Heurística Escalável Para o Problema de Alocação de vBBU e Comprimento de Onda em Cloud-Fog RAN

Matias R. P. dos Santos[1,2], **Rodrigo I. Tinini**[3], **Tiago Januario**[1,4], **Gustavo B. Figueiredo**[1]

[1]Universidade Federal do Bahia - UFBA - Brasil

{matiasrps,januario,gustavobf}@ufba.br

[2] Instituto Federal do Ceará - Campus Acopiara

[3] Universidade Federal do ABC - UFABC - Brasil

rodrigo.tinini@ufabc.edu.br

[4] Boston University - BU - Estados Unidos

januario@bu.edu

Abstract. *Centralized baseband processing imposes high bandwidth and fronthaul delay requirements on Cloud Radio Access Networks (CRAN) deployments. To address these issues, we propose a hybrid architecture called Cloud-Fog RAN (CF-RAN) which uses fog computing and network functions virtualization (NFV) to place virtualized Baseband Processing Units (BBUs) on fog nodes closer to users, thus alleviating fronthaul constraints. In this article, we utilize integer linear programming (ILP) and linear relaxation to determine the optimal location for virtual BBU processing (vBBU) with a focus on energy efficiency through the minimal activation of processing elements in the network. Our main objective is to present a scalable alternative to the optimal solution with reduced execution time. The results indicate that a heuristic approach based on linear relaxation significantly reduces the use of computational resources and execution time.*

Resumo. *O processamento de banda base centralizado impõe requisitos de alta largura de banda e atraso de fronthaul em implantações de Cloud Radio Access Networks (CRAN). Para resolver esses problemas, consideramos uma arquitetura híbrida chamada Cloud-Fog RAN (CF-RAN). Ele alivia as restrições de fronthaul usando computação em fog e virtualização de funções de rede (NFV) para colocar Unidades de processamento de banda base (BBUs) virtualizadas em nós de fog mais próximos dos usuários. Neste artigo, utilizamos programação linear inteira (PLI) e relaxação linear para resolver o problema da colocação de processamento de BBU virtual (vBBU) voltado para eficiência energética a partir da ativação mínima de elementos de processamento na rede. O principal objetivo deste artigo é apresentar uma alternativa escalável para a solução ótima com redução no tempo de execução. Os resultados mostram que uma abordagem heurística baseada no relaxamento linear reduz significativamente a utilização de recursos computacionais e o tempo de execução.*

1. Introdução

Propostas arquiteturais, como a Cloud Radio Access Network (CRAN), surgiram para contornar os desafios do CAPEX e OPEX das redes móveis 5G e Beyond 5G (B5G). A

CRAN reduz esses custos centralizando os elementos de processamento de banda base em uma instalação de nuvem chamada de BBU pool [Wu et al. 2015]. Nessa arquitetura, as BaseBand Units (BBUs) são desacopladas das células da rede e centralizadas em uma nuvem, comumente chamada de BBU pool. Antenas de baixo consumo energético chamadas de 5G Radio Units (RUs) são utilizadas nas células da rede para se conectar com os usuários móveis e com a nuvem. A conexão das RUs e da nuvem é feita por meio de uma rede de transporte chamada de fronthaul. Em relação ao 6G, o uso de frequências mais altas, computação de borda e inteligência artificial são tecnologias-chave para fornecer menor latência e maior velocidade. A literatura e a indústria já estão propondo soluções e tecnologias-chave para fornecer os serviços desejados.

Embora a CRAN forneça economia de energia devido à centralização, surgem problemas quando a rede experimenta altas cargas de tráfego. Quanto maior a carga na nuvem, maior será a latência e a largura de banda demandada pelo fronthaul [Figueiredo et al. 2016]. Altas cargas na nuvem podem impedir que novas RUs sejam processadas nas BBUs, diminuindo assim a cobertura da rede. Esses desafios estimularam o estudo e o aprimoramento de tecnologias e técnicas que possam contornar tais limitações.

Como alternativa para contornar os problemas de consumo energético, alta latência e cobertura limitada introduzidos pela CRAN, foi proposta a arquitetura híbrida Cloud-Fog RAN (CF-RAN) [Tinini et al. 2020]. A CF-RAN traz o conceito de computação em névoa (*Fog Computing*) para instalar nós de processamento chamados de nós de névoa, ou fog nodes, próximos aos usuários da rede. Os nós de névoa são utilizados para receber novas BBUs quando a nuvem estiver sobrecarregada. A inclusão de mais BBUs aumenta a cobertura da rede ao possibilitar mais processamento de banda base. Além disso, a CF-RAN se utiliza do paradigma de Virtualização de Funções de Rede para implementar BBUs virtualizadas, que são ativadas e desativadas sob demanda em função da carga da rede. Dessa forma, os nós de processamento, seja a nuvem ou nós de névoa, podem ser ativados ou desativados dinamicamente. Naturalmente, a inclusão de nós de névoa acarreta um maior consumo de energia para a rede.

Observa-se que haverá consumos adicionais de energia ou bloqueio de demandas caso as virtualized BBUs não sejam devidamente ativadas em função da carga da rede. Note que existem diversas abordagens para o dimensionamento da rede em uma arquitetura CF-RAN. Uma possibilidade é utilizar algoritmos de otimização para determinar a melhor alocação de recursos de processamento e de energia para cada nó de processamento na rede. Outra possibilidade é utilizar técnicas de aprendizado de máquina para prever a carga da rede e ajustar dinamicamente a alocação de recursos. Além disso, a escolha da localização dos nós de névoa também é importante, pois eles devem estar localizados de forma estratégica para garantir uma cobertura adequada da rede e minimizar o consumo de energia.

Devido aos desafios de latência e altas demandas de largura de banda impostas pela centralização, apresentou-se pela indústria e academia o uso de particionamento de funções de banda base como técnica chave para contornar esses desafios. Nesta técnica ocorre a quebra da cadeia de processamento de banda base em duas ou mais porções para aliviar a carga no *fronthaul* e nos nós de processamento intermediários. Assim, o processamento dos sinais de banda-base dos RUs pode ser realizados em vBBUs localizadas em

múltiplos nós de processamento. Observou-se que o alívio das restrições no *fronthaul* e nos nós de processamento é, de fato, alcançado com o particionamento correto, dado o estado da rede [Larsen et al. 2019]. Estas duas soluções, arquiteturas híbridas e particionamento de funções, são abordagens promissoras para a resolução do dimensionamento de rede com alívio das restrições impostas pela CRAN.

No trabalho anterior [Santos et al. 2021a], apresentou-se uma solução ótima desenvolvida mediante formulação matemática de programação linear inteira (PLI) para dimensionar a rede e propor o particionamento funcional. Nela, foram descritas as expressões que modelam a operação da rede e todas a suas restrições de funcionamento. A função de otimização buscava centralizar, ao máximo, as funções de processamento de banda base no *BBU Pool* para promover minimização do consumo energético. No entanto, observou-se que a solução de PLI sofre problemas de escalabilidade, inviabilizando seu uso em cenários de redes de larga escala. À medida que novos elementos de processamento precisam ser ativados para acomodar novas demandas, o tempo de execução do modelo de PLI cresce exponencialmente. Isto é, o tempo de execução da PLI para encontrar a solução ótima cresceu em função do crescimento da demanda por largura de banda [Santos et al. 2021a].

Dado o problema de escalabilidade do PLI, este trabalho apresenta como contribuição a aplicação da relaxação linear para decidir o particionamento flexível de funções de processamento e o dimensionamento de rede em uma arquitetura híbrida CF-RAN. Isto é, foca na correta ativação das vBBUs para receber as funções de processamento particionadas. A relaxação linear alivia o tempo de execução do modelo de PLI ao quebrar a obrigatoriedade da integralidade das variáveis. No entanto, essa quebra da integralidade apresenta soluções não inteiras. Decorrente disso, nós utilizando a relaxação linear (*relax and round*) das variáveis e a validação de funções de pós processamento que verificam se todas as restrições do PLI foram respeitadas [Santos et al. 2021b]. Como resultados, este trabalho apresenta soluções que escalam em redes de larga escala, apresenta resultados próximos ao ótimo na função objetivo e apresenta resultados viáveis de dimensionamento da rede.

O restante deste trabalho está organizado da seguinte forma: A seção 2 discute os principais trabalhos relacionados. A seção 3 apresenta a arquitetura CF-RAN considerada neste trabalho. A seção 4 discute o particionamento das funções de banda-base e dimensionamento dos recursos da rede. A formulação da relaxação linear proposta é apresentada na seção 5. Os resultados são apresentados na seção 6 e a conclusão na seção 7.

2. Trabalhos Relacionados

Esta seção apresenta os trabalhos relacionados diretamente com este trabalho. Mais especificamente, esta seção apresenta trabalhos relacionados ao *fronthaul*, uso de PLI e/ou relaxação para viabilizar a escalabilidade da PLI. Um dos temas mais estudados pela indústria e pela academia é o uso e aplicação de redes ópticas em arquiteturas de redes móveis.

2.1. Fronthaul óptico e questões associadas

A rede óptica é considerada uma tecnologia chave para as arquiteturas de redes móveis 5G e B5G, como CRAN e CF-RAN, por sua capacidade em atender os altos requisitos de largura de banda e de latência [Chadha 2019, Mukherjee 2006]. Como resultado dessas capacidades apresentadas das redes ópticas, os pesquisadores abordam e utilizam redes ópticas como forma de solucionar vários dos principais problemas do *fronthaul*, como os alto requisitos de largura de banda de aplicação. Considerando o *fronthaul* óptico, há várias soluções propostas pela literatura para solucionar problemas de dimensionamento de rede, de particionamento de processamento de banda base e de arquiteturas híbridas. Ademais, as pesquisas buscam a otimalidade da solução.

Alguns problemas identificados na literatura correspondem à dificuldade de provisionamento de recursos e dimensionamento da rede. Em suma, esse problema é apresentado como forma eficiente de promover o menor desperdício de recursos ao realizar o dimensionamento corretamente.

2.1.1. O problema de dimensionamento de redes

Uma forma de auxilia na modelagem e projeto de redes mais eficientes, a aplicação de PLI fornecerá recursos suficientes para operação com a máxima eficiência energética. A PLI pode ser aplicada em vários áreas de estudos, sendo amplamente utilizada em matemática, transporte, roteamento e pesquisas em redes e otimização em geral. Exemplo disso está as pesquisas de [Figueiredo et al. 2016, Mohammed Mikaeil et al. 2019], no qual os autores usam PLI no problema de dimensionamento de rede e alocação de recursos em arquiteturas C-RAN. Em suma, os autores a aplicam para prover gerenciamento de recursos ótimo e obter o melhor provisionamento no *pool* de BBU objetivando maximizar a cobertura enquanto minimizam as restrições ópticas no *fronthaul*. Esta mesma abordagem pode ser aplicada em arquiteturas híbridas, como ocorre em [Tinini et al. 2019, Nassar and Yilmaz 2019]. Nessas pesquisas, os autores fazem uso da PLI em uma arquitetura baseada em *fog* para alocação de recursos visando resolver problemas de latência e de largura de banda.

Diversos outros trabalhos apresentam diferentes metodologias para o gerenciamento de recursos em redes 5G e B5G categorizadas em gerenciamento de recursos computacionais e de rádio. O trabalho de [Rodoshi et al. 2020] discute e apresenta a avaliação de métricas de desempenho e técnicas de validação centradas no gerenciamento de recursos de redes 5G. Além disso, os autores discutem alguns dos principais desafios e questões de pesquisa em aberto com intuito de fornecer orientação de pesquisa futura. Ainda relacionado ao gerenciamento e dimensionamento de rede, os autores em [Aqeeli et al. 2018] afirmam que a alocação de recursos computacionais de unidades de banda base da RU e do *pool* de BBU para pequenas células densamente implantadas incorrem em desperdício de recursos e aumento no gasto com energia. Dessa forma, os autores afirmam que a melhor solução para esse problema é a utilização de otimizadores que garantam otimalidade. Assim, os autores propuseram o uso de formulações matemáticas de PLI usando modelo de decomposição formulando dois subproblemas; um para cada nível da rede. O trabalho em [dos Santos et al. 2022] apresenta uma abordagem híbrida que trata de prever antecipadamente a demanda de rede das próximas horas (*multi step time series forecasting*) para dar ao PLI mais tempo para resolver o problema. Porém, os resultados mostram que quanto maior for o salto da previsão temporal, maior é o erro. Isso implica que a

solução proposta promove um dimensionamento ótimo para uma demanda de rede que não corresponde ao real, implicando em bloqueio ou desperdício de recursos.

Apesar da otimalidade garantida com o PLI, a sua aplicação está limitada para redes de pequeno porte por não ser uma técnica escalável. Assim, pesquisadores utilizam outras abordagens para redes maiores, o que será discutido nas subseções a seguir.

2.1.2. Dimensionamento de rede utilizando heurísticas

Muitas pesquisas relacionadas ao problema de escalabilidade de PLI para solucionar problemas operacionais em redes 5G/B5G utilizam técnicas alternativas por meio de relaxação linear, meta heurísticas e aprendizado de máquina. Cada uma das abordagens permite a aplicação em redes grandes, diferente da PLI. A literatura apresenta diversos trabalhos que tratam do tempo de execução e da escalabilidade da PLI utilizando a relaxação linear. As soluções propostas aplicam diversos tipos de abordagens de relaxação para obter valores próximos ao ótimo em uma infinidade de aplicação de redes e de logística [Noor-E-Alam and Doucette 2012, Zaky Kasem et al. 2012, Baruah et al. 2019, dos Santos et al. 2022]. Essas soluções mostram a busca por alternativas ou outras abordagens para resolver o problema de escalabilidade do PLI sem grande disparidade em comparação ao resultado ótimo.

Relacionada à relaxação da PLI, os autores em [Tang et al. 2017] investigam a colocação de *cache* conjunta e alocação de subportadora para dois casos de uso de *slice* de rede com atenção ao atraso e aos dados em uma arquitetura *Fog*-RAN. Devido à baixa escalabilidade e não convexidade do PLI, os autores apresentaram um método interativo baseado em relaxamento e convexificação para resolver o problema mencionado. Os resultados das simulações mostram que houve convergência e uma resposta rápida que alcançou uma solução quase ótima.

Com relação ao uso de outras heurísticas, os autores em [Gao et al. 2019] apresentam uma política de aprendizado por reforço profundo para roteamento e alocação de BBU em uma arquitetura C-RAN para aumentar a utilização de recursos. Os autores em [Gkatzios et al. 2019] introduziram conceitos de dessegregação de recursos de rede em RAN virtualizado centralizado. Eles propuseram uma alocação individual de funções de processamento para vários servidores de acordo com a carga de tráfego da rede. Finalmente, os autores em [Mo et al. 2018] implementaram uma rede de memória longa de curto prazo (LSTM) para prever as demandas de tráfego de rede de *pools* de BBU em redes C-RAN ROADM. O artigo se concentra na realocação de recursos com antecedência para prever os requisitos futuros de recursos da rede.

Diferentemente dos trabalhos anteriores, nós apresentamos um abordagem escalável que utiliza relaxação linear e particionamento flexível de processamento de banda base para dimensionar a rede *on-demand*.

3. Arquitetura CF-RAN

A Figura 1apresenta a arquitetura CF-RAN. Ela é composta por nós de computação em nuvem e *fog*. O nó de nuvem centraliza o processamento de banda base para economizar o consumo de energia. No entanto, a centralização do tráfego pode exceder a capacidade do *fronthaul* de transmitir tráfego ou pode experimentar uma sobrecarg na infraestrutura

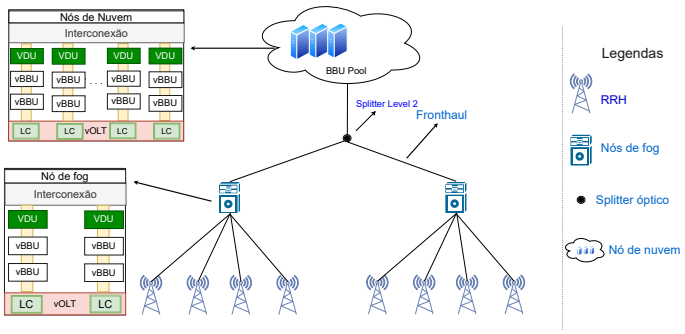


Figura 1. Arquitetura CF-RAN

elétrica. Portanto, nesse caso, os nós de *fog* devem ser ativados para receber a carga de trabalho excedida da nuvem. É importante destacar que na CF-RAN o consumo de energia cresce enquanto a cobertura da rede aumenta à medida que nós de *fog* são ativados para receber vBBUs.

Seja em nós de nuvem ou de *fog*, um conjunto de Unidades Digitais Virtuais (do inglês *Virtualized Digital Unit* – VDUs), ou seja, contêineres de processamento de banda base, implementa um conjunto de Funções de Processamento Virtualizadas (do inglês *Virtualized Processing Functions* – VPFs). Um desses VPFs é o vBBU, que recebe e processa os sinais de banda base de um determinado RU. Assim, para cada RU transmitindo tráfego, uma VDU e uma vBBU precisam ser ativadas em algum nó de processamento para realizar o processamento de banda base para o RU.

O *fronthaul* do 4G LTE utilizando o protocolo de interface de rádio pública comum (do inglês *Common Public Radio Interface*–CPRI) requer 2.5 Gb/s de velocidade de acesso óptico em esquema MIMO *2x2*. Já para a rede 5G, exige-se taxa de dados de pico de até 20 Gb/s, o que leva a velocidade de acesso óptico acima de 25 Gb/s por comprimento de onda, mesmo que o particionamento funcional seja empregado no *fronthaul* [Zhang and Jia 2022]. A 3GPP (*3rd Generation Partnership Project*), ITU-T (do inglês *International Telecommunications Union - Telecommunication Standardization Sector*) e a IEEE (do inglês *Institute of Electrical and Electronics Engineers*) vem apresentando relatórios e levantamentos técnicos (*releases*) especificando como a PON pode suportar *fronthaul* 5G e como os elementos de rede de rádio (CU, DU e RU) podem ser mapeados para os elementos de transporte [Consortium et al. 2018, IEEE 2019, ITU 2018].

Para dar esse suporte, o *fronthaul* CF-RAN opera em um TWDM-PON (do inglês *Time wavelength division multiplexing PON*) que fornece baixo consumo de energia e baixa latência [Tinini et al. 2020]. Na arquitetura existe uma Unidade de Rede Óptica (ONU) que conecta cada RU, e cada nó de processamento implementa um Terminal de Linha Óptica (do inglês *Optical Line Terminal* – OLT). A OLT implementa um conjunto de *Line Cards* (LC), além de permitir comprimentos de onda para ONUs. Os VDUs conectam cada LC e o tráfego encaminhado de um LC para seu VDU relacionado. Os VBBUs de cada VDU se comunicam internamente mediante um *switch* interno para alternar o tráfego quando necessário. Assumimos o uso de canais PON virtualizados (VPON), pois cada ONU pode sintonizar em qualquer um dos comprimentos de onda disponíveis, de modo que várias ONUs podem compartilhar o mesmo canal óptico de forma TDM (*Time*

Division Multiplexing) para transmitir para um nó de processamento comum. Os VPONs também podem ser ativados dinamicamente para suportar demandas de rede específicas nesse sentido.

Em relação à operação CF-RAN, para um determinado conjunto de RUs ativos, um número adequado de vBBUs deve ser colocado na rede, e VPONs precisam ser alocados para promover transmissões de RUs para os vBBUs colocados. Uma decisão importante é definir onde colocar vBBUs para promover a eficiência energética, seja na nuvem ou em nós de névoa. Na próxima seção, apresentamos um relaxamento LP de uma formulação PLI para resolver o problema de alocação de vBBU e atribuição de VPON em CF-RAN.

4. Particionamento das Funções de Banda-Base e Dimensionamento de Rede

O particionamento do processamento de banda base permite que funções da RAN sejam separadas e ainda mantenham as conexões entre as partes, a saber: RU, DU (do inglês *Distributed Unit*) e CU (do inglês *Centralized Unit*). Com essa flexibilidade, uma questão chave é onde colocar as funções de rede dada uma premissa de eficiência energética.

Com relação ao desempenho da rede *fronthaul*, a escolha de implantação pode fazer uso de rede comutada por circuito ou de rede comutada por pacote, impactando diretamente no uso dos recursos de rede. Por exemplo, redes comutadas por circuito, como as Redes Óptica de Transporte (do inglês *Optical Transport Network* – OTN), possuem atribuição estática e com uso constante de recursos. Isso significa que problemas como filas serão muito raros, uma vez que os recursos são reservados antecipadamente e há o fornecimento de uma conexão estável.

Com relação ao dimensionamento do CF-RAN, temos que vBBUs podem ser instanciadas na nuvem ou na *fog* para acomodar processamento de banda base. Outro ponto são as VPONs que podem ser instanciadas para os nós de processamento, sendo exclusiva de cada uma delas. Outro ponto que deve-se considerar é o total de instâncias de processamento que deverão ser ativas para acomodar corretamente a rede. Assim, o dimensionamento na CF-RAN considera a ativação, alocação e instanciação dos elementos de processamento corretamente, objetivando a redução do consumo energético e a minimização dos desperdícios de recursos computacionais.

5. Formulação PLI considerada para a relaxação

A seguir apresentamos a formulação do PLI considerado que fornece a solução ótima para o problema estudado [Santos et al. 2021a]. A formulação do PLI proposto realiza o dimensionamento do fronthaul objetivando a minimização do consumo de energia. As notações utilizadas na formulação do problema são apresentadas na Tabela 1.

Função Objetivo

$$Min \sum_{n=1}^{N} x_n.C_n + C_{lc} \sum_{w=1}^{W} \sum_{n=1}^{N} z_{wn} + C_{split} \sum_{i=1}^{I} \sum_{s=1}^{S} \delta_{is} \tag{1}$$

A função objetivo (1) minimiza a ativação geral dos elementos de processamento do CF-RAN, como nós de processamento, VDUs, vBBUs, Line Cards, colocando ao máximo as demandas de RUs em um único VPONs para processamento ao máximo na

Tabela 1. Notação utilizada para o modelo

Símbolos	Definições
Conjuntos	
$i \in R$	conjuntos de demandas de tráfego da RU
$n \in N$	conjunto de nós de processamento
$w \in W$	conjuntos de VPONs disponíveis
$s \in S$	conjunto de divisões funcionais disponibilizados
Parâmetros	
B_i^s	demanda de largura de banda da RU i de acordo com a divisão funcional s
B_w	capacidade do VPON w
P_i^s	demanda de processamento de demandas do RU i de acordo com a divisão funcional s
P_n	capacidade de processamento do nó n
M	número muito grande
C_n	custo energético do nó n
C_{lc}	custo energético da *LineCard* LC
C_{split}	valor atribuido a cada divisão funcional s
Variáveis	
y_{wns}^i	1 se a demanda i é processada no nó n delimitada por s sendo transmitida no VPON w, 0 caso contrário.
z_{wn}	1 se o VPON w está alocado no nó n, 0 caso contrário.
x_n	1 se funções de processamento e elementos do nó n estão ativados, 0 caso contrário.
d_{in}	1 se as demandas de i estão alocadas no nó n, 0 caso contrário.
δ_{is}	1 se as demandas de i estão associadas com a opção de divisão s, 0 caso contrário.
t_{isn}	1 se as demandas de i tem sua opção de divisão s sendo processada no nó n, 0 caso contrário.

nuvem. Nossa formulação utiliza valores decrescentes nos custos da opção de divisão funcional para promover a escolha da máxima centralização na nuvem para melhor eficiência energética. A descentralização das funções de banda base ocorrerá apenas quando a nuvem saturar a sua capacidade de processamento.

Restrições

$$\sum_{s=1}^{S} \delta_{is} = 1, \forall i \in R \tag{2}$$

A restrição (2) garante que cada RU ativará apenas uma das opções das divisões funcionais disponíveis. Esta restrição está associada à escolha flexível para cada RU.

$$\sum_{n=1}^{N} d_{in} = 2, \forall i \in R \tag{3}$$

$$\sum_{w=1}^{W} \sum_{n=1}^{N} \sum_{s=1}^{S} y_{wns}^i = 2, \forall i \in R \tag{4}$$

Restrições (3) e (4) garantem que dois VPONs serão criados para transmitir a demanda i para diferentes nós de processamento n dada a opção de divisão decidida pela formulação, sendo uma direcionada para o nós de fog e outra para a nuvem.

$$\sum_{n=1}^{N} z_{wn} \leq 1, \forall w \in W \tag{5}$$

A restrição (5) garante que um VPON seja alocado para no máximo um nó de processamento.

$$\sum_{w=1}^{W} \sum_{n=1}^{N} y_{wns}^i \leq \delta_{is}, \forall i \in R, \forall s \in S \tag{6}$$

$$\sum_{n=1}^{N} t_{isn} \leq \delta_{is}, \forall i \in R, \forall s \in S \tag{7}$$

$$\sum_{w=1}^{W} y_{wns}^{i} \leq t_{isn}, \forall i \in R, \forall s \in S, \forall n \in N \tag{8}$$

Restrições (6), (7), (8) garantem que a demanda i seja mapeada usando a restrição (7) e, assim, ocorre a divisão correta do processamento entre nuvem e fog e a correta alocação da demanda para seu respectivo nó de processamento.

$$\sum_{i=1}^{R} \sum_{n=1}^{N} \sum_{s=1}^{S} (y_{wns}^{i} \times B_{i}^{s}) \leq B_{w}, \forall w \in W \tag{9}$$

$$\sum_{i=1}^{R} \sum_{w=1}^{W} \sum_{s=1}^{S} (y_{wns}^{i} \times P_{i}^{s}) \leq P_{n}, \forall n \in N \tag{10}$$

Restrições (9), (10) garantem que a capacidade total dos VPONs e dos nós de processamento sejam respeitados para o total de demanda vigente, o que torna possível o envio de demandas apenas se houver capacidade de processamento para isso.

$$M \times x_{n} \geq \sum_{i=1}^{R} \sum_{w=1}^{W} \sum_{s=1}^{S} y_{wns}^{i}, \forall n \in N \tag{11}$$

$$x_{n} \leq \sum_{i=1}^{R} \sum_{w=1}^{W} \sum_{s=1}^{S} y_{wns}^{i}, \forall n \in N \tag{12}$$

$$M \times z_{wn} \geq \sum_{i=1}^{R} \sum_{s=1}^{S} y_{wns}^{i}, \forall w \in W, \forall n \in N \tag{13}$$

$$z_{wn} \leq \sum_{i=1}^{R} \sum_{s=1}^{S} y_{wns}^{i}, \forall n \in N, \forall w \in W \tag{14}$$

Restrições (11), (12), (13), (14) executam a ativação dos nós de processamento e VPON para reagir corretamente às demandas vigentes e, assim, executar a correta atribuição destes.

$$M \times d_{in} \geq \sum_{w=1}^{W} \sum_{s=1}^{S} y_{wns}^{i}, \forall n \in N, \forall i \in R \tag{15}$$

$$d_{in} \leq \sum_{w=1}^{W} \sum_{s=1}^{S} y_{wns}^{i}, \forall n \in N, \forall i \in R \tag{16}$$

Restrições (15), (16) garantem a ativação de nós de processamento e outros elementos de rede para decisões de redirecionamento de tráfego excedente dado escolhas realizadas na opção de divisão funcional e na largura de banda disponível. Esta restrição garante a ativação de novos elementos de processamento quando não houver recursos suficientes disponíveis relacionados às restrições (9) e (10).

5.1. Relaxação da PLI

A relaxação linear proposta segue os mesmo passos de [Santos et al. 2021b]. A proposta consiste em retirar a integralidade obrigatória de algumas variáveis e realizar o arredondamento das variáveis fracionárias quando o *solver* retornar uma solução. A viabilidade da solução é testada mediante um algoritmo que verifica se todas as restrições foram respeitadas. Assim, resolver o problema de colocação de vBBU e atribuição de VPONs em CF-RAN de forma escalável.

Após extensa análise, nós consideramos as seguintes variáveis para serem relaxadas:

$$0 \leq y^{i}_{wns} \leq 1, \forall i, n, w, s \in R, N, W, S \tag{17}$$

$$0 \leq k_{wns} \leq 1, \forall w, n, s \in W, N, S \tag{18}$$

$$0 \leq t_{isn} \leq 1, \forall i, s, n \in R, S, N \tag{19}$$

$$s_{is}\{0\} \cup [1, \infty], \forall i, s \in R, S \tag{20}$$

$$y_{in}\{0\} \cup [1, \infty], \forall i, n \in R, N \tag{21}$$

A análise extensa realizada para identificar as variáveis suscetíveis a quebra de integralidade apresentou a seguinte configuração: as variáveis y^{i}_{wns}, k_{wns} e t_{isn} de binária para contínua e y_{in} e s_{is} de binário para semi-contínua. Como mencionado anteriormente, a solução ótima para a relaxação linear não é necessariamente uma solução integral. No entanto, como a região viável da relaxação é maior que a região viável da PLI, o valor ótimo da primeira não é pior que o valor ótimo da segunda.

6. Exemplos Numéricos

Tabela 2. Parâmetros de Simulação

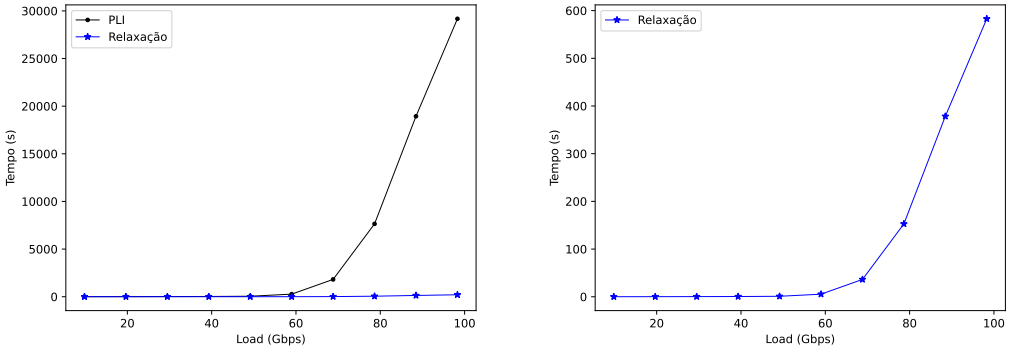
Parâmetros	**Valor**
Topologia	1 cloud, 3 fog
Configuração do RU	10 MHz, 1x1 MIMO
Comunicação entre RU-vBBU	eCPRI
Custo da *Cloud*	600 watts
Custo da *Fog*	300 watts
Line Card	20 watts
Custo por Cloud VDU	100 watts
Custo por Fog VDU	50 watts

Nesta seção, são apresentados os resultados na avaliação de dois tipos diferentes de cenários de tráfego. O primeiro cenário é o estático, que nos permite avaliar se a relaxação escala em comparação ao PLI. O segundo cenário é o dinâmico, que permite verificar métricas de redes, como probabilidade de bloqueio, consumo de energia e disponibilidade dos serviços de rede. Os parâmetros de simulação usados para simulações são apresentados em Tabela 2.

6.1. Cenário de Tráfego Estático

Realizou-se testes em simulação com demandas de tráfego crescentes para avaliar a escalabilidade da solução em termos do tempo de execução. Para os testes, foi utilizado a *API DOCPLEX Python* com *CPLEX* 12.10 com configurações padrão.

Os resultados da Figura 2 mostram o tempo de execução entre a relaxação linear e a PLI. Observa-se na Figura 2 (a) que a relaxação apresenta comportamento linear se comparado a PLI. A PLI apresentou comportamento com crescimento exponencial no tempo de execução à medida que mais demandas foram geradas para se fazer o dimensionamento e o particionamento de funções. Os resultados da PLI reforçam o já conhecido problema de escalabilidade da solução, tornando-a inviável para soluções *on-demand*. Em contra partida, a Figura 2 (b) mostra o tempo de execução individual da relaxação linear. Observa-se que o comportamento de crescimento também mostrou-se exponencial, mas com tempo de execução muito inferior ao PLI.



(a) Tempo de execução PLI e Relaxação

(b) Tempo de execução da Relaxação

Figura 2. Comparativo entre o tempo de execução da PLI e da Relaxação

6.2. Cenário de Tráfego Dinâmico

Foi utilizado um simulador de eventos discretos, 5GPy [Tinini et al. 2020], para avaliar o desempenho da relaxação linear e a formulação PLI em um cenário de tráfego dinâmico usando a mesma topologia do cenário de tráfego estático, com carga máxima de tráfego (ε/160), onde ε é o *erlang* de cada hora do dia [Tinini et al. 2020]. As requisições de tráfego chegam seguindo um processo de *Poisson*, cuja média é igual ao ε da hora atual de operação, e têm um tempo de atendimento seguindo uma distribuição exponencial negativa. Esse comportamento de tráfego é baseada em cenários de rede de acesso comercial detalhado em [Peng et al. 2011]. Os resultados mostram valores médios adquiridos de 40 execuções com nível de confiança de 95%.

Os resultados apresentados na Figura 3 foram obtidos por meio da simulação da relaxação linear, observando a ativação de elementos de processamento de rede e o posicionamento das funções com base no particionamento funcional. Essa figura representa a métrica de corretude, que avalia se a relaxação linear conseguiu gerar resultados ideais em termos de consumo de energia, ativação de elementos para processamento e particionamento funcional para alocar todas as demandas. Mais especificamente, os resultados da Programação Linear Inteira (PLI) foram utilizados como base de comparação na avaliação dos resultados. Observa-se que, para demandas de até 100 Gbps, os resultados de corretude são estatisticamente iguais aos da PLI, mas à medida que a demanda aumenta, os valores de corretude também diminuem.

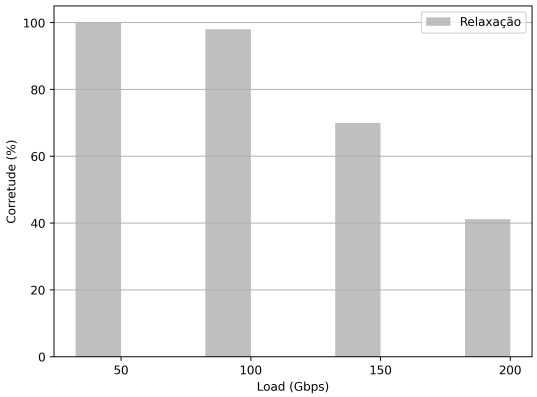
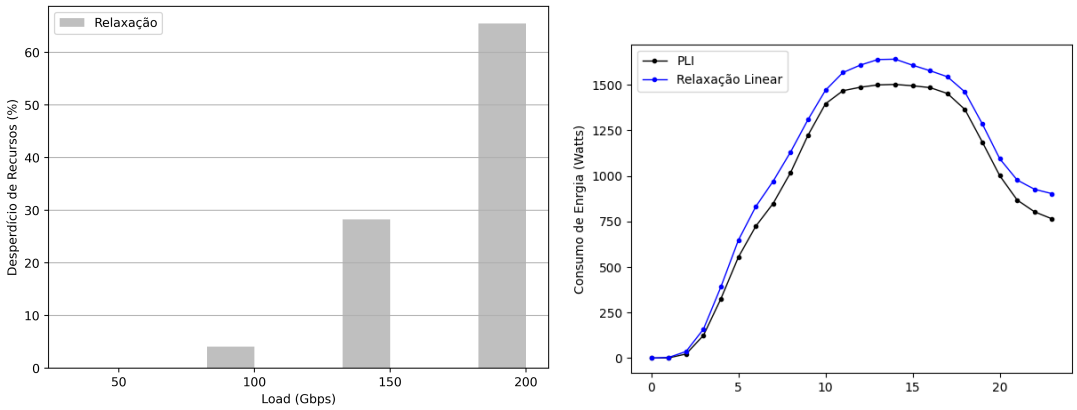


Figura 3. Corretude da ativação de nós de processamento e colocação das funções de banda base



(a) Desperdício de Recursos de Rede **(b)** Consumo de Energia

Figura 4. Resultados Ilustrativos da relaxação linear

A Figura 4 (a) apresenta o total de recursos desperdiçados pela técnica de relaxação linear. Os desperdícios estão relacionados à não alocação completa de demandas de VPONs e de nós de névoa, bem como à ativação desnecessária de elementos de processamento para as demandas geradas. Esse comportamento é semelhante ao observado anteriormente, onde, à medida que a demanda aumenta, a relaxação linear acaba acomodando ou ativando incorretamente os elementos de processamento de banda base. O desperdício de recursos não está relacionado apenas aos bloqueios, mas também ao subdimensionamento da rede, um comportamento observado na Fig. 4 (a). Observa-se que houve um consumo de energia maior pela relaxação linear, o que é esperado. Esse resultado reforça que houve um sobredimensionamento da rede e que, mesmo assim, não impediu que demandas fossem alocadas incorretamente. A Fig. 4 (b) está relacionada à nossa função objetivo. Observa-se que, em comparação com a programação linear inteira (PLI), a técnica de relaxação linear obteve um maior consumo de energia. Esse comportamento reforça o sobredimensionamento da rede. Neste caso, podemos afirmar que mais elementos de rede para o processamento de banda base foram ativados do que o ideal, que foi apresentado pela PLI..

7. Conclusão e Trabalhos Futuros

Neste trabalho, foi proposta uma relaxação linear escalável para resolver o problema de escalabilidade da PLI. A relaxação obteve resultados próximos ao PLI em redes moderadas, com uma redução significativa no tempo de execução. No entanto, em redes maiores, a relaxação apresentou perdas na otimalidade, indicando a necessidade de buscar outras estratégias, como heurísticas e aprendizado de máquina, para resolver o problema da escalabilidade. Para trabalhos futuros, será importante explorar outras técnicas para alcançar uma solução escalável sem sacrificar a qualidade ótima.

Referências

Aqeeli, E., Moubayed, A., and Shami, A. (2018). Power-aware optimized rrh to bbu allocation in c-ran. *IEEE Transactions on Wireless Communications*, 17(2):1311–1322.

Baruah, S. K., Bonifaci, V., Bruni, R., and Marchetti-Spaccamela, A. (2019). Ilp models for the allocation of recurrent workloads upon heterogeneous multiprocessors. *Journal of Scheduling*, 22(2):195–209.

Chadha, D. (2019). *Optical WDM Networks: From Static to Elastic Networks*. John Wiley & Sons.

Consortium, G. et al. (2018). Ng-ran; architecture description. Technical report, Technical Report TR-38.401 Release 16.

dos Santos, M. R., Tinini, R. I., Januario, T. O., and Figueiredo, G. B. (2022). Deep recurrent neural network for optical fronthaul dimensioning and proactive vbbu placement in cf-ran. *Photonic Network Communications*, 43(1):59–73.

Figueiredo, G. B., Wang, X., Meixner, C. C., Tornatore, M., and Mukherjee, B. (2016). Load balancing and latency reduction in multi-user comp over twdm-vpons. In *IEEE Intl. Conf. on Communications (ICC)*.

Gao, Z., Zhang, J., Yan, S., Xiao, Y., Simeonidou, D., and Ji, Y. (2019). Deep reinforcement learning for bbu placement and routing in c-ran. In *Optical Fiber Communications*.

Gkatzios, N., Anastasopoulos, M., Tzanakaki, A., and Simeonidou, D. (2019). Efficiency gains in 5g softwarised radio access networks. *EURASIP Journal on Wireless Communications and Networking*, 2019(1):183.

IEEE (2019). Ieee draft standard for packet-based fronthaul transport networks. *IEEE P1914.1/D5.0, April 2019*, pages 1–89.

ITU, T. S. S. O. (2018). Series g: Transmission systems and media, digital systems and networks: 5g wireless fronthaul requirements in a passive optical network context. Technical report, ITU-T G-series Recommendations – Supplement 66.

Larsen, L. M. P., Checko, A., and Christiansen, H. L. (2019). A survey of the functional splits proposed for 5g mobile crosshaul networks. *IEEE Communications Surveys Tutorials*, 21(1):146–172.

Mo, W., Gutterman, C. L., Li, Y., Zussman, G., and Kilper, D. C. (2018). Deep neural network based dynamic resource reallocation of bbu pools in 5g c-ran roadm networks. In *Optical Fiber Communications*.

Mohammed Mikaeil, A., Hu, W., and Li, L. (2019). Joint allocation of radio and fronthaul resources in multi-wavelength-enabled c-ran based on reinforcement learning. *Journal of Lightwave Technology*.

Mukherjee, B. (2006). *Optical WDM networks*. Springer Science & Business Media.

Nassar, A. and Yilmaz, Y. (2019). Reinforcement learning for adaptive resource allocation in fog ran for iot with heterogeneous latency requirements. *IEEE Access*, 7:128014–128025.

Noor-E-Alam, M. and Doucette, J. (2012). Relax-and-fix decomposition technique for solving large scale grid-based location problems. *Computers & Industrial Engineering*, 63(4):1062–1073.

Peng, C., Lee, S.-B., Lu, S., Luo, H., and Li, H. (2011). Traffic-driven power saving in operational 3g cellular networks. In *Proceedings of the 17th annual international conference on Mobile computing and networking*, pages 121–132.

Rodoshi, R. T., Kim, T., and Choi, W. (2020). Resource management in cloud radio access network: Conventional and new approaches. *Sensors*, 20(9):2708.

Santos, M., Tinini, R., Januario, T., and Figueiredo, G. (2021a). Dimensionamento ótimo de fronthaul óptico com divisão flexível de funções de processamento em cf-ran. In *Anais do XXXIX Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*, pages 210–223, Porto Alegre, RS, Brasil. SBC.

Santos, M., Tinini, R., Januario, T., and Figueiredo, G. (2021b). Posicionamento quase ideal de vbbu e atribuição de comprimento de onda em cloud fog ran. In *Anais do XXXIX Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*, pages 448–461, Porto Alegre, RS, Brasil. SBC.

Tang, L., Zhang, X., Xiang, H., Sun, Y., and Peng, M. (2017). Joint resource allocation and caching placement for network slicing in fog radio access networks. In *2017 IEEE 18th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, pages 1–6.

Tinini, R. I., Batista, D. M., Figueiredo, G. B., Tornatore, M., and Mukherjee, B. (2019). Low-latency and energy-efficient bbu placement and vpon formation in virtualized cloud-fog ran. *IEEE/OSA Journal of Optical Communications and Networking*, 11(4):B37–B48.

Tinini, R. I., dos Santos, M. R. P., Figueiredo, G. B., and Batista, D. M. (2020). 5GPy: A SimPy-based simulator for performance evaluations in 5G hybrid Cloud-Fog RAN architectures. *Simulation Modelling Practice and Theory*, 101:102030.

Wu, J., Zhang, Z., Hong, Y., and Wen, Y. (2015). Cloud radio access network (c-ran): a primer. *IEEE Network*, 29(1):35–41.

Zaky Kasem, A., Doucette, J., et al. (2012). Ilp model and relaxation-based decomposition approach for incremental topology optimization in p-cycle networks. *Journal of Computer Networks and Communications*, 2012.

Zhang, J. and Jia, Z. (2022). Coherent passive optical networks for 100g/λ-and-beyond fiber access: Recent progress and outlook. *IEEE Network*, 36(2):116–123.