Proposta de um Modelo ILP para Alocação de Funções na Arquitetura Open-RAN

K. D. R. Assis, A. F. dos Santos e R. C. Almeida Jr

***Resumo*— A 5ª geração de redes móveis (5G), baseadas em redes de acesso de rádio virtualizadas, exigirá soluções econômicas e flexíveis para satisfazer os requisitos de alta taxa de transferência e limites de latência. Uma das principais candidatas para alcançar estes objetivos é a arquitetura de interface "fronthaul". Nessa arquitetura, conhecida Open RAN, que significa "Rede de Acesso de Rádio Aberta" (*Open Radio Access Networks*) ou (O-RAN), há uma desagregação dos componentes tradicionalmente integrados em uma estação base de celular; as funcionalidades da estação são virtualizadas como funções de rede e divididas ao longo dos nós de rede. Neste artigo, nós abordamos o problema de posicionamento de funções sujeitas à restrições de capacidade e latência em uma arquitetura O-RAN, com o objetivo de minimizar custos de ativação dessas funções. Para tanto, utilizamos um formulação de programação linear inteira (ILP), com caminhos pré-definidos, para modelar o problema de otimização e resolvê-lo. Para avaliar a eficácia do método ILP e analisar o desempenho da rede, executamos um amplo conjunto de experimentos em diferentes cenários de rede.**

***Palavras-Chave*— Open-RAN, Rede 5G, Programação Linear.**

***Abstract*— The 5th generation mobile networks (5G) based on virtualized and centralized radio access networks will require cost-effective and flexible solutions for satisfying high-throughput and latency requirements. The next generation fronthaul or Open Radio Access Networks (O-RAN), baseband processing is split and performed in functions units. In this paper, we address the latency-aware and node capacity of functions placement problem in O-RAN. To this end, we make use of integer linear programming (ILP) with predefined paths in order to formulate the optimization problem and to solve it. To assess the effectiveness of the MILP method and analyze the network performance, we run a broad set of experiments in different network scenarios.**

***Keywords*— Open-RAN, 5G network, Linear Programming.**

I. Introdução

As redes 5G tem três cenários de uso mais comuns associados, são eles: *Enhanced Mobile Broadband* (eMBB), *Massive Machine-Type Communication* (mMTC) e *Ultra-Reliable and Low Latency Communications* (URLLC). Um quarto cenário de uso, particularmente importante para o Brasil, é o cenário *Enhanced Remote Area Communications* (eRAC). Cada um destes quatro cenários tem suas características, em termos de requisitos de desempenho e aplicações avançadas. Em resposta a essas aplicações, o provisionamento desses serviços em redes

Karcius Day Rosário Assis, Departamento de Engenharia Elétrica e de Computação, Universidade Federal da Bahia (UFBA), Salvador, BA, Brasil, e-mail karcius.assis@ufba.br; Alex Ferreira dos Santos, Universidade Federal do Recôncavo da Bahia (UFRB) e Programa de Pós-Graduação em Ciência da Computação da Universidade Estadual de Feira de Santana (UEFS), Feira de Santana, BA, Brasil, e-mail: alex.ferreira@ufrb.edu.br. Raul Camelo de Andrade Almeida Júnior, Departamento de Eletrônica e Sistemas, Universidade Federal de Pernambuco (UFPE), Recife, PE, Brasil, e-mail: ralmeida.ufpe@gmail.com.

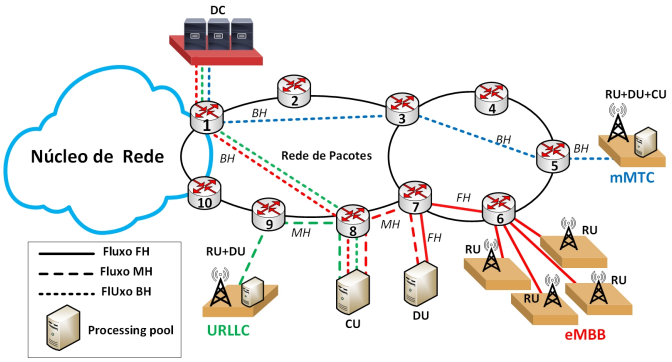

Fig. 1. Exemplo de uma O-RAN, inspirada na Figura de [8]

5G exigirá a implementação de arquiteturas de rede de acesso de rádio (RAN) virtualizadas para estabelecer conexões de resposta rápida e reduzir as latências de acesso dos usuários aos serviços necessários [1].

Alinhada com os padrões 3GPP [2], as (RANs) de última geração se diferenciam das soluções anteriores e dos padrões de particionamento de banda base que antes eram integrados à Unidade de Rádio Remota (RRU) e à Unidade de Banda Base (BBU). Recentemente, houve evolução para funções mais desagregadas, incluindo Unidade de Rádio (RU), Unidade Distribuída (DU) e Unidade Central (CU). Esses módulos podem ser separados por uma variedade de opções de divisão de função de banda base adaptadas às diferentes necessidades de serviço. As unidades DU e CU são virtualizadas, o que permite executar o processamento de RF em processadores de uso geral em uma instalação de pool de processamento (PP) [3].

Com essa proposta de desagregação, viu-se a necessidade em deixar as interfaces de comunicação entre os elementos da RAN abertas e inteligentes, criando interoperabilidade entre elas [4], a qual é chamada de Open-RAN (O-RAN), que estão atualmente em desenvolvimento dentro da iniciativa *Alliance* O-RAN, [5]. Veja a Figura 1 para um exemplo de arquitetura O-RAN. Na arquitetura O-RAN, três seções podem ser distinguidas, a saber, *fronthaul* (FH), *midhaul* (MH) e *backhaul* (BH) [6], [7].

Fronthaul (FH) é a parte da rede entre a RU e a DU, *midhaul* (MH) é entre a DU e a CU, e o *backhaul* (BH) é entre a CU e um data center (DC). No cenário O-RAN considerado, os dados entre os RUs, DUs, CUs e DCs são encapsulados e transmitidos como pacotes.

Um problema desafiador na O-RAN é o provisionamento de qualidade de serviço (QoS) para os fluxos de tráfego, que muitas vezes têm requisitos de latência rigorosos. Esses requisitos estão relacionados a serviços 5G de baixa tolerância de latência no processamento (DU/CU) na rede de acesso

de rádio. Portanto, o cuidado com restrições de latência é fundamental para uma adequada entrega de fluxos de tráfego. Adicionalmente, os requisitos de largura de banda exigidos por alguns serviços tornam a alocação de DUs e CUs uma tarefa de otimização desafiadora.

A literatura atual apresenta alguns problemas na concepção de modelos ou políticas de posicionamento de DU/CU nas modernas arquiteturas O-RAN. Para solucionar problemas em relação aos modelos, nossas principais contribuições neste artigo são:

- O ILP com caminhos pre-definidos, proposto neste artigo, diminui a complexidade computacional e também permite introduzir um conjunto de restrições que representam limites de latências para certos serviços.
- O ILP proposto busca remediar a falta de estratégias para calcular a latência entre funções. Estudos anteriores de latência não abordaram esta questão ou a trataram com restrições não lineares.
- Nosso ILP, por meio do provisionamento eficiente de posicionamento DU/CU para cada solicitação na rede, pode ajudar um operador de rede a evitar o alto consumo de recursos; isso se torna particularmente importante se alguma resiliência é desejada. Neste estudo, nós usamos um fator de resiliência que permite proteção dedicada para as requisições.

O restante deste artigo é organizado da seguinte forma. A Seção II apresenta os principais detalhes e características de um exemplo de O-RAN e introduz a metodologia usada. A Seção III apresenta a formulação ILP com caminhos pré-definidos. A Seção IV apresenta simulações e resultados numéricos e, finalmente, A Seção V faz um resumo dos principais resultados e propõe estudos futuros.

II. Cenário de desagregação na O-RAN

A especificação O-RAN tem como base a desagregação das funcionalidades da estação rádio base em unidades [10], que nós chamaremos de funções f (ver Figuras 1 e 2), representadas nesse artigo como:

$$f = \begin{cases} 1 & \text{se é uma RU} \\ 2 & \text{se é uma DU} \\ 3 & \text{se é uma CU} \\ 4 & \text{se é um DC} \end{cases} \quad (1)$$

Nós decidimos formar a *chain* com a (RU) e (DC) sendo ingresso e egresso da requisição, respectivamente. Mas, de fato as funções desagregadas são a DU e CU. Nós decidimos usar dois tipos de requisições: eMBB, que executa funções na *chain* (RU - DU - CU - DC) e uRLLC, onde RU é alocado junto a DU, resultando na *chain* (RU/DU - CU - DC).

Logo, a arquitetura consiste de uma *service function chain* (SFC) da fonte RU para o destino DC. As RUs agregam todos os serviços da sua área de cobertura para dar início à formação da SFC. Cada SFC terá requisitos diferentes de máxima latência e largura de banda. Especificamente, URLCC está associado a aplicações que demandam latência muito baixa e alta confiabilidade da rede, como por exemplo, aplicações de Internet Tátil. Já eMBB é o cenário voltado para aplicações que demandam a oferta de maiores taxas de transmissão de dados para os usuários e, consequentemente, uma maior capacidade da rede para escoar tráfego. Trata-se do cenário comumente associado à evolução das redes de comunicações.

Para facilitar a leitura deste artigo, dois exemplos são apresentados para examinar a nossa formulação ILP. No exemplo da Figura 2, há quatro requisições, r=1, r=2, r=3 e r=4 dos nós 1, 2, 8 e 9, respectivamente, todas do tipo eMBB, exceto a requisição r=3, que é URLCC. O termo L_{FH} refere-se à latência da seção *fronthaul* (se nós assumimos que uma latência máxima de 10 ms é permitida entre funções, e que essa latência é devido apenas à propagação do fluxo pelos links de fibra óptica, onde a velocidade de transmissão do sinal é $c = 200000$ km/s, isso equivale a assumir uma distância máxima de transmissão de 2000 km entre funções).

Assumindo que cada link da rede tenha 2000 km, nós podemos usar saltos em vez de tempo para a latência máxima entre funções. Logo, neste exemplo o número de saltos irá ser considerado como um tipo de latência. Então, nós definimos uma latência de um salto (L_{FH}=1) para FH e não há restrições de latência para MH e BH.

No exemplo da Figura 2, um único caminho ($|P_r|$=1) é disponível para cada uma das quatro requisições, ou seja há apenas p=1. Por outro lado, na Figura 3, dois caminhos alternativos ($|P_r|$=2) são disponíveis para cada requisição, isto é, as requisições podem usar p=1 ou p=2. A Tabela I mostra os caminhos pré-definidos para os exemplos.

Nós consideramos que um nó é ativo quando ele executa qualquer função DU ou CU.

É possível perceber na parte inferior da Figura 2 que o SFC formado para cada requisição exige mais ativação de nós do que no caso da Figura 3 visto que na primeira, 5 nós foram ativados no total com funções f=2 e/ou f=3, enquanto que na Figura 3, apenas 4 nós foram ativados com funções f=2 e/ou f=3, demonstrando as vantagens de ter mais caminhos alternativos, $|P_r|$=2, para reduzir o custo de ativação dos nós. Também fica evidente que restrições de latência podem mudar completamente a configuração de nós ativados, e essas questões serão discutidas ao longo do artigo.

TABELA I
2 CAMINHOS ALTERNATIVOS PARA CADA REQUISIÇÃO

requisição r	p=1	p=2
1	(1-2-3-4-5-6)	(1-11-10-9-8-7-6)
2	(2-3-4-5-6)	(2-1-11-10-9-8-7-6)
3	(8-7-6)	(8-9-10-4-5-6)
4	(9-8-7-6)	(9-10-4-5-6)

III. Modelo ILP

Nós consideramos um grafo bi-direcional $G = (N, E)$ representando N nós e E arcos ou links. Para cada arco (m, n) em E, nós denotamos d_{mn} como a sua distância física. Cada nó $u \in N$ tem associado a ele uma capacidade c_u, referindo-se ao número máximo de funções que podem ser instaladas no nó u. O conjunto de todas as funções é denotada por F. O custo de ativação das funções DU e CU é denotado por ζ_u. Nós definimos como R o conjunto de todas as requisições; cada requisição $r \in R$ é definida por um nó fonte $s_r \in N$ e um nó destino $d_r \in N$, e também pela máxima latência permitida entre as funções $L_{FH}, L_{MH}, L_{BH} \in R_+$ e um conjunto de

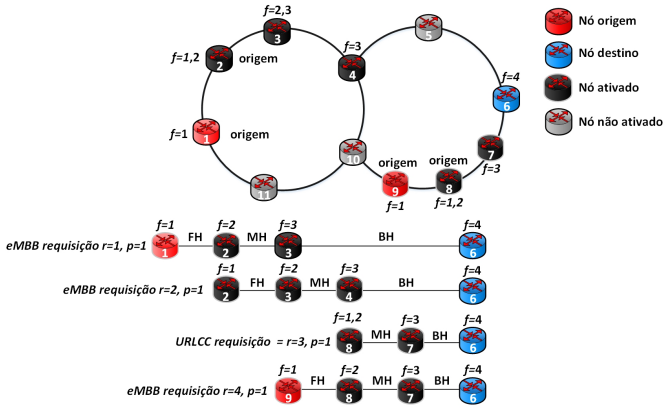

Fig. 2. Arquitetura O-RAN. Rede com 11 nós e 4 requisições, dos nós 1, 2, 8 e 9 para o nó 6 (DC). Também é mostrado o SFC de cada requisição r com $|P_r|$=1 e L_{FH}=1.

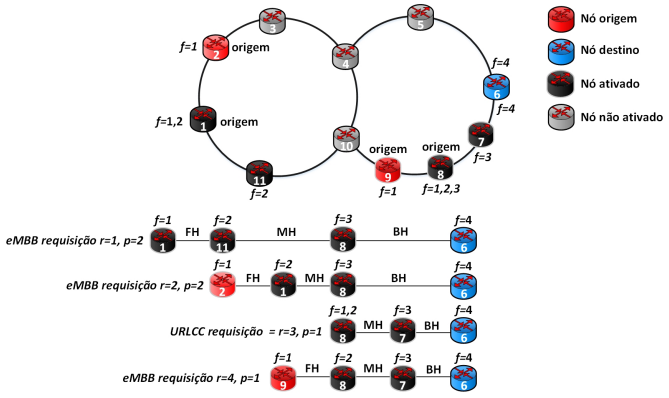

Fig. 3. Arquitetura O-RAN. Rede com 11 nós e 4 requisições, dos nós 1, 2, 8 e 9 para o nó 6 (DC). Também, é mostrado o SFC de cada requisição r com $|P_r|$=2 e L_{FH}=1.

restrições sobre a colocação das funções, dependendo do tipo de requisição r. $F^r \subseteq F$: nós denotamos $f \prec g$, o fato de que uma função f deve ser visitada antes de uma função g para a requisição r. Se f e g são instaladas em um mesmo nó u, as restrições de precedência ainda são satisfeitas. Todas as notações são descritas na Tabela II.

Como usual, uma solução do problema é chamada de viável se ela satisfaz todas as restrições. Neste modelo, o custo de uma solução é definido como a soma de todos os custos, ζ_u, $u \in N$, dos nós ativados por funções DU e CU.

Lembrando que, neste artigo, uma *chain* é estruturada da maneira representada por (1).

A. Formulação Matemática

- **Função objetivo:**

$$Minimize : \sum_u w_u \zeta_u \tag{2}$$

O custo das ativações dos nós, para atender todas as SFCs, com as funções (f=2 e/ou f=3) na rede deve ser minimizado.

- **Estabelecimento da requisição:**

$$\textstyle\sum_p B^{r,p} = \alpha^r h^r \ , \qquad \forall r \in R \tag{3}$$

$$B^{r,p} \leq h^r \quad \forall r \in R, p \in P_r \tag{4}$$

Cada requisição de *commodity* h^r deve ser roteada por um caminho $p \in P_r$ se α^r=1. Entretanto, se α^r=2, o *commodity* é roteada por 2 caminhos disjuntos ($p_1 \in P_r$ e $p_2 \in P_r$) para fins de backup.

- **Restrições de capacidade das funções:**

TABELA II
NOTAÇÕES USADAS EM TODO O ARTIGO

	Notações	Descrição
Parâmetros	R	conjunto de requisições a serem alocadas na rede. r = 1, 2, ..., $\|R\|$ é uma das requisições existentes.
	E	conjunto de links da rede. e = 1, 2, ..., $\|E\|$ são os links existentes da rede.
	P_r	conjunto pré-definido de caminhos candidatos para a requisição r, $p \in \{1, 2, \ldots, \|P_r\|\}$. Em caso de resiliência, esses caminhos são disjuntos.
	$q_{mn}^{r,p}$: = 1	se o caminho p da requisição r passa pelo nó m e logo posteriormente pelo nó n, i.e., se o link físico entre (m, n) pertence ao caminho p da requisição r.
	B_w	máximo número de requisições que pode passar por qualquer link físico.
	h^r	*commodity* da requisição r (binário).
	s_r	nó fonte da requisição r.
	d_r	nó destino da requisição r.
	c_u	máximo número de núcleos no nó u (capacidade do nó).
	$p_r(f, g)$	processamento da precedência das VNFs. $p_r(f, g) = 1$ (resp. -1) se f deve ser processada antes (resp. depois) g na cadeia de funções de serviço (*service function chain* - SFC) da requisição r; 0 caso contrário.
	$dist_{m,n}$	distância entre os nós m e n.
	L_{FH}	máxima latência de *fronthaul*.
	L_{MH}	máxima latência de *midhaul* .
	L_{BH}	máxima latência de *backhaul*.
	ζ_u	custo de instalação do nó u. $F_r \subset F$: conjunto de funções virtualizadas requisitadas por r.
	$\delta_{f,g}^r$	indica uma *anti-affinity* lei entre f e g, i.e. f e g podem ambas serem processadas (ou não) no mesmo nó. $\delta_{f,g}^r = 1$ em caso da função f e g da requisição r poder ser executada no mesmo nó e $\delta_{f,g}^r = 0$ caso contrário. Por definição, $\delta_{f,f}^r = 1$.
	ξ_u^{rp} = 1	se o nó u pertencente ao caminho p executa a requisição r.
	α^r	*overhead* para implementação da proteção.
	M	um número grande.
Variáveis	$B^{r,p}$ Binária	serviço da requisição r que é roteado através do caminho p.
	x_u^f	denota o número de VNFs f requeridas no nó u.
	$y_u^{r,p,f}$ Binária	informa se o tráfego da requisição r no caminho p é designado para executar a VNF f no nó u ($y_u^{r,p,f} = 1$) ou não ($y_u^{r,p,f} = 0$).
	$Y_u^{r,p,f}$ Binário	variável auxiliar para endereçar se a VNF f já foi executada no nó u ou em algum nó anterior a ele. $Y_u^{r,p,f} = 1$ quando a função f já foi executada em u ou em algum nó anterior a u ao longo de sua rota; $Y_u^{r,p,f} = 0$ caso contrário.
	w_u Binário	indica se o nó u tem alguma função ativa.
	$t^{r,p,f}$	atraso da origem até o nó onde a função f é executada.

$$\sum_r \sum_p y_u^{r,p,f} = x_u^f, \qquad \forall\, u \in N, f \in F, \tag{5}$$

$$\sum_{f=2,3} x_u^f \leq c_u w_u, \qquad \forall\, u \in N, \tag{6}$$

A restrição de capacidade de nó garante que as funções f=2 (DU) e f=3 (CU) ativadas em um nó u não exceda a capacidade c_u daquele nó.

- **Restrições de capacidade de link**:

$$\sum_{r,p} q_{m,n}^{r,p} B^{r,p} \leq B_w \qquad \forall m, n \tag{7}$$

Ela garante que as requisições passando por um link não excedam a capacidade do link.

- **Restrições *Anti-affinity*:**

Anti-affinity implica que funções de uma requisição r não podem ser alocadas na mesma localização. Logo, se f e g obedecem a essa lei, $\delta_{f,g}^r = 0$ e f ou g da

requisição r podem ser designadas a um mesmo nó u. Isto irá depender do tipo de tráfego e deve existir somente entre funções de um mesmo caminho p.
Para os dois tipos de requisição r neste artigo, (eMBB and URLCC), nós temos as seguintes leis.

a) **(DU-CU-DC *chain*)**. DU, CU e DC devem obedecer à regra *anti-affinity*. Logo, de acordo com a restrição 2 e valores de f e g nós temos:

$$y_u^{r,p,2} + y_u^{r,p,3} \leq 1 + \delta_{2,3}^r \quad \forall r \in R, \forall p \in P^r, \forall u \in N, \tag{8}$$

$$y_u^{r,p,3} + y_u^{r,p,4} \leq 1 + \delta_{3,4}^r \quad \forall r \in R, \forall p \in P^r, \forall u \in N. \tag{9}$$

b) **eMBB lei (RU-DU-CU-DC *chain*)**.
Se nós definimos uma requisição r como eMBB, nós devemos adicionar as restrições abaixo para aquela requisição:

$$y_u^{r,p,1} + y_u^{r,p,2} \leq 1 + \delta_{1,2}^r \quad \forall r \in R, \forall p \in P^r, \forall u \in N. \tag{10}$$

c) **URLCC lei (RU/DU-CU-DC *chain*)**.

Também, se nós definimos qualquer requisição r como URLCC, em vez de (10) nós devemos implementar a restrição (11) abaixo para a requisição:

$$y_{s_r}^{r,p,2} = 1 \qquad \forall r \in R, \forall p \in P^r, \tag{11}$$

- **Lei de precedência f para g, [11]**:

$$(Y_u^{r,p,f} - Y_u^{r,p,g}) P_r(f,g) \geq 0 \quad \forall r \in R,\ \forall p \in P^r,\ \forall u \in N,\ \forall f,g \in F_r. \tag{12}$$

- **Restrições de Latência**:

$$t^{r,p,f} = \sum_{m,n} q_{m,n}^{r,p} \left(1 - Y_u^{r,p,f}\right) dist_{m,n} \tag{13}$$

a) **Atraso entre RU e DU (latência FH), DU e CU (latência MH), e CU e DC (latência BH) na rota p**:

$$|t^{r,p,f} - t^{r,p,g}| \leq \begin{cases} L_{FH} & \text{se } f\text{=1 e } g\text{ =2} \\ L_{MH} & \text{se } f\text{=2 e } g\text{ =3} \\ L_{BH} & \text{se } f\text{=3 e } g\text{ =4} \end{cases} \tag{14}$$

O atraso entre as funções f e g (de um mesmo caminho) não pode exceder o máximo valor entre elas. Note que o tempo para calcular a latência fim a fim já é garantido nas rotas de entrada.

- **Restrições de Fonte/Destino:**
As restrições seguintes definem as restrições da fonte e destino.

$$y_u^{rpf} \leq \xi_u^{rp} \tag{15}$$

$$\sum_u y_u^{rpf} \xi_u^{rp} \leq 1 \tag{16}$$

$$\sum_u y_u^{rpf} \xi_u^{rp} \geq B^{r,p} \tag{17}$$

$$q_{uv}^{rp} \left(Y_v^{rpf} - Y_u^{rpf} - y_v^{rpf}\right) = 0 \quad \forall\, r, p, f, u, v \tag{18}$$

a) **RU é designado para a fonte** :.

$$y_{s_r}^{r,p,1} \leq 1 \quad \forall r \in R, \forall p \in P^r, \tag{19}$$

$$y_u^{r,p,1} = 0 \quad \forall r \in R, \forall p \in P^r, \forall u \neq s_r \in N \tag{20}$$

b) **DC é designado para o destino:**.

$$y_{d_r}^{r,p,4} \leq 1 \quad \forall r \in R, \forall p \in P^r, \tag{21}$$

$$y_u^{r,p,4} = 0 \quad \forall r \in R, \forall p \in P^r, \forall u \neq d_r \tag{22}$$

IV. Exemplos Numéricos

Nesta seção apresentamos alguns resultados computacionais para comparar a ILP proposta com a formulação tradicional de nó-link (NL) (o termo ILP-NL foi adotado para se referir à formulação de nó-link). Na formulação ILP-NL, os caminhos não são pré-definidos, logo há mais possibilidades de escolha. Entretanto, o custo computacional aumenta.

Nesta comparação dos dois modelos, focamos sobre o tempo de CPU e na qualidade das soluções obtidas. Todos os testes foram realizados utilizando o *solver* comercial CPLEX e uma máquina com Intel (R) Xeon (R) CPU E5-2650 v2 processador com clock de 2.60GHz e 252GB de RAM no sistema operacional Windows.

Duas redes de tamanhos diferentes foram usadas no experimento: uma rede de anel duplo de 11 nós (DRING-11), apresentada anteriormente na Fig.2, e uma rede em malha de 37 nós (MESH-37), apresentada na Figura 6.

Assumimos que há um PP conectado a cada nó switch; portanto, o número total de PPs é de N nas redes consideradas. Assumimos que uma dada função RU (fonte da requisição) está conectada a um switch, e todos os rádios conectados a uma função RU constituem um cluster. Assumimos que o comprimento de um link é de 2000 km (1 salto) para as duas redes e que os links têm capacidade de largura de banda ilimitada $B_w = M$. Também assumimos que os nós têm capacidade de processamento ilimitada $c_u = M$, pois o artigo se concentra no posicionamento das funções e latência máxima das requisições.

Nós definimos o custo do nó como ζ_u=10. No entanto, esse valor pode variar com as características nodais ou a localização geográfica da rede. Por exemplo, nós em locais sensíveis podem ser mais caros.

Para fins de análise, definimos Ru como o número de solicitações URLLC e Re como o número de solicitações eMBBB. Os dois cenários analisados são:

- **Cenário I**. Requisições Ru e Re com α^r=1.
- **Cenário II**. Requisições Ru e Re com α^r=2 (*Dedicated Path Protection* - DPP).

A. Planejamento da O-RAN com o ILP (Resiliência e Latência)

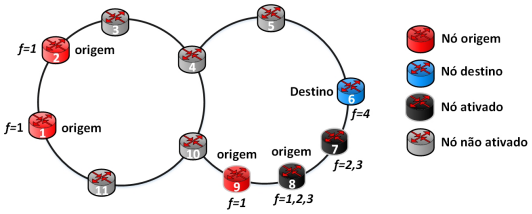

Fig. 4. O-RAN. Rede de 11 nós, Re=4 requisições dos nós 1, 2, 8 e 9 para 6 (DC), $|P_r|$=2 e sem restrições de latência ou resiliência, α^r=1.

Embora a O-RAN traga vários benefícios, o provisionamento de serviços de rede sensíveis à latência em uma infraestrutura baseada em virtualização continua sendo um desafio, pois exigem prazos de serviço rigorosos. Para o D-RING11, analisaremos o melhor caso em termos de posicionamento de funções sem limites de latência, mas com possibilidade de resiliência. Em seguida, mostramos a simulação para o MESH-37 com restrições de latência.

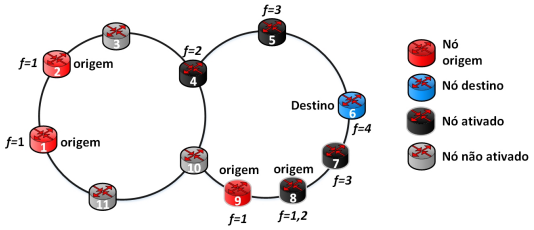

Fig. 5. O-RAN. Rede de 11 nós, Re=4 requisições dos nós 1, 2, 8 e 9 para 6 (DC), $|P_r|$=2 e sem restrições de latência, mas com resiliência, α^r=2.

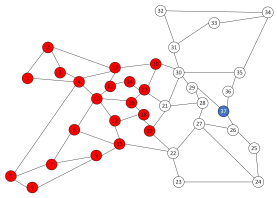

Fig. 6. (MESH-37)

1) D-RING11: No primeiro exemplo, Figura 4, temos o Cenário I com Re=3 (nós 1, 2 e 9) e Ru=1 (nó 8), sem limites de latência e o nó DC é o 6. A simulação ILP encontra apenas dois nós ativados. Embora a ênfase na reutilização de VNFs já implementadas para novas solicitações de serviço possa reduzir os custos, ela pode não ser adequada para serviços com prazos de serviço rigorosos.

Para a rede com resiliência, ou seja proteção dedicada para as requisições, podemos observar na Figura 5 que, no Cenário II, em que se fornece DPP para todas as requisições, o planejamento proposto consome o dobro de recursos do Cenário I, e 4 nós são ativados.

2) MESH-37: Na Figura 7, mostramos os resultados para o Cenário I com Re=20 (do nó 1 até o nó 20) e o nó 37 definido como o DC. Também limitamos a latência do FH, mas com uma rede mais complexa e com diferentes conjuntos de caminhos. A partir do gráfico, pode-se ver que, de longe, o relaxamento da latência diminui o número de nós ativados. Por exemplo, com $|P_r|$=4 rotas alternativas para cada demanda, aumentar L_{FH} de 1 para 2, 2 para 3 e 3 para 4 é suficiente para reduzir em 1/3, 1/3 e 1/4, respectivamente, o custo dos nós instalados.

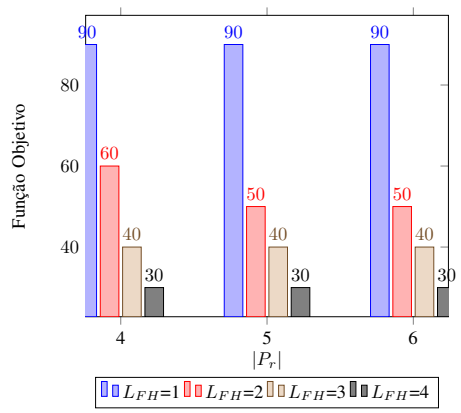

Fig. 7. (MESH-37), (DC) = nó 37. Custo de nós ativados em função da latência de FH para diferentes quantidades de $|P_r|$.

B. Análise da Complexidade Computacional

A formulação de link e nó (ILP-NL) de [11], adaptada para O-RAN, e a formulação de caminho ILP proposta, foram configuradas sem restrições de latência e com requisições Re do nó 1 até o nó 20 usando o Cenário I. A Tabela III mostra o tempo de execução da formulação por caminho e da formulação ILP-NL. Como pode ser visto, o tempo de execução aumenta um pouco com $|P_r|$ devido ao maior número de variáveis na formulação. No entanto, a nova formulação por caminho torna possível resolver instâncias realistas em um curto período de tempo (cerca de segundos). Portanto, a formulação por caminho mostrou-se ser capaz de superar o desempenho em termos de tempo de CPU. Foi notado que com $|P_r|$=6, a formulação por caminho alcança o mesmo valor ótimo atingido por ILP-NL. A formulação ILP-NL encontra a solução ótima com mais de 2h de simulação com a topologia MESH-37. Logo, a formulação por caminho tem um resultado muito bom em comparação com o *benchmark* ILP-NL, que tem uma complexidade elevada, [11].

TABELA III
TEMPO DE SIMULAÇÃO, MESH-37 COM 20 REQUISIÇÕES DO TIPO EMBBB

	$\|P_r\|$=1	$\|P_r\|$=2	$\|P_r\|$=3	$\|P_r\|$=4	$\|P_r\|$=5	$\|P_r\|$=6	ILP-NL
time	0.14s	0.25s	0.34s	0.45s	0.63s	1.39s	> 2h

V. CONCLUSÃO

Este artigo propõe uma formulação ILP com caminhos pré-definidos que minimiza o custo de alocações de funções (DU/CU) na arquitetura O-RAN. No mesmo modelo, restrições de latência entre funções são introduzidas como uma medida adaptável para projetos de redes com características diferentes de serviços. Nas simulações, a formulação proposta mostrou vantagem em termos de complexidade computacional em relação a formulações que não adotam caminhos pré-definidos.

REFERÊNCIAS

[1] A. M. Alberti *et al.* “OPEN RAN: A Conexão do Futuro". Disponível online: https://inatel.br/cxsc/documents/white-paper-open-ran.pdf. accessado em 06 Abril de 2024).

[2] 3GPP. “Study on new radio access technology: Radio access architecture and interfaces". Technical Report, 38.801, 3GPP. Release 14. Disponível online: https://www.etsi.org/deliver/etsi_tr/138900_138999/138912/14.01.00_60 /tr_138912v140100p.pdf. accessado em 06 Abril de 2024).

[3] A. Garcia-Saavedra *et al.* “WizHaul: On the centralization degree of cloud RAN next generation fronthaul."*IEEE Transactions on Mobile Computing* v.17, n.10, pp. 2452-2466, 2018.

[4] V. Ferro, G. Monteiro, J. G. Oliveira e L. F. Silva. “Implementação de um Ambiente Open RAN LTE Conteinerizado e Virtualizado."XL Simpósio Brasileiro de Telecomunicações e Processamento de Sinais, pp. 25-28, 2022.

[5] D. A. L Marques *et al.* “Desagregando e Softwarizando as Redes de Celulares e o Programa OpenRAN Brasil."Sociedade Brasileira de Computação, pp. 209-254, 2023.

[6] M. Klinkowski. “Optimized Planning of DU/CU Placement and Flow Routing in 5G Packet Xhaul Networks."*IEEE Transactions on Network and Service Management*, v. 21, n. 1, pp. 232-248, 2024.

[7] H. Li *et al.* “NetMind: Adaptive RAN Baseband Function Placement by GCN Encoding and Maze-solving DRL."IEEE Wireless Communications and Networking Conference (WCNC), 2024.

[8] M. Klinkowski. “Latency-aware DU/CU placement in convergent packet-based 5G fronthaul transport networks."*Applied Sciences*, v. 10, n.21:7429, 2020.

[9] IBM. “CPLEX Optimizer."Disponível online: http://www.ibm.com/. Accessado em 06 Abril de 2024).

[10] R. Wang, J. Zhang, Z. Gu, S. Yan, Y. Xiao e Y. Ji. “Edge-enhanced graph neural network for DU-CU placement and lightpath provision in X-Haul networks."*Journal of Optical Communications and Networking*, v. 14, n.10, pp. 828-839, 2022.

[11] Z. Allybokus, N. Perrot, J. L. L. Maggi e E. Gourdin. “Virtual function placement for service chaining with partial orders and anti-affinity rules."*Networks*, v. 71, n.2, pp. 97-106, 2018.