

UNIVERSIDADE FEDERAL FLUMINENSE

DIEGO CANIZIO LOPES

**Escalonamento de Recursos em Redes 5G: Um
Mecanismo Baseado em Aprendizado por Reforço**

NITERÓI

2026

DIEGO CANIZIO LOPES

Escalonamento de Recursos em Redes 5G: Um Mecanismo Baseado em Aprendizado por Reforço

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Computação da Universidade Federal Fluminense como requisito para a obtenção do Grau de Doutor em Computação. Área de concentração: Ciência da Computação

Orientador:

Igor Monteiro Moraes

Coorientador:

André Luiz Nasserla Pires

Coorientador:

Ian Vilar Bastos

NITERÓI

2026

Ficha catalográfica automática - SDC/BEE
Gerada com informações fornecidas pelo autor

L864e Lopes, Diego Canizio
Escalonamento de recursos em redes 5G : Um mecanismo baseado
em aprendizado por reforço / Diego Canizio Lopes. - 2026.
178 f.: il.

Orientador: Igor Monteiro Moraes.
Coorientador: André Luiz Nasserela Pires; Ian Vilar Bastos.
Tese (doutorado)-Universidade Federal Fluminense, Instituto
de Computação, Niterói, 2026.

1. Escalonamento de recursos. 2. Aprendizado por reforço.
3. Justiça. 4. Redes móveis. 5. Produção intelectual. I.
Moraes, Igor Monteiro, orientador. II. Pires, André Luiz
Nasserela, coorientador. III. Bastos, Ian Vilar, coorientador.
IV. Universidade Federal Fluminense. Instituto de
Computação. V. Título.

CDD - XXX

DIEGO CANIZIO LOPES

Escalonamento de Recursos em Redes 5G:
Um Mecanismo Baseado em Aprendizado por Reforço

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Computação da Universidade Federal Fluminense como requisito para a obtenção do Grau de Doutor em Computação. Área de concentração: Ciência da Computação

BANCA EXAMINADORA

Prof. Igor Monteiro Moraes, D.Sc. - Orientador, UFF

Prof. Célio Vinicius Neves de Albuquerque, Ph.D., UFF

Prof. Flávio Luiz Seixas, D.Sc., UFF

Prof. Rodrigo de Souza Couto, D.Sc., UFRJ

Prof. Weverton Luis da Costa Cordeiro, D.Sc., UFRGS

Prof. André Luiz Nasseralla Pires, D.Sc., UFAC

Prof. Ian Vilar Bastos, D.Sc., UERJ

Niterói

2026

À minha mãe, que fez da minha educação uma prioridade e dos meus sonhos uma extensão dos seus próprios sonhos. Esta conquista também é sua.

Agradecimentos

A Deus, minha eterna gratidão. Por me sustentar nos momentos mais difíceis, por me dar força quando o cansaço parecia maior que a vontade de continuar, por me conceder coragem diante das incertezas e por iluminar meu caminho até aqui.

À minha família, por todo apoio, paciência e compreensão ao longo desta jornada. Foram muitos momentos de ausência, preocupação e renúncia, e ainda assim encontrei em vocês acolhimento, incentivo e amor. Cada conquista alcançada neste caminho também pertence a vocês, que estiveram ao meu lado mesmo quando a dedicação a este trabalho exigiu silêncio, distância e tempo.

À minha filha, Rayanne, por ser uma das razões que me motivaram a seguir adiante nesta trajetória. Sua presença em minha vida reforçou meu senso de responsabilidade e propósito, especialmente nos períodos em que a realização deste trabalho exigiu maior dedicação, concentração e renúncia.

À minha companheira, Sara, por estar ao meu lado durante esta caminhada, por compreender minhas ausências, por me incentivar nos momentos de dúvida e por tornar esta jornada mais humana e possível. Sua presença, apoio e confiança foram fundamentais para que eu chegasse até aqui.

Ao meu orientador, professor Igor, meu profundo agradecimento pela orientação, pela confiança, pela paciência e pelas contribuições ao longo do desenvolvimento desta tese. Sua condução foi essencial para que este trabalho amadurecesse e chegasse à forma apresentada.

Aos meus coorientadores, professores André e Ian, agradeço pela disponibilidade, pelas discussões, pelas sugestões e pelo apoio técnico e acadêmico em diferentes etapas desta pesquisa. As contribuições de vocês foram decisivas para a evolução deste trabalho.

À professora Francesca, minha sincera gratidão pelo apoio, pelo encorajamento e pela generosidade com que contribuiu para o meu trabalho. Sua ajuda e suas palavras de incentivo tiveram grande importância em um período muito significativo desta trajetória.

Aos professores Célio, Flávio, Rodrigo e Weverton que gentilmente aceitaram fazer parte da minha banca e contribuir com este trabalho através de suas críticas e revisões.

À família que se formou no terceiro andar da Maison du Brésil, em Paris, meu agradecimento especial. A convivência, as conversas, o companheirismo e o apoio compartilhado tornaram aquele período mais leve, acolhedor e inesquecível. Em meio aos desafios de estar longe de casa, encontrei ali amizade, pertencimento e afeto.

Aos amigos, colegas, professores e a todos aqueles que, direta ou indiretamente, fizeram parte desta jornada, deixo minha sincera gratidão. Cada palavra de apoio, cada gesto de confiança, cada conversa e cada incentivo contribuíram para que este momento se tornasse realidade.

Por fim, agradeço a todos que compreenderam minhas ausências, respeitaram meus limites e torceram por mim ao longo deste percurso. Esta tese é resultado de muito trabalho, mas também de muitas pessoas que caminharam comigo, me apoiaram e me ajudaram a continuar.

Resumo

O escalonamento de recursos de rádio na quinta geração de redes móveis (5G), em particular no padrão *New Radio* (NR), deve conciliar eficiência e justiça no atendimento a equipamentos de usuário (UEs, do inglês *User Equipment*) submetidos a diferentes condições de canal e demandas de tráfego. Escalonadores como *Round Robin* (RR), *Proportional Fair* (PF) e *Maximum Rate* (MR) utilizam critérios previamente definidos e ocupam regiões distintas desse compromisso. Esta tese propõe o *In-Action Fair Max Rate* (IA-FMR), um mecanismo de escalonamento baseado em aprendizado por reforço no qual o compromisso entre eficiência e justiça integra a própria ação da política. A cada decisão, o agente produz conjuntamente os valores utilizados na distribuição dos recursos entre os UEs e o coeficiente α_t , empregado durante o treinamento para ponderar os termos de justiça e eficiência da recompensa. Essa formulação permite que o compromisso seja representado como parte da decisão da política. Na realização temporal analisada para a política treinada em 40 MHz, α_t permaneceu próximo do limite superior do intervalo de ponderação. As políticas são treinadas com o algoritmo *Proximal Policy Optimization* (PPO) e integradas ao ns-3/5G-LENA por meio do ns3-ai.

A avaliação experimental compreende três eixos: Avaliação por Largura de Banda, Avaliação de Generalização Estrutural por Número de UEs e Avaliação de Disponibilidade Temporal em Tráfego Sintético de Vídeo. Na primeira avaliação, são consideradas doze larguras de banda da *Frequency Range 1* (FR1), entre 10 e 100 MHz, com 100 sementes sistêmicas por configuração. Em 40 MHz, por exemplo, o IA-FMR alcança vazão agregada média de 97,53 Mbps, acima de RR e PF, com 50,93 Mbps e 34,36 Mbps, respectivamente, e abaixo do MR, com 120,40 Mbps. Na mesma configuração, o Índice de Jain médio da alocação de RBGs é de 0,9943 no IA-FMR, frente a 0,9988 no RR, 0,5926 no PF e 0,3711 no MR. O IA-FMR também apresenta maior média do percentil 5 da vazão por fluxo entre 15 e 100 MHz, embora existam sobreposições dos intervalos de confiança em algumas configurações.

Na avaliação com 3, 5, 7 e 9 UEs em 40 MHz, a mesma arquitetura treinada para nove posições é aplicada a conjuntos menores por meio de preenchimento das posições não

utilizadas. Os Índices de Jain médios variam entre 0,9727 e 0,9943, e não são observadas ocorrências de demanda pendente sem alocação nas configurações avaliadas. Ao mesmo tempo, o aumento da cardinalidade é acompanhado por redução da vazão por UE e do percentil 5 da vazão, além de aumento de perdas e atraso, mostrando que a generalização estrutural não implica invariância de desempenho. Na avaliação com tráfego sintético de vídeo e seis UEs, 66,88% das unidades geradas com o IA-FMR são recebidas dentro do prazo de 60 ms, em comparação com 53,10% no RR, 47,12% no PF e 46,41% no MR. Essas métricas temporais são calculadas após as simulações e não integram o treinamento do agente. Em conjunto, os resultados mostram que o IA-FMR mantém elevada justiça na distribuição dos recursos, atendimento amplo dos UEs e um ponto de operação distinto dos escalonadores de referência, sem caracterizar superioridade universal entre todos os critérios avaliados.

Palavras-chave: 5G NR, escalonamento de recursos, aprendizado por reforço, PPO, justiça, disponibilidade temporal.

Abstract

Radio resource scheduling in fifth-generation mobile networks (5G), particularly in the *New Radio* (NR) standard, must reconcile efficiency and fairness when serving user equipment (UEs) subject to different channel conditions and traffic demands. Schedulers such as *Round Robin* (RR), *Proportional Fair* (PF), and *Maximum Rate* (MR) rely on pre-defined criteria and operate at different regions of this trade-off. This thesis proposes the *In-Action Fair Max Rate* (IA-FMR), a reinforcement-learning-based scheduling mechanism in which the efficiency–fairness trade-off is incorporated directly into the policy action. At each decision, the agent jointly produces the values used to distribute resources among the UEs and the coefficient α_t , which is used during training to weight the fairness and efficiency terms of the reward. This formulation allows the trade-off to be represented as part of the policy decision. In the temporal realization analyzed for the policy trained at 40 MHz, α_t remained close to the upper bound of the weighting interval. The policies are trained using *Proximal Policy Optimization* (PPO) and integrated with ns-3/5G-LENA through ns3-ai.

The experimental evaluation comprises three axes: Bandwidth Evaluation, Structural Generalization Evaluation by Number of UEs, and Temporal Availability Evaluation with Synthetic Video Traffic. The first evaluation considers twelve *Frequency Range 1* (FR1) bandwidths between 10 and 100 MHz, with 100 system-level seeds per configuration. At 40 MHz, for example, IA-FMR achieves a mean aggregate throughput of 97.53 Mbps, above RR and PF, with 50.93 Mbps and 34.36 Mbps, respectively, and below MR, with 120.40 Mbps. In the same configuration, the mean Jain fairness index of the RBG allocation is 0.9943 for IA-FMR, compared with 0.9988 for RR, 0.5926 for PF, and 0.3711 for MR. IA-FMR also presents the largest mean fifth percentile of per-flow throughput from 15 to 100 MHz, although confidence intervals overlap in some configurations.

In the evaluation with 3, 5, 7, and 9 UEs at 40 MHz, the same architecture trained for nine positions is applied to smaller sets by zero-padding unused positions. The mean Jain fairness indices range from 0.9727 to 0.9943, and no occurrences of pending demand without resource allocation are observed in the evaluated configurations. At the same

time, increasing the number of UEs is accompanied by reductions in per-UE throughput and in the fifth percentile of flow throughput, as well as increases in packet loss and delay, showing that structural generalization does not imply performance invariance. In the synthetic video traffic evaluation with six UEs, 66.88% of the units generated under IA-FMR are received within the 60 ms deadline, compared with 53.10% for RR, 47.12% for PF, and 46.41% for MR. These temporal metrics are computed after the simulations and are not part of the agent training process. Overall, the results show that IA-FMR maintains high fairness in resource allocation, broad UE service, and an operating point distinct from those of the reference schedulers, without implying universal superiority across all evaluated criteria.

Keywords: 5G NR, resource scheduling, reinforcement learning, PPO, fairness, temporal availability.

Lista de Figuras

1	Exemplos de domínios de aplicação e dispositivos conectados por redes 5G.	12
2	Arquitetura simplificada de uma rede 5G NR com destaque para o escalonamento de recursos na camada MAC da gNodeB.	21
3	Diferença conceitual entre OFDM e OFDMA na utilização dos recursos de tempo e frequência.	22
4	Estrutura temporal do 5G NR e relação entre o índice de numerologia, o espaçamento entre subportadoras, a duração dos <i>slots</i> e a quantidade de <i>slots</i> por subquadro.	23
5	Organização conceitual dos recursos de rádio no plano tempo e frequência, com exemplo de sua distribuição entre usuários.	24
6	Representação do problema de escalonamento de recursos como um Processo de Decisão de Markov.	35
7	Estrutura <i>actor-critic</i> com atualização para PPO.	37
8	Síntese qualitativa das abordagens de escalonamento consideradas na organização do capítulo.	46
9	Organização temática dos trabalhos relacionados ao escalonamento de recursos em redes 5G.	48
10	Mapeamento conceitual das abordagens analisadas segundo a explicitação do compromisso e a flexibilidade operacional.	55
11	Ciclo de decisão do IA-FMR, destacando os logits utilizados na distribuição dos recursos e o componente escalar transformado em α_t	69
12	Composição da recompensa do IA-FMR a partir da justiça na alocação, da eficiência normalizada, da ponderação por α_t e da penalização por déficit de vazão.	74

13	Integração entre o escalonador IA-FMR no ns-3/5G-LENA, a memória compartilhada do ns3-ai e o processo Python responsável pela inferência da política PPO.	80
14	Fluxo da metodologia experimental, desde a seleção das políticas previamente treinadas até a execução dos três cenários, a coleta dos registros e o tratamento estatístico dos resultados.	88
15	Evolução da recompensa média durante o treinamento das políticas IA-FMR. Cada curva corresponde a uma política treinada durante 600.000 passos com a semente 211382627 e posteriormente utilizada na avaliação sistêmica.	105
16	Evolução do coeficiente α_t durante o treinamento da política IA-FMR para 40 MHz. A curva representa o valor médio de α_t e a região sombreada corresponde à faixa entre os percentis 5 e 95 dos valores observados em cada janela de agregação.	106
17	Vazão agregada média em função da largura de banda. As regiões sombreadas representam os intervalos de confiança de 95%.	108
18	Índice de Jain calculado sobre a alocação de RBGs em função da largura de banda. As regiões sombreadas representam os intervalos de confiança de 95%.	110
19	Número médio de UEs servidos por <i>slot</i> em função da largura de banda. As regiões sombreadas representam os intervalos de confiança de 95%. . . .	111
20	Percentual de registros UE- <i>slot</i> com demanda pendente nos quais a alocação aplicada foi nula, em função da largura de banda. As regiões sombreadas representam os intervalos de confiança de 95%.	113
21	Percentil 5 da vazão por fluxo em função da largura de banda. As regiões sombreadas representam os intervalos de confiança de 95%.	115
22	Taxa média de perda de pacotes em função da largura de banda. As regiões sombreadas representam os intervalos de confiança de 95%.	117
23	Atraso médio dos pacotes recebidos em função da largura de banda. As regiões sombreadas representam os intervalos de confiança de 95%.	118
24	Participação temporal dos UEs na alocação de RBGs pelo RR em 40 MHz para a semente 47889.	120

25	Participação temporal dos UEs na alocação de RBGs pelo PF em 40 MHz para a semente 47889.	121
26	Participação temporal dos UEs na alocação de RBGs pelo MR em 40 MHz para a semente 47889.	121
27	Participação temporal dos UEs na alocação de RBGs pelo IA-FMR em 40 MHz para a semente 47889.	122
28	Número efetivo de UEs na distribuição temporal de RBGs em 40 MHz para a semente 47889.	123
29	Maior participação individual na distribuição temporal de RBGs em 40 MHz para a semente 47889.	124
30	Número de UEs servidos por <i>slot</i> em 40 MHz para a semente 47889.	126
31	Evolução temporal do <i>backlog</i> agregado em 40 MHz para a semente 47889, expresso em Mbit.	128
32	Evolução temporal de α_t produzido pelo IA-FMR em 40 MHz para a semente 47889. O eixo vertical utiliza escala ampliada para evidenciar as pequenas variações observadas ao longo da execução.	130
33	Vazão agregada média do IA-FMR em 40 MHz em função do número de UEs. A região sombreada representa o intervalo de confiança de 95%.	133
34	Índice de Jain sobre a alocação de RBGs do IA-FMR em 40 MHz em função do número de UEs. A região sombreada representa o intervalo de confiança de 95%.	134
35	Número médio de UEs servidos por <i>slot</i> pelo IA-FMR em 40 MHz em função do número de UEs. A região sombreada representa o intervalo de confiança de 95%.	135
36	Percentil 5 da vazão por fluxo do IA-FMR em 40 MHz em função do número de UEs. A região sombreada representa o intervalo de confiança de 95%.	136
37	Taxa média de perda do IA-FMR em 40 MHz em função do número de UEs. A região sombreada representa o intervalo de confiança de 95%.	137
38	Atraso médio dos pacotes recebidos pelo IA-FMR em 40 MHz em função do número de UEs. A região sombreada representa o intervalo de confiança de 95%.	138

-
- 39 Composição das unidades sintéticas de vídeo considerando os prazos de 60 e 80 ms. As barras de erro representam os intervalos de confiança de 95% da TUP e da proporção total de unidades recebidas. 141
- 40 Frequência dos eventos de congelamento inferidos para os prazos de 60 e 80 ms, com intervalo de confiança de 95% calculado sobre as 100 sementes. 142
- 41 Duração P95 dos eventos de congelamento inferidos para os prazos de 60 e 80 ms, com intervalo de confiança de 95% calculado sobre as sementes em que houve pelo menos um evento de congelamento. 143
- 42 Percentual de usuários temporalmente satisfeitos para os prazos de 60 e 80 ms, considerando $TUP \geq 99\%$ e intervalo de confiança de 95%. 145

Lista de Tabelas

1	Significado dos principais parâmetros de treinamento do PPO utilizados nesta tese.	40
2	Procedimento de busca e organização temática da literatura	48
3	Resumo comparativo dos principais trabalhos relacionados.	56
4	Formas de incorporação do coeficiente de ponderação consideradas na evolução do FMR.	59
5	Exemplo de conversão dos logits em uma alocação inteira pelo método dos maiores restos.	71
6	Ambiente de software e arquitetura neural utilizados no treinamento do IA-FMR.	75
7	Parâmetros da configuração final de treinamento do IA-FMR.	76
8	Resultados preliminares do FMR original no ambiente Python.	78
9	Resultados preliminares da formulação <i>in-action</i> no ambiente Python. . . .	78
10	Origem das informações utilizadas na reconstrução do estado durante a inferência.	81
11	Resumo das avaliações experimentais.	89
12	Parâmetros comuns das simulações sistêmicas.	90
13	Quantidade de PRBs considerada para as larguras de banda FR1 avaliadas com SCS de 30 kHz.	90
14	Fases de tráfego utilizadas na Avaliação por Largura de Banda.	92
15	Configuração física e espacial da avaliação com tráfego sintético de vídeo. .	94
16	Configuração do tráfego e das execuções da avaliação com tráfego sintético de vídeo.	95

17	Métricas sistêmicas utilizadas nas avaliações.	96
18	Comparação entre a alocação alvo produzida pelo IA-FMR e a alocação aplicada no ns-3/5G-LENA.	114
19	Síntese da avaliação por largura de banda na configuração de 40 MHz. . . .	146
20	Desempenho do IA-FMR em função do número de UEs.	147
21	Distribuição e atendimento do IA-FMR em função do número de UEs. . . .	147
22	Principais resultados da avaliação com tráfego sintético de vídeo para o prazo de 60 ms.	148

Lista de Abreviaturas e Siglas

3GPP *3rd Generation Partnership Project*

5G *Fifth Generation*

5GC *5G Core*

6G *Sixth Generation*

B5G *Beyond 5G*

BWP *Bandwidth Part*

CAPES *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior*

COFECUB *Comitê Francês de Avaliação da Cooperação Universitária com o Brasil*

CQI *Channel Quality Indicator*

CSV *Comma-Separated Values*

D2D *Device-to-Device*

DDPG *Deep Deterministic Policy Gradient*

DDQN *Double Deep Q-Network*

DL *Downlink*

DQN *Deep Q-Network*

DRL *Deep Reinforcement Learning*

EESM *Exponential Effective SINR Mapping*

eMBB *enhanced Mobile Broadband*

FMR *Fair Max Rate*

FR1 *Frequency Range 1*

GAE *Generalized Advantage Estimation*

gNB *gNodeB*

HARQ *Hybrid Automatic Repeat Request*

IA-FMR *In-Action Fair Max Rate*

IC *Intervalo de Confiança*

IR *Incremental Redundancy*

MAC *Medium Access Control*

MCS *Modulation and Coding Scheme*

MDP *Markov Decision Process*

MIMO *Multiple-Input Multiple-Output*

mMTC *massive Machine-Type Communications*

MOS *Mean Opinion Score*

MR *Maximum Rate*

NG-RAN *Next Generation Radio Access Network*

NGBR *Non-Guaranteed Bit Rate*

NR *New Radio*

O-RAN *Open Radio Access Network*

OFDM *Orthogonal Frequency-Division Multiplexing*

OFDMA *Orthogonal Frequency-Division Multiple Access*

OWA *Ordered Weighted Average*

PDCCP *Packet Data Convergence Protocol*

PDSCH *Physical Downlink Shared Channel*

PF *Proportional Fair*

PHY *Physical Layer*

PPO *Proximal Policy Optimization*

PRB *Physical Resource Block*

QoE *Quality of Experience*

QoS *Quality of Service*

RBG *Resource Block Group*

RE *Resource Element*

RL *Reinforcement Learning*

RLC *Radio Link Control*

RNTI *Radio Network Temporary Identifier*

RR *Round Robin*

SCS *Subcarrier Spacing*

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

SLA *Service Level Agreement*

TDD *Time Division Duplexing*

TUP *Taxa de Unidades no Prazo*

UDP *User Datagram Protocol*

UE *User Equipment*

UL *Uplink*

UM *Unacknowledged Mode*

UMi *Urban Micro*

URLLC *Ultra-Reliable Low-Latency Communications*

XR *Extended Reality*

Sumário

1	Introdução	12
1.1	Motivação	14
1.2	Problema de Pesquisa	15
1.3	Objetivos	16
1.3.1	Objetivo Geral	16
1.3.2	Objetivos Específicos	16
1.4	Contribuições	16
1.5	Metodologia Resumida	17
1.6	Organização da Tese	18
2	Fundamentação Teórica	19
2.1	Redes Móveis 5G	19
2.1.1	Arquitetura e Serviços do 5G	20
2.1.2	OFDM e OFDMA no 5G NR	21
2.1.3	Numerologia e Estrutura Temporal	22
2.1.4	Organização dos Recursos de Rádio	24
2.1.5	Modulação, Codificação, CQI e MCS	25
2.2	Escalonamento de Recursos em Redes 5G	26
2.2.1	Escalonadores de Referência	26
2.3	Métricas de Avaliação, QoS e Requisitos Temporais	27
2.3.1	Vazão, Percentis, Perdas e Atraso	28
2.3.2	Justiça e Distribuição dos Recursos	29

2.3.3	Atendimento, Demanda Pendente e Backlog	30
2.3.4	QoS, Requisitos e Critérios de Treinamento	31
2.3.5	Requisitos Temporais no Nível da Aplicação	31
2.4	Aprendizado por Reforço	33
2.4.1	Estado, Ação, Política, Recompensa e Retorno	33
2.4.2	Processos de Decisão de Markov	34
2.4.3	Funções de Valor e Vantagem	35
2.4.4	Função de Recompensa e Problemas Multiobjetivo	36
2.4.5	Métodos Baseados em Gradiente de Política	36
2.4.6	Arquiteturas Actor-Critic	37
2.5	Proximal Policy Optimization	38
2.5.1	Objetivo com Clipping	38
2.5.2	Generalized Advantage Estimation	38
2.5.3	Função de Valor, Entropia e Otimização	39
2.5.4	Significado dos Parâmetros de Treinamento do PPO	39
2.5.5	Ações Contínuas e Transformações da Saída	42
2.6	Aprendizado por Reforço em Escalonamento de Recursos	42
2.7	Considerações Finais do Capítulo	43
3	Trabalhos Relacionados	45
3.1	Procedimento de Seleção e Organização da Literatura	46
3.2	Escalonadores Tradicionais para Redes Móveis	49
3.3	Escalonadores Baseados em Aprendizado por Reforço	50
3.4	Justiça no Escalonamento de Recursos	51
3.5	Representação do Compromisso entre Justiça e Eficiência	53
3.6	Posicionamento da Proposta	54
3.7	Considerações Finais do Capítulo	57

4	O Mecanismo IA-FMR	58
4.1	Evolução do FMR para o IA-FMR	58
4.2	Modelo do Sistema e Definições do Problema	60
4.3	Formulação do IA-FMR como Processo de Decisão	62
4.3.1	Ambiente de Treinamento e Dinâmica Temporal	63
4.3.2	Espaço de Estados	64
4.3.3	Espaço de Ações e Coeficiente de Ponderação	67
4.3.4	Conversão da Ação em Alocação Alvo	69
4.3.5	Função de Recompensa	71
4.4	Treinamento da Política IA-FMR	74
4.4.1	Configuração do Treinamento e Arquitetura da Política	75
4.4.2	Evidências Preliminares da Evolução da Formulação	77
4.5	Integração do IA-FMR com o ns-3/5G-LENA	79
4.5.1	Visão Geral da Arquitetura de Integração	79
4.5.2	Construção e Comunicação da Observação	80
4.5.3	Identidade dos UEs, Posições da Política e Inferência	82
4.5.4	Da Alocação Alvo à Alocação Aplicada na Camada MAC	83
4.6	Considerações Finais	84
5	Metodologia Experimental	86
5.1	Planejamento da Avaliação Experimental	87
5.2	Ambiente e Parâmetros Comuns das Simulações	87
5.2.1	Larguras de Banda e Quantidade de Recursos	89
5.3	Cenários Experimentais	90
5.3.1	Avaliação por Largura de Banda	90
5.3.2	Avaliação de Generalização Estrutural por Número de UEs	93
5.3.3	Avaliação de Disponibilidade Temporal em Tráfego Sintético de Vídeo	94

5.4	Métricas e Tratamento Estatístico	96
5.4.1	Métricas Sistêmicas	96
5.4.2	Comparação entre Alocação Alvo e Alocação Aplicada	98
5.4.3	Processamento das Métricas Temporais em 40 MHz	99
5.4.4	Métricas de Disponibilidade Temporal no Cenário de Vídeo	100
5.4.5	Tratamento Estatístico	102
5.5	Considerações Finais	102
6	Resultados e Discussão	104
6.1	Treinamento do IA-FMR	104
6.2	Avaliação por Largura de Banda	107
6.2.1	Vazão Agregada	108
6.2.2	Justiça na Alocação de RBGs	109
6.2.3	Atendimento dos UEs e Demanda sem Alocação	111
6.2.4	Vazão dos Fluxos Menos Favorecidos	115
6.2.5	Perdas e Atraso	117
6.2.6	Análise Temporal em 40 MHz	119
6.2.6.1	Distribuição Temporal dos RBGs	120
6.2.6.2	Concentração da Alocação	123
6.2.6.3	Atendimento dos UEs	125
6.2.6.4	Evolução do <i>Backlog</i>	127
6.2.6.5	Evolução de α_t	129
6.2.7	Síntese da Avaliação por Largura de Banda	131
6.3	Avaliação de Generalização Estrutural por Número de UEs	132
6.3.1	Vazão Agregada	132
6.3.2	Justiça na Alocação de RBGs	133
6.3.3	Atendimento dos UEs	135

6.3.4	Vazão dos Fluxos Menos Favorecidos	136
6.3.5	Perdas e Atraso	137
6.4	Avaliação de Disponibilidade Temporal em Tráfego Sintético de Vídeo . . .	139
6.4.1	Entrega Temporal das Unidades de Vídeo	140
6.4.2	Frequência e Severidade dos Congelamentos Inferidos	141
6.4.3	Usuários Temporalmente Satisfeitos	144
6.4.4	Síntese da Avaliação de Disponibilidade Temporal	145
6.5	Considerações Finais do Capítulo	146
7	Conclusões e Trabalhos Futuros	150
7.1	Considerações Iniciais	150
7.2	Conclusões Gerais	151
7.3	Principais Contribuições	154
7.4	Internacionalização e Produção Científica	155
7.5	Limitações	156
7.6	Trabalhos Futuros	158
	REFERÊNCIAS	160

1 Introdução

As redes móveis de quinta geração (5G), em particular o padrão *New Radio* (NR), foram concebidas para atender aplicações com requisitos distintos de vazão, atraso, confiabilidade e densidade de conexões. Esses requisitos são organizados em categorias como *enhanced Mobile Broadband* (eMBB), *Ultra-Reliable Low-Latency Communications* (URLLC) e *massive Machine Type Communications* (mMTC). Aplicações multimídia imersivas, como realidade virtual e realidade aumentada, ampliam essa diversidade ao demandar altas taxas de transmissão e continuidade no atendimento ([Andrews et al., 2014](#); [Dahlman; Parkvall; Skold, 2023](#); [3GPP, 2026c](#)).



Figura 1: Exemplos de domínios de aplicação e dispositivos conectados por redes 5G.

A Figura 1 reúne aplicações associadas a cidades, indústrias, dispositivos e sistemas de transporte conectados. Esses domínios combinam diferentes níveis de mobilidade, densidade de equipamentos, volume de dados, confiabilidade e sensibilidade ao atraso. Como os dispositivos compartilham a infraestrutura de acesso, a rede precisa distribuir

os recursos de rádio entre demandas simultâneas.

Na rede de acesso, os recursos nos domínios do tempo e da frequência são compartilhados entre os equipamentos de usuário (UEs, do inglês *User Equipment*). No *downlink*, a estação rádio base, ou *gNodeB* (gNB), executa uma política de escalonamento que determina quais UEs serão atendidos e quantos recursos serão atribuídos a cada um em cada intervalo de decisão. Essa política de escalonamento pode considerar condições de canal, dados nas filas, histórico de atendimento, requisitos de serviço e recursos disponíveis (Dahlman; Parkvall; Skold, 2023; 3GPP, 2026a,c,d).

As decisões do escalonador influenciam a utilização da capacidade da célula e sua distribuição entre os usuários. Uma política de escalonamento pode privilegiar as melhores condições de canal, a regularidade do atendimento ou um compromisso entre esses objetivos. Por isso, a avaliação deve combinar métricas: uma vazão agregada elevada pode coexistir com concentração de recursos, enquanto uma distribuição uniforme pode limitar o aproveitamento das diferenças de eficiência espectral entre os UEs (Sedqi *et al.*, 2013; Guo *et al.*, 2014).

Vazão, atraso, perdas e confiabilidade caracterizam a Qualidade de Serviço (*Quality of Service*, QoS) oferecida pela rede (Ferreira *et al.*, 2017). Em aplicações de vídeo, essas métricas podem ser complementadas por indicadores temporais no nível da aplicação (Gobatto *et al.*, 2024). Uma unidade recebida após o prazo considerado pode ser contabilizada como recebida pela rede sem integrar a parcela entregue dentro desse prazo. Esses indicadores, isoladamente, não constituem uma caracterização completa da Qualidade de Experiência (*Quality of Experience*, QoE).

Entre os algoritmos de escalonamento utilizados como referência encontram-se *Round Robin* (RR), *Proportional Fair* (PF) e *Maximum Rate* (MR). O RR distribui oportunidades de atendimento de forma regular, o MR prioriza usuários com melhor eficiência espectral e o PF combina usuários com melhor eficiência espectral estimada e histórico de vazão. Embora representem critérios distintos, esses algoritmos seguem regras definidas previamente (Mamane *et al.*, 2022).

A variação das condições de canal, da carga de tráfego e da quantidade de recursos disponíveis dificulta a escolha de um único critério para todas as situações. O aprendizado por reforço permite formular o escalonamento como um processo de decisão sequencial, no qual uma política de escalonamento é aprendida a partir da interação com o ambiente (Sutton; Barto, 2018; You *et al.*, 2019; Galdino *et al.*, 2024). Esta tese propõe o *In-Action Fair Max Rate* (IA-FMR), mecanismo de escalonamento em que a ação contém

os valores usados para definir a distribuição dos recursos e um coeficiente de ponderação entre justiça e eficiência, utilizado na recompensa durante o treinamento e incorporado ao estado da decisão seguinte durante a inferência.

1.1 Motivação

A motivação desta tese está no compromisso entre eficiência e justiça no escalonamento. O RR preserva a regularidade das oportunidades de atendimento, mas não utiliza as diferenças instantâneas de eficiência espectral como critério principal. O MR explora essas diferenças para elevar a vazão agregada, porém pode concentrar recursos em um subconjunto de UEs. O PF combina taxa instantânea e histórico de vazão, mas permanece associado a uma regra previamente definida (Mamane *et al.*, 2022; Sediq *et al.*, 2013).

Quando a demanda aumenta, as decisões de priorização se tornam mais frequentes. Nessas condições, a vazão agregada isolada pode ocultar atendimento insuficiente, redução do desempenho dos fluxos menos favorecidos ou concentração persistente de recursos (Sediq *et al.*, 2013; Guo *et al.*, 2014). A análise deve, portanto, relacionar eficiência, justiça, atendimento e desempenho individual.

Em aplicações de vídeo, o problema também possui uma dimensão temporal. Uma vazão média elevada não assegura que as unidades sintéticas de vídeo sejam entregues dentro do prazo considerado. Essas unidades, quando recebidas tardiamente, podem ser contabilizadas como recebidas pela rede sem integrar a parcela entregue no prazo. Médias de atraso também podem ocultar sequências de unidades tardias ou não recebidas. Por isso, a avaliação pode classificar as unidades segundo sua entrega temporal e inferir eventos de congelamento a partir da sequência dessas classificações (3GPP, 2024a,b).

O aprendizado por reforço oferece uma forma de obter políticas de escalonamento a partir dos estados observados e das consequências acumuladas das decisões (Sutton; Barto, 2018). Entretanto, o comportamento aprendido permanece condicionado à definição do estado, da ação, da recompensa e dos parâmetros estabelecidos para o treinamento ou para a execução (Al-Tam; Correia; Rodriguez, 2020; Gu *et al.*, 2021; Paz-Pérez *et al.*, 2024).

No IA-FMR, α_t integra o espaço de ações e é produzido juntamente com os valores que determinam a distribuição dos recursos (Lopes *et al.*, 2026). A proposta investiga se essa formulação pode aumentar o aproveitamento da célula sem reproduzir a concentração associada a políticas de escalonamento orientadas exclusivamente à taxa. O parâmetro

faz parte do modelo aprendido, mas seu efeito não é analisado de forma isolada dos demais componentes da política.

A tese também examina se as decisões produzidas pelo IA-FMR se relacionam com a entrega temporal em um cenário de tráfego sintético de vídeo. Essa avaliação é externa ao treinamento, pois o estado, a ação e a recompensa não incluem identificadores de unidades, prazos de apresentação, estado de reprodução ou eventos de congelamento inferidos.

1.2 Problema de Pesquisa

O problema investigado pode ser expresso pela seguinte questão:

Como projetar e avaliar um mecanismo de escalonamento de recursos para redes 5G NR capaz de conciliar vazão agregada e justiça, buscando preservar o atendimento dos usuários e o aproveitamento dos recursos disponíveis?

A questão envolve objetivos potencialmente conflitantes. A priorização de UEs com melhores condições de canal pode elevar a vazão agregada e concentrar a alocação. Em sentido oposto, uma distribuição uniforme pode preservar a justiça, mas limitar o aproveitamento das diferenças de eficiência espectral (Sediq *et al.*, 2013; Guo *et al.*, 2014).

Nesta tese, o escalonamento é formulado como um processo de decisão sequencial. A cada intervalo, o agente observa informações sobre os UEs e produz uma ação composta pelos valores que definem a distribuição dos recursos e por α_t . O parâmetro associado ao compromisso entre vazão e justiça é, assim, produzido conjuntamente com a distribuição, em vez de ser apenas estabelecido externamente como um valor fixo da formulação (Lopes *et al.*, 2026).

A avaliação combina métricas nos níveis da rede, do escalonamento e da aplicação. Nas avaliações sistêmicas, são considerados eficiência, justiça, atendimento, desempenho dos fluxos e distribuição dos recursos, permitindo diferenciar aumento de vazão com atendimento amplo de ganhos associados à concentração em parte dos usuários (Jain; Chiu; Hawe, 1998). Na avaliação com tráfego sintético de vídeo, acrescentam-se indicadores de entrega temporal e satisfação por usuário.

1.3 Objetivos

1.3.1 Objetivo Geral

O objetivo geral desta tese é propor e avaliar o IA-FMR, um mecanismo de escalonamento baseado em aprendizado por reforço para redes 5G NR, no qual o parâmetro associado ao compromisso entre vazão agregada e justiça integra a ação produzida pelo agente juntamente com os valores usados para definir a distribuição dos recursos.

1.3.2 Objetivos Específicos

Para alcançar o objetivo geral, são definidos os seguintes objetivos específicos:

- Desenvolver um mecanismo de escalonamento baseado em aprendizado por reforço capaz de conciliar justiça e eficiência na alocação de recursos em redes 5G NR;
- Integrar o mecanismo proposto ao ns-3/5G-LENA e avaliá-lo comparativamente em relação a escalonadores de referência;
- Avaliar o comportamento do mecanismo sob diferentes condições de disponibilidade de recursos, carga e número de UEs, considerando métricas de eficiência, justiça e atendimento;
- Investigar a aplicação do mecanismo em tráfego sintético de vídeo, considerando indicadores objetivos de entrega temporal das unidades.

1.4 Contribuições

As contribuições desta tese são:

1. O IA-FMR, com a inclusão de α_t no espaço de ações. A política de escalonamento produz conjuntamente os valores associados à distribuição alvo dos recursos e o componente usado na ponderação entre justiça e eficiência durante o treinamento;
2. A formulação do escalonamento como um processo de decisão cujo estado reúne MCS, alocações anteriores, vazão anterior, Índice de Jain e o valor anterior do coeficiente de ponderação. A recompensa combina justiça, eficiência em relação a uma referência de taxa e penalização do déficit de vazão. O limiar utilizado nessa penalização atua como referência de treinamento e não como garantia de QoS;

3. A implementação do IA-FMR no ns-3/5G-LENA e sua integração com uma política PPO executada em Python por meio do ns3-ai, permitindo empregar modelos treinados previamente nas decisões da camada MAC (Henderson *et al.*, 2008; Koutlia *et al.*, 2022; Yin *et al.*, 2020);
4. A definição de um mecanismo de compatibilidade para aplicar políticas de escalonamento treinadas com nove posições a conjuntos menores de UEs. As posições inativas são preenchidas com zeros e apenas os valores associados aos usuários elegíveis participam da distribuição dos recursos. A generalização estrutural para cardinalidades inferiores foi avaliada com 3, 5, 7 e 9 UEs, e a mesma representação foi utilizada no cenário de vídeo com 6 UEs;
5. A avaliação sistêmica com 100 sementes comuns nas configurações comparáveis, considerando vazão, justiça, atendimento, percentil 5 da vazão por fluxo (P5), perdas, atraso, demanda nas filas e distribuição temporal dos recursos;
6. A comparação quantitativa entre a alocação alvo produzida pelo modelo e a alocação aplicada no simulador, permitindo verificar quanto da decisão aprendida é preservado na execução;
7. A avaliação do IA-FMR com diferentes quantidades de usuários dentro da dimensionalidade prevista pela política, caracterizando justiça, atendimento e a degradação de desempenho observada à medida que o conjunto de UEs é alterado;
8. A instrumentação por unidade sintética de vídeo e a análise de indicadores objetivos de entrega temporal, incluindo taxa de unidades no prazo (TUP), composição entre unidades no prazo, tardias e não recebidas, frequência e duração dos congelamentos inferidos e satisfação temporal por usuário.

1.5 Metodologia Resumida

O IA-FMR foi modelado e treinado em ambiente Python com o algoritmo *Proximal Policy Optimization* (PPO) (Schulman *et al.*, 2017). Uma política de escalonamento foi treinada para cada largura de banda e, depois do treinamento, os modelos foram mantidos congelados durante as simulações. Para cada largura de banda foi realizada uma única execução final de treinamento, com a semente 211382627. A integração com o ns-3/5G-LENA foi realizada por meio do ns3-ai (Henderson *et al.*, 2008; Koutlia *et al.*, 2022; Yin *et al.*, 2020).

A avaliação experimental foi realizada por meio de simulações no ns-3/5G-LENA e organizada em três eixos. O primeiro comparou IA-FMR, RR, PF e MR em diferentes larguras de banda e incluiu a análise entre a distribuição alvo produzida pelo modelo e a alocação aplicada. O segundo avaliou a generalização estrutural do IA-FMR para diferentes números de UEs dentro da dimensionalidade prevista pela política. O terceiro examinou indicadores de entrega temporal em tráfego sintético de vídeo. As comparações sistêmicas utilizaram 100 sementes comuns aos mecanismos avaliados. Essas sementes caracterizam a variabilidade das simulações e não a variabilidade entre treinamentos independentes.

1.6 Organização da Tese

O restante desta tese está organizado da seguinte forma. O Capítulo 2 reúne os conceitos relacionados às redes 5G NR, à organização dos recursos de rádio, ao escalonamento na camada MAC, aos escalonadores de referência, às métricas de avaliação e aos fundamentos de aprendizado por reforço e PPO, incluindo o significado de seus principais parâmetros de treinamento. O Capítulo 3 discute trabalhos sobre aprendizado por reforço e técnicas inteligentes aplicadas à alocação de recursos, ao escalonamento e ao gerenciamento de redes 5G. O Capítulo 4 descreve o IA-FMR, incluindo a formulação do ambiente de aprendizado, o espaço de estados, o espaço de ações, a recompensa, o coeficiente de ponderação entre justiça e eficiência e a integração com o ns-3/5G-LENA. O Capítulo 5 apresenta o ambiente de simulação, os parâmetros, os cenários, os escalonadores de referência, a instrumentação de vídeo e as métricas. O Capítulo 6 apresenta os resultados das avaliações por largura de banda, por número de UEs e de disponibilidade temporal de vídeo. O Capítulo 7 sintetiza os resultados e as contribuições, discute as limitações e apresenta direções para trabalhos futuros.

2 Fundamentação Teórica

Este capítulo reúne os conceitos necessários para compreender a formulação do IA-FMR e interpretar as avaliações apresentadas nos capítulos seguintes. Inicialmente, são descritos a arquitetura das redes 5G, a interface rádio NR, a organização dos recursos de tempo e frequência, a relação entre qualidade de canal, MCS e eficiência espectral e o papel do escalonador na camada MAC. Em seguida, são discutidos os critérios utilizados para avaliar escalonamento, distinguindo métricas de desempenho, métricas de distribuição dos recursos e requisitos temporais de aplicação. Por fim, são apresentados os fundamentos de aprendizado por reforço, Processos de Decisão de Markov, métodos baseados em gradiente de política, arquiteturas *actor-critic*, o algoritmo Proximal Policy Optimization e o significado dos principais parâmetros empregados em seu treinamento.

2.1 Redes Móveis 5G

As redes móveis 5G foram desenvolvidas para ampliar a capacidade dos sistemas celulares e atender aplicações com requisitos distintos de taxa de transmissão, atraso, confiabilidade e densidade de conexões. Essa diversidade exige flexibilidade na configuração da interface rádio e no gerenciamento dos recursos disponíveis ([Andrews *et al.*, 2014](#); [Dahlman; Parkvall; Skold, 2023](#)).

No padrão New Radio (NR), parâmetros como largura de banda, espaçamento entre subportadoras, duração dos *slots*, modulação e codificação podem ser configurados de acordo com a faixa de frequência e com os requisitos de operação. Essa flexibilidade amplia as possibilidades de adaptação da transmissão, mas também aumenta a quantidade de decisões relacionadas à utilização e à distribuição dos recursos de rádio entre os usuários ([Dahlman; Parkvall; Skold, 2023](#); [3GPP, 2026c](#)).

2.1.1 Arquitetura e Serviços do 5G

As redes 5G combinam tecnologias de múltiplas antenas, formação de feixes, virtualização de funções de rede, densificação de células e operação em diferentes faixas de frequência. Em conjunto, essas tecnologias permitem ampliar a capacidade do sistema e atender perfis heterogêneos de comunicação (Andrews *et al.*, 2014; Dahlman; Parkvall; Skold, 2023).

O MIMO massivo (*massive Multiple-Input Multiple-Output*) utiliza arranjos com vários elementos de antena, principalmente na gNB, para explorar ganho de arranjo e multiplexação espacial. A formação de feixes (*beamforming*) complementa esse funcionamento ao ajustar os sinais aplicados aos elementos do arranjo, concentrando energia em determinadas direções e modificando as condições do enlace com os UEs (Andrews *et al.*, 2014; Dahlman; Parkvall; Skold, 2023).

Os requisitos do 5G são comumente organizados em três cenários de uso de referência. O eMBB prioriza capacidade e elevadas taxas de transmissão. O URLLC está associado a aplicações que exigem baixa latência e elevada confiabilidade. O mMTC está relacionado a cenários com grande quantidade de dispositivos e transmissões de menor volume por equipamento. Essas categorias representam perfis de referência e não classes mutuamente exclusivas de aplicações (Dahlman; Parkvall; Skold, 2023; 3GPP, 2026c).

A padronização do sistema é conduzida pelo 3GPP, responsável pelas especificações do NR e da NG-RAN. A arquitetura 5G é composta pelo núcleo da rede, denominado 5G Core (5GC), e pela rede de acesso rádio. O 5GC executa funções relacionadas a autenticação, mobilidade, sessões, encaminhamento de dados e aplicação de políticas, enquanto a NG-RAN realiza a comunicação entre o núcleo e os UEs (3GPP, 2026c).

Na NG-RAN, as gNBs implementam as camadas PHY, MAC, RLC e PDCP. A camada PHY está associada à transmissão física. As camadas RLC e PDCP executam funções de tratamento, segmentação e entrega dos dados. A camada MAC realiza multiplexação, adaptação entre canais lógicos e de transporte e escalonamento dos recursos de rádio (3GPP, 2026c).

A Figura 2 apresenta essa organização. Os UEs acessam a gNB pela interface rádio NR, em transmissões de *uplink* ou *downlink*. A gNB implementa as camadas de protocolo do acesso rádio, enquanto o 5GC executa funções de controle e conexão com redes de dados externas.

A camada MAC, destacada na figura, contém o escalonador investigado nesta tese. Em cada decisão, esse mecanismo determina quais UEs receberão recursos e em que

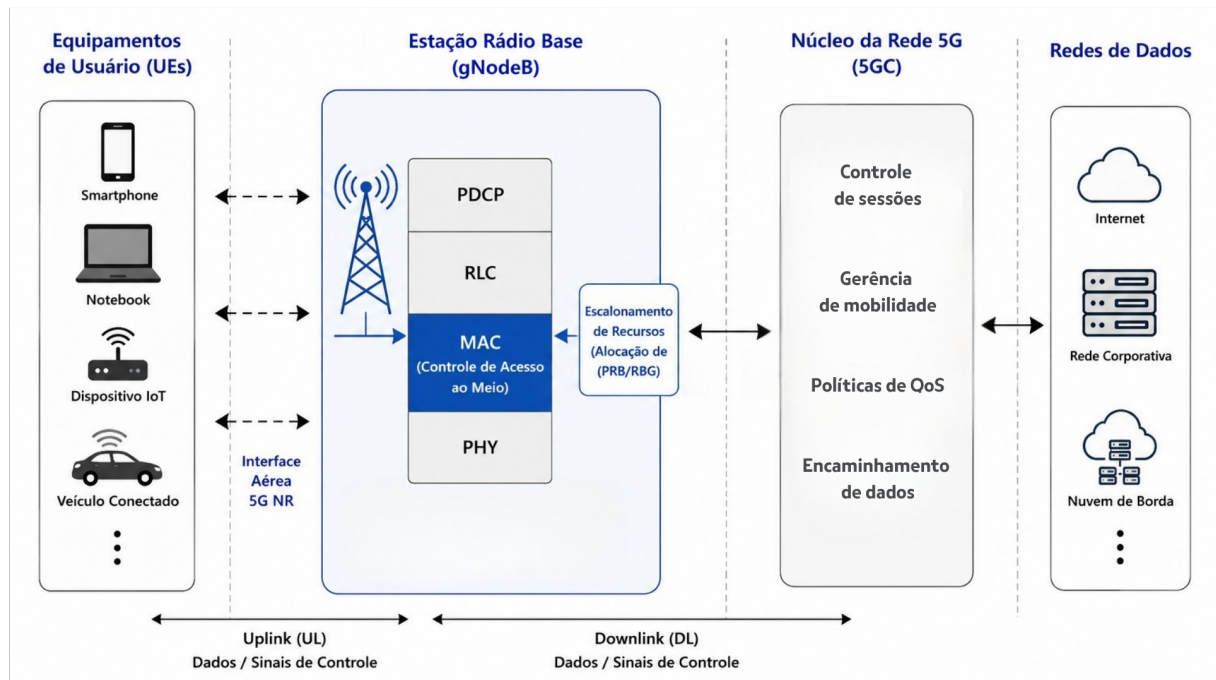


Figura 2: Arquitetura simplificada de uma rede 5G NR com destaque para o escalonamento de recursos na camada MAC da gNodeB.

quantidade. Dependendo da política utilizada, a decisão pode considerar condições de canal, dados pendentes nas filas, histórico de atendimento e parâmetros associados aos fluxos. Essas escolhas afetam métricas como vazão, justiça, atendimento, perdas e atraso.

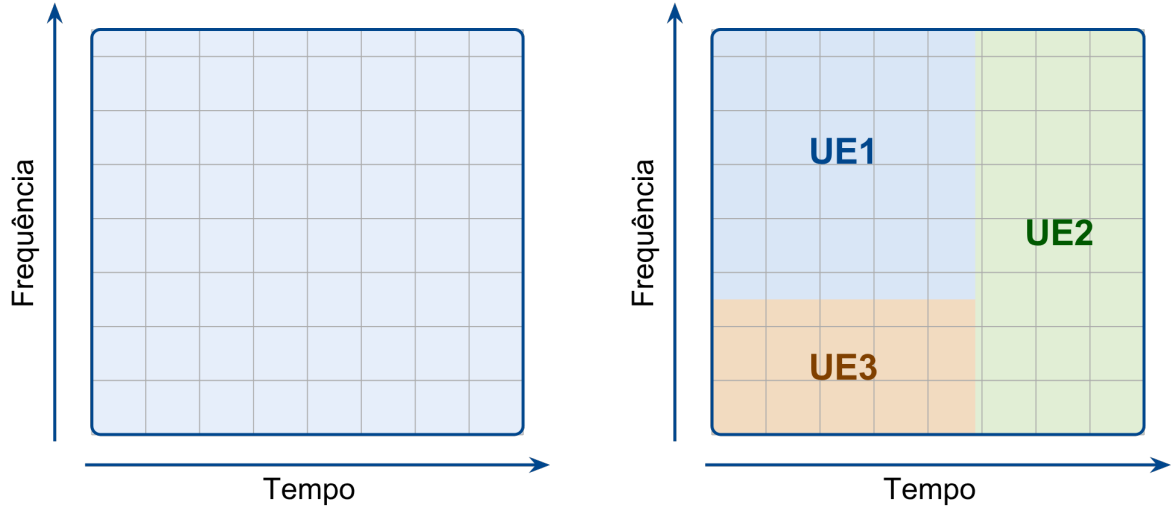
2.1.2 OFDM e OFDMA no 5G NR

O NR utiliza a multiplexação por divisão de frequência ortogonal, ou OFDM, como base para a transmissão. No OFDM, a largura de banda é dividida em subportadoras ortogonais, permitindo a transmissão paralela de símbolos e facilitando o tratamento de efeitos de propagação por múltiplos percursos (Dahlman; Parkvall; Skold, 2023; 3GPP, 2026a).

A ortogonalidade permite que as subportadoras ocupem intervalos espectrais sobrepostos sem interferência mútua nas condições ideais de sincronização previstas pela formulação. A divisão da transmissão entre várias subportadoras também reduz a taxa de símbolos em cada uma delas, tornando o sistema mais adequado ao tratamento do desvanecimento seletivo em frequência.

A OFDMA estende esse princípio para o acesso de múltiplos usuários. Enquanto o OFDM descreve a organização da transmissão em subportadoras ortogonais, a OFDMA permite atribuir subconjuntos dos recursos de tempo e frequência a usuários diferentes (Dahlman; Parkvall; Skold, 2023; 3GPP, 2026a).

A Figura 3 resume essa diferença. A Figura 3(a) representa a organização OFDM em subportadoras ortogonais, enquanto a Figura 3(b) mostra regiões do plano de tempo e frequência atribuídas a UEs distintos.



(a) Representação conceitual do OFDM, com a transmissão organizada em subportadoras ortogonais no plano de tempo e frequência.

(b) Representação conceitual do OFDMA, com regiões do plano de tempo e frequência atribuídas a diferentes UEs.

Figura 3: Diferença conceitual entre OFDM e OFDMA na utilização dos recursos de tempo e frequência.

Dessa forma, a gNB pode multiplexar vários UEs sobre a mesma largura de banda. O escalonador da camada MAC determina como esses recursos serão distribuídos ao longo das decisões de transmissão.

2.1.3 Numerologia e Estrutura Temporal

A numerologia do NR define parâmetros associados à organização temporal e espectral da transmissão. O espaçamento entre subportadoras é dado por:

$$\Delta f = 15 \times 2^\mu \text{ kHz}, \quad (2.1)$$

em que μ representa o índice da numerologia. Cada incremento de μ duplica o espaçamento entre as subportadoras. Para os valores padronizados, são obtidos espaçamentos de 15, 30, 60, 120 e 240 kHz (3GPP, 2026a).

Considerando prefixo cíclico normal, cada *slot* contém 14 símbolos OFDM. Sua duração nominal é inversamente proporcional a 2^μ :

$$T_{\text{slot}} = \frac{1}{2^\mu} \text{ ms.} \quad (2.2)$$

Assim, $\mu = 0$ corresponde a *slots* de 1 ms, $\mu = 1$ produz *slots* de 0,5 ms e $\mu = 2$ produz *slots* de 0,25 ms (3GPP, 2026a). Como a duração de um subquadro permanece fixa em 1 ms, a quantidade de *slots* contida em cada subquadro é:

$$N_{\text{slots/subquadro}} = 2^\mu. \quad (2.3)$$

A estrutura temporal do NR utiliza quadros de 10 ms, formados por dez subquadros de 1 ms. A Figura 4 relaciona quadro, subquadro e *slot* e compara as numerologias $\mu = 0$, $\mu = 1$ e $\mu = 2$. O subquadro contém, respectivamente, um *slot* de 1 ms, dois *slots* de 0,5 ms ou quatro *slots* de 0,25 ms.

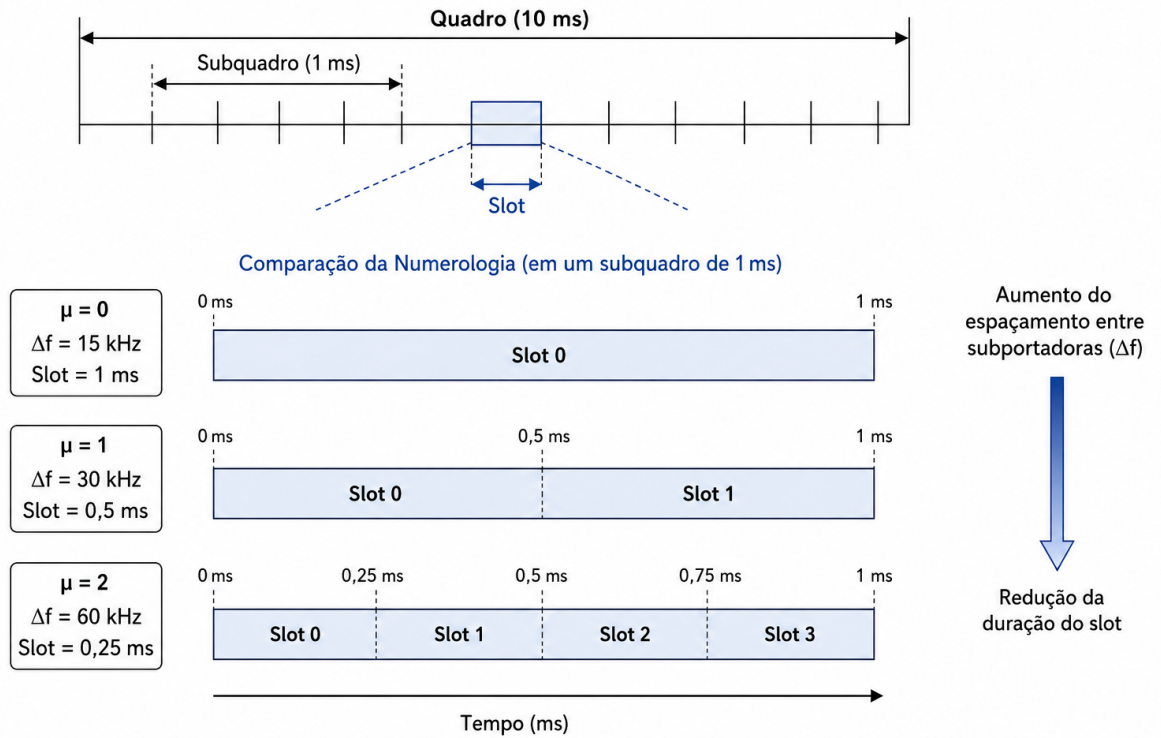


Figura 4: Estrutura temporal do 5G NR e relação entre o índice de numerologia, o espaçamento entre subportadoras, a duração dos *slots* e a quantidade de *slots* por subquadro.

O aumento do espaçamento entre subportadoras reduz a duração dos símbolos e dos *slots*. Isso permite intervalos temporais menores para determinadas operações, mas também altera a relação do sistema com efeitos de propagação, como dispersão temporal e deslocamento Doppler. A escolha da numerologia envolve, portanto, características tem-

porais, espectrais e de propagação (3GPP, 2026a).

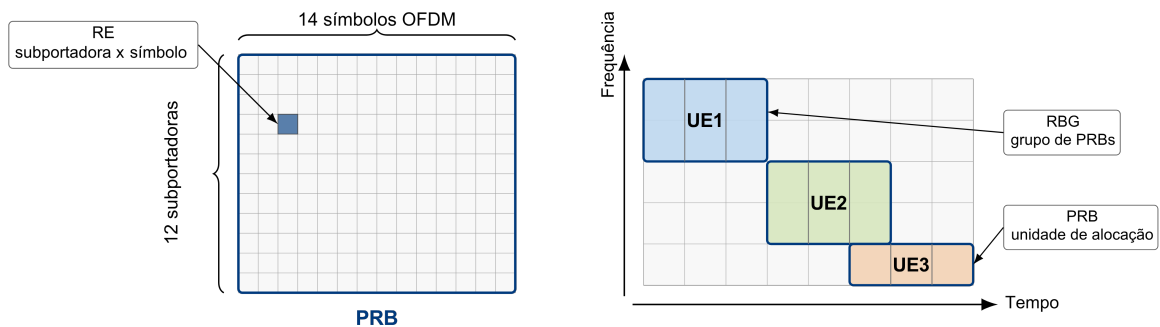
2.1.4 Organização dos Recursos de Rádio

Os recursos de rádio no NR são representados em uma grade bidimensional nos domínios do tempo e da frequência. O eixo temporal é organizado em símbolos OFDM, enquanto o eixo da frequência é dividido em subportadoras. A menor unidade dessa grade é o elemento de recurso, ou *Resource Element* (RE), definido pela interseção entre uma subportadora e um símbolo OFDM (3GPP, 2026a).

Um bloco de recursos corresponde a 12 subportadoras consecutivas no domínio da frequência. Quando referenciado à posição física ocupada dentro da largura de banda utilizada, esse bloco é denominado *Physical Resource Block* (PRB). Considerando um *slot* com prefixo cíclico normal, formado por 14 símbolos OFDM, a região associada a um PRB contém até $12 \times 14 = 168$ posições de RE. Nem todos esses elementos são necessariamente utilizados para dados do usuário, pois parte deles pode ser destinada a sinais de referência, controle e outros procedimentos físicos (3GPP, 2026a).

Para determinadas formas de alocação, PRBs consecutivos são reunidos em *Resource Block Groups* (RBGs). O tamanho de cada RBG depende do número de PRBs da parte de largura de banda, ou *Bandwidth Part* (BWP), e da configuração de alocação adotada (3GPP, 2026d). O RBG é, portanto, uma unidade de agrupamento empregada pelo procedimento de alocação e não deve ser confundido com o PRB em todos os cenários.

A Figura 5 apresenta essas estruturas no plano de tempo e frequência.



(a) Formação dos REs e representação das 12 subportadoras associadas a um PRB ao longo dos símbolos OFDM.

(b) Agrupamento de PRBs consecutivos em RBGs e exemplo de distribuição entre diferentes UEs.

Figura 5: Organização conceitual dos recursos de rádio no plano tempo e frequência, com exemplo de sua distribuição entre usuários.

Na Figura 5(a), cada célula representa um RE e o conjunto indicado de 12 subpor-

tadoras corresponde à extensão de um PRB em frequência. A Figura 5(b) mostra PRBs agrupados em RBGs e distribuídos entre três UEs.

Nesta tese, a decisão de escalonamento é representada pela quantidade de RBGs destinada a cada UE em cada decisão. A mesma unidade é utilizada para caracterizar a distribuição dos recursos. A correspondência específica entre RBG e PRB adotada nos experimentos é apresentada no Capítulo 5.

2.1.5 Modulação, Codificação, CQI e MCS

A quantidade de dados que pode ser transportada por uma parcela de recursos depende, entre outros fatores, das condições do canal e do esquema de modulação e codificação. No NR, diferentes combinações de ordem de modulação e taxa de codificação são representadas por índices de *Modulation and Coding Scheme* (MCS) (3GPP, 2026d).

Os UEs podem reportar informações relacionadas à qualidade do canal, incluindo o *Channel Quality Indicator* (CQI). Essas informações participam do processo pelo qual a gNB seleciona uma configuração de transmissão compatível com as condições estimadas do enlace (3GPP, 2026d). CQI e MCS não são a mesma grandeza. O CQI informa uma avaliação da condição do canal, enquanto o MCS identifica a combinação de modulação e codificação utilizada na transmissão.

Valores de MCS associados a modulações de maior ordem e taxas de codificação maiores permitem transportar mais bits por unidade de recurso quando as condições do enlace suportam essa escolha. Configurações mais conservadoras reduzem a quantidade de informação transportada por recurso, mas oferecem maior robustez. Como consequência, dois UEs que recebem a mesma quantidade de recursos podem alcançar vazões diferentes.

A eficiência espectral expressa essa capacidade de transporte em relação à largura de banda utilizada. Em forma genérica,

$$\eta = \frac{R}{B}, \quad (2.4)$$

em que R representa uma taxa de transmissão e B a largura de banda correspondente. A unidade usual é bit/s/Hz. Nesta tese, a relação entre MCS e eficiência espectral é utilizada tanto na formulação simplificada do ambiente de treinamento quanto na interpretação de políticas que exploram condições instantâneas de canal. A forma específica utilizada pelo IA-FMR é definida no Capítulo 4.

2.2 Escalonamento de Recursos em Redes 5G

O escalonamento de recursos é executado pela camada MAC da gNB e define como os recursos disponíveis serão distribuídos entre os UEs elegíveis. A política de escalonamento determina quais informações serão utilizadas e como elas serão combinadas para selecionar os usuários atendidos e a quantidade de recursos destinada a cada um (Mamane *et al.*, 2022; Dahlman; Parkvall; Skold, 2023).

Entre as informações que podem subsidiar a decisão estão as estimativas de taxa alcançável derivadas da qualidade do canal, a quantidade de dados pendentes nas filas de transmissão, o histórico de vazão ou de atendimento e parâmetros de QoS associados aos fluxos. O conjunto efetivamente utilizado depende da política implementada (Mamane *et al.*, 2022; Dahlman; Parkvall; Skold, 2023).

Em transmissões OFDMA, a decisão corresponde à atribuição de unidades da grade de tempo e frequência aos diferentes UEs. Uma política que prioriza taxas instantâneas elevadas tende a explorar UEs em condições favoráveis de canal. Uma política orientada à regularidade do atendimento distribui oportunidades de transmissão entre um conjunto maior de usuários. Essas escolhas afetam simultaneamente o aproveitamento dos recursos e sua distribuição (Guo *et al.*, 2014; Sediq *et al.*, 2013).

A qualidade de uma política não pode ser inferida a partir de uma única métrica. Vazão agregada elevada pode coexistir com concentração de recursos. Justiça elevada na distribuição de RBGs não implica igualdade de vazão, pois os UEs podem utilizar os mesmos recursos com eficiências espectrais distintas. Baixo atraso médio também não demonstra, isoladamente, atendimento equilibrado, pois a média é calculada sobre os dados efetivamente recebidos. Por esse motivo, a avaliação do escalonamento utiliza métricas complementares (Guo *et al.*, 2014; Sediq *et al.*, 2013).

2.2.1 Escalonadores de Referência

Entre as políticas utilizadas como referência nesta tese estão Round Robin (RR), Max Rate (MR) e Proportional Fair (PF) (Mamane *et al.*, 2022).

O RR distribui oportunidades de atendimento de maneira cíclica entre os UEs elegíveis. Como seu critério principal não prioriza diretamente as diferenças de taxa alcançável, tende a produzir atendimento regular entre os usuários, embora possa deixar de explorar condições favoráveis de canal.

O MR prioriza os usuários com maiores taxas instantâneas estimadas. Esse critério procura aumentar o aproveitamento da capacidade instantânea ao direcionar os recursos aos UEs capazes de utilizá-los com maior eficiência espectral. Usuários que permanecem em condições desfavoráveis podem, contudo, receber poucos recursos ou permanecer sem atendimento por intervalos prolongados (Guo *et al.*, 2014).

O PF utiliza uma métrica que relaciona a taxa instantânea estimada de cada UE ao seu histórico de vazão. Em uma representação geral, a prioridade do usuário i no instante t pode ser expressa por:

$$M_i(t) = \frac{R_i(t)}{\bar{R}_i(t)}, \quad (2.5)$$

em que $R_i(t)$ representa a taxa alcançável no instante atual e $\bar{R}_i(t)$ representa uma média histórica de atendimento. O critério favorece UEs que apresentam boa oportunidade instantânea em relação ao seu histórico, combinando informação corrente e desempenho passado. A forma exata de atualização do histórico depende da implementação do escalonador (Mamane *et al.*, 2022).

Além dessas políticas, o escalonamento pode ser formulado como um problema de decisão aprendido por interação com o ambiente. Essa possibilidade é retomada nas Seções 2.4 e 2.6.

2.3 Métricas de Avaliação, QoS e Requisitos Temporais

Uma métrica descreve uma propriedade observada na execução. Um requisito, por outro lado, estabelece uma condição, alvo ou limite que se deseja atender. Essa distinção é importante porque valores utilizados em uma função de recompensa ou em um critério de análise não se tornam, automaticamente, garantias oferecidas pela rede.

Na avaliação desta tese, as métricas são organizadas em três grupos. O primeiro descreve desempenho de entrega, como vazão, perdas e atraso. O segundo descreve a distribuição e o atendimento dos recursos, como Índice de Jain, quantidade de UEs servidos, demanda pendente sem alocação e *backlog*. O terceiro caracteriza o cumprimento de prazos no cenário sintético de vídeo. As métricas de rede e justiça seguem conceitos consolidados na literatura de comunicação e escalonamento (Dahlman; Parkvall; Skold, 2023; Jain; Chiu; Hawe, 1998; Guo *et al.*, 2014; Sediq *et al.*, 2013). As métricas operacionais derivadas dos registros do escalonador e os critérios temporais específicos da tese

são definidos explicitamente a seguir e detalhados metodologicamente no Capítulo 5.

2.3.1 Vazão, Percentis, Perdas e Atraso

A vazão mede a taxa efetivamente entregue durante um intervalo de observação. Para um fluxo f , uma forma genérica de cálculo é

$$T_f = \frac{8B_f^{\text{rx}}}{\Delta t}, \quad (2.6)$$

em que B_f^{rx} representa a quantidade de bytes recebidos e Δt o intervalo de observação. Quando o resultado é expresso em Mbps, aplica-se o fator de conversão correspondente. A vazão agregada é obtida pela soma das vazões dos fluxos considerados. Essa métrica caracteriza a quantidade total de dados entregue por unidade de tempo, mas não descreve como esse desempenho se distribui entre os fluxos (Dahlman; Parkvall; Skold, 2023; Sediq *et al.*, 2013).

Para observar a região inferior dessa distribuição, esta tese utiliza o percentil 5 da vazão por fluxo. O P5 é o valor abaixo do qual se encontram aproximadamente 5% das observações da distribuição considerada. Diferentemente do valor mínimo, o percentil não depende de uma única observação extrema. No contexto desta tese, um P5 baixo indica que a região inferior da distribuição contém fluxos com vazões pequenas, mesmo que a vazão agregada seja elevada.

A taxa de perda caracteriza a fração de unidades transmitidas que não são recebidas. Em uma representação por pacotes,

$$L_f = \frac{N_f^{\text{tx}} - N_f^{\text{rx}}}{N_f^{\text{tx}}} \times 100, \quad (2.7)$$

em que N_f^{tx} e N_f^{rx} representam as quantidades transmitida e recebida. A taxa de perda informa ausência de entrega, mas não revela em que momento ocorreram as perdas nem se os pacotes recebidos chegaram dentro de um prazo relevante para a aplicação.

O atraso de um pacote recebido é a diferença entre seus instantes de recepção e transmissão. Para os N_f^{rx} pacotes recebidos de um fluxo,

$$\overline{D}_f = \frac{1}{N_f^{\text{rx}}} \sum_{k=1}^{N_f^{\text{rx}}} (t_k^{\text{rx}} - t_k^{\text{tx}}). \quad (2.8)$$

A média é condicionada aos pacotes efetivamente recebidos. Por isso, um mecanismo que entrega apenas uma parcela dos pacotes pode apresentar atraso médio baixo entre os pacotes recebidos e, ao mesmo tempo, perda elevada. A interpretação conjunta das métricas evita atribuir ao atraso uma informação que pertence à taxa de entrega.

2.3.2 Justiça e Distribuição dos Recursos

A justiça caracteriza o grau de uniformidade de uma distribuição. Em problemas de escalonamento, ela pode ser aplicada à vazão, ao tempo de atendimento ou à quantidade de recursos destinada aos usuários. Uma métrica amplamente utilizada para essa finalidade é o Índice de Jain ([Jain; Chiu; Hawe, 1998](#)):

$$J(\mathbf{x}) = \frac{(\sum_{i=1}^n x_i)^2}{n \sum_{i=1}^n x_i^2}, \quad (2.9)$$

em que x_i representa o valor atribuído ao usuário i e n a quantidade de usuários considerados. Para um vetor não negativo com pelo menos um componente positivo, $J(\mathbf{x})$ pertence ao intervalo $[1/n, 1]$. O valor 1 ocorre quando todos os componentes são iguais. Valores mais próximos de $1/n$ correspondem a distribuições mais concentradas. Quando todos os componentes são nulos, a expressão possui denominador nulo e a implementação deve adotar uma convenção para esse caso.

O significado do índice depende da grandeza utilizada em $\mathbf{x}$. Um Jain calculado sobre vazões mede uniformidade das vazões. Um Jain calculado sobre RBGs mede uniformidade da distribuição dos recursos. Nesta tese, o indicador empregado na recompensa e na avaliação principal é calculado sobre os RBGs atribuídos aos UEs ([Lopes et al., 2026](#); [Guo et al., 2014](#)).

Essa escolha separa duas dimensões que não são equivalentes. Se dois UEs recebem a mesma quantidade de RBGs, o Jain dos recursos pode ser igual a 1 para esse par, mesmo que suas vazões sejam diferentes devido ao MCS. Consequentemente, justiça na alocação não representa garantia de vazão mínima, ausência de perdas ou atendimento de um requisito de QoS. Por essa razão, o Jain é sempre interpretado em conjunto com as demais métricas.

2.3.3 Atendimento, Demanda Pendente e Backlog

O número de UEs servidos em uma decisão é uma métrica operacional que conta quantos usuários receberam alocação positiva. Seja $A_i(t)$ a quantidade de RBGs aplicada ao UE i na decisão t . Então,

$$N_{\text{servidos}}(t) = \sum_{i=1}^N \mathbb{I}[A_i(t) > 0], \quad (2.10)$$

em que $\mathbb{I}[\cdot]$ é a função indicadora. Essa métrica mostra a amplitude do atendimento em uma decisão, mas não informa a quantidade de recursos recebida por cada UE nem a vazão resultante.

Para diferenciar ausência de alocação de ausência de demanda, utiliza-se a proporção de ocorrências em que existe dado pendente, mas nenhum recurso é aplicado. Seja $q_i(t)$ a demanda pendente informada ao escalonador. A proporção é:

$$P_{\text{demanda sem alocação}} = 100 \frac{\sum_{t,i} \mathbb{I}[q_i(t) > 0 \wedge A_i(t) = 0]}{\sum_{t,i} \mathbb{I}[q_i(t) > 0]}. \quad (2.11)$$

O denominador considera somente ocorrências em que havia demanda pendente. Dessa forma, um UE sem dados para transmitir não é contabilizado como falta de atendimento. Essa métrica foi definida para a análise operacional desta tese a partir dos registros do escalonador. Uma ocorrência isolada com demanda e alocação nula não caracteriza, por si só, inanição. A interpretação de *starvation* exige observar persistência temporal e seus efeitos sobre vazão e filas (Guo *et al.*, 2014).

O *backlog* representa dados acumulados aguardando atendimento no ponto da pilha observado pelo escalonador. Na avaliação sistêmica, a grandeza é obtida a partir da ocupação informada pelas filas RLC/LCG e agregada entre os UEs. Se $q_i(t)$ é expresso em bytes, então:

$$B(t) = \frac{8}{10^6} \sum_{i=1}^N q_i(t), \quad (2.12)$$

produz o resultado em Mbit. Um *backlog* elevado indica acúmulo de dados pendentes, mas não determina sozinho sua causa. A grandeza pode refletir a relação entre carga oferecida, capacidade disponível, condições de canal e decisões de escalonamento.

2.3.4 QoS, Requisitos e Critérios de Treinamento

A Qualidade de Serviço, ou *Quality of Service* (QoS), reúne propriedades mensuráveis relevantes ao atendimento dos fluxos, como taxa de transmissão, atraso, perdas e confiabilidade. A importância relativa desses requisitos depende do serviço considerado ([Dahlman; Parkvall; Skold, 2023](#); [3GPP, 2026c](#)).

É importante distinguir um requisito de serviço de um valor utilizado internamente por um algoritmo de aprendizado. Um limiar inserido em uma função de recompensa pode incentivar determinado comportamento durante o treinamento, mas não constitui uma reserva de recursos ou uma garantia de QoS, a menos que exista um mecanismo que imponha explicitamente essa condição durante a execução. Essa distinção é utilizada no Capítulo 4 para interpretar o limiar de vazão empregado na penalização da recompensa do IA-FMR.

Da mesma forma, uma métrica de justiça não substitui um requisito de atendimento. Um Jain elevado informa que a grandeza avaliada está distribuída de forma uniforme. Ele não estabelece que cada fluxo alcançou uma vazão específica, que nenhum pacote foi perdido ou que um prazo de entrega foi atendido.

2.3.5 Requisitos Temporais no Nível da Aplicação

A Qualidade de Experiência (*Quality of Experience*, QoE) adota uma perspectiva centrada no usuário e pode ser influenciada pelo desempenho técnico, pelas características da aplicação, pelo contexto de uso e pelas expectativas do usuário ([International Telecommunication Union, 2017](#)). QoS e QoE são relacionadas, mas não são equivalentes. Além disso, a avaliação temporal realizada nesta tese não constitui uma caracterização completa de QoE.

Em serviços de mídia, o instante de entrega pode ser relevante para a utilidade da informação. Uma unidade recebida após o prazo considerado pela aplicação pode ter sido entregue pela rede e, ainda assim, não satisfazer o critério temporal adotado. Referências do 3GPP para mídia e XR consideram requisitos temporais como parte da avaliação desses serviços ([3GPP, 2024a,b](#)).

Nesta tese, uma unidade sintética de vídeo corresponde a uma unidade de tráfego para a qual são registrados instante de geração, instante de recepção e atraso de entrega. Para um prazo d , cada unidade pertence a uma entre três categorias: recebida dentro do

prazo, recebida após o prazo ou não recebida até o encerramento da simulação. Essas categorias são mutuamente exclusivas e totalizam as unidades geradas.

A proporção de unidades recebidas dentro do prazo é denominada Taxa de Unidades no Prazo (TUP):

$$\text{TUP}_d = \frac{N_{\text{prazo},d}}{N_{\text{geradas}}} \times 100. \quad (2.13)$$

em que $N_{\text{prazo},d}$ representa a quantidade de unidades recebidas com atraso menor ou igual a d , e N_{geradas} corresponde ao total de unidades produzidas. As unidades recebidas após o prazo e as unidades não recebidas permanecem no denominador e não integram o numerador.

A TUP utilizada nesta tese é uma métrica operacional de pós-processamento inspirada na necessidade de distinguir entrega de cumprimento de prazo. Ela não é apresentada como uma métrica subjetiva de QoE nem como requisito universal definido pelo 3GPP. Os prazos concretos adotados nos experimentos são escolhas metodológicas fundamentadas nas referências de serviço discutidas no Capítulo 5.

A ordem temporal das unidades também permite inferir eventos de congelamento. Depois que a reprodução inferida é iniciada por uma unidade recebida no prazo, um evento corresponde a uma sequência contínua de unidades que deixam de satisfazer o prazo e termina quando uma nova unidade volta a satisfazê-lo. A frequência informa quantos eventos ocorrem em determinado intervalo, enquanto a duração caracteriza sua extensão temporal. Essas duas métricas devem ser analisadas conjuntamente, pois uma frequência menor pode coexistir com episódios mais longos.

Nesta tese, um UE é classificado como temporalmente satisfeito quando:

$$\text{TUP}_d^{(u)} \geq 99\%. \quad (2.14)$$

Esse valor de 99% é um critério operacional adotado para comparar os mecanismos sob o mesmo prazo. Ele não representa limiar de percepção humana, requisito normativo universal ou garantia oferecida pela rede.

A avaliação não executa um reprodutor de vídeo completo e não calcula qualidade visual, *Mean Opinion Score*, estado de *playout buffer* adaptativo ou métricas de interação. Também não representa uma cadeia completa de Realidade Estendida, que poderia incluir

captura de pose, seleção de *viewport*, renderização, codificação, decodificação e dispositivo de apresentação. Os indicadores caracterizam, portanto, propriedades temporais da entrega das unidades sintéticas.

2.4 Aprendizado por Reforço

O aprendizado por reforço, ou *Reinforcement Learning* (RL), baseia-se na interação entre um agente e um ambiente. O agente observa informações do ambiente, seleciona uma ação e recebe uma recompensa associada às consequências da decisão, buscando uma política que maximize o retorno acumulado (Sutton; Barto, 2018). Diferentemente do aprendizado supervisionado, não existe uma ação correta previamente rotulada para cada estado. O aprendizado decorre das transições e recompensas observadas (Sutton; Barto, 2018).

Quando políticas e funções de valor são aproximadas por redes neurais profundas, utiliza-se a denominação aprendizado por reforço profundo, ou *Deep Reinforcement Learning* (DRL). A rede neural atua como aproximador de funções e seus parâmetros são ajustados a partir das experiências produzidas pela interação com o ambiente.

2.4.1 Estado, Ação, Política, Recompensa e Retorno

Em cada instante de decisão t , o agente observa um estado s_t e seleciona uma ação $\mathbf{a}_t$ de acordo com sua política. Após a execução da ação, o ambiente produz uma recompensa e evolui para um novo estado (Sutton; Barto, 2018).

O estado é a representação das informações disponibilizadas ao agente para decidir. Ele não precisa reproduzir todo o ambiente físico, mas deve conter as variáveis utilizadas pela política. A ação representa a decisão produzida pelo agente. Dependendo do problema, ela pode ser discreta, contínua ou composta por vários componentes.

A política estabelece como as ações são escolhidas. Uma política determinística associa cada estado a uma ação. Uma política estocástica define uma distribuição de probabilidade condicionada ao estado, representada por $\pi(\mathbf{a} \mid s)$ (Sutton; Barto, 2018). Durante o treinamento, a natureza estocástica da política pode contribuir para exploração do espaço de ações. Na inferência, a implementação pode utilizar uma regra determinística derivada da política aprendida.

A recompensa r_{t+1} é o sinal produzido pelo ambiente após a decisão. Ela não corresponde necessariamente a uma métrica física única. Em problemas multiobjetivo, pode

combinar diferentes termos para orientar o comportamento aprendido.

O retorno acumulado a partir do instante t é:

$$G_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}, \quad (2.15)$$

em que $\gamma \in [0,1]$ é o fator de desconto. Valores próximos de 1 atribuem maior peso relativo às consequências futuras. Valores menores fazem o retorno decair mais rapidamente com a distância temporal (Sutton; Barto, 2018).

O processo de aprendizado envolve um compromisso entre exploração e aproveitamento (*exploration-exploitation*). A exploração produz experiências em regiões ainda pouco conhecidas do espaço de ações. O aproveitamento utiliza o conhecimento acumulado para selecionar ações já associadas a bons retornos. O equilíbrio entre essas duas tendências influencia o conjunto de experiências utilizado para atualizar a política (Sutton; Barto, 2018).

2.4.2 Processos de Decisão de Markov

Problemas de RL são frequentemente modelados como Processos de Decisão de Markov, ou *Markov Decision Processes* (MDPs). Um MDP pode ser representado pela tupla

$$\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, P, R, \gamma \rangle, \quad (2.16)$$

em que $\mathcal{S}$ é o espaço de estados, $\mathcal{A}$ o espaço de ações, P a dinâmica de transição, R a função de recompensa e γ o fator de desconto (Sutton; Barto, 2018).

A propriedade de Markov estabelece que, dado o estado atual e a ação executada, a distribuição do próximo estado pode ser determinada sem depender explicitamente de toda a sequência anterior de estados. Isso não significa que o passado seja irrelevante. Quando informações históricas são necessárias à decisão, elas podem ser incorporadas à representação do estado atual.

A Figura 6 representa esse ciclo no contexto do escalonamento. O ambiente contém UEs, condições de canal, dados pendentes e recursos disponíveis. O estado s_t contém a parcela dessas informações apresentada ao agente. A política produz $\mathbf{a}_t$, o ambiente processa a decisão e fornece a recompensa e o estado seguinte.

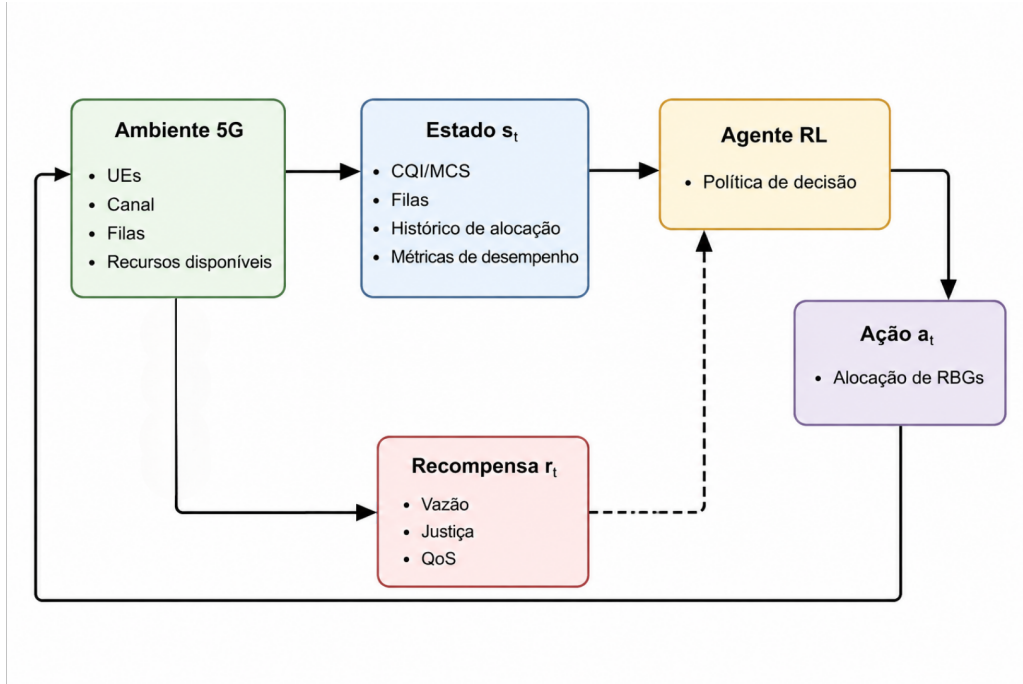


Figura 6: Representação do problema de escalonamento de recursos como um Processo de Decisão de Markov.

2.4.3 Funções de Valor e Vantagem

A função de valor de estado sob uma política π representa o retorno esperado ao iniciar no estado s e seguir essa política:

$$V^\pi(s) = \mathbb{E}_\pi [G_t \mid s_t = s]. \quad (2.17)$$

A função de valor ação-estado representa o retorno esperado quando a ação a é executada em s e a política é seguida posteriormente:

$$Q^\pi(s, a) = \mathbb{E}_\pi [G_t \mid s_t = s, a_t = a]. \quad (2.18)$$

A vantagem compara o valor esperado de uma ação ao valor de referência do estado:

$$A^\pi(s, a) = Q^\pi(s, a) - V^\pi(s). \quad (2.19)$$

Uma vantagem positiva indica que a ação é estimada como melhor que a referência representada por $V^\pi(s)$. Uma vantagem negativa indica resultado estimado abaixo dessa referência. Essa grandeza é utilizada em métodos de gradiente de política para orientar a

direção das atualizações (Sutton; Barto, 2018; You *et al.*, 2019).

2.4.4 Função de Recompensa e Problemas Multiobjetivo

A formulação da recompensa traduz para o processo de aprendizado os objetivos considerados durante o treinamento. Em escalonamento, a recompensa pode incorporar vazão, distribuição dos recursos, atendimento, atraso, perdas ou penalizações relacionadas a condições indesejadas.

Quando diferentes objetivos são combinados em uma única recompensa, os coeficientes de ponderação estabelecem sua importância relativa naquela formulação. Isso não significa que a política resolva uma otimização multiobjetivo sem preferências. A recompensa escalar define uma forma específica de combinar os objetivos. Se um coeficiente também for produzido pela própria política, a formulação passa a permitir que esse componente varie entre decisões, mas o comportamento efetivamente aprendido precisa ser observado empiricamente.

Uma penalização inserida na recompensa atua durante o treinamento como sinal de aprendizado. Ela não implica, por si só, que a mesma condição seja verificada ou imposta durante a inferência. Essa distinção é relevante para a interpretação da penalização por déficit de vazão definida para o IA-FMR no Capítulo 4.

2.4.5 Métodos Baseados em Gradiente de Política

Métodos de gradiente de política ajustam diretamente os parâmetros de uma política parametrizada para aumentar o retorno esperado (Sutton; Barto, 2018). Seja $\pi_{\theta}(\mathbf{a} \mid s)$ uma política com parâmetros θ . A atualização procura modificar θ de forma que ações associadas a vantagens positivas se tornem mais prováveis e ações associadas a vantagens negativas tenham sua probabilidade reduzida.

Essa abordagem é particularmente útil quando a ação é contínua, pois não exige enumerar todas as ações possíveis em uma tabela. Entretanto, estimativas de gradiente obtidas a partir de trajetórias amostradas podem apresentar variância elevada. Funções de valor, estimativas de vantagem e mecanismos de controle das atualizações são utilizados para tornar o processo de treinamento mais controlado (Sutton; Barto, 2018).

2.4.6 Arquiteturas Actor-Critic

Arquiteturas *actor-critic* combinam uma representação da política com uma função de valor. O *actor* representa a política parametrizada e produz a distribuição das ações. O *critic* estima o valor dos estados ou pares estado–ação e fornece uma referência utilizada durante o treinamento (Sutton; Barto, 2018; You *et al.*, 2019).

Em implementações baseadas em redes neurais, o estado s_t pode ser processado por uma representação compartilhada e encaminhado a ramos com objetivos distintos. O ramo do *actor* determina os parâmetros da política $\pi_\theta(\mathbf{a} \mid s)$. O ramo do *critic* estima $V_\phi(s)$, que representa o retorno esperado a partir do estado.

A diferença entre o retorno estimado e o valor previsto pelo *critic* é utilizada na construção da vantagem. Dessa forma, o *critic* não escolhe a ação. Sua função é fornecer uma referência para avaliar as decisões durante o treinamento. Na inferência, a ação é produzida a partir do *actor* e a estimativa do *critic* não precisa participar da seleção da ação.

A Figura 7 apresenta essa organização. O *actor* define a política, enquanto o *critic* estima o valor do estado. No PPO, as experiências coletadas são reutilizadas em atualizações controladas pelo mecanismo de *clipping*.

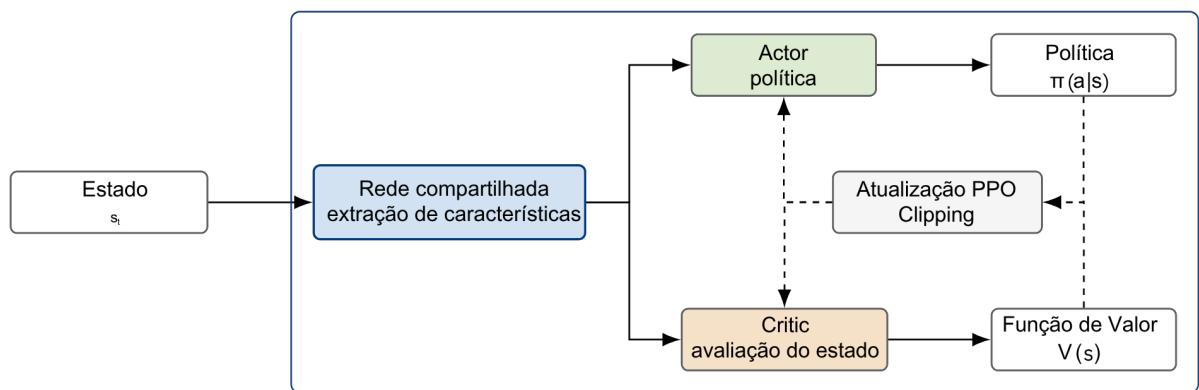


Figura 7: Estrutura *actor-critic* com atualização para PPO.

2.5 Proximal Policy Optimization

O Proximal Policy Optimization (PPO) é um método de gradiente de política que procura limitar o incentivo a mudanças excessivas entre a política utilizada para coletar as experiências e a política atualizada (Schulman *et al.*, 2017). O algoritmo é frequentemente empregado com arquitetura *actor-critic*, combinando uma função objetivo da política, uma função de valor e, quando configurado, um termo de entropia.

2.5.1 Objetivo com Clipping

Seja:

$$r_t(\theta) = \frac{\pi_\theta(\mathbf{a}_t \mid s_t)}{\pi_{\theta_{\text{old}}}(\mathbf{a}_t \mid s_t)} \quad (2.20)$$

a razão entre a probabilidade ou densidade atribuída à ação pela política atual e pela política utilizada na coleta. A função objetivo com *clipping* é (Schulman *et al.*, 2017):

$$L^{\text{CLIP}}(\theta) = \widehat{\mathbb{E}}_t \left[\min \left(r_t(\theta) \widehat{A}_t, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon) \widehat{A}_t \right) \right], \quad (2.21)$$

em que $\widehat{A}_t$ é a estimativa de vantagem e ϵ define o intervalo de *clipping*. O recorte não impede que os parâmetros da rede mudem além de um valor específico. Seu papel é limitar a contribuição, na função objetivo, de mudanças da razão de probabilidade que ultrapassem a faixa estabelecida. Por isso, o *clipping* deve ser interpretado como mecanismo de controle da atualização e não como garantia absoluta de distância entre políticas (Schulman *et al.*, 2017).

2.5.2 Generalized Advantage Estimation

O *Generalized Advantage Estimation* (GAE) combina erros de diferença temporal em diferentes horizontes para produzir uma estimativa de vantagem (Schulman *et al.*, 2015). O erro de diferença temporal é:

$$\delta_t = r_{t+1} + \gamma V(s_{t+1}) - V(s_t), \quad (2.22)$$

e a vantagem pode ser estimada por:

$$\hat{A}_t^{\text{GAE}(\gamma, \lambda)} = \sum_{l=0}^{\infty} (\gamma \lambda)^l \delta_{t+l}. \quad (2.23)$$

O parâmetro λ controla como erros de diferença temporal de horizontes distintos participam da estimativa. Valores menores dão maior peso a estimativas de horizonte curto e tendem a reduzir variância ao custo de maior viés. Valores próximos de 1 incorporam informação de horizontes mais longos e tendem a reduzir viés ao custo de maior variância (Schulman *et al.*, 2015). O efeito de λ deve ser interpretado em conjunto com γ , pois ambos aparecem no fator $(\gamma \lambda)^l$.

2.5.3 Função de Valor, Entropia e Otimização

Além do objetivo da política, implementações *actor-critic* do PPO treinam uma função de valor. Um termo de erro da função de valor mede a diferença entre o valor estimado e um alvo construído a partir das experiências. Seu coeficiente determina quanto esse erro contribui para a função de otimização conjunta (Schulman *et al.*, 2017).

Um bônus de entropia pode ser utilizado para favorecer políticas menos concentradas durante o treinamento. A entropia mede a dispersão da distribuição de ações. Quando seu coeficiente é positivo, o processo de otimização recebe incentivo para preservar maior diversidade de ações, contrapondo-se à redução prematura da exploração (Schulman *et al.*, 2017).

Os parâmetros das redes são atualizados por um otimizador. Nesta tese é utilizado o Adam, que mantém estimativas adaptativas dos momentos do gradiente para ajustar as atualizações dos parâmetros (Kingma; Ba, 2017). A taxa de aprendizado controla a escala base dessas atualizações.

2.5.4 Significado dos Parâmetros de Treinamento do PPO

Os valores concretos empregados pela tese são apresentados no Capítulo 4. Nesta seção, o objetivo é explicar o significado de cada parâmetro e evitar confundir parâmetros que atuam em partes diferentes do treinamento. A Tabela 1 resume essas funções com base na formulação do PPO, no GAE e no processo de otimização (Schulman *et al.*, 2017, 2015; Kingma; Ba, 2017).

A relação entre `n_steps`, `batch_size` e `n_epochs` merece atenção. Primeiro, a política interage com o ambiente e coleta `n_steps` experiências. Em seguida, essas experiências

Tabela 1: Significado dos principais parâmetros de treinamento do PPO utilizados nesta tese.

Parâmetro	Significado no treinamento
Taxa de aprendizado	Define a escala base das atualizações realizadas pelo otimizador nos parâmetros das redes. Valores maiores produzem passos de otimização maiores, enquanto valores menores produzem mudanças mais graduais.
<code>n_steps</code>	Quantidade de interações coletadas com o ambiente antes de iniciar uma etapa de atualização do PPO. Não corresponde ao total de passos de treinamento.
<code>batch_size</code>	Quantidade de amostras usada em cada <i>minibatch</i> durante a otimização. Os dados coletados por <code>n_steps</code> são divididos em lotes menores para os cálculos de gradiente.
<code>n_epochs</code>	Quantidade de passagens de otimização realizadas sobre o conjunto de experiências coletado antes de uma nova coleta. Aumentar esse valor reutiliza mais vezes os mesmos dados.
γ	Fator de desconto do retorno. Controla quanto recompensas futuras contribuem para o valor atribuído às decisões correntes.
λ_{GAE}	Parâmetro do GAE. Controla o compromisso entre viés e variância na estimativa da vantagem em conjunto com γ .
Normalização da vantagem	Reescala as vantagens do lote antes da atualização, tipicamente para média próxima de zero e desvio padrão unitário. Seu objetivo é controlar a escala numérica utilizada na otimização sem redefinir a função de recompensa.
Intervalo de <i>clipping</i>	Define ϵ na função objetivo do PPO e, portanto, a faixa $[1 - \epsilon, 1 + \epsilon]$ usada para limitar a contribuição de mudanças acentuadas na razão entre a política atual e a anterior.
Coefficiente de entropia	Define o peso do bônus de entropia. Valores positivos aumentam o incentivo à dispersão da política durante o treinamento.
Coefficiente da função de valor	Define o peso do erro do <i>critic</i> na função de otimização conjunta. Ele controla a importância relativa da aproximação de $V(s)$ durante o treinamento.
Norma máxima do gradiente	Estabelece um limite para a norma do vetor de gradientes antes da etapa do otimizador. Esse recorte atua sobre os gradientes e é diferente do <i>clipping</i> da razão de probabilidades da função objetivo do PPO.

são divididas em *minibatches* de tamanho `batch_size`. O conjunto é reutilizado por `n_epochs` passagens de otimização. Somente depois o processo retorna à coleta de novas experiências. Essa reutilização de dados é parte do procedimento de atualização do PPO (Schulman *et al.*, 2017).

A normalização da vantagem pode ser representada por:

$$\hat{A}_t^{\text{norm}} = \frac{\hat{A}_t - \mu_A}{\sigma_A + \varepsilon}, \quad (2.24)$$

em que μ_A e σ_A são, respectivamente, a média e o desvio padrão das vantagens no conjunto utilizado para a atualização, e ε evita divisão por zero. A operação modifica a escala numérica apresentada ao otimizador. Ela não altera os valores originais da recompensa armazenada na trajetória.

O recorte da norma do gradiente também não deve ser confundido com o *clipping* do PPO. Se $\mathbf{g}$ representa o vetor de gradientes e c a norma máxima configurada, uma representação do procedimento é:

$$\mathbf{g}' = \mathbf{g} \min \left(1, \frac{c}{\|\mathbf{g}\|_2} \right). \quad (2.25)$$

Quando a norma permanece abaixo de c , o gradiente não é alterado. Quando excede o limite, seu vetor é reescalado antes da atualização. O *clipping* de $r_t(\theta)$ atua na função objetivo da política, enquanto o limite de gradiente atua posteriormente sobre a magnitude do gradiente usado pelo otimizador.

O número total de passos de treinamento constitui outro parâmetro experimental importante, mas possui significado diferente de `n_steps`. O total de passos define o orçamento de interação destinado ao treinamento completo. Já `n_steps` define quantas interações são acumuladas entre atualizações sucessivas do PPO.

A semente pseudoaleatória também deve ser distinguida dos hiperparâmetros anteriores. A semente inicializa geradores de números pseudoaleatórios que afetam elementos estocásticos da execução, como inicialização de parâmetros e amostragem durante o treinamento. Fixar uma semente favorece reprodutibilidade de uma execução. Entretanto, uma única semente não permite estimar a variabilidade que poderia ocorrer entre treinamentos independentes com inicializações distintas.

2.5.5 Ações Contínuas e Transformações da Saída

O PPO pode parametrizar políticas em espaços de ações contínuos, o que é relevante quando a decisão não corresponde à escolha de uma única ação entre categorias discretas. Nesse caso, a rede produz valores reais que podem ser transformados antes de sua aplicação ao ambiente (Schulman *et al.*, 2017).

Uma função *softmax* pode converter um vetor de valores reais $\mathbf{z}$ em proporções não negativas cuja soma é igual a 1:

$$p_i = \frac{\exp(z_i/\tau)}{\sum_j \exp(z_j/\tau)}, \quad (2.26)$$

em que $\tau > 0$ é a temperatura. Os valores z_i podem ser interpretados como escores relativos antes da normalização. Eles não são probabilidades nem quantidades físicas de recursos por si mesmos. Valores menores de τ tornam as diferenças entre os escores mais acentuadas na distribuição resultante. Valores maiores produzem uma distribuição mais suave.

Uma transformação sigmoidal pode mapear um valor real para o intervalo aberto $(0,1)$:

$$\sigma(u) = \frac{1}{1 + e^{-u}}. \quad (2.27)$$

Essas transformações são operações matemáticas distintas do PPO. O algoritmo de aprendizado produz os componentes contínuos da ação, enquanto a formulação do problema define como esses componentes serão convertidos em variáveis com significado para o escalonamento. O Capítulo 4 apresenta a forma específica utilizada pelo IA-FMR.

2.6 Aprendizado por Reforço em Escalonamento de Recursos

O escalonamento pode ser formulado como um problema sequencial no qual o agente observa informações da rede e seleciona uma distribuição de recursos. Estados podem incluir indicadores de canal, dados das filas, histórico de atendimento e disponibilidade de recursos. Ações podem representar seleção de usuários, quantidades de recursos ou parâmetros que controlam a política de distribuição (Seguin *et al.*, 2023; Mamane *et al.*,

2022).

O uso de DRL permite aproximar políticas e funções de valor em problemas com relações não lineares e espaços de maior dimensão. Essa flexibilidade tem motivado aplicações em escalonamento e alocação de recursos. Entretanto, ela também introduz questões relacionadas à estabilidade do treinamento, custo computacional, interpretação das decisões e generalização para condições diferentes das utilizadas durante o aprendizado (Seguin *et al.*, 2023; Mamane *et al.*, 2022).

A recompensa condiciona o comportamento aprendido e precisa ser interpretada de acordo com sua formulação. Um termo associado à justiça atua sobre a grandeza utilizada para medir justiça. Um termo associado à eficiência atua sobre a definição de eficiência adotada. Uma penalização por déficit influencia o treinamento, mas não equivale automaticamente a uma garantia de serviço. Essas distinções são fundamentais para interpretar a proposta do IA-FMR.

A formulação desta tese inclui na ação um componente associado ao coeficiente de ponderação entre justiça e eficiência, além dos valores utilizados para formar a distribuição alvo de recursos. A política produz esses componentes conjuntamente. A possibilidade de variar o coeficiente entre decisões pertence à formulação do espaço de ações. O comportamento efetivamente aprendido pela política é uma questão empírica examinada nos resultados (Lopes *et al.*, 2026).

2.7 Considerações Finais do Capítulo

Este capítulo apresentou a organização da interface rádio 5G NR, a relação entre numerologia, recursos, MCS e eficiência espectral e o papel do escalonamento na camada MAC. Foram também definidos os princípios dos escalonadores RR, MR e PF e as métricas utilizadas para interpretar os experimentos, distinguindo desempenho agregado, região inferior da distribuição, justiça dos RBGs, atendimento, demanda pendente, *backlog*, perdas, atraso e cumprimento de prazos no cenário sintético de vídeo.

A discussão diferenciou métricas observadas, requisitos de serviço e sinais utilizados durante o treinamento. Essa separação evita interpretar um limiar da recompensa como garantia de QoS ou um valor elevado de Jain como evidência suficiente de proteção de vazão. Também foi esclarecido que a avaliação temporal de vídeo utiliza métricas objetivas de entrega e não constitui uma caracterização completa de QoE.

Por fim, foram apresentados os fundamentos de RL, MDPs, funções de valor, vantagem, gradiente de política, arquiteturas *actor-critic* e PPO. Os principais parâmetros de treinamento foram descritos de acordo com a função que desempenham na coleta de experiências, na estimação da vantagem, na atualização da política, no treinamento do *critic* e na otimização numérica. Esses conceitos fornecem a base teórica para a análise dos trabalhos relacionados e para a formulação do IA-FMR apresentada no Capítulo 4.

3 Trabalhos Relacionados

Este capítulo apresenta uma revisão temática estruturada e não exaustiva da literatura sobre escalonamento de recursos em redes móveis. A discussão reúne abordagens representativas para examinar cinco aspectos: critérios tradicionais de decisão, políticas baseadas em aprendizado por reforço, tratamento da justiça, formas de representar ou ajustar o compromisso entre justiça e eficiência e ambientes de validação relacionados às redes 5G. O objetivo é construir uma comparação temática e posicionar o *In-Action Fair Max Rate* (IA-FMR) em relação aos trabalhos analisados, sem antecipar a descrição detalhada do mecanismo apresentada no Capítulo 4.

Inicialmente, são descritos o procedimento de busca e a organização da literatura. Em seguida, discutem-se os escalonadores tradicionais empregados como referências, as propostas baseadas em aprendizado por reforço e os mecanismos que incorporam justiça à formulação. A seção posterior distingue critérios fixos, pesos definidos na recompensa, parâmetros fornecidos como entrada, regras externas de adaptação e parâmetros incorporados à ação. Por fim, o IA-FMR é posicionado diante dessas alternativas.

A Figura 8 sintetiza a perspectiva de análise adotada. Os escalonadores tradicionais utilizam critérios definidos previamente. As abordagens baseadas em aprendizado por reforço produzem decisões condicionadas ao estado observado, segundo a formulação de estado, ação e recompensa adotada no treinamento. A justiça pode aparecer na recompensa, em restrições, em métricas de prioridade ou na própria representação da decisão. As formas de ajuste do compromisso entre objetivos incluem pesos fixos, informações fornecidas como entrada, regras externas e parâmetros incorporados à ação.

Essa organização é qualitativa. Ela não constitui uma taxonomia universal, não estabelece uma sequência obrigatória de evolução e não implica superioridade de uma classe sobre outra. O IA-FMR é incluído nesse panorama porque produz conjuntamente os componentes usados na distribuição alvo dos recursos e o coeficiente de ponderação, associado à justiça e eficiência durante o treinamento.

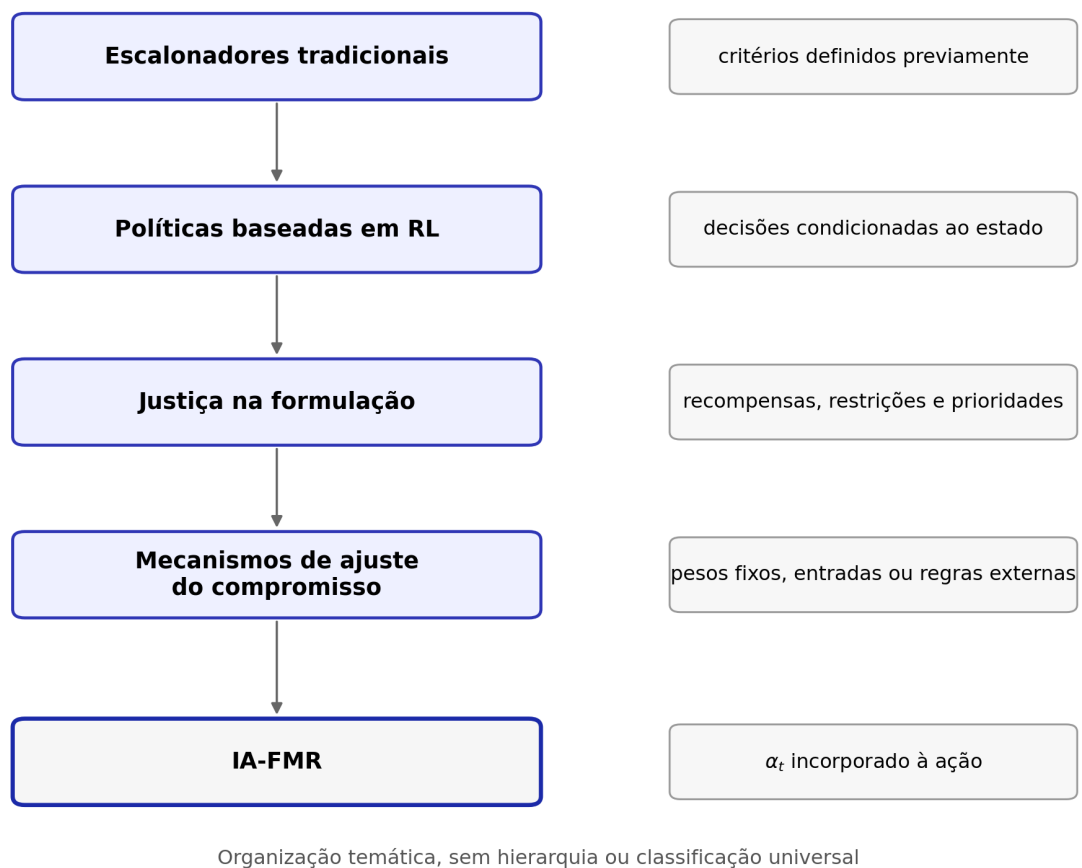


Figura 8: Síntese qualitativa das abordagens de escalonamento consideradas na organização do capítulo.

3.1 Procedimento de Seleção e Organização da Literatura

A literatura foi organizada por meio de uma revisão temática estruturada, destinada a identificar abordagens representativas e construir a comparação apresentada neste capítulo. A revisão não foi conduzida como levantamento sistemático, não pretende cobrir exaustivamente o campo e não oferece a reprodutibilidade quantitativa de um protocolo sistemático.

Os estudos foram localizados em consultas ao Google Scholar, IEEE Xplore, ScienceDirect, ACM Digital Library e SpringerLink. A busca foi complementada pela análise das referências e citações dos trabalhos identificados e pelo uso das ferramentas Consensus¹ e ResearchRabbit² para localizar estudos relacionados.

¹Ferramenta de busca acadêmica assistida por inteligência artificial. Disponível em: <https://consensus.app/>. Acesso em: 12 jun. 2026.

²Ferramenta de descoberta de literatura e exploração de redes de citações. Disponível em: <https://www.researchrabbit.ai>. Acesso em: 12 jun. 2026.

As consultas combinaram termos associados a escalonamento de recursos, aprendizado por reforço e justiça em redes móveis, incluindo “5G scheduling”, “resource allocation”, “reinforcement learning”, “deep reinforcement learning”, “fairness-aware scheduling”, “dynamic fairness”, “PPO”, “O-RAN” e “network slicing”. Os termos foram refinados durante o desenvolvimento da pesquisa conforme novos conceitos, autores e referências eram identificados.

Como a busca foi construída progressivamente, não foram definidos um período rígido de publicação, um *corpus* fechado ou um fluxo quantitativo de identificação e exclusão. Também não foram registrados totais de resultados recuperados, duplicatas ou estudos excluídos. Esses elementos não são apresentados porque não fizeram parte do procedimento adotado.

A pertinência temática foi estabelecida pela relação dos estudos com pelo menos um dos seguintes aspectos: aplicação de aprendizado por reforço ao escalonamento ou à alocação de recursos; tratamento explícito da justiça; combinação de justiça e eficiência; ajuste de políticas, prioridades ou parâmetros; atuação na camada MAC de redes 5G; ou validação em ambientes relacionados a O-RAN, ns-3 e 5G-LENA.

Não foram incluídos na comparação direta trabalhos dedicados exclusivamente a roteamento, posicionamento de funções de rede, gerenciamento isolado de recursos computacionais ou problemas sem decisão de distribuição de recursos de comunicação entre usuários ou serviços. Estudos de revisão e trabalhos conceituais foram utilizados para organizar o campo e fundamentar a análise, mas não foram tratados como propostas diretamente comparáveis ao IA-FMR.

Os trabalhos discutidos neste capítulo foram examinados quanto à técnica de aprendizado, ao objeto da ação, aos objetivos considerados, ao tratamento da justiça, à forma de representar o compromisso entre objetivos, às métricas e ao ambiente de validação. A Tabela 2 resume esse procedimento.

O quarto eixo distingue formas de representar ou ajustar o compromisso entre objetivos, como pesos fixados na recompensa, parâmetros fornecidos à política, regras externas de adaptação e parâmetros incorporados à ação. O quinto reúne ambientes e arquiteturas empregados na validação, como 5G NR, O-RAN e ns-3/5G-LENA.

A Figura 9 apresenta a organização temática usada para relacionar os estudos. O primeiro eixo reúne os escalonadores tradicionais RR, MR e PF. O segundo contempla abordagens baseadas em aprendizado por reforço, nas quais o problema é definido por

Tabela 2: Procedimento de busca e organização temática da literatura

Elemento	Descrição
Natureza da revisão	Revisão temática estruturada e não exaustiva, sem protocolo de revisão sistemática.
Fontes de busca	Google Scholar, IEEE Xplore, ScienceDirect, ACM Digital Library e SpringerLink.
Termos de busca	Combinações relacionadas a 5G, escalonamento, alocação de recursos, aprendizado por reforço, justiça, PPO, O-RAN e <i>network slicing</i> .
Expansão da busca	Consensus, ResearchRabbit, referências cruzadas e análise das citações dos estudos identificados.
Crítérios de pertinência	Estudos sobre distribuição de recursos de comunicação que abordam aprendizado por reforço, justiça, múltiplos objetivos, ajuste de políticas ou parâmetros, camada MAC 5G, O-RAN, ns-3 ou 5G-LENA.
Não inclusão na comparação	Trabalhos dedicados exclusivamente a roteamento, posicionamento de funções, gerenciamento isolado de recursos computacionais ou problemas sem distribuição de recursos de comunicação entre usuários ou serviços.
Dimensões de análise	Técnica de aprendizado, objeto da ação, objetivos, tratamento da justiça, representação do compromisso, métricas, ambiente de validação e relação com o IA-FMR.
Organização temática	Escalonadores tradicionais, aprendizado por reforço, justiça no escalonamento, representação do compromisso e validação em redes 5G.

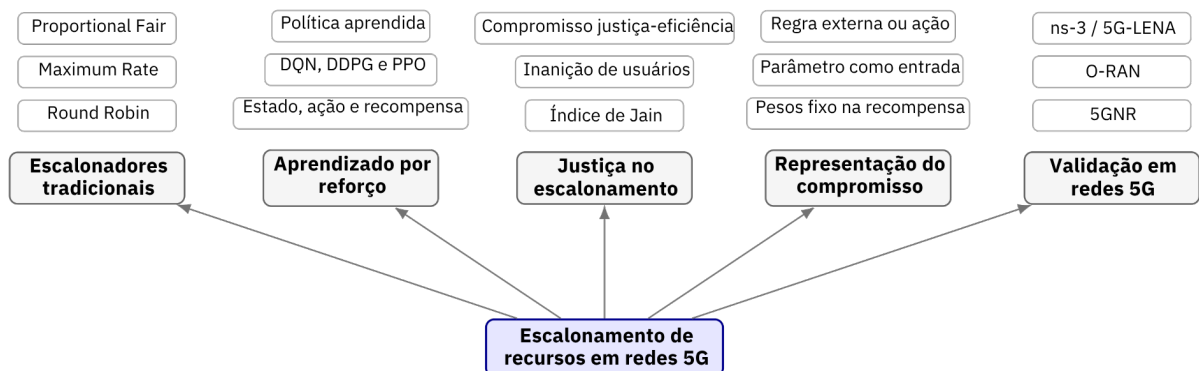


Figura 9: Organização temática dos trabalhos relacionados ao escalonamento de recursos em redes 5G.

estados, ações, recompensas e políticas aprendidas. O terceiro agrupa trabalhos que tratam justiça, inanição e o compromisso entre justiça e eficiência.

A figura funciona como um mapa de leitura do capítulo. Uma mesma proposta pode pertencer a mais de um eixo, e a posição em uma categoria não representa avaliação quantitativa. A Tabela 3, apresentada na Seção 3.6, sintetiza os trabalhos mais próximos do IA-FMR entre as abordagens discutidas, sem constituir uma relação exaustiva da literatura consultada.

3.2 Escalonadores Tradicionais para Redes Móveis

Os escalonadores RR, MR e PF são empregados como referências porque representam critérios distintos de priorização (Mamane *et al.*, 2022). Os fundamentos desses mecanismos foram apresentados no Capítulo 2. Nesta seção, a discussão se limita às características necessárias para compará-los com políticas aprendidas.

O RR distribui oportunidades de atendimento de forma cíclica entre os usuários elegíveis. Como a condição instantânea de canal não constitui seu critério principal, a política tende a regularizar o acesso aos recursos, mas pode utilizar parte da capacidade em enlaces com menor eficiência espectral. O MR prioriza os usuários com maiores taxas instantâneas estimadas. Essa escolha favorece a vazão agregada e pode concentrar a alocação nos UEs com melhores condições de propagação (Mamane *et al.*, 2022).

O PF utiliza a relação entre a taxa instantânea estimada e uma média histórica de vazão ou atendimento. A atualização desse histórico modifica as prioridades ao longo da execução, embora a regra de decisão permaneça definida previamente. O comportamento obtido depende das condições de canal, das filas, da dinâmica do tráfego e da forma de atualização do histórico (Seguin *et al.*, 2023; Mamane *et al.*, 2022).

Essas políticas não são inteiramente estáticas, pois suas decisões respondem às informações consideradas em cada intervalo. A distinção relevante é que o critério de priorização não é aprendido e a preferência entre regularidade, oportunismo e equilíbrio proporcional permanece vinculada à regra estabelecida. Essa característica motiva abordagens que aprendem decisões de alocação a partir da interação com o ambiente.

3.3 Escalonadores Baseados em Aprendizado por Reforço

Em redes 5G, a variação do canal, das filas, do tráfego e dos requisitos de serviço dificulta a definição de uma única regra de escalonamento para todas as condições. O aprendizado por reforço permite formular o problema como decisão sequencial e obter uma política a partir das consequências acumuladas das ações (Seguin *et al.*, 2023; Mamane *et al.*, 2022).

O agente observa informações do sistema, como qualidade do canal, ocupação das filas, histórico de atendimento e requisitos de QoS, e produz uma ação relacionada à distribuição dos recursos. A recompensa associa os efeitos dessa decisão aos objetivos definidos para o treinamento (Sutton; Barto, 2018; Seguin *et al.*, 2023). O objeto da ação varia entre os trabalhos e pode corresponder à seleção de usuários, à distribuição direta de recursos, ao controle de potência, à divisão de capacidade entre serviços ou à reconfiguração de parâmetros de outro escalonador.

Métodos de aprendizado por reforço profundo aproximam políticas e funções de valor em espaços de maior dimensão (Seguin *et al.*, 2023). Entre as técnicas utilizadas nos estudos analisados estão *Deep Q-Network* (DQN), *Double Deep Q-Network* (DDQN), *Deep Deterministic Policy Gradient* (DDPG), PPO e arquiteturas multiagente.

Al-Tam, Correia e Rodriguez apresentam o LEASCH, baseado em DDQN, para decisões de escalonamento em redes 5G. O mecanismo considera condições de canal e requisitos de QoS, e os resultados reportados pelos autores mostram ganhos de vazão e atraso nas condições avaliadas (Al-Tam; Correia; Rodriguez, 2020). Abbas propõe um escalonador baseado em aprendizado por reforço para reduzir o atraso no enlace de subida, utilizando informações de filas e tráfego em situações de congestionamento (Abbas, 2025).

Gu *et al.* utilizam DDPG com assistência de conhecimento prévio sobre o ambiente para tratar restrições temporais associadas ao escalonamento em redes 5G (Gu *et al.*, 2021). A proposta ilustra como informações do domínio podem orientar o processo de aprendizagem e reduzir a dependência de exploração não orientada.

Outros estudos abordam problemas correlatos. Budhiraja, Kumar e Tyagi desenvolvem um mecanismo proporcional baseado em aprendizado profundo para comunicação direta entre dispositivos (*Device-to-Device*, D2D) (Budhiraja; Kumar; Tyagi, 2020). Vishnoi *et al.* combinam alocação de recursos e controle de potência entre pares D2D, considerando soma das taxas e justiça (Vishnoi *et al.*, 2023). Embora esses cenários sejam distintos do enlace de descida investigado nesta tese, eles mostram formas de incorporar objetivos concorrentes à formulação.

O PPO também é empregado na distribuição de capacidade entre serviços ou fatias. Shezi *et al.* propõem uma abordagem para O-RAN que considera ocupação de recursos e atendimento de acordos de nível de serviço (*Service Level Agreement*, SLA) (Shezi *et al.*, 2025). Li utiliza PPO para distribuir recursos entre serviços a partir do volume de tráfego, da utilização das fatias e de requisitos de QoS (Li, 2026). Nesses casos, o agente atua em um nível de gerenciamento diferente da distribuição direta de RBGs entre UEs.

Paz-Pérez *et al.* apresentam um escalonador integrado ao ns-3/5G-LENA e projetado para operar com variações no número de usuários (Paz-Pérez *et al.*, 2024). Essa flexibilidade é relevante porque políticas neurais podem depender de estados e ações com dimensão previamente definida. Villegas *et al.* utilizam aprendizado por reforço profundo para reconfigurar parâmetros de um escalonador MAC em um ambiente O-RAN integrado ao ns-3/5G-LENA (Villegas *et al.*, 2026). Nesse caso, a ação modifica a configuração de outro mecanismo, em vez de produzir diretamente a parcela de recursos destinada a cada usuário.

Abordagens multiagente estendem o problema para decisões distribuídas. Kim e Lim investigam o gerenciamento de recursos em fatiamento fim a fim por meio de agentes cooperativos (Kim; Lim, 2021). Cui *et al.* revisam o emprego de aprendizado por reforço profundo multiagente em alocação de recursos e fatiamento de rede para cenários B5G e 6G (Cui *et al.*, 2026). Essas arquiteturas permitem decisões descentralizadas ou hierárquicas, mas acrescentam requisitos de coordenação e observabilidade.

Em conjunto, os estudos mostram que políticas aprendidas podem produzir ações diferentes diante de estados distintos. Essa capacidade não equivale, por si só, à alteração da preferência relativa entre os objetivos do treinamento. Em muitas formulações, a relação entre vazão, atraso, QoS e justiça permanece estabelecida pelos termos, pesos e restrições da recompensa (Seguin *et al.*, 2023; Patil *et al.*, 2026). A política pode variar suas alocações sem modificar a forma pela qual esse compromisso foi definido.

3.4 Justiça no Escalonamento de Recursos

A justiça pode ser analisada sobre recursos, oportunidades de atendimento, vazões ou níveis de serviço. A interpretação depende da variável considerada. Uma distribuição uniforme de recursos não produz necessariamente vazões semelhantes quando os UEs apresentam condições de canal distintas. Da mesma forma, vazões próximas podem exigir quantidades diferentes de recursos (Mamane *et al.*, 2022; Fossati *et al.*, 2020).

A priorização de usuários com melhores condições de canal pode elevar a vazão agregada e concentrar o atendimento. Em sentido oposto, uma distribuição uniforme pode reduzir o aproveitamento das diferenças de eficiência espectral. A literatura trata essa relação por funções de utilidade, restrições, métricas de justiça e formulações multiobjetivo (Saraiva *et al.*, 2020; Fossati *et al.*, 2020).

Fossati *et al.* analisam esse compromisso na alocação de múltiplos recursos para fatiamento de rede. O trabalho emprega o conceito de *Price of Fairness* e operadores de média ponderada ordenada (*Ordered Weighted Average*, OWA) para representar soluções intermediárias entre critérios utilitaristas e igualitários (Fossati *et al.*, 2020). Embora o problema seja diferente do escalonamento MAC desta tese, a análise evidencia que o aumento da justiça pode envolver perda de eficiência e que diferentes parametrizações produzem pontos de operação distintos.

No PF, a justiça é considerada indiretamente pela relação entre taxa instantânea e histórico de atendimento. Essa regra reduz a prioridade de usuários recentemente favorecidos, mas não estabelece, por si só, quantidade mínima de recursos, taxa mínima ou distribuição uniforme (Seguin *et al.*, 2023; Mamane *et al.*, 2022).

Nas abordagens baseadas em aprendizado por reforço, a justiça pode integrar a recompensa, as restrições, as métricas de prioridade ou a representação da política. Budhiraja, Kumar e Tyagi tratam a proporcionalidade na comunicação D2D (Budhiraja; Kumar; Tyagi, 2020). Vishnoi *et al.* combinam soma das taxas e justiça entre pares D2D (Vishnoi *et al.*, 2023). Baheti e Khunteta investigam escalonamento orientado à justiça para aplicações mMTC em redes 5G NR (Baheti; Khunteta, 2026). E Patil *et al.* analisam abordagens destinadas a conciliar eficiência e equidade (Patil *et al.*, 2026).

A inclusão de justiça na formulação não implica garantia de um valor específico em todas as condições. O resultado depende da métrica escolhida, dos pesos, das restrições, do estado, da ação e das condições de treinamento e avaliação (Seguin *et al.*, 2023; Patil *et al.*, 2026). Também é necessário combinar indicadores agregados, individuais e temporais. Um índice global pode revelar uniformidade, mas não identificar quais usuários receberam pouco atendimento, por quanto tempo essa condição persistiu ou qual custo de eficiência acompanhou a distribuição observada. Métricas como P5, número de UEs atendidos e demanda pendente sem alocação complementam essa análise (Mamane *et al.*, 2022; Baheti; Khunteta, 2026; Patil *et al.*, 2026).

3.5 Representação do Compromisso entre Justiça e Eficiência

As abordagens analisadas diferem não apenas quanto às ações produzidas, mas também quanto à forma de definir a preferência entre objetivos. Essa distinção evita tratar toda decisão condicionada ao estado como alteração do compromisso entre justiça e eficiência.

Em políticas tradicionais, o compromisso decorre de um critério fixo de priorização. RR, MR e PF podem produzir decisões diferentes a cada intervalo, mas a regra que orienta essas decisões permanece estabelecida previamente. Em formulações de aprendizado por reforço, uma alternativa recorrente é combinar objetivos por meio de pesos fixados na recompensa. A política resultante responde ao estado, porém aprende sob a preferência representada por esses pesos (Seguin *et al.*, 2023; Patil *et al.*, 2026).

Outra possibilidade é fornecer requisitos, prioridades ou parâmetros como parte da entrada da política. Essa estratégia pode condicionar a decisão ao valor informado, desde que a política tenha sido treinada para interpretar essa informação. O valor continua sendo definido externamente ao agente. Trabalhos que consideram estado de fatias, requisitos de QoS ou atendimento de SLA ilustram políticas condicionadas por informações operacionais, embora não empreguem necessariamente uma entrada específica para o compromisso justiça-eficiência (Shezi *et al.*, 2025; Li, 2026).

O ajuste também pode ocorrer por uma regra ou controlador externo. Raja *et al.* combinam classificação de tráfego e PPO para adaptar a distribuição de recursos em O-RAN (Raja *et al.*, 2026). Villegas *et al.* empregam o agente para reconfigurar parâmetros de um escalonador MAC (Villegas *et al.*, 2026). Nesses casos, a ação aprendida modifica prioridades ou configurações aplicadas por outro mecanismo. A adaptação é operacional, mas não corresponde necessariamente a uma variável definida para representar diretamente a ponderação entre justiça e eficiência.

Formulações multiobjetivo em áreas correlatas reforçam essa distinção. Zhang e Ou combinam custo computacional, consumo energético e QoS em escalonamento de nuvem e borda (Zhang; Ou, 2025). Yu *et al.* consideram energia e custo no escalonamento de tarefas (Yu *et al.*, 2025), enquanto Sravan e Shaik combinam latência, energia e atendimento de SLA em recursos de borda e nuvem (Sravan; Shaik, 2026). Esses trabalhos mostram que múltiplos objetivos podem ser agregados ou ajustados sem que o parâmetro de preferência integre a ação.

Entre as abordagens representativas analisadas, a forma mais comum de estabelecer

o compromisso permanece associada a critérios fixos, à recompensa, a requisitos fornecidos externamente ou à reconfiguração de outro mecanismo. A diferença de formulação investigada nesta tese não está na ausência de adaptação na literatura. Ela está na incorporação, ao espaço de ações, de um parâmetro associado à ponderação entre justiça e eficiência, produzido conjuntamente com os componentes usados para definir a alocação alvo.

3.6 Posicionamento da Proposta

Lopes *et al.* propõem o *Fair Max Rate* (FMR), um escalonador baseado em aprendizado por reforço voltado ao aproveitamento das diferenças de eficiência espectral sem deixar de considerar o atendimento dos UEs em condições menos favoráveis (Lopes *et al.*, 2025). Na formulação inicial, a recompensa considera a vazão agregada em relação a uma referência de eficiência e uma penalização associada a vazões individuais abaixo de um limiar estabelecido. A justiça é tratada de forma indireta por essa proteção contra atendimento insuficiente, enquanto o Índice de Jain é utilizado na avaliação da distribuição produzida pelo escalonador.

O IA-FMR representa uma etapa posterior dessa linha de desenvolvimento ao incorporar ao espaço de ações um coeficiente associado à ponderação entre justiça e eficiência (Lopes *et al.*, 2026). A política produz conjuntamente os valores utilizados para determinar a distribuição alvo dos recursos e esse coeficiente. Durante o treinamento, o coeficiente participa da ponderação dos termos de justiça e eficiência da recompensa. Durante a inferência, a alocação alvo corrente decorre dos valores produzidos para a distribuição, enquanto o valor do coeficiente produzido na decisão anterior integra o estado seguinte.

Essa formulação representa explicitamente o compromisso no espaço de ações, mas não implica que o coeficiente controle isoladamente a alocação corrente ou apresente ampla variação temporal. O comportamento observado deve ser atribuído à política aprendida em seu conjunto. A região ocupada pelo coeficiente e sua relação com as alocações são examinadas no Capítulo 6, com as limitações correspondentes.

A implementação utiliza treinamento *offline* em Python e inferência integrada ao escalonador do ns-3/5G-LENA por meio do ns3-ai. Os espaços de estados e ações, a recompensa e a aplicação da decisão na camada MAC são detalhados no Capítulo 4.

A Figura 10 utiliza dois eixos conceituais. A “Explicitação do compromisso” indica em que medida a relação entre justiça e eficiência aparece diretamente na formulação



Figura 10: Mapeamento conceitual das abordagens analisadas segundo a explicitação do compromisso e a flexibilidade operacional.

da decisão. A “Flexibilidade operacional” representa a capacidade de produzir decisões condicionadas ao estado observado ou de alterar parâmetros durante a operação.

As posições decorrem de uma leitura qualitativa das formulações descritas nos trabalhos, e não de uma medição comum de desempenho. Critérios fixos podem explicitar uma preferência sem permitir sua alteração, enquanto políticas aprendidas podem apresentar elevada flexibilidade operacional com a preferência ainda definida na recompensa. O IA-FMR ocupa a região em que o parâmetro associado ao compromisso integra a ação juntamente com os valores utilizados para determinar a distribuição dos recursos. Essa posição caracteriza uma diferença de formulação, não superioridade universal ou demonstração de seleção ampla de pontos de operação.

A Tabela 3 sintetiza características dos trabalhos mais próximos da proposta. A coluna “RL” indica o emprego de aprendizado por reforço. A coluna “PPO”, o uso desse algoritmo ou de formulação diretamente associada. Na coluna “Justiça”, a presença explícita ou indireta de equidade na recompensa, nas restrições, nas prioridades ou na avalia-

ção. A coluna “Ação J/E”, a inclusão de variável associada ao compromisso entre justiça e eficiência no espaço de ações. A coluna “MAC 5G”, representa a atuação direta sobre decisões de escalonamento ou alocação na camada MAC. Por fim, a coluna “ns-3/O-RAN”, o uso de um desses ambientes. O símbolo ✓ indica presença identificada diretamente, — indica presença parcial ou indireta e ✗ indica que a característica não foi identificada na formulação descrita.

Tabela 3: Resumo comparativo dos principais trabalhos relacionados.

Trabalho	RL	PPO	Justiça	Ação J/E	MAC 5G	ns-3/O-RAN
(Al-Tam; Correia; Rodriguez, 2020)	✓	✗	✓	✗	✓	✗
(Gu <i>et al.</i> , 2021)	✓	✗	✓	✗	✓	✗
(Shezi <i>et al.</i> , 2025)	✓	✓	—	✗	✗	✓
(Li, 2026)	✓	✓	✓	✗	✗	✗
(Raja <i>et al.</i> , 2026)	✓	✓	—	✗	✗	✓
(Villegas <i>et al.</i> , 2026)	✓	—	✓	—	✓	✓
(Abbas, 2025)	✓	✗	✓	✗	✓	✗
(Budhiraja; Kumar; Tyagi, 2020)	✓	✗	✓	✗	✗	✗
(Baheti; Khunteta, 2026)	✗	✗	✓	✗	✓	✗
(Paz-Pérez <i>et al.</i> , 2024)	✓	✓	✓	✗	✓	✓
(Lopes <i>et al.</i> , 2026) (IA-FMR)	✓	✓	✓	✓	✓	✓

A tabela mostra que aprendizado por reforço e justiça aparecem em diferentes combinações. Shezi *et al.* (Shezi *et al.*, 2025), Li (Li, 2026), Raja *et al.* (Raja *et al.*, 2026), Paz-Pérez *et al.* (Paz-Pérez *et al.*, 2024) e o IA-FMR (Lopes *et al.*, 2026) empregam PPO. A marca parcial atribuída a Villegas *et al.* (Villegas *et al.*, 2026) registra proximidade com essa classe de métodos sem equiparar sua formulação às propostas marcadas integralmente.

A coluna “Justiça” não significa que os trabalhos usem a mesma métrica, ofereçam a mesma garantia ou atuem sobre a mesma variável. Em propostas voltadas a fatias e serviços, a presença pode ser indireta e associada ao atendimento de requisitos, o que justifica as marcações parciais.

Na coluna “Ação J/E”, o IA-FMR é o único trabalho marcado integralmente entre os estudos comparados. Villegas *et al.* (Villegas *et al.*, 2026) recebem marca parcial porque o agente reconfigura parâmetros que alteram o comportamento multiobjetivo de outro

escalonador, mas não utiliza uma variável definida especificamente como ponderação entre justiça e eficiência. Nos demais trabalhos, as decisões variam segundo o estado, enquanto a preferência entre objetivos permanece associada à recompensa, às restrições, às entradas ou à configuração estabelecida.

As últimas colunas distinguem o nível de atuação e o ambiente de validação. Parte das propostas atua diretamente sobre o escalonamento MAC, enquanto outras distribuem recursos entre fatias, serviços ou entidades em outro nível. O uso de ns-3 ou O-RAN também não é comum a todos os estudos. O IA-FMR reúne atuação no escalonamento MAC em 5G, integração com ns-3/5G-LENA e incorporação de um coeficiente de ponderação ao espaço de ações. A tabela descreve diferenças de formulação e implementação, não uma comparação experimental de desempenho entre cenários distintos.

3.7 Considerações Finais do Capítulo

Este capítulo organizou a literatura em torno de critérios tradicionais, políticas baseadas em aprendizado por reforço, tratamento da justiça e formas de representar o compromisso entre justiça e eficiência. As abordagens analisadas incluem métodos adaptativos, condicionados ao estado, multiobjetivo e orientados por QoS. Portanto, a lacuna não pode ser formulada como ausência geral de adaptação ou de justiça na literatura.

Entre os estudos usados na comparação temática, a preferência entre justiça e eficiência é usualmente definida por um critério fixo, por pesos da recompensa, por informações fornecidas externamente ou por regras que reconfiguram outro mecanismo. Não foi identificada, nesse conjunto, outra proposta que incorporasse à ação uma variável especificamente associada a essa ponderação e a produzisse conjuntamente com os componentes da alocação alvo.

O IA-FMR investiga essa diferença de formulação por meio de uma ação composta pelos valores utilizados para determinar a distribuição dos recursos e pelo coeficiente de ponderação entre justiça e eficiência. Durante o treinamento, esse coeficiente participa da ponderação da recompensa. Durante a inferência, a alocação alvo corrente decorre dos valores produzidos para a distribuição, enquanto o valor do coeficiente produzido na decisão anterior integra o estado seguinte. O Capítulo 4 apresenta a formulação completa, e os Capítulos 5 e 6 descrevem a avaliação e os resultados sem atribuir efeitos isolados a um único componente da política.

4 O Mecanismo IA-FMR

Este capítulo apresenta o *In-Action Fair Max Rate* (IA-FMR), mecanismo de escalonamento baseado em aprendizado por reforço para o *downlink* de redes 5G NR. A proposta resulta da evolução do *Fair Max Rate* (FMR) (Lopes *et al.*, 2025) e incorpora ao espaço de ações um coeficiente de ponderação entre justiça e eficiência. São apresentadas a evolução da proposta, sua formulação, o treinamento da política e a integração com o ns-3/5G-LENA.

4.1 Evolução do FMR para o IA-FMR

O IA-FMR deriva do *Fair Max Rate* (FMR), apresentado em trabalho anterior (Lopes *et al.*, 2025). O FMR introduziu uma política de aprendizado por reforço para produzir uma distribuição de recursos a partir das condições dos UEs. Naquela formulação, a recompensa privilegiava a vazão agregada em relação a uma referência de eficiência e incluía uma penalização destinada a desencorajar vazões individuais muito baixas. A justiça aparecia, portanto, de forma indireta, por meio dessa penalização, e não como um termo explícito ponderado na recompensa.

A evolução para o IA-FMR introduz explicitamente dois componentes na recompensa: um termo de justiça na distribuição dos recursos e um termo de eficiência da decisão. A importância relativa desses componentes é representada por um coeficiente de ponderação α . Quando α aumenta, o termo de justiça recebe maior peso e o termo de eficiência recebe, de forma complementar, o peso $1 - \alpha$. O coeficiente não mede justiça, eficiência ou qualidade de serviço. Ele apenas define como esses dois termos participam da recompensa utilizada durante o treinamento.

Durante o desenvolvimento da proposta, foram consideradas quatro formas de fornecer esse coeficiente. A diferença entre elas está em quem determina α e em que ponto ele entra no processo de decisão. A Tabela 4 resume essas alternativas.

Tabela 4: Formas de incorporação do coeficiente de ponderação consideradas na evolução do FMR.

Modo	Origem do coeficiente	Papel na evolução
Fixo (<i>fixed</i>)	Valor definido externamente	Referência fixa
Como entrada (<i>input</i>)	Valor externo incluído no estado	Alternativa exploratória
Adaptativo (<i>adaptive</i>)	Regra externa baseada em métricas	Alternativa exploratória
Na ação (<i>in-action</i>)	Componente produzido pela política	Formulação IA-FMR

No modo fixo, α é um parâmetro externo e não se altera durante a execução. No modo como entrada, o valor também é definido externamente, mas é apresentado à política como parte do estado. No modo adaptativo, uma regra programada fora da política altera o coeficiente a partir das métricas observadas. Essas três alternativas mantêm, em diferentes graus, a definição de α fora da ação aprendida.

Na formulação *in-action*, adotada no IA-FMR, a própria política produz o componente utilizado para obter α_t juntamente com os valores que determinam a distribuição dos recursos (Lopes *et al.*, 2026). Ao longo deste capítulo, t indexa os instantes sucessivos de decisão. O símbolo α_t representa o coeficiente produzido na decisão t .

É importante distinguir o papel de α_t no treinamento e na inferência. Durante o treinamento, ele pondera os termos de justiça e eficiência da recompensa da decisão corrente. Durante a inferência no ns-3/5G-LENA, a recompensa não é calculada e nenhum aprendizado é realizado. Nessa etapa, α_t é preservado como parte do estado fornecido à decisão seguinte. Ele não é aplicado posteriormente para corrigir ou redimensionar a alocação que já foi produzida.

A formulação permite, portanto, que um novo valor de α_t seja produzido a cada decisão. Essa possibilidade não implica que uma política treinada necessariamente explore uma faixa ampla de valores. O comportamento efetivamente aprendido é uma propriedade empírica da política final e é examinado no Capítulo 6.

Para evitar ambiguidade terminológica, neste capítulo o termo *agente* é utilizado para a entidade que interage com o ambiente durante o treinamento, a *política* corresponde à função parametrizada aprendida pelo PPO e o *IA-FMR* corresponde ao mecanismo completo de escalonamento, incluindo a política, a conversão da ação em recursos e sua integração à camada MAC.

4.2 Modelo do Sistema e Definições do Problema

Considera-se o *downlink* de uma rede 5G NR composta por uma gNB e por UEs que compartilham os recursos de rádio disponíveis. A política do IA-FMR possui uma quantidade fixa de posições destinada à representação dos usuários. Essa quantidade é denotada por N . Na configuração final desta tese, utiliza-se:

$$N = 9. \quad (4.1)$$

O valor nove corresponde à dimensionalidade máxima escolhida para a arquitetura da política e não a um limite do ns-3/5G-LENA ou da tecnologia 5G NR.

Em uma decisão específica, nem todas as posições precisam estar ocupadas. Seja $\mathcal{U}_t$ o conjunto de UEs elegíveis no instante de decisão t e:

$$N_t = |\mathcal{U}_t| \quad (4.2)$$

sua cardinalidade, com $N_t \leq N$. Um UE elegível é um usuário considerado pelo escalonador como candidato ao recebimento de recursos naquela decisão. No ambiente final de treinamento, todas as nove posições são utilizadas e, portanto, $N_t = N = 9$. Na inferência, conjuntos menores podem ser representados por preenchimento com zeros nas posições não ocupadas, conforme detalhado posteriormente.

Seja B_t a quantidade de unidades de alocação disponível para a decisão. Na configuração empregada nesta tese, o tamanho do RBG é definido como um PRB. Assim, para a implementação considerada,

$$1 \text{ RBG} = 1 \text{ PRB}. \quad (4.3)$$

Essa igualdade é uma escolha da configuração utilizada e não uma definição geral de RBG em 5G NR. A denominação RBG é mantida no restante do capítulo por corresponder à unidade manipulada pela interface do escalonador.

A política não produz diretamente quantidades inteiras de RBGs. A ação contínua é transformada em uma distribuição inteira denominada *alocação alvo*:

$$\tilde{\mathbf{r}}_t = [\tilde{r}_{1,t}, \tilde{r}_{2,t}, \dots, \tilde{r}_{N_t,t}], \quad (4.4)$$

em que $\tilde{r}_{i,t}$ representa a quantidade inteira de RBGs solicitada para o UE i . Quando existem UEs elegíveis e recursos disponíveis, a conversão da ação preserva o total disponível:

$$\sum_{i=1}^{N_t} \tilde{r}_{i,t} = B_t, \quad \tilde{r}_{i,t} \in \mathbb{Z}_{\geq 0}. \quad (4.5)$$

No ambiente Python de treinamento, essa alocação é utilizada diretamente pelo modelo simplificado para calcular as grandezas da decisão. Não existe nessa etapa uma camada MAC adicional capaz de modificar a distribuição. Por esse motivo, todas as métricas da recompensa e os históricos do estado de treinamento são calculados a partir de $\tilde{\mathbf{r}}_t$.

Na integração sistêmica, a alocação alvo orienta a concessão de recursos, mas pode ser modificada pelos procedimentos da camada MAC. Para distinguir as duas grandezas, denota-se por:

$$\mathbf{r}_t = [r_{1,t}, r_{2,t}, \dots, r_{N_t,t}] \quad (4.6)$$

a *alocação aplicada*, na qual $r_{i,t}$ corresponde à quantidade de RBGs efetivamente concedida ao UE i pelo escalonador no ns-3/5G-LENA. Em geral,

$$\mathbf{r}_t \neq \tilde{\mathbf{r}}_t \quad (4.7)$$

pode ocorrer, embora as duas distribuições sejam frequentemente próximas. Os mecanismos responsáveis por essa diferença são apresentados na Seção 4.5.4.

A condição de canal utilizada pelo modelo de treinamento é representada pelo índice de MCS de *downlink* do UE i no instante t , denotado por $mcs_{i,t}$. Esse índice é convertido em eficiência espectral por uma tabela baseada na especificação 3GPP utilizada pela implementação. A partir do MCS e da alocação alvo, o ambiente estima a vazão individual por:

$$T_{i,t} = \eta(mcs_{i,t}) \tilde{r}_{i,t} C, \quad (4.8)$$

em que $\eta(mcs_{i,t})$ é a eficiência espectral associada ao MCS e C converte a quantidade de recursos e a eficiência espectral em Mbps. Na aproximação utilizada,

$$C = \frac{\rho 12 \Delta f}{10^6}, \quad (4.9)$$

com $\rho = 1,0$ e $\Delta f = 30$ kHz. O fator 12 corresponde ao número de subportadoras de um PRB. Para essa configuração, $C = 0,36$.

Essa expressão não é um modelo completo da pilha 5G NR. Ela estabelece apenas uma relação simplificada entre eficiência espectral, quantidade de recursos e vazão. O ambiente de treinamento não reproduz formação de blocos de transporte, HARQ, modelo de erro, filas de protocolo, perdas de pacotes ou retransmissões. Consequentemente, $T_{i,t}$ é uma grandeza interna do treinamento e não deve ser confundida com a vazão fim a fim medida posteriormente no ns-3/5G-LENA.

Essa distinção é importante porque o IA-FMR não é treinado diretamente sobre atraso, perda, ocupação real das filas ou métricas da aplicação. O treinamento fornece uma política a partir de uma representação simplificada do problema. A avaliação sistêmica, apresentada nos capítulos seguintes, verifica como essa política se comporta quando passa a interagir com uma pilha de simulação mais detalhada.

A partir dessas definições, o objetivo de projeto do IA-FMR é produzir distribuições de recursos que considerem simultaneamente três aspectos: a uniformidade da alocação entre os UEs, o aproveitamento das diferenças de eficiência espectral e a ocorrência de vazões estimadas abaixo de uma referência interna de treinamento. Esses aspectos são representados explicitamente na função de recompensa descrita posteriormente.

4.3 Formulação do IA-FMR como Processo de Decisão

Para o treinamento com PPO, o problema é tratado como um processo de decisão sequencial representado por um MDP. Em cada instante t , o agente recebe o estado observado $\mathbf{s}_t$, produz uma ação contínua $\mathbf{a}_t$ e recebe uma recompensa R_t calculada a partir da consequência imediata dessa ação no ambiente de treinamento.

A representação adotada pode ser escrita como:

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, R, \gamma), \quad (4.10)$$

em que $\mathcal{S}$ representa o espaço de estados fornecidos à política, $\mathcal{A}$ o espaço de ações, P a dinâmica de transição entre decisões sucessivas, R a função de recompensa e γ o fator de

desconto utilizado pelo PPO.

Essa formulação é uma abstração do ambiente de treinamento. O estado não contém todas as variáveis físicas de uma rede 5G e a evolução do canal é fornecida por traços previamente registrados. Assim, o MDP representa o problema tal como apresentado ao algoritmo de aprendizado e não uma descrição completa do estado interno da rede.

O objetivo do treinamento é obter uma política parametrizada $\pi_\theta(\mathbf{a}_t | \mathbf{s}_t)$ que maximize o retorno descontado esperado ao longo das interações:

$$\pi_\theta^* = \arg \max_{\pi_\theta} \mathbb{E}_{\pi_\theta} \left[\sum_{t=0}^{H-1} \gamma^t R_t \right], \quad (4.11)$$

em que H representa o horizonte do episódio e θ reúne os parâmetros aprendidos da política.

Depois do treinamento, os parâmetros θ são congelados. Na avaliação sistêmica, o ns-3/5G-LENA fornece as informações necessárias para reconstruir o mesmo formato de entrada e a política produz ações por inferência determinística. Nessa etapa, a recompensa não é utilizada e os parâmetros não são atualizados.

4.3.1 Ambiente de Treinamento e Dinâmica Temporal

O ambiente Python fornece as interações utilizadas para treinar a política do IA-FMR. O tempo é representado por uma sequência discreta de instantes de decisão. Em cada instante, o ambiente apresenta as condições correntes de canal, recebe a ação da política, converte essa ação em uma alocação alvo e calcula as grandezas utilizadas na recompensa e no estado seguinte.

As condições de canal são obtidas de traços temporais de MCS provenientes de um conjunto público da plataforma O-RAN Colosseum (Lopes *et al.*, 2025, 2026). Para a configuração final desta tese, são utilizados traços correspondentes a nove UEs. O pré-processamento remove valores nulos, indefinidos ou não inteiros, preservando a ordem temporal das amostras válidas.

Como esses traços foram registrados previamente, a evolução do MCS é exógena à ação do agente. A decisão tomada em t não modifica os valores de MCS apresentados em $t+1$. O que a ação modifica são as grandezas derivadas da própria decisão, como a fração de recursos atribuída a cada posição, a vazão estimada, o Índice de Jain e o coeficiente de

ponderação. Essas informações são preservadas para compor o estado seguinte juntamente com a próxima amostra de MCS.

O ambiente de treinamento não simula pacotes, filas das camadas de protocolo, formação de blocos de transporte, HARQ, modelos de erro ou mobilidade. A relação entre MCS, recursos e vazão segue a aproximação apresentada na Seção 4.2. O treinamento concentra-se, portanto, no efeito da distribuição dos recursos sobre justiça, eficiência normalizada e déficit de vazão estimada.

Em cada instante de decisão, o ciclo do ambiente pode ser resumido da seguinte forma:

1. o ambiente obtém os valores correntes de MCS a partir dos traços.
2. esses valores são combinados com as informações preservadas da decisão anterior para construir o estado $\mathbf{s}_t$.
3. a política PPO recebe o estado e produz a ação contínua $\mathbf{a}_t$.
4. os componentes da ação associados à distribuição são convertidos em uma alocação inteira de recursos.
5. o ambiente estima as vazões e calcula os termos de justiça, eficiência e déficit.
6. a recompensa R_t é calculada e as informações necessárias ao estado seguinte são atualizadas.

4.3.2 Espaço de Estados

O vetor fornecido à política combina informação corrente de canal com um histórico de uma decisão. Para uma arquitetura com N posições, o estado observado no instante t é:

$$\mathbf{s}_t = [\mathbf{m}_t, \mathbf{b}_{t-1}, \mathbf{x}_{t-1}, J_{t-1}, \alpha_{t-1}], \quad (4.12)$$

em que $\mathbf{m}_t$ descreve as condições relativas de canal na decisão corrente. Os demais componentes representam informações derivadas da decisão precedente. Essa estrutura fornece à política uma memória de curto prazo sem utilizar uma rede recorrente.

O primeiro bloco é:

$$\mathbf{m}_t = [m_{1,t}, m_{2,t}, \dots, m_{N,t}], \quad (4.13)$$

com:

$$m_{i,t} = \frac{mcs_{i,t}}{\max(1, \max_j mcs_{j,t})}. \quad (4.14)$$

O denominador é o maior MCS presente na observação, com piso igual a 1 para evitar divisão por zero. Assim, cada componente permanece em $[0,1]$ e representa a condição do UE de forma relativa às demais posições ativas naquela decisão. Essa normalização não transforma MCS em eficiência espectral. Ela apenas reescala os índices antes de apresentá-los à rede neural.

Durante o treinamento final, as nove posições estão ocupadas. Durante a inferência com $N_t < N$, os UEs elegíveis ocupam as primeiras N_t posições e os componentes restantes de $\mathbf{m}_t$ são preenchidos com zero. O estado não inclui uma máscara explícita de posições ativas. A quantidade de logits efetivamente utilizada na distribuição dos recursos é controlada externamente pelo número de UEs elegíveis recebido pela interface.

O segundo bloco descreve a participação de cada posição na alocação alvo anterior:

$$\mathbf{b}_{t-1} = [b_{1,t-1}, b_{2,t-1}, \dots, b_{N,t-1}], \quad (4.15)$$

em que:

$$b_{i,t-1} = \frac{\tilde{r}_{i,t-1}}{\sum_{j=1}^N \tilde{r}_{j,t-1}}. \quad (4.16)$$

Quando uma posição está inativa, sua alocação alvo é representada por zero. Quando nenhum recurso foi distribuído, todos os componentes de $\mathbf{b}_{t-1}$ são definidos como zero. O valor $b_{i,t-1} = 0,20$, por exemplo, significa que a posição i recebeu 20% dos recursos da alocação alvo anterior.

O terceiro bloco registra uma aproximação normalizada da vazão associada à decisão precedente:

$$\mathbf{x}_{t-1} = [x_{1,t-1}, x_{2,t-1}, \dots, x_{N,t-1}], \quad (4.17)$$

com:

$$x_{i,t-1} = \frac{T_{i,t-1}}{T_{t-1}^{\text{ideal}}}. \quad (4.18)$$

O numerador é a vazão estimada a partir do MCS e da alocação alvo anterior. O denominador T_{t-1}^{ideal} é a referência de eficiência definida na Subseção 4.3.5. Dessa forma, $x_{i,t-1}$ não representa a vazão fim a fim do UE. Ele informa à política a contribuição estimada daquela posição em relação à referência usada pelo próprio ambiente simplificado.

Os dois últimos componentes são escalares. J_{t-1} é o Índice de Jain calculado sobre a representação da alocação alvo anterior e α_{t-1} é o coeficiente produzido pela política na decisão anterior. O primeiro fornece um resumo da uniformidade da distribuição. O segundo devolve à política o contexto de ponderação produzido por ela própria na decisão precedente.

A dimensão do estado é, portanto,

$$\dim(\mathcal{S}) = 3N + 2. \quad (4.19)$$

Para $N = 9$, a política recebe 29 componentes:

$$\begin{aligned} \dim(\mathcal{S}) &= 9 \text{ MCSs normalizados} + 9 \text{ frações de alocação} \\ &\quad + 9 \text{ vazões normalizadas} + 1 \text{ Índice de Jain} + 1 \text{ coeficiente} \\ &= 29. \end{aligned} \quad (4.20)$$

Na primeira observação de uma execução, os históricos de alocação, vazão e Jain são inicializados com zero. Na configuração final, o contexto inicial do coeficiente também é zero. A partir da primeira ação, os componentes históricos passam a refletir a decisão anterior.

Há duas distinções importantes. Primeiro, o estado da política não contém diretamente demanda em fila, identificadores dos UEs, quantidade de pacotes, atraso ou informações da aplicação. Segundo, durante a inferência os históricos de $\mathbf{b}_{t-1}$, $\mathbf{x}_{t-1}$ e J_{t-1} são reconstruídos a partir da alocação alvo produzida pelo processo Python, e não da alocação eventualmente modificada pela camada MAC. Essa escolha mantém a escala e a estrutura esperadas pela política treinada, mas significa que o histórico observado pela política representa sua própria decisão anterior em forma simplificada, e não todo o estado efetivamente realizado na rede.

4.3.3 Espaço de Ações e Coeficiente de Ponderação

A ação do IA-FMR possui dois tipos de componentes. Os primeiros N valores determinam a participação relativa dos UEs na distribuição dos recursos. O último valor controla o coeficiente de ponderação entre justiça e eficiência utilizado durante o treinamento.

No instante t , a ação é escrita como:

$$\mathbf{a}_t = [z_{1,t}, z_{2,t}, \dots, z_{N,t}, u_t], \quad (4.21)$$

em que $z_{i,t}$ é o componente associado à posição i e u_t é o escalar utilizado para obter α_t .

Os valores $z_{i,t}$ são denominados *logits*. Um logit não é uma quantidade de RBGs e também não é, por si só, uma probabilidade. Ele é um valor contínuo cuja importância depende de sua relação com os demais logits da mesma decisão. A função *softmax*, apresentada na próxima subseção, transforma esse conjunto de valores em proporções normalizadas. Em termos intuitivos, um UE cujo logit seja maior que os demais tende a receber uma participação maior na distribuição, enquanto logits menores tendem a produzir participações menores.

O componente u_t também é um valor contínuo. Ele não é utilizado diretamente como α_t . Na configuração final da tese, aplica-se:

$$\alpha_t = \sigma \left(\frac{u_t}{T_\alpha} \right), \quad (4.22)$$

em que:

$$\sigma(v) = \frac{1}{1 + \exp(-v)} \quad (4.23)$$

é a função sigmoide e:

$$T_\alpha = 0,70. \quad (4.24)$$

A sigmoide transforma o escalar em um coeficiente entre 0 e 1. O coeficiente pode representar todo o domínio natural da ponderação entre os dois termos, desde valores próximos de 0 até valores próximos de 1. A função sigmoide possui extremos assintóticos, de modo que 0 e 1 são limites do domínio de ponderação e não valores atingidos exatamente

por uma entrada finita.

O parâmetro T_α é denominado temperatura da transformação. Ele não representa uma temperatura física. Seu papel é controlar quão rapidamente a sigmoide se desloca entre valores próximos de 0 e de 1 quando u_t varia. Valores menores tornam a transição mais acentuada, enquanto valores maiores produzem uma mudança mais gradual. Na configuração adotada, $T_\alpha = 0,70$ caracteriza a sensibilidade utilizada nas políticas finais.

Durante o treinamento, o espaço de ações contínuas é limitado a:

$$-5 \leq a_{k,t} \leq 5, \quad k = 1, \dots, N + 1. \quad (4.25)$$

Esse limite é uma configuração numérica do ambiente e restringe a magnitude dos logits e de u_t antes das transformações. Ele não representa uma quantidade física de recursos.

A dimensão do espaço de ações é:

$$\dim(\mathcal{A}) = N + 1. \quad (4.26)$$

Para $N = 9$, a ação possui dez componentes: nove logits e o escalar u_t .

Durante o treinamento, α_t participa diretamente da recompensa da decisão corrente. Durante a inferência, a transformação continua sendo executada para manter o comportamento da política treinada, mas a recompensa não é calculada. O valor de α_t é armazenado no estado utilizado na decisão seguinte. Por esse motivo, não se deve interpretar α_t como um comando externo aplicado depois dos logits ou como um fator que multiplica a alocação já calculada.

A Figura 11 apresenta o encadeamento entre estado, ação, alocação e recompensa.

A partir de $\mathbf{s}_t$, a política produz $\mathbf{a}_t$. Os logits seguem o caminho de conversão em proporções e, posteriormente, em uma alocação inteira. Em paralelo, u_t é transformado em α_t . No treinamento, a alocação permite calcular os termos da recompensa e atualizar os históricos do estado seguinte. Na inferência, os mesmos componentes da ação são produzidos, mas sem atualização dos parâmetros da política.

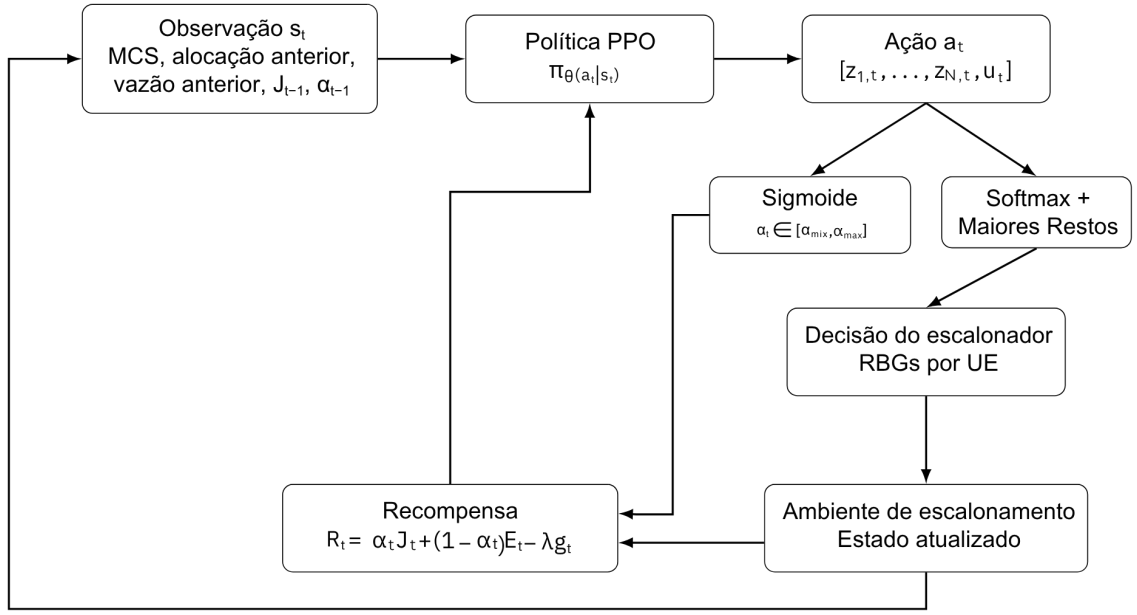


Figura 11: Ciclo de decisão do IA-FMR, destacando os logits utilizados na distribuição dos recursos e o componente escalar transformado em α_t .

4.3.4 Conversão da Ação em Alocação Alvo

Os logits precisam ser convertidos em uma distribuição utilizável pelo escalonador. Essa conversão ocorre em duas etapas. Primeiro, os logits são transformados em proporções contínuas por uma função *softmax*. Depois, essas proporções são convertidas em quantidades inteiras de RBGs pelo método dos maiores restos.

Para os N_t UEs elegíveis na decisão, a proporção associada ao UE i é:

$$p_{i,t} = \frac{\exp(z_{i,t}/\tau)}{\sum_{j=1}^{N_t} \exp(z_{j,t}/\tau)}, \quad (4.27)$$

em que $\tau > 0$ é a temperatura da *softmax*. Na configuração final,

$$\tau = 0,65. \quad (4.28)$$

Assim como T_α , τ não representa uma temperatura física. Esse parâmetro controla o contraste entre as proporções produzidas. Valores menores acentuam as diferenças entre os logits e tendem a concentrar mais a distribuição. Valores maiores suavizam essas diferenças e aproximam as proporções entre si.

As proporções resultantes satisfazem:

$$p_{i,t} \geq 0, \quad \sum_{i=1}^{N_t} p_{i,t} = 1. \quad (4.29)$$

Matematicamente, esse vetor possui as propriedades de uma distribuição de probabilidade. Operacionalmente, porém, o IA-FMR não sorteia um UE a partir dessa distribuição. Cada $p_{i,t}$ é utilizado diretamente como a participação relativa do UE i no total de recursos.

Durante o treinamento, $N_t = N = 9$. Na inferência, somente os logits correspondentes às posições ocupadas por UEs elegíveis participam do denominador da *softmax*. Os logits das posições preenchidas com zero no estado continuam existindo como saídas da rede neural, mas são excluídos da normalização e não recebem recursos.

A quantidade contínua associada a cada UE é:

$$y_{i,t} = p_{i,t} B_t. \quad (4.30)$$

O valor $y_{i,t}$ expressa a participação idealizada em uma escala contínua. Se, por exemplo, $y_{i,t} = 2,7$, a proporção produzida corresponde a 2,7 unidades de recurso. Como a decisão precisa utilizar quantidades inteiras de RBGs, esse valor ainda não é uma alocação aplicável.

A conversão para inteiros utiliza o método dos maiores restos. Inicialmente, cada UE recebe a parte inteira de $y_{i,t}$:

$$\tilde{r}_{i,t}^0 = \lfloor y_{i,t} \rfloor. \quad (4.31)$$

A quantidade ainda não distribuída é:

$$B_t^{\text{restante}} = B_t - \sum_{i=1}^{N_t} \tilde{r}_{i,t}^0. \quad (4.32)$$

Os recursos restantes são atribuídos, uma unidade por vez, aos UEs com as maiores partes fracionárias de $y_{i,t}$. O procedimento preserva o total disponível:

$$\sum_{i=1}^{N_t} \tilde{r}_{i,t} = B_t, \quad (4.33)$$

produzindo o vetor inteiro $\tilde{\mathbf{r}}_t$.

A Tabela 5 apresenta um exemplo com três UEs, $B_t = 10$ e $\tau = 0,65$.

Tabela 5: Exemplo de conversão dos logits em uma alocação inteira pelo método dos maiores restos.

UE	Logit	Proporção	$y_{i,t}$	Parte inteira	Alocação alvo
UE 1	1,20	0,710	7,102	7	7
UE 2	0,40	0,207	2,074	2	2
UE 3	-0,20	0,082	0,824	0	1
Total		1,000	10,000	9	10

As proporções são exibidas com três casas decimais, razão pela qual pequenas diferenças podem aparecer na soma dos valores impressos. No exemplo, a parte inteira distribui inicialmente nove RBGs. O UE 3 possui a maior parte fracionária e recebe a unidade restante, resultando em:

$$\tilde{\mathbf{r}}_t = [7, 2, 1]. \quad (4.34)$$

A saída desse procedimento é a alocação alvo da política. No treinamento, ela é utilizada diretamente pelo ambiente simplificado. Na integração com o ns-3/5G-LENA, ela é enviada ao escalonador e ainda pode ser modificada pelos procedimentos necessários à concessão efetiva dos recursos.

4.3.5 Função de Recompensa

A função de recompensa fornece ao PPO um valor escalar que resume três propriedades da decisão produzida no ambiente simplificado: a uniformidade da distribuição dos recursos, a eficiência de vazão em relação a uma referência e o déficit de vazão estimada dos UEs. Esses três componentes têm papéis diferentes e não devem ser interpretados como métricas equivalentes.

A recompensa no instante t é:

$$R_t = \alpha_t J_t + (1 - \alpha_t) E_t - \lambda g_t^{\text{TP}}, \quad (4.35)$$

em que J_t representa a justiça na distribuição dos RBGs, E_t representa a eficiência normalizada da decisão e g_t^{TP} representa o déficit médio de vazão estimada. O coeficiente α_t pondera apenas os dois primeiros termos. A penalização por déficit é subtraída separadamente.

A justiça é calculada pelo Índice de Jain aplicado à alocação alvo da decisão corrente (Jain; Chiu; Hawe, 1998):

$$J_t = \frac{\left(\sum_{i=1}^N \tilde{r}_{i,t}\right)^2}{N \sum_{i=1}^N \tilde{r}_{i,t}^2}. \quad (4.36)$$

Na configuração final de treinamento, $N = 9$ e todas as posições estão ativas. Para alocações não negativas e com ao menos um recurso distribuído, J_t pertence a $[1/N, 1]$. O valor 1 corresponde à distribuição uniforme. Valores menores indicam maior concentração dos RBGs em parte dos UEs.

Esse índice mede justiça na *quantidade de recursos*, e não igualdade de vazão. Dois UEs podem receber a mesma quantidade de RBGs e obter vazões diferentes porque suas eficiências espectrais são distintas. Por essa razão, um valor elevado de Jain não constitui, isoladamente, garantia de vazão mínima ou ausência de *starvation*.

O termo de eficiência é medido pela vazão agregada estimada no ambiente, calculada como:

$$T_t = \sum_{i=1}^N T_{i,t}, \quad (4.37)$$

com $T_{i,t}$ calculado pela aproximação da Equação 4.8. Para normalizar essa grandeza, define-se uma referência que atribui todos os recursos ao UE com maior eficiência espectral na decisão:

$$T_t^{\text{ideal}} = B_t C \max_i \eta(mcs_{i,t}). \quad (4.38)$$

O termo “ideal” é utilizado apenas como nome da referência de normalização empregada na implementação. T_t^{ideal} não representa a capacidade física da célula, uma solução ótima do problema multiobjetivo ou um ponto de operação que considere simultaneamente justiça e atendimento. Ele corresponde somente à vazão estimada pelo modelo simplificado caso todos os recursos sejam atribuídos ao UE de maior eficiência espectral naquele instante.

A eficiência normalizada é:

$$E_t = \frac{T_t}{T_t^{\text{ideal}}}, \quad (4.39)$$

com o valor limitado numericamente ao intervalo $[0,1]$. Valores próximos de 1 significam que a vazão agregada estimada da decisão está próxima da referência definida acima. Isso não implica que a distribuição seja justa.

Os termos J_t e E_t são globais. Uma distribuição pode apresentar Jain elevado ou eficiência agregada elevada e, ainda assim, algum UE obter vazão estimada muito baixa. Para introduzir uma pressão adicional de treinamento sobre esse comportamento, utiliza-se uma penalidade por *deficit* de vazão:

$$g_t^{\text{TP}} = \frac{1}{N} \sum_{i=1}^N \frac{\max(0, T_{\min} - T_{i,t})}{T_{\min}}. \quad (4.40)$$

Para um UE com $T_{i,t} \geq T_{\min}$, a contribuição é zero. Quando $T_{i,t} < T_{\min}$, o déficit cresce de forma proporcional à distância em relação à referência. Se $T_{i,t} = T_{\min}/2$, a contribuição daquele UE é 0,5. Se $T_{i,t} = 0$, a contribuição é 1. O termo g_t^{TP} é a média desses déficits entre as nove posições utilizadas no treinamento.

Na configuração final,

$$T_{\min} = 0,40 \text{ Mbps} \quad (4.41)$$

e:

$$\lambda = 1,2. \quad (4.42)$$

O valor $T_{\min}$ é uma referência interna do ambiente de treinamento. Ele não reserva recursos, não estabelece uma taxa mínima garantida, não é requisito de QoS e não corresponde a um requisito da avaliação com tráfego sintético de vídeo. Da mesma forma, λ apenas controla quanto o déficit reduz a recompensa. Esses valores foram definidos durante o desenvolvimento da proposta sem uma busca sistemática de hiperparâmetros.

A penalização não é multiplicada por α_t . Portanto, mesmo quando α_t se aproxima de 1, o déficit de vazão continua contribuindo para a recompensa por meio do termo $-\lambda g_t^{\text{TP}}$.

A Figura 12 resume essa composição. α_t distribui o peso entre justiça e eficiência, enquanto o déficit constitui um terceiro termo independente dessa ponderação. A recom-

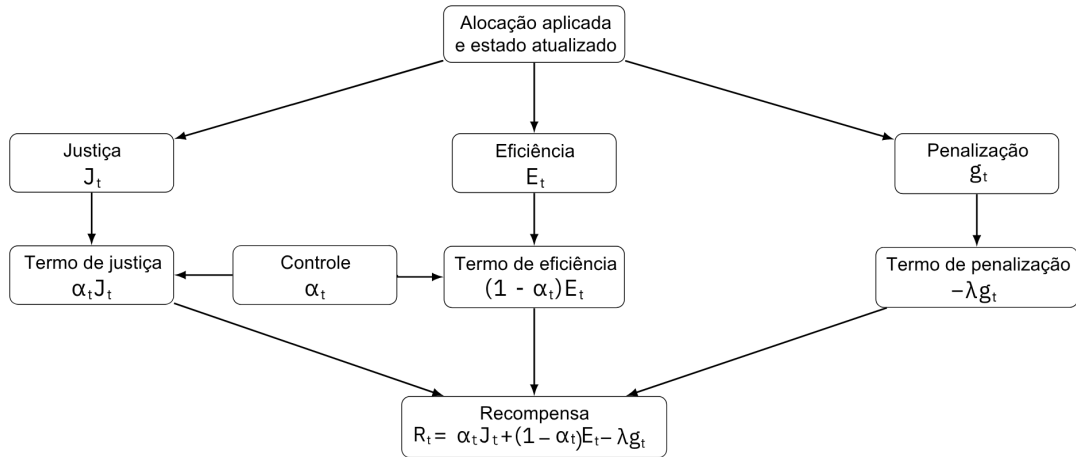


Figura 12: Composição da recompensa do IA-FMR a partir da justiça na alocação, da eficiência normalizada, da ponderação por α_t e da penalização por déficit de vazão.

pensa é utilizada somente durante o treinamento. Na inferência com a política congelada, J_t , E_t , g_t^{TP} e R_t não são recalculados para atualizar a política.

4.4 Treinamento da Política IA-FMR

A política do IA-FMR é treinada de forma *offline* no ambiente Python descrito anteriormente. O treinamento utiliza o algoritmo *Proximal Policy Optimization* (PPO) (Schulman et al., 2017). Ao final de cada execução, os parâmetros aprendidos são armazenados e posteriormente utilizados sem atualização durante a avaliação sistêmica.

A escolha do PPO está relacionada a duas características da formulação. Primeiro, a ação do IA-FMR é contínua, pois tanto os logits quanto o componente u_t são produzidos como valores reais antes das transformações descritas anteriormente. Segundo, o PPO utiliza um objetivo com *clipping* para limitar alterações excessivamente grandes da política entre atualizações sucessivas. Esse mecanismo oferece uma forma prática de controlar o processo de atualização em um problema sequencial.

Essa escolha é metodológica. A tese não realiza uma comparação sistemática entre PPO e outros algoritmos de aprendizado por reforço e, portanto, não conclui que o PPO seja a alternativa ótima para o problema. A fundamentação do algoritmo e do mecanismo de *clipping* é apresentada no capítulo de conceitos e a configuração utilizada é registrada a seguir.

4.4.1 Configuração do Treinamento e Arquitetura da Política

A implementação utiliza a interface de ambientes do Gymnasium e o PPO da biblioteca Stable-Baselines3 (Lopes *et al.*, 2025, 2026; Schulman *et al.*, 2017). A política é implementada por uma `ActorCriticPolicy` com `MlpPolicy`. O termo *actor* identifica o ramo responsável pela distribuição de ações. O *critic* estima a função de valor utilizada no cálculo das atualizações durante o treinamento.

O estado de 29 componentes é processado por um `FlattenExtractor`. Em seguida, *actor* e *critic* utilizam ramos independentes, cada um com duas camadas densas de 64 neurônios e ativação *Tanh*. No PPO da Stable-Baselines3, o *actor* parametriza uma distribuição contínua de ações. Durante o treinamento, as ações são amostradas dessa distribuição. Na inferência determinística utilizada posteriormente, emprega-se a ação central da distribuição aprendida, respeitando os limites definidos pelo espaço de ações.

A Tabela 6 registra o ambiente de software e a arquitetura neural utilizados.

Tabela 6: Ambiente de software e arquitetura neural utilizados no treinamento do IA-FMR.

Elemento	Configuração
Python	3.10.12
Gymnasium	1.3.0
Stable-Baselines3	2.9.0
PyTorch	2.12.1+cu130
NumPy	2.2.6
Política	<code>ActorCriticPolicy</code> com <code>MlpPolicy</code>
Extrator de características	<code>FlattenExtractor</code>
Ramo do <i>actor</i>	Duas camadas densas de 64 neurônios, com ativação <i>Tanh</i>
Ramo do <i>critic</i>	Duas camadas densas de 64 neurônios, com ativação <i>Tanh</i>
Dimensão da ação contínua	10 componentes
Saída de valor do <i>critic</i>	1 componente

A implementação utilizada no desenvolvimento é organizada em módulos que separam o ambiente de treinamento da configuração das execuções. O ambiente contém as operações de normalização, conversão MCS-eficiência espectral, transformação da ação em recursos e cálculo da recompensa. O módulo de construção das execuções define as configurações experimentais, instancia o PPO e armazena as políticas obtidas. Essa separação

permite utilizar a mesma formulação do ambiente em diferentes larguras de banda.

Cada política final foi treinada durante 600.000 passos de interação. Um passo corresponde a uma transição completa do ambiente. O PPO coleta 2.048 passos antes de realizar cada conjunto de atualizações. A duração de um episódio é determinada pelo percurso do segmento de traço de MCS utilizado pelo ambiente.

Como a quantidade de recursos varia com a largura de banda, foi treinada uma política independente para cada largura de banda empregada na avaliação. Todas possuem a mesma arquitetura, o mesmo espaço de estados e o mesmo espaço de ações. O que muda entre esses treinamentos é a quantidade de recursos B_t associada à configuração.

Para cada largura de banda foi realizada uma única execução de treinamento com a semente 211382627. Essa semente identifica a instância de treinamento utilizada na tese. As 100 sementes empregadas posteriormente nas simulações sistêmicas têm função diferente e não medem variabilidade entre treinamentos independentes. Consequentemente, os resultados desta tese não permitem quantificar a sensibilidade da política final a diferentes inicializações do PPO.

A Tabela 7 reúne a configuração final utilizada.

Tabela 7: Parâmetros da configuração final de treinamento do IA-FMR.

Parâmetro	Valor
Número de UEs no treinamento	9
Passos de treinamento	600.000
Semente de treinamento	211382627
Execuções de treinamento	Uma por largura de banda
Espaçamento entre subportadoras	30 kHz
Mapeamento de recursos no ambiente	1 RBG = 1 PRB
Dimensão do estado	29 componentes
Dimensão da ação	10 componentes
Intervalo da ação contínua	$[-5,5]$
Temperatura da <i>softmax</i> (τ)	0,65
Transformação do coeficiente	$\alpha_t = \sigma(u_t/0,70)$
Domínio de ponderação de α_t	$[0,1]$
Métrica de justiça na recompensa	Jain sobre os RBGs da decisão corrente

Continua na próxima página

Tabela 7 – continuação

Parâmetro	Valor
Limiar de vazão por UE ($T_{\min}$)	0,40 Mbps
Peso da penalização (λ)	1,2
Taxa de aprendizado	0,0003, constante
Passos por coleta do PPO (n_steps)	2048
Tamanho do lote (batch_size)	64
Épocas por atualização (n_epochs)	10
Fator de desconto (γ)	0,99
Parâmetro GAE (λ_{GAE})	0,95
Normalização da vantagem	Habilitada
Intervalo de <i>clipping</i>	0,2
Coefficiente de entropia	0,005
Coefficiente da função de valor	0,5
Norma máxima do gradiente	0,5

Os valores de τ , T_α , $T_{\min}$, λ e do coeficiente de entropia foram definidos durante testes preliminares de desenvolvimento. Não foi realizada uma busca sistemática de hiperparâmetros, nem uma análise completa de sensibilidade individual desses valores. Os demais parâmetros do PPO correspondem à configuração utilizada pela implementação adotada. A tabela caracteriza as políticas efetivamente avaliadas nesta tese e não estabelece uma combinação universal ou ótima.

As políticas finais utilizam a transformação sigmoide para produzir α_t no domínio natural de ponderação $[0,1]$. Assim, concentrações em regiões específicas desse domínio, quando observadas nos resultados, caracterizam o comportamento produzido pela política nas realizações analisadas.

4.4.2 Evidências Preliminares da Evolução da Formulação

Antes da configuração final utilizada na avaliação sistêmica, diferentes etapas da evolução FMR-IA-FMR foram examinadas no ambiente Python. Esses resultados têm finalidade histórica e metodológica: mostram como a formulação foi desenvolvida, mas não substituem os experimentos finais da tese e não devem ser interpretados como comparação direta entre versões, pois empregam cenários e configurações distintos.

A primeira etapa corresponde ao FMR original (Lopes *et al.*, 2025). Ela considera

três UEs, 10 MHz, numerologia $\mu = 1$, SCS de 30 kHz e 25 unidades de alocação. O treinamento utilizou 204.800 passos e separação temporal entre os dados de treinamento e avaliação. A Tabela 8 reproduz os resultados dessa etapa.

Tabela 8: Resultados preliminares do FMR original no ambiente Python.

Escalonador	Índice de Jain	Vazão agregada média (Mbps)
RR	0,9247	38,48
PF	0,7362	43,06
MR	0,3411	53,62
FMR	0,4696	48,05

Nessa configuração, RR e MR ocupam extremos distintos entre uniformidade da alocação e vazão agregada. O FMR apresenta valores intermediários entre esses comportamentos. Os três UEs recebem recursos na execução avaliada, mas esse resultado descreve apenas aquela configuração e não constitui garantia de atendimento em outros cenários.

Uma etapa posterior introduziu a formulação *in-action* (Lopes *et al.*, 2026). Essa avaliação utilizou nove UEs, 100 MHz, SCS de 30 kHz e 273 unidades de alocação, com 30 sementes e segmentos temporais distintos para treinamento e avaliação. Além da vazão agregada e do Índice de Jain, foram analisadas a menor vazão média entre os UEs e a fração dos instantes em que pelo menos um usuário apresentou vazão nula.

Tabela 9: Resultados preliminares da formulação *in-action* no ambiente Python.

Escalonador	Jain	Vazão agregada (Mbps)	Vazão mínima por UE (Mbps)	Fração com vazão nula
RR	1,000	129,51	2,53	0,000
PF	0,686	215,23	0,08	0,000
MR	0,111	536,76	0,00	0,889
IA-FMR	0,722	209,83	0,50	0,000

Nessa etapa, o IA-FMR apresenta 209,83 Mbps de vazão agregada e Índice de Jain de 0,722. Em relação ao PF, a vazão média é aproximadamente 2,5% menor, enquanto o Índice de Jain e a menor vazão média por UE são maiores. O MR alcança a maior vazão agregada, mas apresenta forte concentração da alocação e ocorrência frequente de UE com vazão nula.

Esses valores serviram como evidência de desenvolvimento para prosseguir com a integração ao ns-3/5G-LENA. Eles não caracterizam a configuração final das políticas

utilizadas nos Capítulos 5 e 6. Em particular, a configuração final foi novamente treinada com o domínio completo de ponderação de α_t em $[0,1]$ e com os parâmetros registrados na Tabela 7.

As duas etapas preliminares utilizam uma representação simplificada baseada em MCS e quantidade de recursos. Elas não reproduzem pacotes, filas das camadas de protocolo, blocos de transporte, HARQ, modelos de erro ou a dinâmica de tráfego empregada posteriormente. Por isso, seus valores absolutos não devem ser comparados diretamente aos resultados sistêmicos nem utilizados para inferir ganho de uma versão sobre a outra.

4.5 Integração do IA-FMR com o ns-3/5G-LENA

Após o treinamento, a política é integrada ao ns-3/5G-LENA por meio do ns3-ai (Henderson *et al.*, 2008; Koutlia *et al.*, 2022; Yin *et al.*, 2020). A integração tem dois objetivos: disponibilizar à política informações produzidas pelo simulador e converter sua resposta em uma solicitação de alocação compreensível pelo escalonador da camada MAC.

Durante essa etapa, a política não aprende. Seus parâmetros permanecem congelados e a execução do modelo corresponde apenas à inferência. O ns-3/5G-LENA continua responsável pela evolução da rede, pelas filas, pela formação das transmissões e pela concessão efetiva dos recursos.

4.5.1 Visão Geral da Arquitetura de Integração

A arquitetura reúne três componentes principais. O escalonador `NrMacSchedulerOfdmaFmr`, implementado em C++ no 5G-LENA, reúne as informações da decisão corrente e realiza as concessões na camada MAC. O processo Python carrega a política PPO treinada, reconstrói o estado de 29 componentes e executa a inferência. O ns3-ai fornece a memória compartilhada utilizada para a comunicação entre esses dois processos.

O fluxo é síncrono. Quando uma decisão é necessária, o escalonador disponibiliza as informações correntes na memória compartilhada e aguarda a resposta do processo Python. O Python constrói o estado, executa a política, transforma a ação em alocação alvo e devolve essa distribuição juntamente com α_t . Somente depois dessa resposta o escalonador prossegue com a aplicação dos recursos.

Esse encadeamento pode ser representado por:

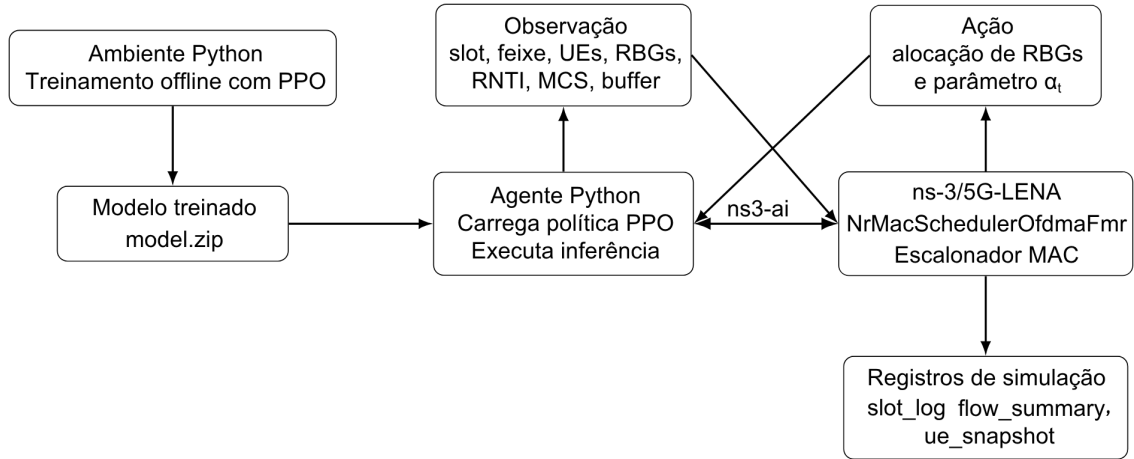


Figura 13: Integração entre o escalonador IA-FMR no ns-3/5G-LENA, a memória compartilhada do ns3-ai e o processo Python responsável pela inferência da política PPO.

$$\begin{aligned}
 \text{ns-3/5G-LENA} &\rightarrow \text{dados da decisão} \rightarrow \text{ns3-ai} \\
 &\rightarrow \text{processo Python} \rightarrow \text{estado} \rightarrow \text{política PPO} \\
 &\rightarrow \text{alocação alvo e } \alpha_t \rightarrow \text{ns3-ai} \rightarrow \text{camada MAC}.
 \end{aligned} \tag{4.43}$$

A alocação alvo não deve ser confundida com a alocação aplicada. A primeira é a distribuição inteira derivada diretamente da ação da política. A segunda resulta da tentativa de realizar essa distribuição dentro dos procedimentos do escalonador 5G-LENA.

Como a comunicação é síncrona, o tempo de execução do processo Python aumenta o tempo de parede necessário para executar a simulação. Esse tempo não é acrescentado ao relógio simulado como atraso experimentado pelos pacotes. A integração demonstra, portanto, o funcionamento conjunto entre política e simulador, mas não constitui uma avaliação de viabilidade em tempo real. O custo de inferência é tratado como limitação e possibilidade de avaliação futura.

4.5.2 Construção e Comunicação da Observação

As informações transferidas pela memória compartilhada não formam diretamente o estado de 29 componentes. Para distinguir os dois conceitos, denomina-se *observação da interface* a mensagem bruta enviada pelo ns-3/5G-LENA e *estado da política* o vetor efetivamente fornecido à rede neural.

No instante t , a observação da interface pode ser representada, de forma simplificada, por:

$$\mathbf{o}_t = \left[slot_t, beam_t, N_t, B_t, \{rnti_i, mcs_{i,t}, q_{i,t}\}_{i=1}^{N_t} \right], \quad (4.44)$$

em que $slot_t$ identifica o *slot* associado à decisão, $beam_t$ identifica o contexto de feixe, N_t informa a quantidade de UEs elegíveis e B_t a quantidade de recursos disponível. Para cada UE, são enviados o RNTI, o MCS de *downlink* e a demanda de buffer $q_{i,t}$.

Na implementação utilizada, essa demanda corresponde à ocupação informada ao escalonador a partir das estruturas RLC/LCG e pode incluir dados aguardando transmissão, retransmissão e informação de estado. Ela é necessária ao tratamento realizado no lado do escalonador, mas não integra diretamente os 29 componentes fornecidos à política.

O processo Python reconstrói o estado da Equação 4.12 combinando a informação corrente da interface com os históricos preservados internamente. A origem de cada componente é resumida na Tabela 10.

Tabela 10: Origem das informações utilizadas na reconstrução do estado durante a inferência.

Componente	Origem	Tratamento
$\mathbf{m}_t$	MCS corrente dos UEs elegíveis	Normalização pelo maior MCS e preenchimento das posições restantes com zero
$\mathbf{b}_{t-1}$	Alocação alvo anterior	Conversão das quantidades alvo em frações da distribuição
$\mathbf{x}_{t-1}$	MCS e alocação alvo da decisão anterior	Estimativa compatível com a escala do ambiente de treinamento
J_{t-1}	Alocação alvo anterior	Índice de Jain sobre o vetor posicional preservado
α_{t-1}	Saída anterior da política	Preservação do coeficiente para a decisão seguinte

A reconstrução de $\mathbf{x}_{t-1}$ merece atenção. A interface não retorna ao Python a vazão fim a fim efetivamente alcançada por cada UE na decisão anterior. O processo utiliza, portanto, o MCS e a alocação alvo anterior para reproduzir a mesma aproximação normalizada utilizada no treinamento. Essa informação existe apenas para compor o estado da política. As métricas sistêmicas de vazão são calculadas posteriormente a partir dos fluxos simulados e possuem significado diferente.

O mesmo vale para J_{t-1} . O escalar do estado é calculado a partir da representação posicional da alocação alvo. Quando $N_t < N$, as posições inativas permanecem com zero e participam dessa representação de dimensão fixa. Esse valor interno não deve

ser confundido com o Índice de Jain reportado na avaliação sistêmica, que é calculado sobre a alocação efetivamente aplicada aos UEs considerados na decisão. A distinção é especialmente relevante na avaliação com 3, 5 e 7 UEs.

As demais informações da observação da interface, como RNTI, identificador do *slot*, contexto de feixe e demanda de buffer, auxiliam a comunicação e a aplicação da decisão, mas não ampliam o vetor de entrada da política.

4.5.3 Identidade dos UEs, Posições da Política e Inferência

A política foi treinada com $N = 9$ posições. Quando $N_t < N$, os UEs elegíveis ocupam as primeiras N_t posições e as posições restantes são preenchidas com zeros. Esse procedimento preserva a dimensão esperada pela rede neural e permite utilizar a mesma arquitetura com conjuntos menores de usuários.

As posições, entretanto, não são identificadores permanentes. A identidade do UE é mantida pelo *Radio Network Temporary Identifier* (RNTI) enviado pela interface. Depois que a política produz a alocação alvo, cada quantidade é associada novamente ao RNTI correspondente antes de retornar ao escalonador.

Embora a rede neural sempre produza nove logits, somente os N_t logits associados às posições ativas participam da *softmax*. Os logits das posições inativas são ignorados na normalização. Dessa forma, o preenchimento com zeros preserva a dimensão da rede sem forçar a atribuição de recursos a posições sem UE.

A inferência é determinística. A política carregada permanece congelada e a ação utilizada corresponde à decisão determinística da distribuição aprendida pelo PPO. O escalar u_t é transformado em α_t pelo mesmo procedimento empregado no treinamento.

Os históricos de alocação e vazão são mantidos por posição, e não por RNTI. Consequentemente, se a composição ou a ordem do conjunto elegível mudar entre duas decisões, uma posição pode carregar temporariamente o histórico associado a outro usuário. A associação da alocação corrente ao UE continua correta porque o RNTI acompanha a resposta, mas o componente histórico do estado pode apresentar esse desalinhamento. Uma implementação que indexasse os históricos diretamente pelo RNTI eliminaria essa dependência posicional.

Nas execuções avaliadas, não foram observadas reordenações quando o conjunto elegível permaneceu inalterado. Considerando as posições comparáveis entre decisões sucessi-

vas, o desalinhamento do histórico correspondeu a apenas 0,041% dos casos. A associação da decisão corrente ao UE permaneceu preservada pelo RNTI. Esse resultado caracteriza as execuções analisadas e não elimina a limitação estrutural do armazenamento por posição.

Uma implementação em que os históricos sejam indexados pelo próprio RNTI eliminaria essa dependência posicional e preservaria diretamente a correspondência entre as informações correntes e históricas de cada UE.

4.5.4 Da Alocação Alvo à Alocação Aplicada na Camada MAC

A alocação alvo $\tilde{\mathbf{r}}_t$ orienta a concessão dos recursos, mas o escalonador precisa convertê-la em uma distribuição realizável na camada MAC. Na implementação utilizada, quando a concessão é possível, o procedimento tenta inicialmente atribuir uma unidade de recurso a cada UE elegível. Os recursos restantes são distribuídos buscando aproximar a quantidade concedida do alvo correspondente a cada usuário.

Se ainda houver recursos residuais, a ordem de concessão considera a quantidade já atribuída, o MCS e o RNTI, de forma determinística. Em seguida, as concessões precisam resultar em blocos de transporte válidos. Concessões que produzam blocos inferiores a 10 bytes são removidas e os respectivos recursos retornam ao conjunto disponível para redistribuição.

Esse limiar de 10 bytes pertence exclusivamente ao procedimento da camada MAC. Ele não é parâmetro aprendido, não constitui requisito de QoS e não possui relação com $T_{\min} = 0,40$ Mbps.

Como resultado, a alocação aplicada pode diferir da alocação alvo:

$$\mathbf{r}_t \neq \tilde{\mathbf{r}}_t. \quad (4.45)$$

A diferença não significa que a decisão da política tenha sido descartada. A alocação alvo define a referência utilizada pelo procedimento de concessão, enquanto a alocação aplicada incorpora as restrições e validações necessárias à camada MAC.

Também é importante observar que a política não recebe de volta a alocação aplicada para reconstruir seus históricos. O processo Python atualiza $\mathbf{b}_t$, $\mathbf{x}_t$ e J_t a partir da alocação alvo que acabou de produzir. Assim, as correções realizadas posteriormente pela camada MAC afetam o sistema simulado, mas não retroalimentam diretamente o estado histórico

da política na decisão seguinte.

As métricas sistêmicas refletem o funcionamento conjunto da política e desses procedimentos de aplicação. Como não foi realizada uma ablação que isolasse cada componente, os resultados não são atribuídos separadamente aos logits, a α_t ou à tentativa inicial de concessão de uma unidade por UE. Por essa razão, as avaliações registram separadamente alocação alvo e alocação aplicada e comparam as duas no Capítulo 6.

4.6 Considerações Finais

Este capítulo apresentou o IA-FMR como um mecanismo de escalonamento que combina uma política aprendida, a conversão de sua ação contínua em uma alocação alvo inteira e um procedimento de integração com a camada MAC do ns-3/5G-LENA.

A principal diferença em relação à formulação anterior do FMR é a introdução explícita do coeficiente α_t como parte da ação. Na configuração final desta tese, esse coeficiente é obtido por uma transformação sigmoide no domínio completo de ponderação $[0,1]$. A formulação permite que um valor seja produzido a cada decisão, mas não impõe que a política explore amplamente todo esse domínio. O comportamento efetivamente aprendido é analisado no Capítulo 6.

A recompensa distingue três funções. O Índice de Jain representa a uniformidade da alocação de RBGs. A eficiência E_t compara a vazão estimada da decisão com uma referência construída a partir do UE de maior eficiência espectral. O termo de déficit utiliza $T_{\min} = 0,40$ Mbps apenas como referência interna do treinamento. Nenhum desses elementos constitui, isoladamente, garantia de QoS ou de vazão mínima.

O treinamento é realizado *offline* com PPO. Uma política independente é obtida para cada largura de banda e seus parâmetros permanecem congelados durante as avaliações. A escolha do PPO é motivada pelo espaço de ações contínuo e pelo mecanismo de *clipping*, sem que a tese estabeleça superioridade do algoritmo em relação a outras alternativas de aprendizado por reforço.

Na integração sistêmica, a política não observa diretamente filas, perdas, atraso ou informações da aplicação. O estado é reconstruído a partir do MCS corrente e dos históricos simplificados derivados da alocação alvo anterior. A camada MAC pode modificar essa alocação ao realizar as concessões, e essas modificações não são retroalimentadas diretamente no histórico da política. Essa separação entre alocação alvo e aplicada é

considerada explicitamente na metodologia e na análise dos resultados.

O Capítulo 5 apresenta os cenários, as métricas e o procedimento experimental utilizados para avaliar o mecanismo. Questões que pertencem à configuração das simulações, ao tratamento estatístico e à instrumentação dos experimentos são detalhadas naquele capítulo.

5 Metodologia Experimental

Este capítulo descreve a metodologia utilizada para avaliar o IA-FMR no ns-3/5G-LENA. A avaliação é organizada em três perspectivas: variação da largura de banda, aplicação da mesma arquitetura a diferentes números de UEs e disponibilidade temporal em tráfego sintético de vídeo. Em todas as avaliações sistêmicas, a política correspondente já foi treinada e é utilizada somente para inferência, com seus parâmetros congelados. A recompensa definida no Capítulo 4 não é recalculada no ns-3 e não ocorre atualização da política durante as simulações.

Os parâmetros empregados no treinamento final foram definidos durante o desenvolvimento da proposta descrita no Capítulo 4. Até a configuração adotada nesta tese, diferentes combinações dos parâmetros do PPO, da transformação da ação e da função de recompensa foram submetidas a sucessivas execuções de treinamento, análise e reteste. Esse processo teve caráter de calibração da proposta e não foi estruturado como um experimento formal de otimização de hiperparâmetros. Por essa razão, os valores utilizados caracterizam a configuração selecionada para a avaliação e não são tratados como parâmetros ótimos ou universais.

Também é necessário distinguir a aleatoriedade do treinamento daquela associada à avaliação sistêmica. Para cada uma das doze larguras de banda foi realizada uma execução final de treinamento com a semente 211382627, produzindo uma política específica para aquela configuração de recursos. Após o treinamento, essas políticas permaneceram congeladas. As simulações no ns-3/5G-LENA utilizam, por sua vez, uma lista de 100 sementes sistêmicas, comum aos escalonadores comparados em cada avaliação. Portanto, essas 100 sementes caracterizam a variabilidade das realizações do cenário simulado e não a variabilidade entre treinamentos independentes.

5.1 Planejamento da Avaliação Experimental

Na Avaliação por Largura de Banda, IA-FMR, RR, PF e MR são comparados em doze larguras de banda da FR1. Na Avaliação de Generalização Estrutural por Número de UEs, a política para a banda de 40 MHz treinada com uma arquitetura de nove posições é aplicada a conjuntos com 3, 5, 7 e 9 usuários. Na Avaliação de Disponibilidade Temporal em Tráfego Sintético de Vídeo, os quatro escalonadores são comparados utilizando a banda de 100 MHz, com seis UEs e tráfego sintético baseado em um modelo de geração de vídeo disponibilizado no 5G-LENA.

O IA-FMR utiliza uma política específica para cada largura de banda. Nas avaliações com menos de nove UEs, as posições não ocupadas do estado são preenchidas com zeros e apenas os logits associados aos usuários elegíveis participam da conversão da ação em recursos, conforme descrito no Capítulo 4. A avaliação com tráfego sintético de vídeo reutiliza a política treinada para 100 MHz. Nenhuma das políticas é treinada a partir das métricas temporais utilizadas nessa avaliação.

A Figura 14 resume o fluxo experimental. As políticas previamente treinadas são selecionadas segundo a largura de banda e integradas ao ns-3/5G-LENA pelo ns3-ai (Henderson *et al.*, 2008; Koutlia *et al.*, 2022; Yin *et al.*, 2020). Nas execuções do IA-FMR, falhas de comunicação ou respostas inválidas interrompem a execução, sem substituição por outro escalonador. Os registros são consolidados por semente antes do cálculo das estatísticas, de modo que UEs, fluxos, pacotes e unidades de uma mesma execução não sejam tratados como repetições independentes.

A Tabela 11 resume as três avaliações e suas respectivas finalidades. A primeira examina como o comportamento dos escalonadores se modifica quando a quantidade de recursos disponíveis varia. A segunda verifica se a representação de dimensão fixa do IA-FMR pode ser utilizada com cardinalidades menores do que a dimensão máxima prevista pela política. A terceira examina se diferenças produzidas pelo escalonamento aparecem em métricas de entrega temporal no nível da aplicação, mesmo sem essas métricas integrarem o estado, a ação ou a recompensa do IA-FMR.

5.2 Ambiente e Parâmetros Comuns das Simulações

As simulações utilizam o ns-3.46 com o módulo 5G-LENA (Henderson *et al.*, 2008; Koutlia *et al.*, 2022), uma gNB e tráfego de *downlink*. O tráfego é gerado por um *host* remoto

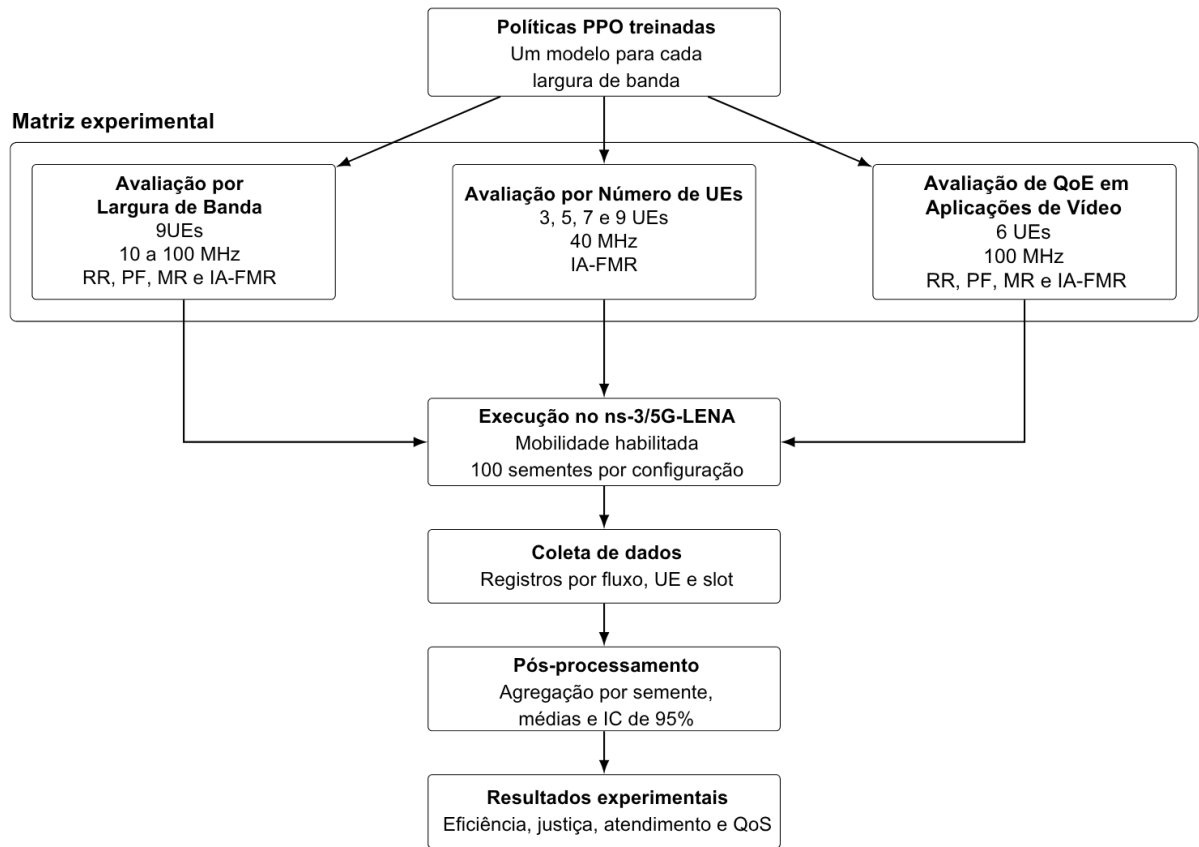


Figura 14: Fluxo da metodologia experimental, desde a seleção das políticas previamente treinadas até a execução dos três cenários, a coleta dos registros e o tratamento estatístico dos resultados.

modelado no ns-3 e conectado ao núcleo da rede simulada. Esse *host* atua como fonte dos fluxos, enquanto os UEs são seus destinos. Os dados percorrem a infraestrutura de núcleo até a gNB e, a partir dela, são transmitidos aos UEs pela interface rádio.

A frequência central é 4 GHz, a potência de transmissão da gNB é 43 dBm e a numerologia é $\mu = 1$, com espaçamento entre subportadoras de 30 kHz e *slots* de 0,5 ms. O padrão TDD é:

$$DL | DL | DL | DL | UL | DL | DL | DL | DL | UL | .$$

Cada ciclo contém dez *slots*, dos quais oito são destinados ao *downlink* e dois ao *uplink*. Como cada *slot* dura 0,5 ms, o padrão se repete a cada 5 ms. O mecanismo estudado atua somente nos *slots* de *downlink*, nos quais os recursos disponíveis são distribuídos entre os UEs elegíveis.

O ambiente de propagação segue o cenário *Urban Micro* (UMi) e os modelos de canal definidos pelo 3GPP disponibilizados no 5G-LENA. O somreamento é desativado nas

Tabela 11: Resumo das avaliações experimentais.

Avaliação	Escalonadores	Finalidade
Avaliação por Largura de Banda	RR, PF, MR e IA-FMR	Examinar vazão, justiça, atendimento e ocupação das filas sob diferentes quantidades de recursos.
Avaliação de Generalização Estrutural por Número de UEs	IA-FMR	Aplicar a mesma arquitetura de 40 MHz a 3, 5, 7 e 9 usuários, dentro da dimensionalidade prevista pela política.
Avaliação de Disponibilidade Temporal em Tráfego Sintético de Vídeo	RR, PF, MR e IA-FMR	Analisar entrega dentro do prazo, congelamentos inferidos e satisfação temporal dos usuários.

avaliações. Essa configuração remove uma fonte de variação lenta da potência recebida associada a obstáculos de maior escala. Os demais efeitos representados pelo modelo de canal permanecem sujeitos à configuração e às realizações do simulador.

A adaptação de enlace utiliza a Tabela MCS 2. O modelo de erro empregado é o EESM IR T2. O *Exponential Effective SINR Mapping* (EESM) reúne as condições observadas nos recursos atribuídos em uma SINR efetiva utilizada pelo simulador no modelo de decodificação. A indicação IR refere-se à parametrização associada à redundância incremental e T2 identifica a configuração vinculada à Tabela MCS 2.

A Tabela 12 reúne os parâmetros compartilhados pelas três avaliações. Configurações de antenas, alturas, posicionamento, mobilidade, largura de banda, número de UEs, tráfego e duração são apresentadas nas subseções específicas, porque esses elementos não são idênticos em todos os cenários.

5.2.1 Larguras de Banda e Quantidade de Recursos

Foram avaliadas as bandas de 10, 15, 20, 25, 30, 40, 50, 60, 70, 80, 90 e 100 MHz na FR1 (3GPP, 2026b). Na configuração utilizada, o agrupamento tratado como RBG pelo IA-FMR contém um único PRB. A mesma granularidade foi adotada no ambiente de treinamento, de modo que:

$$1 \text{ RBG} = 1 \text{ PRB.} \quad (5.1)$$

Assim, a quantidade de unidades usada no treinamento coincide com a quantidade

Tabela 12: Parâmetros comuns das simulações sistêmicas.

Parâmetro	Configuração
Simulador	ns-3.46 com 5G-LENA
Topologia e direção	Uma gNB, múltiplos UEs e <i>downlink</i>
Frequência e potência	4 GHz e 43 dBm
Numerologia	$\mu = 1$, SCS de 30 kHz e <i>slot</i> de 0,5 ms
Padrão TDD	Oito <i>slots</i> de <i>downlink</i> e dois de <i>uplink</i> por ciclo
Canal	UMi, modelos 3GPP e sombreamento desativado
Adaptação de enlace	Tabela MCS 2 e EESM IR T2
Sementes sistêmicas	100 valores por configuração, comuns aos mecanismos comparados

disponível na simulação sistêmica. Como esse total varia entre as bandas, foi treinada uma política para cada configuração.

Tabela 13: Quantidade de PRBs considerada para as larguras de banda FR1 avaliadas com SCS de 30 kHz.

Largura de banda (MHz)	Número de PRBs
10	24
15	38
20	51
25	65
30	78
40	106
50	133
60	162
70	189
80	217
90	245
100	273

5.3 Cenários Experimentais

5.3.1 Avaliação por Largura de Banda

A primeira avaliação compara RR, PF, MR e IA-FMR quando a quantidade de recursos de rádio é modificada pela largura de banda. O cenário contém uma gNB e nove UEs.

As posições iniciais seguem um perfil fixo com raios de 20, 30, 40, 60, 80, 100, 120, 140 e 150 m e ângulos de 0° , 30° , 60° , 90° , 120° , 150° , 180° , 240° e 300° . A gNB e os UEs são posicionados a 1,5 m e utilizam arranjos 1×1 com elementos isotrópicos e formação de feixe pelo caminho direto.

A cardinalidade de nove UEs deve ser interpretada como a dimensão do conjunto de usuários concorrentes considerada no cenário experimental, e não como uma estimativa da quantidade total de dispositivos que podem estar associados a uma célula 5G. Em uma rede operacional, a população de UEs conectados pode ser substancialmente maior que o conjunto efetivamente elegível para receber recursos em uma determinada decisão de escalonamento, uma vez que a disputa naquele instante depende, entre outros fatores, da existência de tráfego pendente e da elegibilidade do usuário para a transmissão. Portanto, a cardinalidade considerada nesta tese caracteriza um conjunto controlado de usuários que competem pelos recursos de *downlink*, e não um limite de capacidade da célula ou do ns-3/5G-LENA.

O valor de nove UEs foi inicialmente adotado como configuração de referência no trabalho que antecedeu esta tese e deu origem à formulação do IA-FMR. Essa escolha não resultou de um dimensionamento analítico destinado a estabelecer uma cardinalidade ótima ou representativa de todas as redes 5G. Ao longo do desenvolvimento, entretanto, ela foi mantida como referência experimental, preservando a comparabilidade entre as diferentes versões da proposta e passando também a definir a dimensionalidade máxima da representação utilizada pela política. Para reduzir a dependência das conclusões em relação a essa cardinalidade específica, a Avaliação de Generalização Estrutural por Número de UEs aplica posteriormente a mesma arquitetura a conjuntos com 3, 5, 7 e 9 UEs.

Após a inicialização, cada UE realiza um passeio aleatório implementado no cenário. A velocidade é sorteada entre 0,5 e 1,5 m/s, a direção é renovada a cada atualização e a distância de referência é 5 m. O intervalo de atualização é $5/1,5 \approx 3,33$ s. O movimento permanece no quadrado $[-200, 200] \times [-200, 200]$ m. Candidatos externos são limitados à borda. As sementes sistêmicas controlam as variáveis pseudoaleatórias da execução, incluindo aquelas associadas à mobilidade, enquanto as posições iniciais permanecem definidas pelo perfil apresentado.

Cada UE mantém quatro fluxos UDP, totalizando 36 fluxos. Os pacotes possuem 3000 bytes. A geração de tráfego inicia em 0,4 s, encerra em 30,4 s e a simulação termina no mesmo instante. Dessa forma, não é acrescentado período de drenagem após o encerramento das aplicações nesta avaliação. O RLC opera em modo UM. O HARQ não foi

alterado em relação à configuração do 5G-LENA utilizada na campanha.

Para uma taxa λ em pacotes por segundo, a carga oferecida por fluxo é:

$$L_{\text{fluxo}} = \frac{8S\lambda}{10^6}, \quad (5.2)$$

em que $S = 3000$ bytes. A carga total oferecida é:

$$L_{\text{total}} = N_f L_{\text{fluxo}}, \quad (5.3)$$

com $N_f = 36$. Os 36 fluxos permanecem configurados durante toda a execução. O que se modifica entre as fases é a taxa de geração de pacotes. A Tabela 14 apresenta os cinco intervalos utilizados. A taxa permanece constante dentro de cada fase e muda em degraus nas transições.

Tabela 14: Fases de tráfego utilizadas na Avaliação por Largura de Banda.

Fase	Intervalo (s)	λ (pacotes/s)	Carga por fluxo (Mbps)	Carga total (Mbps)
Baixa carga	0,4 a 6,4	200	4,8	172,8
Aumento de carga	6,4 a 12,4	350	8,4	302,4
Primeira rajada	12,4 a 18,4	550	13,2	475,2
Recuperação	18,4 a 24,4	250	6,0	216,0
Segunda rajada	24,4 a 30,4	650	15,6	561,6

A carga oferecida representa o volume de dados produzido pelas aplicações e pode ser superior à capacidade de entrega da célula. Quando isso ocorre, dados permanecem pendentes nas filas. Seja $Q_i(t)$ o volume pendente associado ao UE i . O *backlog* agregado é:

$$Q_{\text{agg}}(t) = \sum_{i=1}^N Q_i(t). \quad (5.4)$$

O perfil de tráfego foi definido para produzir diferentes níveis de carga ao longo de uma mesma execução e manter competição por recursos entre os UEs. Não é utilizada uma métrica específica de desperdício de recursos. Decisões em que um UE não apresenta demanda pendente são tratadas pelo critério de elegibilidade do escalonador e não são

classificadas como ausência de atendimento.

As doze larguras de banda são submetidas às mesmas cinco fases e à mesma lista de 100 sementes sistêmicas. Em cada largura de banda, o IA-FMR utiliza a política final treinada especificamente para a quantidade correspondente de recursos. Essa política permanece congelada nas 100 execuções e não é novamente treinada para cada semente do ns-3.

5.3.2 Avaliação de Generalização Estrutural por Número de UEs

A segunda avaliação examina a aplicação da mesma arquitetura do IA-FMR a diferentes cardinalidades dentro da dimensão prevista pela política. A largura de banda é fixada em 40 MHz e são considerados conjuntos com 3, 5, 7 e 9 UEs. A política utilizada foi treinada com $N = 9$ posições. Nas configurações menores, as posições não ocupadas são preenchidas com zeros e somente os logits associados aos UEs elegíveis participam da transformação em recursos.

Essa avaliação não mede escalabilidade para cardinalidades superiores à dimensão da rede neural. Também não avalia custo computacional em função do número de UEs. Seu objetivo é verificar se a representação de dimensão fixa permite utilizar a política com conjuntos menores sem alterar seus parâmetros ou sua arquitetura. Por essa razão, somente o IA-FMR é considerado nesta avaliação.

Os cenários utilizam as primeiras N posições do perfil espacial definido na avaliação anterior. A mobilidade segue a mesma configuração. Cada UE mantém quatro fluxos UDP de 3000 bytes e as cinco fases da Tabela 14. Portanto, o número de fluxos e a carga oferecida dependem da cardinalidade:

$$N_f = 4N, \quad (5.5)$$

$$L_{\text{total}}(N, \lambda) = 4N \left(\frac{8S\lambda}{10^6} \right). \quad (5.6)$$

As configurações com 3, 5, 7 e 9 UEs possuem 12, 20, 28 e 36 fluxos, respectivamente. Dessa forma, a alteração do número de UEs também modifica a carga total oferecida e o conjunto de posições espaciais presentes no cenário. Os resultados não são interpretados como efeito isolado da cardinalidade sobre o desempenho.

As configurações com 3, 5 e 7 UEs são executadas em campanha específica com a mesma lista de 100 sementes sistêmicas. Para 9 UEs, utiliza-se a configuração de 40 MHz da Avaliação por Largura de Banda, que possui os mesmos parâmetros aplicáveis a esse ponto. Em todos os casos, a política de 40 MHz permanece congelada.

5.3.3 Avaliação de Disponibilidade Temporal em Tráfego Sintético de Vídeo

A terceira avaliação examina a entrega temporal de unidades sintéticas de vídeo. O tráfego utiliza o gerador genérico de vídeo disponível no 5G-LENA e parâmetros associados a modelos de mídia e XR do 3GPP (3GPP, 2024a,b, 2022). O cenário é utilizado como uma condição de tráfego com requisito temporal no pós-processamento. Não são simulados conteúdo visual, captura de pose, seleção de *viewport*, renderização, codificação, decodificação, *playout buffer* adaptativo ou percepção subjetiva. Assim, as métricas descritas nesta seção não constituem uma avaliação completa de QoE.

Tabela 15: Configuração física e espacial da avaliação com tráfego sintético de vídeo.

Parâmetro	Configuração
Largura de banda e UEs	100 MHz e 6 UEs
Alturas	10 m para a gNB e 1,5 m para os UEs
Posicionamento inicial	Raio uniforme entre 20 e 150 m e ângulo uniforme entre -60° e $+60^\circ$
Mobilidade	Passeio aleatório restrito ao setor, entre 0,5 e 1,5 m/s
Arranjo da gNB	8×8 , orientação de 0°
Arranjo dos UEs	2×2 , orientação de 180°
Elementos de antena	ThreeGppAntennaModel
Feedback MIMO e <i>rank</i>	Feedback habilitado e <i>rank</i> máximo 1

A distância e o ângulo iniciais são sorteados independentemente. A cada atualização de mobilidade, uma velocidade e uma direção candidatas são sorteadas. O deslocamento é aplicado quando a nova posição permanece no setor anular definido pelos limites de distância e ângulo. Caso contrário, novas direções são testadas e o UE pode manter a posição anterior.

O tamanho médio nominal resulta da combinação de 30 Mbps com 60 unidades/s. O gerador varia o tamanho ao redor dessa média segundo o modelo adotado, com limites de 50% e 150% do valor médio, e acrescenta variação temporal ao intervalo nominal de 16,67 ms. Cada unidade recebe um cabeçalho `SeqTsSizeHeader`, que registra sequência,

Tabela 16: Configuração do tráfego e das execuções da avaliação com tráfego sintético de vídeo.

Parâmetro	Configuração
Fluxos	Um fluxo UDP de <i>downlink</i> por UE
Taxa nominal	30 Mbps por UE e 180 Mbps no total
Frequência nominal	60 unidades por segundo
Tamanho médio nominal	62.500 bytes por unidade
Período de geração	0,4 a 30,4 s
Quantidade nominal	Aproximadamente 1.800 unidades por UE
Encerramento e drenagem	Simulação até 32,4 s, com 2 s após o fim da geração
Pilha de transporte	UDP, RLC UM e HARQ habilitado
Perfil do fluxo	NGBR_LOW_LAT_EMBB
Política IA-FMR	Política de 100 MHz e nove posições, aplicada a seis UEs
Repetições sistêmicas	100 sementes comuns aos quatro escalonadores

instante de geração e tamanho. A recepção é registrada quando o *PacketSink* entrega a unidade identificada. Na ausência desse evento até 32,4 s, a unidade é classificada como não recebida. O critério considera, portanto, a entrega completa da unidade no nível da aplicação.

O RLC opera em UM e o HARQ permanece habilitado. O perfil NGBR_LOW_LAT_EMBB define o *bearer* empregado pelo 5G-LENA, mas não determina a taxa nominal nem os prazos utilizados na análise. No pós-processamento, cada unidade é avaliada considerando prazos de 60 e 80 ms, identificados nas figuras pela nomenclatura PDB adotada no cenário. O prazo de 60 ms é utilizado como referência principal e o prazo de 80 ms permite examinar a sensibilidade dos resultados a um limite temporal menos restritivo.

Esses prazos não integram a decisão do IA-FMR. A política não recebe identificadores das unidades, prazo de entrega, estado de reprodução ou informação sobre congelamentos. O valor $T_{\min} = 0,40$ Mbps utilizado no treinamento também não possui relação com os prazos ou com a taxa nominal deste cenário.

Para cada semente sistêmica, os quatro escalonadores utilizam a mesma realização das variáveis aleatórias correspondentes ao cenário. A política IA-FMR empregada nas 100 repetições é a mesma política congelada de 100 MHz.

5.4 Métricas e Tratamento Estatístico

As métricas sistêmicas são obtidas a partir dos resumos por fluxo e dos registros de escalonamento por UE e *slot*. Os resumos por fluxo fornecem as informações utilizadas no cálculo de vazão, atraso e perdas. Os registros de escalonamento contêm, entre outros campos, o RNTI, a demanda pendente, a alocação alvo, a alocação aplicada e, para o IA-FMR, o valor de α_t . As métricas são calculadas inicialmente por semente sistêmica e somente depois consolidadas entre as repetições.

5.4.1 Métricas Sistêmicas

As métricas sistêmicas caracterizam a vazão alcançada, a região inferior da distribuição das vazões por fluxo, a distribuição dos recursos, o atendimento dos UEs e o comportamento das filas, das perdas e do atraso. A Tabela 17 apresenta uma síntese.

Tabela 17: Métricas sistêmicas utilizadas nas avaliações.

Métrica	Definição
Vazão agregada	Soma das vazões dos fluxos de uma execução.
Percentil 5 da vazão	Percentil 5 das vazões dos fluxos de uma execução.
Índice de Jain dos RBGs	Uniformidade da distribuição dos RBGs aplicados entre os UEs em cada <i>slot</i> .
UEs servidos	Número de UEs que recebem ao menos um RBG aplicado no <i>slot</i> .
Demanda sem alocação	Percentual das ocorrências com demanda pendente nas quais o respectivo UE não recebe RBG no <i>slot</i> .
<i>Backlog</i>	Volume de dados pendente informado ao escalonador.
Atraso	Tempo entre a geração e a recepção dos pacotes entregues.
Perda de pacotes	Fração dos pacotes transmitidos que não são recebidos até o encerramento da execução.

A vazão é calculada a partir dos dados recebidos por cada fluxo. Para o fluxo f ,

$$V_f = \frac{8B_f}{\Delta t 10^6}, \quad (5.7)$$

em que B_f é a quantidade de bytes recebidos durante o período de medição Δt , resultando em Mbps. A vazão agregada da execução corresponde à soma das vazões dos N_f fluxos:

$$V_{\text{agg}} = \sum_{f=1}^{N_f} V_f. \quad (5.8)$$

O percentil 5 da vazão, denotado por P5, é calculado sobre as vazões dos fluxos de cada execução. Ele caracteriza a região inferior da distribuição sem substituir a análise individual dos fluxos. Um valor reduzido de P5 indica que parte das vazões se encontra próxima da extremidade inferior da distribuição.

O atraso é obtido a partir dos pacotes recebidos e corresponde ao tempo decorrido entre sua geração e sua recepção. Consequentemente, o atraso médio caracteriza apenas os pacotes efetivamente entregues. Pacotes não recebidos participam da métrica de perdas, e não da média de atraso. Para P_{tx} pacotes transmitidos e P_{rx} recebidos, a perda é:

$$P_{\text{loss}} = \frac{P_{\text{tx}} - P_{\text{rx}}}{P_{\text{tx}}}. \quad (5.9)$$

As métricas de distribuição e atendimento são calculadas a partir dos registros por UE e *slot*. Seja $A_i(t)$ a quantidade de RBGs efetivamente aplicada ao UE i no *slot* t . O Índice de Jain da distribuição é:

$$J_{\text{RBG}}(t) = \frac{\left(\sum_{i=1}^N A_i(t)\right)^2}{N \sum_{i=1}^N A_i(t)^2}. \quad (5.10)$$

Todos os UEs configurados são mantidos no vetor usado para essa métrica. Para usuários sem recursos aplicados, considera-se $A_i(t) = 0$. Quando não há RBGs aplicados e o denominador da Equação 5.10 é nulo, define-se $J_{\text{RBG}}(t) = 0$. O índice mede a uniformidade da distribuição dos RBGs aplicados. Ele não mede igualdade de vazão e não determina, isoladamente, se um UE permaneceu sem atendimento durante uma sequência de decisões.

Para a semente sistêmica k , o valor utilizado na avaliação corresponde à média entre os *slots* registrados:

$$\bar{J}_{\text{RBG},k} = \frac{1}{|\mathcal{T}_k|} \sum_{t \in \mathcal{T}_k} J_{\text{RBG}}(t). \quad (5.11)$$

O número de UEs servidos em cada *slot* corresponde à quantidade de usuários com $A_i(t) > 0$. Sua média ao longo da execução indica quantos UEs recebem recursos em uma

decisão típica dentro daquela realização.

A métrica de demanda sem alocação considera somente ocorrências nas quais existe demanda pendente. Seja $q_i(t)$ o volume pendente informado ao escalonador para o UE i no *slot* t . O percentual é calculado por:

$$D_{\text{sem aloc.}} = 100 \frac{\sum_t \sum_i \mathbb{I}[q_i(t) > 0 \wedge A_i(t) = 0]}{\sum_t \sum_i \mathbb{I}[q_i(t) > 0]}, \quad (5.12)$$

em que $\mathbb{I}[\cdot]$ é a função indicadora. O denominador contém apenas registros UE-*slot* com demanda pendente. Uma ocorrência nessa métrica indica ausência de alocação naquela decisão apesar da existência de dados pendentes. Uma ocorrência isolada não caracteriza *starvation*, pois esse conceito requer persistência do não atendimento ao longo do tempo.

O campo de demanda utilizado nessa análise corresponde à ocupação do buffer RLC/LCG informada ao escalonador, em bytes. Para as análises temporais, os valores dos UEs são somados e convertidos para Mbit:

$$Q_{\text{agg}}^{\text{Mbit}}(t) = \frac{8}{10^6} \sum_{i=1}^N q_i(t). \quad (5.13)$$

Esse valor representa dados pendentes no ponto da pilha observado pelo escalonador. Ele não corresponde, por si só, à vazão entregue ou à quantidade total de dados gerados pela aplicação.

5.4.2 Comparação entre Alocação Alvo e Alocação Aplicada

Para o IA-FMR, os registros permitem distinguir a alocação alvo $\tilde{\mathbf{r}}_t$, obtida a partir da ação da política, da alocação $\mathbf{r}_t$ efetivamente realizada pela camada MAC, conforme definido no Capítulo 4. Essa distinção é utilizada para verificar em que medida a distribuição solicitada pela política é preservada durante sua aplicação.

O pós-processamento considera três aspectos. O primeiro é a correspondência exata entre os vetores alvo e aplicado. O segundo é a preservação do conjunto de UEs com alocação positiva. O terceiro compara propriedades das duas distribuições, como número de UEs contemplados e Índice de Jain. Essas medidas não alteram a execução e são calculadas a partir dos registros produzidos pelo IA-FMR.

A correspondência exata em uma decisão pode ser representada por:

$$I_{\text{exato}}(t) = \mathbb{I}[\tilde{\mathbf{r}}_t = \mathbf{r}_t]. \quad (5.14)$$

A preservação do conjunto de UEs atendidos é verificada comparando:

$$\tilde{\mathcal{U}}_t^+ = \{i : \tilde{r}_{i,t} > 0\}, \quad \mathcal{U}_t^+ = \{i : r_{i,t} > 0\}. \quad (5.15)$$

Essa análise permite separar diferenças introduzidas na realização da decisão de diferenças já presentes na saída da política. Ela é empregada de forma descritiva no Capítulo 6.

5.4.3 Processamento das Métricas Temporais em 40 MHz

A análise temporal apresentada no Capítulo 6 possui finalidade diferente das estatísticas agregadas sobre 100 sementes. Para visualizar a evolução das decisões ao longo das cinco fases de tráfego, é utilizada uma única realização em 40 MHz, correspondente à semente sistêmica 47889. A mesma semente é utilizada para RR, PF, MR e IA-FMR.

Os registros por UE e *slot* dessa realização são agrupados em intervalos não sobrepostos de 0,1 s. Em cada intervalo, são calculadas as médias das grandezas correspondentes. Para reduzir variações de curta duração na representação gráfica, as curvas utilizam média móvel centrada de três intervalos, equivalente a 0,3 s. Como essa análise utiliza uma única semente, as curvas temporais não possuem intervalo de confiança e não representam a variabilidade entre as 100 execuções.

Seja $\bar{A}_i(b)$ a média de RBGs aplicados ao UE i no intervalo temporal b . Sua participação na distribuição é:

$$p_i(b) = \frac{\bar{A}_i(b)}{\sum_{j=1}^N \bar{A}_j(b)}. \quad (5.16)$$

A partir dessas participações, são calculados o número efetivo de UEs e a maior participação individual:

$$N_{\text{ef}}(b) = \frac{1}{\sum_{i=1}^N p_i(b)^2}, \quad (5.17)$$

$$p_{\text{max}}(b) = \max_{1 \leq i \leq N} p_i(b). \quad (5.18)$$

O valor N_{ef} pode ser interpretado como o número de usuários equivalentes que produziram o mesmo nível de concentração caso possuísem participações iguais. Para nove UEs, valores próximos de 9 correspondem a uma distribuição próxima da uniformidade, enquanto valores menores indicam maior concentração. Já p_{max} identifica diretamente a maior fração dos recursos atribuída a um único UE no intervalo.

O número de UEs servidos é obtido pela contagem dos usuários com alocação positiva e agregado nos mesmos intervalos de 0,1 s. O *backlog* utiliza a soma da demanda pendente dos UEs, convertida para Mbit pela Equação 5.13. Para o IA-FMR, α_t é processado no mesmo eixo temporal. As transições entre as cinco fases de tráfego são indicadas nas figuras para relacionar a evolução temporal das métricas às mudanças da carga oferecida. Essa leitura é descritiva e permanece restrita à realização escolhida.

5.4.4 Métricas de Disponibilidade Temporal no Cenário de Vídeo

Para cada prazo temporal d , uma unidade sintética é classificada em uma entre três categorias: recebida completamente dentro do prazo, recebida após o prazo ou não recebida até o encerramento da simulação. Sejam $N_{prazo,d}$, $N_{tardia,d}$ e N_{nr} essas quantidades. Para cada UE:

$$N_{geradas} = N_{prazo,d} + N_{tardia,d} + N_{nr}. \quad (5.19)$$

A Taxa de Unidades no Prazo (TUP) é definida por:

$$TUP_d = \frac{N_{prazo,d}}{N_{geradas}} \times 100. \quad (5.20)$$

Unidades recebidas após o prazo permanecem classificadas como tardias mesmo que sejam entregues antes do encerramento da simulação. Unidades que não são recebidas até esse encerramento permanecem no denominador da TUP e não contribuem para o numerador. As três categorias permitem distinguir atraso temporal de ausência completa de recepção.

O prazo de 60 ms é utilizado como referência principal com base no limite superior da faixa de latência de *downlink* considerada no modelo de referência adotado (3GPP, 2024b). O prazo de 80 ms é empregado como análise de sensibilidade. Esses limites são aplicados no pós-processamento e não constituem parâmetros da política IA-FMR.

Após a primeira unidade recebida dentro do prazo, uma sequência contínua de unidades que não satisfazem esse prazo é tratada como um evento de congelamento inferido. Sequências que permanecem abertas até o término da observação são contabilizadas como eventos terminais. A utilização do termo *inferido* indica que o evento é reconstruído a partir da sequência temporal das unidades e não observado em um reproduutor de vídeo real.

A frequência de congelamentos é expressa em eventos por minuto. Para um UE que iniciou a reprodução inferida, ela é obtida a partir do número de eventos e do período observado após a primeira unidade recebida no prazo:

$$F_{u,k,d} = \frac{60N_{\text{eventos},u,k,d}}{T_{\text{obs},u,k,d}}, \quad (5.21)$$

em que u identifica o UE, k a semente sistêmica e d o prazo considerado. No pós-processamento, a métrica é consolidada por semente. As 100 sementes participam do cálculo da média e do IC95 da frequência. A métrica admite valor zero e os casos em que nenhum UE inicia a reprodução inferida também permanecem nesse conjunto.

A severidade dos eventos é caracterizada pela duração P95. Em cada semente, os eventos dos seis UEs são reunidos e o percentil 95 das durações é calculado por interpolação linear. Eventos terminais participam desse conjunto. Quando há apenas um evento, o P95 coincide com sua duração. Pelo procedimento definido, uma semente com reprodução iniciada e ausência de eventos receberia P95 igual a zero. Quando nenhum UE inicia a reprodução inferida, o P95 não é definido e essa semente é excluída somente dessa métrica. Dessa forma, a média da duração P95 e seu IC95 são calculados sobre o conjunto de sementes válidas de cada escalonador e de cada prazo. O número de sementes válidas é informado junto aos resultados.

Um UE é classificado como temporalmente satisfeito quando:

$$\text{TUP}_d^{(u)} \geq 99\%. \quad (5.22)$$

O indicador reportado em cada semente corresponde à proporção dos seis UEs que satisfazem esse critério. O limiar de 99% é uma escolha operacional para esta análise e não um requisito normativo nem uma garantia fornecida pelo escalonador.

Essas métricas descrevem propriedades temporais da entrega das unidades no cenário sintético. Como não são simulados reprodução real, decodificação, *playout buffer* ou

adaptação de vídeo, elas não constituem uma caracterização completa da experiência de reprodução ou da QoE do usuário.

5.4.5 Tratamento Estatístico

As métricas agregadas são calculadas inicialmente em cada semente sistêmica. Para as avaliações baseadas nas 100 repetições, os resultados são resumidos por média e intervalo de confiança de 95%. O P5 da vazão é calculado dentro de cada execução sobre os fluxos e, em seguida, os valores obtidos por semente são consolidados. A duração P95 dos congelamentos segue o tratamento específico de sementes válidas definido anteriormente.

O IC95 da média é calculado por:

$$IC_{95\%} = \bar{x} \pm 1,96 \frac{s}{\sqrt{n}}, \quad (5.23)$$

em que $\bar{x}$ é a média dos valores por semente, s é o desvio padrão amostral e n é o número de sementes válidas para a métrica. Para a maior parte das métricas sistêmicas, $n = 100$. Para a duração P95 dos congelamentos, n pode ser menor porque sementes nas quais nenhum UE inicia a reprodução inferida não possuem valor definido para essa métrica.

A mesma lista de sementes sistêmicas é utilizada pelos mecanismos comparados. Essa organização mantém correspondência entre as realizações de uma mesma configuração, mas as comparações apresentadas nesta tese permanecem descritivas. Não são realizados testes de hipótese, intervalos de confiança para diferenças pareadas ou afirmações de significância estatística baseadas na sobreposição dos IC95 individuais.

A análise temporal de 40 MHz constitui uma exceção ao tratamento agregado. Ela utiliza somente a semente 47889 e, portanto, não apresenta IC95. Da mesma forma, as 100 sementes sistêmicas não são utilizadas para construir intervalos de confiança do treinamento. Como cada política final foi obtida a partir de uma única execução de treinamento, não há estimativa de variabilidade entre políticas treinadas com sementes diferentes.

5.5 Considerações Finais

Este capítulo apresentou a metodologia empregada para avaliar o IA-FMR no ns-3/5G-LENA. Foram distinguidas a etapa de calibração realizada durante o desenvolvimento, o

treinamento final das políticas e a avaliação sistêmica com 100 sementes. Também foram descritos os parâmetros comuns, os três cenários experimentais, as métricas e o tratamento estatístico.

A Avaliação por Largura de Banda compara os quatro escalonadores sob diferentes quantidades de recursos. A Avaliação de Generalização Estrutural por Número de UEs examina a utilização da arquitetura do IA-FMR com cardinalidades menores dentro da dimensão prevista. A avaliação com tráfego sintético de vídeo acrescenta métricas de entrega temporal sem incorporá-las ao treinamento da política. O Capítulo 6 apresenta os resultados obtidos nessas três avaliações.

6 Resultados e Discussão

Este capítulo apresenta e discute os resultados do treinamento e das três avaliações experimentais. Inicialmente, analisa-se o treinamento das políticas IA-FMR. Em seguida, os escalonadores são comparados na Avaliação por Largura de Banda, o IA-FMR é examinado na Avaliação de Generalização Estrutural por Número de UEs e, por fim, são apresentados os resultados da Avaliação de Disponibilidade Temporal em Tráfego Sintético de Vídeo.

A Avaliação por Largura de Banda considera conjuntamente vazão, justiça, atendimento dos usuários, perdas, atraso e evolução temporal das decisões e das filas. A Avaliação de Generalização Estrutural examina vazão, justiça, atendimento, desempenho dos fluxos menos favorecidos, perdas e atraso para diferentes cardinalidades. Na avaliação com tráfego sintético de vídeo, são analisadas a entrega das unidades dentro do prazo, a frequência e a duração dos congelamentos inferidos e a satisfação temporal dos usuários.

Os experimentos sistêmicos utilizam 100 sementes de avaliação. As métricas agregadas são calculadas sobre essas execuções, exceto quando a própria definição da métrica estabelece um subconjunto de sementes válidas, como no cálculo da duração P95 dos congelamentos inferidos. A análise é descritiva: as figuras apresentam médias e intervalos de confiança individuais para um nível de confiança de 95%, sem testes de hipótese ou afirmações de significância estatística. As diferenças percentuais reportadas servem para quantificar os contrastes observados no conjunto avaliado.

6.1 Treinamento do IA-FMR

Esta seção apresenta a evolução do treinamento das políticas IA-FMR posteriormente utilizadas nos experimentos sistêmicos. Para cada largura de banda, foi realizada uma única execução de treinamento, com a semente 211382627, durante 600.000 passos. Ao final de cada execução, a política obtida foi exportada e posteriormente utilizada na integração com o ns-3/5G-LENA. Dessa forma, as curvas apresentadas nesta seção correspondem

diretamente aos modelos avaliados nos experimentos.

Como o treinamento não foi repetido com diferentes inicializações, os resultados desta seção não caracterizam a variabilidade do processo de otimização em função da semente de treinamento. A semente 211382627 identifica a instância de treinamento utilizada e permite sua reprodução. Sua finalidade é, portanto, distinta das 100 sementes empregadas posteriormente na avaliação sistêmica, utilizadas para produzir diferentes realizações dos cenários de simulação sob condições pseudoaleatórias comuns aos escalonadores comparados.

Um passo de treinamento corresponde a uma interação entre o agente e o ambiente. A cada passo, o agente recebe o estado corrente, produz uma ação e recebe a recompensa resultante da decisão realizada. Esse processo é repetido até atingir o orçamento de 600.000 passos definido para cada política.

A Figura 15 apresenta a evolução da recompensa média ao longo do treinamento. O eixo horizontal representa o número de passos de treinamento e o eixo vertical apresenta a recompensa média. Cada curva corresponde à política associada a uma das larguras de banda avaliadas.

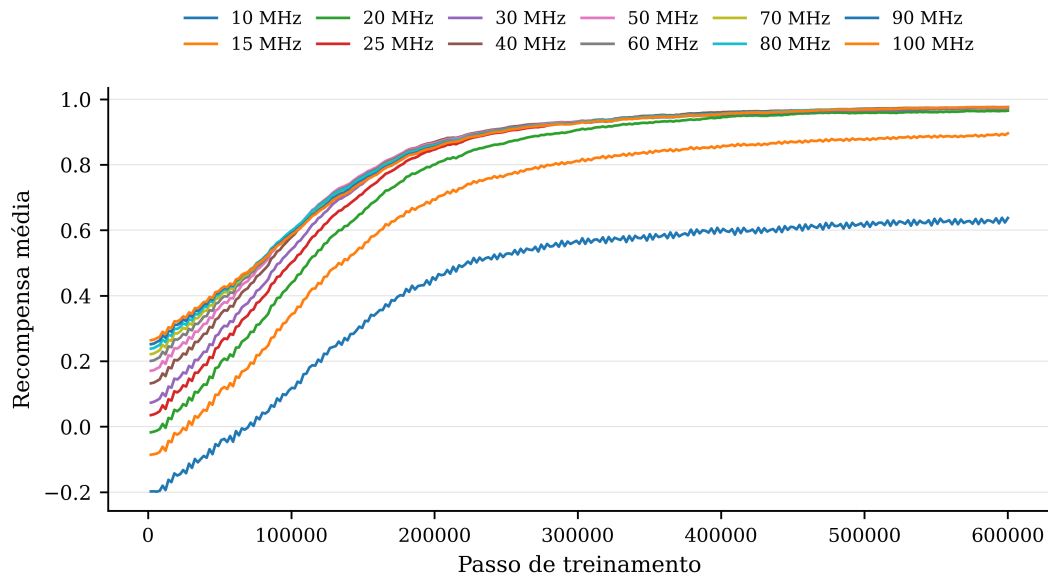


Figura 15: Evolução da recompensa média durante o treinamento das políticas IA-FMR. Cada curva corresponde a uma política treinada durante 600.000 passos com a semente 211382627 e posteriormente utilizada na avaliação sistêmica.

De forma geral, as curvas apresentam aumento da recompensa ao longo do treinamento, com crescimento mais acentuado nas etapas iniciais e progressivamente menor à medida que o número de passos aumenta. A maior parte das políticas passa a ocupar

regiões próximas nas etapas finais, embora as políticas associadas a 90 e 100 MHz permaneçam em patamares inferiores, com diferença mais pronunciada em 90 MHz. Essas diferenças caracterizam as instâncias de treinamento obtidas e, isoladamente, não permitem estabelecer sua causa.

A interpretação dessas curvas deve considerar a composição da função de recompensa. Seu valor resulta da combinação entre o termo associado à justiça na distribuição dos recursos, o termo de eficiência e a penalização por déficit de vazão. Portanto, uma recompensa mais elevada representa uma combinação mais favorável desses componentes segundo o objetivo definido para o treinamento, mas não corresponde diretamente a uma medida isolada de vazão, justiça ou atendimento dos UEs.

Além da recompensa, o comportamento do coeficiente α_t permite observar como a política passa a representar o compromisso entre os termos de justiça e eficiência durante o aprendizado. A Figura 16 apresenta essa evolução para a política de 40 MHz, utilizada como caso representativo para examinar com maior detalhe o comportamento interno da política. O eixo horizontal representa os passos de treinamento e o eixo vertical apresenta α_t . A curva corresponde ao valor médio e a região sombreada delimita os percentis 5 e 95 dos valores observados em cada janela de agregação.

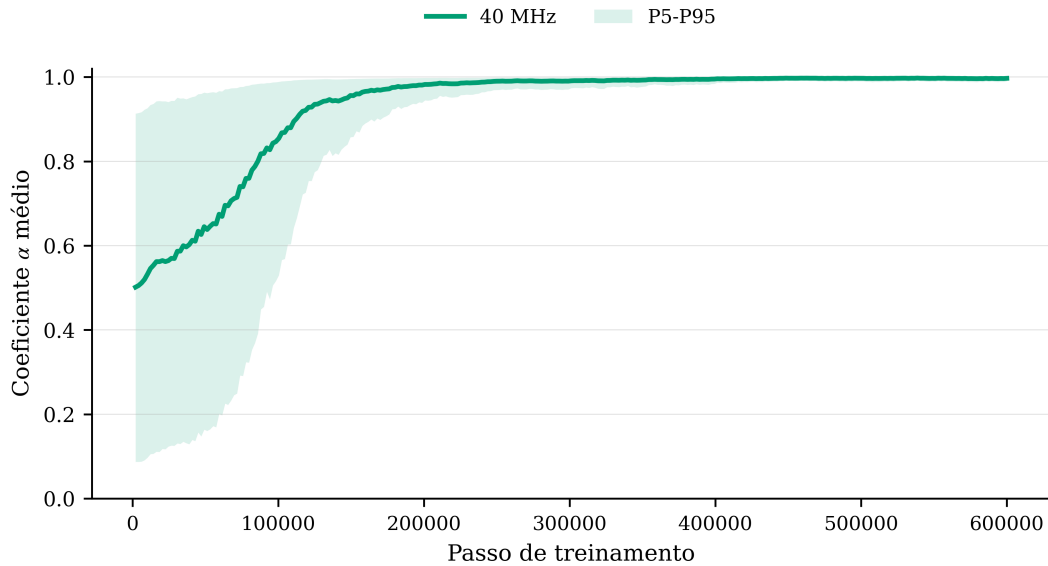


Figura 16: Evolução do coeficiente α_t durante o treinamento da política IA-FMR para 40 MHz. A curva representa o valor médio de α_t e a região sombreada corresponde à faixa entre os percentis 5 e 95 dos valores observados em cada janela de agregação.

A curva mostra que, nas etapas iniciais do treinamento, os valores de α_t apresentam maior dispersão e ocupam uma região ampla do domínio disponível ao coeficiente. À medida que o treinamento avança, o valor médio aumenta e a faixa de dispersão se estreita

progressivamente, concentrando-se em valores elevados nas etapas finais.

A região sombreada da Figura 16 deve ser interpretada como uma medida descritiva da dispersão dos valores produzidos pela política. Em cada janela de agregação, seus limites correspondem aos percentis 5 e 95, de modo que aproximadamente 90% das amostras centrais de α_t encontram-se entre esses limites. Essa faixa não representa um intervalo de confiança da média, nem caracteriza variabilidade entre treinamentos independentes. Ela descreve exclusivamente a distribuição dos valores de α_t produzidos ao longo da única execução de treinamento apresentada.

O comportamento observado indica que a política de 40 MHz, embora disponha do domínio completo definido para o coeficiente, passa a concentrar suas decisões em valores elevados de α_t . Considerando a formulação da recompensa, esse resultado caracteriza uma preferência aprendida por maior ponderação do componente de justiça em relação ao componente de eficiência. Ao mesmo tempo, a figura mostra que a possibilidade de produzir α_t a cada decisão não implica que a política treinada explore amplamente todo o domínio do coeficiente após o aprendizado. Deve-se, portanto, distinguir a capacidade oferecida pela formulação do comportamento efetivamente aprendido pela política.

Essa distinção será retomada posteriormente na avaliação sistêmica, quando o comportamento de α_t será observado durante a execução da política congelada no ns-3/5G-LENA. A comparação permite verificar se a concentração identificada no ambiente de treinamento também se manifesta diante dos estados produzidos pelo simulador, sem pressupor equivalência direta entre as duas condições de execução.

Em conjunto, os resultados desta seção mostram aumento da recompensa ao longo do orçamento de treinamento e estabilização progressiva do comportamento de α_t no caso analisado. Os 600.000 passos correspondem ao orçamento adotado para produzir as políticas posteriormente avaliadas nesta tese. Esses resultados não demonstram convergência para um ótimo global, nem permitem estimar a variabilidade do treinamento diante de outras inicializações. O objetivo desta etapa foi obter as instâncias de política utilizadas na avaliação sistêmica, e não caracterizar estatisticamente o procedimento de otimização entre diferentes sementes de treinamento.

6.2 Avaliação por Largura de Banda

Esta seção compara RR, PF, MR e IA-FMR nas doze larguras de banda definidas no Capítulo 5. Inicialmente, são apresentados os resultados agregados. Em seguida, a con-

figuração de 40 MHz é analisada no domínio temporal para explicar como as decisões de alocação produzem os comportamentos observados.

6.2.1 Vazão Agregada

A Figura 17 apresenta a vazão agregada média da célula em função da largura de banda. O eixo horizontal representa as doze larguras de banda avaliadas, em MHz, e o eixo vertical apresenta a vazão agregada média, em Mbps. Cada curva corresponde a um escalonador e a região sombreada representa o intervalo de confiança de 95% da respectiva média. Como o tráfego é transmitido em *downlink*, a vazão de cada fluxo é calculada a partir dos dados efetivamente recebidos pelo UE de destino. A vazão agregada corresponde à soma dessas vazões e representa, portanto, a taxa total de dados que alcança os UEs, e não a carga de tráfego oferecida pelo *host* remoto.

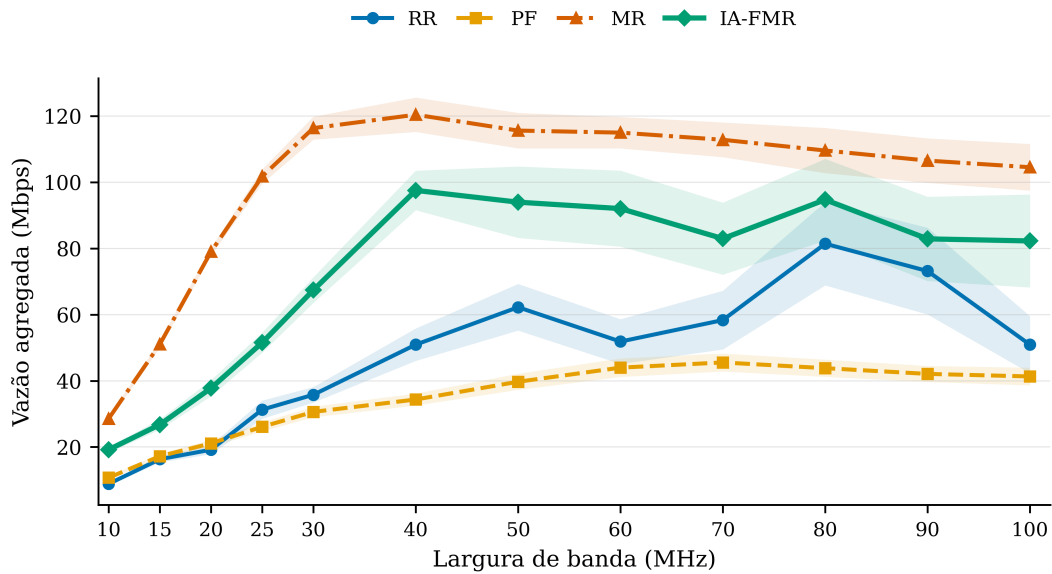


Figura 17: Vazão agregada média em função da largura de banda. As regiões sombreadas representam os intervalos de confiança de 95%.

Em termos dos valores médios, o MR apresenta a maior vazão agregada em todas as larguras de banda avaliadas, comportamento compatível com seu critério de priorizar os UEs com melhores condições instantâneas de transmissão. As médias do IA-FMR situam-se abaixo das observadas para o MR e acima das médias de RR e PF ao longo das configurações avaliadas. Essa ordenação, entretanto, deve ser interpretada de forma descritiva, uma vez que os intervalos de confiança apresentam sobreposição em parte das larguras de banda e não permitem concluir, a partir desta figura isoladamente, que exista separação estatística entre os métodos em todos os pontos.

Em 40 MHz, configuração utilizada posteriormente em análises mais detalhadas, o MR alcança vazão média de 120,40 Mbps e o IA-FMR, 97,53 Mbps. Nessa mesma configuração, RR e PF apresentam médias de 50,93 Mbps e 34,36 Mbps, respectivamente. Assim, considerando exclusivamente os valores médios, a vazão do IA-FMR corresponde a 1,92 vezes a do RR e a 2,84 vezes a do PF, permanecendo 18,99% abaixo da média do MR.

A evolução da vazão com o aumento da largura de banda não é monotônica para todos os escalonadores. No caso do IA-FMR, a vazão aumenta até as configurações intermediárias e apresenta oscilações nas larguras de banda superiores. Comportamentos não monotônicos também são observados, em diferentes magnitudes, para outros escalonadores.

Esse resultado deve ser interpretado considerando que a largura de banda determina a quantidade de recursos disponíveis, mas não estabelece isoladamente a quantidade de dados efetivamente recebidos pelos UEs. A vazão resultante depende também das condições de canal experimentadas pelos usuários, da mobilidade, da evolução das filas, das perdas e das decisões tomadas pelo escalonador ao longo da simulação. Os experimentos realizados não isolam individualmente esses fatores e, portanto, não permitem atribuir o comportamento observado a uma única causa.

As regiões correspondentes aos intervalos de confiança mostram ainda que a incerteza varia entre as configurações e se torna particularmente ampla em algumas das maiores larguras de banda. Em determinados pontos, os intervalos associados ao IA-FMR e a outros escalonadores apresentam sobreposição. Dessa forma, a figura permite caracterizar a tendência e a ordenação das médias, mas não sustenta, por si só, uma afirmação de diferença estatisticamente significativa entre os métodos em todas as larguras de banda. A análise desta métrica é complementada nas subseções seguintes por indicadores relacionados à distribuição dos recursos e ao atendimento dos fluxos.

6.2.2 Justiça na Alocação de RBGs

A Figura 18 apresenta o Índice de Jain calculado sobre os RBGs alocados aos nove UEs. O eixo horizontal representa a largura de banda, em MHz, e o eixo vertical apresenta o valor médio do índice. Cada curva corresponde a um escalonador e as regiões sombreadas representam os intervalos de confiança de 95%. Valores próximos de 1 indicam uma distribuição mais uniforme dos recursos, enquanto valores menores indicam maior concentração da alocação em parte dos usuários.

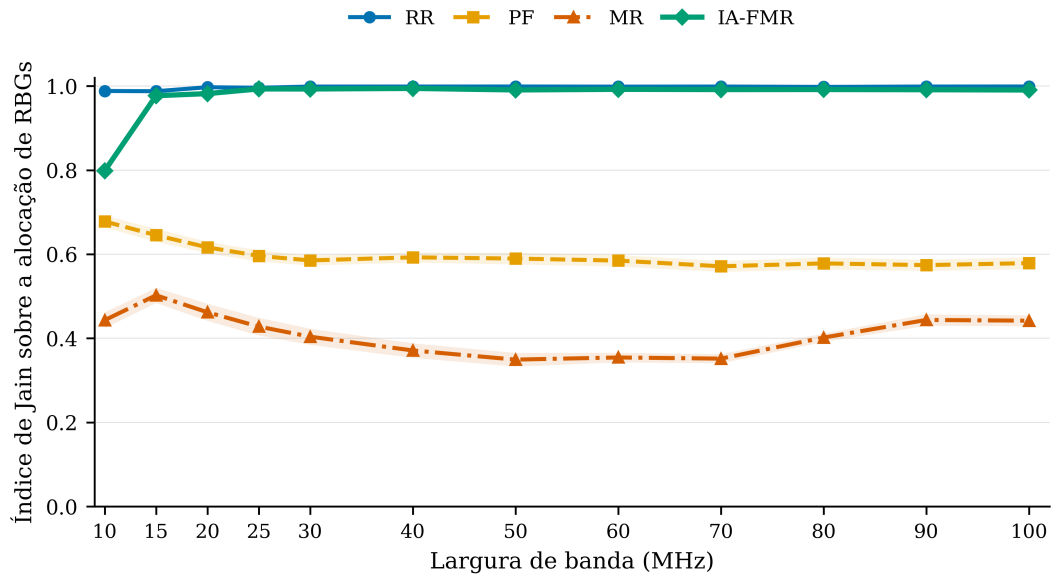


Figura 18: Índice de Jain calculado sobre a alocação de RBGs em função da largura de banda. As regiões sombreadas representam os intervalos de confiança de 95%.

O RR mantém valores médios próximos de 1 em toda a faixa avaliada, comportamento compatível com seu critério de distribuição alternada dos recursos entre os usuários. O IA-FMR também apresenta valores elevados de justiça na maior parte das configurações e se aproxima do RR a partir de 15 MHz. A principal diferença ocorre em 10 MHz, condição na qual o Índice de Jain médio do IA-FMR é reduzido para 0,7985, enquanto o RR apresenta 0,9882.

Em 40 MHz, configuração utilizada posteriormente nas análises mais detalhadas, o Índice de Jain médio é de 0,9988 para o RR e de 0,9943 para o IA-FMR. Na mesma configuração, PF e MR apresentam valores de 0,5926 e 0,3711, respectivamente. Assim, RR e IA-FMR produzem distribuições de RBGs próximas da uniformidade nessa configuração, enquanto PF e, principalmente, MR concentram uma parcela maior dos recursos em determinados usuários.

O comportamento do MR é coerente com seu critério de priorizar os UEs que apresentem melhores condições instantâneas de transmissão. Essa escolha favorece a eficiência espectral, mas resulta em menor uniformidade na distribuição dos RBGs. O PF ocupa uma região intermediária, refletindo o compromisso introduzido por seu critério entre a taxa instantânea e o histórico de atendimento dos usuários.

A configuração de 10 MHz constitui a principal exceção ao comportamento observado para o IA-FMR. Nessa largura de banda, seu Índice de Jain médio é de 0,7985, superior aos valores observados para PF e MR, de 0,6780 e 0,4432, respectivamente, mas inferior

ao RR. O resultado evidencia que a elevada uniformidade observada para o IA-FMR nas demais configurações não é preservada na mesma intensidade quando a quantidade de recursos é mais restrita. A figura, isoladamente, não permite atribuir essa redução a um único mecanismo, razão pela qual essa condição deve ser interpretada em conjunto com as métricas de atendimento apresentadas posteriormente.

É importante observar que o Índice de Jain utilizado nesta análise caracteriza a uniformidade da *alocação de recursos*, e não a igualdade das vazões obtidas pelos UEs. Usuários que recebem quantidades semelhantes de RBGs podem alcançar vazões distintas em razão de diferenças nas condições de canal e no MCS. Por esse motivo, valores elevados de Jain não constituem, por si só, evidência de proteção contra vazões muito baixas ou *starvation*. Essa distinção motiva a análise conjunta com as métricas de vazão por fluxo e atendimento apresentadas nas subseções seguintes.

6.2.3 Atendimento dos UEs e Demanda sem Alocação

A Figura 19 apresenta o número médio de UEs que receberam ao menos um RBG em cada *slot*. O eixo horizontal representa a largura de banda, em MHz, e o eixo vertical apresenta o número médio de UEs servidos por *slot*. Cada curva corresponde a um escalonador e as regiões sombreadas representam os intervalos de confiança de 95%. Essa métrica complementa o Índice de Jain ao indicar quantos usuários participam efetivamente da alocação realizada em cada decisão.

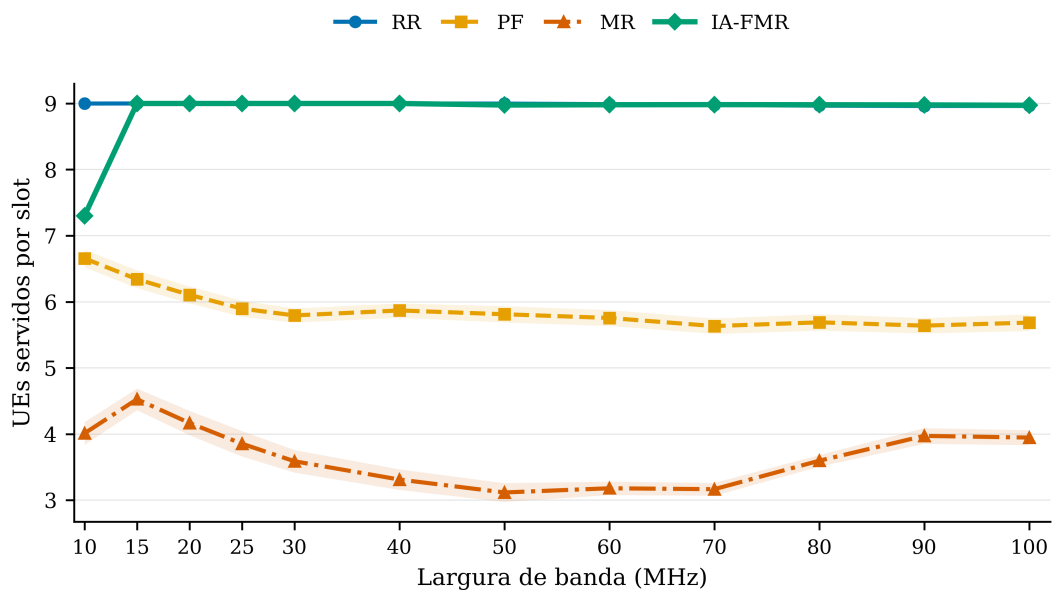


Figura 19: Número médio de UEs servidos por *slot* em função da largura de banda. As regiões sombreadas representam os intervalos de confiança de 95%.

O RR mantém média próxima de nove UEs servidos por *slot* em todas as larguras de banda avaliadas. O IA-FMR apresenta comportamento semelhante a partir de 15 MHz, enquanto em 10 MHz atende, em média, 7,30 UEs por *slot*. Nessa mesma configuração, PF e MR atendem, respectivamente, 6,66 e 4,01 UEs. Em 40 MHz, RR e IA-FMR atendem, em média, os nove UEs, enquanto PF e MR apresentam médias de 5,87 e 3,31 UEs servidos por *slot*.

Esses resultados complementam a análise do Índice de Jain. Em 40 MHz, a elevada uniformidade da alocação observada para RR e IA-FMR ocorre juntamente com a participação de todo o conjunto de UEs nas decisões. PF e, principalmente, MR distribuem os recursos entre subconjuntos menores de usuários em cada *slot*, comportamento coerente com os menores valores de justiça observados anteriormente.

A configuração de 10 MHz novamente se diferencia das demais para o IA-FMR. Nessa condição, o número médio de UEs efetivamente servidos é inferior aos nove usuários considerados pelo escalonador. Esse comportamento será retomado adiante na comparação entre a alocação produzida pela política e aquela efetivamente aplicada pela camada MAC.

A Figura 20 apresenta a frequência de demanda pendente sem alocação. O eixo horizontal representa a largura de banda, em MHz, e o eixo vertical apresenta o percentual de registros UE-*slot* com demanda positiva nos quais a alocação aplicada foi nula. Cada curva corresponde a um escalonador e as regiões sombreadas representam os intervalos de confiança de 95%. A métrica é calculada considerando apenas os registros UE-*slot* em que havia demanda pendente, identificada por `buf_req > 0`, e corresponde à fração desses registros nos quais a alocação aplicada foi nula, `alloc_rbg == 0`. Formalmente, o indicador representa a frequência com que um UE que possuía dados aguardando transmissão permaneceu sem recursos naquele *slot*.

O RR mantém esse indicador próximo de zero em 10 MHz, com média de 0,012%, e igual a zero em 40 MHz. Para o IA-FMR, a demanda sem alocação alcança 18,89% em 10 MHz, mas não é observada a partir de 15 MHz. O PF apresenta 26,04% em 10 MHz e 34,79% em 40 MHz, enquanto o MR alcança 55,44% e 61,96%, respectivamente.

Em 10 MHz, portanto, o IA-FMR apresenta frequência de demanda sem alocação inferior às observadas para PF e MR, mas superior à do RR. Esse resultado é coerente com a redução do número médio de UEs servidos pelo IA-FMR nessa configuração. Em 40 MHz, tanto RR quanto IA-FMR apresentam valor nulo para o indicador e atendem, em média, os nove UEs por *slot*, enquanto PF e MR mantêm simultaneamente menor número de UEs servidos e maior frequência de demanda pendente sem alocação.

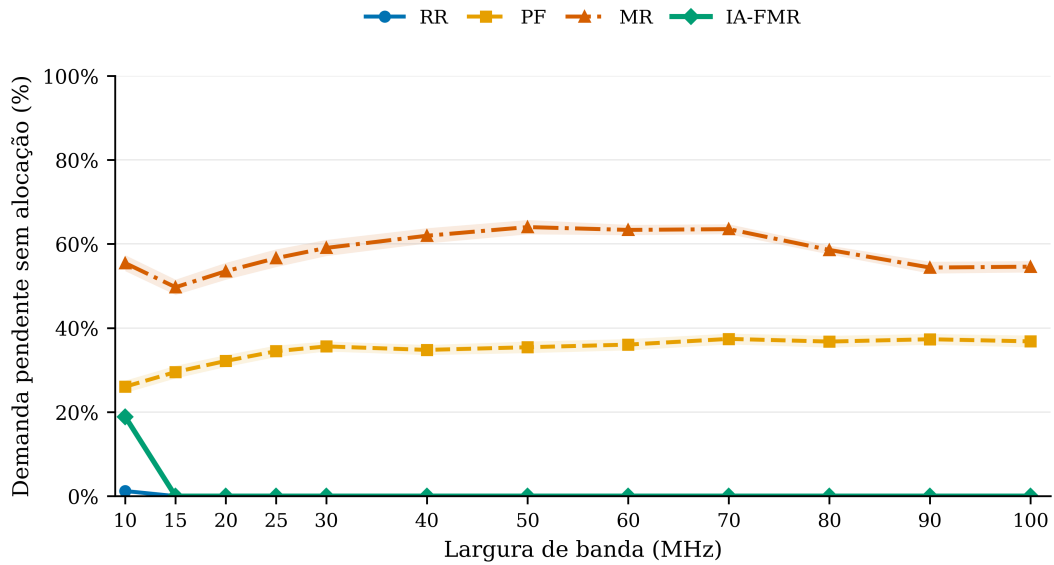


Figura 20: Percentual de registros UE-slot com demanda pendente nos quais a alocação aplicada foi nula, em função da largura de banda. As regiões sombreadas representam os intervalos de confiança de 95%.

Uma ocorrência isolada de demanda pendente sem alocação não caracteriza *starvation*. Um UE pode permanecer sem recursos em determinado *slot* e ser atendido nas decisões seguintes. O indicador quantifica, portanto, a frequência de ausência de atendimento instantâneo entre os registros com demanda positiva. A caracterização de subatendimento persistente requer sua interpretação conjunta com a evolução temporal do serviço e com as métricas de vazão dos fluxos.

Como discutido na Seção 4.5.4, a alocação aplicada pela camada MAC pode diferir da alocação alvo produzida pelo IA-FMR. A Tabela 18 quantifica essa diferença e permite distinguir a decisão produzida pela política daquela efetivamente realizada no simulador.

A coluna “decisões idênticas” representa o percentual de decisões em que a quantidade de RBGs aplicada a cada UE coincide integralmente com a alocação alvo. A coluna “mesmo conjunto de UEs” considera apenas quais usuários receberam alocação positiva, independentemente de pequenas alterações nas quantidades atribuídas. A divergência normalizada quantifica a diferença entre as distribuições, sendo igual a zero quando alvo e aplicação são idênticos. As duas colunas seguintes comparam o número médio de UEs com alocação positiva e o Índice de Jain antes e depois da realização da decisão na camada MAC. As métricas foram calculadas separadamente em cada semente e posteriormente agregadas entre as 100 execuções.

Entre 15 e 30 MHz, a distribuição aplicada coincide integralmente com a alocação

Tabela 18: Comparação entre a alocação alvo produzida pelo IA-FMR e a alocação aplicada no ns-3/5G-LENA.

Banda (MHz)	Decisões idênticas (%)	Mesmo conjunto de UEs (%)	Divergência normalizada	UEs com alocação alvo → aplicada	Índice de Jain alvo → aplicado
10	0,00	0,00	0,19618	9,000 → 7,300	0,9630 → 0,7985
15	100,00	100,00	0,00000	9,000 → 9,000	0,9769 → 0,9769
20	100,00	100,00	0,00000	9,000 → 9,000	0,9820 → 0,9820
25	100,00	100,00	0,00000	9,000 → 9,000	0,9931 → 0,9931
30	100,00	100,00	0,00000	9,000 → 9,000	0,9929 → 0,9929
40	98,54	100,00	0,00095	9,000 → 9,000	0,9946 → 0,9943
50	98,74	100,00	0,00195	8,977 → 8,977	0,9946 → 0,9930
60	99,29	100,00	0,00120	8,981 → 8,981	0,9951 → 0,9941
70	99,54	100,00	0,00070	8,982 → 8,982	0,9938 → 0,9933
80	99,41	100,00	0,00102	8,981 → 8,981	0,9946 → 0,9936
90	99,47	100,00	0,00077	8,978 → 8,978	0,9940 → 0,9934
100	99,36	100,00	0,00103	8,973 → 8,973	0,9941 → 0,9932

alvo em todas as decisões. Nessas configurações, não são observadas alterações nem nas quantidades de RBGs atribuídas aos UEs nem no conjunto de usuários atendidos.

A partir de 40 MHz, pequenas diferenças quantitativas passam a aparecer, mas o conjunto de UEs atendidos permanece preservado em 100% das decisões. Em 40 MHz, 98,54% das distribuições são exatamente idênticas, e a divergência normalizada média é de apenas 0,00095. O Índice de Jain passa de 0,9946 no alvo para 0,9943 após a aplicação, enquanto os nove UEs continuam atendidos. Entre 50 e 100 MHz, as decisões exatamente idênticas permanecem acima de 98,7%, e as divergências normalizadas mantêm-se próximas de zero. Os valores médios ligeiramente inferiores a nove UEs nessas bandas estão presentes tanto no alvo quanto na aplicação e, portanto, não decorrem de remoções introduzidas pela camada MAC.

O comportamento de 10 MHz é distinto. Nessa configuração, a política produz alocação positiva para os nove UEs, com Índice de Jain médio de 0,9630. Após a realização da decisão pela camada MAC, o número médio de UEs com alocação positiva diminui para 7,30 e o Índice de Jain para 0,7985. Nenhuma das decisões apresenta distribuição integralmente idêntica, e o mesmo conjunto de UEs atendidos também não é preservado. A divergência normalizada média aumenta para 0,1962, valor substancialmente superior ao observado nas demais larguras de banda.

Esses resultados mostram que a redução de atendimento observada em 10 MHz não está presente com a mesma intensidade na distribuição alvo produzida pela política. A política solicita recursos para os nove UEs e produz uma distribuição com elevado Índice de Jain, mas parte dessa distribuição é modificada durante sua realização na camada

MAC. A configuração de 10 MHz, que possui a menor quantidade de recursos entre as avaliadas, é a única em que essas modificações alteram de forma expressiva o conjunto de UEs efetivamente atendidos.

Nas demais larguras de banda, a alocação aplicada acompanha de perto a decisão produzida pelo IA-FMR. Entre 15 e 30 MHz, a correspondência é integral e, entre 40 e 100 MHz, as diferenças são predominantemente quantitativas e residuais, sem alteração do conjunto de usuários selecionados para atendimento.

6.2.4 Vazão dos Fluxos Menos Favorecidos

A Figura 21 apresenta o percentil 5 da vazão por fluxo. O eixo horizontal representa a largura de banda, em MHz, e o eixo vertical apresenta o P5 médio, em Mbps. Cada curva corresponde a um escalonador e as regiões sombreadas representam os intervalos de confiança de 95%. Essa métrica caracteriza a região inferior da distribuição e permite observar o desempenho dos fluxos menos favorecidos, complementando a análise baseada na vazão agregada da célula.

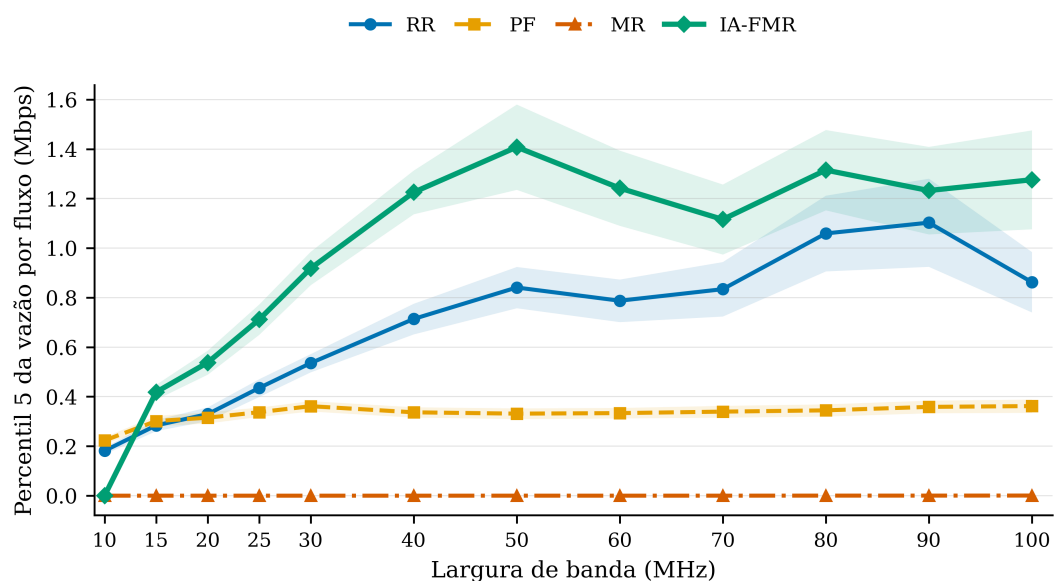


Figura 21: Percentil 5 da vazão por fluxo em função da largura de banda. As regiões sombreadas representam os intervalos de confiança de 95%.

Em 10 MHz, o IA-FMR e o MR apresentam P5 nulo, enquanto RR e PF alcançam médias de 0,181 Mbps e 0,223 Mbps, respectivamente. Para o IA-FMR, esse comportamento é coerente com os resultados anteriores dessa configuração, na qual foram observados menor número médio de UEs servidos, redução do Índice de Jain aplicado e ocorrência de demanda pendente sem alocação. A análise da alocação alvo e aplicada mostrou ainda

que, nessa largura de banda, parte da distribuição produzida pela política é modificada durante sua realização na camada MAC.

A partir de 15 MHz, o IA-FMR apresenta a maior média de P5 entre os escalonadores avaliados em todas as larguras de banda. Em 40 MHz, por exemplo, o P5 médio é de 1,225 Mbps para o IA-FMR, 0,714 Mbps para o RR e 0,337 Mbps para o PF, enquanto o MR permanece próximo de zero. Considerando os valores médios, o IA-FMR apresenta nessa configuração P5 71,67% maior que o RR e 263,96% maior que o PF.

O MR mantém valores de P5 nulos ou muito próximos de zero ao longo de toda a faixa avaliada. Esse comportamento contrasta com sua elevada vazão agregada e indica que o aumento da taxa total da célula é acompanhado por forte concentração do desempenho em parte dos fluxos. O resultado é consistente com o menor Índice de Jain e com o menor número médio de UEs servidos observados anteriormente para esse escalonador.

O PF apresenta variação relativamente pequena do P5 com o aumento da largura de banda. Suas médias permanecem próximas de 0,3 Mbps na maior parte das configurações, indicando crescimento limitado da região inferior da distribuição mesmo quando a quantidade de recursos disponível aumenta.

Nas configurações de 80 e 90 MHz, os intervalos de confiança individuais do IA-FMR e do RR apresentam sobreposição, embora a média do IA-FMR permaneça maior. Essa observação é descritiva e não deve ser interpretada como teste formal de diferença entre os escalonadores.

Em conjunto, os resultados mostram que, exceto em 10 MHz, o IA-FMR combina elevada uniformidade na distribuição dos RBGs com valores médios de P5 superiores aos observados para PF e MR e, em termos das médias, também superiores aos do RR. Dessa forma, a maior vazão agregada obtida pelo IA-FMR em relação a RR e PF em parte das configurações não ocorre apenas pelo aumento da taxa total da célula, mas é acompanhada por maior vazão na região inferior da distribuição dos fluxos.

É importante distinguir essa métrica do limiar $T_{\min}$ utilizado durante o treinamento. O P5 apresentado nesta análise é calculado sobre a vazão dos fluxos observada no ns-3/5G-LENA, enquanto $T_{\min}$ atua apenas como referência para a penalização por déficit de vazão no ambiente de treinamento. Portanto, valores de P5 superiores ou inferiores a $T_{\min}$ não devem ser interpretados como cumprimento ou violação de uma garantia de QoS.

6.2.5 Perdas e Atraso

A Figura 22 apresenta a taxa média de perda de pacotes em função da largura de banda. O eixo horizontal representa a largura de banda, em MHz, e o eixo vertical apresenta a taxa média de perda, em percentual. Cada curva corresponde a um escalonador e as regiões sombreadas representam os intervalos de confiança de 95%.

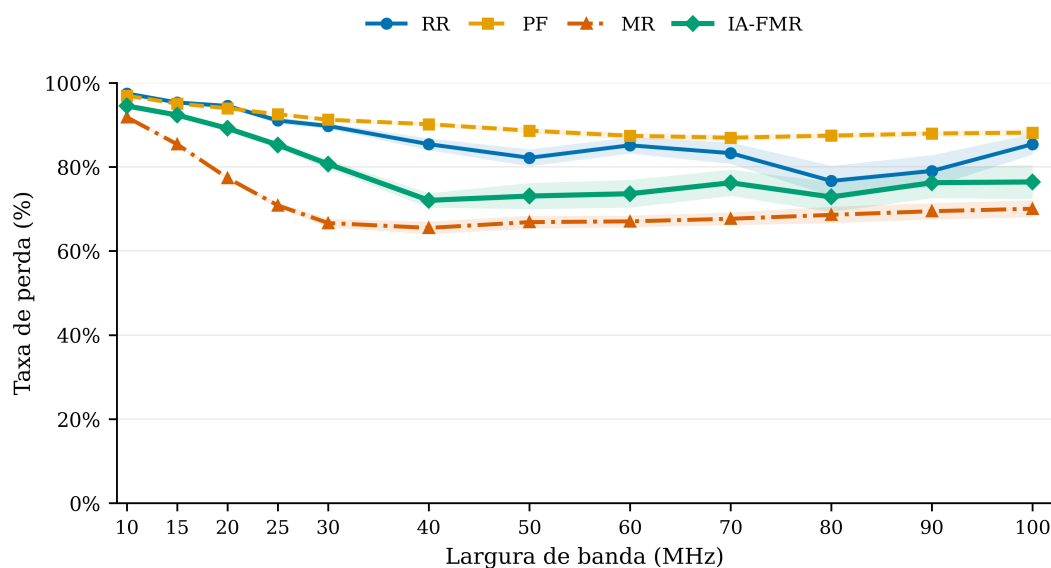


Figura 22: Taxa média de perda de pacotes em função da largura de banda. As regiões sombreadas representam os intervalos de confiança de 95%.

As perdas permanecem elevadas nas configurações avaliadas, comportamento compatível com um cenário submetido a períodos de carga superior à capacidade efetiva de atendimento da célula.

Em termos das médias observadas, o MR apresenta as menores taxas de perda ao longo da faixa avaliada, enquanto PF e RR ocupam as regiões superiores do gráfico. O IA-FMR permanece entre esses dois grupos, apresentando médias menores que as de RR e PF e maiores que as do MR nas diferentes configurações.

Em 40 MHz, a taxa média de perda é de 90,15% para o PF, 85,40% para o RR, 72,04% para o IA-FMR e 65,48% para o MR. Assim, nessa configuração, a menor perda observada para o IA-FMR em relação a RR e PF acompanha a maior vazão agregada e o maior P5 apresentados anteriormente.

A menor taxa de perda do MR deve, entretanto, ser interpretada em conjunto com as demais métricas. O escalonador concentra os recursos em um número menor de UEs, apresenta Índice de Jain reduzido, elevada frequência de demanda pendente sem alocação

e P5 praticamente nulo. Dessa forma, sua menor perda média e sua maior vazão agregada não caracterizam atendimento uniforme dos fluxos, mas refletem o desempenho obtido sob uma política que prioriza usuários com melhores condições instantâneas de transmissão.

A Figura 23 apresenta o atraso médio dos pacotes efetivamente recebidos pelos UEs. O eixo horizontal representa a largura de banda, em MHz, e o eixo vertical apresenta o atraso médio, em milissegundos. Cada curva corresponde a um escalonador e as regiões sombreadas representam os intervalos de confiança de 95%.

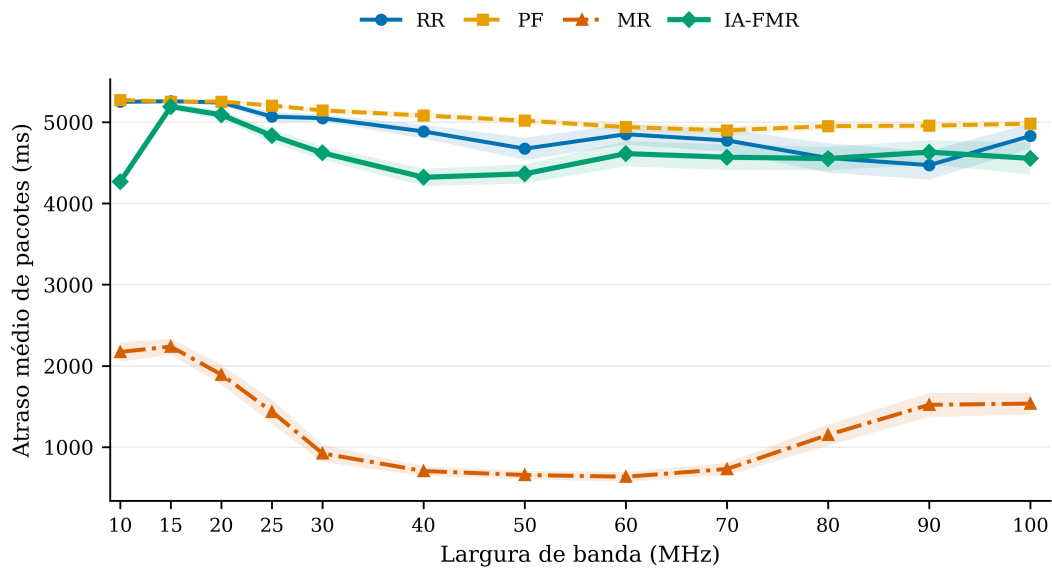


Figura 23: Atraso médio dos pacotes recebidos em função da largura de banda. As regiões sombreadas representam os intervalos de confiança de 95%.

O MR apresenta as menores médias de atraso ao longo das configurações avaliadas. Esse resultado deve ser interpretado considerando a própria definição da métrica: somente os pacotes efetivamente recebidos participam do cálculo do atraso. Pacotes descartados ou que não chegam ao destino não contribuem para essa média. Como o MR concentra o atendimento nos usuários selecionados segundo suas condições instantâneas de transmissão, seu baixo atraso médio descreve o subconjunto de pacotes que efetivamente alcança os UEs e não o desempenho temporal de toda a carga oferecida.

Em 40 MHz, o atraso médio é de 704,60 ms para o MR, 4.321,06 ms para o IA-FMR, 4.885,64 ms para o RR e 5.081,83 ms para o PF. Nessa configuração, portanto, a média observada para o IA-FMR é inferior às médias de RR e PF, mas permanece substancialmente acima daquela obtida pelo MR.

A mesma tendência geral aparece ao longo das demais larguras de banda: as médias de atraso do IA-FMR permanecem abaixo das observadas para RR e PF, enquanto o

MR ocupa uma região consideravelmente inferior. Nas configurações com maior largura de banda, entretanto, os intervalos de confiança tornam-se mais amplos e apresentam sobreposição em alguns pontos, especialmente entre RR e IA-FMR. Por essa razão, a ordenação das médias é interpretada de forma descritiva e não como evidência, por si só, de diferença estatisticamente significativa em todas as configurações.

Em conjunto, perda e atraso reforçam a necessidade de interpretar as métricas de desempenho de forma combinada. O MR apresenta as menores médias para esses dois indicadores e a maior vazão agregada, mas simultaneamente concentra recursos e apresenta desempenho reduzido na região inferior da distribuição dos fluxos. O IA-FMR ocupa uma região distinta: suas médias de perda e atraso são menores que as de RR e PF nas configurações avaliadas, ao mesmo tempo em que preserva maior uniformidade na distribuição dos RBGs e maiores valores de P5 a partir de 15 MHz.

6.2.6 Análise Temporal em 40 MHz

A análise temporal em 40 MHz complementa os resultados agregados apresentados anteriormente ao examinar, em maior detalhe, o comportamento dos escalonadores ao longo de uma realização específica do cenário. Para essa análise, foi utilizada a semente 47889, comum aos quatro escalonadores, de forma que RR, PF, MR e IA-FMR sejam observados sob a mesma realização pseudoaleatória da simulação.

Diferentemente das figuras agregadas da Avaliação por Largura de Banda, que utilizam as 100 sementes de avaliação, as curvas apresentadas nesta subseção são construídas a partir de uma única execução por escalonador. Por esse motivo, não são apresentados intervalos de confiança e os resultados não devem ser interpretados como estimativas da variabilidade entre sementes. Seu objetivo é permitir a inspeção da dinâmica das decisões ao longo das cinco fases de tráfego definidas na Seção 5.3.1.

Os registros foram agrupados em janelas não sobrepostas de 0,1 s. Para reduzir oscilações visuais de curta duração, foi aplicada posteriormente uma média móvel de três janelas, equivalente a 0,3 s. As linhas verticais pontilhadas indicam as transições entre as cinco fases de tráfego.

Os identificadores UE 1 a UE 9 seguem a ordem definida na configuração experimental. Durante o processamento, os registros são associados pelo RNTI, preservando a correspondência entre os usuários ao longo da execução.

As métricas N_{ef} e p_{max} , apresentadas posteriormente, também são calculadas a partir

dessa única realização. Seus valores caracterizam, portanto, a evolução temporal da concentração dos recursos na semente 47889 e não correspondem a médias calculadas entre as 100 sementes do experimento.

6.2.6.1 Distribuição Temporal dos RBGs

As Figuras 24 a 27 apresentam a participação de cada UE no total de RBGs alocados ao longo da execução. Em todas elas, o eixo horizontal representa o tempo de simulação, em segundos, e o eixo vertical apresenta a participação de cada UE na alocação de RBGs, em percentual. Cada curva corresponde a um dos nove UEs. A linha horizontal tracejada indica a participação correspondente à divisão uniforme dos recursos entre os nove UEs, aproximadamente 11,11%, e as linhas verticais pontilhadas marcam as transições entre as fases de tráfego.

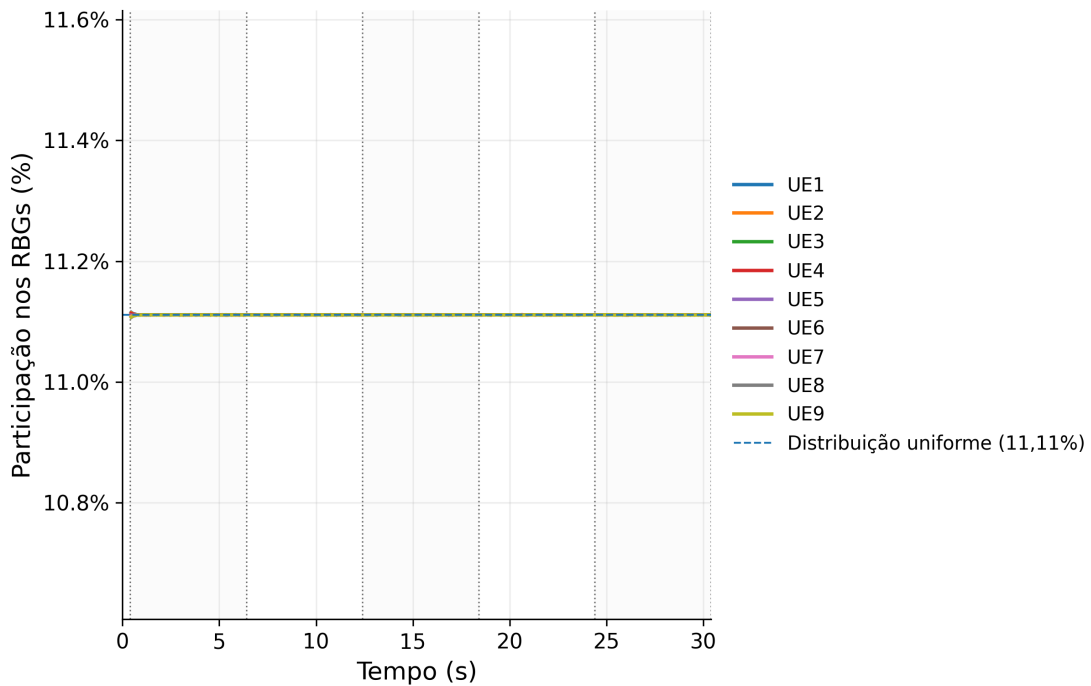


Figura 24: Participação temporal dos UEs na alocação de RBGs pelo RR em 40 MHz para a semente 47889.

Na Figura 24, as participações dos nove UEs permanecem praticamente sobrepostas à referência de divisão uniforme ao longo de toda a execução. O comportamento é coerente com o mecanismo do RR, que distribui os recursos de forma alternada entre os usuários e, nessa realização, produz participações acumuladas praticamente iguais.

Na Figura 25, as participações dos UEs oscilam em torno da referência uniforme durante toda a execução. Diferentemente do RR, as curvas não permanecem sobrepostas,

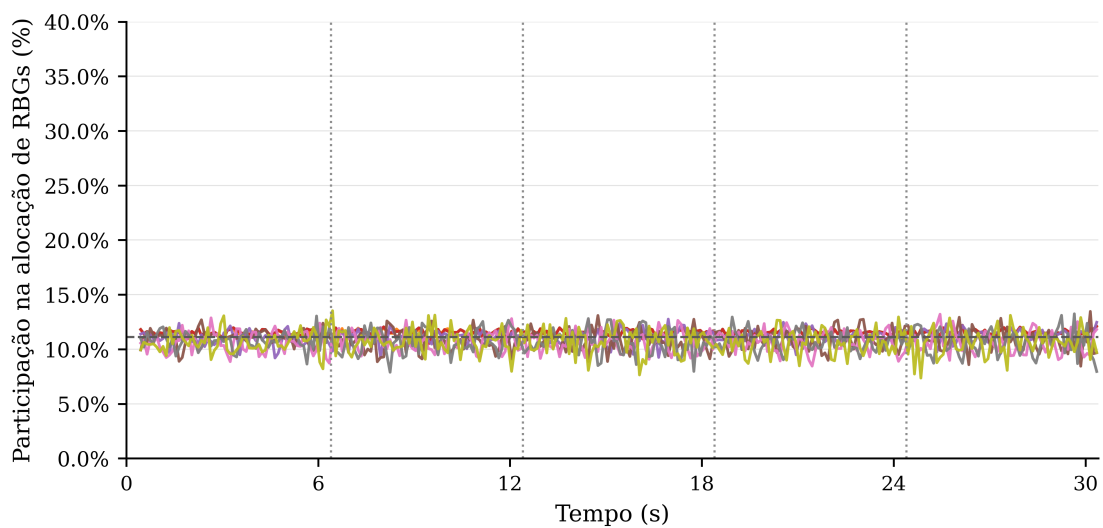


Figura 25: Participação temporal dos UEs na alocação de RBGs pelo PF em 40 MHz para a semente 47889.

mas também não se observa a manutenção prolongada de um mesmo usuário em posição dominante.

Esse comportamento não significa que os nove UEs recebam recursos simultaneamente em cada *slot*. Como mostrado anteriormente, o PF atende, em média, um subconjunto menor de usuários por decisão. A proximidade das participações observada nessa escala temporal resulta da alternância dos UEs atendidos ao longo da execução.

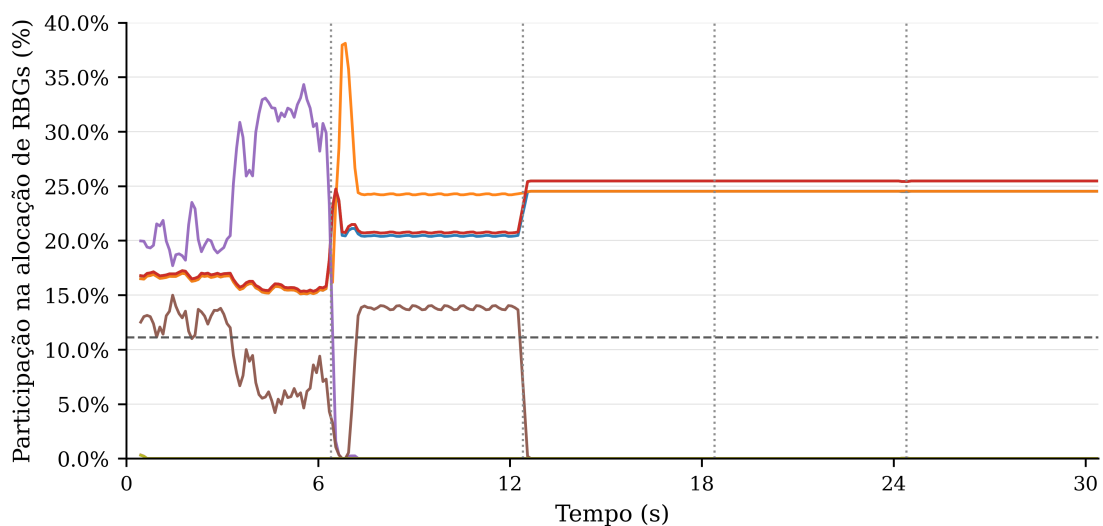


Figura 26: Participação temporal dos UEs na alocação de RBGs pelo MR em 40 MHz para a semente 47889.

O MR, apresentado na Figura 26, exibe comportamento substancialmente mais concentrado. Já na primeira fase há diferenças relevantes entre as participações dos UEs.

Após as transições de carga, parte dos usuários passa a concentrar parcelas maiores dos RBGs, enquanto outros permanecem com participação reduzida ou nula durante intervalos prolongados.

A mudança é particularmente evidente após a primeira transição, quando alguns UEs concentram parcelas próximas ou superiores a 20% dos recursos, enquanto outros deixam de receber participação relevante. Nas fases posteriores, a distribuição permanece concentrada em um subconjunto de usuários. Esse comportamento é coerente com os resultados agregados do MR, que combinaram maior vazão agregada com menor Índice de Jain, menor número médio de UEs servidos e P5 praticamente nulo.

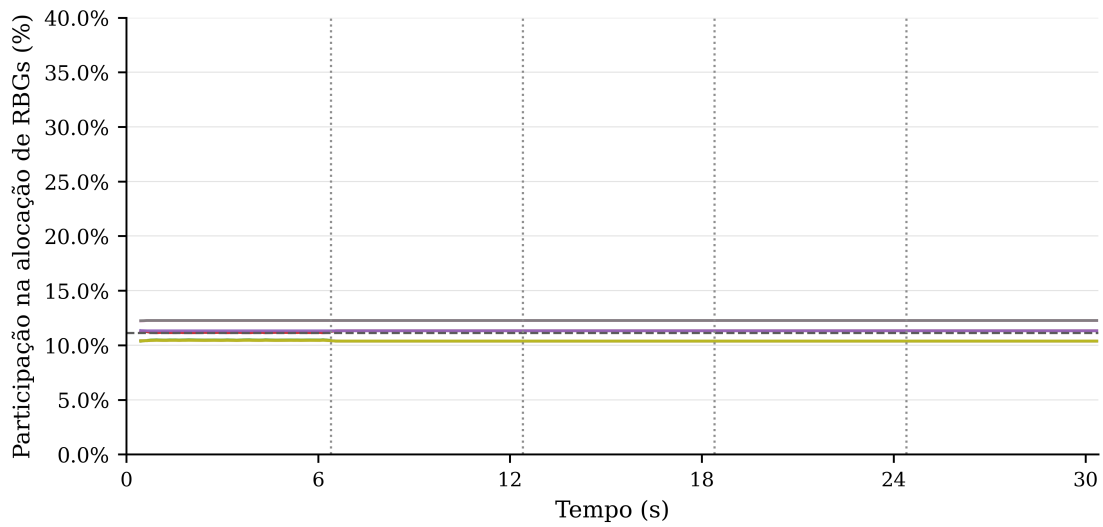


Figura 27: Participação temporal dos UEs na alocação de RBGs pelo IA-FMR em 40 MHz para a semente 47889.

A Figura 27 mostra comportamento distinto para o IA-FMR. Os nove UEs permanecem representados na distribuição durante toda a execução, mas as participações não são rigidamente iguais como no RR. As curvas permanecem próximas da referência de 11,11%, com pequenas diferenças persistentes entre os usuários.

Nessa realização, o IA-FMR preserva, portanto, uma distribuição ampla dos recursos sem impor uma divisão exatamente uniforme. Esse comportamento é coerente com os resultados agregados em 40 MHz, nos quais o mecanismo apresentou Índice de Jain elevado e atendimento médio dos nove UEs, ao mesmo tempo em que sua vazão agregada diferiu da observada para o RR.

As quatro figuras evidenciam mecanismos temporalmente distintos. O RR mantém uma divisão praticamente uniforme. O PF produz oscilações em torno dessa referência ao alternar os usuários atendidos. O MR concentra recursos de forma persistente em parte

dos UEs, enquanto o IA-FMR mantém os nove usuários representados e permite pequenas diferenças de participação ao longo da execução. Essas observações caracterizam especificamente a semente 47889 e complementam, sem substituir, os resultados estatísticos obtidos com as 100 sementes.

6.2.6.2 Concentração da Alocação

A Figura 28 apresenta o número efetivo de UEs na distribuição de RBGs ao longo da execução. O eixo horizontal representa o tempo de simulação, em segundos, e o eixo vertical apresenta N_{ef} . As linhas verticais pontilhadas indicam as transições entre as fases de tráfego. Para as participações $p_i(b)$ dos N usuários no intervalo temporal b , a métrica é calculada por

$$N_{\text{ef}}(b) = \frac{1}{\sum_{i=1}^N p_i(b)^2}. \quad (6.1)$$

Essa métrica representa o número equivalente de usuários que participariam da distribuição caso os recursos fossem divididos uniformemente entre eles. Seu valor se aproxima de nove quando as participações são semelhantes e diminui à medida que os RBGs se concentram em um conjunto menor de UEs.

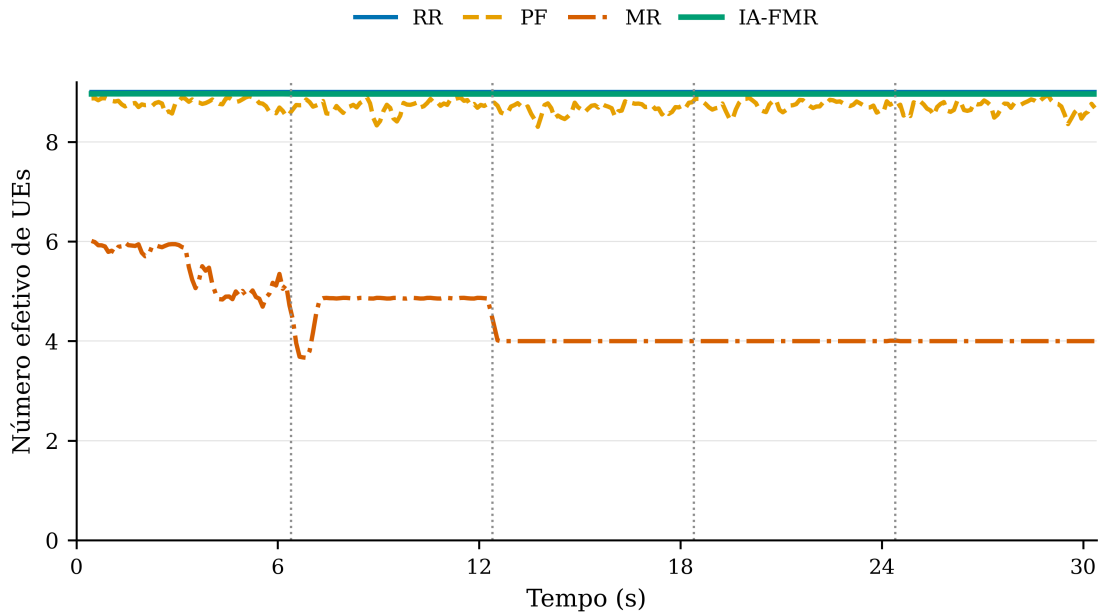


Figura 28: Número efetivo de UEs na distribuição temporal de RBGs em 40 MHz para a semente 47889.

Na realização analisada, o RR permanece praticamente no valor máximo durante toda

a execução, com N_{ef} médio de 9,00. O IA-FMR apresenta comportamento semelhante, com média de 8,96 e pequena variação temporal. O PF também mantém valor elevado, embora inferior aos dois anteriores, com média de 8,72 e oscilações entre aproximadamente 8,30 e 8,94.

A interpretação do PF deve considerar que N_{ef} é calculado sobre a distribuição de recursos agregada em cada janela temporal. Como mostrado anteriormente, o mecanismo não necessariamente atende os nove UEs em um mesmo *slot*. A proximidade de N_{ef} com nove decorre da alternância dos usuários atendidos ao longo das decisões contidas em cada janela.

O MR apresenta comportamento distinto. Seu N_{ef} médio é de 4,44, com valores entre aproximadamente 3,66 e 6,01 ao longo da execução. Após as primeiras transições de tráfego, a métrica diminui e passa a permanecer próxima de quatro, evidenciando uma distribuição concentrada em um subconjunto menor de usuários nessa realização.

A Figura 29 complementa essa análise ao apresentar a maior participação individual na distribuição de RBGs em cada intervalo temporal, denotada por p_{max} . O eixo horizontal representa o tempo de simulação, em segundos, e o eixo vertical apresenta essa participação máxima, em percentual. As linhas verticais pontilhadas indicam as transições entre as fases de tráfego. A métrica corresponde à parcela dos recursos atribuída ao UE com maior participação naquele intervalo, que não necessariamente corresponde ao mesmo usuário ao longo de toda a execução.

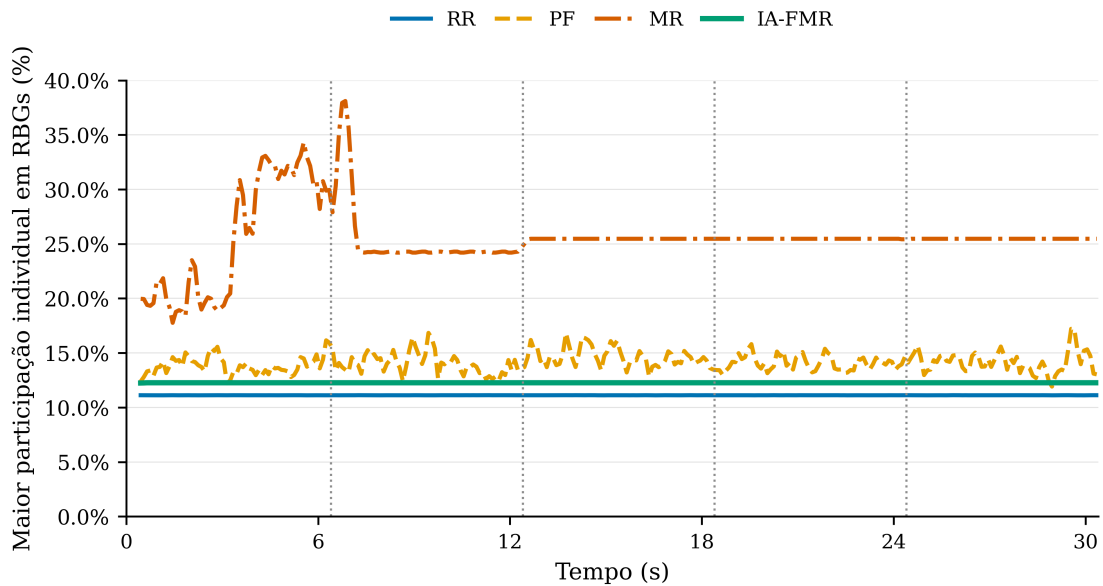


Figura 29: Maior participação individual na distribuição temporal de RBGs em 40 MHz para a semente 47889.

No RR, a maior participação individual permanece praticamente sobre a referência de divisão uniforme entre nove usuários, com média de 11,12%. O IA-FMR também mantém baixa concentração, com $p_{\max}$ médio de 12,26% e pequena variação ao longo da execução. Esse resultado é consistente com a distribuição observada anteriormente, na qual os nove UEs permanecem representados, mas não recebem exatamente a mesma parcela dos recursos.

O PF apresenta maior variação temporal, com $p_{\max}$ médio de 14,16% e valores entre aproximadamente 11,92% e 17,21%. Essa oscilação é coerente com a alternância dos usuários priorizados pelo mecanismo, embora a distribuição agregada em cada janela permaneça relativamente próxima do equilíbrio.

A concentração é substancialmente maior no MR. Sua maior participação individual apresenta média de 25,46% e alcança 38,11% no ponto de maior concentração da execução. Esse pico ocorre próximo à primeira transição de tráfego. Nas fases posteriores, a participação máxima permanece em um patamar elevado, mas não cresce monotonicamente a cada mudança de carga.

Os dois indicadores fornecem, portanto, uma leitura consistente da realização selecionada. O RR mantém uma distribuição praticamente uniforme. O IA-FMR preserva um número efetivo próximo de nove, permitindo pequenas diferenças entre as participações individuais. O PF permanece relativamente próximo do equilíbrio na escala temporal utilizada, mas apresenta maior oscilação na participação do usuário mais favorecido. Já o MR concentra uma parcela substancial dos recursos em um conjunto reduzido de UEs.

Os valores apresentados nesta análise são descritivos da semente 47889. Eles foram obtidos após a agregação dos registros em janelas de 0,1 s e a aplicação de média móvel centralizada de três janelas, equivalente a 0,3 s, e não representam médias ou intervalos de confiança calculados entre as 100 sementes da Avaliação por Largura de Banda.

6.2.6.3 Atendimento dos UEs

A Figura 30 apresenta a evolução temporal do número de UEs que receberam ao menos um RBG aplicado em cada *slot*, considerando a semente 47889. O eixo horizontal representa o tempo de simulação, em segundos, e o eixo vertical apresenta o número médio de UEs servidos por *slot* em cada janela. As linhas verticais pontilhadas indicam as transições entre as fases de tráfego. Assim como nas demais figuras temporais, os valores são agregados em janelas de 0,1 s e suavizados por média móvel de 0,3 s.

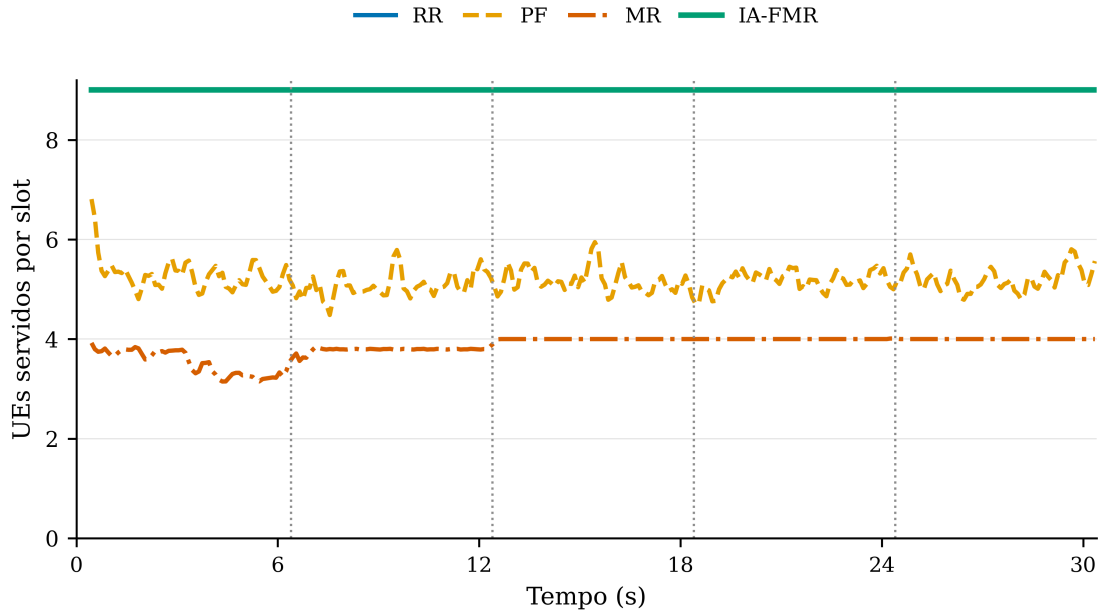


Figura 30: Número de UEs servidos por *slot* em 40 MHz para a semente 47889.

RR e IA-FMR mantêm os nove UEs atendidos ao longo da execução. No caso do IA-FMR, esse comportamento é consistente com a comparação entre alocação alvo e aplicada apresentada anteriormente. Em 40 MHz, a política produz alocação positiva para os nove usuários e a camada MAC preserva o mesmo conjunto de UEs atendidos em 100% das decisões, embora pequenas alterações nas quantidades de RBGs ocorram em uma parcela reduzida dos casos.

Esse resultado também ajuda a interpretar as diferenças observadas nas participações individuais do IA-FMR. Os valores distintos atribuídos aos UEs não decorrem, nessa configuração, da exclusão de usuários da distribuição, mas de pequenas diferenças na quantidade de recursos atribuída a cada um deles.

O PF apresenta comportamento distinto. Após uma redução inicial, o número de UEs servidos oscila predominantemente em torno de cinco usuários por *slot*, com variações ocasionais para valores superiores. Essa dinâmica explica como o mecanismo pode produzir, na escala temporal agregada, participações relativamente próximas entre os nove UEs mesmo sem atendê-los simultaneamente em cada decisão. O subconjunto servido varia ao longo da execução, distribuindo o atendimento entre diferentes usuários.

O MR mantém o menor número de UEs servidos. Nas primeiras fases, o valor oscila aproximadamente entre três e quatro usuários por *slot* e, a partir da transição em torno de 12 s, estabiliza-se próximo de quatro. Esse comportamento confirma que a concentração observada nas análises de N_{ef} e p_{max} não resulta apenas de diferenças na quantidade de

RBGs atribuída a cada usuário. Nessa realização, o MR também restringe efetivamente o conjunto de UEs atendidos em cada decisão.

A comparação temporal evidencia, portanto, mecanismos distintos de distribuição. RR e IA-FMR mantêm os nove usuários atendidos durante a realização analisada. O PF alterna subconjuntos de aproximadamente cinco usuários ao longo do tempo, enquanto o MR concentra o atendimento em um conjunto ainda menor. Essas observações referem-se especificamente à semente 47889 e complementam os resultados agregados obtidos com as 100 sementes.

6.2.6.4 Evolução do *Backlog*

A Figura 31 apresenta a evolução temporal da ocupação agregada dos buffers informada ao escalonador na realização correspondente à semente 47889. O eixo horizontal representa o tempo de simulação, em segundos, e o eixo vertical apresenta a ocupação agregada dos buffers, em Mbit. As linhas verticais pontilhadas indicam as transições entre as fases de tráfego. O indicador é obtido a partir do campo `buf_req`, que representa a ocupação de buffer associada aos UEs e inclui dados pendentes de transmissão, retransmissão e mensagens de estado.

Para cada instante, os valores de `buf_req`, originalmente expressos em bytes, são somados entre os UEs e convertidos para Mbit por $\text{buf_req} \times 8/10^6$. Em seguida, os registros são agregados em janelas não sobrepostas de 0,1 s e suavizados por média móvel centralizada de três janelas, equivalente a 0,3 s. O gráfico representa, portanto, a demanda pendente visível ao escalonador ao longo das cinco fases de tráfego.

O *backlog* cresce nos quatro escalonadores ao longo da execução, com mudanças de inclinação associadas às transições entre as fases de tráfego. Os maiores acréscimos ocorrem nos períodos de maior carga, enquanto nas fases de menor intensidade o crescimento desacelera. A redução da carga oferecida, entretanto, não é suficiente para eliminar o volume acumulado nas fases anteriores, de modo que a ocupação dos buffers permanece crescente nessa realização.

RR e PF apresentam os maiores volumes pendentes durante praticamente toda a execução. Em 6 s, o *backlog* é de aproximadamente 639,94 Mbit para o RR e 607,53 Mbit para o PF. Em 18 s, esses valores alcançam 4.544,11 Mbit e 4.408,47 Mbit, respectivamente. Próximo de 30 s, chegam a 8.536,10 Mbit para o RR e 8.281,73 Mbit para o PF.

MR e IA-FMR permanecem abaixo de RR e PF durante grande parte da realização,

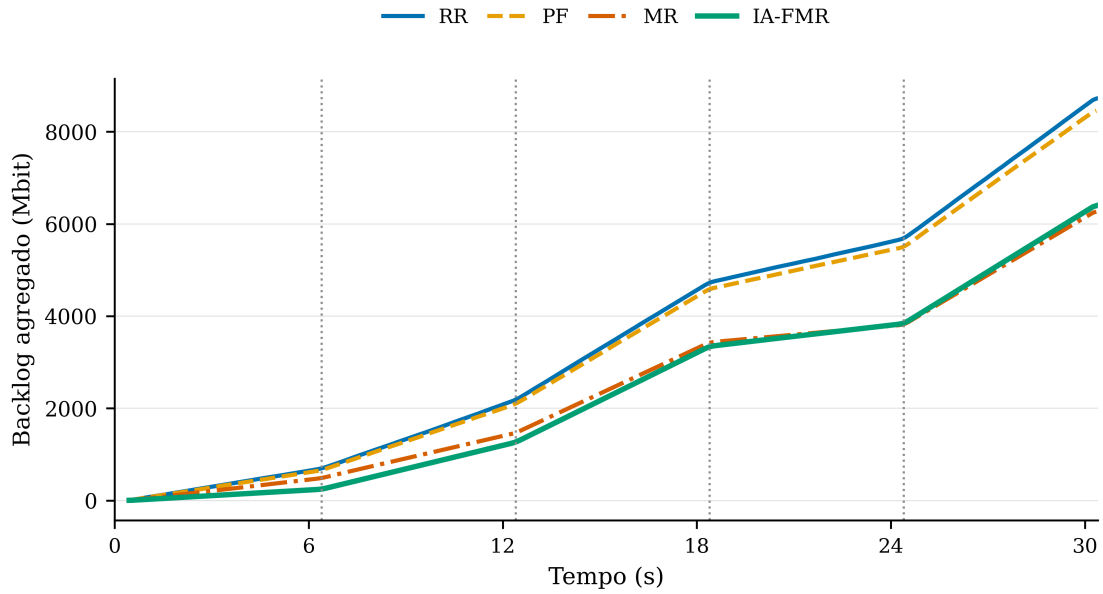


Figura 31: Evolução temporal do *backlog* agregado em 40 MHz para a semente 47889, expresso em Mbit.

mas a ordenação entre os dois varia ao longo do tempo. Nas fases iniciais, o IA-FMR apresenta menor ocupação agregada. Em 6 s, são observados 227,16 Mbit para o IA-FMR e 449,82 Mbit para o MR. Em 12 s, os valores alcançam 1.187,43 Mbit e 1.393,46 Mbit, respectivamente, e, em 18 s, 3.189,63 Mbit para o IA-FMR e 3.284,87 Mbit para o MR.

As curvas se aproximam nas fases posteriores. Em 24 s, o MR apresenta 3.785,04 Mbit e o IA-FMR, 3.800,79 Mbit. Em 30 s, os valores são 6.120,40 Mbit para o MR e 6.243,44 Mbit para o IA-FMR. Dessa forma, a vantagem inicial do IA-FMR em termos de menor ocupação de buffer deixa de ser observada nas etapas finais, quando o MR passa a apresentar valores ligeiramente menores.

A ocorrência de menor *backlog* no IA-FMR durante parte da execução, mesmo com vazão agregada inferior à do MR, não constitui uma contradição. As duas métricas descrevem aspectos distintos do sistema. A vazão agregada mede a quantidade de dados efetivamente recebida pelos UEs, enquanto o *backlog* representa a ocupação instantânea dos buffers visível ao escalonador. O MR obtém elevada vazão ao concentrar recursos nos usuários com melhores condições instantâneas de transmissão, mas deixa um conjunto menor de UEs atendido em cada decisão. Como mostrado anteriormente, em 40 MHz o MR atende, em média, aproximadamente 3,31 UEs por *slot* e apresenta elevada frequência de demanda pendente sem alocação.

O IA-FMR, por outro lado, mantém os nove UEs atendidos nessa configuração. Assim,

embora cada recurso não seja necessariamente convertido na mesma quantidade de dados entregues que no MR, a distribuição mais ampla do atendimento pode impedir que as filas de determinados usuários cresçam tão rapidamente. Consequentemente, é possível que o IA-FMR apresente menor ocupação agregada dos buffers em determinados períodos mesmo entregando menor vazão total que o MR. Nas fases finais, o MR apresenta backlog ligeiramente inferior ao do IA-FMR.

Essa interpretação deve permanecer associada às demais métricas, pois `buf_req` não mede diretamente o volume de dados entregue fim a fim. Ele representa a ocupação dos buffers informada ao escalonador em determinado instante. Dados transmitidos, removidos das filas ou processados pelos mecanismos da camada RLC deixam de contribuir para esse valor. Portanto, menor *backlog* não implica, isoladamente, maior vazão, menor perda ou melhor desempenho global.

Por fim, os valores desta análise correspondem exclusivamente à semente 47889 e possuem finalidade descritiva. Eles permitem observar a evolução interna das filas durante uma realização específica e complementam, sem substituir, os resultados estatísticos obtidos anteriormente com as 100 sementes da Avaliação por Largura de Banda.

6.2.6.5 Evolução de α_t

A Figura 32 apresenta a evolução temporal de α_t produzido pela política IA-FMR treinada para 40 MHz durante a execução correspondente à semente 47889. O eixo horizontal representa o tempo de simulação, em segundos, e o eixo vertical apresenta α_t em escala ampliada para evidenciar suas pequenas variações. As linhas verticais pontilhadas indicam as transições entre as fases de tráfego. A curva foi obtida a partir da coluna `alpha` do `slot_log_fmr_rl.csv`. Como α_t é um valor global por decisão e aparece repetido nos registros dos UEs associados ao mesmo instante, os valores foram inicialmente agregados por `time_s`. Em seguida, foram agrupados em janelas de 0,1 s e suavizados por média móvel centralizada de três janelas, equivalente a 0,3 s.

Nessa realização, α_t permanece muito próximo do limite superior de seu domínio durante toda a execução. A média temporal é de 0,9982, com mínimo de 0,9974 e máximo de 0,9983. Após uma pequena variação no início da simulação, o coeficiente se estabiliza próximo de 0,9980 e apresenta apenas pequenos incrementos nas fases posteriores, alcançando aproximadamente 0,9983 ao final da execução.

As transições entre as cinco fases de tráfego produzem, portanto, alterações muito

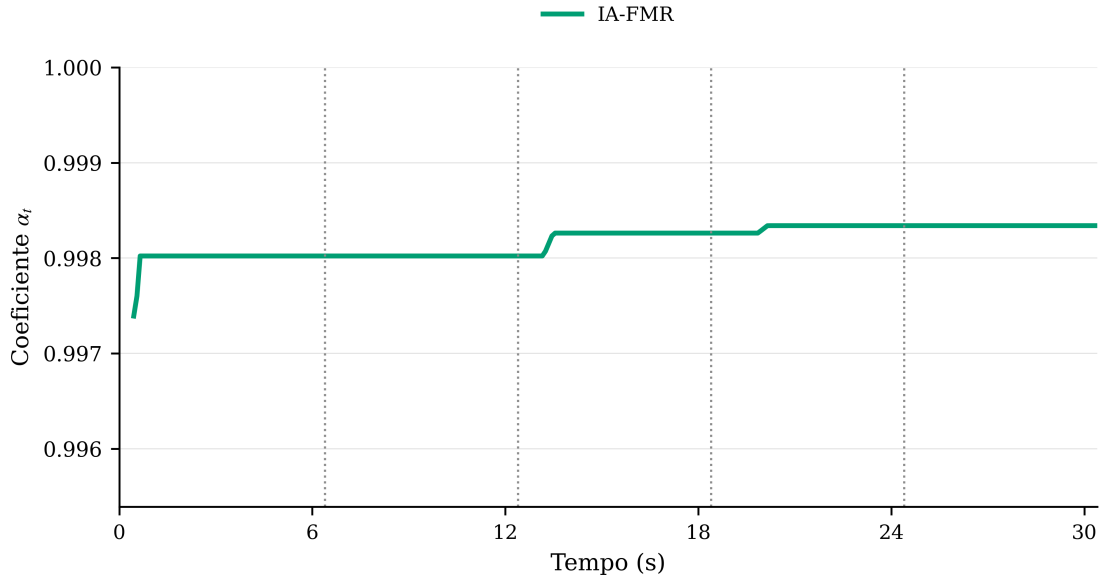


Figura 32: Evolução temporal de α_t produzido pelo IA-FMR em 40 MHz para a semente 47889. O eixo vertical utiliza escala ampliada para evidenciar as pequenas variações observadas ao longo da execução.

pequenas em α_t nessa realização. Embora o coeficiente seja produzido a cada decisão e possa ocupar o domínio definido para a política, o comportamento observado em 40 MHz permanece concentrado em uma região muito próxima de 1.

Esse resultado é consistente com a evolução observada durante o treinamento da política de 40 MHz, apresentada na Figura 16. Durante o aprendizado, o valor médio de α_t deslocou-se progressivamente para valores elevados e a faixa entre os percentis 5 e 95 tornou-se progressivamente mais estreita. Na execução da política congelada no ns-3/5G-LENA, essa preferência aprendida também se manifesta, com o coeficiente permanecendo concentrado na mesma região do domínio.

A comparação entre treinamento e inferência deve, entretanto, ser interpretada de forma qualitativa. No treinamento, a política atua sobre os estados produzidos pelo ambiente utilizado para o aprendizado, enquanto na avaliação sistêmica os estados são construídos a partir da execução no ns-3/5G-LENA. A proximidade dos valores mostra que a preferência aprendida por valores elevados de α_t permanece presente nessa realização do simulador, mas não implica equivalência entre as duas condições.

Considerando a formulação da recompensa, valores de α_t próximos de 1 atribuem maior peso ao termo associado à justiça em relação ao termo de eficiência. O comportamento observado é coerente com o elevado Índice de Jain, com o número efetivo de UEs próximo de nove e com a participação ampla dos usuários na distribuição de RBGs

em 40 MHz. Essa associação, entretanto, não permite estabelecer uma relação causal isolada entre α_t e esses resultados, uma vez que os logits responsáveis pela distribuição dos recursos e o coeficiente α_t são produzidos conjuntamente pela mesma política.

Assim, a formulação do IA-FMR permite que o parâmetro associado ao compromisso entre justiça e eficiência seja produzido a cada decisão, mas a política treinada para 40 MHz não explora de forma ampla diferentes pontos desse domínio na realização analisada. Seu comportamento permanece fortemente concentrado em valores elevados de α_t , caracterizando um ponto de operação fortemente orientado ao termo de justiça.

Além disso, α_{t-1} integra o estado utilizado na decisão seguinte. Sem uma ablação que fixe, remova ou controle independentemente esse componente, não é possível isolar sua contribuição causal sobre as alocações subsequentes. Os resultados devem, portanto, ser interpretados como comportamento da política aprendida como um todo.

Por fim, a Figura 32 corresponde exclusivamente à semente 47889. Ela não contém intervalo de confiança e não representa a média de α_t entre as 100 sementes da Avaliação por Largura de Banda. Seu objetivo é caracterizar temporalmente o comportamento do coeficiente em uma realização específica da política de 40 MHz.

6.2.7 Síntese da Avaliação por Largura de Banda

A avaliação por largura de banda mostra comportamentos distintos entre os escalonadores. O MR privilegia a vazão agregada, mas apresenta maior concentração dos recursos e menor atendimento dos fluxos menos favorecidos. O RR mantém distribuição próxima da uniformidade, enquanto o PF alterna os usuários atendidos ao longo do tempo e ocupa uma região intermediária em termos de justiça e concentração.

O IA-FMR mantém, na maior parte das configurações, distribuição de RBGs próxima do equilíbrio e atendimento amplo dos UEs, ao mesmo tempo em que apresenta médias de vazão agregada superiores às observadas para RR e PF e, a partir de 15 MHz, a maior média de P5 entre os mecanismos avaliados. A comparação entre a alocação alvo e a aplicada mostra ainda que, entre 15 e 100 MHz, a decisão produzida pela política é preservada integralmente ou sofre apenas pequenas alterações durante sua realização na camada MAC.

A configuração de 10 MHz constitui a principal exceção. Nessa condição, parte da alocação produzida pela política não é preservada na aplicação, reduzindo o número de UEs atendidos e o Índice de Jain e aumentando a frequência de demanda pendente sem

alocação. A análise temporal de 40 MHz, por sua vez, confirma em uma realização específica o comportamento amplo de atendimento do IA-FMR e mostra que α_t permanece concentrado em valores próximos de 1, caracterizando um ponto de operação fortemente orientado à justiça.

6.3 Avaliação de Generalização Estrutural por Número de UEs

Esta seção avalia o comportamento do IA-FMR em 40 MHz quando o número de UEs varia entre 3, 5, 7 e 9. A política utilizada mantém a arquitetura definida para uma dimensão máxima de nove usuários, enquanto as posições não ocupadas são tratadas conforme o mecanismo de preenchimento descrito no Capítulo 4.

O objetivo desta avaliação é verificar se a mesma política continua operando quando aplicada a conjuntos menores de usuários dentro da dimensão para a qual foi projetada. O experimento caracteriza, portanto, uma avaliação de generalização estrutural para cardinalidades inferiores a nove UEs, e não uma análise de escalabilidade para quantidades arbitrariamente maiores de usuários. A avaliação considera exclusivamente o IA-FMR, pois o objetivo não é realizar uma nova comparação entre escalonadores, mas examinar o comportamento da política quando a quantidade de usuários ativos é alterada.

6.3.1 Vazão Agregada

A Figura 33 apresenta a vazão agregada média do IA-FMR em função do número de UEs. O eixo horizontal representa as quatro cardinalidades avaliadas e o eixo vertical apresenta a vazão agregada média, em Mbps. A região sombreada representa o intervalo de confiança de 95%.

A vazão agregada média aumenta de 101,72 Mbps com 3 UEs para 149,75 Mbps com 5 UEs. A partir desse ponto, a tendência se inverte: a média diminui para 127,39 Mbps com 7 UEs e para 97,53 Mbps com 9 UEs. Dessa forma, o aumento do número de usuários não produz crescimento monotônico da vazão agregada nessa configuração.

Quando a vazão agregada é normalizada pelo número de UEs, observa-se redução progressiva do valor médio por usuário: aproximadamente 33,91 Mbps com 3 UEs, 29,95 Mbps com 5 UEs, 18,20 Mbps com 7 UEs e 10,84 Mbps com 9 UEs. Assim, o maior valor agregado observado com 5 UEs não corresponde a melhoria do desempenho individual à

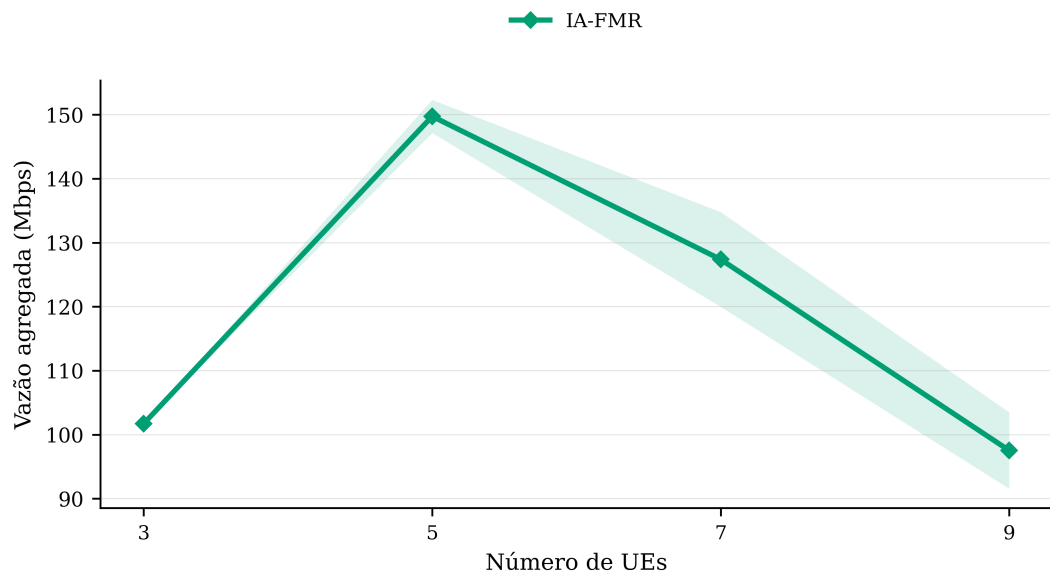


Figura 33: Vazão agregada média do IA-FMR em 40 MHz em função do número de UEs. A região sombreada representa o intervalo de confiança de 95%.

medida que novos usuários são adicionados.

Os intervalos de confiança são mais amplos nas configurações de 5, 7 e 9 UEs do que no caso de 3 UEs, com maior semiamplitude na configuração com 7 UEs. Esse comportamento caracteriza maior variabilidade entre as realizações nessas configurações, sem permitir identificar sua causa.

Esses resultados mostram que a política treinada para uma dimensão máxima de nove UEs continua produzindo decisões operacionais quando aplicada a conjuntos menores. A variação da vazão, entretanto, evidencia que a generalização estrutural da política não implica invariância de desempenho quando o número de usuários é alterado.

6.3.2 Justiça na Alocação de RBGs

A Figura 34 apresenta o Índice de Jain calculado sobre os RBGs distribuídos entre os UEs considerados em cada configuração. O eixo horizontal representa o número de UEs e o eixo vertical apresenta o valor médio do índice. A região sombreada representa o intervalo de confiança de 95%. Valores próximos de 1 indicam maior uniformidade na distribuição dos recursos.

O IA-FMR mantém valores elevados do Índice de Jain nas quatro configurações avaliadas. Os índices médios foram de 0,9766, 0,9727 e 0,9731 para 3, 5 e 7 UEs, respectivamente. A configuração com 9 UEs apresentou o maior valor, de 0,9943. Dessa forma, a

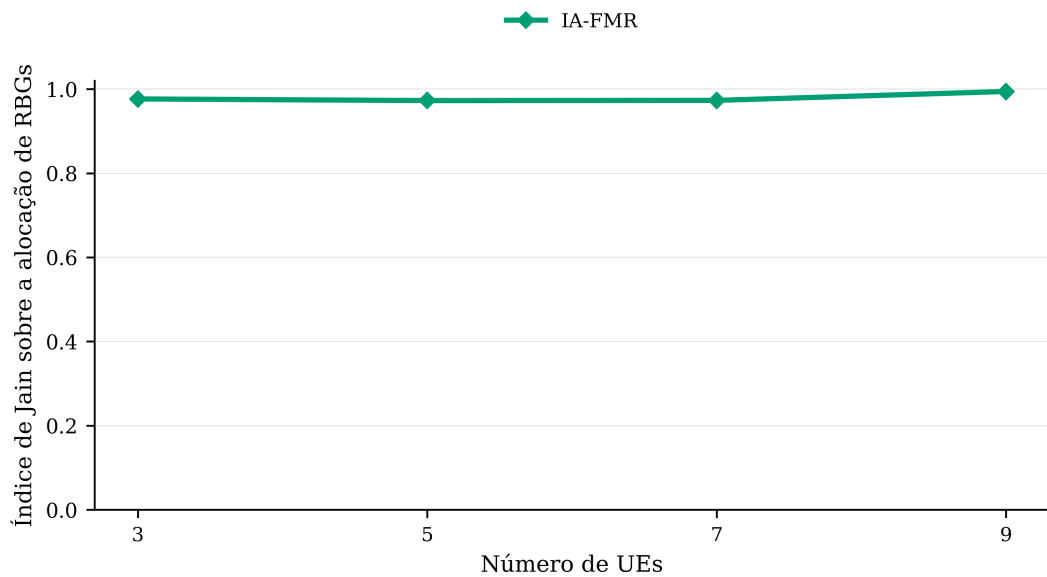


Figura 34: Índice de Jain sobre a alocação de RBGs do IA-FMR em 40 MHz em função do número de UEs. A região sombreada representa o intervalo de confiança de 95%.

redução do conjunto de usuários não produz concentração acentuada dos RBGs, embora a distribuição seja ligeiramente menos uniforme nas configurações com menos de nove UEs.

Entre 3, 5 e 7 UEs, as diferenças entre as médias são pequenas e não apresentam tendência monotônica com o aumento do número de usuários. Esse comportamento contrasta com a variação mais pronunciada observada na vazão agregada, que alcança seu maior valor com 5 UEs e diminui nas configurações seguintes. Portanto, as alterações de vazão observadas no experimento não são acompanhadas por uma redução progressiva da justiça na distribuição dos recursos.

Os intervalos de confiança permanecem estreitos nas quatro configurações, com semi-amplitudes entre aproximadamente 0,0003 e 0,0022. Eles caracterizam a incerteza associada às médias obtidas a partir das 100 realizações de cada configuração e não são utilizados, nesta análise, como teste formal de diferença entre as diferentes quantidades de UEs.

Os resultados são consistentes com o objetivo desta avaliação de generalização estrutural. A mesma arquitetura de política, definida para uma dimensão máxima de nove UEs, mantém elevada uniformidade na distribuição dos RBGs quando aplicada a conjuntos de 3, 5 e 7 usuários. Essa observação se restringe às cardinalidades avaliadas dentro da dimensão prevista pela política e não demonstra generalização para quantidades superiores a nove UEs.

6.3.3 Atendimento dos UEs

A Figura 35 apresenta o número médio de UEs que receberam ao menos um RBG por *slot* em cada configuração. O eixo horizontal representa o número de UEs considerados e o eixo vertical apresenta o número médio de UEs servidos por *slot*. A região sombreada representa o intervalo de confiança de 95%.

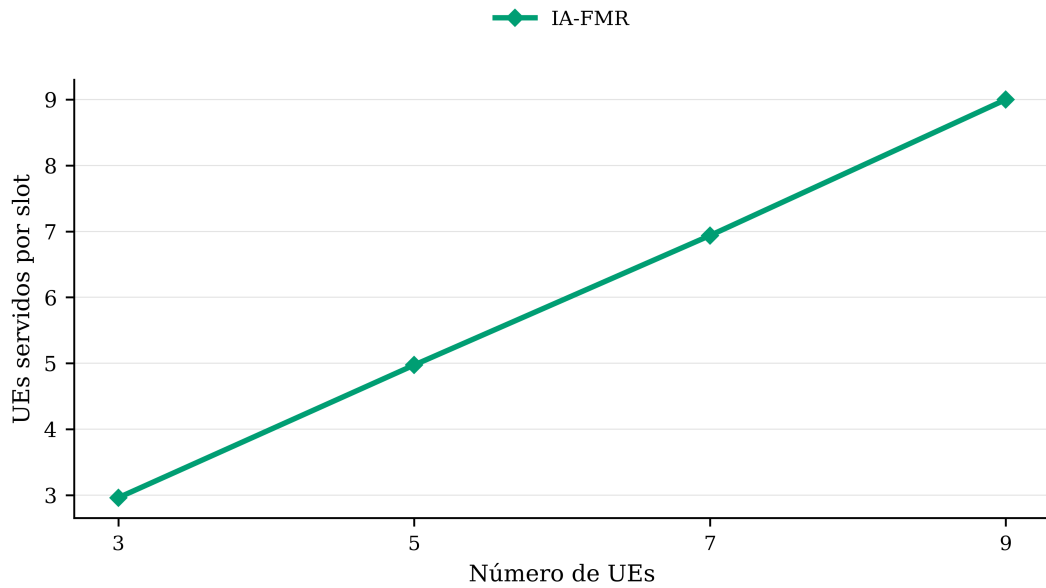


Figura 35: Número médio de UEs servidos por *slot* pelo IA-FMR em 40 MHz em função do número de UEs. A região sombreada representa o intervalo de confiança de 95%.

O número médio de UEs servidos acompanha de perto a quantidade de usuários considerada em cada configuração. As médias são de 2,960, 4,972, 6,937 e 9,000 UEs servidos por *slot* para os casos com 3, 5, 7 e 9 UEs, respectivamente. Assim, o IA-FMR mantém atendimento amplo do conjunto de usuários nas quatro cardinalidades avaliadas.

A análise de demanda pendente sem alocação complementa esse resultado. Nas quatro configurações, o percentual de registros com `buf_req > 0` e `alloc_rbg == 0` é igual a zero. Portanto, não foram observados registros em que um UE com demanda positiva permanecesse sem alocação aplicada nas condições avaliadas.

Em conjunto com os valores elevados do Índice de Jain, esses resultados mostram que a aplicação da mesma arquitetura a conjuntos de 3, 5, 7 e 9 UEs preserva uma distribuição ampla do atendimento. A redução da vazão agregada observada nas configurações com 7 e 9 UEs não ocorre acompanhada pela exclusão de usuários com demanda pendente da alocação aplicada.

6.3.4 Vazão dos Fluxos Menos Favorecidos

A Figura 36 apresenta o percentil 5 da vazão por fluxo em função do número de UEs. O eixo horizontal representa as quatro cardinalidades avaliadas e o eixo vertical apresenta o P5 médio, em Mbps. A região sombreada representa o intervalo de confiança de 95%. Essa métrica caracteriza a região inferior da distribuição e complementa a análise da vazão agregada da célula.

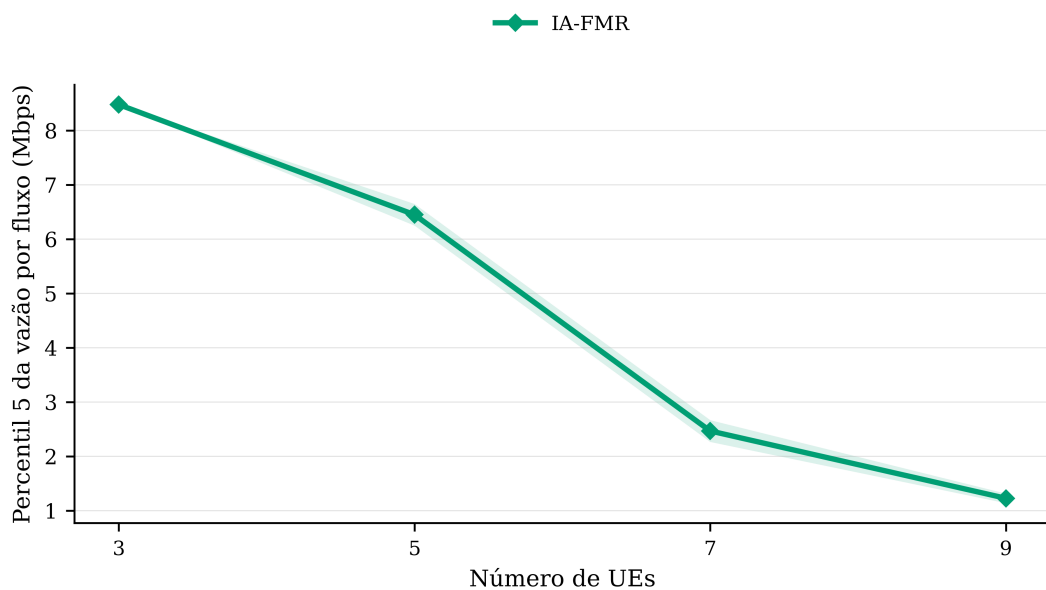


Figura 36: Percentil 5 da vazão por fluxo do IA-FMR em 40 MHz em função do número de UEs. A região sombreada representa o intervalo de confiança de 95%.

O P5 diminui progressivamente à medida que o número de UEs aumenta. A média passa de 8,477 Mbps com 3 UEs para 6,449 Mbps com 5 UEs, correspondendo a uma redução de 23,92%. Com 7 UEs, o P5 diminui para 2,466 Mbps, uma queda de 61,76% em relação à configuração anterior. Com 9 UEs, o valor alcança 1,225 Mbps, representando redução adicional de 50,30%.

Entre as configurações com 3 e 9 UEs, o P5 diminui 85,54%. Esse comportamento mostra que a preservação do acesso aos recursos não implica manutenção do mesmo nível de vazão na região inferior da distribuição. A política continua atribuindo recursos aos UEs, mas a vazão dos fluxos menos favorecidos diminui conforme a cardinalidade do conjunto avaliado aumenta.

A configuração com 5 UEs evidencia essa distinção. Embora apresente a maior vazão agregada média do experimento, de 149,75 Mbps, seu P5 é inferior ao observado com 3 UEs. Dessa forma, o aumento da taxa total entre essas duas configurações não se distribui

de maneira uniforme por toda a distribuição de vazão dos fluxos.

Os intervalos de confiança também mostram baixa incerteza no caso de 3 UEs e maior dispersão nas configurações seguintes. As semi-amplitudes são de aproximadamente 0,00002 Mbps, 0,201 Mbps, 0,203 Mbps e 0,088 Mbps para 3, 5, 7 e 9 UEs, respectivamente. Esses intervalos caracterizam a incerteza de cada média e não são utilizados como teste formal de diferença entre as configurações.

Os resultados reforçam a distinção entre generalização estrutural e invariância de desempenho. A política mantém elevada justiça e atendimento amplo quando aplicada a 3, 5, 7 e 9 UEs, mas o desempenho dos fluxos menos favorecidos se reduz de forma acentuada conforme aumenta o número de usuários considerados.

6.3.5 Perdas e Atraso

As Figuras 37 e 38 apresentam, respectivamente, a taxa média de perda e o atraso médio dos pacotes recebidos em função do número de UEs. Em ambas, o eixo horizontal representa as quatro cardinalidades avaliadas e a região sombreada representa o intervalo de confiança de 95%. Na Figura 37, o eixo vertical apresenta a taxa média de perda, em percentual. Na Figura 38, o eixo vertical apresenta o atraso médio dos pacotes recebidos, em milissegundos.

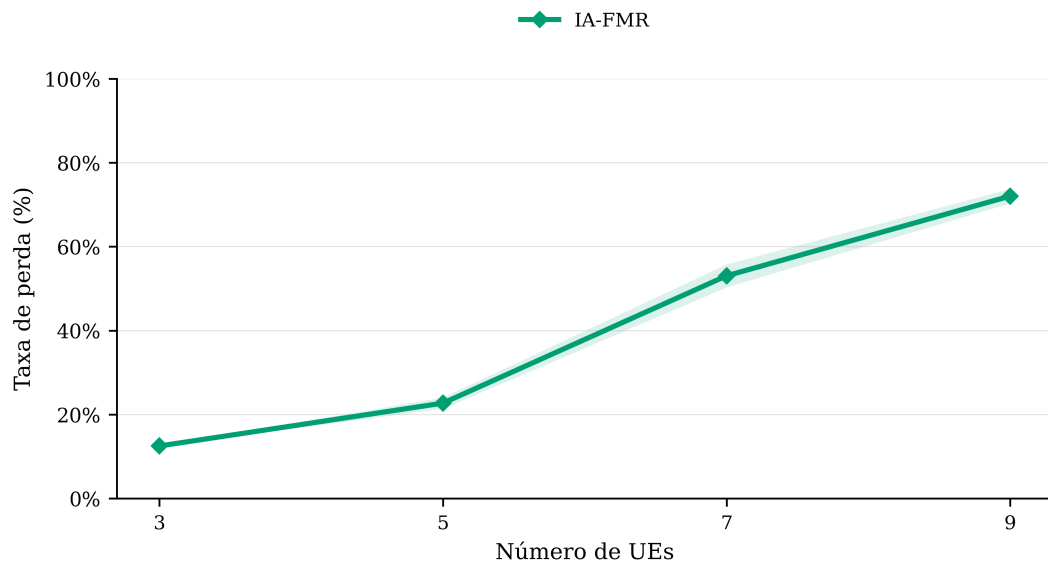


Figura 37: Taxa média de perda do IA-FMR em 40 MHz em função do número de UEs. A região sombreada representa o intervalo de confiança de 95%.

Na Figura 37, a taxa média de perda cresce continuamente com o número de usuários.

O valor passa de 12,51% com 3 UEs para 22,73% com 5 UEs. Com 7 UEs, a perda alcança 53,05% e aumenta para 72,04% na configuração com 9 UEs.

O maior aumento entre configurações consecutivas ocorre na transição de 5 para 7 UEs, com elevação de 30,32 pontos percentuais. Entre 3 e 9 UEs, a taxa média de perda aumenta 59,53 pontos percentuais. Essa evolução acompanha a redução do P5 e, a partir de 5 UEs, também a diminuição da vazão agregada observada anteriormente.

O aumento da perda não é acompanhado pela exclusão de usuários com demanda pendente da alocação aplicada. Como mostrado na subseção anterior, o IA-FMR continua atendendo praticamente todo o conjunto considerado em cada *slot*, e a frequência de demanda pendente sem alocação permanece nula nas quatro configurações. Assim, preservar o acesso aos recursos não impede que a parcela de pacotes não recebidos aumente à medida que a cardinalidade do conjunto avaliado cresce.

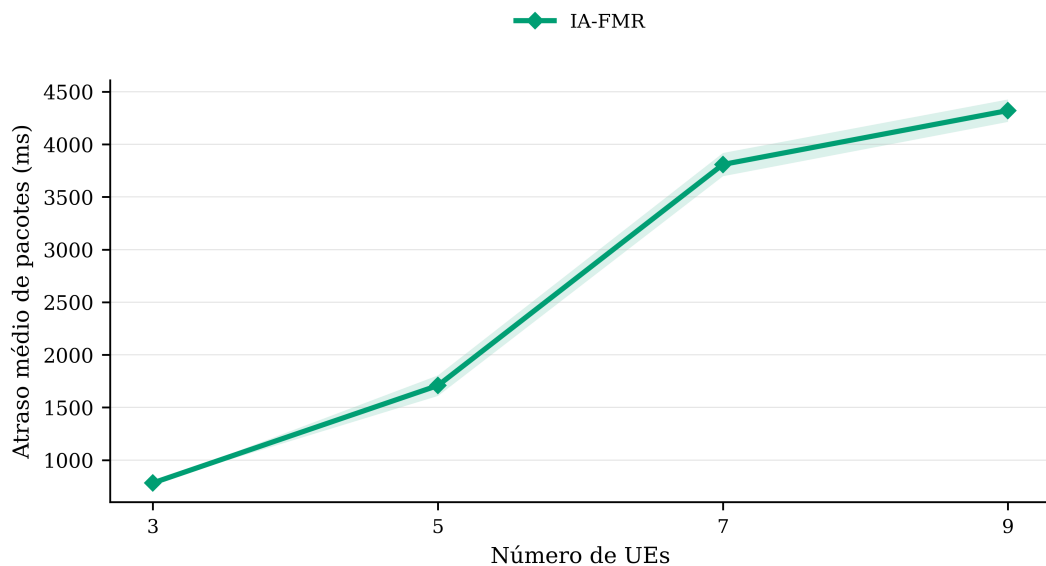


Figura 38: Atraso médio dos pacotes recebidos pelo IA-FMR em 40 MHz em função do número de UEs. A região sombreada representa o intervalo de confiança de 95%.

O atraso médio também aumenta com o número de usuários. A média passa de 782,98 ms com 3 UEs para 1.708,40 ms com 5 UEs. Na configuração com 7 UEs, o atraso alcança 3.808,77 ms e aumenta para 4.321,06 ms com 9 UEs.

Assim como na taxa de perda, a maior alteração ocorre entre 5 e 7 UEs. Nessa transição, o atraso médio aumenta 2.100,37 ms. Entre 7 e 9 UEs, o crescimento é menor, de 512,29 ms, embora a taxa de perda continue aumentando de forma expressiva.

Essa leitura deve considerar que o atraso médio é calculado apenas para os pacotes efetivamente recebidos. Os pacotes que não chegam ao destino não participam dessa

média e são refletidos pela taxa de perda. Dessa forma, o crescimento mais moderado do atraso entre 7 e 9 UEs não indica estabilização do desempenho temporal, pois nessa mesma transição aumenta a parcela de pacotes que não é recebida durante a execução.

Em conjunto, perda e atraso reforçam a distinção entre generalização estrutural e invariância de desempenho. O IA-FMR continua produzindo alocações com elevada justiça e atendimento amplo nas quatro cardinalidades, mas o aumento do número de UEs é acompanhado por redução do P5, aumento da perda e crescimento do atraso dos pacotes recebidos. A capacidade da política de operar com conjuntos menores dentro da dimensão prevista não implica, portanto, manutenção do mesmo nível de desempenho dos fluxos.

6.4 Avaliação de Disponibilidade Temporal em Tráfego Sintético de Vídeo

Esta seção avalia como as decisões dos escalonadores repercutem na entrega temporal de unidades sintéticas de vídeo. Diferentemente dos experimentos anteriores, que analisam diretamente métricas da rede, esta avaliação considera se as unidades geradas são recebidas dentro dos prazos temporais definidos para a aplicação.

A configuração é utilizada como um cenário de alta carga para produzir contenção entre os usuários e observar as consequências temporais das decisões de escalonamento. Ela não representa um dimensionamento destinado a garantir a qualidade de uma aplicação XR ou de um serviço comercial de vídeo. O cenário considera seis UEs, largura de banda de 100 MHz e geração de 60 unidades sintéticas por segundo. Cada UE possui taxa nominal de 30 Mbps, resultando em carga nominal agregada de 180 Mbps.

A comparação envolve RR, PF, MR e IA-FMR sob as mesmas condições experimentais e utiliza 100 sementes comuns aos quatro mecanismos. São considerados dois prazos de entrega, indicados nas figuras como *Packet Delay Budget* (PDB): 60 ms e 80 ms. O modelo de tráfego e a referência temporal adotada são fundamentados nas recomendações do 3GPP para avaliação de serviços de mídia e XR (3GPP, 2024b,a). O prazo de 60 ms é utilizado como referência principal por corresponder ao limite superior da faixa de latência de *downlink* considerada no modelo de referência adotado (3GPP, 2024b). O prazo de 80 ms constitui uma escolha metodológica para avaliar a sensibilidade dos resultados a um limite temporal menos restritivo.

Uma unidade sintética de vídeo é considerada temporalmente útil quando sua recepção completa é concluída dentro do prazo definido a partir do respectivo instante de geração.

Unidades recebidas após esse limite são classificadas como tardias. As unidades que não chegam ao destino até o término da simulação compõem a terceira categoria.

A principal métrica de entrega temporal utilizada nesta avaliação é a *Taxa de Unidades no Prazo* (TUP). Para um prazo de entrega d , a métrica representa a proporção das unidades sintéticas geradas cuja recepção completa é concluída em até d após o respectivo instante de geração. A TUP é definida por:

$$\text{TUP}_d = \frac{N_{\text{prazo},d}}{N_{\text{geradas}}} \times 100, \quad (6.2)$$

em que $N_{\text{prazo},d}$ representa o número de unidades recebidas dentro do prazo d e N_{geradas} corresponde ao número total de unidades geradas. Nesta tese, a TUP é calculada para $d = 60$ ms e $d = 80$ ms. As unidades recebidas após o prazo e aquelas não recebidas até o término da simulação permanecem no denominador, mas não contribuem para o numerador.

As métricas desta seção são calculadas após a simulação a partir dos registros das unidades sintéticas e não integram o treinamento do IA-FMR. Dessa forma, a avaliação verifica se as diferenças produzidas pelo escalonamento também se manifestam na entrega temporal dos dados no nível da aplicação, embora esse objetivo não tenha sido incorporado à função de recompensa.

6.4.1 Entrega Temporal das Unidades de Vídeo

A Figura 39 apresenta a composição das unidades sintéticas de vídeo nos dois prazos avaliados. O painel (a) corresponde a 60 ms e o painel (b) a 80 ms. Em ambos, o eixo horizontal identifica os escalonadores e o eixo vertical representa a proporção das unidades geradas. Cada barra é dividida entre unidades recebidas no prazo, unidades tardias e unidades não recebidas. As três parcelas totalizam 100% para cada escalonador. As barras de erro posicionadas na fronteira entre as unidades recebidas no prazo e as unidades tardias representam o intervalo de confiança de 95% da TUP. As barras localizadas na fronteira seguinte representam o intervalo de confiança de 95% da proporção total de unidades recebidas até o término da simulação, correspondente à soma das unidades recebidas no prazo e tardias. Não há intervalo no topo das colunas porque as três categorias totalizam, por definição, 100%.

Para o prazo de 60 ms, o IA-FMR apresenta a maior TUP média, com 66,88%. RR,

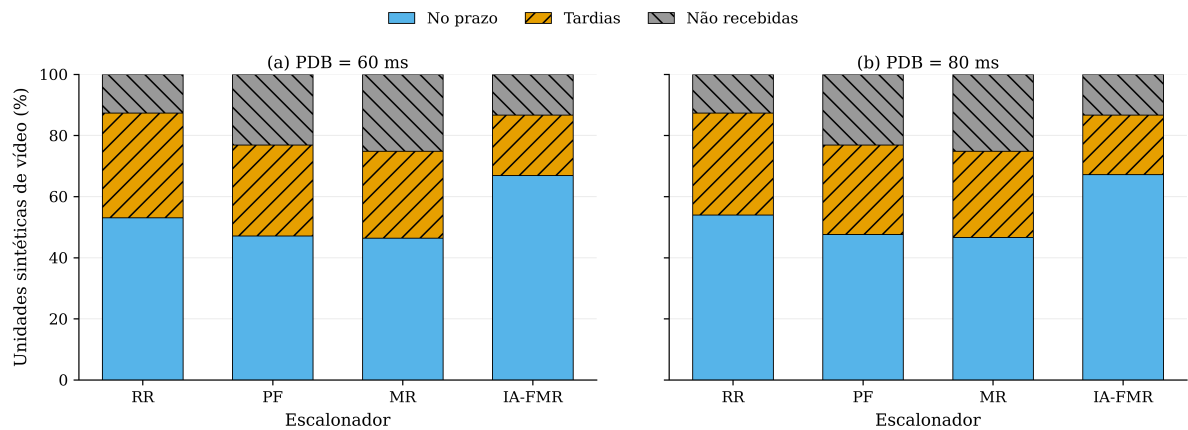


Figura 39: Composição das unidades sintéticas de vídeo considerando os prazos de 60 e 80 ms. As barras de erro representam os intervalos de confiança de 95% da TUP e da proporção total de unidades recebidas.

PF e MR alcançam 53,10%, 47,12% e 46,41%, respectivamente. No IA-FMR, 20,00% das unidades são recebidas após o prazo e 13,12% não são recebidas. No RR, essas parcelas correspondem a 34,30% e 12,60%. PF apresenta 29,78% de unidades tardias e 23,10% de unidades não recebidas, enquanto no MR os valores são 28,39% e 25,20%.

A comparação entre RR e IA-FMR ajuda a interpretar a diferença na TUP. A parcela de unidades não recebidas é semelhante nos dois mecanismos, 12,60% no RR e 13,12% no IA-FMR. A maior diferença aparece entre as unidades tardias, cuja proporção diminui de 34,30% no RR para 20,00% no IA-FMR. Assim, a maior TUP média do IA-FMR nessa comparação está associada principalmente a uma parcela maior das unidades recebidas que conclui sua entrega dentro do prazo.

Quando o prazo é ampliado para 80 ms, as médias da TUP passam para 54,00% no RR, 47,58% no PF, 46,64% no MR e 67,20% no IA-FMR. As alterações em relação a 60 ms são pequenas nos quatro mecanismos. Como a classificação de unidade não recebida independe desse limite temporal, a ampliação do prazo atua apenas sobre a divisão entre unidades no prazo e unidades tardias. Os resultados mostram, portanto, que somente uma pequena parcela das unidades que ultrapassam 60 ms conclui sua recepção nos 20 ms adicionais.

6.4.2 Frequência e Severidade dos Congelamentos Inferidos

A TUP descreve a quantidade de unidades que satisfaz o prazo, mas não informa como as unidades que deixam de satisfazê-lo se distribuem ao longo da execução. Para caracterizar

essa dimensão, são analisados eventos de congelamento inferidos. Após a primeira unidade recebida dentro do prazo, uma sequência contínua de unidades que não satisfazem o limite temporal é classificada como um evento de congelamento. Como não é modelado um reprodutor real, esses eventos constituem uma aproximação construída a partir da sequência temporal das unidades.

A Figura 40 apresenta a frequência média desses eventos. O eixo horizontal identifica os escalonadores e o eixo vertical representa o número de eventos por minuto. Para cada escalonador, a barra sem hachura corresponde ao prazo de 60 ms e a barra hachurada ao prazo de 80 ms. A métrica é agregada sobre as 100 sementes de cada configuração, incluindo os casos em que nenhum UE inicia a reprodução inferida, para os quais a frequência de eventos é nula. As barras de erro representam os intervalos de confiança de 95% calculados sobre essas 100 sementes.

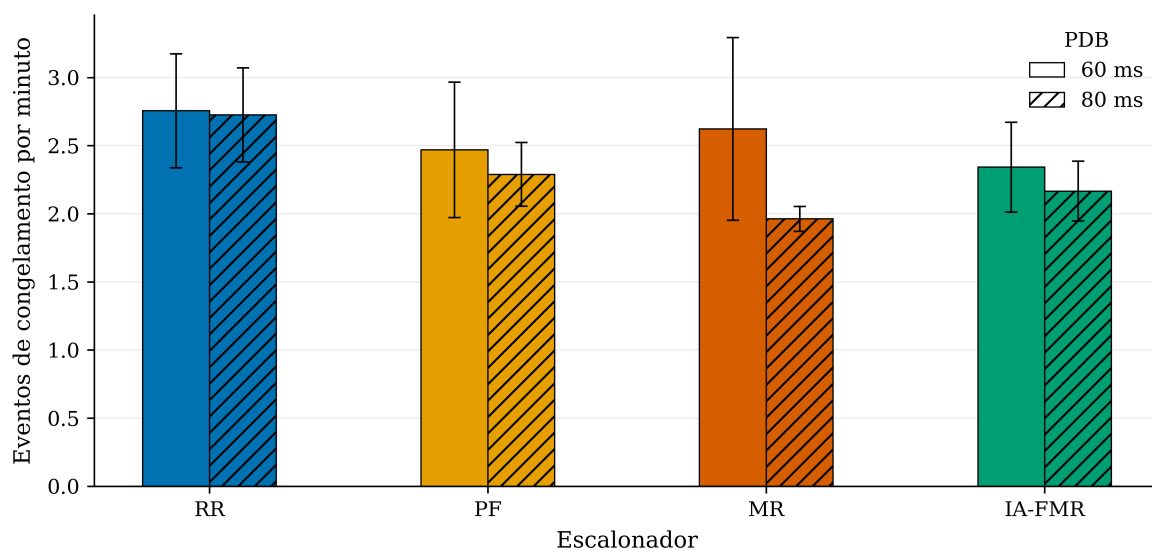


Figura 40: Frequência dos eventos de congelamento inferidos para os prazos de 60 e 80 ms, com intervalo de confiança de 95% calculado sobre as 100 sementes.

Para 60 ms, as médias são de 2,76 eventos por minuto no RR, 2,47 no PF, 2,62 no MR e 2,34 no IA-FMR. As diferenças entre os mecanismos são menores nessa métrica do que na TUP. Para 80 ms, os valores passam para 2,73 eventos por minuto no RR, 2,29 no PF, 1,96 no MR e 2,17 no IA-FMR. Nesse prazo, o MR apresenta a menor frequência média de eventos.

A frequência, entretanto, não caracteriza sozinha a severidade dos congelamentos. Um mecanismo pode apresentar menos eventos, mas produzir episódios de maior duração. A

Figura 41 complementa essa análise com a duração P95 dos eventos.

O eixo horizontal novamente identifica os escalonadores e o eixo vertical apresenta a duração P95 em milissegundos. As barras sem hachura representam 60 ms e as barras hachuradas representam 80 ms. Para cada semente, os eventos observados nos seis UEs são reunidos antes do cálculo do percentil 95. Sequências que permanecem sem satisfazer o prazo até o término da simulação são contabilizadas como eventos terminais. Pelo procedimento definido, sementes com reprodução iniciada e ausência de congelamentos receberiam P95 igual a zero, enquanto sementes nas quais nenhum UE inicia a reprodução não participam da média. No conjunto avaliado, não houve sementes com reprodução iniciada e ausência completa de congelamentos. Assim, o conjunto efetivamente utilizado nessa métrica coincide com as sementes em que ocorreu pelo menos um evento de congelamento.

Para o prazo de 60 ms, a duração P95 é calculada sobre 90 sementes no RR, 90 no PF, 96 no MR e 91 no IA-FMR. Para 80 ms, são utilizadas 96 sementes no RR, 97 no PF, 99 no MR e 97 no IA-FMR. A barra apresentada na figura corresponde à média dos percentis obtidos nessas sementes válidas e as barras de erro representam o intervalo de confiança de 95% calculado sobre o mesmo conjunto. O percentil é obtido por interpolação linear e, quando existe apenas um evento, o P95 coincide com sua duração.

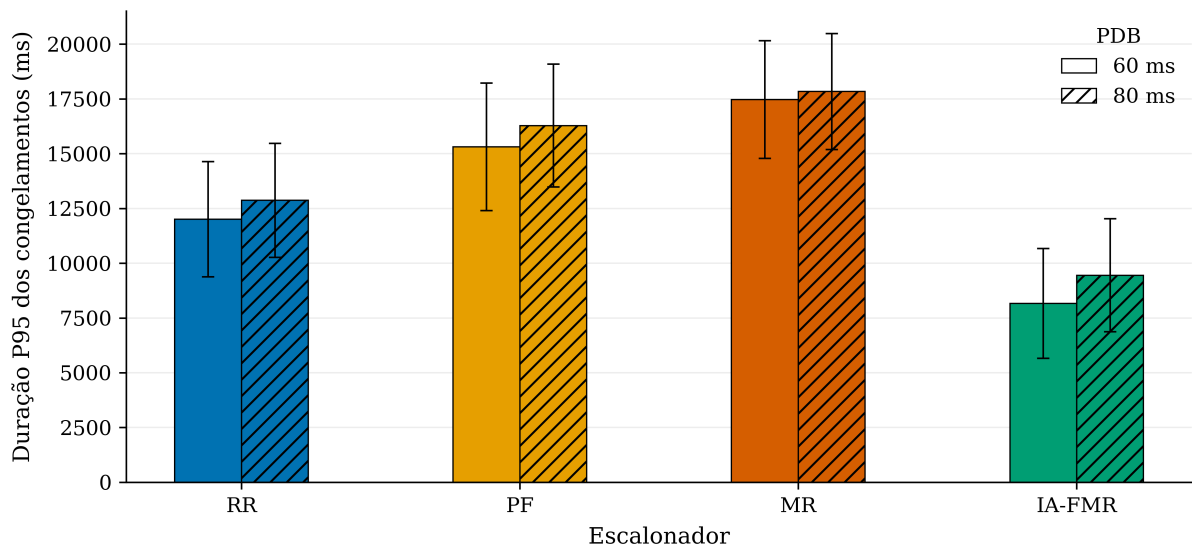


Figura 41: Duração P95 dos eventos de congelamento inferidos para os prazos de 60 e 80 ms, com intervalo de confiança de 95% calculado sobre as sementes em que houve pelo menos um evento de congelamento.

Para o prazo de 60 ms, o IA-FMR apresenta a menor duração P95 média, com 8.171,26 ms, equivalente a aproximadamente 8,17 s. O RR alcança 12.016,90 ms, enquanto PF e MR apresentam 15.315,15 ms e 17.477,98 ms, respectivamente.

No prazo de 80 ms, as durações P95 médias são de 12.875,52 ms no RR, 16.285,40 ms no PF, 17.843,26 ms no MR e 9.454,52 ms no IA-FMR. Nesse caso, o MR combina a menor frequência média de eventos com a maior duração P95 média. Os congelamentos inferidos são, portanto, menos frequentes nessa métrica, mas os episódios que permanecem classificados como congelamento incluem ocorrências prolongadas.

O aumento da duração P95 entre 60 e 80 ms não significa que a ampliação do prazo tenha piorado o comportamento temporal. Ao elevar o limite, parte dos episódios mais curtos deixa de ser classificada como congelamento. O conjunto remanescente passa a conter proporcionalmente mais eventos longos, o que pode elevar o percentil calculado sobre essa distribuição condicionada.

Mesmo no IA-FMR, os valores de P95 permanecem elevados. O resultado deve, portanto, ser interpretado como uma redução relativa da duração dos episódios mais severos no cenário avaliado, e não como eliminação dos congelamentos.

6.4.3 Usuários Temporalmente Satisfeitos

As métricas anteriores caracterizam a entrega temporal das unidades e a frequência e a duração dos congelamentos inferidos. Para verificar quantos usuários alcançam individualmente um critério mais restritivo, um UE é classificado como temporalmente satisfeito quando

$$\text{TUP}_d^{(u)} \geq 99\%, \quad (6.3)$$

em que $\text{TUP}_d^{(u)}$ representa a TUP do UE u para o prazo d . A métrica é calculada em cada semente como a proporção dos seis UEs que satisfazem esse critério.

A Figura 42 apresenta essa proporção. O eixo horizontal identifica os escalonadores e o eixo vertical representa o percentual de UEs classificados como temporalmente satisfeitos. As barras sem hachura correspondem a 60 ms e as barras hachuradas a 80 ms. As barras de erro apresentam os intervalos de confiança de 95% entre as sementes.

Para o prazo de 60 ms, as médias são de 40,00% no RR, 39,67% no PF e 41,17% no MR. No IA-FMR, a proporção média é de 59,17%. Para 80 ms, os valores passam para

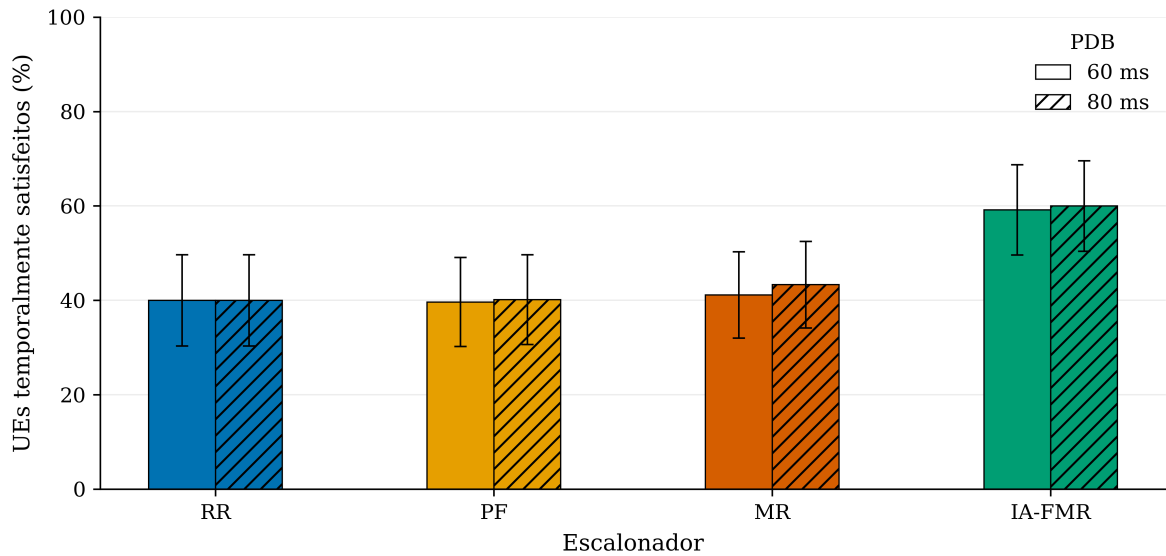


Figura 42: Percentual de usuários temporalmente satisfeitos para os prazos de 60 e 80 ms, considerando $TUP \geq 99\%$ e intervalo de confiança de 95%.

40,00% no RR, 40,17% no PF, 43,33% no MR e 60,00% no IA-FMR.

O critério de 99% é mais restritivo do que a observação da TUP média da célula. Por esse motivo, uma média elevada de unidades recebidas no prazo não implica que todos os usuários satisfaçam individualmente o limiar definido. Os intervalos de confiança também permanecem amplos, indicando variação entre as sementes. As comparações são, portanto, utilizadas de forma descritiva.

6.4.4 Síntese da Avaliação de Disponibilidade Temporal

A avaliação com tráfego sintético de vídeo mostra diferenças entre os escalonadores também na dimensão temporal. Nas médias observadas, o IA-FMR apresenta a maior TUP e a maior proporção de UEs que atendem ao critério de satisfação temporal nos dois prazos avaliados. A duração P95 média dos congelamentos inferidos também permanece inferior à observada nos três mecanismos de referência.

A frequência dos congelamentos apresenta diferenças menores e não mantém a mesma ordenação entre os escalonadores nos dois prazos. Em particular, o MR apresenta a menor frequência média para 80 ms, mas simultaneamente a maior duração P95 média. Esse comportamento mostra que frequência e duração descrevem aspectos distintos da sequência temporal das unidades.

A ampliação do prazo de 60 para 80 ms produz mudanças pequenas na TUP e na proporção de usuários temporalmente satisfeitos. Os resultados caracterizam a entrega temporal das unidades no cenário sintético de alta carga considerado. Como não são modelados reprodução real, decodificação, *playout buffer* ou adaptação de vídeo, essa avaliação não deve ser interpretada como uma caracterização completa da QoE percebida pelo usuário.

6.5 Considerações Finais do Capítulo

Esta seção reúne os principais resultados das três avaliações apresentadas neste capítulo. O objetivo não é reproduzir as análises anteriores, mas destacar os comportamentos observados nas diferentes configurações. As comparações permanecem descritivas e devem ser interpretadas considerando as condições experimentais e os intervalos de confiança apresentados ao longo do capítulo.

Na avaliação por largura de banda, os escalonadores apresentaram pontos de operação distintos entre vazão agregada, distribuição dos recursos e atendimento dos fluxos menos favorecidos. Para evitar a combinação de configurações fisicamente distintas em uma média única, a Tabela 19 utiliza 40 MHz como referência descritiva. Essa configuração também foi utilizada na análise temporal e na avaliação por número de UEs.

Tabela 19: Síntese da avaliação por largura de banda na configuração de 40 MHz.

Escalonador	Vazão agregada (Mbps)	Índice de Jain dos RBGs	P5 por fluxo (Mbps)
RR	50,93	0,9988	0,714
PF	34,36	0,5926	0,337
MR	120,40	0,3711	≈ 0
IA-FMR	97,53	0,9943	1,225

Em 40 MHz, o MR apresenta a maior vazão agregada média, mas também o menor Índice de Jain e P5 praticamente nulo. O RR mantém distribuição praticamente uniforme dos RBGs, com vazão agregada inferior à observada para MR e IA-FMR. O IA-FMR combina Índice de Jain próximo ao do RR com a maior média de P5 entre os mecanismos avaliados nessa configuração e vazão agregada média situada entre RR e MR.

A avaliação completa entre 10 e 100 MHz mostra que esse comportamento não é uniforme em todas as bandas. A configuração de 10 MHz constitui a principal exceção para o IA-FMR. Nessa condição, a política produz alocação positiva para os nove UEs, mas

parte das concessões não é preservada durante sua realização na camada MAC. O número médio de UEs atendidos e o Índice de Jain da alocação aplicada são, conseqüentemente, inferiores aos observados na distribuição alvo.

A partir de 15 MHz, a correspondência entre a alocação produzida pela política e sua realização torna-se elevada. Entre 15 e 30 MHz, alvo e aplicação são integralmente idênticos. Entre 40 e 100 MHz, o conjunto de UEs atendidos é preservado e as diferenças nas quantidades de RBGs são residuais. Assim, na maior parte das configurações avaliadas, a distribuição ampla dos recursos já está presente na saída da política e é preservada durante sua aplicação no simulador.

A segunda avaliação examinou a generalização estrutural do IA-FMR quando a mesma arquitetura, definida para uma dimensão máxima de nove UEs, é aplicada a conjuntos com 3, 5, 7 e 9 usuários. As Tabelas 20 e 21 resumem os resultados obtidos em 40 MHz.

Tabela 20: Desempenho do IA-FMR em função do número de UEs.

UEs	Vazão agregada (Mbps)	P5 por fluxo (Mbps)	Perda (%)	Atraso médio (ms)
3	101,72	8,477	12,51	782,98
5	149,75	6,449	22,73	1.708,40
7	127,39	2,466	53,05	3.808,77
9	97,53	1,225	72,04	4.321,06

Tabela 21: Distribuição e atendimento do IA-FMR em função do número de UEs.

UEs	Índice de Jain	UEs servidos por <i>slot</i>	Demanda sem alocação (%)
3	0,9766	2,960	0,00
5	0,9727	4,972	0,00
7	0,9731	6,937	0,00
9	0,9943	9,000	0,00

A política continua operando nas quatro cardinalidades avaliadas e mantém Índice de Jain elevado e atendimento próximo ao número total de UEs considerado. A demanda pendente sem alocação permanece nula nas quatro configurações. Esses resultados mostram que a redução da cardinalidade dentro da dimensão prevista pela arquitetura não provoca concentração acentuada do atendimento.

A preservação da distribuição dos recursos, entretanto, não implica invariância de desempenho. A vazão agregada cresce entre 3 e 5 UEs e diminui nas configurações seguintes. O P5 apresenta redução contínua, passando de 8,477 Mbps com 3 UEs para 1,225 Mbps

com 9 UEs. No mesmo intervalo, a perda aumenta de 12,51% para 72,04% e o atraso médio dos pacotes recebidos passa de 782,98 ms para 4.321,06 ms. A avaliação demonstra, portanto, generalização estrutural para conjuntos menores dentro da dimensão prevista, mas não manutenção do mesmo nível de desempenho quando a cardinalidade é alterada.

A terceira avaliação examinou a entrega temporal em um cenário sintético de vídeo de alta carga. A Tabela 22 resume o prazo principal de 60 ms. A TUP representa a parcela das unidades geradas cuja recepção completa ocorre dentro do prazo, enquanto as parcelas tardias e não recebidas permitem identificar como as demais unidades se distribuem. A duração P95 apresentada na tabela corresponde à média calculada sobre as sementes válidas de cada escalonador, conforme o critério definido na Subseção 6.4.2, e não necessariamente sobre as 100 sementes completas.

Tabela 22: Principais resultados da avaliação com tráfego sintético de vídeo para o prazo de 60 ms.

Escalonador	TUP (%)	Tardias (%)	Não recebidas (%)	Duração P95 (s)	UEs satisfeitos (%)
RR	53,10	34,30	12,60	12,02	40,00
PF	47,12	29,78	23,10	15,32	39,67
MR	46,41	28,39	25,20	17,48	41,17
IA-FMR	66,88	20,00	13,12	8,17	59,17

Para 60 ms, o IA-FMR apresenta a maior TUP média e a maior proporção média de UEs que satisfazem o critério temporal de 99%. Também apresenta a menor duração P95 média dos eventos de congelamento inferidos. A comparação com o RR mostra que a principal diferença na composição das unidades está na parcela tardia, que passa de 34,30% no RR para 20,00% no IA-FMR, enquanto as proporções de unidades não recebidas permanecem próximas.

A ampliação do prazo para 80 ms produz mudanças pequenas na TUP e na proporção de usuários temporalmente satisfeitos. A frequência dos eventos de congelamento não mantém a mesma ordenação entre os mecanismos nos dois prazos. Em 80 ms, o MR apresenta a menor frequência média, mas também a maior duração P95 média. Essa combinação reforça que frequência e duração caracterizam aspectos distintos da sequência temporal das unidades.

Os três experimentos evidenciam, em conjunto, que justiça na distribuição dos RBGs, vazão agregada, atendimento dos UEs e desempenho temporal não evoluem necessariamente na mesma direção. O MR privilegia a vazão agregada ao concentrar recursos em um subconjunto menor de usuários. O RR mantém uma distribuição próxima da uni-

formidade. O PF atende subconjuntos menores de usuários por decisão e, na realização temporal analisada, alterna os usuários atendidos ao longo do tempo. O IA-FMR apresenta, nas configurações avaliadas, uma distribuição ampla dos recursos e elevada justiça, ao mesmo tempo em que mantém médias de vazão e de P5 distintas das observadas pelos demais mecanismos.

Os resultados também delimitam o alcance da proposta. Em condições de maior escassez de recursos ou com maior número de usuários, o atendimento amplo não impede o aumento da perda, do atraso ou a redução da vazão dos fluxos menos favorecidos. Da mesma forma, no cenário sintético de vídeo, mesmo o IA-FMR apresenta eventos de congelamento inferidos com duração elevada em parte das sementes. A política reorganiza a distribuição dos recursos, mas não elimina as limitações de capacidade impostas pelo cenário.

Por fim, a avaliação com tráfego sintético de vídeo mostra que diferenças produzidas pelo escalonamento também aparecem em métricas temporais que não foram utilizadas no treinamento do IA-FMR. Essa observação complementa os resultados de vazão, justiça e atendimento, mas deve permanecer restrita ao cenário avaliado. Como não são modelados reprodução real, decodificação, *playout buffer* ou adaptação de vídeo, os resultados não constituem uma caracterização completa da QoE percebida pelo usuário.

7 Conclusões e Trabalhos Futuros

Este capítulo sintetiza as conclusões da tese, explicita suas principais contribuições, registra a experiência de internacionalização e a produção científica diretamente associada ao desenvolvimento da proposta, delimita as limitações do estudo e apresenta possibilidades de continuidade da pesquisa.

7.1 Considerações Iniciais

O escalonamento de recursos de rádio em redes 5G NR envolve objetivos que não evoluem necessariamente na mesma direção. A priorização de UEs com melhores condições de transmissão pode aumentar a vazão agregada, mas tende a concentrar os recursos. Uma distribuição mais uniforme amplia a justiça da alocação, embora possa atribuir parte da capacidade a usuários com menor eficiência espectral. RR, PF e MR representam critérios distintos para realizar essas decisões e fornecem referências importantes para caracterizar esse compromisso.

Esta tese investigou uma formulação baseada em aprendizado por reforço na qual o compromisso entre justiça e eficiência também integra a decisão produzida pela política. A proposta resultante, denominada IA-FMR, evolui o FMR ao incluir na ação, além dos logits utilizados para definir a distribuição alvo dos recursos, o coeficiente α_t . Durante o treinamento, esse coeficiente determina a ponderação entre os termos de justiça e eficiência da recompensa. O valor anterior, α_{t-1} , também integra o estado utilizado na decisão seguinte. Durante a inferência com a política congelada, a recompensa não é recalculada, mas α_t continua sendo produzido e passa a compor o estado subsequente.

O treinamento foi realizado offline em Python e as políticas resultantes foram integradas ao ns-3/5G-LENA por meio do ns3-ai. A avaliação sistêmica contemplou doze larguras de banda, conjuntos com 3, 5, 7 e 9 UEs e um cenário de alta carga com seis UEs e tráfego sintético de vídeo. Os experimentos sistêmicos utilizaram 100 sementes comuns às configurações comparáveis. Essas sementes caracterizam a variabilidade das realizações

de simulação e não a variabilidade do treinamento, que foi executado uma única vez para cada largura de banda.

As comparações apresentadas na tese são descritivas. As médias e os intervalos de confiança de 95% permitem caracterizar os comportamentos observados nas configurações avaliadas, mas não foram acompanhados por testes formais de hipótese. Por essa razão, as conclusões a seguir permanecem restritas às políticas treinadas, aos cenários e aos parâmetros utilizados.

7.2 Conclusões Gerais

O treinamento produziu políticas capazes de serem executadas posteriormente no simulador, mas seus resultados também evidenciaram um aspecto importante da formulação proposta. Na política de 40 MHz, utilizada para a análise detalhada de α_t , o coeficiente deslocou-se progressivamente para valores elevados durante o aprendizado e permaneceu concentrado nessa região. Na execução temporal da política congelada no ns-3/5G-LENA para a semente 47889, α_t apresentou média de 0,9982, com variações muito pequenas ao longo das fases de tráfego. Assim, a formulação permite que o coeficiente seja produzido a cada decisão, mas a política aprendida nesse caso não explorou amplamente diferentes pontos do domínio disponível. A capacidade introduzida pela formulação e o comportamento efetivamente aprendido devem, portanto, ser interpretados separadamente.

Na Avaliação por Largura de Banda, o MR apresentou as maiores médias de vazão agregada ao longo das configurações, acompanhado por maior concentração dos recursos e desempenho reduzido na região inferior da distribuição dos fluxos. O RR manteve alocação próxima da uniformidade. O PF atendeu subconjuntos menores de usuários por decisão e apresentou valores intermediários de justiça. O IA-FMR, por sua vez, manteve distribuição ampla dos recursos na maior parte das bandas e apresentou médias de vazão agregada acima das observadas para RR e PF e abaixo das observadas para MR.

A configuração de 40 MHz ilustra esse ponto de operação sem combinar resultados de diferentes larguras de banda em uma única média. Nessa condição, a vazão agregada média foi de 97,53 Mbps no IA-FMR, 50,93 Mbps no RR, 34,36 Mbps no PF e 120,40 Mbps no MR. O Índice de Jain sobre os RBGs foi de 0,9943 no IA-FMR e 0,9988 no RR, enquanto PF e MR apresentaram 0,5926 e 0,3711. O P5 da vazão por fluxo alcançou 1,225 Mbps no IA-FMR, 0,714 Mbps no RR e 0,337 Mbps no PF, permanecendo próximo de zero no MR. A partir de 15 MHz, o IA-FMR apresentou a maior média de P5 entre os quatro

mecanismos em todas as larguras de banda avaliadas, embora os intervalos de confiança individuais se sobreponham em alguns pontos.

A comparação entre a alocação alvo produzida pelo IA-FMR e a alocação efetivamente realizada pela camada MAC mostrou que a distribuição ampla dos recursos já estava presente na saída da política. Entre 15 e 30 MHz, alvo e aplicação coincidiram integralmente. Entre 40 e 100 MHz, o mesmo conjunto de UEs foi preservado em 100% das decisões e as diferenças observadas nas quantidades de RBGs foram residuais. A configuração de 10 MHz foi a principal exceção. Nessa banda, a política produziu alocação positiva para os nove UEs, com Índice de Jain médio de 0,9630, mas a aplicação efetiva resultou em 7,30 UEs atendidos em média e Jain de 0,7985. O resultado mostra que a redução de atendimento observada em 10 MHz não estava presente com a mesma intensidade na decisão produzida pela política.

A análise temporal de 40 MHz, realizada sobre uma única semente comum aos quatro mecanismos, complementou os resultados agregados sem substituí-los. Nessa realização, RR e IA-FMR mantiveram os nove UEs atendidos. O IA-FMR apresentou número efetivo médio de UEs igual a 8,96 e maior participação individual média de 12,26%, valores compatíveis com uma distribuição ampla, porém não rigidamente uniforme. O PF alterou os subconjuntos atendidos ao longo do tempo. O MR manteve maior concentração, com número efetivo médio de 4,44 e maior participação individual média de 25,46%. A evolução do *backlog* também mostrou que nenhuma ordenação isolada descreve todo o comportamento: o IA-FMR apresentou menor ocupação que o MR nas fases iniciais da realização, enquanto o MR passou a apresentar valores ligeiramente inferiores nas fases finais.

A Avaliação de Generalização Estrutural por Número de UEs mostrou que a mesma arquitetura, definida para uma dimensão máxima de nove usuários, continua produzindo decisões quando aplicada a conjuntos menores dentro dessa dimensão. O Índice de Jain permaneceu elevado, com médias de 0,9766, 0,9727, 0,9731 e 0,9943 para 3, 5, 7 e 9 UEs. O número médio de UEs servidos acompanhou de perto a cardinalidade de cada configuração e não foram observados registros com demanda positiva e alocação aplicada nula.

Essa generalização estrutural não implicou invariância de desempenho. A vazão agregada média aumentou de 101,72 Mbps com 3 UEs para 149,75 Mbps com 5 UEs e diminuiu para 127,39 Mbps com 7 UEs e 97,53 Mbps com 9 UEs. No mesmo intervalo entre 3 e 9 UEs, o P5 caiu de 8,477 para 1,225 Mbps, a perda média aumentou de 12,51% para

72,04% e o atraso médio dos pacotes recebidos passou de 782,98 para 4.321,06 ms. Os resultados distinguem, portanto, a capacidade da política de operar com cardinalidades menores da capacidade de preservar o mesmo nível de desempenho quando o número de usuários é alterado.

A avaliação com tráfego sintético de vídeo examinou uma dimensão temporal que não foi utilizada no treinamento do IA-FMR. Para o prazo principal de 60 ms, a TUP média foi de 66,88% no IA-FMR, 53,10% no RR, 47,12% no PF e 46,41% no MR. Na comparação entre IA-FMR e RR, a principal diferença ocorreu na parcela de unidades tardias, que foi de 20,00% no IA-FMR e 34,30% no RR, enquanto as proporções de unidades não recebidas permaneceram próximas.

A duração P95 média dos congelamentos inferidos para 60 ms foi de 8,17 s no IA-FMR, 12,02 s no RR, 15,32 s no PF e 17,48 s no MR, considerando as sementes válidas para essa métrica. A proporção média de UEs que atingiu o critério de $TUP \geq 99\%$ foi de 59,17% no IA-FMR, 40,00% no RR, 39,67% no PF e 41,17% no MR. A ampliação do prazo para 80 ms produziu alterações pequenas na TUP e na proporção de usuários temporalmente satisfeitos. A frequência de congelamentos, por outro lado, não preservou a mesma ordenação entre os mecanismos, reforçando que frequência e duração caracterizam aspectos diferentes da sequência temporal das unidades.

Esses resultados não caracterizam o IA-FMR como um escalonador orientado a QoS ou QoE. A política não observa prazos, unidades de vídeo, estado de reprodução ou eventos de congelamento, e o limiar de vazão utilizado durante o treinamento constitui apenas uma referência para penalização de déficit, não uma garantia de serviço. A avaliação de vídeo mostra somente que, no cenário sintético de alta carga considerado, diferenças produzidas pelo escalonamento também se manifestam em métricas objetivas de entrega temporal que não integraram a função de recompensa.

Em conjunto, os experimentos mostram que vazão agregada, uniformidade na distribuição dos RBGs, atendimento dos UEs e desempenho dos fluxos menos favorecidos não evoluem necessariamente na mesma direção. Nas configurações avaliadas, o IA-FMR produziu um ponto de operação caracterizado por elevada justiça e atendimento amplo, médias de vazão acima das observadas para RR e PF e desempenho de cauda distinto do MR. Ao mesmo tempo, os resultados de 10 MHz, da variação do número de UEs e do cenário de vídeo evidenciam que a política não elimina limitações de capacidade nem estabelece superioridade universal sobre os mecanismos de referência.

7.3 Principais Contribuições

As principais contribuições desta tese são:

1. A evolução do FMR para o IA-FMR por meio de uma formulação *in-action*, na qual a ação da política inclui tanto os logits associados à distribuição alvo dos recursos quanto o coeficiente α_t . Durante o treinamento, α_t participa da ponderação entre justiça e eficiência, e seu valor anterior integra o estado utilizado na decisão seguinte.
2. A formulação do problema de decisão com um estado que reúne informações de canal, histórico recente de alocação e vazão, Índice de Jain e o valor anterior de α . A função de recompensa combina justiça na distribuição dos RBGs, eficiência em relação a uma referência instantânea de vazão e uma penalização por déficit abaixo de $T_{\min}$. Esse limiar atua como referência de treinamento e não como garantia de QoS.
3. O treinamento offline de políticas PPO e sua integração ao ns-3/5G-LENA por meio do ns3-ai, permitindo utilizar modelos previamente treinados nas decisões de escalonamento da camada MAC. A integração inclui o intercâmbio do estado, a produção da alocação alvo e o retorno da decisão ao simulador.
4. A implementação de um mecanismo de dimensionalidade fixa para até nove UEs, com preenchimento das posições inativas e aplicação da transformação apenas aos usuários elegíveis. A generalização estrutural para cardinalidades inferiores foi avaliada explicitamente com 3, 5, 7 e 9 UEs, sem interpretar nove usuários como limite do ns-3/5G-LENA ou como evidência de escalabilidade para cardinalidades superiores.
5. A instrumentação necessária para distinguir a alocação alvo produzida pela política da alocação efetivamente aplicada pela camada MAC. Essa comparação permitiu mostrar que, entre 15 e 100 MHz, o conjunto de UEs contemplado pela política foi preservado e que a principal divergência ocorreu em 10 MHz.
6. Uma avaliação sistêmica com 100 sementes comuns que combina métricas de vazão, justiça, atendimento, P5 por fluxo, perdas, atraso e demanda nas filas, complementada por uma análise temporal de uma realização comum em 40 MHz. Essa combinação permitiu distinguir resultados agregados entre sementes de comportamentos internos observados ao longo de uma execução específica.

7. A avaliação da política em cardinalidades menores dentro da dimensão prevista, caracterizando separadamente generalização estrutural, justiça, atendimento e degradação de desempenho à medida que o conjunto de usuários é alterado.
8. A instrumentação de tráfego sintético de vídeo e a análise de métricas objetivas de entrega temporal não utilizadas durante o treinamento, incluindo TUP, composição entre unidades no prazo, tardias e não recebidas, frequência e duração de congelamentos inferidos e satisfação temporal por usuário.

7.4 Internacionalização e Produção Científica

Parte da pesquisa foi desenvolvida durante estágio de doutorado sanduíche no Laboratoire d'Informatique de Sorbonne Université (LIP6), em Paris, entre junho de 2025 e março de 2026, no âmbito do programa CAPES/COFECUB. As atividades desenvolvidas nesse período envolveram discussões sobre escalonamento de recursos, aprendizado por reforço e representação explícita do compromisso entre justiça e eficiência, contribuindo para o amadurecimento da formulação que resultou no IA-FMR.

Entre a produção científica desenvolvida durante o doutorado, dois trabalhos estão diretamente vinculados à evolução do FMR e do IA-FMR:

1. Diego Canizio Lopes, André Luiz Nasseralla Pires, Ian Vilar Bastos e Igor Monteiro Moraes. *Fair Max Rate: Um Escalonador de Recursos Baseado em Aprendizado por Reforço para Redes 5G*. Publicado nos anais do Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos, SBRC 2025.
2. Diego Canizio Lopes, André Luiz Nasseralla Pires, Ian Vilar Bastos, Igor Monteiro Moraes e Francesca Fossati. *In-action Fairness Control for 5G Resource Scheduling Using Reinforcement Learning*. Publicado nos anais do IEEE Conference on Network Softwarization, IEEE NetSoft 2026.

O primeiro trabalho registra uma etapa anterior da pesquisa, centrada na formulação do FMR e em sua avaliação no ambiente de treinamento em Python. O segundo introduz a formulação *in-action*, incorporando α_t ao espaço de ações. A tese amplia essas etapas ao consolidar a integração com o ns-3/5G-LENA, a avaliação com 100 sementes, a análise entre alocação alvo e aplicada, a generalização estrutural para diferentes cardinalidades e a avaliação temporal com tráfego sintético de vídeo.

7.5 Limitações

Os resultados desta tese devem ser interpretados considerando o escopo do modelo e da avaliação experimental.

A primeira limitação decorre do uso de simulação. O ns-3/5G-LENA permite controlar condições de canal, mobilidade, tráfego e configuração da camada de acesso, mas não reproduz todas as particularidades de uma rede implantada. As avaliações também se restringem a uma célula, tráfego em *downlink* e um conjunto específico de modelos de propagação, mobilidade, antenas e parâmetros de rádio. Os resultados não estabelecem comportamento equivalente em cenários multicélula, no enlace de subida ou durante procedimentos de *handover*.

As políticas foram treinadas separadamente para cada largura de banda e para uma dimensão máxima de nove UEs. O mecanismo de preenchimento permite utilizar a arquitetura com conjuntos menores, mas a avaliação realizada caracteriza apenas cardinalidades dentro dessa dimensão. Ela não demonstra generalização para quantidades superiores a nove UEs nem para qualquer disposição espacial, perfil de carga ou condição de canal.

O treinamento foi executado uma única vez para cada largura de banda, utilizando a mesma semente de treinamento. Dessa forma, as 100 sementes da avaliação sistêmica caracterizam a variabilidade das simulações, mas não permitem estimar a variabilidade entre treinamentos independentes. Também não foi conduzida uma busca sistemática de hiperparâmetros nem uma comparação experimental abrangente entre PPO e outros algoritmos de aprendizado por reforço. A escolha do PPO foi motivada pela ação contínua e pelo mecanismo de atualizações controladas, mas a tese não demonstra que ele seja a alternativa ótima para o problema.

A política permaneceu congelada durante as simulações. A adaptação observada decorre da produção de ações diferentes em resposta aos estados apresentados ao modelo e não de atualização online dos pesos. Consequentemente, a avaliação não caracteriza aprendizado contínuo diante de mudanças persistentes na distribuição dos usuários, nas condições de canal ou nos perfis de tráfego.

A formulação permite produzir α_t a cada decisão, mas a política analisada em 40 MHz permaneceu fortemente concentrada em valores próximos de 1 durante a inferência. Não foi realizada uma ablação que fixasse, removesse ou controlasse independentemente α_t e seu histórico. Como os logits de alocação e o coeficiente são produzidos conjuntamente e α_{t-1} integra o estado, não é possível isolar a contribuição causal do parâmetro sobre

as decisões subsequentes. A tese demonstra a formulação e caracteriza o comportamento aprendido, mas não demonstra exploração ampla de diferentes pontos de compromisso.

A função de recompensa emprega $T_{\min}$ como referência para penalizar déficit de vazão durante o treinamento. Esse valor não representa taxa reservada, garantia de QoS ou requisito da aplicação de vídeo. Da mesma forma, o termo de eficiência compara a vazão obtida a uma referência instantânea baseada na melhor eficiência espectral e não deve ser interpretado como solução ótima de um problema multiobjetivo. Essas distinções limitam a interpretação direta dos componentes da recompensa como garantias operacionais.

A alocação aplicada pela camada MAC pode diferir da distribuição solicitada pela política. Essa diferença foi expressiva em 10 MHz e residual nas demais bandas avaliadas. A instrumentação utilizada permite quantificar a divergência entre alvo e aplicação, mas os experimentos não foram desenhados para atribuir cada ajuste a uma etapa interna específica da realização da concessão.

A representação histórica utiliza posições associadas ao conjunto de UEs observado em cada decisão. O mapeamento da ação corrente por RNTI assegura sua aplicação aos usuários correspondentes, mas alterações no conjunto elegível podem produzir desalinhamento transitório de componentes históricos organizados por posição. Uma representação inteiramente indexada pela identidade persistente dos UEs evitaria essa dependência.

A análise temporal detalhada de 40 MHz foi realizada com uma única semente comum aos quatro escalonadores. Ela permite inspecionar a dinâmica interna da realização selecionada, mas não caracteriza a variabilidade temporal entre as 100 sementes. As conclusões estatísticas do experimento permanecem apoiadas nas métricas agregadas entre sementes.

A avaliação com tráfego sintético de vídeo não executa uma cadeia completa de produção e reprodução. Não são simulados captura, renderização, codificação real, decodificação, *playout buffer*, adaptação de taxa, dispositivo de apresentação ou interação do usuário. Os congelamentos são inferidos a partir da sequência temporal das unidades que satisfazem ou não o prazo. Por isso, as métricas apresentadas caracterizam entrega temporal no cenário sintético considerado e não constituem uma avaliação completa de QoE ou de uma aplicação XR operacional.

Por fim, o treinamento foi realizado offline e seu custo não faz parte do caminho de decisão durante a inferência. Entretanto, o custo computacional da inferência, a comunicação entre ns-3 e Python e o tempo total necessário para produzir e aplicar uma decisão não foram caracterizados como restrições de tempo real. A implementação experimen-

tal demonstra integração funcional, mas não permite concluir que a mesma arquitetura satisfaça os requisitos temporais ou computacionais de uma estação rádio base comercial.

7.6 Trabalhos Futuros

Uma primeira direção consiste em ampliar a avaliação sistêmica para cenários multicélula, incluindo interferência intercelular, mobilidade entre gNBs e procedimentos de *handover*. Também devem ser considerados o enlace de subida, fluxos bidirecionais e combinações de serviços com requisitos distintos, permitindo verificar em que medida os comportamentos observados no cenário atual permanecem sob condições mais heterogêneas.

A estrutura da política pode ser modificada para representar conjuntos variáveis de usuários sem depender de uma dimensão máxima fixa e de preenchimento das posições inativas. Arquiteturas baseadas em atenção, agregação sobre conjuntos ou redes neurais para grafos constituem possibilidades para representar cardinalidades variáveis. Em paralelo, os históricos podem ser mantidos por identidade persistente, como o RNTI, evitando dependência da posição ocupada pelo usuário no vetor de estado.

A robustez do processo de treinamento também deve ser investigada. Repetições com diferentes sementes de treinamento permitiriam separar a variabilidade do aprendizado da variabilidade do cenário sistêmico. Uma avaliação sistemática de hiperparâmetros e a comparação com outros algoritmos adequados a ações contínuas poderiam ainda esclarecer a sensibilidade da solução às escolhas realizadas durante o desenvolvimento. No nível da avaliação sistêmica, análises pareadas entre sementes comuns podem complementar os intervalos de confiança individuais quando a pergunta experimental exigir inferência formal sobre as diferenças entre mecanismos.

Outra possibilidade é investigar adaptação ou treinamento online. Essa extensão permitiria atualizar os pesos diante de mudanças persistentes no canal, na distribuição dos usuários ou no perfil de tráfego, mas exigiria mecanismos para limitar o efeito de exploração durante a operação e preservar estabilidade e segurança das decisões.

O papel de α_t constitui uma direção específica de investigação. Estudos futuros podem realizar ablações que fixem o coeficiente, removam seu histórico ou controlem separadamente sua produção, permitindo estimar sua contribuição para as alocações. Também podem ser avaliadas outras funções de recompensa, temperaturas de transformação, regularizações e cenários capazes de induzir pontos de operação distintos. O objetivo não é apenas ampliar a variação observada de α_t , mas compreender em quais condições sua

presença no espaço de ações altera efetivamente a política aprendida.

A função de recompensa pode ser estendida para incorporar informações diretamente relacionadas ao serviço, como ocupação das filas, idade dos pacotes, prazos, classes de tráfego ou requisitos específicos de fluxo. Essas extensões produziriam versões diferentes da proposta atual e poderiam ser orientadas explicitamente a QoS ou QoE. A definição de $T_{\min}$ também pode ser estudada como parâmetro de treinamento, avaliando sua sensibilidade sem reinterpretá-lo como garantia operacional.

Na avaliação de vídeo, trabalhos futuros podem utilizar traços de codificação real, dependências entre quadros e um reprodutor com *playout buffer*. A inclusão de métricas de qualidade visual, adaptação de taxa e avaliações perceptuais permitiria relacionar as decisões de escalonamento a uma representação mais ampla da experiência do usuário. Uma avaliação XR completa exigiria ainda componentes como *viewport*, rastreamento de pose, renderização, codificação, decodificação e decomposição do atraso entre as diferentes etapas da cadeia.

Por fim, a política pode ser incorporada a uma implementação em linguagem compilada ou a um ambiente de teste com equipamentos de rede. A medição do tempo de inferência, consumo de memória, sobrecarga de comunicação e tempo disponível por decisão é necessária para avaliar a viabilidade de execução fora da simulação. Essa etapa deve considerar separadamente o custo do treinamento offline e o custo do caminho de inferência utilizado durante a operação.

REFERÊNCIAS

3GPP. **3GPP TS 38.211: NR; Physical channels and modulation.** [*S. l.*], fev. 2026. Disponível em: https://www.etsi.org/deliver/etsi_ts/138200_138299/138211/18.08.00_60/ts_138211v180800p.pdf.

3GPP. **NR; Base Station (BS) radio transmission and reception.** [*S. l.*], fev. 2026. Disponível em: https://www.etsi.org/deliver/etsi_ts/138100_138199/138104/19.03.00_60/ts_138104v190300p.pdf.

3GPP. **NR; NR and NG-RAN Overall description; Stage-2.** [*S. l.: s. n.*], 2026. Technical Specification (TS). (3GPP TS 38.300). Disponível em: <https://www.3gpp.org/dynareport/38300.htm>.

3GPP. **NR; Physical layer procedures for data.** [*S. l.*]: 3GPP, 2026. 3GPP TS 38.214. (TS 38.214). Release 19. Disponível em: <https://www.3gpp.org/dynareport/38214.htm>.

3GPP. **QoE parameters and metrics relevant to the Virtual Reality (VR) user experience.** [*S. l.*], mai. 2024. (Release 18, 3GPP TR 26.929). Disponível em: <https://www.3gpp.org/dynareport/26929.htm>.

3GPP. **Study on XR (Extended Reality) evaluations for NR.** [*S. l.*], jan. 2022. (Release 17, 3GPP TR 38.838). Disponível em: <https://www.3gpp.org/dynareport/38838.htm>.

3GPP. **Traffic Models and Quality Evaluation Methods for Media and XR Services in 5G Systems.** [*S. l.*], mar. 2024. (Release 18, 3GPP TR 26.926 Version 18.2.0). Disponível em: <https://www.3gpp.org/dynareport/26926.htm>.

ABBAS, Ali Haider. A Reinforcement Learning–Driven Scheduler for Minimizing Uplink Delay in 5G Networks. **Journal of Al-Qadisiyah for Computer Science and Mathematics**, v. 17, n. 4, p. 274–284, 2025.

ANDREWS, Jeffrey G; BUZZI, Stefano; CHOI, Wan; HANLY, Stephen V; LOZANO, Angel; SOONG, Anthony CK; ZHANG, Jianzhong Charlie. What will 5G be? **IEEE Journal on Selected Areas in Communications**, IEEE, v. 32, n. 6, p. 1065–1082, 2014.

BAHETI, Prashant Kumar; KHUNTETA, Ajay. Fairness-Aware Radio Resource Scheduling for 5G NR mMTC Networks. **Physical Communication**, Elsevier, p. 103103, 2026.

BUDHIRAJA, Ishan; KUMAR, Neeraj; TYAGI, Sudhanshu. Deep-reinforcement-learning-based proportional fair scheduling control scheme for underlay D2D communication. **IEEE Internet of Things Journal**, IEEE, v. 8, n. 5, p. 3143–3156, 2020.

CUI, Zhiyi; QAMAR, Faizan; KAZMI, Syed Hussain Ali; ARIFFIN, Khairul Akram Zainol; SAFDAR, Ghazanfar Ali; UR REHMAN, Muhammad Habib. A review of multi-agent deep reinforcement learning for resource allocation in beyond 5G network slicing: solutions, challenges and future research directions. **PeerJ Computer Science**, PeerJ Inc., v. 12, e3728, 2026.

DAHLMAN, Erik; PARKVALL, Stefan; SKOLD, Johan. **5G/5G-advanced: the new generation wireless access technology**. [S. l.]: Elsevier, 2023.

FERREIRA, Vinicius C.; BALBI, Helga; SEIXAS, Flavio L.; ALBUQUERQUE, Celio; MUCHALUAT-SAADE, Debora C. Wireless Body Area Networks: An Overview. *In*: ANAIS do XXXV Simpósio Brasileiro de Telecomunicações e Processamento de Sinais (SBrT). São Pedro, SP, Brazil: [s. n.], 2017. p. 65–78.

FOSSATI, Francesca; MORETTI, Stefano; PERNY, Patrice; SECCI, Stefano. Multi-resource allocation for network slicing. **IEEE/ACM Transactions on Networking**, IEEE, v. 28, n. 3, p. 1311–1324, 2020.

GALDINO, Caio P.; COUTO, Rodrigo S.; MEDEIROS, Dianne S. V. de; MORAES, Igor M.; MATTOS, Diogo M. F. Reinforcement Learning for Radio Resource Management in O-RAN. *In*: 2024 IEEE 13th International Conference on Cloud Networking (CloudNet). Rio de Janeiro, Brazil: IEEE, 2024. p. 1–9. DOI: [10.1109/CLOUDNET62863.2024.10815702](https://doi.org/10.1109/CLOUDNET62863.2024.10815702).

GOBATTO, Leonardo; SAQUETTI, Mateus; DINIZ, Claudio; ZATT, Bruno; CORDEIRO, Weverton; AZAMBUJA, Jose R. Improving Video Streaming Quality in Congested Networks with In-Network Computing. *In*: 2024 IEEE 15th Latin America Symposium on Circuits and Systems (LASCAS). Punta del Este, Uruguay: IEEE, 2024. p. 1–5. DOI: [10.1109/LASCAS60203.2024.10506184](https://doi.org/10.1109/LASCAS60203.2024.10506184).

GU, Zhouyou; SHE, Changyang; HARDJAWANA, Wibowo; LUMB, Simon; MCKECHNIE, David; ESSERY, Todd; VUCETIC, Branka. Knowledge-assisted deep reinforcement learning in 5G scheduler design: From theoretical framework to implementation. **IEEE Journal on Selected Areas in Communications**, IEEE, v. 39, n. 7, p. 2014–2028, 2021.

GUO, Chongtao; SHENG, Min; WANG, Xijun; ZHANG, Yan. Throughput maximization with short-term and long-term Jain's index constraints in downlink OFDMA systems. **IEEE Transactions on Communications**, IEEE, v. 62, n. 5, p. 1503–1517, 2014.

HENDERSON, Thomas R.; LACAGE, Mathieu; RILEY, George F.; DOWELL, Craig; KOPENA, Joseph. Network simulations with the ns-3 simulator. *In: PROCEEDINGS of the 1st International Conference on Simulation Tools and Techniques for Communications, Networks and Systems (SIMUTools)*. Marseille, France: [s. n.], 2008. p. 1–7. DOI: [10.4108/ICST.SIMUTOLS2008.3031](https://doi.org/10.4108/ICST.SIMUTOLS2008.3031).

INTERNATIONAL TELECOMMUNICATION UNION. **P.10/G.100: Vocabulary for performance, quality of service and quality of experience**. [S. l.], set. 2017. Disponível em: <https://www.itu.int/rec/T-REC-P.10>.

JAIN, Rajendra K.; CHIU, Dah-Ming W.; HAWKES, William R. A Quantitative Measure Of Fairness And Discrimination For Resource Allocation In Shared Computer Systems. **CoRR**, cs.NI/9809099, jan. 1998.

KIM, Yohan; LIM, Hyuk. Multi-agent reinforcement learning-based resource management for end-to-end network slicing. **IEEE Access**, IEEE, v. 9, p. 56178–56190, 2021.

KINGMA, Diederik P.; BA, Jimmy. **Adam: A Method for Stochastic Optimization**. [S. l.: s. n.], 2017. arXiv: [1412.6980](https://arxiv.org/abs/1412.6980) [cs.LG]. Disponível em: <https://arxiv.org/abs/1412.6980>.

KOUTLIA, Katerina; BOJOVIC, Biljana; ALI, Zoraze; LAGÉN, Sandra. Calibration of the 5G-LENA system level simulator in 3GPP reference scenarios. **Simulation Modelling Practice and Theory**, Elsevier, v. 119, p. 102580, 2022.

LI, Qingyang. QoS Assurance Mechanism for 5G Network Slicing Based on the Deep Reinforcement Learning PPO Algorithm. **arXiv preprint arXiv:2605.03345**, 2026.

LOPES, Diego Canizio; NASSERALA, André; BASTOS, Ian Vilar; MORAES, Igor M. Fair Max Rate: Um Escalonador de Recursos Baseado em Aprendizado por Reforço para Redes 5G. *In: SBC. SIMPÓSIO Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC)*. [S. l.: s. n.], 2025. p. 434–447.

LOPES, Diego Canizio; PIRES, André L. Nasserála; BASTOS, Ian Vilar; MORAES, Igor M.; FOSSATI, Francesca. In-Action Fairness Control for 5G Resource Scheduling Using Reinforcement Learning. *In: 2026 IEEE 12th International Conference on Network Softwarization (NetSoft)*. [S. l.: s. n.], 2026. p. 198–206. DOI: [10.1109/NetSoft70012.2026.11603482](https://doi.org/10.1109/NetSoft70012.2026.11603482).

MAMANE, Asmae; FATTAH, Mohammed; EL GHAZI, Mohammed; EL BEKKALI, Moulhime; BALBOUL, Younes; MAZER, Said. Scheduling Algorithms for 5G Networks and Beyond: Classification and Survey. **IEEE Access**, IEEE, v. 10, p. 51643–51661, 2022.

PATIL, Suvarna; KRISHNAN, Sivaneasan Bala; SAINI, Dilip Kumar Jang Bahadur; CHAKRABARTI, Prasun. Systematic Analysis of Fairness-Aware Deep Reinforcement Learning for Equitable and Efficient Resource Allocation. *In: IEEE. 2026 International Conference on Emerging Smart Computing and Informatics (ESCI). [S. l.: s. n.], 2026. p. 1–7.*

PAZ-PÉREZ, Aurora; TATO, Anxo; ESCUDERO-GARZÁS, J Joaquín; GÓMEZ-CUBA, Felipe. Flexible reinforcement learning scheduler for 5G networks. *In: IEEE. 2024 IEEE International Conference on Machine Learning for Communication and Networking (ICMLCN). [S. l.: s. n.], 2024. p. 566–572.*

RAJA, Gunasekaran; SANJEEV, A; RAVISHANKAR, Krithika; ARUNACHALAM, Karthikeyan. TRIP-B5G: Traffic classification and Resource allocation using Intelligent PPO in B5G O-RAN. *In: IEEE. 2026 IEEE 23rd Consumer Communications & Networking Conference (CCNC). [S. l.: s. n.], 2026. p. 1–4.*

SARAIVA, Juno Vitorino; BRAGA JUNIOR, Iran Mesquita; MONTEIRO, Victor Farias; LIMA, Franciso Rafael Marques; MACIEL, Tarcisio Ferreira; FREITAS JUNIOR, Walter da Cruz; CAVALCANTI, Francisco Rodrigo Porto. Deep Reinforcement Learning for QoS-Constrained Resource Allocation in Multiservice Networks. **Journal of Communication and Information Systems**, v. 35, n. 1, p. 66–76, abr. 2020. DOI: [10.14209/jcis.2020.7](https://doi.org/10.14209/jcis.2020.7). Disponível em: <https://jcis.sbrt.org.br/jcis/article/view/711>.

SCHULMAN, John; MORITZ, Philipp; LEVINE, Sergey; JORDAN, Michael; ABBEEL, Pieter. High-Dimensional Continuous Control Using Generalized Advantage Estimation. **arXiv preprint**, jun. 2015. DOI: [10.48550/arXiv.1506.02438](https://doi.org/10.48550/arXiv.1506.02438).

SCHULMAN, John; WOLSKI, Filip; DHARIWAL, Prafulla; RADFORD, Alec; KLIMOV, Oleg. Proximal policy optimization algorithms. **arXiv preprint arXiv:1707.06347**, 2017.

SEDIQ, Akram Bin; GOHARY, Ramy H.; SCHOENEN, Rainer; YANIKOMEROGLU, Halim. Optimal Tradeoff Between Sum-Rate Efficiency and Jain's Fairness Index in Resource Allocation. **IEEE Transactions on Wireless Communications**, v. 12, n. 7, p. 3496–3509, 2013. DOI: [10.1109/TWC.2013.061413.121703](https://doi.org/10.1109/TWC.2013.061413.121703).

SEGUIN, Michael; OMER, Anjali; KOOSHA, Mohammad; MALANDRA, Filippo; MASTRONARDE, Nicholas. Deep Reinforcement Learning for Downlink Scheduling in 5G and Beyond Networks: A Review. *In: IEEE. 2023 IEEE 34th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC).* [S. l.: s. n.], 2023. p. 1–6.

SHEZI, Nokwanda; MERAYO, Noemí; NLEYA, Bakhe; AGUADO, Juan C. Adaptive Proximal Policy Optimization for Efficient and SLA-Compliant Dynamic O-RAN Slice Resource Allocation. *In: IEEE. 2025 International Conference on Computer, Information and Telecommunication Systems (CITS).* [S. l.: s. n.], 2025. p. 1–7.

SRAVAN, Padala; SHAIK, Mohammed Ali. DRL-based multi-objective task scheduling for edge-cloud computing: latency, energy, and SLA optimisation. **Scientific Reports**, Nature Publishing Group UK London, 2026.

SUTTON, Richard S; BARTO, Andrew G. **Reinforcement learning: An introduction**. Second. Cambridge, Massachusetts: MIT press Cambridge, 2018. p. 552. (Adaptive Computation and Machine Learning series). A Bradford Book. ISBN 978-0-262-03924-6. Disponível em: <http://incompleteideas.net/book/the-book-2nd.html>.

AL-TAM, Farooq; CORREIA, Noélia; RODRIGUEZ, Jonathan. Learn to Schedule: LEASCH: A Deep Reinforcement Learning Approach for Radio Resource Scheduling in the 5G MAC Layer. **IEEE Access**, IEEE, v. 8, p. 143063–143076, 2020.

VILLEGAS, Neco; HERRERA, Juan Luis; DIEZ, Luis; SCOTECE, Domenico; FOSCHINI, Luca; AGÜERO, Ramón. ISAC-Assisted DRL for Dynamic MAC Scheduler Reconfiguration in O-RAN. **IEEE Open Journal of the Communications Society**, IEEE, 2026.

VISHNOI, Vineet; BUDHIRAJA, Ishan; GUPTA, Suneet; KUMAR, Neeraj. A deep reinforcement learning scheme for sum rate and fairness maximization among D2D pairs underlying cellular network with NOMA. **IEEE Transactions on Vehicular Technology**, IEEE, v. 72, n. 10, p. 13506–13522, 2023.

YIN, Hao; LIU, Pengyu; LIU, Keshu; CAO, Liu; ZHANG, Lytianyang; GAO, Yayu; HEI, Xiaojun. Ns3-Ai: Fostering Artificial Intelligence Algorithms for Networking Research. *In: PROCEEDINGS of the 2020 Workshop on Ns-3*. Gaithersburg, MD, USA: Association for Computing Machinery, 2020. (WNS3 2020), p. 57–64. ISBN 9781450375375. DOI: [10.1145/3389400.3389404](https://doi.org/10.1145/3389400.3389404). Disponível em: <https://doi.org/10.1145/3389400.3389404>.

YOU, Xiaohu; ZHANG, Chuan; TAN, Xiaosi; JIN, Shi; WU, Hequan. AI for 5G: research directions and paradigms. **Science China Information Sciences**, Springer, v. 62, p. 1–13, 2019.

YU, Xiaomo; MI, Jie; TANG, Ling; LONG, Long; QIN, Xiao. Dynamic multi objective task scheduling in cloud computing using reinforcement learning for energy and cost optimization. **Scientific Reports**, Nature Publishing Group UK London, 2025.

ZHANG, Wenfan; OU, Haijiao. Reinforcement learning based multi objective task scheduling for energy efficient and cost effective cloud edge computing. **Scientific Reports**, Nature Publishing Group UK London, v. 15, n. 1, p. 41716, 2025.