogonais para os M APs para estimação de canal. Então, cada AP envia $K \times \frac{N_{\text{SC}}}{C_{\text{BW}}}$ canais estimados para a CU, onde N_{SC} é o número de subportadoras do sinal multiportadora e C_{BW} é o número de subportadoras na largura de banda de coerência. Então, a CU calcula os coeficientes de pré-codificação, a alocação de potência, realiza a pré-codificação dos símbolos e envia os símbolos pré-codificados para cada AP. Finalmente, o AP envia os símbolos pré-codificados para os UEs pelo ar. Em resumo, a pré-codificação de símbolos, o transporte *fronthaul* e a transmissão aérea são repetidos para cada símbolo OFDM durante o intervalo de coerência. Em contraste, a estimação de CSI e seu transporte para a CU são feitos uma vez a cada intervalo de coerência.

Considerando um AP que implementa o *functional split* 7.1 [33], onde apenas a representação no domínio da frequência é trocada com a CU, o tráfego *fronthaul* necessário para transmitir *downlink* ou *uplink* IQ amostras para cada AP podem ser escritas como $R_{IQ}^{DL} = R_{IQ}^{UL} = \frac{N_{SC} b_{\text{IQ}}}{T_{\text{sym}}}$, onde b_{IQ} é o número de bits para representar cada amostra IQ e T_{sym} é o período de um único símbolo OFDM. Além disso, no início de cada intervalo de coerência, o AP também precisa enviar o CSI para a CU, o que requer um tráfego *uplink* adicional de $R_{\text{CSI}} = \frac{K \times b_{\text{CSI}} \times \frac{N_{\text{SC}}}{C_{\text{BW}}}}{T_{\text{sym}}}$, onde b_{CSI} é o número de bits usados para representar cada amostra CSI. Neste caso, considera-se que o AP utiliza um período de símbolo OFDM para enviar o CSI para a CU para evitar latência extra na cadeia *uplink* . Assim, o tráfego no *fronthaul uplink* por AP teria um pico no início de cada intervalo de coerência dado por $R_P = R_{IQ}^{UL} + R_{CSI}$, e no tempo restante, o tráfego seria R_{IQ}^{UL}. Considerando M APs, a taxa de dados no *fronthaul* agregada é $R_a = M R_P$.

Com base na discussão anterior, as Figuras 2.17 - 2.22 mostram o tráfego *uplink fronthaul* para dados IQ, para dados CSI e a taxa agregada de pico para cada *link* de *fronthaul* que conecta

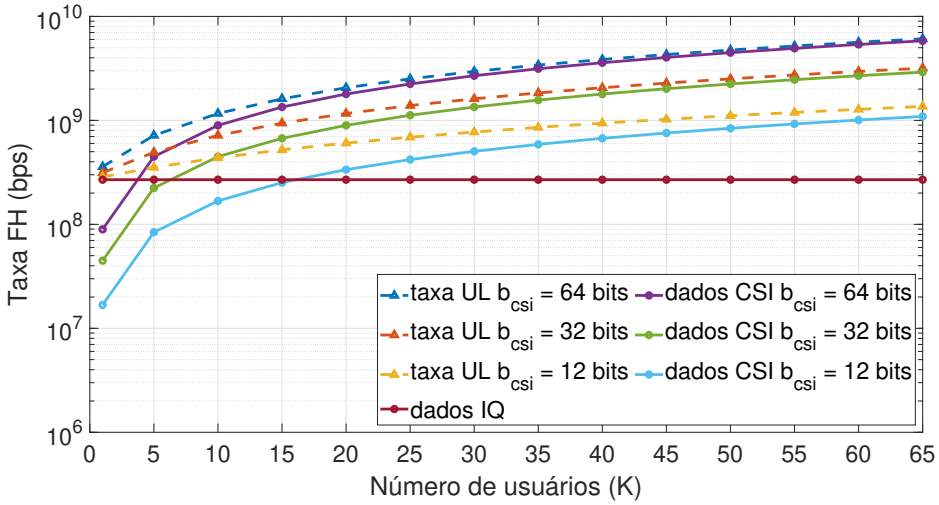


Figura 2.17: Tráfego *fronthaul uplink* para dados IQ (R_{IQ}^{UL}), dados CSI (R_{CSI}) e taxa agregada de pico (R_P) para cada *link* de *fronthaul* que conecta o AP a CU para diferentes números de UEs e com $N_{SC} = 1200$ subportadoras.

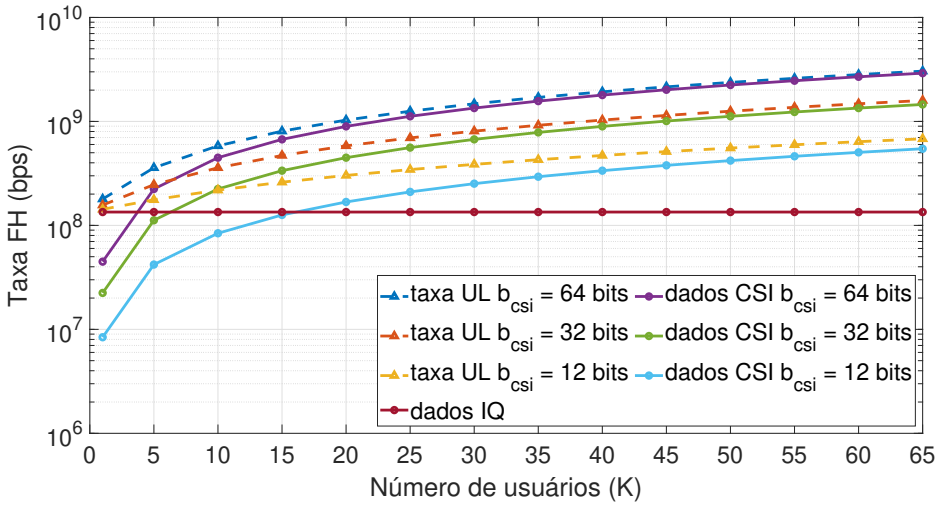


Figura 2.18: Tráfego *fronthaul uplink* para dados IQ (R_{IQ}^{UL}), dados CSI (R_{CSI}) e taxa agregada de pico (R_P) para cada *link* de *fronthaul* que conecta o AP a CU para diferentes números de UEs e com $N_{SC} = 600$ subportadoras.

o AP a CU, considerando diferentes números de UEs e diferentes números de APs conectados à rede distribuída. Além disso, foi considerado $N_{SC} = 1200, 600\ e\ 180$ subportadoras, $b_{IQ} = 16$ bits, $T_{sym} = 1/14$ ms e $C_{BW} = 12$ subportadoras. Existem três configurações com diferentes b_{CSI}: 64, 32 e 12 bits.

Na Figura 2.17 é usada a configuração de $N_{SC} = 1200$ subportadoras para diferentes

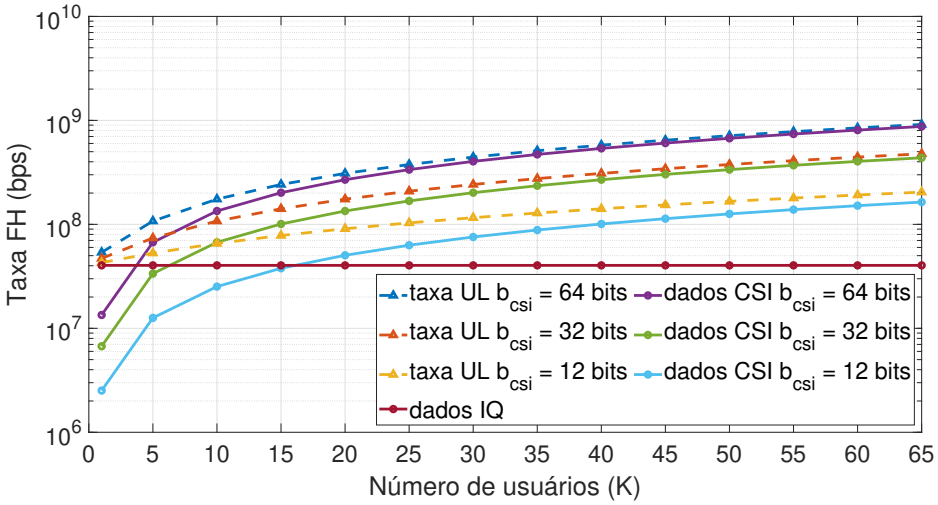


Figura 2.19: Tráfego *fronthaul uplink* para dados IQ (R_{IQ}^{UL}), dados CSI (R_{CSI}) e taxa agregada de pico (R_P) para cada *link* de *fronthaul* que conecta o AP a CU para diferentes números de UEs e com $N_{SC} = 180$ subportadoras.

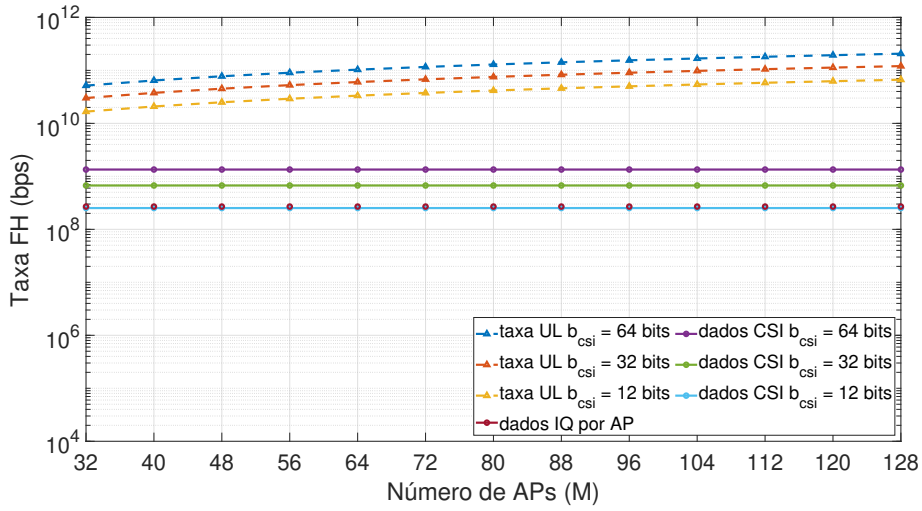


Figura 2.20: Tráfego *fronthaul uplink* para dados IQ (R_{IQ}^{UL}), dados CSI (R_{CSI}) e taxa agregada de pico (R_P) de todos os APs somadas para cada *link* de *fronthaul* que conecta o AP a CU para diferentes números de APs e com $N_{SC} = 1200$ subportadoras, considerando $K = 15$ UEs.

números de UEs. Com o gráfico é possível verificar que a taxa aumenta conforme o aumento do número de UEs e de b_{CSI}. Se o link *fronthaul* tiver uma restrição determinada, um b_{CSI} inferior permitiria implantações de rede distribuída com mais usuários. Por exemplo, considerando um *link Ethernet* comum de 1 Gbps para conectar cada AP a CU, a configuração com $b_{csi} = 12$ bits permitiria que 40 usuários fossem conectados na rede distribuída usando os mesmos recursos

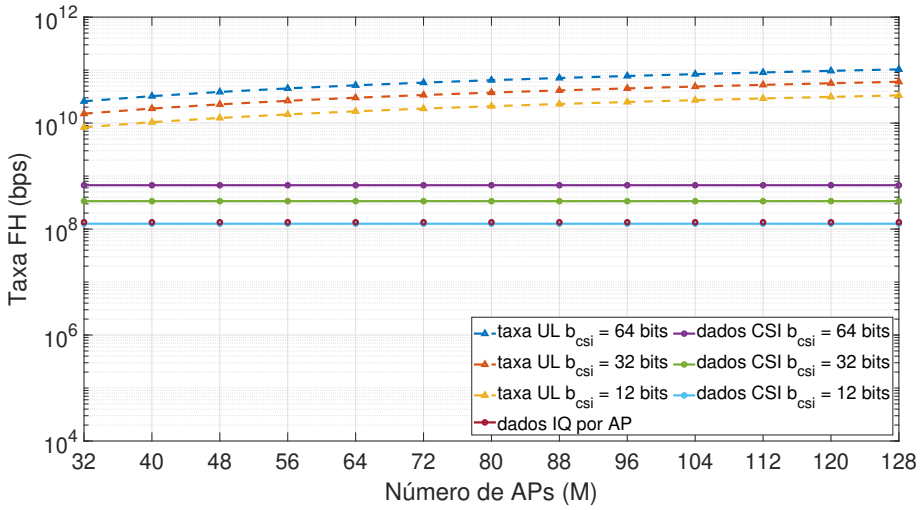


Figura 2.21: Tráfego *fronthaul uplink* para dados IQ (R_{IQ}^{UL}), dados CSI (R_{CSI}) e taxa agregada de pico (R_P) de todos os APs somadas para cada *link* de *fronthaul* que conecta o AP a CU para diferentes números de APs e com $N_{SC} = 600$ subportadoras, considerando $K = 15$ UEs.

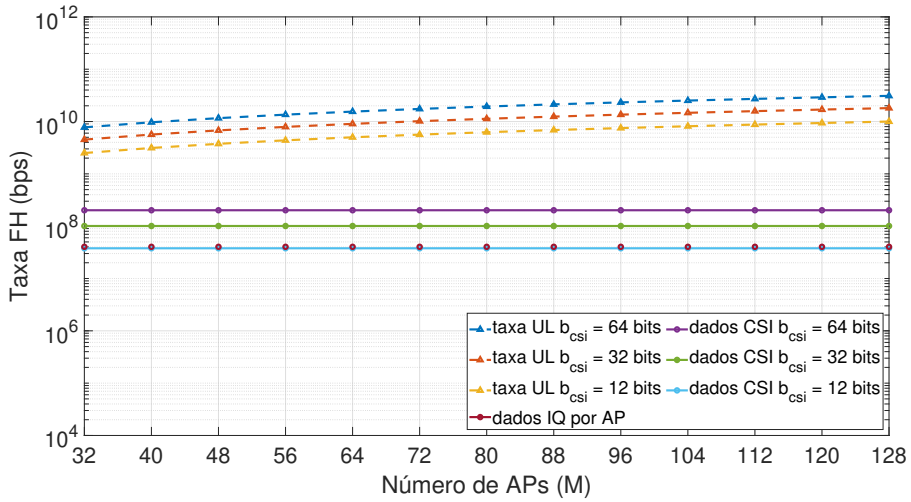


Figura 2.22: Tráfego *fronthaul uplink* para dados IQ (R_{IQ}^{UL}), dados CSI (R_{CSI}) e taxa agregada de pico (R_P) de todos os APs somadas para cada *link* de *fronthaul* que conecta o AP a CU para diferentes números de APs e com $N_{SC} = 180$ subportadoras, considerando $K = 15$ UEs.

de tempo-frequência. Por outro lado, $b_{csi} = 32$ bits e $b_{csi} = 64$ bits permitiriam apenas 16 e 8 usuários, respectivamente.

Na Figura 2.18 é usada a configuração de $N_{SC} = 600$ subportadoras para diferentes números de UEs, mostrando que a diminuição do N_{SC} influencia na diminuição da taxa no *fronthaul*. Por exemplo, considerando o *link Ethernet* comum de 1 Gbps, a configuração com

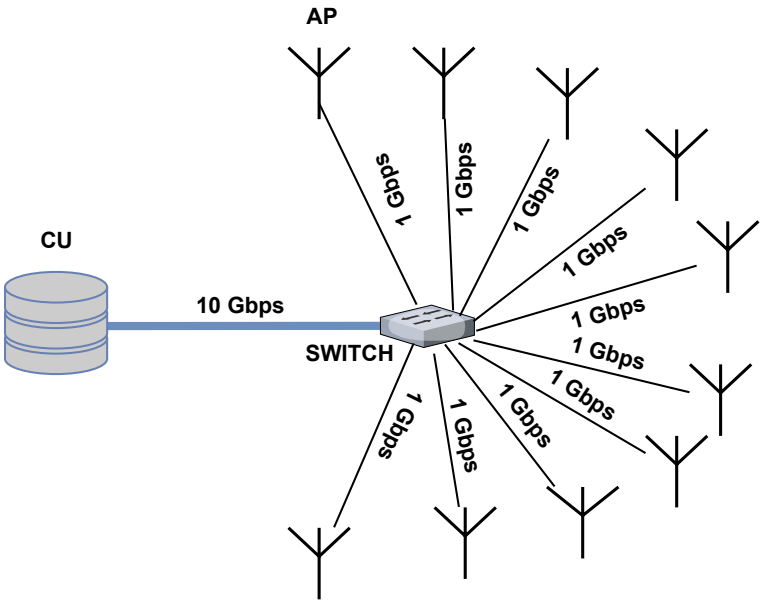


Figura 2.23: Arquitetura de compartilhamento de dados pelo *link ethernet*.

$b_{csi} = 12$ bits permitiria que mais de 65 usuários fossem conectados na rede distribuída usando os mesmos recursos de tempo-frequência. Por outro lado, $b_{csi} = 32$ bits e $b_{csi} = 64$ bits permitiriam 45 e 20 usuários, respectivamente.

A análise também é feita para $N_{SC} = 180$ subportadoras conforme mostrado na Figura 2.19. Considerando o mesmo *link Ethernet* comum de 1 Gbps, as configurações com $b_{csi} = 12, 32 \; e \; 64$ bits permitiriam que mais de 65 usuários fossem conectados na rede distribuída usando os mesmos recursos de tempo-frequência.

Além disso, a taxa no *fronthaul* foi avaliada para diferentes números de APs na rede. É possível verificar o aumento da taxa agregada no *fronthaul* com o aumento do número de APs. As Figuras 2.20 - 2.22 mostram a taxa no *fronthaul* para valores de $N_{SC} = 1200, 600 \; e \; 180$. No estudo, foi considerado o número de $K = 15$ e o número de APs variando de $32 - 128$. Com os gráficos, é possível verificar que nenhuma das opções adotadas de b_{csi} seria possível para servir a todos os UEs com um *link ethernet* de 1 Gbps, ou seja, é impossível atender vários APs com um único *link ethernet*. Contudo, uma arquitetura possível para esse cenário seria construída com a utilização de um equipamento *switch*, o qual conecta uma CU por meio de um *link* de 10 Gbps, ou com velocidade superior, e conecta os APs em outras portas de 1 Gbps. Essa arquitetura é mostrada na Figura 2.23. Dessa maneira, seria possível atender a todos os

APs. Além disso, com os gráficos pode-se perceber que ao considerar o número de UEs fixo, a taxa de dados CSI se mantém a mesma independente do número de APs adotado. As variações ocorrem apenas sobre as taxas de *uplink*.

A Figura 2.20 mostra que a menor taxa de *fronthaul uplink* alcançada com $N_{SC} = 1200$ e $b_{csi} = 12$ é de 16 Gbps, enquanto que para $b_{csi} = 32\ e\ 64$ a taxa é de 30 Gbps e 51 Gbps, respectivamente. A Figura 2.21 mostra que a menor taxa de *fronthaul uplink* alcançada com $N_{SC} = 600$ e $b_{csi} = 12$ é de 8.3 Gbps, enquanto que para $b_{csi} = 32\ e\ 64$ a taxa é de 15 Gbps e 25 Gbps, respectivamente. Por fim, a Figura 2.22 mostra que a menor taxa de *fronthaul uplink* alcançada com $N_{SC} = 180$ e $b_{csi} = 12$ é de 2.5 Gbps, enquanto que para $b_{csi} = 32\ e\ 64$ a taxa é de 4.5 Gbps e 7.7 Gbps, respectivamente.

Os números apresentados nos gráficos mostram que estratégias para comprimir as amostras CSI de forma eficiente são muito importantes em D-MIMO para diminuir o pico de tráfego *fronthaul* durante o *uplink* ou permitir implantações com mais usuários.

Capítulo 3

Esquemas de Compressão do O-RAN para D-MIMO

Este capítulo visa investigar a implementação de alguns métodos de compressão especificados na documentação da aliança O-RAN que apresentam baixa complexidade computacional, os quais foram adaptados para serem usados no cenário de compressão de CSI em sistemas D-MIMO. Além disso, neste capítulo é avaliado o desempenho desses métodos por meio da análise da taxa por usuário e da taxa de pico no *fronthaul* em comparação aos casos onde não há compressão.

3.1 Aliança O-RAN

A Aliança O-RAN é um grande esforço mundial para alcançar novos níveis de abertura em *Virtualized Radio Access Networks* (vRANs) de próxima geração. Lançado inicialmente por cinco grandes operadoras de telefonia móvel há alguns anos, é hoje apoiado por mais de 160 empresas (incluindo 24 operadoras de telefonia móvel em 4 continentes), representando um excelente exemplo de como operadoras e fornecedores em todo o mundo podem colaborar de forma construtiva para definir novos padrões técnicos [39].

O-RAN é um grupo liderado pelas operadoras para definir as vRANs da próxima geração para implantações de vários fornecedores. O objetivo é atualizar o ecossistema vRAN, quebrando o aprisionamento dos fornecedores e abrindo um mercado que tem sido tradicionalmente dominado por um pequeno conjunto de participantes. A alinça O-RAN poderá desencadear um nível de inovação sem precedentes no espaço *Radio Access Network* (RAN), reduzindo

a barreira de entrada no mercado para novos concorrentes [39].

A estrutura de comunicação sem fio tem crescido exponencialmente devido a grande demanda massiva da *Internet of Things* (IoT) e de aplicativos mais avançados que são usados em tempo real para se comunicar com outros equipamentos e usuários. Em virtude disso, grandes empresas estão desenvolvendo tecnologias com capacidade para atender a tais requisitos. A aliança O-RAN surgiu com o conceito de rede de acesso de rádio aberta e possui como grande objetivo oferecer abertura e inteligência aos fornecedores de RAN tradicionais, os quais permitirão que vários fornecedores reformulem a estrutura de RAN e otimizem a rede [40]. Para facilitar o gerenciamento e estruturação da rede, a aliança O-RAN criou dez grupos de trabalho para cobrir toda estrutura da rede.

3.2 Métodos de Compressão em Blocos

Os algoritmos de compressão propostos são baseados nos métodos de compressão de bloco BFP e BS da especificação O-RAN [14], mas com comprimento de informação lateral e tamanho de bloco modificados. Enquanto os métodos do O-RAN são aplicados a vetores de amostras de banda base de rádio 12 IQ ou blocos CSI de antenas co-localizadas, os algoritmos propostos são usados para compactar K amostras complexas de estimação de canal entre K UEs e uma única antena AP, que são enviadas do AP para a CU. Finalmente, neste trabalho, as amostras CSI são assumidas como representadas em pontos fixos.

Os métodos apresentam baixa complexidade, podendo ser aplicados em um tempo muito inferior ao intervalo de coerência dos sistemas de comunicação. Outra vantagem é que pode ser útil reutilizar recursos de *hardware* ou *software* para compressão de dados IQ, ou seja, o recurso previamente desenvolvido para compressão em O-RAN pode ser reutilizado ou facilmente adaptado. Desta forma, os métodos propostos seriam facilmente adaptados ao novo cenário com D-MIMO.

3.2.1 Compressão Block Floating Point

No método de compressão BFP, os K CSIs estimados no m-ésimo AP representado pelo vetor $\dot{\mathbf{g}}_m = [\dot{g}_{1m}, \dot{g}_{2m}, \cdots, \dot{g}_{Km}]$ são convertidos para uma representação de ponto flutuante, onde um único expoente é compartilhado entre todos os coeficientes CSI de cada AP. Além disso, os bits da mantissa são reduzidos a um valor alvo, de modo que os dados compactados

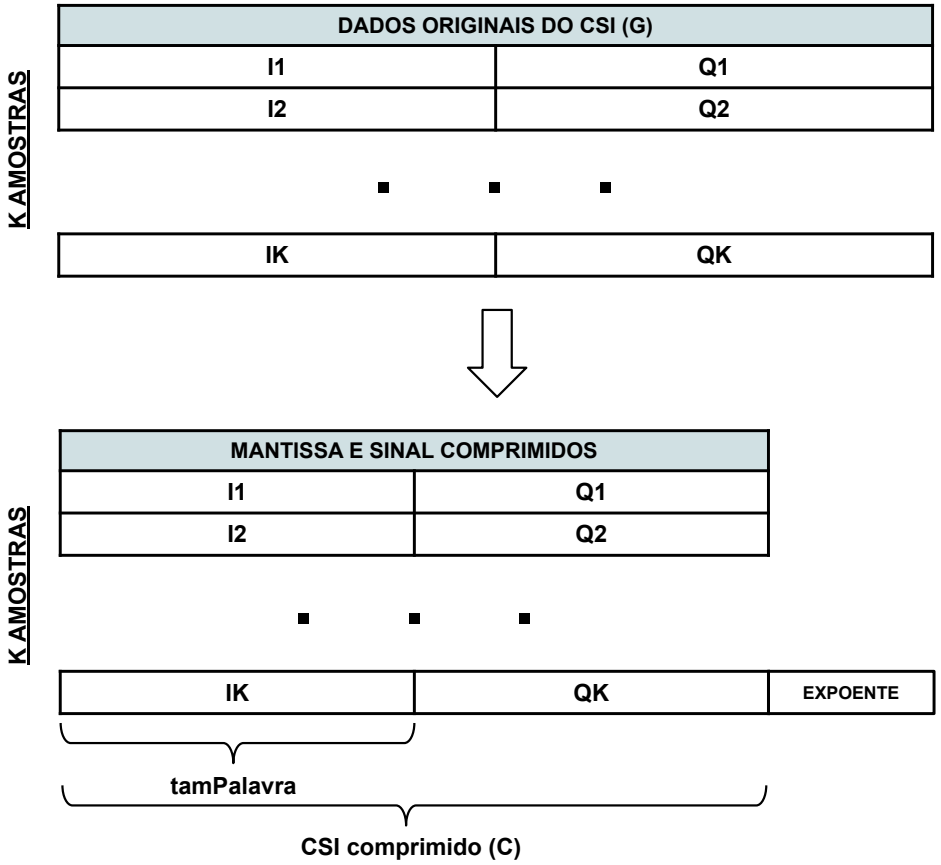


Figura 3.1: Representação do método de compressão BFP.

usam menos bits do que os coeficientes CSI originais. A Figura 3.1 mostra uma visão geral de como o método BFP é realizado.

O Algoritmo 1 mostra uma descrição detalhada do método de compressão. Os componentes reais e imaginários são representados por uma mantissa, um expoente e um sinal. O primeiro passo é encontrar o valor absoluto máximo entre todas as amostras reais e imaginárias dos coeficientes CSI (linha 3). Em seguida, calcula-se um expoente para esse número máximo (linhas 4 e 5). A seguir, o mesmo expoente é usado para representar todos os outros números do CSI como ponto flutuante. Finalmente, os bits do sinal e da mantissa são representados com o número alvo de bits B_{CSI}. As operações de multiplicação nas linhas 7 e 8 significam deslocamento bit a bit para a direita pelo valor do expoente, e as linhas 9 e 10 obtêm o $tamPalavra$ *least significant bits* (LSBs) da informação processada.

A saída do BFP é composta pelo vetor $c_{m1}, c_{m2}, \ldots, c_{mK}$ correspondente às amostras CSI comprimidas K e um expoente de 5 bits como informações secundárias. Assim, o número efetivo de bits por amostra CSI é $b_{CSI,BFP} = B_{CSI} + 5/K$, onde o segundo termo corresponde

Algoritmo 1: Algoritmo de compressão do *Block Floating Point* no m-ésimo AP

Entrada: $\dot{g}_{m1}, \dot{g}_{m2}, \ldots, \dot{g}_{mK}, B_{CSI}$

Saída: $c_{m1}, c_{m2}, \ldots, c_{mK}$, *expoente*

1 **início**
2 *tamPalavra* $\leftarrow B_{CSI}/2$
3 *maxValor* $\leftarrow \max(\max(|\text{Re}(\dot{g}_m)|), \max(|\text{Im}(\dot{g}_m)|))$
4 *arredExp* $\leftarrow \lfloor \log_2(\textit{maxValor}) + 1 \rfloor$
5 *expoente* $\leftarrow \max(\textit{arredExp} - \textit{tamPalavra}, 0)$
6 **para todo** *i* $\leftarrow 1$ *até* K **faça**
7 *tmpRe* $\leftarrow \text{Re}(\dot{g}_{mi}) \times 2^{-\textit{expoente}}$
8 *tmpIm* $\leftarrow \text{Im}(\dot{g}_{mi}) \times 2^{-\textit{expoente}}$
9 $\text{Re}(c_{mi}) \leftarrow$ *tmpRe{tamPalavra:1}*
10 $\text{Im}(c_{mi}) \leftarrow$ *tmpIm{tamPalavra:1}*
11 **fim**
12 **fim**

Algoritmo 2: Algoritmo de descompressão do Block Floating Point em CU recebendo o CSI comprimido do m-ésimo AP

Entrada: $c_{m1}, c_{m2}, \ldots, c_{mK}$, *expoente*

Saída: $\hat{g}_{m1}, \hat{g}_{m2}, \ldots, \hat{g}_{mK}$

1 **início**
2 **para todo** *i* $\leftarrow 1$ *até* K **faça**
3 $\text{Re}(\hat{g}_{mi}) \leftarrow \text{Re}(c_{mi}) \times 2^{\textit{expoente}}$
4 $\text{Im}(\hat{g}_{mi}) \leftarrow \text{Im}(c_{mi}) \times 2^{\textit{expoente}}$
5 **fim**
6 **fim**

à informação secundária dividido pelo número de amostras em um bloco. Observe a diferença entre b_{CSI} e B_{CSI}.

Em seguida, os elementos comprimidos e o expoente são enviados ao decodificador, que realiza todas as operações reversas. O vetor $\hat{g}_{m1}, \hat{g}_{m2}, \ldots, \hat{g}_{mK}$ correspondente ao K descompactado estimado CSIs pode ser recuperado com um reescalonamento, conforme mostrado nas linhas 3 e 4 em Algoritmo 2.

Uma característica notável da compressão BFP é que ela pode ser sem perdas se o expoente for zero. Isso acontece quando a amplitude dos elementos no CSI original é relativamente baixa, ou B_{CSI} é relativamente alta.

3.2.2 Compressão Block Scaling

O método de compressão BS escalona todos os K CSIs complexos estimados no m-ésimo AP representado pelo vetor $\dot{\mathbf{g}}_m = [\dot{g}_{1m}, \dot{g}_{2m}, \cdots, \dot{g}_{Km}]$ (semelhante ao método anterior) e quantiza as amostras escalonadas, conforme mostrado no Algoritmo 3. O primeiro passo é encontrar o valor absoluto máximo entre todos os componentes reais e imaginários do CSI, ao qual nos referiremos daqui em diante como *maxValor*. O segundo passo é quantizar o *maxValor* (linha 3) com 10 bits. Um *lookup table* (LUT) pode implementar eficientemente a quantização com duas saídas: o *escalarQuant* (a versão quantizada de *maxValor*) e o *escalar* representando o *escalarQuant* com 10 bits .

As seguintes etapas no Algoritmo 3 são: escalar todos os símbolos IQ do CSI (linhas 6 e 7) e quantizar a versão escalonada do CSI (linhas 8 e 9). Cada parte real e imaginária é quantizada com *tamPalavra* bits que é metade do B_{CSI}. Assim, cada amostra comprimida é representada com B_{CSI} bits. Como *escalarQuant* $\approx 1/$*maxValor*, todos os valores escalonados estão entre -1 e 1. Assim, um único intervalo de quantização de $[-1, 1]$ pode ser usado em todos os cenários. Após a quantização, os dados IQ comprimidos e as informações secundárias são transportados pelo *fronthaul*, onde o número efetivo de bits por amostras CSI é $b_{CSI,BS} = B_{CSI} + 10/K$.

No decodificador, todas as etapas são invertidas, conforme mostrado em Algoritmo 4. O primeiro passo é recuperar o valor de escala que trará de volta os valores aos níveis originais. Então, todos os valores são multiplicados pelo valor de escala e uma aproximação dos valores originais é recuperada. Uma visão completa do método de compressão BS é mostrada na Figura 3.2.

3.3 Resultados da Simulação com os Métodos de Compressão

Dois cenários D-MIMO diferentes foram implementados neste trabalho. No primeiro cenário, avaliamos o impacto no desempenho do sistema da compressão do CSI com diferentes números de bits sob uma configuração específica do sistema de $M = 128$ APs com única antena e $K = 16$ UEs distribuídos aleatoriamente na área. Na segunda, com a análise geral do impacto

Algoritmo 3: Algoritmo de compressão *Block Scaling* no m-ésimo AP

Entrada: $\dot{g}_{m1}, \dot{g}_{m2}, \ldots, \dot{g}_{mK}, B_{CSI}$

Saída: $c_{m1}, c_{m2}, \ldots, c_{mK}$, *escalar*

início
  tamPalavra $\leftarrow B_{CSI}/2$
  maxValor $\leftarrow \max(\max(|\text{Re}(\dot{g}_m)|), \max(|\text{Im}(\dot{g}_m)|))$
  $[$*escalarQuant*, *escalar*$]$ $\leftarrow$ LUT(*maxValor*)
  para todo *i* $\leftarrow 1$ *até* K **faça**
    reEscalonado $\leftarrow \text{Re}(\dot{g}_{mi})/$*escalarQuant*
    imEscalonado $\leftarrow \text{Im}(\dot{g}_{mi})/$*escalarQuant*
    $\text{Re}(c_{mi}) \leftarrow$ quantiza(*reEscalonado,tamPalavra*)
    $\text{Im}(c_{mi}) \leftarrow$ quantiza(*imEscalonado,tamPalavra*)
  fim
fim

Algoritmo 4: Algoritmo de descompressão Block Scaling em CU recebendo o CSI compactado do m-ésimo AP

Entrada: $c_{m1}, c_{m2}, \ldots, c_{mK}$, *escalar*

Saída: $\hat{g}_{m1}, \hat{g}_{m2}, \ldots, \hat{g}_{mK}$

início
  escalarQuant $\leftarrow$ LUT2(*escalar*)
  para todo *i* $\leftarrow 1$ *até* K **faça**
    $\text{Re}(\hat{g}_{mi}) \leftarrow \text{Re}(c_{mi}) \times$ *escalarQuant*
    $\text{Im}(\hat{g}_{mi}) \leftarrow \text{Im}(c_{mi}) \times$ *escalarQuant*
  fim
fim

no desempenho já abordada, avaliamos a escalabilidade do sistema fixando o número de bits utilizados durante a compressão, variando o número de UEs (K) e o número de APs (M) na simulação.

Em ambas as simulações, consideramos configurações semelhantes a [7, 11] com uma área de $2 \times 2\ \text{km}^2$ para evitar efeitos de contorno. As posições AP são fixas durante a simulação, mas as UEs mudam de posição a cada realização do canal. Para o coeficiente de desvanecimento

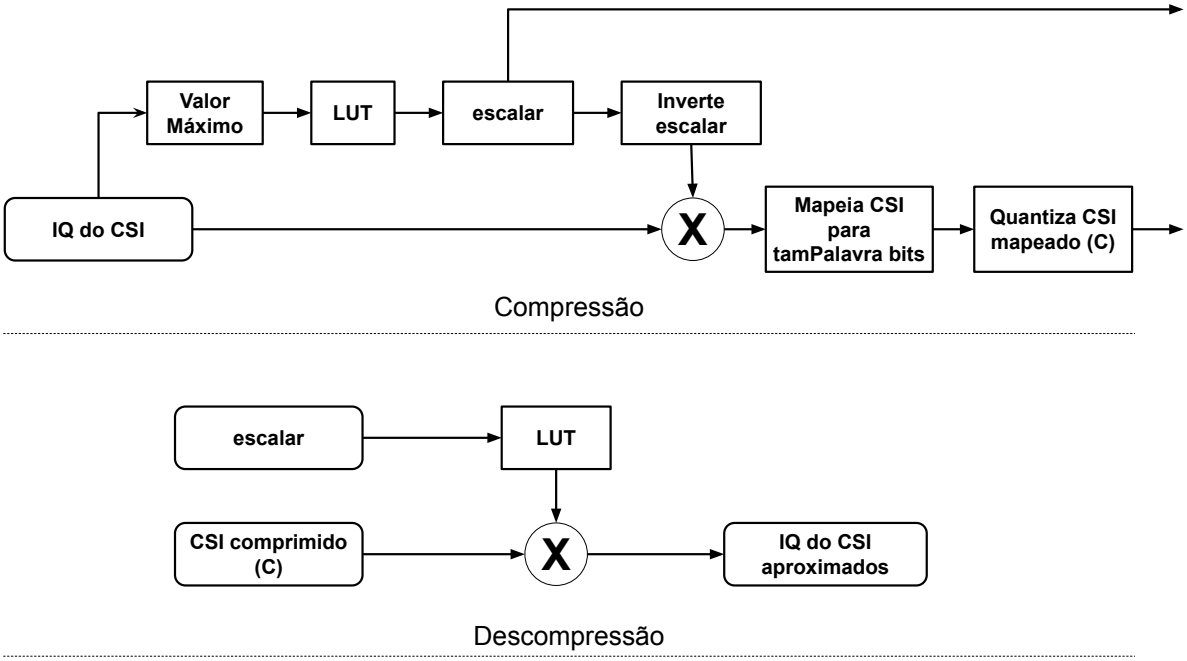


Figura 3.2: Representação do método de compressão BS.

em grande escala, o modelo *COST Hata* é usado: $10\log_{10}(\beta_{km}) = -136 - 35\log_{10}(d_{km}) + X_{km}$, onde d_{mk} é a distância entre a antena m e UE k em km, $X_{km} \sim \mathcal{N}(0, \sigma^2_{\text{shadow}})$, com $\sigma^2_{\text{shadow}} = 8\,\mathrm{dB}$. A variação do ruído no receptor é considerada $\sigma^2_w = 290 \times B_{\text{c}} \times \text{BW} \times \text{NF}$, onde $290\,\mathrm{K}$ é a temperatura ambiente, $B_{\text{c}} = 1.3807 \times 10^{-23}\,\mathrm{J\,K^{-1}}$ é a constante de Boltzmann, $\text{BW} = 20\,\mathrm{MHz}$ é a largura de banda disponível e $\text{NF} = 9\,\mathrm{dB}$ é o valor do ruído, respectivamente. A pré-codificação e a alocação de potência utilizadas na simulação são descritas em (2.2) e (2.3), respectivamente. A potência máxima transmitida em cada AP é $\rho_d = 200\,\mathrm{mW}$, o esquema de modulação digital usado é 16-QAM com potência unitária média, e a sequência do piloto transmitida é mutuamente ortogonal para evitar a contaminação do piloto. Para a simulação, a taxa alcançável *downlink* do k-ésimo usuário é $R_k = \log_2(1 + \text{SINR}_k)$ in $\mathrm{bit/s/Hz}$, e é avaliado em termos de *Cumulative Distribution Function* (CDF). Assumimos uma estimação de canal perfeita no UEs.

A Figura 3.3 apresenta a análise do *Normalised Mean Square Error* (NMSE) dos métodos de compressão BFP e BS para diferentes valores de bits de compressão por amostra CSI (b_{CSI}) destacando a distorção de compressão inserida no sistema. O NMSE em dB é calculado conforme a equação $NMSE_{dB} = 10\log_{10}\left(\frac{\|\dot{G} - \tilde{G}\|_2^2}{\|\tilde{G}\|_2^2}\right)$, onde $\dot{G}$ representa o canal antes de ser comprimido e o $\tilde{G}$ é o canal após a compressão. A avaliação abrange a faixa de 2 a 32 bits. A

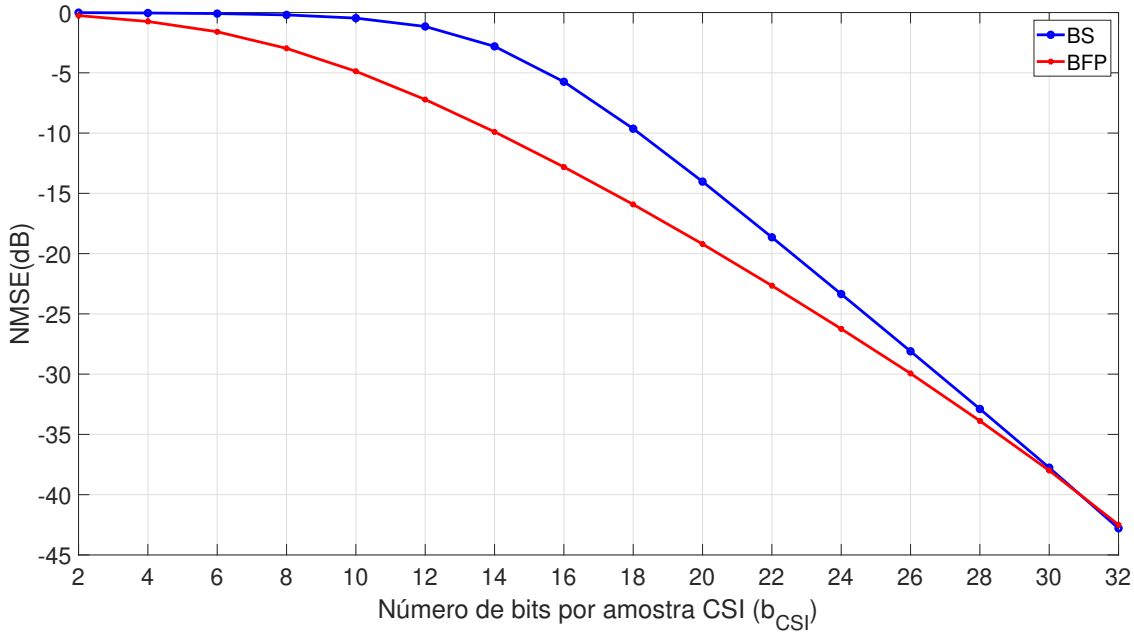


Figura 3.3: Análise do NMSE de compressão dos métodos de compressão BFP e BS para diferentes números de bits de compressão por amostra CSI.

análise revela que o método BFP geralmente exibe menor erro em comparação com o método BS, especialmente para valores menores de bits. Por exemplo, para 14 bits, a diferença de erro chega a cerca de 7 dB. Para valores mais altos, os desempenhos dos métodos se aproximam.

As Figuras 3.4 - 3.7 mostram a CDF da taxa de usuário alcançável sem compressão e para diferentes configurações de compressão, considerando diferentes números de APs. As Figuras mostram que a eficiência espectral tende a diminuir com a diminuição do número de APs na rede. Além disso, é possível verificar que o método BFP é virtualmente sem perda quando usa um B_{CSI} relativamente alto (16 bits) ou (12 bits) dependendo do número de APs adotados na rede. Já o método BS consegue ser sem perda adotando B_{CSI} igual ou superior a 16 bits. Curiosamente, o desempenho do BFP é sempre superior ao do BS ao usar a mesma quantidade de bits para comprimir as amostras, o que é mais perceptível quando $B_{CSI} = 12$ e 8 bits.

Por exemplo, na Figura 3.4, a taxa do 5º percentil sem compressão é aproximadamente $5.2\,\mathrm{bit/s/Hz}$, e para o mesmo percentil, o BFP atinge 5.2, 5.1 e 4.6, e o BS atinge aproximadamente 5.2, 4.8 e 2.3, para $B_{CSI} = 16, 12$ e 8 bits, respectivamente, considerando $M = 128$. Para $M = 96$ (Figura 3.5), a taxa do 5º percentil sem compressão é aproximadamente $4.4\,\mathrm{bit/s/Hz}$, e para o mesmo percentil, o BFP atinge 4.4, 4.4 e 3.8, e o BS atinge aproximadamente 4.4, 3.8 e 1.2, para $B_{CSI} = 16, 12$ e 8 bits, respectivamente. Para $M = 64$ (Fi-

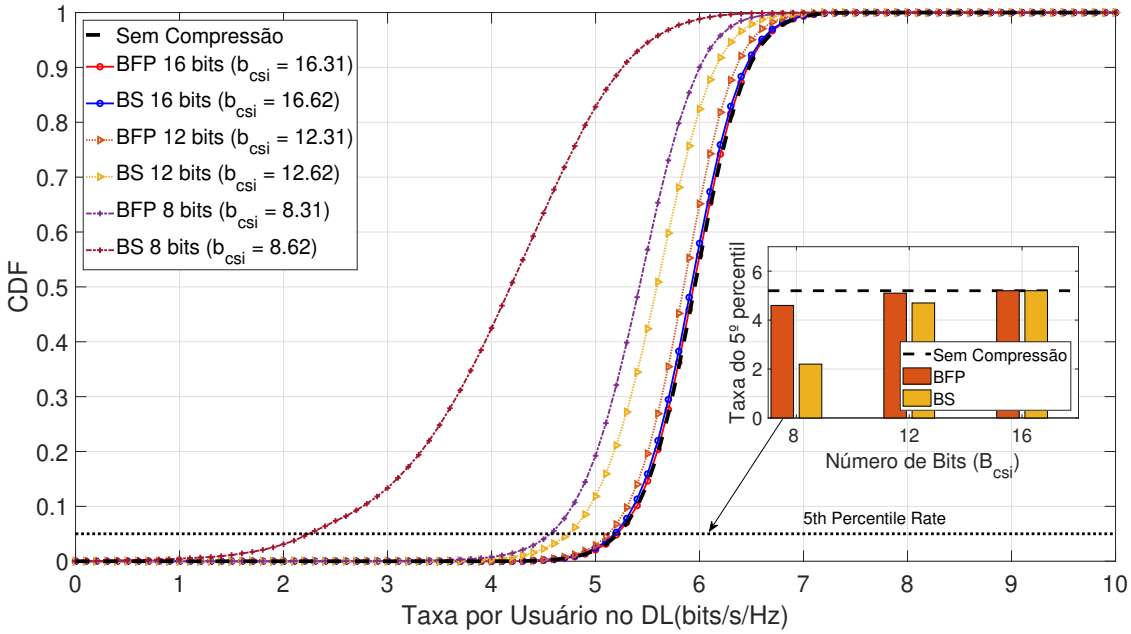


Figura 3.4: CDF da taxa de usuário alcançável para cada método de compressão e para $B_{CSI} = 16, 12$ e 8 bits considerando $M = 128$ APs.

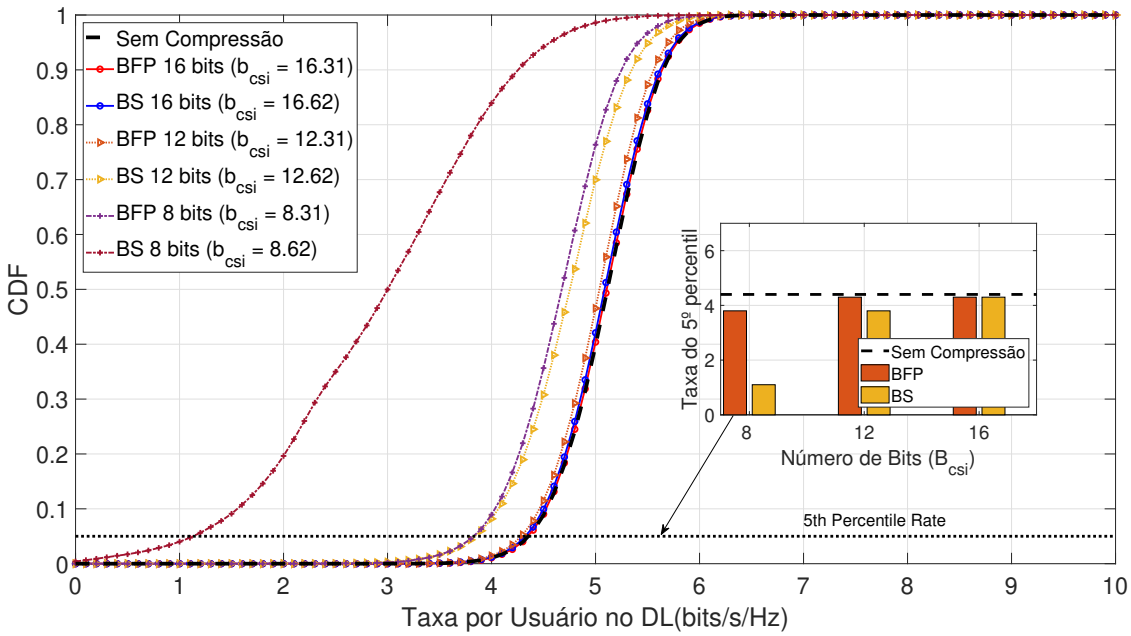


Figura 3.5: CDF da taxa de usuário alcançável para cada método de compressão e para $B_{CSI} = 16, 12$ e 8 bits considerando $M = 96$ APs.

gura 3.6), a taxa do 5º percentil sem compressão é aproximadamente $3.2\,\mathrm{bit/s/Hz}$, e para o mesmo percentil, o BFP atinge 3.2, 3.2 e 2.8, e o BS atinge aproximadamente 3.2, 2.6 e 0.4,

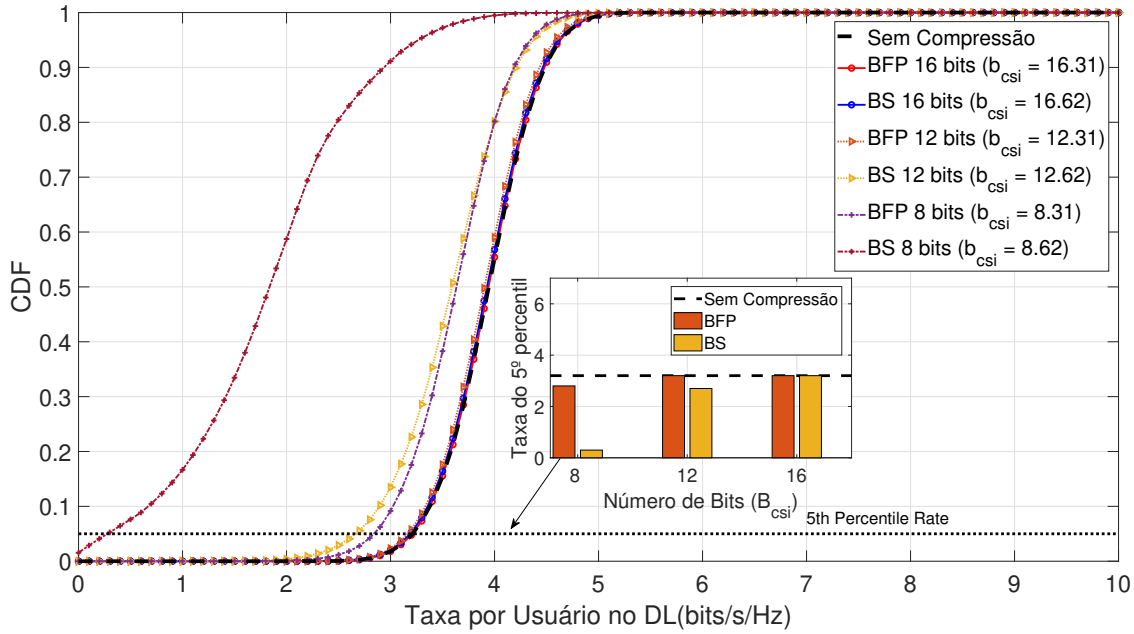


Figura 3.6: CDF da taxa de usuário alcançável para cada método de compressão e para $B_{CSI} = 16, 12$ e 8 bits considerando $M = 64$ APs.

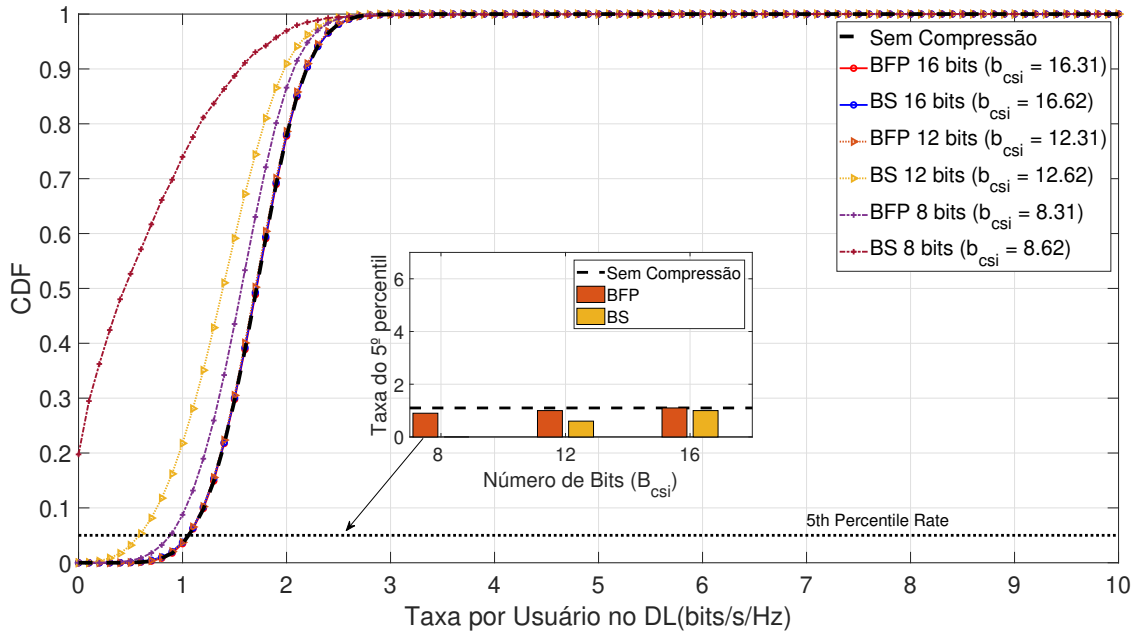


Figura 3.7: CDF da taxa de usuário alcançável para cada método de compressão e para $B_{CSI} = 16, 12$ e 8 bits considerando $M = 32$ APs.

para $B_{CSI} = 16, 12$ e 8 bits, respectivamente. Para $M = 32$ (Figura 3.7), a taxa do 5º percentil sem compressão é aproximadamente $1.1\,\mathrm{bit/s/Hz}$, e para o mesmo percentil, o BFP atinge 1.1,

1 e 0.9, e o BS atinge aproximadamente 1, 0.6 e 0, para $B_{CSI} = 16, 12$ e 8 bits, respectivamente.

Os resultados do segundo experimento são mostrados na Figura 3.8, onde comparamos o desempenho do sistema D-MIMO com os métodos de compressão propostos e sem comprimir o CSI. Na Figura, os gráficos de *boxplot* mostram as estatísticas da taxa por usuário no *downlink* não agregada (superior) e agregada (inferior), onde a mediana é a linha horizontal dentro da caixa que abrange o intervalo interquartil (do 25º ao 75º percentil). Os pontos da curva se estendem do 5º ao 95º percentil da taxa de usuários *downlink*. Os resultados são agrupados pelo número de UEs (superior) e pelo número de APs (inferior) na simulação. A Figura mostra que o BFP supera o BS sob a mesma configuração B_{CSI}, o que é válido ao aumentar o número de UEs. Além disso, à medida que o número de UEs cresce e o número de APs diminui, o desempenho geral da rede diminui continuamente, mesmo sem compressão. Ao empregar compressão, para valores de $B_{CSI} \geq 16$, o desempenho torna-se próximo ao caso sem compressão para baixos valores de UEs e de APs. A diferença começa a ficar mais acentuada com o aumento de UEs e de APs na rede. É perceptível que com muitos UEs e com muitos APs na rede, a taxa de usuário alcançável pode ser menor do que no caso sem compressão. Por exemplo, com $K = 35$ UEs e $B_{CSI} = 16$ bits, a compressão BS causa uma perda de aproximadamente 0.8 bit/s/Hz no 5º percentil quando comparada ao caso sem compressão. Com $M = 80$ APs e $B_{CSI} = 16$ bits, a compressão BS causa uma perda de aproximadamente 0.4 bit/s/Hz no 5º percentil quando comparada ao caso sem compressão.

Por fim, a Figura 3.9 também mostra o tráfego de pico no *fronthaul* por AP versus o número de UEs conectados no sistema (superior) e o tráfego de pico no *fronthaul* agregado versus o número de APs (inferior), considerando os mesmos parâmetros da Seção 2.5.2 e os valores de $B_{CSI} = 16$ e 12 bits. Na Figura inferior, foi considerado $K = 15$ UEs. Nessas análises, assume-se que o CSI sem compressão é representado com 64 bits, ou seja, as partes real e imaginária são representadas em ponto flutuante de precisão simples (32 bits). Na Figura superior, se considerarmos o caso com $K = 25$ UEs e BFP com 16 bits, a taxa no rádio *downlink* foi aproximadamente a mesma do caso sem compressão, e o tráfego de pico no *fronthaul* foi reduzido de aproximadamente 1.6 Gbps. Na Figura inferior, para o caso com $M = 80$ UEs e BFP com 16 bits, a taxa no rádio *downlink* também foi aproximadamente a mesma do caso sem compressão, e o tráfego de pico no *fronthaul* foi reduzido de aproximadamente 79 Gbps.

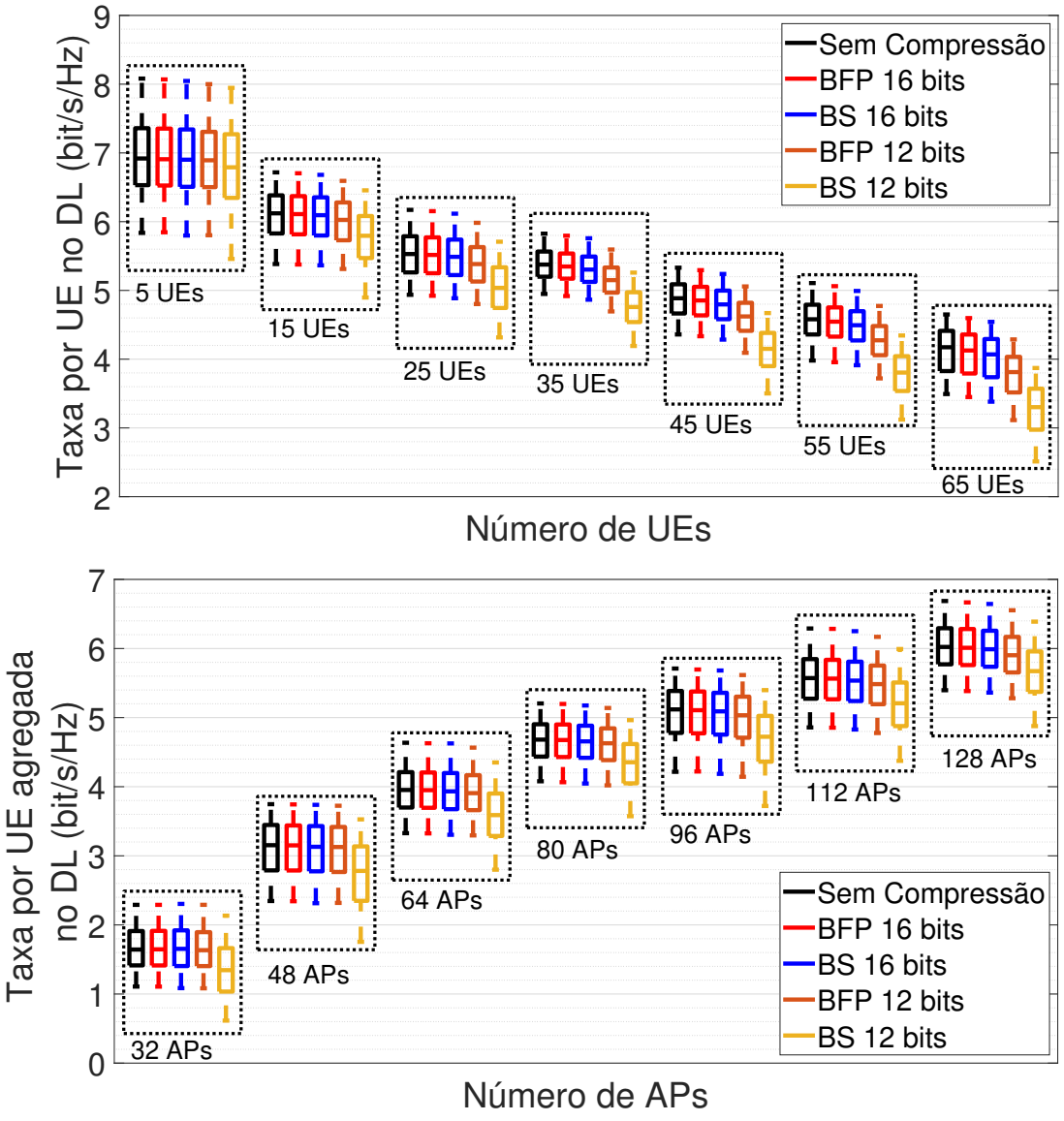


Figura 3.8: Taxa de usuário alcançável em cenários com diferentes números de UEs (superior) e taxa de usuário agregada alcançável em cenários com diferentes números de APs (inferior) sem compressão e com diferentes configurações de compressão.

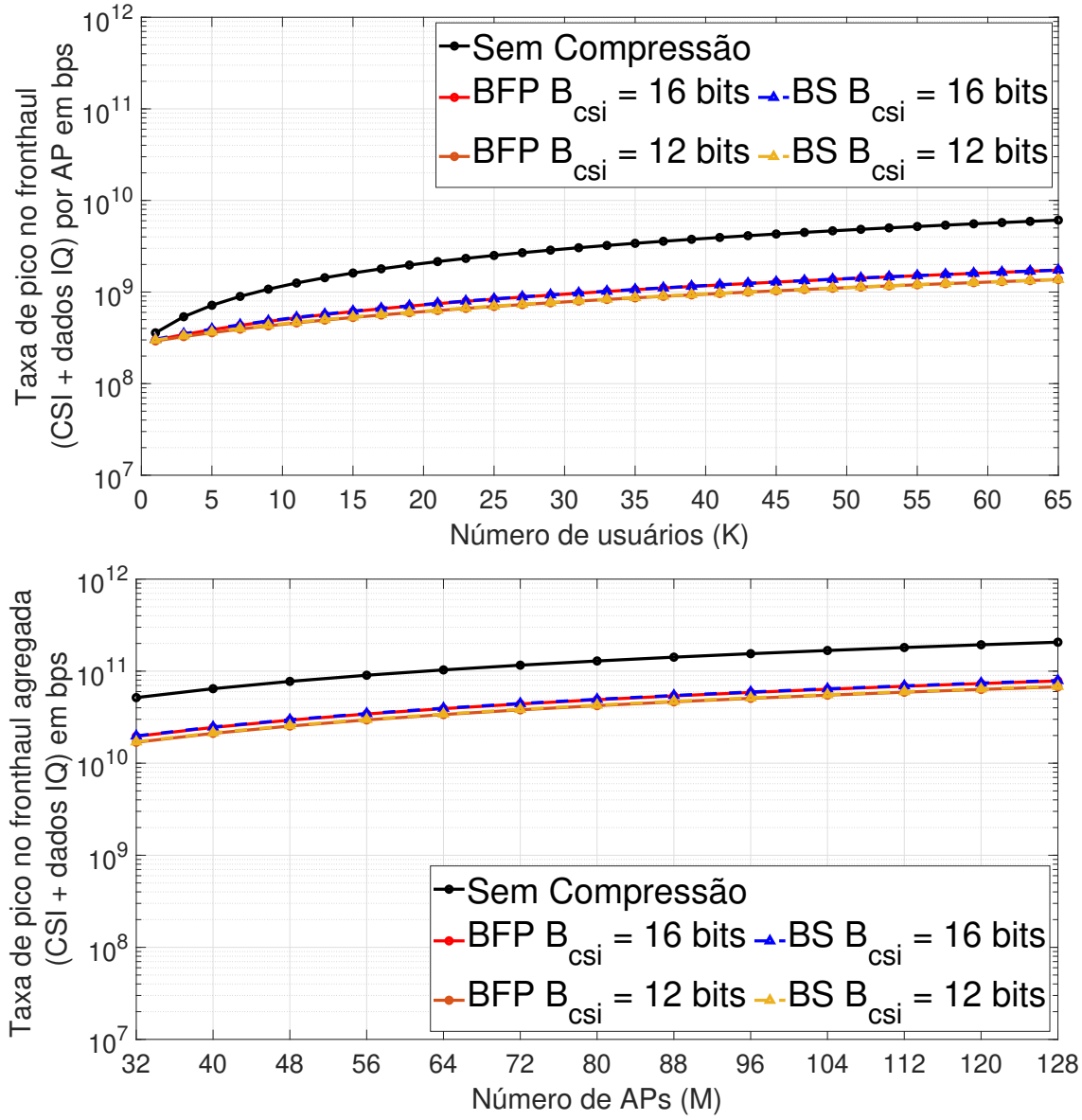


Figura 3.9: Tráfego no *fronthaul* alcançável por AP em cenários com diferentes números de UEs (superior) e tráfego no *fronthaul* alcançável agregado em cenários com diferentes números de APs (inferior) sem compressão e com diferentes configurações de compressão.

Capítulo 4

Desempenho do D-MIMO em Sistemas 5G/NR

Este capítulo visa investigar a implementação do D-MIMO na arquitetura do 5G/NR, assim como avaliar o seu desempenho com a aplicação de compressão de CSI em comparação ao cenário sem compressão, abrangendo taxas de dados no *fronthaul* e taxa de transferência de rádio. No capítulo é investigada a implementação dos métodos de compressão de CSI sobre o D-MIMO em sistema 5G/NR. O capítulo mostrará o modelo do sistema adotado, incluindo o cenário de implantação, *functional split*, pré-codificação, alocação de potência e transmissão dos símbolos pré-codificados. Por fim, serão mostrados resultados de desempeno baseados em simulações realizadas no cenário considerado.

4.1 Modelo do Sistema

Considerando a arquitetura O-RAN [14], onde uma O-DU pode controlar múltiplas O-RUs. As CUs e APs comumente referidos em redes *cell-free* podem ser implementadas com O-DUs e O-RUs, respectivamente. Assim, as antenas distribuídas são denominadas O-RU e a entidade centralizada responsável pela pré-codificação e alocação de potência é referida como O-DU. Neste cenário, o O-DU, juntamente com as O-RUs, implementa a camada PHY de um *Next Generation NodeB* (gNB).

Assumimos uma abordagem centrada em célula, onde as O-RUs são divididas em células disjuntas, cada uma coordenada por uma única O-DU. Consequentemente, os UEs percebem células maiores, onde a interferência entre as antenas da mesma célula é diminuída através da

pré-codificação pela O-DU.

A gNB é implementada com *functional split* 7.2 [14], no qual a O-RU lida com o processamento de PHY baixo para *uplink*/*downlink*, abrange tarefas como FFT e *Inverse Fast Fourier Transform* (IFFT), remoção/inserção de *Cyclic Prefix* (CP), conversão de analógico para digital / digital para analógico e todos os outros circuitos de *Radio Frequency* (RF) restantes.

Este estudo novamente considera um cenário onde uma célula consiste em M antenas distribuídas (M O-RUs) controladas por uma única O-DU que atende K UEs com única antena sobre o mesmo recurso de tempo-frequência, onde $K \ll M$. Focando na transmissão *downlink* de símbolos OFDM, os símbolos recebidos no n-ésimo RE podem ser representados como

$$\mathbf{y} = \mathbf{G}\mathbf{x} + \mathbf{w}, \tag{4.1}$$

$$\mathbf{y} = \mathbf{G}\mathbf{A}\mathbf{s} + \mathbf{w}, \tag{4.2}$$

onde $\mathbf{y} \in \mathbb{C}^{K\times 1}$ representa os símbolos, $\mathbf{G} \in \mathbb{C}^{K\times M}$ é a matriz de canal entre as portas de antena distribuídas da gNB e as portas de antena dos UEs, $\mathbf{x} \in \mathbb{C}^{M\times 1}$ são os símbolos transmitidos pré-codificados no *downlink* em cada porta de antena, e $\mathbf{w} \sim \mathcal{N}_{\mathbb{C}}(0, I_K\sigma_w^2)$ é o vetor de ruído. Assumindo K símbolos por RE para serem transmitidos para os usuários, uma matriz de pré-codificação $\mathbf{A} \in \mathbb{C}^{M\times K}$ executa alocação de potência e mitiga interferências entre os símbolos *downlink*. Assim, os símbolos pré-codificados transmitidos nas portas de antena são dadas por $\mathbf{x} = \mathbf{A}\times\mathbf{s}$, onde $\mathbf{s} \in \mathbb{C}^{K\times 1}$ representam os símbolos *Quadrature Amplitude Modulation* (QAM) antes da pré-codificação, designados para transmissão sobre a n-ésima subportadora.

Cada elemento de $\mathbf{G}$ pode ser modelado como uma combinação de *large scale fading* β_{km} e *small scale fading* h_{kmn}, entre o k UE e m O-RU sobre a subportadora n: $g_{kmn} = \sqrt{\beta_{km}}h_{kmn}$, onde $h_{kmn} \sim \mathcal{N}_{\mathbb{C}}(0, 1)$. O subscrito n é omitido no componente *large scale* porque é o mesmo em todas as subportadoras. Além disso, existe um nível de correlação entre o *small scale fading* de subportadoras vizinhas devido à largura de banda de coerência. Na discussão restante, o subscrito n que representa o índice de subportadoras será omitido.

O pré-codificador é derivado do canal $\mathbf{G}$ e, no contexto do NR, um único pré-codificador pode ser aplicado para um conjunto de REs ou um grupo de *Physical Resource Blocks* (PRBs) [4]. A gNB pode encontrar a matriz $\mathbf{G}$ utilizando o conceito de reciprocidade de canal, que aproveita a simetria entre os canais *uplink* e *downlink*, possibilitando a estimação do canal *downlink* com base no canal *uplink* medido. Isto pode ser conseguido em NR implementando TDD. Neste

caso, o *slot* NR é dividido em 14 símbolos OFDM, onde o escalonador define o número de símbolos OFDM alocados para *downlink*, *uplink* e intervalo de guarda. Neste trabalho, assume-se uma configuração TDD, com nove símbolos *downlink* OFDM, quatro *uplink* e um símbolo de guarda posicionados entre as transmissões *downlink* e *uplink*.

Deve-se notar que o padrão NR acomoda diversas configurações de s*lots*, incluindo aquelas que consistem exclusivamente em símbolos *downlink* ou *uplink*, bem como configurações com diferentes números de símbolos OFDM *downlink* e *uplink* dentro do mesmo *slot*. A escolha da configuração adotada, ilustrada na Figura 4.1, garante que os pilotos *uplink* estejam próximos no tempo dos símbolos OFDM pré-codificados no *downlink*, minimizando assim a discrepância entre a estimação do canal e as condições reais do canal. Na investigação proposta, os símbolos *downlink* são usados para transmitir dados do usuário *downlink* através do PDSCH e transmissão de DM-RS para demodulação coerente do PDSCH. No cenário investigado, os símbolos OFDM de *uplink* são usados para transmitir apenas pilotos na forma de SRSs. Dessa maneira, o tráfego de dados de rádio do *uplink* não é avaliado.

Uma visão geral abrangente de todo o processo, desde a transmissão piloto em *uplink* até a transmissão de sinais *downlink* pré-codificados, é mostrada na Figura 4.2. Inicialmente, todos os UEs conectados à célula transmitem pilotos na forma de SRS em *uplink*. Em seguida, é feito o processamento da PHY baixa, abrangendo tarefas como remoção de CP e FFT. Posteriormente, com base nos sinais dos K UEs, as M O-RUs realizam a estimação de canal para a matriz $\mathbf{G}$, representando os enlaces de comunicação entre o gNB e os UEs. Então, os K canais estimados pelas O-RUs são comprimidos com os métodos de compressão detalhados na Seção 3.2 e são enviados para a O-DU. Posteriormente, na O-DU, os canais estimados são descomprimidos e, em seguida, o pré-codificador *downlink*

$$\mathbf{A} = \mathbf{P}\mathbf{A}', \tag{4.3}$$

é derivado, onde o componente $\mathbf{A}' \in \mathbb{C}^{M \times K}$ aborda a mitigação das interferências, enquanto $\mathbf{P} \in \mathbb{R}^{M \times M}$ controla a potência sobre as portas de antena. É importante destacar na Seção 2.5, o pré-codificador *downlink* é calculado conforme $\mathbf{A} = \mathbf{A}'\mathbf{P}$, dessa maneira, o produto matricial entre a matriz de canal $\mathbf{G}$ e $\mathbf{A}'$ resulta na matriz identidade $\mathbf{I}_K$ ($\mathbf{G}\mathbf{A}' = \mathbf{I}_K$).

Neste novo cenário, o cálculo do pré-codificador *downlink* é calculado conforme $\mathbf{A} = \mathbf{P}\mathbf{A}'$. Ao aplicar o produto matricial do pré-codificador com o canal, o cálculo não resulta na matriz identidade $\mathbf{G}\mathbf{P}\mathbf{A}' \neq \mathbf{I}_K$.

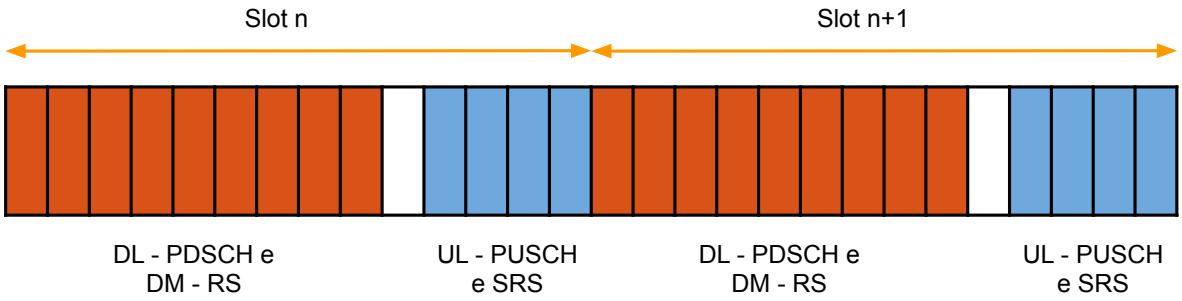


Figura 4.1: Exemplo de Slot TDD NR composto por 14 símbolos OFDM, onde 9 símbolos são alocados para *downlink* e 4 para *uplink*.

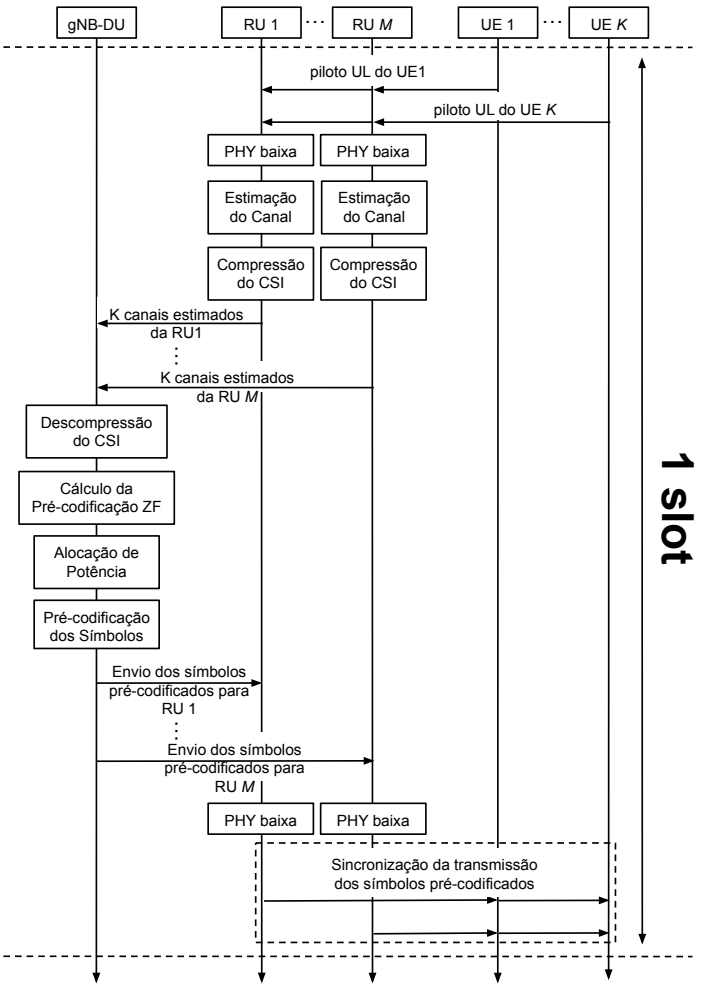


Figura 4.2: Diagrama de sequência ilustrando as tarefas e sinalização na rede.

O passo final realizado na O-DU é a pré-codificação

$$\mathbf{x} = \mathbf{A}\mathbf{s}, \tag{4.4}$$

onde $\mathbf{s}$ representa os símbolos QAM a serem transmitidos para os UEs. Cada elemento de $\mathbf{s}$ mantém unidade média de potência ($\mathbf{E}[|s_k|^2]$) para $\mathbf{k} = 1, \cdots, K$. O mesmo pré-codificador é empregado na parte do *downlink* do restante do *slot* do TDD.

Alguns algoritmos existem para realizar a mitigação de interferência e alocação de potência [7], requerendo avaliação de ambos desempenho e complexidade computacional para garantir viabilidade dos sistemas. Além disso, considerando a centralização na computação de ambos os componentes do pré-codificador, a O-DU tem o potencial de melhorar o desempenho geral do sistema.

Por exemplo, algoritmos como *Zero-Forcing* podem ser implementados, muitas vezes resultando em melhor desempenho em termos de SE. No caso do *Zero-Forcing*, o pré-codificador pode ser calculado como a pseudo-inversa da matriz de canal estimada:

$$\mathbf{A}' = \hat{\mathbf{G}}^H(\hat{\mathbf{G}}\hat{\mathbf{G}}^H)^{-1}, \tag{4.5}$$

onde $\hat{\mathbf{G}}$ é o canal estimado e descomprimido.

Depois de determinar a matriz $\mathbf{A}'$, os coeficientes da matriz $\mathbf{P}$ podem ser encontrados para ajustar a potência de transmissão em cada porta da antena. Neste trabalho, empregamos uma alocação de potência de baixa complexidade, onde todas as antenas transmitem com igual potência ρ_d:

$$\mathbf{P} = \mathbf{diag}(p_1, \cdots, p_M), \tag{4.6a}$$

$$p_m = \sqrt{\frac{\rho_d}{\sum_{i=1}^{K} |a'_{mk}|^2}}, m = 1, \cdots, M \tag{4.6b}$$

onde a'_{mk} são os elementos do pré-codificador *Zero-Forcing* $\mathbf{A}'$.

Após o cálculo de (4.5) e da alocação de potência em (4.6), a matriz de pré-codificação completa pode ser determinada usando (4.3), e a pré-codificação dos símbolos QAM é implementada com (4.4). O NR já suporta as operações de pré-codificação descritas. Na camada NR PHY, considerando o *Functional Split* 7.2, o O-DU lida com níveis mais elevados da camada física (por exemplo, mapeador de modulação, mapeador de camada e pré-codificação), enquanto

o O-RU gerencia funções inferiores. Neste caso, K *codewords* são convertidas em símbolos QAM. Posteriormente, a pré-codificação é realizada e os K símbolos QAM pré-codificados são transportados através de M portas de antena usando a mesma subportadora e símbolo OFDM.

Considerando um cenário com sinais OFDM e o D-MIMO com pré-codificação e alocação de potência totalmente centralizadas, no início de cada intervalo de coerência, os K UEs enviam pilotos ortogonais para os M O-RUs para estimação de canal. Então, cada O-RU envia $K \times \frac{N_{\text{SC}}}{C_{\text{BW}}}$ canais estimados para a O-DU, onde N_{SC} é o número de subportadoras do sinal multiportadora e C_{BW} é o número de subportadoras na largura de banda de coerência. Então, a O-DU calcula os coeficientes de pré-codificação, a alocação de potência, realiza a pré-codificação dos símbolos e envia os símbolos pré-codificados para cada O-RU. Finalmente, a O-RU envia os símbolos pré-codificados para os UEs pelo ar. Em resumo, a pré-codificação de símbolos, o transporte *fronthaul* e a transmissão aérea são repetidos para cada símbolo OFDM durante o intervalo de coerência. Em contraste, a estimação de CSI e seu transporte para a O-DU são feitos uma vez a cada intervalo de coerência.

Considerando uma O-RU que implementa o *functional split* 7.1 [33], onde apenas a representação no domínio da frequência é trocada com a O-DU, o tráfego *fronthaul* necessário para transmitir *downlink* ou *uplink* IQ amostras para cada O-RU podem ser escritas como $R_{IQ}^{DL} = R_{IQ}^{UL} = \frac{N_{SC} b_{\text{IQ}}}{T_{\text{sym}}}$, onde b_{IQ} é o número de bits para representar cada amostra IQ e T_{sym} é o período de um único símbolo OFDM. Além disso, no início de cada intervalo de coerência, a O-RU também precisa enviar o CSI para a O-DU, o que requer um tráfego *uplink* adicional de $R_{\text{CSI}} = \frac{K \times b_{\text{CSI}} \times \frac{N_{\text{SC}}}{C_{\text{BW}}}}{T_{\text{sym}}}$, onde b_{CSI} é o número de bits usados para representar cada amostra CSI. Neste caso, considera-se que a O-RU utiliza um período de símbolo OFDM para enviar o CSI para a O-DU para evitar latência extra na cadeia *uplink* . Assim, o tráfego no *fronthaul uplink* por O-RU teria um pico no início de cada intervalo de coerência dado por $R_P = R_{IQ}^{UL} + R_{CSI}$, e no tempo restante, o tráfego seria R_{IQ}^{UL}. Considerando M O-RUs, a taxa de dados no *fronthaul* agregada é $R_a = M R_P$.

Considerando uma resolução de bits suficiente de b_{IQ} = 20 bits, b_{CSI} = 64 bits e K = 16 UEs, a Tabela 4.1 mostra a taxa de dados IQ, taxa de dados de CSI e taxa de pico no *fronthaul* por O-RU em diferentes BWs. Foi assumido um espaçamento de subportadoras de Δf = 30kHz (T_{slot} = 0,5ms), e o número de subportadoras foi derivado da configuração máxima da largura de banda de transmissão, que é dada em número máximo de PRBs (N_{PRB}) por largura de banda [41].

Tabela 4.1: Tráfego no *Fronthaul* por O-RU para Várias Configurações de Bandwidth com b_{IQ} = 20 bits, b_{CSI} = 64 bits e K = 16 UEs.

BW (MHz)	N_{PRB}	$N_{SC} = 12N_{PRB}$	Dados IQ (Mbps)	Dados CSI (Mbps)	Taxa de Dados no Fronthaul por O-RU (Mbps)
5	11	132	73.92	315.39	389.31
10	24	288	161.28	688.13	849.41
20	51	612	342.7	1462.3	1805
40	106	1272	712.3	3039.2	3751.6
80	217	2604	1458.2	6221.8	7680.1
100	273	3276	1834.6	7827.5	9662

A Tabela 4.1 representa o pior cenário, onde todos os PRBs estão em uso no rádio. Contudo, a taxa de dados instantânea depende do número de subportadoras ativas sendo transmitidas aos UEs. Durante períodos de baixo tráfego de rádio, a taxa de dados no *fronthaul* correspondente é reduzida proporcionalmente.

O número de bits das amostras CSI e o número de usuários na rede contribuem para as altas taxas de dados das amostras CSI no *fronthaul* em comparação com as taxas de dados das amostras IQ. Isso ocorre porque as taxas de dados das amostras CSI são proporcionais tanto ao número de bits de CSI quanto ao número de usuários na rede.

4.2 Compressão em Sistemas NR

Para avaliar o cenário descrito na Seção 4.1, experimentos numéricos foram conduzidos usando sinais 5G/NR compatíveis com o padrão, incorporando *Multiple User - Multiple Input Multiple Output* (MU-MIMO). O objetivo é avaliar o desempenho da rede D-MIMO em sistemas 5G/NR comparando o D-MIMO com a presença de compressão e sem compressão do CSI.

Neste cenário foi considerada uma área medindo $500 \times 500\ m^2$. A multiplexação espacial é empregada para servir simultaneamente K UEs em 51 PRBs. A configuração do *slot* TDD é ilustrada na Figura 4.1, abrangendo 9 símbolos OFDM para *downlink*, 1 período de guarda e 4 símbolos OFDM de *uplink*. Para cada UE, os SRSs são configurados para transmitir na seção

uplink de todos os *slots*, cobrindo toda a largura de banda a cada 4 subportadoras dentro de um único símbolo OFDM.

O modelo COST Hata é usado para o coeficiente de desvanecimento em grande escala:

$$10\log_{10}(\beta_{km}) = -136 - 35\log_{10}(d_{km}) + X_{km}, \tag{4.7}$$

onde d_{km} é a distância entre a antena m e o UE k em km, $X_{km} \sim \mathcal{N}(0, \sigma^2_{\text{shadow}})$, com $\sigma^2_{\text{shadow}} = 8\,\text{dB}$.

A variação do ruído no receptor é considerada $\sigma_w^2 = 290 \times B_\text{c} \times \text{BW} \times \text{NF}$, onde $290\,\text{K}$ é a temperatura ambiente, $B_\text{c} = 1.3807 \times 10^{-23}\,\text{J K}^{-1}$ é a constante de Boltzmann, $\text{BW} = 20\,\text{MHz}$ é a largura de banda disponível e $\text{NF} = 9\,\text{dB}$ é o valor da figura de ruído, respectivamente.

A simulação implementa *Modulation and Coding Scheme* (MCS) adaptativo, onde a relação *Signal-to-Interference-plus-Noise Ratio* (SINR) de limite para uma determinada *Block Error Rate* (BLER) é constantemente atualizada com o algoritmo *Outer Loop Link Adaptation* (OLLA) de [42], e as tabelas iniciais que mapeiam SINR para BLER consideram sinais NR. O MCS pode assumir qualquer valor da tabela 2 de índice MCS para PDSCH [[43], Tabela 5.1.3.1-2], onde a modulação pode variar de *Quadrature Phase Shift Keying* (QPSK) a 256QAM e a taxa de codificação pode ser consultada na tabela.

O MCS muda com a condição do canal, visando uma taxa de erro de bloco alvo de $BLER_T = 0.1$. A taxa de dados através do rádio é calculada como o número de bits do bloco de transporte decodificados com sucesso nos UEs dividido pelo tempo dos quadros NR.

4.2.1 Resultados Numéricos

Neste trabalho foram avaliados os métodos de compressão de CSI detalhados na Seção 3.2 no sistema NR. O objetivo consiste em avaliar o desempenho da rede D-MIMO sem compressão e com compressão no cenário específico. O modelo de sistema adotado nas análises é o mesmo detalhado na Seção 4.1 e a compressão é implementada conforme descrito na Seção 2.5.2 na qual é mostrado que a compressão ocorre sobre o canal estimado visando a diminuição do tráfego no *fronthaul*.

Um resumo dos parâmetros de simulação é apresentado na Tabela 4.2. Eles foram utilizados variando o número de usuários (K) atendidos simultaneamente nos mesmos recursos tempo-frequência. Para um determinado número de usuários, as simulações são executadas 10 vezes, onde a posição dos UEs e, consequentemente, os ganhos do canal são diferentes. Em

Tabela 4.2: Resumo dos parâmetros da simulação e taxa de dados no *fronthaul*.

Parâmetros	Especificação
Número de *slots* NR	200
Subcarrier spacing	30 kHz
Período do *slot*	$0.5ms$
Número de PRBs	51
Número de símbolos OFDM por *slot*	14
Número de símbolos OFDM no DL por *slot*	9
Número de símbolos OFDM no UL por *slot*	4
Opções para MCS durante adaptação do *link*	[[43], Tabela 5.1.3.1-2]
BLER alvo	0.1
Área de simulação	$0.5 \times 0.5\ km^2$
Número de UEs	$2, \cdots, 16$
Número de antenas gNB	32
Número de antenas nos UEs	1
Taxa de dados no *fronthaul* por O-RU	1.805 Gbps
Taxa de dados no *fronthaul* agregada	$\approx 57.76\ Gbps$
b_{CSI}	$24, 16$ bits

seguida, a taxa de dados de rádio para cada usuário é calculada e as seguintes métricas foram capturadas para cada valor de K: a média da taxa agregada de dados no rádio (média do *sum rate*) e as taxas de dados de rádio individuais. A Figura 4.3 mostra a taxa média de soma no rádio para os cenários com D-MIMO sem compressão e com compressão de CSI. O desempenho nos cenários cresce com o número de usuários devido à multiplexação espacial.

Considerando a métrica *sum rate*, a Figura 4.3 mostra que o desempenho do D-MIMO com compressão $b_{\text{CSI}} = 24$ bits permanece semelhante ao caso sem compressão para todo o intervalo de valores de K analisado. Para $b_{\text{CSI}} = 16$ bits, o desempenho fica próximo ao caso sem compressão para valores menores de K, contudo, para valores maiores de K, a diferença torna-se maior. Por exemplo, para $K = 9$, a diferença entre o caso sem compressão e o método BS é em torno de 22 Mbps. Para o método BFP, a diferença é em torno de 32 Mbps.

A Figura 4.4 mostra estatísticas das taxas de dados individuais para diferentes números de

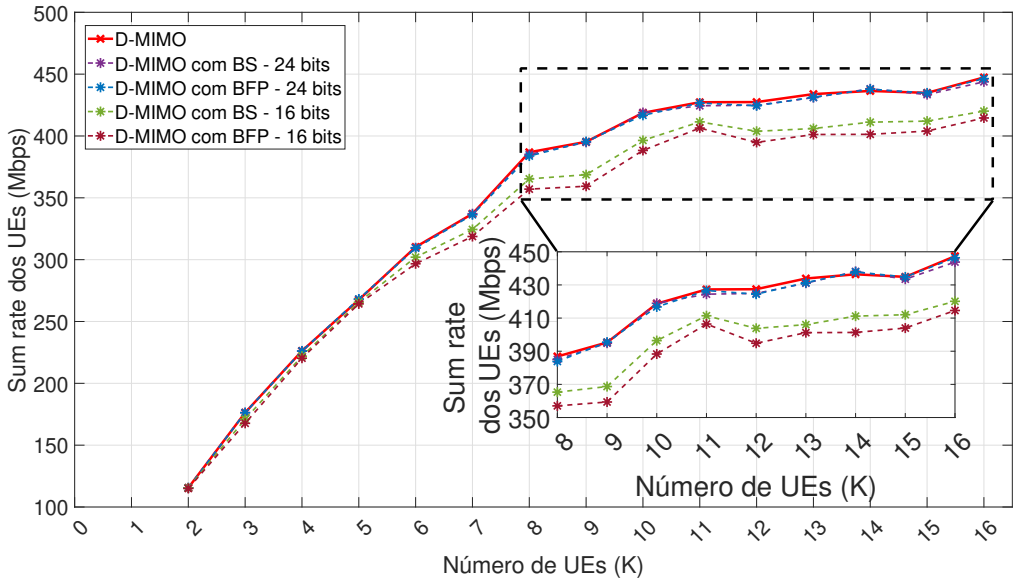


Figura 4.3: Estatísticas de taxa de dados individuais para diferentes números de UEs, considerando o cenário com MIMO distribuído sem compressão e com compressão.

UEs e diferentes valores de b_{CSI} na forma de *boxplots*, onde cada caixa representa o *Interquantile Range* (IQR) com as bordas inferior e superior representando o 25º e o 75º percentil, respectivamente. Os *whiskers* acima e abaixo das caixas são definidos para um máximo de $1.5IQR$ e as amostras fora dessa faixa são representadas com círculos. Cada subfigura compara o desempenho do D-MIMO sem compressão e com compressão para b_{CSI} = 24 e 16 bits.

Em primeiro lugar, todas as amostras de taxa de dados são limitadas a aproximadamente $63\ Mbps$ devido ao número limitado de PRBs e símbolos OFDM alocados para *downlink* no *slot* TDD. Por outro lado, a taxa de dados mais baixa pode ser próximo de zero quando a SINR é muito baixa, levando a erros de bloco de transporte mesmo com MCS de ordem mais baixa.

Em segundo lugar, percebe-se que à medida que o número de UEs cresce na rede, o desempenho geral da rede diminui continuamente, mesmo sem compressão. O desempenho do D-MIMO com os métodos de compressão BS e BFP, para b_{CSI} = 24 bits, mantém-se próximo do caso sem compressão para todos os valores de K analisados. A diferença é maior para b_{CSI} = 16 bits, principalmente ao comparar com valores maiores de K. Por exemplo, com $K = 10$ UEs e $b_{\text{CSI}} = 16$ bits, a compressão BS causa uma perda de aproximadamente 10 Mbps no *whisker* abaixo da caixa quando comparada ao caso sem compressão, enquanto que a compressão BFP causa uma perda de aproximadamente 16 Mbps no mesmo *whisker* analisado. De modo geral, o método BS apresenta desempenho superior ao BFP com pequenas variações entre os métodos.

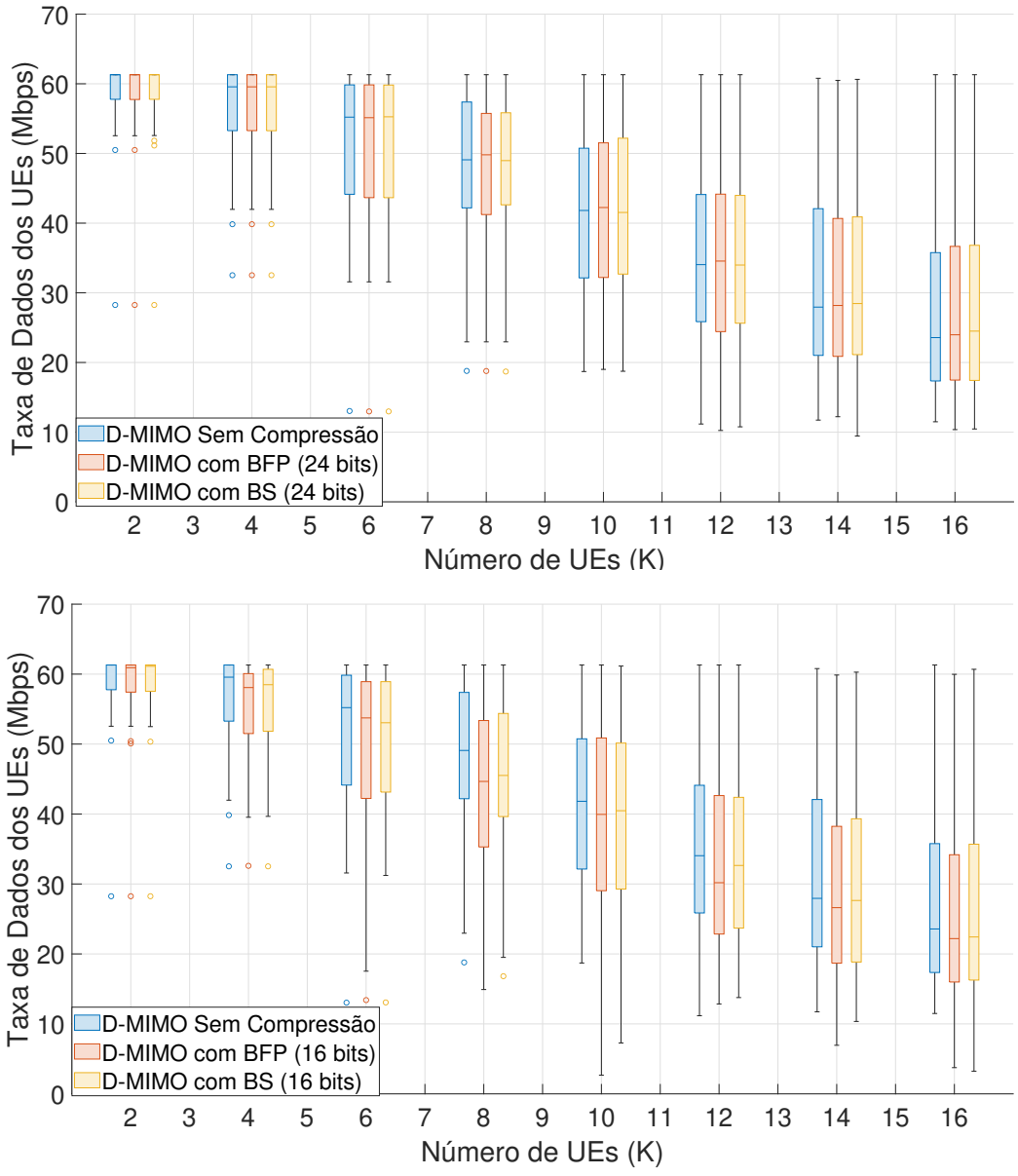


Figura 4.4: Estatísticas de taxa de dados individuais para diferentes números de UEs, considerando o cenário com MIMO distribuído sem compressão e com compressão com b_{CSI} = 24 e 16 bits.

Capítulo 5

Conclusão e Trabalhos Futuros

5.1 Conclusão

Neste trabalho, foram detalhados aspectos sobre a arquitetura da tecnologia MIMO, assim como fundamentos sobre a implementação desta tecnologia em sistemas 4G/LTE-5G/NR. Em especial, foram abordados aspectos de implementação da arquitetura D-MIMO em sistemas 4G/LTE-5G/NR e foram feitas análises de desempenho da tecnologia distribuída sem compressão em comparação à abordagem com compressão.

Além disso, fornecemos *insights* sobre como os métodos de compressão de CSI podem habilitar sistemas D-MIMO com alocação de potência centralizada e pré-codificação *Zero-Forcing*. Os resultados obtidos são comparados com o caso sem compressão de CSI. Foi demonstrado que as abordagens de compressão propostas apresentam um baixo impacto na eficiência espectral da rede distribuída, especialmente se o número de bits utilizados na compressão for suficientemente elevado. Além disso, os métodos de compressão reduzem a taxa de pico no *fronthaul*, o que tem sido um grande desafio para a implementação do D-MIMO com processamento centralizado. Assim, os algoritmos de compressão podem permitir implementações práticas de sistemas D-MIMO centralizados, que possuem maior eficiência espectral do que as abordagens distribuídas.

Além disso, foi avaliado o desempenho do D-MIMO com sinais 5G/NR sem compressão e com a presença de compressão. Resultados numéricos foram apresentados, indicando que a eficiência espectral pode ser tão boa quanto no caso sem compressão dependendo da configuração de compressão. Além disso, foram fornecidos *insights* para implantação prática, como a implementação distribuída com camada PHY/NR e tráfego *fronthaul*. É esperado que este

esforço possa produzir *insights* úteis para a implementação do D-MIMO em sistemas 5G/B5G práticos.

5.2 Trabalhos Futuros

Algumas sugestões para investigações futuras são fornecidas a seguir:

- Avaliar os esquemas propostos utilizando simulações em nível de sistema onde o desempenho pode ser avaliado em termos de *throughput* do usuário em diferentes cenários de tráfego.
- Avaliar outros métodos de compressão como os mencionados no Capítulo 1 e assim verificar a relação de desempenho, complexidade computacional e taxa de transmissão de dados.
- Avaliar os sistemas NR com várias células com D-MIMO e comparar os resultados obtidos com o cenário sem D-MIMO.
- Avaliar o desempenho do sistema D-MIMO com mais antenas por AP e por UE comparando com o desempenho com uso de uma única antena por AP e por UE.

Referências Bibliográficas

[1] G. Interdonato, P. Frenger, and E. G. Larsson, "Scalability Aspects of Cell-Free Massive MIMO," in *ICC 2019 - 2019 IEEE International Conference on Communications (ICC)*, 2019, pp. 1–6.

[2] G. Interdonato, E. Björnson, H. Quoc Ngo, P. Frenger, and E. G. Larsson, "Ubiquitous Cell-Free Massive MIMO Communications," *EURASIP Journal on Wireless Communications and Networking*, vol. 2019, no. 1, Aug. 2019. [Online]. Available: http://dx.doi.org/10.1186/s13638-019-1507-0

[3] A. Nakamura, L. Ramalho, and A. Klautau, "Fronthaul Requirements Analysis for Cell-Free MIMO," *School on Systems and Networks (SSN)*, December 2020.

[4] E. Dahlman, S. Parkvall, and J. Skold, *5G NR: The Next Generation Wireless Access Technology*, 1st ed. USA: Academic Press, Inc., 2018.

[5] 3GPP, "NR; Physical Channels and Modulation," *3GPP TS 38.211 version 16.2.0 Release 16*, 2020.

[6] O. Haliloglu, H. Yu, C. Madapatha, H. Guo, F. E. Kadan, A. Wolfgang, R. Puerta, P. Frenger, and T. Svensson, "Distributed MIMO Systems for 6G," in *2023 Joint European Conference on Networks and Communications 6G Summit (EuCNC/6G Summit)*, 2023, pp. 156–161.

[7] E. Nayebi, A. Ashikhmin, T. L. Marzetta, H. Yang, and B. D. Rao, "Precoding and Power Optimization in Cell-Free Massive MIMO Systems," *IEEE Trans. Wirel. Commun.*, vol. 16, no. 7, pp. 4445–4459, July 2017.

[8] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-Free Massive MIMO Versus Small Cells," *IEEE Trans. Wirel. Commun.*, vol. 16, no. 3, pp. 1834–1850, March 2017.

[9] G. Interdonato, M. Karlsson, E. Björnson, and E. G. Larsson, "Local Partial Zero-Forcing Precoding for Cell-Free Massive MIMO," *IEEE Trans. Wirel. Commun.*, vol. 19, no. 7, pp. 4758–4774, 2020.

[10] L. D. Nguyen, T. Q. Duong, H. Q. Ngo, and K. Tourki, "Energy Efficiency in Cell-Free Massive MIMO with Zero-Forcing Precoding Design," *IEEE Commun. Letters*, vol. 21, no. 8, pp. 1871–1874, 2017.

[11] M. N. Boroujerdi, A. Abbasfar, and M. Ghanbari, "Cell Free Massive MIMO with Limited Capacity Fronthaul," *Wireless Personal Communications*, vol. 104, no. 2, pp. 633–648, 2019.

[12] L. Sun, J. Hou, and T. Shu, "Bandwidth-Efficient Precoding in Cell-Free Massive MIMO Networks with Rician Fading Channels," in *2021 18th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*, 2021, pp. 1–9.

[13] E. Björnson and L. Sanguinetti, "Scalable Cell-Free Massive MIMO Systems," *IEEE Trans Commun*, vol. 68, no. 7, pp. 4247–4261, 2020.

[14] O-RAN Alliance, "Control, User and Synchronization Plane Specification 9.0," *O-RAN Fronthaul Working Group, ORAN-WG4.CUS.0-v09.00*, 2022.

[15] Y. Liao, H. Yao, Y. Hua, and C. Li, "CSI Feedback Based on Deep Learning for Massive MIMO Systems," *IEEE Access*, vol. 7, pp. 86 810–86 820, 2019.

[16] D. Maryopi and A. Burr, "Few-Bit CSI Acquisition for Centralized Cell-Free Massive MIMO with Spatial Correlation," in *2019 IEEE Wireless Communications and Networking Conference (WCNC)*, 2019, pp. 1–6.

[17] M. Bashar, K. Cumanan, A. G. Burr, H. Q. Ngo, M. Debbah, and P. Xiao, "Max–Min Rate of Cell-Free Massive MIMO Uplink With Optimal Uniform Quantization," *IEEE Trans Commun*, vol. 67, no. 10, pp. 6796–6815, 2019.

[18] S. Chen, J. Zhang, E. Björnson, J. Zhang, and B. Ai, "Structured Massive Access for Scalable Cell-Free Massive MIMO Systems," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 4, pp. 1086–1100, 2021.

[19] F. Li, Q. Sun, X. Ji, and X. Chen, "Scalable Cell-Free Massive MIMO with Multiple CPUs," *Mathematics*, vol. 10, no. 11, 2022. [Online]. Available: https://www.mdpi.com/2227-7390/10/11/1900

[20] G. Femenias, F. Riera-Palou, and E. Björnson, "Another Twist to the Scalability in Cell-Free Massive MIMO Networks," *IEEE Transactions on Communications*, vol. 71, no. 11, pp. 6793–6804, 2023.

[21] TechTarget. (2021) MIMO (Multiple Input, Multiple Output). [Online]. Available: https://www.techtarget.com/searchmobilecomputing/definition/MIMO

[22] S. Biswas and P. Vijayakumar, "AP Selection in Cell-Free Massive MIMO System Using Machine Learning Algorithm," in *2021 Sixth International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, 2021, pp. 158–161.

[23] Y. Xiaohu, D. Wang, and J. Wang, *Distributed MIMO and Cell-Free Mobile Communication*, 01 2021.

[24] F. Riera-Palou and G. Femenias, "Trade-offs in Cell-free Massive MIMO Networks: Precoding, Power Allocation and Scheduling," in *2019 14th International Conference on Advanced Technologies, Systems and Services in Telecommunications (TELSIKS)*, 2019, pp. 158–165.

[25] H. A. Ammar, R. Adve, S. Shahbazpanahi, G. Boudreau, and K. V. Srinivas, "User-Centric Cell-Free Massive MIMO Networks: A Survey of Opportunities, Challenges and Solutions," *IEEE Communications Surveys Tutorials*, vol. 24, no. 1, pp. 611–652, 2022.

[26] E. Björnson and L. Sanguinetti, "Scalable Cell-Free Massive MIMO Systems," *IEEE Transactions on Communications*, vol. 68, no. 7, pp. 4247–4261, 2020.

[27] B. Hassan, S. Baig, and S. Aslam, "On Scalability of FDD-Based Cell-Free Massive MIMO Framework," *Sensors*, vol. 23, no. 15, 2023. [Online]. Available: https://www.mdpi.com/1424-8220/23/15/6991

[28] T. Kim, H. Kim, S. Choi, and D. Hong, "How Will Cell-Free Systems Be Deployed?" *IEEE Communications Magazine*, vol. 60, no. 4, pp. 46–51, 2022.

[29] D. Subitha and R. Vani, "Analysis of Linear Precoding Techniques for Massive MIMO-OFDM Systems under various scenarios," *IOP Conference Series: Materials Science and Engineering*, vol. 1084, no. 1, p. 012053, mar 2021. [Online]. Available: https://dx.doi.org/10.1088/1757-899X/1084/1/012053

[30] A. Wiesel, Y. C. Eldar, and S. Shamai, "Zero-Forcing Precoding and Generalized Inverses," *IEEE Transactions on Signal Processing*, vol. 56, no. 9, pp. 4409–4418, 2008.

[31] R. Rajashekar, C. Xu, N. Ishikawa, L.-L. Yang, and L. Hanzo, "Multicarrier Division Duplex Aided Millimeter Wave Communications," *IEEE Access*, vol. PP, pp. 1–1, 07 2019.

[32] Y. Huang, B. Jalaian, S. Russell, and H. Samani, "Reaping the Benefits of Dynamic TDD in Massive MIMO," *IEEE Systems Journal*, vol. PP, pp. 1–8, 06 2018.

[33] L. M. P. Larsen, A. Checko, and H. L. Christiansen, "A Survey of the Functional Splits Proposed for 5G Mobile Crosshaul Networks," *IEEE Communications Surveys Tutorials*, vol. 21, no. 1, pp. 146–172, 2019.

[34] 3GPP, "NR; Physical Layer; General Description," *3GPP TS 38.201 version 16.0.0 Release 16*, 2020.

[35] ——, "5G; NR; Multiplexing and Channel Coding ," *3GPP TS 38.212 version 15.2.0 Release 15*, 2018.

[36] S. Nagul, "A Review on 5G Modulation Schemes and their Comparisons for Future Wireless Communications," in *2018 Conference on Signal Processing And Communication Engineering Systems (SPACES)*, 2018, pp. 72–76.

[37] C. Yang and A. Soloviev, "Pinch the Correlation Function: A Method to Improve Delay Estimation in Multipath," 03 2017, pp. 347–364.

[38] 3GPP, "NR; Services Provided by the Physical Layer," *3GPP TS 38.202 version 16.1.0 Release 16*, 2020.

[39] A. Garcia-Saavedra and X. Costa-Pérez, "O-RAN: Disrupting the Virtualized RAN Ecosystem," *IEEE Communications Standards Magazine*, vol. 5, no. 4, pp. 96–103, 2021.

[40] M. Mohsin, J. M. Batalla, E. Pallis, G. Mastorakis, E. K. Markakis, and C. X. Mavromoustakis, "On Analyzing Beamforming Implementation in O-RAN 5G," *Electronics*, vol. 10, no. 17, 2021. [Online]. Available: https://www.mdpi.com/2079-9292/10/17/2162

[41] 3GPP, "NR; User Equipment (UE) Radio Transmission and Reception," *3GPP TS 38.101 version 17.5.0*, 2022.

[42] K. I. Pedersen, G. Monghal, I. Z. Kovacs, T. E. Kolding, A. Pokhariyal, F. Frederiksen, and P. Mogensen, "Frequency Domain Scheduling for OFDMA with Limited and Noisy Channel Feedback," in *2007 IEEE 66th Vehicular Technology Conference*, 2007, pp. 1792–1796.

[43] 3GPP, "NR; Physical Layer Procedures for Data," *3GPP TS 38.214 version 16.2.0*, 2020.