Alocação Eficiente de vBBUs e VPONs em uma Arquitetura Cloud-Fog RAN Virtualizada sobre TWDM-PON

Rodrigo Izidoro Tinini[1], Matias R. P. dos Santos[3], Carlos Kamienski[2], Gustavo Bittencourt Figueiredo[3], Daniel Macêdo Batista[1]

[1]Instituto de Matemática e Estatística - Universidade de São Paulo

[2]Centro de Matemática, Computação e Cognição - Universidade Federal do ABC

[3]Departamento de Ciência da Computação – Universidade Federal da Bahia

{rtinini, batista}@ime.usp.br, matiasrps@ufba.br,

cak@ufabc.edu.br, gustavo@dcc.ufba.br

Abstract. *To promote power efficiency and low latency in 5G networks, operators are adopting Cloud Radio Access Networks (CRAN) with a TWDM-PON fronthaul. Although reducing power consumption, CRAN imposes heavy loads on the fronthaul, which could increase latency and reduce network coverage. So, in this paper we consider an architecture called Cloud-Fog RAN (CF-RAN) that uses fog nodes to alleviate the demands on the fronthaul. We propose a graph-based heuristics to plan a dynamic activation of the fog nodes and to dimension the bandwidth on the TWDM-POM fronthaul. Experiments show that our proposal benchmarks an ILP model, mitigates blocking probability, and increases power-efficiency in 58% in comparison to baseline heuristics.*

Resumo. *Para diminuir a latência e o consumo de energia em redes 5G, operadores têm adotado a arquitetura Cloud Radio Access Networks (CRAN) com um fronthaul TWDM-PON. Apesar de reduzir o consumo de energia, a carga de dados no fronthaul da CRAN é alta, podendo aumentar as latências e diminuir a cobertura da rede. Assim, neste artigo, consideramos a arquitetura Cloud-Fog RAN (CF-RAN), que usa fog nodes para diminuir a carga do fronthaul. Para esta arquitetura, nós propomos uma heurística baseada em grafos que dinamicamente ativa fog nodes e dimensiona a banda do fronthaul. A proposta provê soluções ótimas de um ILP, é capaz de mitigar a probabilidade de bloqueio e reduz o consumo de energia em até 58% em comparação com heurísticas base.*

1. Introdução

A Rede de Acesso a Rádio baseada em Computação em Nuvem (Cloud Radio Access Networks - CRAN) é uma recente arquitetura de rede considerada por operadores de telecomunicações para aumentar a capacidade de redes móveis, ao mesmo tempo em que o consumo de energia é reduzido [Wu et al. 2015] em emergentes redes 5G. A arquitetura CRAN difere de RANs distribuídas (DRAN) ao desacoplar as Unidades de Processamento de Banda-Base (BaseBand Unit) das antenas das células da rede e centralizá-las em um *pool* de BBUs em um nó de nuvem. Desta forma, apenas antenas de baixo consumo energético, chamadas de Remote Radio-Head (RRH), são deixadas nas células da rede e uma única infraestrutura é utilizada para realizar o processamento dos sinais de

banda base gerados pelos RRHs, reduzindo os custos provenientes de infraestruturas de refrigeração e alimentação energética. Além disso, a CRAN implementa uma rede de transporte, tipicamente óptica, para realizar a transmissão de sinais de banda-base entre os RRHs e a nuvem, i.e., o fronthaul [Chanclou et al. 2013].

Apesar da arquitetura CRAN proporcionar reduções no consumo energético, ela também impõe altas demandas por largura de banda e rígidas restrições de latência no fronthaul óptico. Os sinais de banda-base gerados pelos RRHs são transmitidos por meio do protocolo Common Public Radio Interface (CPRI) [de la Oliva et al. 2016], que impõe diferentes linhas de transmissão que variam de $614,4$Mbps até $24,3$Gbps dependendo da configuração de Múltiplas Entradas-Saídas (Multiple Input-Output (MIMO)) de um RRH. Além disso, uma latência máxima de 3μs é permitida entre os RRHs e a pool de BBUs em decorrência do protocolo Hybrid Automatic Repeat Request (HARQ) [Chitimalla et al. 2017]. Em cenários de ultra densidade de RRHs, o fronthaul tenderá a se tornar ainda mais congestionado, o que poderá levar ao bloqueio de transmissões CPRI e enfileiramento de processamento de banda-base na pool de BBUs, aumentando ainda mais a latência do fronthaul e do processamento de banda-base [Figueiredo et al. 2016].

A fim de lidar com os rigorosos requisitos do protocolo CPRI e aliviar o tráfego no fronthaul, foi proposta uma arquitetura chamada Cloud-Fog RAN (CF-RAN) [Tinini et al. 2019]. CF-RAN é uma arquitetura híbrida que se baseia em Computação em Névoa (Fog Computing) [Bonomi et al. 2012] para estender a capacidade de processamento de banda-base da nuvem em nós mais próximos das células da rede, chamados de fog nodes. Se o fronthaul ou a nuvem tornam-se congestionados, os fog nodes podem ser utilizados para realizar o processamento de banda-base de requisições que não poderiam ser atendidas pela nuvem. Entretanto, ao utilizar-se de fog nodes, o custo de operação da rede aumentará. Desta forma, a CF-RAN implementa a Virtualização de Funções de Rede (Network Functions Virtualization (NFV)) [Hawilo et al. 2014] para que o processamento nas BBUs possa ser virtualizado (virtualized BBU (vBBU)) em máquinas virtuais. Assim, vBBUs e fog nodes podem ser dinamicamente ativados ou desativados em função da demanda de tráfego da rede. Além disso, o fronthaul da CF-RAN é implementado sobre uma Rede Óptica Passiva Multiplexada por Divisão de Comprimento de Onda e Tempo (Time-and-Wavelength Division Multiplexed Passive Optical Network (TWDM-PON)) [Luo et al. 2013]. Com a TWDM-PON, redes ópticas passivas dedicadas podem ser dinamicamente criadas por meio da virtualização dos canais ópticos (virtualized PON (VPON))[Wang et al. 2016] e compartilhadas por um grupo de RRHs para transmitirem, através de um comprimento de onda dedicado, a um nó de processamento.

A operação eficiente da arquitetura CF-RAN traz alguns desafios. Como o processamento de banda-base é realizado por meio da instanciação dinâmica de vBBUs para os RRHs ativos de uma rede, o operador deve decidir quantas vBBUs devem ser instanciadas e quando ativar os fog nodes para receber vBBUs, de forma a aliviar a demanda no fronthaul em função da demanda de tráfego. Além disso, enquanto na CRAN todos os comprimentos de onda da TWDM-PON são utilizados para transmitir apenas para o pool de BBUs na nuvem, na CF-RAN os comprimentos de onda devem ser dimensionados de forma que VPONs sejam criados para suportar transmissões para vBBUs alocadas tanto

na nuvem quanto nos fog nodes. Devido às restrições de colisão [Tinini et al. 2019], cada VPON só pode ser utilizado para transmitir a um único nó de processamento por vez, o que impede que dois nós de processamento diferentes partilhem a banda de um mesmo comprimento de onda.

Neste artigo, nós propomos um algoritmo de escalonamento para dinamicamente ativar vBBUs e decidir quando os fog nodes devem ser ativados, ao mesmo tempo em que os comprimentos de onda disponíveis são dimensionados, a fim de que VPONs sejam criados e transmissões possam ser realizadas para os fog nodes recentemente ativados. O algoritmo proposto é implementado sobre uma heurística baseada em um modelo de grafos. Resultados de simulações mostram que a alocação e dimensionamento eficiente dos recursos de processamento e de rede proporcionam uma operação energeticamente eficiente, ao mesmo tempo em que a probabilidade de bloqueio é reduzida. Por exemplo, observou-se que o algoritmo é capaz de prover as mesmas soluções ótimas fornecidas por um Problema de Programação Linear Inteira (Integer Linear Programming (ILP)); com a vantagem de um menor tempo de execução, é também capaz de mitigar a probabilidade de bloqueio e reduzir o consumo de energia em até 58% em comparação com heurísticas base.

O restante desse artigo está organizado da seguinte forma: na Seção 2, o estado da arte é apresentado; a Seção 3 apresenta a arquitetura proposta; o problema de ativação dos fog nodes e dimensionamento dos comprimentos de onda é apresentado na Seção 4; na Seção 5, apresentamos o modelo em grafo e a heurística proposta; na seção 6, um modelo de ILP usado para comparações e os detalhes das simulações executadas são apresentados; na Seção 7, os resultados são discutidos; e o artigo é concluído na Seção 8.

2. Trabalhos Relacionados

O uso conjunto de computação em nuvem e em névoa é um tema de pesquisa emergente para a concepção das futuras redes móveis. Muitos trabalhos focaram no desenvolvimento do fronthaul e em algoritmos de escalonamento de recursos em tais redes.

A proposta de uma camada de computação em névoa em redes móveis foi introduzida em [Peng et al. 2014]. Neste trabalho, uma CRAN Heterogênea (Heterogeneous CRAN (H-CRAN)) foi proposta para aumentar a eficiência espectral e energética de redes móveis. HetNets foram consideradas para hospedar aplicações e estender a capacidade da nuvem. Entretanto, os autores atestam que a operação do fronthaul tem um impacto crucial no desempenho da rede. Por exemplo, um fronthaul congestionado pode deteriorar o desempenho de técnicas de mitigação de interferência como algoritmos de transmissão coordenada entre multipontos (Coordinated Multi-Point (CoMP)). Assim, além da implementação de um fronthaul espectralmente eficiente, um escalonamento eficiente de seus canais de transmissão também deve ser realizado para que o desempenho do fronthaul seja maior. Os autores de [Iida et al. 2013] propuseram o uso de uma rede TWDM-PON para a implantação de um fronthaul flexível em CRANs. Por meio da configuração dinâmica de comprimentos de onda nas Optical Network Units (ONU) da rede, VPONs podem ser usados para agregar a transmissão de múltiplos RRHs a um único canal óptico. Comparado com uma rede Time Division Multiplexing PON (TDM-PON), a utilização da banda total disponível foi otimizada.

Em [Wang et al. 2016], os autores propuseram a configuração dinâmica de

VPONs para uma completa virtualização das estações base no pool de BBUs da CRAN. Uma formulação de ILP e heurísticas foram propostas para realizar a configuração dos VPONs e gerenciar os recursos de processamento da nuvem. Os resultados mostraram que o consumo de energia da RAN pode ser diminuído quando a configuração dinâmica de VPONs é utilizada. Em [Wang et al. 2017], os autores exploraram o uso de uma camada de fog computing para receber processamentos de banda-base quando o fronthaul não possui capacidade suficiente. Entretanto, tanto a rede de transporte utilizada na camada de fog computing quanto o escalonamento de seus canais de transmissão não foram propostos e foram deixados como um problema em aberto.

Em relação a algoritmos de escalonamento de recursos, a capacidade de cachê cooperativo em uma RAN baseada em fog (F-RAN) para a implementação de *offloading* de conteúdo foi explorada pelos autores em [Cui et al. 2018]. Um algoritmo baseado em grafos foi proposto para identificar possíveis *clusters* de fog nodes que poderiam cooperar entre si para realizar o *offload* de conteúdo. Essa proposta permitiu o aumento do *offloading*, mas com uma complexidade de tempo menor que um algoritmo ótimo de força bruta. Os autores em [Wang et al. 2018] propuseram o escalonamento de tarefas na CRAN considerando requisitos de tempo. Nesse trabalho, os autores consideraram a capacidade do fronthaul como um fator limitante para o escalonamento das tarefas na nuvem e modelaram sua capacidade máxima em função do consumo de energia de cada fluxo de dados transmitido por ele.

Apesar de alguns dos trabalhos supracitados apresentarem o escalonamento de VPONs em uma rede CRAN, nenhum desses trabalhos considerou como os comprimentos de onda de uma rede TWDM-PON devem ser dimensionados para permitirem transmissões a nós de nuvem e fog nodes em uma arquitetura híbrida como a CF-RAN. Além disso, os trabalhos que propuseram algoritmos de escalonamento de recursos para F-RAN negligenciaram a rede de transporte dos fog nodes e como seus canais de transmissão devem ser escalonados.

Em [Tinini et al. 2019], o uso de grafos foi proposto para ativar vBBUs e criar VPONs dinamicamente na CF-RAN. Nessa proposta, a ativação de vBBUs e criação de VPONs nos fog nodes da rede baseava-se em priorizar a ativação de vBBUs e criação de VPONs primeiramente nos fog nodes com maior carga de trabalho ou nos fog nodes com menor carga de trabalho. Assim, não era possível ativar os recursos de uma maneira mais uniforme entre os fog nodes. Assim, neste trabalho, nossa nova proposta busca propor a ativação dinâmica de vBBUs e criação de VPONs de uma forma mais balanceada em relação à carga de todos os fog nodes da rede.

3. Arquitetura CF-RAN Virtualizada

A arquitetura CF-RAN é composta de um nó de nuvem e fog nodes responsáveis por implementar o processamento de banda-base virtualizado em vBBUs. Cada nó de processamento possui um servidor dedicado, onde um conjunto de Virtual Digital Units (VDUs) é implementado. Um VDU é um *container* em que um conjunto de vBBUs podem ser instanciadas para realizarem o processamento de banda-base dos RRHs. Utilizando-se de NFV, cada VDU pode ser instanciado apenas quando um RRH solicita por uma vBBU; de outra forma, ele permanece desativado para economizar energia.

Em relação ao fronthaul, os RRHs são conectados aos fog nodes e à nuvem através

de uma topologia TWDM-PON em árvore (Figura 1a)). Cada RRH é conectado a uma ONU, que pode ser configurada para transmitir em qualquer comprimento de onda da rede. Cada ONU é conectada a um fog node através de uma fibra dedicada. Em cada nó de processamento, há um Optical Line Terminal (OLT), que é responsável por configurar um comprimento de onda em ONUs que desejam transmitir a seu nó de processamento por meio de uma VPON, além de receber os sinais transmitidos em diferentes VPONs em um conjunto de Line Cards (LC), i.e., *transceivers* presentes na OLT. Cada LC recebe o tráfego de um comprimento de onda específico e o encaminha para uma VDU diretamente associada a si. A Figura 1b) ilustra um exemplo de VDUs ativadas na nuvem e recebendo seus dados por LC exclusivos.

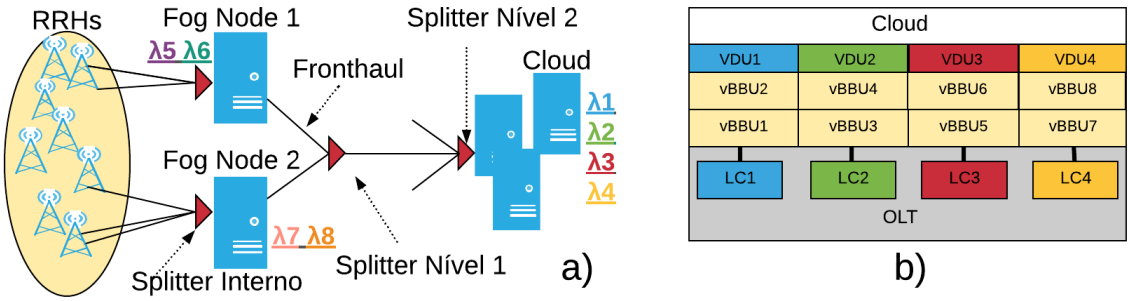

Figura 1. a) Visão geral da arquitetura CF-RAN; b) Detalhes internos do nó de processamento da nuvem

Diferentes níveis de multiplexação são implementados na rede por meio de *splitters* ópticos (Figura 1a)). No primeiro nível, as fibras dedicadas dos RRHs são multiplexadas rumo à nuvem em uma fibra de distribuição ou rumo aos LC de um fog node, através de um *splitter* interno no fog node. Um *splitter* nível 1 multiplexa múltiplas fibras de distribuição em uma fibra de alimentação. Finalmente, *splitter* nível 2 localizado na nuvem pode multiplexar múltiplas fibras de alimentação.

Como as VDUs, vBBUs e os próprios fog nodes são ativados em função da demanda de tráfego da rede. Sua ativação deve ser cuidadosamente planejada para que uma operação energeticamente eficiente possa ser mantida. Além disso, quando os fog nodes começam a ser ativados, a quantidade necessária de largura de banda do fronthaul destinada a transmissões locais deve ser dimensionada. Este dimensionamento é realizado ao alocar comprimentos de onda disponíveis aos OLTs de fog nodes ativos, para que VPONs sejam configuradas por eles a fim de receberem transmissões. Na próxima seção, o problema de ativação dos recursos de processamento e dimensionamento dos comprimentos de onda é apresentado.

4. Alocação de Recursos e Dimensionamento do Fronthaul Energeticamente Eficientes

Durante a operação de uma rede móvel, diferentes números de RRHs podem estar ativos em função da flutuação do tráfego [Peng et al. 2011]. Assim, o escalonamento ideal dos recursos, tanto de rede quanto de processamento, pode variar ao longo do dia, e para cada momento do dia um novo escalonamento pode ser necessário para se manter uma operação energeticamente eficiente, atendendo ao maior número de usuários.

Desta forma, para cada RRH ativo, uma vBBU deve ser instanciada na nuvem ou em um fog node para realizar o processamento de seus dados. Para economizar energia, nós consideramos que a nuvem está sempre ativa e, enquanto ela possui capacidade de processamento disponível, seus VDUs são ativados para o instanciamento de novas vBBUs. Após as vBBUs serem ativadas, o fronthaul deve prover banda suficiente para suportar transmissões CPRI para a nova vBBU instanciada.

Na CF-RAN, múltiplos RRHs podem transmitir as suas vBBUs compartilhando um mesmo comprimento de onda através de uma VPON. Uma VPON compreende uma PON dedicada, que é estabelecida entre um grupo de RRHs e um nó de processamento em comum. Assim, RRHs em uma VPON compartilham o mesmo canal óptico por meio de TDM.

Para que um VPON seja estabelecido em um nó de processamento, um comprimento de onda deve ser alocado ao OLT desse nó. Assim, o OLT pode configurar esse comprimento de onda nos ONUs de um grupo de RRHs e ativar o LC correspondente para receber o tráfego do VPON. Note que cada comprimento de onda só pode ser alocado para um OLT por vez. Se dois OLTs diferentes usam o mesmo comprimento de onda, os sinais transmitidos podem colidir nos enlaces ópticos caso um gerenciamento da transmissão de múltiplas OLTs não seja utilizado. Tal gerenciamento poderia aumentar a capacidade da rede, mas aumentaria também a complexidade do gerenciamento de cada VPON [Tinini et al. 2019]. Desta forma, assim que vBBUs são instanciadas na nuvem, seu OLT deve configurar a quantidade necessária de comprimentos de onda para suportar as transmissões a seus VDUs.

Quando a demanda de tráfego aumenta, se a capacidade de processamento da nuvem ou seus VPONs não possuírem capacidade suficiente, os fog nodes são ativados. Assim, o operador deve decidir quantos comprimentos de onda devem ser configurados pelos OLTs dos fog nodes. Note que, se todos comprimentos de onda forem alocados para o OLT da nuvem, não será possível estabelecer VPONs nos fog nodes. Assim, o dimensionamento do fronthaul é crucial para o provisionamento da banda necessária em cada fog node. Para maximizar o uso das VPONs entre os nós de processamento, quanto mais RRHs têm suas vBBUs instanciadas em um único nó, menos VPONs serão exigidas para suportar o tráfego em cada nó. A Figura 1a) ilustra um exemplo de dimensionamento dos comprimentos de onda. Os comprimentos de onda $1, 2, 3$ e 4 foram alocados para o OLT da nuvem e os restantes foram divididos entre os fog nodes. Assim, apenas VPONs que utilizam-se desses comprimentos de onda podem ser criados nestes nós.

5. Heurística Baseada em Grafo para a Ativação de vBBUs e Dimensionamento do Fronthaul

Para eficientemente instanciar as vBBUs, ativar os fog nodes e dimensionar a quantidade específica de comprimentos de onda para cada OLT, nós apresentamos uma heurística baseada em teoria dos grafos. Essa heurística é baseada na modelagem da CF-RAN como um grafo direcionado, que deve operar como uma rede de fluxo.

Vértices são utilizados para representarem os RRHs, os nós de nuvem e de fog. Arcos direcionados entre os vértices dos RRHs e dos nós de processamento representam os enlaces ópticos do fronthaul TWDM-PON. Cada arco possui um custo e uma capacidade, utilizados para representar os custos energéticos e as capacidades de processamento

e de banda (quantidade de VPONs configuradas pela OLT) de cada nó, respectivamente. O grafo direcionado é descrito a seguir.

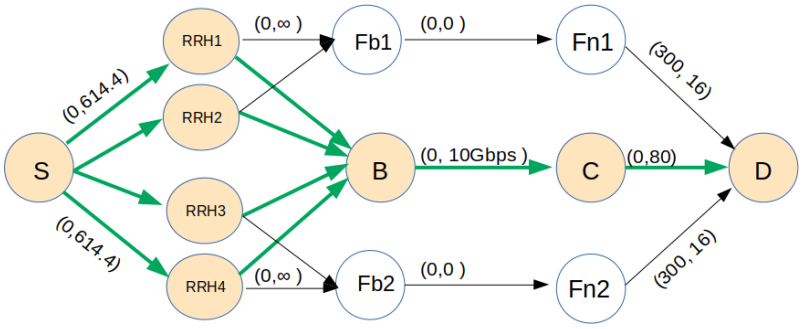

Figura 2. Fluxo em um grafo direcionado representando uma rede CF-RAN

Seja $G = (V, E)$ um grafo direcionado, onde $V(G)$ é o conjunto de vértices de G e $E(G)$ o conjunto de arcos. Cada arco e possui uma capacidade e um custo. Dado $e = (u, v)$, um fluxo é definido partindo de u para v por um arco que é direcionado de u para v, com capacidade maior do que 0. Seja $R \in V$ o conjunto de vértices que representam os RRHs; $F \in V$ o conjunto de vértices que representam os fog nodes; e $FB \in V$ um vértice auxiliar e intermediário chamado de *fog bridge*, responsável por interconectar múltiplos RRHs em um único fog node em nossa implementação. Um vértice $C \in V$ representa a nuvem e um vértice $B \in V$ é um vértice auxiliar e intermediário chamado de *cloud bridge*, usado para representar um enlace entre os RRHs e a nuvem. O vértice $S \in V$ é um vértice de origem direcionado a cada $r \in R$ responsável por injetar tráfego CPRI nos RRHs; o vértice $D \in V$ é responsável por receber todo tráfego CPRI que flui através da rede.

O grafo direcionado é construído da seguinte forma: Para cada $r \in R$, um arco direcionado é colocado de S até r com custo 0 e capacidade 0. Para cada RRH r conectado a um fog node n, um arco direcionado de r até a *fog bridge* $fb \in FB$ é colocado com custo 0 e capacidade ∞. Para cada $fb \in FB$, um arco direcionado de fb até $f \in F$ é colocado com custo 0 e capacidade 0. Para cada $r \in R$, um arco direcionado até B é colocado com custo 0 e capacidade ∞, e um arco direcionado de B até C também é colocado com custo 0 e capacidade 0. Por fim, para cada $f \in F$ um arco direcionado até D é colocado com custo fog_{cost} e capacidade $fog_{capacity}$, e um arco direcionado de C até D é colocado com custo 0 e capacidade $cloud_{capacity}$.

O objetivo da heurística é decidir quantos VPONs devem ser alocados para a OLT em cada nó de processamento, de forma que o fluxo máximo entre os vértices S e D seja maximizado. Isso é feito decidindo quanta capacidade deve ser dada aos arcos que incidem dos vértices de *cloud bridge* e *fog bride*. Note que a capacidade desses arcos será um múltiplo da quantidade de comprimentos de onda alocada para seu nó de processamento. Para que isso seja realizado, nossa heurística é dividida em duas sub-rotinas: na primeira, dada a demanda de tráfego, uma heurística para o dimensionamento do fronthaul é executada para determinar a capacidade de cada arco, considerando os comprimentos de onda livres. Na segunda, um algoritmo de fluxo máximo de custo mínimo (FM-CM) é executado (*max flow-min cost*). O algoritmo FM-CM procura maximizar o fluxo entre os vértices S e D minimizando a soma dos custos dos arcos atravessados. Um exem-

plo de fluxo é ilustrado na Figura 2, onde cada RRH transmite à nuvem através de um VPON com capacidade de 10Gbps, configurado pelo OLT da nuvem e representado pela capacidade do arco entre B e C.

Algorithm 1 Heurística CF-LLB

Input: Grafo direcionado G, tráfego do RRH $r \in RB_r$, demanda de tráfego da rede, comprimentos de onda disponíveis $w \in W$
Output: Processamento do RRH i e dimensionamento do fronthaul

```
for all RRH r requisitando transmissão do
    if A demanda da rede é menor que a capacidade de processamento da nuvem then
        Verifique se o tráfego da rede pode ser suportado pela nuvem
        if Não há VPONs disponíveis ou a demanda da rede é maior que a banda disponível then
            while A capacidade dos VPONs da nuvem for menor que a demanda da rede do
                if Há comprimentos de onda disponíveis then
                    Aloque um comprimento de onda para a nuvem
        Execute um algoritmo de fluxo máximo de custo mínimo
    else if Tráfego da rede é maior que a capacidade da nuvem then
        Dimensione os comprimentos de onda para a nuvem primeiramente e então para os fog nodes
        if Há comprimentos de onda disponíveis then
            while Capacidade dos VPONs da nuvem é menor que a capacidade de processamento do
                Aloque um comprimento de onda para a nuvem
        Ordene os fog nodes em ordem crescente da quantidade de RRHs ativos conectados a ele, se
    e somente se, os fog nodes não tiverem VPONs suficientes para suportar a transmissão de todos esses
    RRHs
        while A demanda da rede for maior do que a capacidade de todos VPONs disponíveis e houver
    comprimentos de onda não alocados a nenhum nó do
            Escolha o fog node com menos RRHs ativos
            Aloque um comprimento de onda para esse fog node
    Execute um algoritmo de fluxo máximo de custo mínimo
```

Para dimensionar os comprimentos de onda para os nós de processamento, é feita a proposta da heurística Cloud First-Least Loaded Bandwidth (CF-LLB), formalmente descrita no Algoritmo 1. Dado um ou mais RRHs requisitando a instanciação de uma vBBU em algum nó e um VPON, enquanto há capacidade na nuvem, CF-LLB primeiramente aloca capacidade no arco (C, D) enquanto o total de carga de processamento da nuvem é maior que a capacidade do arco (B, C). O total de capacidade alocada no arco (B, C) é limitado pela capacidade de processamento da nuvem.

Quando os fog nodes são ativados e demandam VPONs para receberem transmissões dos RRHs, CF-LLB ordena em uma lista em ordem crescente os fog nodes que não têm banda suficiente em função da quantidade de RRHs conectados a eles, que estão ativos. Assim, enquanto a demanda de tráfego é maior do que a banda disponível em todos os nós, CF-LLB aloca um VPON para cada fog node da lista.

6. Análise de Desempenho

Para atestar a eficiência da heurística proposta, utilizamos nosso simulador de redes 5G, 5GPy [Tinini et al. 2020][1]. Ao nosso conhecimento, não há na literatura algoritmos que solucionam o problema proposto neste artigo em arquiteturas híbridas como a CF-RAN. Assim, nós comparamos a heurística CF-LLB a um modelo de ILP proposto em [Tinini et al. 2019] para a solução do problema de dimensionamento do fronthaul na

[1]Disponível em https://github.com/rodrigo-tinini/5GPy

CF-RAN em um cenário de tráfego estático, onde a carga da rede é conhecida de antemão. A heurística CF-LLB também foi comparada com outras heurísticas presentes na literatura em cenários de tráfego dinâmico. O modelo ILP é formalmente descrito a seguir.

6.1. Formulação do ILP Usado para Comparação

Parâmetros de Entrada

R: conjunto de RRHs r, N: conjunto dos nós de processamento n, W: conjunto de comprimentos de onda/VPONs w, B_i: Demanda CPRI de cada RRH r, B_w: capacidade de um comprimento de onda w, $Proc_i$: demanda de processamento de um RRH r, $Proc_n$: capacidade de processamento de um nó n.

Variáveis de Decisão

y^i_{wn}: = 1, se a vBBU do RRH r é instanciado no nó n e utiliza o VPON w, 0 senão; z_{wn}: = 1, se o comprimento de onda/VPON w é alocado ao OLT do nó n, 0 senão; x_n: = 1 se o nó n é ativado, 0 senão.

Função Objetivo

A função objetivo (1) busca reduzir o consumo total de energia da rede, dado por $P_{network} = \sum_{n=1}^{N} x_n.C_n + \sum_{w=1}^{W}\sum_{n=1}^{N} z_{wn}.(C_{lc} + C_{du})$, onde C_n é o consumo de energia de um nó n, C_{du} é o consumo de energia de um VDU e C_{lc} é o consumo de energia de um LC.

$(1) Minimize P_{network}$

Restrições

$(2) \sum_{n=1}^{N} z_{wn} \leq 1 | \forall w \in W; \quad (3) \sum_{w=1}^{W}\sum_{n=1}^{N} y^i_{wn} = 1 | \forall i \in R$

$(4) \sum_{i=1}^{R}\sum_{n=1}^{N} y^i_{wn}.B_i \leq B_w | \forall w \in W; \quad (5) \sum_{i=1}^{R}\sum_{w=1}^{W} y^i_{wn}.Proc_i \leq Proc_n | \forall n \in N$

$(6) y^i_{wn} = 1 \Rightarrow x_n = 1 | \forall i, w, n \in R, W, N$

$(7) y^i_{wn} = 1 \Rightarrow z_{wn} = 1 | \forall i, w, n \in R, W, N$

A restrição (2) limita a alocação de cada comprimento de onda a um único OLT por vez; a restrição (3) garante que cada RRH somente seja processado em um único nó; a restrição (4) limita a quantidade de RRHs em um VPON de acordo com a capacidade do comprimento de onda; a restrição (5) limita a quantidade de vBBUs em um nó à capacidade de processamento do nó; a restrição (6) ativa um nó n assim que um RRH r é alocado a ele; e a Restrição (7) garante que um VPON w seja alocado ao OLT do nó n assim que um RRH é alocado nesse nó.

O modelo ILP foi executado no software IBM CPLEX 12.8. Tanto as simulações quanto o modelo ILP foram executados em um computador Intel i7 com 16GB de RAM, rodando o Ubuntu 16.04. Quanto ao cenário de tráfego dinâmico, ele foi simulado utilizando-se de um padrão diário de tráfego móvel em uma região comercial, utilizando-se dos dados providos em [Peng et al. 2011]. A Figura 3 mostra a carga máxima de processamento das VDUs em diferentes horas do dia. No início da simulação, todos os RRHs estão desativados e começam a ser ativados e solicitar a instanciação de uma vBBU se-

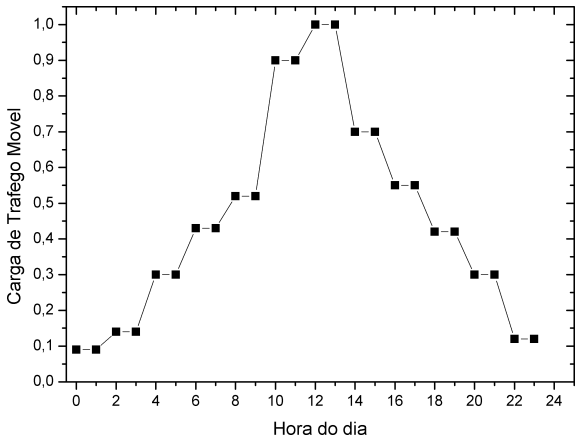

Figura 3. Padrão de tráfego móvel em uma região comercial

gundo um processo de Poisson, com média igual ao *erlang* da rede para aquele momento. O tempo que cada RRH permanece consumindo uma VPON e uma vBBU é uniformemente sorteado no intervalo $(25min., 1hora)$.

Tabela 1. Parâmetros de Simulação

Nós de processamento	1 nuvem e 5 fog nodes
Capacidade de processamento	80 RRHs (nuvem) 16 RRHs (fog node)
Número de RRHs no caso dinâmico	160
Configuração MIMO dos RRHs	2 canais de rádio de 10MHz
Taxa CPRI de cada RRH	614, 4Mbps
Comprimentos de onda	10
Capacidade de um comprimentos de onda	10Gbps
Consumo da nuvem	600 watts
Consumo de um fog node	300 watts
Consumo de uma vBBU	15 watts
Consumo de um line card	5 watts
Consumo do OLT	100 watts

Neste cenário, a heurística CF-LLB foi comparada a duas propostas da literatura [Tinini et al. 2019]: a heurística *fog first* (FF), onde primeiramente os fog nodes são ativados e somente após a nuvem; e com a heurística *most loaded* (ML) para dimensionamento da banda, onde os VPONs são criados primeiramente nos fog nodes que possuem mais RRHs ativos. Os parâmetros utilizados nas simulações são apresentados na Tabela 1. Todos os resultados do caso dinâmico foram obtidos da média de 60 execuções com um nível de confiança de 95%.

7. Resultados Numéricos

As métricas avaliadas foram, para o caso estático, o consumo de energia, o tempo de execução e a latência de propagação na rede. Para o caso dinâmico, foram avaliados a probabilidade de bloqueio, o consumo de energia e o tempo de execução. Foi considerado que tanto os nós de processamento quanto a rede óptica têm capacidade para atender a todos os 160 RRHs. Assim, a probabilidade de bloqueio só irá ocorrer em decorrência

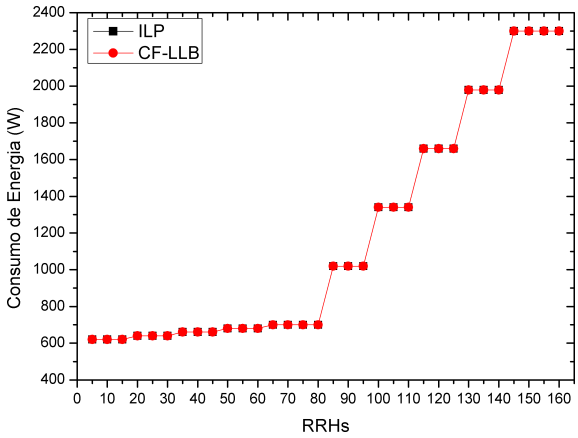

Figura 4. Consumo de energia com tráfego estático

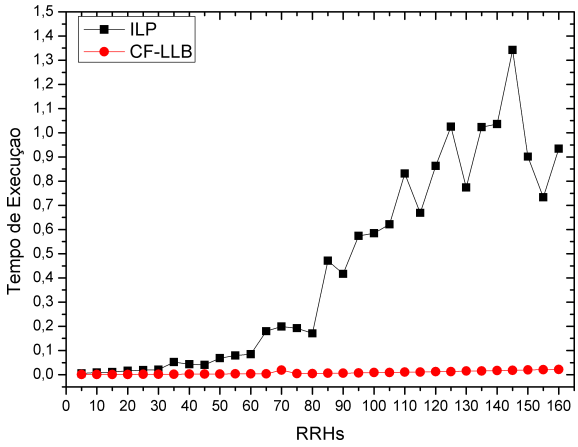

Figura 5. Tempo de execução (em segundos) com tráfego estático

da eficiência das heurísticas. Também assumimos que a ativação dos fog nodes se dá em decorrência do esgotamento da capacidade de processamento da nuvem.

A Figura 4 mostra a comparação das soluções providas pelo ILP e pela heurística proposta. A heurística CF-LLB foi capaz de prover as mesmas soluções ótimas que o modelo ILP, não apresentando nenhuma variação nos resultados. Entretanto, como mostra a Figura 5, o tempo de execução é drasticamente reduzido pela CF-LLB em comparação com o ILP, alcançando uma redução de até $98,7\%$.

A Figura 6 mostra a latência média de propagação, em segundos, na CF-RAN. Para quantidades menores de RRHs, a latência é maior pois todos são processados na nuvem. Conforme os fog nodes são ativados, ao ser esgotada a capacidade de processamento da nuvem, a latência média diminui. Em ambos os casos, a latência provida pelo fronthaul TWDM-PON está abaixo dos 3μs requisitados pelo protocolo CPRI.

Em relação ao caso de tráfego dinâmico, a Figura 7 mostra a probabilidade de bloqueio. Note que a heurística ML possui uma probabilidade de bloqueio muito alta, ao passo que as heurísticas FF e CF-LLB possuem probabilidades próximas de 0.

A Figura 8 mostra o consumo de energia das heurísticas. Note que a política ML possui o melhor desempenho energético, mas ao custo de altas probabilidades de blo-

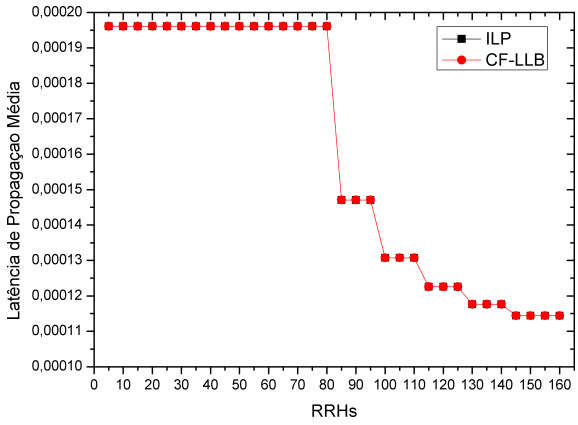

Figura 6. Latência média de propagação (em segundos) com tráfego estático

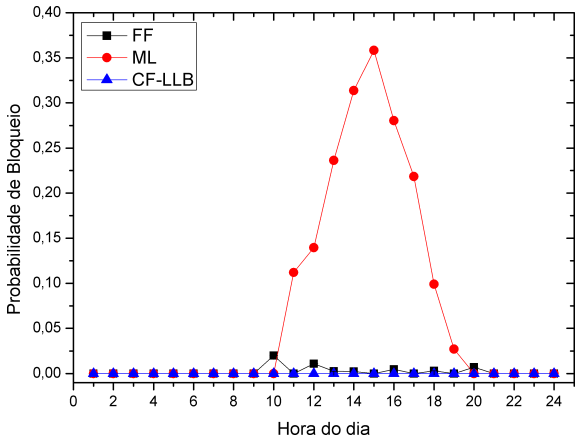

Figura 7. Probabilidade de bloqueio com tráfego dinâmico

queio. Em relação à heurística FF, a CF-LLB possui melhor consumo energético em todos os momentos do dia, aproximando-se da FF nos horários de pico, sendo capaz de reduzir em até 9% o consumo energético em tais horários e em até 58% em horários de tráfego menor. A Figura 9 mostra os tempos de execução para o caso dinâmico. Novamente, a heurística ML possui o melhor tempo, mas ao custo de piores soluções. As heurísticas CF-LLB e FF possuem tempos de execução semelhantes, entretanto, como supracitado, a CF-LLB é capaz de prover um desempenho melhor para a operação da rede.

8. Conclusão

Neste trabalho foi proposta uma heurística para CF-RAN baseada em grafos para realizar a ativação dinâmica de fog nodes e o dimensionamento da banda necessária no fronthaul, a fim de prover transmissões a esses nós em cenários de tráfego dinâmico. A heurística foi capaz de obter as mesmas soluções ótimas que um algoritmo baseado em ILP, porém com reduções de até $98,7\%$ no tempo de execução. Em cenários dinâmicos, a heurística foi capaz de mitigar a probabilidade de bloqueio em comparação a duas heurísticas de ativação dos fog nodes.

Agradecimentos

Esta pesquisa é parte do INCT da Internet do Futuro para Cidades Inteligentes, financiado por CNPq (proc. 465446/2014-0), Coordenação de Aperfeiçoamento de Pessoal de Nível

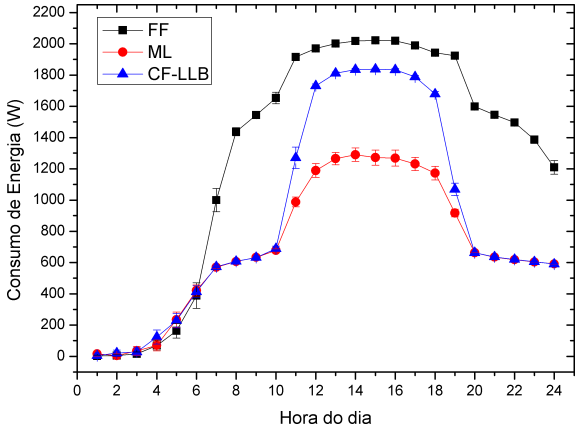

Figura 8. Consumo de energia com tráfego dinâmico

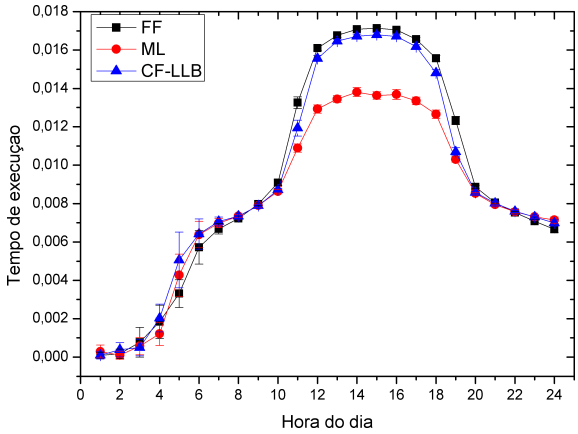

Figura 9. Tempo de execução com tráfego dinâmico

Superior – Brasil (CAPES) – Código de Financiamento 001 e FAPESP (procs. 14/50937-1 e 15/24485-9). Também é parte do projeto FAPESP proc. 18/22979-2.

Referências

Bonomi, F., Milito, R., Zhu, J., and Addepalli, S. (2012). Fog computing and its role in the internet of things. In *Proceedings of the first edition of the MCC workshop on Mobile cloud computing*, pages 13–16. ACM.

Chanclou, P., Pizzinat, A., Le Clech, F., Reedeker, T.-L., Lagadec, Y., Saliou, F., Le Guyader, B., Guillo, L., Deniel, Q., Gosselin, S., et al. (2013). Optical fiber solution for mobile fronthaul to achieve cloud radio access network. In *Future Network and Mobile Summit (FutureNetworkSummit), 2013*, pages 1–11. IEEE.

Chitimalla, D., Kondepu, K., Valcarenghi, L., Tornatore, M., and Mukherjee, B. (2017). 5G fronthaul-latency and jitter studies of CPRI over ethernet. *IEEE/OSA Journal of Optical Communications and Networking*, 9(2):172–182.

Cui, X., Jiang, Y., Chen, X., Zhengy, F., and You, X. (2018). Graph-based cooperative caching in Fog-RAN. In *2018 International Conference on Computing, Networking and Communications (ICNC)*, pages 166–171. IEEE.

de la Oliva, A., Hernández, J. A., Larrabeiti, D., and Azcorra, A. (2016). An overview of the CPRI specification and its application to C-RAN-based LTE scenarios. *IEEE Communications Magazine*, 54(2):152–159.

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 *Communications (ICC), 2016 IEEE International Conference on*, pages 1–6. IEEE.

Hawilo, H., Shami, A., Mirahmadi, M., and Asal, R. (2014). NFV: state of the art, challenges, and implementation in next generation mobile networks (vEPC). *IEEE Network*, 28(6):18–26.

Iida, D., Kuwano, S., Kani, J.-i., and Terada, J. (2013). Dynamic TWDM-PON for mobile radio access networks. *Optics Express*, 21(22):26209–26218.

Luo, Y., Zhou, X., Effenberger, F., Yan, X., Peng, G., Qian, Y., and Ma, Y. (2013). Time- and wavelength-division multiplexed passive optical network (TWDM-PON) for next-generation PON stage 2 (NG-PON2). *Journal of Lightwave Technology*, 31(4):587–593.

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. ACM.

Peng, M., Li, Y., Jiang, J., Li, J., and Wang, C. (2014). Heterogeneous cloud radio access networks: a new perspective for enhancing spectral and energy efficiencies. *IEEE Wireless Communications*, 21(6):126–135.

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. *J. Opt. Commun. Netw.*, 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.

Wang, K., Yang, K., and Magurawalage, C. S. (2018). Joint energy minimization and resource allocation in C-RAN with mobile cloud. *IEEE Transactions on Cloud Computing*, 6(3):760–770.

Wang, X., Alabbasi, A., and Cavdar, C. (2017). Interplay of energy and bandwidth consumption in CRAN with optimal function split. In *Communications (ICC), 2017 IEEE International Conference on, 21-25 May 2017, Paris, France*. IEEE conference proceedings.

Wang, X., Thota, S., Tornatore, M., Chung, H. S., Lee, H. H., Park, S., and Mukherjee, B. (2016). Energy-efficient virtual base station formation in optical-access-enabled cloud-RAN. *IEEE Journal on Selected Areas in Communications*, 34(5):1130–1139.

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