Recursos na Borda: Otimização para o *Open* RAN usando Programação Inteira e Algoritmos Genéticos

Jonathas dos Santos, Fernanda de Castro Fernandes, Marcos Antônio de Sousa, Flávio Geraldo Coelho Rocha

***Resumo*— Este artigo investiga o desafio conjunto de admissão e associação de usuários, bem como o posicionamento otimizado de unidades funcionais na arquitetura *Open* RAN com suporte ao fatiamento de rede. Trata-se de um modelo de otimização formulado como um problema de Programação Linear Inteira, que incorpora restrições de Qualidade de Serviço e limitações de recursos computacionais. A escalabilidade do modelo é avaliada por meio de técnicas que combinam abordagens de solução exata e heurísticas aproximadas. Os resultados das simulações evidenciam não apenas a viabilidade e a eficiência computacional das soluções propostas, mas também sua escalabilidade, robustez e aplicabilidade prática em cenários de larga escala.**

***Palavras-Chave*— Open RAN, Alocação de recursos, Otimização, Simulação.**

***Abstract*— This paper addresses the joint challenge of user admission and association, as well as the optimized placement of functional units in Open RAN architectures with network slicing support. An optimization model is formulated as an Integer Linear Programming problem, which incorporates Quality of Service constraints and computational resource limitations. The model's scalability is assessed through techniques that combine exact solution methods with approximate heuristics. Simulation results demonstrate not only the feasibility and computational efficiency of the proposed solutions but also their scalability, robustness, and practical applicability in large-scale scenarios.**

***Keywords*— Open RAN, Resource allocation, Optimization, Simulation.**

I. Introdução

As redes móveis de quinta geração (5G) e suas evoluções (*Beyond* 5G – B5G) impulsionam a pesquisa e o desenvolvimento de arquiteturas desagregadas, virtualizadas e mais flexíveis [12]. Nesse contexto, a arquitetura Aberta de Rede de Acesso por Rádio (*Open Radio Access Network* – O-RAN), da O-RAN Alliance [10], propõe a separação da *Next Generation Node B* (gNB) tradicional em três componentes funcionais: Unidade Centralizada (*Open Centralized Unit* – O-CU), Unidade Distribuída (*Open Distributed Unit* - O-DU) e Unidade Rádio (*Open Radio Unit* – O-RU), os quais são implementados como funções virtualizadas sobre uma infraestrutura em nuvem (*O-Cloud*) e alocados de forma dinâmica para atender aos requisitos de Qualidade de Serviço (*Quality of Service* - QoS) dos usuários.

Segundo relatório técnico recente da O-RAN *Alliance* [10], a arquitetura *Open* RAN adota a divisão funcional 7.2x, no qual a pilha de protocolos de rádio é distribuída entre as três unidades funcionais, onde a O-RU é responsável pelo processamento da camada Física inferior (*low-PHY*); a O-DU trata as funções da camada Física superior (*high-PHY*), Controle de Acesso ao Meio (*Medium Access Control* – MAC) e Controle de Enlace de Rádio (*Radio Link Control* - RLC); e a O-CU fica encarregada das camadas superiores. Cada unidade pode ser instanciada virtualmente em ambientes de nuvem, viabilizando diferentes estratégias de implantação. A O-RAN *Alliance* define cenários distintos a depender do posicionamento da O-DU e O-CU, dentre os quais destaca-se: o cenário B (com O-CU e O-DU na borda) e o cenário C (com O-CU na nuvem regional e O-DU na borda) [12][10].

Com a introdução do Fatiamento de Rede (*Network Slicing* – NS), tornou-se possível dividir logicamente a infraestrutura física da rede em fatias virtuais (*slices*), cada uma configurada com recursos e políticas específicas para atender a diferentes tipos de serviços [9]. Nesse contexto, as fatias podem ser utilizadas para habilitar os casos de uso para as redes 5G e B5G, como *enhanced Mobile Broadband* (eMBB), voltadas para serviços com alta demanda de largura de banda, como *streaming* de vídeo; *massive Machine Type Communication* (mMTC), que abrange aplicações de Internet das Coisas (IoT), com tráfego esporádico de pequenos pacotes e grande densidade de dispositivos; e *ultra-Reliable Low-Latency Communication* (URLLC), destinado a casos de uso de comunicações de missão crítica com requisitos rigorosos de latência e confiabilidade.

Nesse contexto, um dos principais desafios técnicos está na resolução conjunta do problema de associação entre UEs e O-RUs, bem como no posicionamento eficiente das unidades funcionais O-CU e O-DU na infraestrutura em nuvem [6]. Por isso, diversos estudos têm abordado esses desafios [11],[7], [13]. No entanto, a maioria dessas abordagens desconsidera a heterogeneidade dos requisitos de QoS impostos pelo fatiamento de rede. A escassez de propostas que integrem, de forma conjunta, esses aspectos — especialmente em cenários de larga escala, alta densidade de usuários e diversidade de serviços — ainda representa uma lacuna significativa na literatura.

Neste artigo, objetiva-se maximizar a admissibilidade de UEs em uma rede *Open* RAN, assegurando o atendimento aos seus requisitos mínimos de QoS. Para isso, utiliza-se um modelo de otimização que considera duas decisões principais: o posicionamento das Unidades Funcionais (*Functional Unit* - FU) O-CU e O-DU na nuvem (seja de borda ou regional) e a alocação eficiente de Blocos de Recurso (*Resource Blocks* – RB) por fatia. Inicialmente, o problema é formulado como um modelo de Programação Linear Inteira (*Integer Linear Programming* – ILP) para capturar a complexidade das decisões interdependentes. Subsequentemente, são apresentadas estratégias de solução, tanto exatas quanto heurísticas. O

Jonathas dos Santos, Fernanda de Castro Fernandes, Marcos Antônio de Sousa, Flávio Geraldo Coelho Rocha, EMC, UFG, Goiânia-GO, e-mail: jonathas_santos@discente.ufg.br, castro.fernanda@discente.ufg.br, marcosantonio@ufg.br, flaviogcr@ufg.br.

objetivo é analisar o equilíbrio entre qualidade da solução, escalabilidade e tempo de esforço computacional.

O restante do artigo está organizado da seguinte forma: a Seção II apresenta a modelagem ILP para o problema conjunto de associação e posicionamento; a Seção III descreve as abordagens de solução para o modelo ILP; a Seção IV detalha a avaliação de desempenho conduzida e discute os resultados obtidos; por fim, são apresentadas as conclusões e possíveis direções para trabalhos futuros (Seção V).

II. Modelagem do Sistema

O sistema consiste em um conjunto R de O-RUs localizadas em uma área quadrada fixa de lado L, de modo que cada O-RU $r \in R$ tem uma posição determinada pelas coordenadas $P_r = (X_r, Y_r) \in [0, L]^2$. Um conjunto U de UEs encontra-se distribuído arbitrariamente pelas células (O-RUs), de modo que a posição de cada $u \in U$ é dada pelas coordenadas $P_u = (X_u, Y_u) \in [0, L]^2$.

O sistema de rede *O-Cloud* é modelado como um grafo $G = (H, E)$, onde H é o conjunto de vértices, representando os *hosts* na nuvem, e E é o conjunto de arestas, representando os enlaces físicos que conectam dois *hosts* vizinhos. H é ainda particionado com base na proximidade dos *hosts* ao local da célula em dois domínios: o conjunto de *hosts* da nuvem de borda H_E e o conjunto de *hosts* da nuvem regional H_R, tal que $H = (H_E \cup H_R)$. Para cada *host* h, se $h \in H_E$, então está localizado em $P_h \in [L, L']^2$, caso contrário, se $h \in H_R$, sua localização é $P_h \in [L'', L''']^2$. Cada O-RU pode ser conectada a quaisquer *hosts* da nuvem de borda.

Cada UE solicita o provisionamento de um serviço de comunicação do conjunto de fatias S. Cada fatia tem requisitos de QoS em termos de (i) taxa de dados alcançada (*throughput*) e (ii) atraso de ponta a ponta (*End-to-End* – E2E). Se o sistema é capaz de atender a todos os requisitos de QoS de uma fatia, de um UE específico, então ele admite o UE e fornece o serviço de comunicação solicitado. Essas premissas também são assumidas em [1] e [3]. As características detalhadas do sistema que impactam o provisionamento dos requisitos de QoS dos UEs e, consequentemente, determinam sua admissão, são descritas a seguir.

A. Modelo de Associação dos UEs às O-RUs

Cada UE pode estar simultaneamente ao alcance de múltiplas O-RUs. Caso seja admitido no sistema, ele deverá ser associado a uma dessas O-RUs dentro de sua área de cobertura. A decisão de associação de um UE a uma O-RU é capturada pelas variáveis $x^{RU}_{u,r} \in \{0,1\}, \forall u \in U, \forall r \in R$, indicando se o UE u está associado à O-RU r ($x^{RU}_{u,r} = 1$), ou não, ($x^{RU}_{u,r} = 0$). O vetor de variáveis de associação é definido como $\mathbf{x}^{RU} = [\ x^{RU}_{u,r} : \forall u \in U, \forall r \in R]$.

Utilizando Acesso Múltiplo por Divisão Ortogonal de Frequência (*Orthogonal Frequency Division Multiple Access* – OFDMA), o tempo-largura de banda das O-RUs é dividido em RBs que podem ser atribuídos aos UEs associados. Cada O-RU $r \in R$ tem um total de $M_r \in \mathbb{Z}^+$ RBs que são distribuídos entre todas as fatias em S. Variáveis $\rho_{r,s} \in \mathbb{Z}^+, \forall r \in R, \forall s \in S$ são utilizadas para capturar o número de RBs dedicados à fatia s na O-RU r. A fim de simplificar a notação, utilizam-se RU, DU e CU como abreviações de O-RU, O-DU e O-CU na formalização a seguir.

O número de RBs $RB_{u,r}$ requerido pelo usuário u, se associado à O-RU r, é calculado da seguinte forma:

$$RB_{u,r} \equiv \frac{\lambda_{s(u)}}{\eta_{u,r}}, \forall u \in U, \forall r \in R, \tag{1}$$

onde $s(u) \in S$ é a fatia solicitada pelo UE u, $\lambda_s \in \mathbb{R}^+$ é a taxa de dados requerida pela fatia $s \in S$ e $\eta_{u,r} \in \mathbb{R}^+$ é a capacidade do enlace sem fio, por RB, medida usando o teorema de Shannon para a capacidade do canal. Considera-se um padrão de reuso de frequência, de maneira que as O-RUs vizinhas nao interferem umas com as outras. Nota-se que um usuário u atribuído a uma O-RU r deve obter a quantidade solicitada de RBs para transmitir na taxa de dados requerida. O número de RBs atribuídos a um UE u é determinado por:

$$RB_u(\mathbf{x}^{RU}) \equiv \sum_{r \in R} RB_{u,r} \cdot x^{RU}_{u,r}, \forall u \in U. \tag{2}$$

B. Modelo de Posicionamento das O-DUs e O-CUs

Foi adotado um cenário de implantação híbrido entre os cenários B e C definidos em [8]. As unidades funcionais O-DUs são sempre posicionadas na nuvem de borda, enquanto as O-CUs podem estar na nuvem de borda ou na nuvem regional. Variáveis $x^{DU}_{u,h} \in \{0,1\}, \forall u \in U, \forall h \in H$ indicam se a O-DU da UE u está posicionada no *host* da nuvem h ($x^{DU}_{u,h} = 1$) ou não ($x^{DU}_{u,h} = 0$). O vetor de variáveis de posicionamento das O-DUs é definido como $\mathbf{x}^{DU} = [x^{DU}_{u,h} : \forall u \in U, \forall h \in H]$.

Analogamente, variáveis $x^{CU}_{u,h} \in \{0,1\}, \forall r \in R, \forall h \in H$ indicam se a O-CU do UE u, especificamente seu componente de Plano de Usuário (*User Plan* – UP), está posicionada no *host* da nuvem h ($x^{CU}_{u,h} = 1$) ou não ($x^{CU}_{u,h} = 0$). O vetor de variáveis de posicionamento da O-CU fica $\mathbf{x}^{CU} = [x^{CU}_{u,h} : \forall u \in U, \forall h \in H]$. Portanto, o vetor de variáveis de associação-posicionamento fica definido como: $\mathbf{x} = [\mathbf{x}^{RU}, \mathbf{x}^{DU}, \mathbf{x}^{CU}]$.

C. Modelo de Computação O-Cloud

Cada *host* na nuvem dispõe de capacidade computacional finita, em termos de memória RAM e processamento (CPU), sendo, portanto, capaz de executar apenas um número limitado de instâncias de unidades funcionais. Cada instância de unidade funcional (O-DU e O-CU) tem um custo computacional associado [14], dado em Giga Operações Por Segundo ((*Giga Operations Per Second* – GOPS), que é definido como segue:

$$g^{FU}_u(\mathbf{x}^{RU}) \equiv \alpha^{FU} \cdot \left(3A + A^2 + \frac{MCL}{3}\right) \cdot \frac{RB_u(\mathbf{x}^{RU})}{10}, \tag{3}$$

onde FU é substituída por O-CU ou O-DU, *M* representa os bits de modulação (número de bits por símbolo), *C* denota a taxa de codificação, *L* é o número de camadas MIMO, *A* corresponde ao número de antenas, e $RB_u(\mathbf{x})$ é o número de blocos de recursos atribuídos ao usuário u, conforme definido em (2). As constantes α^{DU} e α^{CU}, definidas para cada FU, servem como um fator de escala representando a carga computacional média das O-DUs e O-CUs, respectivamente, em relação aos seus requisitos computacionais totais.

Especificamente, é adotado o *split* funcional 7.2x entre O-RU e O-DU e o *split* 2 entre O-DU e O-CU. Com base na distribuição da carga computacional descrita em [5], atribui-se $\alpha^{DU} = 50\%$ e $\alpha^{CU} = 10\%$ da carga computacional para a O-DU e a O-CU, respectivamente (a O-RU é responsável pelos 40% restantes). A utilização computacional total no nó fica:

$$g_h(\mathbf{x}^{RU}) \equiv \sum_{u \in U} g_u^{CU}(\mathbf{x}^{RU}) \cdot x_{u,h}^{CU} + g_u^{DU}(\mathbf{x}^{RU}) \cdot x_{u,h}^{DU}, \quad (4)$$

onde as funções $g_u^{CU}(\cdot)$ e $g_u^{DU}(\cdot)$ são definidas em (3).

D. Modelo da Latência Ponta a Ponta

O modelo de latência considerado é o de atraso de propagação entre as unidades funcionais envolvidas na comunicação nas interfaces de *Midhaul* (MH) e *Fronthaul* (FH). Esse modelo de latência foi escolhido para ressaltar o impacto do posicionamento das unidades funcionais, seja próximo ao UE na nuvem de borda, seja distante do UE, na nuvem regional. Portanto, para cada UE u, o atraso MH é medido entre a O-CU e a O-DU e é dado por:

$$d_u^{MH}(\mathbf{x}) \equiv \sum_{h,h' \in H} \frac{||P_h - P_{h'}||}{v_{\text{Fiber}}} \cdot x_{u,h}^{CU} \cdot x_{u,h'}^{DU}, \quad (5)$$

onde $v_{\text{Fiber}} \in \mathbb{R}^+$ é a velocidade de propagação na fibra, e $||\cdot||$ representa a distância euclidiana entre dois *hosts*. Similarmente, para cada UE u, o atraso FH (da O-DU para a O-RU) é:

$$d_u^{FH}(\mathbf{x}) \equiv \sum_{r \in R, h \in H} \frac{||P_r - P_h||}{v_{\text{Fiber}}} \cdot x_{u,h}^{DU} \cdot x_{u,r}^{RU}. \quad (6)$$

E. Formulação do Modelo ILP de Otimização Conjunta

O objetivo é otimizar o desempenho do sistema por meio da maximização do número de UEs admitidos, condicionada à capacidade da infraestrutura, aos requisitos das interfaces do *Open* RAN, e aos requisitos dos serviços demandados pelos UEs em diferentes fatias. A formulação do problema de otimização conjunta de associação e posicionamento é apresentada a seguir. As notações utilizadas estão na Tabela I.

TABELA I
RESUMO DA NOTAÇÃO.

Var.	Definição
R, U	Conjuntos de O-RUs e UEs
H, S	Conjuntos de *Hosts* e Fatias (*Slices*)
M_r	Parâmetro para o número de RBs disponíveis na O-RU r
$x_{u,r}^{RU}$	Variável binária indicando a associação do UE u à O-RU r
$x_{u,h}^{CU}$	Variável binária indicando se o nó h hospeda a O-CU do UE u
$x_{u,h}^{DU}$	Variável binária indicando se o nó h hospeda a O-DU do UE u
$z_{uhh'}$	Variável binária indicando o posicionamento conjunto das unidades O-CU e O-DU para cada UE u
$z_{urhh'}$	Variável binária indicando a associação UE e O-RU conjuntamente com o posicionamento das unidades O-CU e O-DU para cada UE u
$\rho_{r,s}$	Variável inteira para os RBs na O-RU r alocados à fatia s
$\eta_{u,r}$	Taxa de transmissão por RB definida pelo teorema de Shannon
λ_s	Taxa de dados requerida pela fatia s
d_u^{FH}	Atraso máximo permitido no $fronthaul$ para o usuário u
d_u^{MH}	Atraso máximo permitido no $midhaul$ para o usuário u
ϵ_s	Valor de prioridade para a fatia s

$$\max_{\mathbf{x},\rho} \quad \sum_{u \in U} \sum_{r \in R} \sum_{h,h' \in H} \epsilon_{s(u)} \cdot z_{urhh'} \quad (7)$$

sujeito a:

$$z_{urhh'} \geq x_{u,h}^{DU} + x_{u,h'}^{CU} + x_{u,r}^{RU} - 2, \quad \forall h, h' \in H, \forall r \in R, \forall u \in U, \quad (8)$$

$$z_{urhh'} \leq (x_{u,h}^{DU} + x_{u,h'}^{CU} + x_{u,r}^{RU})/3, \quad \forall h, h' \in H, \forall r \in R, \forall u \in U, \quad (9)$$

$$\sum_{h,h' \in H} z_{uhh'} = \sum_{r \in R} x_{u,r}^{RU}, \quad \forall u \in U \quad (10)$$

$$z_{uhh'} \geq x_{u,h}^{DU} + x_{u,h'}^{CU} - 1, \quad \forall h, h' \in H, \forall u \in U, \quad (11)$$

$$z_{uhh'} \leq (x_{u,h}^{DU} + x_{u,h'}^{CU})/2, \quad \forall h, h' \in H, \forall u \in U, \quad (12)$$

$$\sum_{h \in H_R} x_{u,h}^{DU} = 0, \quad \forall u \in U \quad (13)$$

$$\sum_{s \in S} \rho_{r,s} \leq M_r, \quad \forall r \in R \quad (14)$$

$$RB_u(\mathbf{x}^{RU}) \leq \rho_{r,s}, \quad \forall r \in R, \forall s \in S \quad (15)$$

$$g_h(\mathbf{x}^{RU}) \leq G_h, \quad \forall h \in H \quad (16)$$

$$d_u^{MH}(\mathbf{x}) \leq D_u^{MH}, \quad \forall u \in U \quad (17)$$

$$d_u^{FH}(\mathbf{x}) \leq D_u^{FH}, \quad \forall u \in U \quad (18)$$

A função objetivo (7) visa maximizar o número de UEs admitidos ponderados por uma prioridade $\epsilon_{s(u)}$ definida para cada fatia solicitada pelo UE u, $s(u)$. As restrições (10) asseguram que um UE u admitido tenha exatamente uma unidade funcional associada a ele. As variáveis $z_{urhh'}$ e $z_{uhh'}$, virtualmente definidas como $z_{urhh'} = (x_{u,r}^{RU}.x_{u,h}^{DU}.x_{u,h}^{CU})$ e $z_{uhh'} = (x_{u,h}^{DU}.x_{u,h}^{CU})$, em conjunto com as restrições (8), (9), (11) e (12), modelam o fenômeno da associação UE-RU conjuntamente com o posicionamento das unidades funcionais, ao mesmo tempo em que preservam a linearidade do ILP. Devido aos limites de atraso [8], admite-se que as O-DUs devem ser implantadas nas proximidades dos locais das células, no domínio da nuvem de borda. Por outro lado, as O-CUs podem ser implantadas tanto na nuvem de borda quanto na nuvem regional. Esta limitação é capturada por (13). Com (14), assegura-se que a quantidade total de recursos da O-RU r atribuída a cada fatia não exceda seu número total de blocos de recursos M_r. Uma vez que cada fatia tem diferentes requisitos de RB, o número de UEs que a O-RU r pode acomodar é limitado à sua quantidade máxima de RBs $\rho_{r,s}$ dedicados a essa fatia (restrições (15)). Em (16), garante-se que a utilização computacional total em cada nó h não exceda sua capacidade computacional disponível G_h. Finalmente, as restrições (17) e (18) garantem que tanto os atrasos de MH quanto de FH satisfaçam seus valores de tolerância D_u^{MH} e D_u^{FH}, respectivamente. A solução ótima do sistema é representada por $\mathbf{x}^*$ (variáveis de associação e posicionamento) e ρ^* (quantidade de RBs).

III. Técnicas de Solução para o ILP

Para resolver o modelo ILP da Seção II, que otimiza a admissão de usuários e o posicionamento conjunto das unidades funcionais O-DU e O-CU, são avaliadas duas abordagens.

A. Solução Exata via CPLEX

As simulações foram conduzidas utilizando Python e o *solver* IBM CPLEX [2], que implementa o algoritmo *branch-and-bound*. A árvore de decisão é construída a partir das variáveis binárias $\mathbf{x}^{RU}$, $\mathbf{x}^{DU}$e $\mathbf{x}^{CU}$. Parâmetros típicos incluem limites de tempo, tolerância de otimalidade e estratégias de ramificação, que podem ser ajustadas para acelerar a convergência. Essa abordagem tem a vantagem de garantir a solução ótima da função objetivo (7), respeitando as restrições de capacidade (14)–(16) e de latência (17)–(18). Entretanto, ela possui a desvantagem da baixa escalabilidade, tornando-a inadequada para solução do problema NP-difícil em análise neste artigo.

B. Solução via Metaheurísitica GA

Com a metaheurística Algoritmo Genético (*Genetic Algorithm* – GA), cada cromossomo representa uma solução factível, codificando os vetores de decisão $\mathbf{x}^{RU}$, $\mathbf{x}^{DU}$ e $\mathbf{x}^{CU}$. Parâmetros típicos incluem tamanho da população, taxa de cruzamento e de mutação, além do critério de parada baseado no número máximo de gerações ou convergência do *fitness*. O operador de avaliação calcula a função objetivo (7), penalizando soluções que violam restrições, como as de capacidade e latência. Essa abordagem oferece soluções próximas da ótima com custo computacional reduzido. As simulações foram conduzidas utilizando o *software* Matlab com a biblioteca *GA*[4].

IV. Avaliação de Desempenho

A. Estrutura e Cenário de Simulação

O ambiente de simulação contempla uma rede *Open* RAN com ($|R|$ = 4) O-RUs, distribuídas em uma área industrial quadrada de lado $L = 1$ km. Os UEs estão distribuídos aleatoriamente dentro da área definida e pertencem a diferentes tipos de aplicação: eMBB (25%), URLLC (25%) e mMTC (50%). Um exemplo de configuração com 20 UEs é apresentado na Figura 1. Os UEs eMBB e URLLC recebem maior prioridade sobre os UEs mMTC, garantindo uma admissão equilibrada entre os diferentes tipos de fatias. Os parâmetros utilizados nas simulações encontram-se resumidos na Tabela II.

TABELA II
Configuração de Simulação

Configuração Sistema	
Parâmetro	**Valor**
Número de O-RUs	4
Número de UEs	[10, 150]
Número de nós *O-cloud Regional* e *Borda*	1, 3
Localização O-RUs - [x,y]m	[250,250], [250,750], [750,250] e [750,750]
Localização UEs - [x,y]m	*uniform*([0,1000], [0,1000])
Distância *host Borda* e *Regional* - km	*uniform*(5,10), *uniform*(40,80)
Capacidade *host Borda* e *Regional* - GOPS	100, 1000
Largura de banda - MHz	20
RBs disponíveis por TTI em cada O-RU	100
Potência transmissão O-RU - dBm	30
Prioridade (ϵ_s) eMBB, URLLC e mMTC	2, 2, 1
Taxa de dados (λ_s) eMBB, URLLC, mMTC - Mbps	20, 5, 1
Lim. de atraso (ms) MH eMBB, URLLC e mMTC	0,5, *uniform*(0,1, 0,3), 1,0
Lim. de atraso (ms) FH eMBB, URLLC, mMTC	1,0, 1,0, 1,0
Configuração GA	
Tamanho da população	350
Número de gerações	600
Usa processamento paralelo	Não

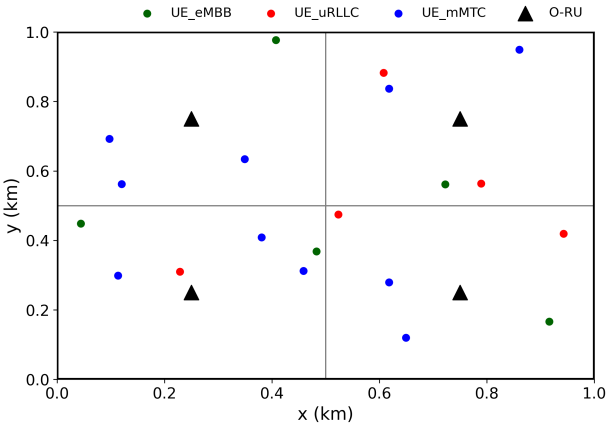

Fig. 1. Topologia de uma rede *Open* RAN com 4 O-RUs e 20 UEs.

B. Resultados e Discussões

Nesta seção, investiga-se o desempenho dos modelos para diferentes densidades de usuários, respeitando a estrutura de simulação descrita na Seção IV.A. Os UEs são dinamicamente associados às O-RUs conforme descrito na Seção II-A. O problema de posicionamento contempla um cenário flexível em que O-CUs e O-DUs podem ser dinamicamente implantados em servidores tanto regionais quanto de borda. Nas simulações, são empregadas as seguintes métricas de desempenho: a taxa ponderada de admissão de UEs, correspondente ao valor da função objetivo, na equação (7), e o *tradeoff* entre qualidade da solução, escalabilidade e tempo computacional.

A Figura 2 apresenta a comparação entre as soluções obtidas pelo modelo ILP e a abordagem GA, considerando como métrica o valor da função objetivo (7) (taxa ponderada de admissão) em função do número de UEs. Observa-se que a solução ótima foi computada apenas para instâncias com um número reduzido de UEs, limitando-se a aproximadamente 20 usuários. Neste intervalo, verifica-se que o valor da função objetivo cresce de forma praticamente linear com o aumento do número de UEs, comportamento esperado dado o acréscimo proporcional de variáveis e restrições no modelo. Por outro lado, a solução obtida com AG acompanha de maneira bastante próxima os valores da solução ótima nas instâncias em que ambas foram computadas, evidenciando a

qualidade das soluções heurísticas produzidas. Além disso, o AG demonstra elevada escalabilidade, mantendo a capacidade de fornecer soluções para cenários significativamente maiores, ultrapassando a marca de 100 UEs.

A Figura 3 descreve o tempo computacional em relação ao número de UEs. Os resultados indicam uma redução expressiva no tempo de execução ao empregar a abordagem baseada em Algoritmo Genético (AG), em comparação à solução ótima, especialmente com o aumento do número de UEs. Observa-se que o tempo de execução da solução ótima cresce rapidamente, ultrapassando 150 segundos com apenas 15 UEs, o que inviabiliza sua aplicação em cenários de maior escala. Por outro lado, embora a solução com AG também apresente aumento no tempo de execução conforme o número de UEs cresce, esse incremento ocorre de forma mais gradual e controlada, mantendo-se significativamente inferior ao da solução ótima. Mesmo com 110 UEs, o tempo de execução da abordagem com AG permanece relativamente reduzido, demonstrando sua maior eficiência e escalabilidade.

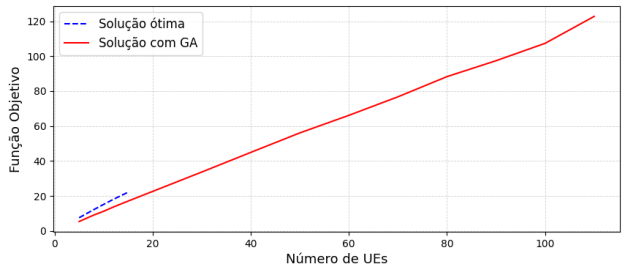

Fig. 2. Taxa de Admissão *versus* Quantidade de UEs.

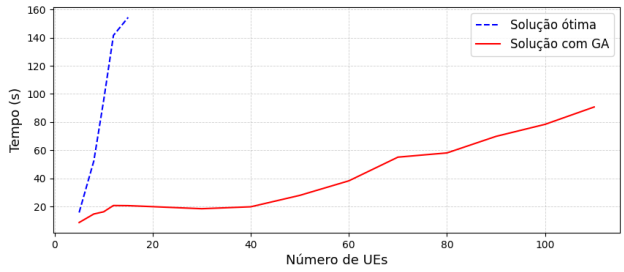

Fig. 3. Tempo computacional *versus* Quantidade de UEs.

V. Conclusões

Este trabalho abordou o problema conjunto de associação de usuários e posicionamento das unidades funcionais O-CU e O-DU em um ambiente de rede compatível com a arquitetura *Open* RAN, considerando a implementação de fatiamento de rede. O objetivo foi alcançar a maior quantidade de associações RU-UEs, conjuntamente com a alocação eficiente das unidades funcionais virtualizadas (O-CUs e O-DUs) nos nós *O-Cloud*, respeitando restrições de recursos da O-RAN, capacidade computacional dos servidores e requisitos de latência ponta a ponta.

O problema foi modelado e resolvido por meio de duas abordagens: Programação Linear Inteira, para obtenção de soluções ótimas, e metaheurística (Algoritmo Genético), com o objetivo de garantir escalabilidade e eficiência computacional. Avaliou-se o impacto da densificação de UEs na rede sobre o processo de alocação de recursos da rede *Open* RAN e sobre a carga computacional nas O-DUs e O-CUs virtualizadas em *hosts O-Cloud*.

Os resultados obtidos demonstram que a abordagem heurística é capaz de aproximar o desempenho da solução ótima com significativa redução no tempo de processamento, evidenciando sua aplicabilidade em cenários de grande escala. Como trabalhos futuros, pretende-se estender a modelagem proposta para cenários mais realistas e dinâmicos, incorporando aspectos da mobilidade dos usuários e da interferência entre fatias utilizando numerologia mista, com o objetivo de aprimorar a robustez e a adaptabilidade do sistema.

Agradecimentos

Os autores agradecem ao Centro de Excelência em Redes Inteligentes Sem Fio e Serviços Avançados (CERISE), à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e à Fundação de Amparo à Pesquisa do Estado de Goiás (FAPEG) pelo apoio e financiamento à pesquisa.

Referências

[1] H. Hojeij, G. I. Ricardo, M. Sharara, S. Hoteit, V. Vèque, and S. Secci, "Flexible association and placement for open-ran," in *IEEE INFOCOM 2024 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, Vancouver, BC, Canada, 2024, pp. 1–6.

[2] IBM Corporation, *User's Manual for CPLEX*, 2022, cplex-22.1.2.0.

[3] R. Joda, T. Pamuklu, P. E. Iturria-Rivera, and M. Erol-Kantarci, "Deep reinforcement learning-based joint user association and cu–du placement in o-ran," *IEEE Transactions on Network and Service Management*, 2022.

[4] MathWorks, "Genetic algorithm (ga)," https://la.mathworks.com/help/gads/ga.html, 2025, accessed: May 20, 2025.

[5] S. Mondal and M. Ruffini, "Optical front/mid-haul with open access edge server deployment framework for sliced o-ran," *IEEE Transactions on Network and Service Management*, vol. 19, no. 3, 2022.

[6] F. Z. Morais, G. M. F. de Almeida, L. Pinto, K. V. Cardoso, L. M. Contreras, R. da Rosa Righi, and C. B. Both, "Placeran: Optimal placement of virtualized network functions in beyond 5g radio access networks," *IEEE Transactions on Mobile Computing*, vol. 22, no. 9, pp. 5434–5448, 2022.

[7] A. Ndao, X. Lagrange, N. Huin, G. Texier, and L. Nuaymi, "Optimal placement of virtualized dus in o-ran architecture," in *2023 IEEE 97th Vehicular Technology Conference (VTC2023-Spring)*, 2023, pp. 1–6.

[8] O-RAN Alliance, "O-ran cloud architecture and deployment scenarios for o-ran virtualized ran 2.02," Tech. Rep., Feb. 2021, [Online]. Available: https://orandownloadsweb.azurewebsites.net/specifications.

[9] O-RAN WG1, "O-ran slicing architecture 13.0," O-RAN Alliance, Tech. Rep. O-RAN.WG1-Slicing Architecture-R003-v13.00, 2024.

[10] O-RAN Working Group 6, "O-ran cloud architecture and deployment scenarios for o-ran virtualized ran 4.0," O-RAN Alliance, Tech. Rep. O-RAN.WG6.CADS-v04.00, Oct. 2022, [Online]. Available: https://orandownloadsweb.azurewebsites.net/specifications.

[11] T. Pamuklu, S. Mollahasani, and M. Erol-Kantarci, "Energy-efficient and delay-guaranteed joint resource allocation and du selection in o-ran," in *2021 IEEE 4th 5G World Forum (5GWF)*. IEEE, Oct. 2021.

[12] M. Polese, L. Bonati, S. D'Oro, S. Basagni, and T. Melodia, "Understanding o-ran: Architecture, interfaces, algorithms, security, and research challenges," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 3, pp. 2480–2524, 2023.

[13] G. I. Ricardo, A. Benhamiche, N. Perrot, and Y. Carlinet, "Latency constrained task distribution in multi-access edge computing systems," in *2022 IEEE 11th International Conference on Cloud Networking (CloudNet)*, 2022.

[14] E. Sarikaya and E. Onur, "Placement of 5g ran slices in multi-tier o-ran 5g networks with flexible functional splits," in *2021 17th International Conference on Network and Service Management (CNSM)*, 2021.