Universidade do Estado do Rio de Janeiro
Centro de Tecnologia e Ciências
Faculdade de Engenharia

Jean Lucas de Lima

Minimização da Latência no Posicionamento de Funções em Cloud RANs

Rio de Janeiro
2018

Jean Lucas de Lima

Minimização da Latência no Posicionamento de Funções em Cloud RANs

Dissertação apresentada, como requisito parcial para obtenção do título de Mestre em Engenharia Eletrônica, ao Programa de Pós-Graduação em Engenharia Eletrônica, da Universidade do Estado do Rio de Janeiro. Área de concentração: Redes de Telecomunicações.

Orientador: Prof. D.Sc. Rodrigo de Souza Couto

Rio de Janeiro
2018

CATALOGAÇÃO NA FONTE

UERJ / REDE SIRIUS / BIBLIOTECA CTC/B

L732 Lima, Jean Lucas de.
Minimização da latência no posicionamento de funções em cloud RANs / Jean Lucas de Lima. – 2018.
51f.

Orientador: Rodrigo de Souza Couto.
Dissertação (Mestrado) – Universidade do Estado do Rio de Janeiro, Faculdade de Engenharia.

1. Engenharia Eletrônica - Teses. 2. Sistemas de comunicação móvel - Teses. 3. Computação em nuvem - Teses. 4. Serviços da Web - Teses. 5. Sistemas de telecomunicação - Teses. I. Couto, Rodrigo de Souza. II. Universidade do Estado do Rio de Janeiro, Faculdade de Engenharia. II. Título.

CDU 004.7

Bibliotecária: Júlia Vieira – CRB7/6022

Assinatura

Data

Jean Lucas de Lima

Minimização da Latência no Posicionamento de Funções em Cloud RANs

Dissertação apresentada, como requisito parcial para obtenção do título de Mestre em Engenharia Eletrônica, ao Programa de Pós-Graduação em Engenharia Eletrônica, da Universidade do Estado do Rio de Janeiro. Área de concentração: Redes de Telecomunicações.

Aprovada em 02 de Fevereiro de 2018.
Banca Examinadora:

Prof. D.Sc. Rodrigo de Souza Couto (Orientador)
Faculdade de Engenharia - UERJ

Prof. D.Sc. Marcelo Gonçalves Rubinstein
Faculdade de Engenharia - UERJ

Prof. D.Sc. Diego Gimenez Passos
IC/UFF

Rio de Janeiro
2018

DEDICATÓRIA

Dedico este trabalho a Deus em primeiro lugar, que me deu suporte trazendo a tranquilidade que precisei nos momentos de dúvida e dificuldade. Aos meus pais e irmãos que estiveram sempre presentes ao longo de toda essa caminhada. Ao meu orientador e a todos os professores que me trouxeram tanto conhecimento durante a vida e também a todos aqueles que estiveram ao meu lado durante todos esses anos e ainda estão, para hoje, comemorarmos com muita alegria este momento de felicidade e vitória.

AGRADECIMENTOS

Agradeço primeiramente a Deus por ter sido o alicerce que me manteve de pé durante todo o tempo ao longo desse curso, sempre me convidando a conhecer o dom de Deus quando eu quis desistir e me mostrando que a vitória chegaria por mais difícil que a trajetória parecesse. Agradeço de maneira imensurável todo o apoio dos meus pais que sempre estiveram presentes, apesar da distância física, por serem minha base e fontes de apoio incondicionais em minha vida. Agradeço aos meus irmãos, por todas as palavras de amizade e carinho, nos momentos em que eu mais precisei de compreensão e, principalmente, por serem os meus melhores amigos. Agradeço ao meu orientador por me acompanhar nos últimos dois anos e meio de mestrado, despertando em mim o interesse pela pesquisa, além de fornecer o conhecimento e sempre me encorajar a seguir adiante. Agradeço aos meus amigos de infância, amigos do Grupo Jura e amigos que a vida profissional me trouxe por acreditarem em meu potencial, por cada palavra de conforto e ombro amigo, mas também por todos os momentos de alegria que compartilhamos e fizeram a vida valer a pena. Agradeço ao apoio dos colegas de classe, que foram sempre prestativos e companheiros durante todos os dias de aula e estudo. Agradeço aos membros da banca examinadora que aceitaram dedicar seu tempo para analisar esta dissertação e contribuíram para o sucesso da mesma. Agradeço a todos os professores, por todos os ensinamentos compartilhados que me desenvolveram como profissional, mas principalmente como indivíduo. Agradeço às agências de fomento envolvidas pelo suporte primordial para que esse trabalho fosse possível.

Se conhecesses o dom de Deus, e quem é que te diz: Dá-me de beber, certamente lhe pedirias tu mesma e Ele te daria a água viva.

João 4, 10

RESUMO

LIMA, J. L. L. *Minimização da Latência no Posicionamento de Funções em Cloud RANs*. 2018. 51 f. Dissertação (Mestrado em Engenharia Eletrônica) – Faculdade de Engenharia, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2018.

Utilizar de forma mais eficiente os recursos de processamento do sinal banda base e proporcionar um gerenciamento centralizado da rede são grandes desafios das operadoras de rede móvel. Para satisfazer o crescimento da demanda por redes móveis, as operadoras de rede celular precisam aumentar a capacidade das estações base, que são responsáveis pela comunicação da rede com os usuários. Um dos caminhos para essa ampliação é aumentar o número de estações base, adicionando células com pequenas áreas de cobertura, criando uma estrutura de rede heterogênea . Além disso, é possível oferecer uma infraestrutura de processamento centralizada, capaz de atribuir recursos às estações de forma dinâmica, reduzindo a quantidade de *hardware* necessária e aumentando a escalabilidade da rede. Com base nessa necessidade, o conceito de C-RAN (*Cloud Radio Access Network*) consiste em executar funções de estações rádio base em uma infraestrutura de nuvem, que pode ser centralizada ou composta por diversos níveis de hierarquia. Assim, as estações base atuam apenas como receptores de sinais, que são posteriormente processados na nuvem. Dada a distância entre a nuvem e as estações, a latência é um fator crítico em C-RAN. Nesta dissertação formula-se um problema de programação linear inteira mista para escolher o posicionamento das funções de rádio em uma C-RAN, de forma a minimizar a latência em uma nuvem com diferentes níveis de hierarquia e diferentes capacidades de processamento e transmissão. Para solução do problema, este trabalho propõe duas heurísticas, uma para redes nas quais todos os enlaces possuem a mesma latência e outra para redes nas quais os enlaces possuem latências diferentes, e mostra situações nas quais essas alcançam o resultado ótimo. A primeira heurística possui complexidade $O(n)$, enquanto a segunda, que é mais geral, possui complexidade $O(n \log n)$.

Palavras-chave: Cloud RAN; Alocação de recursos; BBU; RRH.

ABSTRACT

LIMA, J. L. L. *Minimizing Latency in Cloud RAN Function Placement.* 2018. 51 f. Dissertação (Mestrado em Engenharia Eletrônica) – Faculdade de Engenharia, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2018.

Using the processing resources of baseband signals more efficiently and employing a centralized management of the network are major challenges for mobile network operators. To meet the growth of the mobile network demand, cellular network operators need to increase base stations capacities, which are responsible for the network communication with the users. One alternative to perform this expansion is to increase the number of base stations, adding cells with small coverage areas, creating a heterogeneous network infrastructure. In addition, it is possible to provide a centralized processing infrastructure, able to dynamically allocate resources to the stations, reducing the required amount of hardware and increasing the network scalability. Based on this need, the concept of C-RAN (Cloud Radio Access Network) is to execute base station functions in a cloud infrastructure, which can be centralized or composed of several hierarchy levels. The base stations thus act only as signal receivers, which are later processed in the cloud. Given the distance between the cloud and the stations, latency is a critical factor in C-RAN. This dissertation formulates a Mixed Integer Linear Programming problem to choose the placement of radio functions in a C-RAN, while minimizing the latency in a cloud with different hierarchy levels and different processing and transmission capacities. To solve the problem, this work proposes two heuristics, one for networks in which all the links have the same latency and another one for networks in which the links have different latencies. We then show situations in which they reach the optimal result. The first heuristic has complexity $O(n)$, while the second one, which is more general, has complexity $O(n \log n)$.

Keywords: Cloud RAN; Resource allocation; BBU; RRH.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ALGORITMOS

LISTA DE ABREVIATURAS E SIGLAS

3GPP	*3rd Generation Partnership Project*
API	*Application Programming Interface*
BBU	*Band Base Unit*
BS	*Base Station*
C-RAN	*Cloud Radio Access Network*
ETSI	*European Telecommunications Standards Institute*
ILP	*Integer Linear Programming*
MILP	*Mixed Integer Linear Programming*
NFV	*Network Functions Virtualization*
NFV-RA	*Network Functions Virtualization - Resource Allocation*
NFVI	*Network Functions Virtualization Infrastructure*
NFVO	*Network Functions Virtualization Orchestrator*
OFDMA	*Orthogonal Frequency Division Multiple Access*
RAN	*Radio Access Network*
RRH	*Remote Radio Head*
SDN	*Software Defined Networking*
SLA	*Service Level Agreements*
TI	*Tecnologia da Informação*
VDU	*Virtual Data Units*
VNF	*Virtualized Network Function*
WLAN	*Wireless Local Area Network*

LISTA DE SÍMBOLOS

$\%$	Porcentagem
$\sum$	Somatório
$\in$	Pertence a
$\leq$	Menor ou igual a
$\forall$	Para todos os valores de
$\geq$	Maior ou igual a
$=$	Igual a
$\lvert\ \rvert$	Módulo de

SUMÁRIO

INTRODUÇÃO

Atualmente, nota-se um aumento significativo no consumo de dados de usuários finais devido ao aumento de dispositivos com tecnologia 3G e 4G. Para satisfazer o aumento da demanda de tráfego celular, operadoras são forçadas a buscar novas soluções de custo benefício, permitindo atualizações, melhor gerenciamento e dimensionamento da rede de acesso a rádio (*Radio Access Network* - RAN) (ALYAFAWI et al., 2015). A RAN é uma infraestrutura de telecomunicações projetada para prover conectividade a dispositivos móveis em um sistema de rede celular. Assim, em uma RAN, o usuário conecta seu dispositivo a uma estação base (*Base Station* - BS) que, por sua vez, encaminha o tráfego para o núcleo da rede.

Com o crescente consumo de dados de usuários finais, as RANs têm necessitado de crescente aumento de capacidade. Um dos caminhos para expandir uma RAN é aumentar o número de BSs, criando uma estrutura de rede de pequenas células (HATOUM et al., 2014), ou aumentar a capacidade das BSs já instaladas. Entretanto, esse aumento de capacidade pode reduzir a eficiência do uso de recursos, já que uma determinada BS pode se manter ociosa por longos períodos. Dados indicam que somente de 15% a 20% das BSs operam com mais de 50% da sua capacidade total (ALYAFAWI et al., 2015). Por exemplo, BSs instaladas nas proximidades de grandes estádios recebem uma alta quantidade de usuários em dias de evento. Entretanto, podem estar ociosas nos demais períodos. Uma alternativa é centralizar os recursos que podem ser compartilhados por diversas BSs, semelhante ao que ocorre na Computação em Nuvem. Com isso, uma BS pode solicitar à infraestrutura centralizada apenas a quantidade de recursos necessária em um determinado momento. Essa solução é denominada Cloud RAN, ou C-RAN, e visa oferecer escalabilidade e flexibilidade em um sistema de rede celular (CHECKO et al., 2015).

Para oferecer serviços de rede celular, uma BS possui unidades de rádio (*Remote Radio Head* - RRH) e unidades de banda base (*Band Base Units* - BBUs). A RRH realiza o processamento digital de sinais, a conversão analógico/digital e digital/analógico, e implementa interfaces com os meios de transmissão (CHECKO et al., 2015). A BBU, por sua vez, processa o sinal de banda base. Esse sinal se refere à faixa de frequência original de um sinal de transmissão antes que este seja modulado. Pode se referir também a um tipo de transmissão de dados em que os dados digital ou analógico são enviados por meio de um único canal não multiplexado. Assim, a BBU realiza tarefas de mais alto nível, como controle de acesso ao meio e correção de erros (BARTELT et al., 2015). O sinal banda base é processado pela BBU em sistemas de telecomunicações ou em um *cluster* virtualizado, que consiste em processadores de utilização geral. Uma infraestrutura de BBUs pode ser compartilhada por diferentes operadoras de rede, permitindo o aluguel da

RAN como um serviço na nuvem. Em redes celulares tradicionais, a RRH e a BBU estão na mesma BS, integradas em um mesmo equipamento ou conectadas por fibra óptica. Em uma C-RAN, as RRHs são instaladas nas BSs, que estão espalhadas geograficamente, enquanto as unidades de banda base são centralizadas e podem atender diversas BSs.

Uma desvantagem da C-RAN é a latência inserida entre a BBU e a RRH. Isso é crítico visto que tradicionalmente essas unidades se comunicam por enlaces de poucos metros e, na C-RAN, utilizam enlaces que podem ter comprimentos da ordem de quilômetros. Para solucionar esse problema, utiliza-se o conceito denominado *Dynamic* C-RAN (DALLA-COSTA et al., 2017a). Esse conceito é similar ao de Computação em Névoa (COUTINHO; CARNEIRO; GREVE, 2016), que consiste em uma camada intermediária entre a nuvem e os usuários. Assim, nesse tipo de arquitetura, o processamento das funções de rádio é realizado de forma dinâmica e é distribuído em um conjunto de nuvens, que podem ser organizadas de forma hierárquica. Por exemplo, um conjunto de nuvens de borda, com pouca capacidade computacional, pode ser instalado em locais próximos a algumas BSs. Caso essas BSs necessitem de maior poder computacional, é possível utilizar nuvens mais centrais, com maior capacidade. Mais uma vez, recorre-se ao exemplo do estádio. Em dias comuns, os usuários que estão próximos ao estádio utilizam nuvens locais, que são suficientes para suprir as demandas rotineiras. Em dias de evento, utilizam-se nuvens centrais para processar o tráfego dos usuários excedentes.

Uma forma de implementar uma *Dynamic* C-RAN é utilizar o conceito de Virtualização de Funções de Rede (*Network Functions Virtualization* - NFV) (MIJUMBI et al., 2015). Assim, cada função da BBU pode ser implementada como uma função de rede virtual (*Virtualized Network Function* - VNF). As VNFs são então distribuídas pela hierarquia de nuvens de acordo com a demanda. Essa distribuição deve ser realizada por algoritmos de posicionamento de VNFs. No caso específico de *Dynamic* C-RAN, (DALLA-COSTA et al., 2017b) propõem um modelo de programação linear inteira mista (*Mixed Integer Linear Programming* - MILP) para escolher o posicionamento de VNFs de forma a minimizar o uso de banda na rede. Apesar disso, poucos trabalhos consideram o posicionamento em NFV para o cenário específico de C-RAN (HERRERA; BOTERO, 2016).

Objetivos

Esta dissertação formula um problema MILP baseado em (DALLA-COSTA et al., 2017a; DALLA-COSTA et al., 2017b), para posicionar funções em uma *Dynamic* C-RAN. Nessa direção, os trabalhos (DALLA-COSTA et al., 2017a; DALLA-COSTA et al., 2017b) formulam um problema de otimização para minimizar o uso de banda nos enlaces entre as nuvens e privilegiar as nuvens de borda, que são as nuvens mais próximas das BSs.

Apesar de privilegiar a nuvem de borda reduzir a latência da rede, essa solução é restritiva, uma vez que ignora a baixa latência de nuvens regionais, isto é, que se localizam entre as centrais e as de borda.

Dado o exposto, este trabalho se baseia em (DALLA-COSTA et al., 2017b) para formular um novo problema que considera a latência na função objetivo e trata a banda como uma restrição. Com isso, privilegiam-se também nuvens intermediárias. Tratar a banda como restrição, e não como um objetivo a ser otimizado, é justificado pelo fato de as operadoras de celular possuírem redes bem provisionadas. Além disso, este trabalho propõe heurísticas para solucionar o problema formulado, diferente do trabalho da literatura, no qual apenas a solução por meio de um otimizador de MILP é apresentada.

O problema proposto minimiza a latência média da rede e considera restrições de capacidade das nuvens. Além disso, dada a complexidade de tempo do MILP, este trabalho propõe duas heurísticas para a solução do problema formulado e mostra situações nas quais essas alcançam o resultado ótimo. Para analisar a solução de otimização e as heurísticas propostas, são realizadas análises para medir a latência média e o tempo gasto no posicionamento das funções de rádio dentro de uma hierarquia de nuvens, a partir da variação de parâmetros que configuram diversos cenários pré-estabelecidos. Também analisa-se a complexidade das heurísticas propostas. A heurística que só se aplica em cenários mais restritos, como redes nas quais todos os enlaces possuem a mesma latência, possui complexidade $O(n)$, enquanto a mais geral possui complexidade $O(n \log n)$.

Organização do Texto

Esta dissertação está estruturada da seguinte forma. O Capítulo 1 apresenta conceitos preliminares para entendimento da proposta desta dissertação. O Capítulo 2 trata dos trabalhos relacionados, enquanto o Capítulo 3 formula o problema de otimização, descrevendo a função objetivo e as restrições adotadas. Os Capítulos 4 e 5 descrevem as heurísticas propostas e apresentam os resultados. Finalmente, o Capítulo 6 conclui o trabalho e aponta direções de pesquisa futuras.

1 CONCEITOS PRELIMINARES

O custo de construção e operação de uma nova infraestrutura para oferecer a capacidade de dados necessária, com a crescente demanda de redes celulares, é superior à taxa de crescimento do capital disponível para investimento nas operadoras (CHECKO et al., 2015). A principal razão para o elevado custo de atualização e manutenção da arquitetura da rede de telecomunicações móvel se dá devido ao alto valor dos equipamentos contidos nas estações base, constituídas por componentes de *hardware* e *software*. Assim, o aumento na quantidade de *hardware* da RAN, para atender o crescimento da demanda, impacta de forma direta a receita da operadora.

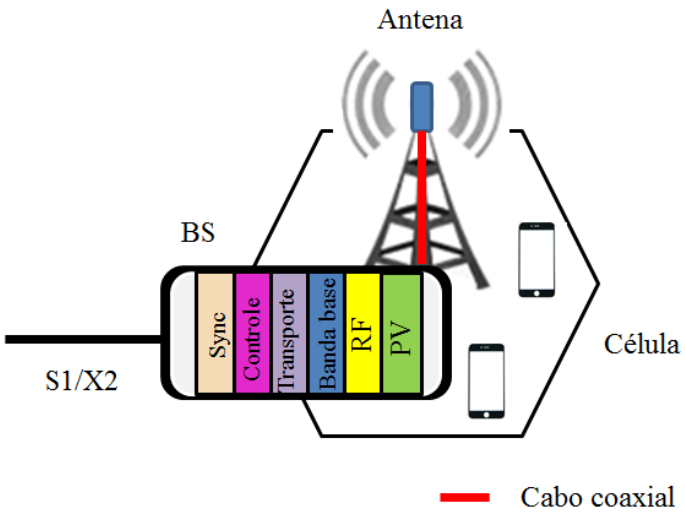
Como fornecedores de telefonia móvel operam em larga escala, são susceptíveis a várias adversidades, como por exemplo, queda do sinal por conta de sobrecarga de uma célula provisionada, falha em *hardwares* do sistema causados pelo tempo de vida útil, ou problemas na atualização da rede devido à incompatibilidade dos equipamentos com as novas soluções de projeto. Então, para que a RAN ofereça máximo desempenho, fornecedores devem oferecer alta densidade de BSs em amplas áreas geográficas. No entanto, a instalação e manutenção de um número alto de BSs integradas tem um preço muito elevado e, com isso, a tendência para o futuro é que sejam implementadas células de tamanhos menores, como as pico células, por exemplo (ALYAFAWI et al., 2015).

Diante do cenário exposto, fizeram-se necessárias diversas mudanças no modelo de implantação da rede de telefonia móvel. Para que essa rede fosse capaz de suportar o aumento das demandas e trabalhasse de forma mais eficiente e mais bem gerenciada, houve uma evolução do modelo de RAN tradicional até a C-RAN com a finalidade de aproveitar de forma mais eficiente os recursos de processamento da rede disponível.

A redução de custos de infraestrutura, em diferentes tecnologias de comunicação ou de computação, é possibilitada pelo NFV e pela tecnologia de computação em nuvem, permitindo migração de *hardwares* específicos e com custo elevado para plataformas de TI (Tecnologia da Informação) de uso geral, balanceamento de carga, rápida implantação e provisionamento de serviços, além de economia de energia. Por exemplo, o processo de conversão de sistemas tradicionais para computação em nuvem tem demonstrado mais de 70% de economia de energia, se comparado a sistemas já existentes (ALYAFAWI et al., 2015). Assim a C-RAN emprega conceitos de NFV e computação em nuvem para oferecer maior eficiência em redes celulares.

As seções a seguir apresentam uma visão geral da evolução da arquitetura RAN e as vantagens da utilização da C-RAN. Discutem-se também a NFV e sua importância na alocação de recursos, a divisão das funções de rádio em C-RANs e o conceito de *Dynamic* C-RAN, que foi utilizado na implementação do modelo desta dissertação.

Figura 1 - Arquitetura RAN tradicional - centralizada.



Fonte: Adaptado de (CHECKO et al., 2015).

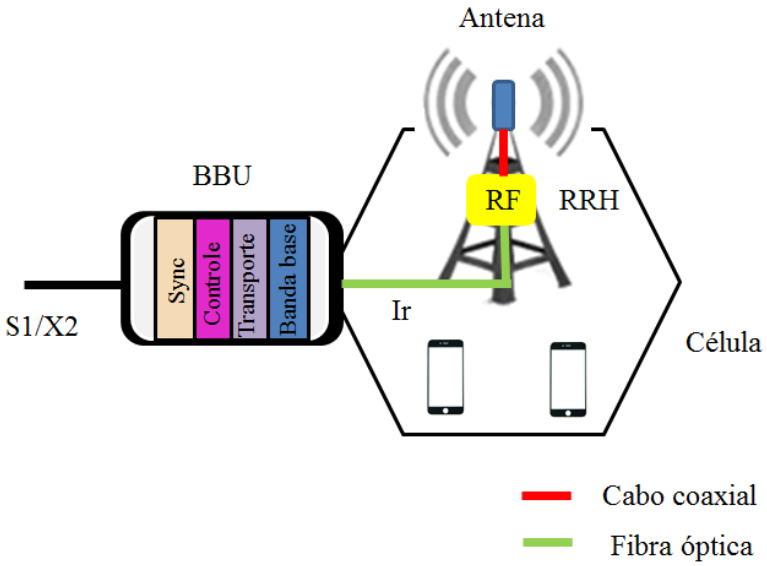
1.1 Evolução da Arquitetura de Rede RAN

Em redes celulares, os usuários se comunicam com a BS, que atende a célula em uma determinada área de cobertura na qual ela está localizada. Na arquitetura RAN tradicional, ilustrada na Figura 1, a RRH e a BBU estão integradas em um único elemento físico na BS. Para essa arquitetura, a conexão entre a antena e a BS é implementada com cabos coaxiais e a interface lógica opcional X2 é definida entre estações rádio base, enquanto a interface S1 conecta os dispositivos móveis à rede celular.

Uma evolução da RAN tradicional é separar a unidade de rádio da unidade de banda base. Essa arquitetura, indicada na Figura 2, contém uma unidade de rádio remota (*Remote Radio Head* - RRH) separada de uma unidade de banda base (*Band Base Unit* - BBU). As RRHs podem ser colocadas em postes ou telhados, minimizando o tamanho dos bastidores nos quais as BBUs são instaladas. Isso também permite que as RRHs aproveitem a refrigeração natural de forma eficiente, proporcionando economia com despesas de ar condicionado nos locais das BBUs. Isso ocorre pois há menos aquecimento nos bastidores com os equipamentos de BBU, em relação aos equipamentos de RRH. A estação base e a unidade de rádio são conectadas por uma interface óptica, denominada Ir.

A C-RAN, mostrada na Figura 3, evolui a arquitetura anterior, centralizando as funções das BBUs em uma nuvem, tornando eficiente a utilização de recursos. Para utilizar as unidades de processamento de banda base, de modo a proporcionar seu melhor

Figura 2 - Arquitetura RAN com RRHs - distribuída.

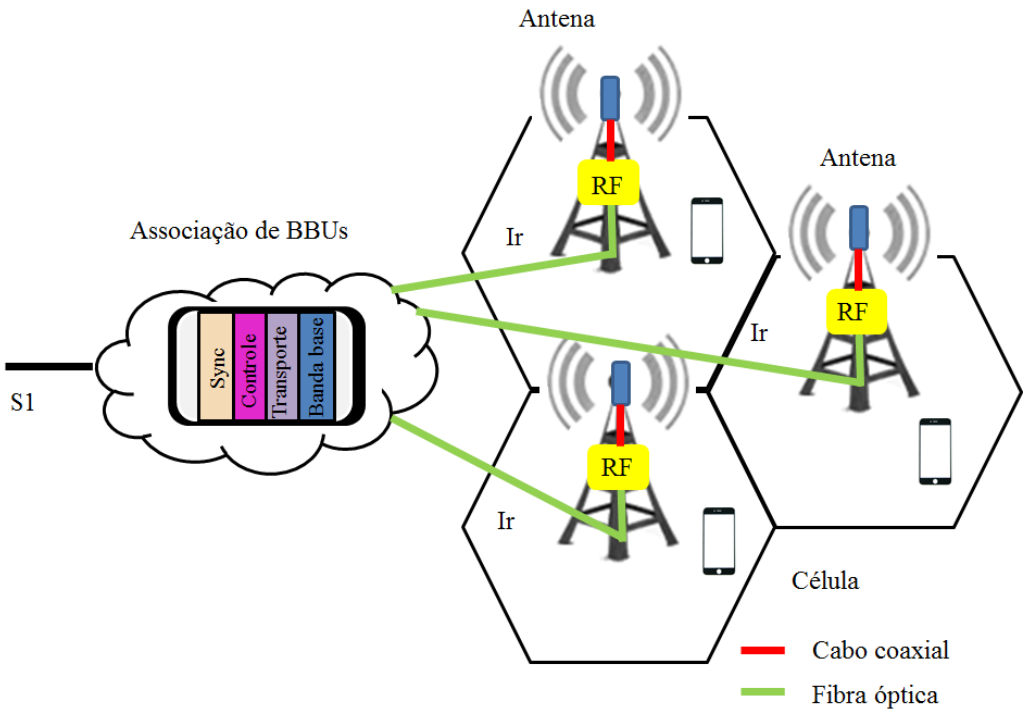


Fonte: Adaptado de (CHECKO et al., 2015).

aproveitamento em questão de capacidade, faz-se necessária sua associação com outras BBUs para que várias RRHs utilizem uma mesma nuvem com BBUs. Assim, evita-se o desperdício de processamento, como ocorre com a arquitetura RAN tradicional. Nessa arquitetura, as funcionalidades RAN de uma estação base são parcialmente ou completamente implementadas em *software*. A fim de se adaptar a vários parâmetros de transmissão, as RRHs devem possuir recursos reconfiguráveis, que podem ser controlados pelos mecanismos de *Cloud-RAN* (BEYENE; JÄNTTI; RUTTIK, 2014). A ideia de separar as RRHs de suas respectivas BBUs é então uma solução atrativa de baixo custo para a realidade do mercado atual, por permitir concentração de processamento de banda base de várias BSs em uma nuvem.

Em C-RANs, as RRHs são conectadas às nuvens de BBUs por meio de fibras ópticas, sendo que as RRHs podem utilizar dinamicamente as BBUs. Assim, usuários de diferentes células podem usar o serviço fornecido por uma BBU, melhorando a utilização de recurso de banda base (WANG; ZHAO; ZHOU, 2014). A associação de BBUs pode ser centralizada ou, como no caso da *Dynamic* C-RAN mostrada mais adiante, pode ser distribuída em diversas nuvens. A centralização de BBUs também facilita a comunicação inter-célula. Como as BBUs de diferentes células estão próximas, elas podem interagir entre si com atrasos mais baixos, facilitando os métodos para implementação do balanceamento de carga entre as células que compartilham de uma mesma infraestrutura (DEMESTICHAS et al., 2013). A separação das BBUs e das unidades de rádio na C-RAN também permite técnicas centralizadas de gerenciamento, possibilitando que as

Figura 3 - Arquitetura C-RAN com RRHs.



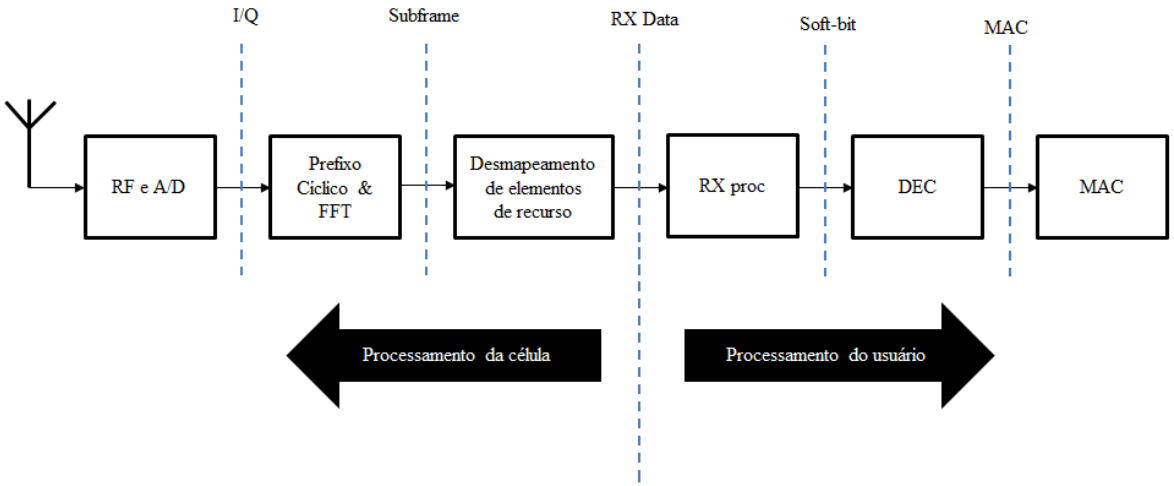
Fonte: Adaptado de (CHECKO et al., 2015).

BBUs na associação possam cooperar para melhorar a capacidade da RAN. Esse gerenciamento centralizado depende da implementação eficiente de algoritmos, que deve lidar com desafios como a latência na comunicação e eficiência do controle de recursos e da capacidade de *fronthaul*, que é uma interface entre a RRH e as BBUs.

Além da redução de custos, a C-RAN propõe também reduzir o consumo de energia, reagir a mudanças no tráfego de dados de usuário e padrões de mobilidade, além de permitir uma articulação sofisticada no processamento de sinais de rádio e um aumento drástico da eficiência espectral (ALYAFAWI et al., 2015).

A BBU pode oferecer os serviços como um só bloco ou divididos em diversos blocos. Nessa última opção, as funcionalidades da BBU são decompostas em diversas funções, que podem ser VNFs de uma infraestrutura NFV. A NFV fornece uma nova abordagem de utilização de recursos de rede que desmembra as funções de rede do *hardware* e permite serviços mais adaptáveis às solicitações dos usuários (LIN et al., 2015). Diversos trabalhos mostram os benefícios de dividir as funções da BBU (WUBBEN et al., 2014; BARTELT et al., 2015). Por exemplo, (BARTELT et al., 2015) mostram que a divisão em funções possibilita reduzir a taxa de dados necessária na rede. No trabalho (DALLA-COSTA et al., 2017b), a proposta abordada processa cinco funções de rádio, sendo elas: MAC, *Soft-bit*, *RX Data*, *Subframe* e *I/Q*, conforme proposto em (WUBBEN et al., 2014) e ilustrado na Figura 4. A divisão de funções de rádio em NFVs faz-se necessária para orquestração

Figura 4 - A divisão de funções de rádio em NFVs.



Fonte: Adaptado de (WUBBEN et al., 2014).

e posicionamento das funções dentro da hierarquia de nuvens. Mais detalhes sobre essas funções estão contidos em (WUBBEN et al., 2014).

Apesar de as BBUs e RRHs estarem em locais distintos, existe um mapeamento lógico um para um entre as BBUs e RRHs, no qual uma BBU é selecionada dinamicamente para gerar ou receber um sinal de uma RRH. Visto isso, é possível utilizar essa arquitetura para suprir as necessidades de tráfego dos usuários e para economizar energia, devido à heterogeneidade da carga de tráfego (SUNDARESAN et al., 2016). A economia de energia na utilização da C-RAN ocorre porque o número de BBUs é reduzido se comparado a uma rede RAN tradicional, na qual existem BBUs que atendem ociosas ou com baixa utilização. Assim, em momentos nos quais o tráfego é reduzido, parte das BBUs podem ser desligadas sem afetar a cobertura da rede, viabilizando a implantação dessa arquitetura.

A C-RAN permite modificar o protocolo de interface de rádio, diferente do que ocorria nas arquiteturas antigas. Além disso, a centralização facilita a manutenção, visto que novas BBUs podem ser inseridas e atualizadas com facilidade, melhorando assim o fator de escalabilidade da rede. Essa é uma vantagem dessa arquitetura, visto que ela é genérica e pode ser ampliada para qualquer número de RRHs. Assim, um ponto positivo em C-RAN é o fato de a implantação de novas células ser mais simples, já que necessita apenas da instalação de uma nova RRH e não de uma estação base completa. Já o aumento da capacidade da rede pode ocorrer melhorando a associação de BBUs, adicionando *hardware* ou substituindo as BBUs existentes por equipamentos de maior capacidade. Vale ressaltar que há rotinas de *software* que implementam funcionalidades de uma estação base e suportam vários RRHs, com diferentes potências de transmissão. Assim, é possível que uma célula seja mais bem provisionada e atenda os usuários de

acordo com as necessidades e características do cenário de cada região.

Apesar de suas vantagens, a arquitetura centralizada também introduz novas limitações práticas que impactam no desenvolvimento da rede de rádio. A configuração C-RAN exige *links* de alta velocidade entre RRH e BBU. Logo, a *Cloud*-RAN tem de lidar não só com a latência do processamento em servidores da nuvem, mas também com latência e taxa de transferência na interface *fronthaul*. Além disso, as componentes de RAN existentes necessitam de mudanças na arquitetura que aceitem o processamento do sinal e as funções de gerenciamento de recursos.

Devido a todas as vantagens mencionadas neste trabalho, pode-se considerar que a C-RAN não é somente aplicada às redes sem fio já existentes, como também é um elemento essencial para os novos sistemas 5G (CHIH-LIN et al., 2014). Isso ocorre pois a centralização da rede e substituição de *hardware* por *software* são necessárias para o crescimento e melhor gerenciamento das redes de nova geração. Assim, será garantido ao usuário um serviço contínuo e rápido, além de custo inferior para as operadoras, que empregarão de forma mais eficiente os recursos disponíveis na infraestrutura da rede. Em alto nível, essas vantagens resultam em adaptabilidade de serviços, maior escalabilidade, flexibilidade e, principalmente, maior eficiência energética.

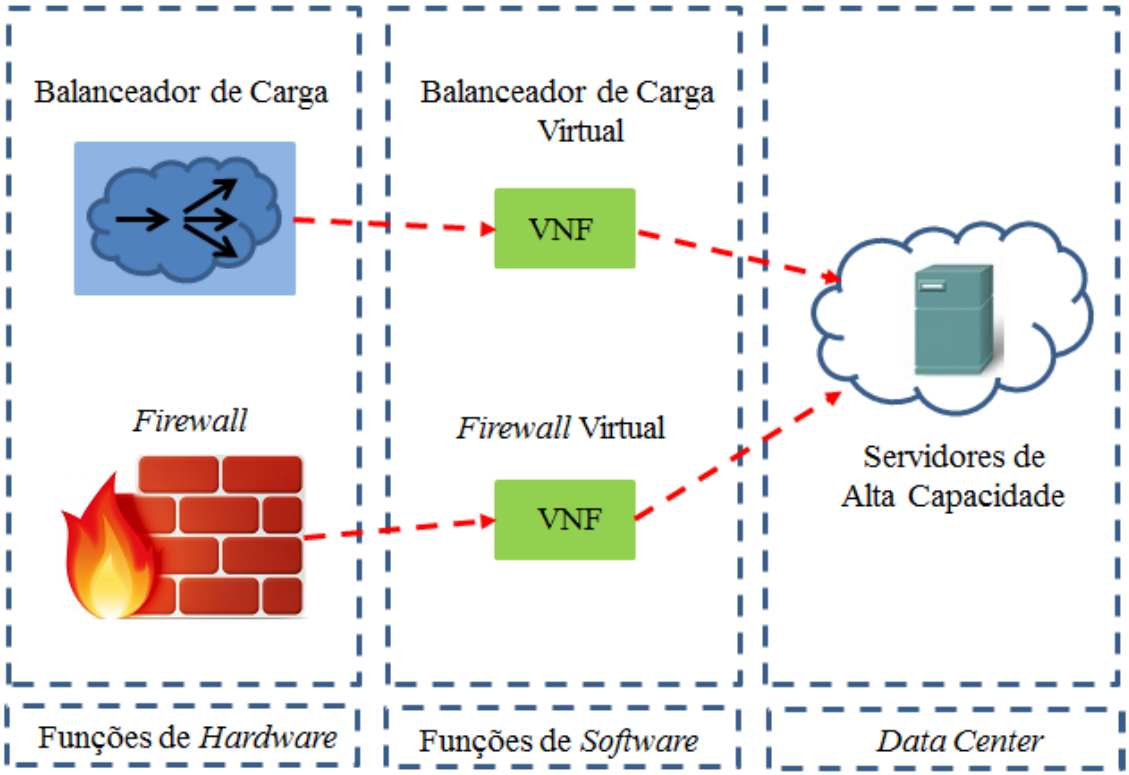
Como já mencionado, o conceito de C-RAN é possível graças à NFV, que é detalhada na próxima seção.

1.2 Gerenciamento de NFV e Orquestração

De modo geral, oferecer aos usuários um novo serviço de rede implica em um processo de instalação, que pode reduzir substancialmente o lucro das operadoras. Isso ocorre pois é necessária a liberação de espaço físico, há custos elevados nas aquisição, instalação e operação de novos equipamentos, além de despesas com energia, devido à grande quantidade de *hardware* envolvida. Além dessas questões, deve-se também levar em consideração a vida útil e o aproveitamento desses equipamentos para suportar novas tecnologias, já que são forçados a receber novos serviços constantemente. Dado o exposto, pode-se dizer que atualizar a infraestrutura da rede tem elevado de forma gradativa os custos das implementações das novas redes. Como agravante, novos serviços necessitam, muitas vezes, de implementação de *hardware* especializado, conhecido como *middlebox*. Para evitar os problemas expostos, foi proposta a NFV, que aproveita a tecnologia de virtualização para, a partir de *hardware* genérico, oferecer uma nova maneira de projetar as redes (HAN et al., 2015).

No conceito de NFV, as *middleboxes* tradicionais são gerenciadas como módulos de *software*, que são programados para atuar como uma Função de Rede Virtual (VNF), permitindo a divisão das funções em partes e isolamento de cada uma dessas funções, para

Figura 5 - Ambiente NFV.



Fonte: Adaptado de (HERRERA; BOTERO, 2016).

que elas possam então ser gerenciadas de forma independente. Além disso, o NFV facilita a instalação e implantação de VNFs em servidores de uso geral, permitindo a migração de VNFs de um servidor para outro, de forma dinâmica (NFV, 2), além de oferecer oportunidades para otimização de rede. Dado o exposto, a NFV permite que as funções de rede, que tradicionalmente usam *hardware* dedicado, como *middleboxes* ou dispositivos específicos de rede por exemplo, passem a ser implementadas em um *software*. Esse *software* é executado em *hardware* genérico, como servidores de alta capacidade (HERRERA; BOTERO, 2016).

A migração de soluções específicas de *hardware* para soluções de *software* vem sendo realizada por meio da consolidação de diferentes tipos de VNFs em servidores de uso geral (localizados em *data centers*, por exemplo), conforme indicado na Figura 5. Assim, uma ou mais máquinas virtuais executam diferentes processos, relacionados às funções de *software*, em servidores de alta capacidade ou em uma infraestrutura de computação em nuvem, ao invés de operar em dispositivos de rede especializados.

Como visto, a NFV aumenta a flexibilidade das implantações e a integração de novos serviços de rede, proporcionando uma maior agilidade dentro das redes das operadoras. Assim obtêm-se reduções significativas nas despesas operacionais e também menores cus-

tos de implantação. Apesar de proporcionar diversas vantagens para o futuro das redes de telecomunicações, um dos principais desafios para a implantação do NFV é a alocação de recursos de serviços de rede na infraestrutura física. Esse desafio é denominado de problema de alocação de recursos do NFV (NFV-RA - *NFV - Resource Allocation*). Vale mencionar que os investimentos futuros na implantação de futuras arquiteturas de rede baseadas em NFV dependerão significativamente do sucesso do NFV-RA, que é o desafio considerado nesta dissertação.

De forma geral, três componentes formam a arquitetura NFV. Como ilustrado na Figura 6, são eles a Infraestrutura de Virtualização de Rede (NFVI - *Network Functions Virtualization Infrastructure*), o Gerenciamento e Orquestração (MANO - *Management and Orchestration*) e Funções Virtuais de Rede (VNFs - *Virtualized Network Functions*), também conhecidas como Serviços:

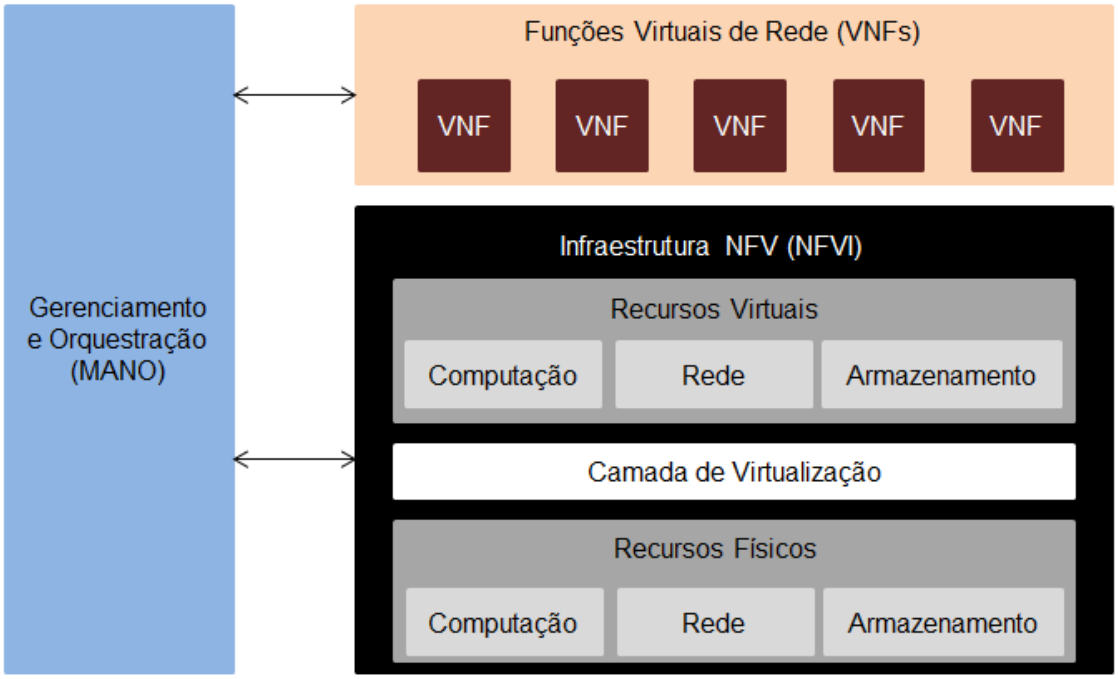
- **VNF:** É a unidade básica de serviço NFV, que pode ser implementada em uma ou várias máquinas virtuais. Há também casos em que as VNFs podem ser executadas diretamente nos sistemas operacionais ou até de forma direta no *hardware*;

- **NFVI:** A infraestrutura NFV cobre todos os recursos de *hardware* e *software* que compõem o ambiente NFV. Os recursos físicos normalmente incluem *hardware* de computação e rede, que fornecem processamento, armazenamento e conectividade para VNFs, por meio da camada de virtualização. Essa camada de virtualização realiza a criação dos recursos virtuais e o desacoplamento das VNFs dos equipamentos, permitindo que as VNFs utilizem a infraestrutura virtual subjacente e forneçam os recursos necessários na execução das funções (ETSI, 2013);

- **MANO:** É composto pelo orquestrador, gerenciador de VNFs e gerenciador de infraestrutura virtualizada. Este componente realiza o gerenciamento de orquestração e ciclo de vida de recursos físicos ou virtuais que suportam a virtualização da infraestrutura, além do gerenciamento do ciclo de vida dos VNFs. É responsável por todas as tarefas de gerenciamento específicas de virtualização necessárias na estrutura NFV.

Todos os blocos funcionais da arquitetura NFV são conectados aos elementos NFV MANO, responsável pelos serviços de gerenciamento e orquestração dos recursos físicos e de *software* que suportam a virtualização da infraestrutura, conforme indicado na Figura 6. Isso ocorre pois ele é o responsável pela conciliação de todos os blocos dessa infraestrutura. Além disso, o MANO possui um componente interno denominado NFVO (Orquestrador NFV), que possui acesso às instâncias quem operam no NFVI, como também aos recursos físicos da NFVI. Com isso, a NFVO é capaz de tomar as decisões sobre os serviços de rede e operação das funções e realizar o melhor posicionamento das funções dentro de uma hierarquia de nuvens, por exemplo, de acordo com as restrições da rede e

Figura 6 - Arquitetura NFV.



Fonte: Adaptado de (ETSI, 2013) e (QUEIROZ; COUTO; SZTAJNBERG, 2017).

da capacidade de processamento dos servidores. Logo, a orquestração dos serviços NFV e as VNFs são fundamentais para o funcionamento dos ambientes baseados em NFV, impactando diretamente no desempenho e controle da rede.

Levando em consideração todas as características citadas, há trabalhos que tiveram suas linhas de pesquisa voltadas para a área de pesquisa de posicionamento de VNFs, como o OpenMANO, que é uma solução de código aberto baseada na implementação da arquitetura exposta acima (MIJUMBI et al., 2016). Esta dissertação, por sua vez, é relacionada ao componente MANO e, mais especificamente, ao seu bloco NFVO. Assim, os algoritmos propostos nesta dissertação devem ser executados como parte do NFVO. Diferente do OpenMANO, focam-se nesta dissertação os algoritmos de posicionamento, e não a alocação real das VNFs na infraestrutura.

1.3 Dynamic C-RAN

A NFV, como exposto anteriormente, permite que as operadoras de telefonia móvel projetem soluções de gerenciamento automatizado em tempo real, facilitando operações como o monitoramento de funções de rede, manutenção e escalabilidade da rede (DALLA-COSTA et al., 2017a). Para proporcionar tais características em C-RAN, é proposto o

conceito de *Dynamic* C-RAN (*Dynamic Cloud Radio Access Network*). Nesse conceito, a infraestrutura é distribuída em diversas nuvens de forma hierárquica, ao invés de uma nuvem central da abordagem original de C-RAN. Isso possibilita, por exemplo, que existam nuvens mais próximas a uma BS, de forma a reduzir a latência na rede. Da mesma forma que exposto anteriormente, as funções de rádio são VNFs. Nesse tipo de arquitetura, as funções de rádio são divididas e seu processamento é feito de forma dinâmica, de acordo com o uso das BSs. Por existirem diversas nuvens, é possível que o processamento seja distribuído em diferentes locais na infraestrutura da rede (DALLA-COSTA et al., 2017b). Consequentemente, a *Dynamic* C-RAN é uma arquitetura de rede sem fio que proporciona agilidade e flexibilidade para redes móveis. Um fator que pode alterar o funcionamento adequado de um ambiente *Dynamic* C-RAN é a mobilidade dos dispositivos móveis na infraestrutura, pois impõe ao NFVO um gerenciamento da rede de acordo com as variações de demanda do usuário. Assim, isso é um fator determinante para o funcionamento adequado da rede. A hierarquia da infraestrutura pode ser composta, por exemplo, por três tipos de nuvens:

- **Nuvem de borda:** É a nuvem que fica mais próxima das BSs e, por isso, possui menor latência das funções de banda base se comparada às demais nuvens contidas na arquitetura;
- **Nuvem regional:** Se comparada à nuvem de borda, possui mais recursos computacionais. No entanto, esse tipo de nuvem está mais distante das BSs, causando uma maior latência das funções, se comparadas às nuvens de borda;
- **Nuvem central:** É a nuvem que tem maior quantidade de recursos para processamento de funções de rádio disponíveis. São as nuvens mais distantes das estações base e, consequentemente, as funções processadas nesse tipo de nuvem sofrem a maior latência da rede.

Se comparada à rede C-RAN tradicional, além de proporcionar mais flexibilidade e adaptabilidade na implantação de serviços, a *Dynamic* C-RAN oferece também diminuição na latência, visto que há a divisão da processamento da funções de rádio em níveis de hierarquia, maior escalabilidade e eficiência energética, dado que menos BBUs são necessárias no processamento de funções de rádio, dependendo da demanda utilizada em determinada região de cobertura. De forma geral, a *Dynamic* C-RAN aproxima as nuvens das BSs, utilizando o conceito de Computação em Névoa, que é um paradigma que estende os recursos computacionais disponíveis em nuvem para a borda da rede (COUTINHO; CARNEIRO; GREVE, 2016). Assim, estende-se a capacidade computacional e o armazenamento da nuvem para os níveis de bordas, permitindo que os dados sejam analisados e transformados em informações ou em ações antes de sua transmissão para

as nuvens em níveis mais altos da hierarquia. O Capítulo seguinte descreve os trabalhos relacionados a esta dissertação.

2 TRABALHOS RELACIONADOS

A fim de posicionar esta dissertação em relação aos trabalhos relacionados disponíveis na literatura, a seguir descrevem-se os trabalhos listados conforme suas contribuições.

O emprego de NFV é essencial para o funcionamento de uma *Dynamic* C-RAN. Assim, é possível mover as funções de rádio para uma nuvem distribuída e hierarquizada, por meio da virtualização dessas funções. Consequentemente, algumas das características de NFV, como migração, escalabilidade e orquestração podem ser utilizadas em cenários dinâmicos de C-RAN (HEIDEKER; KAMIENSKI, 2016) (ABDELWAHAB et al., 2016). Ao utilizar NFV em *Dynamic* C-RAN, é possível constatar dois desafios. O primeiro é posicionar e orquestrar funções virtualizadas na infraestrutura visando otimizar alguma métrica de interesse, como a latência. O segundo é assegurar o desempenho no cálculo desse posicionamento, sem prejudicar a dinamicidade do ambiente. Assim, para a realizar o posicionamento de funções de rádio dentro de uma hierarquia de nuvens, exige-se um algoritmo que deve respeitar restrições, como capacidade das nuvens, tempos de atraso e taxa de dados transmitidos. O posicionamento de VNFs é um assunto bastante estudado na área de NFV (QUEIROZ; COUTO; SZTAJNBERG, 2017; LUIZELLI et al., 2015).

O trabalho (DALLA-COSTA et al., 2017a) propõe a formulação de um problema de programação inteira para um ambiente *Dynamic* C-RAN, capaz de posicionar as funções de rádio em diferentes servidores ao longo de uma hierarquia de nuvens (DALLA-COSTA et al., 2017a). A proposta é um orquestrador, denominado Maestro, e tem como principal objetivo minimizar a taxa de dados transmitida entre as nuvens da hierarquia. O posicionamento proposto pelo Maestro pode então oferecer às operadoras ganhos na transmissão de dados. Esse orquestrador recebe um conjunto de informações sobre a infraestrutura da rede e a sua utilização para que então seja realizado o posicionamento, nas nuvens disponíveis dentro da hierarquia, das funções de rádio pré-estabelecidas. Para tal, foi proposta uma avaliação detalhada do orquestrador Maestro para ambientes sem fio, com o objetivo de estimar as vantagens em aplicar *Dynamic* C-RAN, considerando alguns cenários com diferentes parâmetros, variando a ocupação de cada nuvem composta na hierarquia. Além disso, (DALLA-COSTA et al., 2017a) considera a divisão de funções de uma BBU, assumindo que qualquer divisão validada pode ser considerada no processo de orquestração, como proposto em (ABDELWAHAB et al., 2016) (LIU et al., 2015). Assim como esta dissertação, a proposta do Maestro foca o uso dos elementos de NFV MANO na orquestração de ambientes *Dynamic* C-RAN.

A operação de cada divisão e das próprias funções de rádio são detalhadas em (WUBBEN et al., 2014). Essa divisão de funções dentro dessa hierarquia de nuvens impacta diretamente na taxa de dados transmitida no *fronthaul* e na disponibilidade dos

recursos de computação em nuvem do sistema. Assim, mostra-se em (DALLA-COSTA et al., 2017b) esse impacto por meio de experimentos. Para a avaliação são consideradas diferentes porcentagens de ocupação em cada uma das estações base e diferentes quantidades de recursos computacionais, que variam de acordo com o nível das nuvens.

Como medida de simplificação, (DALLA-COSTA et al., 2017b) considera que cada função de rádio é posicionada consumindo apenas uma unidade de processamento em uma nuvem. Entretanto, foi ressaltado que, em um cenário real, essa demanda de processamento varia com base no percentual de ocupação de cada BS envolvida na rede, visto que essa fator interfere diretamente na taxa de dados transmitida no *fronthaul*. Assim, para a realização dos experimentos e extração dos resultados e análises, foram estimadas porcentagens de ocupação para cada uma das estações base. Além disso, para todos os cenários foram considerados três níveis hierárquicos de nuvem com capacidades de processamento variadas para suportar diversas cargas de trabalho em diferentes cenários, assim como realizado neste trabalho. (DALLA-COSTA et al., 2017b) considera especificações de uma rede uma rede celular do 3GPP (*3rd Generation Partnership Project*). A partir da comparação do desempenho no posicionamento das VNFs nas estações base, com ocupação fixa e variável em ambientes *Dynamic* C-RAN, o trabalho conclui que a transmissão de dados no *fronthaul* é minimizada pelo Maestro. Além disso, a quantidade de recursos utilizados nas nuvens de borda é maximizada, priorizando o uso dos recursos computacionais nas nuvens de borda, que possuem menor latência com as BSs.

Diferente da abordagem feita nesta dissertação, o trabalho (DALLA-COSTA et al., 2017b) otimiza a banda no *fronthaul* da infraestrutura de rede. Já, nesta dissertação, otimiza-se a latência média e não a banda, considerando que os enlaces utilizados pelas operadoras são bem provisionados, sendo portanto a latência um fator mais crítico. Além disso, em (DALLA-COSTA et al., 2017b) só é proposta a solução ótima, não sendo propostas heurísticas para prover escalabilidade, ou seja, para possibilitar a obtenção da solução para problemas maiores. Com o aumento do número de estações base na rede, há um consequente o aumento da complexidade da rede, e mais elementos deverão ser levados em consideração no nos algoritmos de posicionamento de VNFs. Esse aumento no número de BSs tende a aumentar também o tempo de execução do algoritmo orquestrador à medida que a rede cresce. Com isso, faz-se necessária uma heurística para posicionar de forma rápida as funções, como é realizado neste trabalho.

Além do trabalho (DALLA-COSTA et al., 2017b), que foi a base desta dissertação, há alguns trabalhos na área de orquestração em NFV que se destacam no que diz respeito à solução de orquestração. Um deles (MAKAYA et al., 2015) propõe uma plataforma aberta baseada em nuvem que suporta uma API (*Application Programming Interface*) e outros componentes do *framework* ETSI MANO (*European Telecommunications Standards Institute*), que é uma abstração da arquitetura NFV.

Há muitos trabalhos focados no posicionamento de funções virtualizadas. Em (LUI-

ZELLI et al., 2015) foi formalizado o problema de posicionamento de VNFs, sendo proposto um modelo de Programação Linear Inteira (ILP - *Integer Linear Programming*). Para lidar com grandes infraestruturas, foi proposta uma heurística para solucionar o problema. Considerando cargas de trabalho realistas e diferentes cenários, os resultados mostraram que o modelo proposto leva a uma redução de até 25% nos atrasos fim-a-fim, se comparado às infraestruturas tradicionais. Além disso, foi demonstrado que as heurísticas propostas para infraestruturas maiores continuam a encontrar soluções que são muito próximas da otimização, mas sendo executadas em tempo hábil. As métricas otimizadas nesse trabalho são a latência do caminho fim-a-fim e os atrasos de processamento.

Em (MOHAMMADKHAN et al., 2015) é formulado o problema do posicionamento e roteamento de funções de rede como um problema de programação linear inteiro mista (MILP - *Mixed Integer Linear Programming*) que, além de determinar o posicionamento de serviços, busca a minimização da utilização dos recursos, sendo propostas também heurísticas para solução do problema. Já (LIN et al., 2015) apresenta um projeto que implementa de forma ótima as funções de rede e aloca recursos físicos, atendendo às solicitações fim-a-fim de serviços. Isso é realizado a partir da proposta de um MILP que identifica os nós físicos a serem utilizados pelas funções de rede e gera rotas entre essas funções. O posicionamento das funções é então realizado para minimizar os custos de roteamento e de alocação de VNF.

Em (RIGGIO et al., 2016) é proposto um problema de posicionamento de VNFs para redes sem fio como um ILP. É proposta também a WiNE, uma heurística de posicionamento de VNFs para resolver o problema, buscando minimizar a distância entre as VNFs, fazendo com que as VNFs de uma requisição estejam próximas umas das outras. É apresentada também a implementação de uma estrutura de gerenciamento e orquestração de NFV para redes WLAN (*Wireless Local Area Network* - Redes Sem-fio Locais) empresariais. De forma genérica, há outros trabalhos que também levam em conta o posicionamento de funções virtualizadas, como o T-NOVA (XILOURIS et al., 2014) e também o Cloud4NFV (SOARES et al., 2014), que propõem NFVOs para provisionamento de VNF com focos nos serviços fim-a-fim.

Para gerenciar infraestruturas virtuais relacionadas ao posicionamento de NFVs, é necessário introduzir a orquestração de sistemas de alto nível. Com base nisso, em (CLAYMAN et al., 2014) é abordado um problema de gerenciar ambientes de rede e serviços altamente dinâmicos, nos quais nós virtuais e *links* virtuais são criados de acordo com o volume de tráfego e solicitações de serviço. Nele, foi proposta uma arquitetura baseada em um orquestrador que assegura a alocação automática dos nós virtuais e a alocação de serviços de rede sobre eles. Esse orquestrador é suportado por um sistema de monitoramento que analisa o comportamento dos recursos e administra a criação e remoção dos nós virtuais, realizando também a configuração, monitoramento e execução

do *software* sobre cada um deles. Além disso, foi projetado, implementado e testado um protótipo de orquestrador distribuído com resultados de diferentes algoritmos de posicionamento. A orquestração de funções virtuais pode ser estudada também em outras áreas, como em SDN (*Software Defined Networking*). No trabalho (MUÑOZ et al., 2015) é proposto um orquestrador para SDN baseado na migração de controladores SDN virtuais em infraestruturas heterogêneas.

Apesar de os trabalhos descritos anteriormente se concentrarem na orquestração NFV, poucos trabalhos consideram o ambiente C-RAN. Por exemplo, a existência de uma hierarquia de nuvens não é considerada nos trabalhos de NFV, sendo consideradas redes mais genéricas. A utilização de redes genéricas pode tornar mais complexa a execução do problema. Assim, baseou-se esta dissertação no trabalho do orquestrador Maestro (DALLA-COSTA et al., 2017a) para *Dynamic* C-RANs, considerando também três níveis hierárquicos de nuvem com capacidades de processamento diferentes, para suportar diversas cargas de trabalho. Como já exposto, diferente do Maestro, considera-se nesta dissertação a latência média para a tomada de decisão no posicionamento das funções de rádio, além de serem propostas heurísticas para solução do problema.

3 MODELAGEM DO PROBLEMA

No problema proposto, considera-se que, para o funcionamento de uma BS, todas suas funções $f \in \mathcal{F}$ devem ser posicionadas na infraestrutura C-RAN. Cada função pode ser implementada como uma VNF em uma infraestrutura NFV (MIJUMBI et al., 2015). Cada VNF consome, na nuvem que está associada, uma quantidade de banda e uma fatia de seus recursos destinados à hospedagem de VNFs, como memória e processamento.

O problema formulado tem como objetivo minimizar a latência percebida por uma BS. Cada uma das funções de uma BS é executada em uma determinada nuvem, em um determinado nível da hierarquia. A hierarquia das nuvens é uma árvore, como mostram os exemplos das Figuras 7, 8 e 9. Nesses exemplos existem três níveis de nuvem. Quanto mais baixo o nível da nuvem, mais baixa é sua latência para a BS. Além disso, espera-se que, quanto mais baixo o nível, menor a capacidade de processamento e banda da nuvem. Devido à estrutura em árvore, cada BS só pode ser atendida por uma nuvem em cada nível. Por exemplo, na Figura 9 a BS4 pode ser atendida pelas nuvens n2, n5 e n6.

Define-se, neste trabalho, a latência l_a percebida por uma BS a como a latência dessa BS até a função mais distante possível. Por exemplo, na Figura 7 se a BS0 possuir uma função em n0, uma em n1 e outra em n2, a latência l_a percebida por essa BS é a sua latência do seu caminho até a nuvem $n2$. Se a BS só possuísse funções em n0 e n1, sua latência l_a seria a latência até nuvem $n1$.

Dado o exposto, o problema formulado visa minimizar a média entre as latências l_a de todas as BSs da infraestrutura. A Tabela 1 lista as variáveis, parâmetros e conjuntos utilizados na formulação do problema, bem como as notações utilizadas e suas descrições. A formulação do MILP é apresentada a seguir.

Tabela 1 - Notações utilizadas no problema.

Notação	**Descrição**	**Tipo**
$\mathcal{N}$	Nuvens existentes na infraestrutura	Conjunto
$\mathcal{A}$	BSs existentes na infraestrutura	Conjunto
$\mathcal{F}$	Tipos de funções de rede	Conjunto
$\mathcal{D}_a$	Nuvens que podem receber funções da BS a	Conjunto
C_V^n	Capacidade em VDUs da nuvem n	Parâmetro
C_B^n	Capacidade de banda do enlace de saída da nuvem n	Parâmetro
$D_V(f,n,a)$	Demanda de VDUs da função f, da BS a, na nuvem n	Parâmetro
$D_B(f,n,a)$	Demanda de banda da função f, da BS a, na nuvem n	Parâmetro
$D_L(n,a)$	Latência entre a nuvem n e a BS a	Parâmetro
l_{max}	Latência máxima permitida na C-RAN	Parâmetro
$d_{f,n,a}$	Variável binária que indica se a função f, da BS a, deve ser posicionada na nuvem n	Variável
l_a	Latência percebida pela BS a	Variável

Figura 7 - Topologia 1.

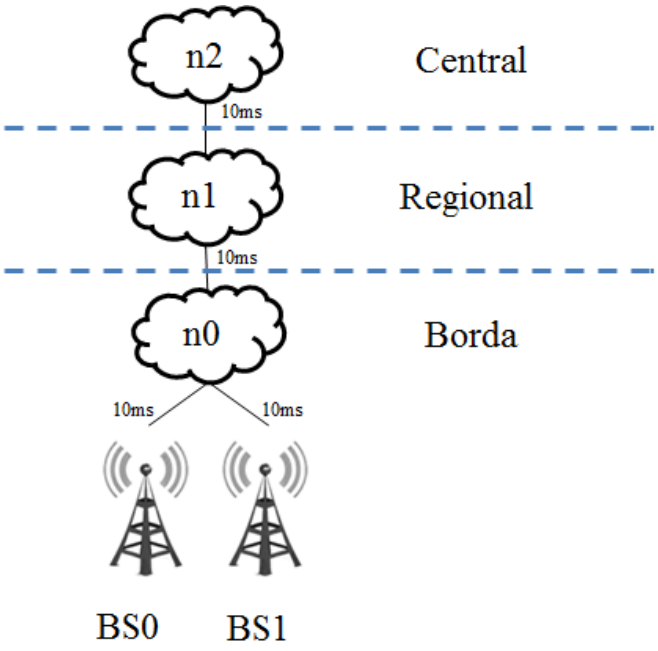


Figura 8 - Topologia 2.

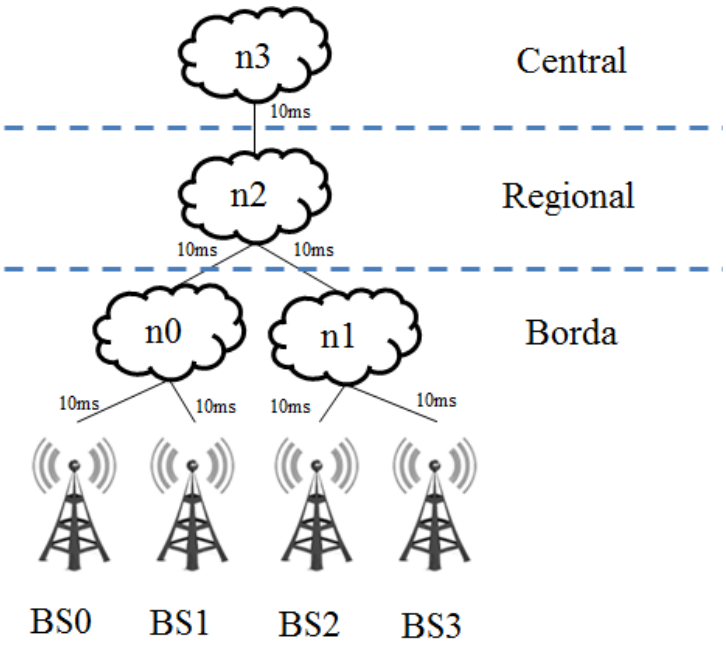
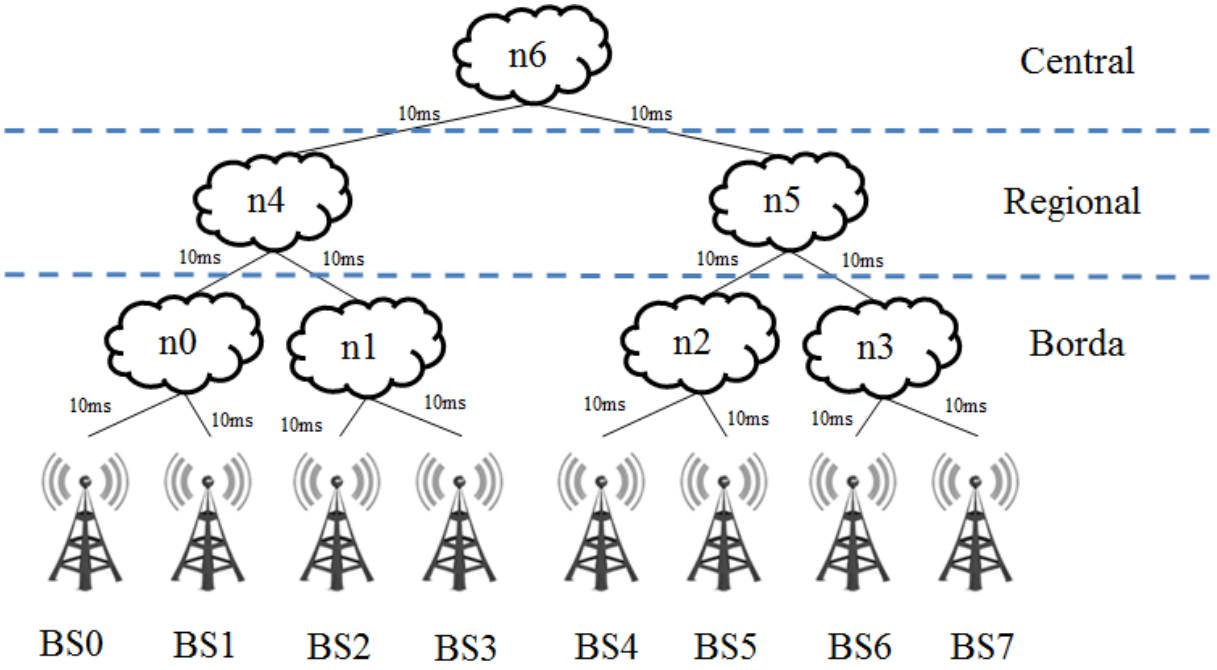


Figura 9 - Topologia 3.



$$\text{minimizar:} \quad \frac{1}{|\mathcal{A}|} \sum_{\forall a \in \mathcal{A}} l_a. \tag{1}$$

$$\text{sujeito a:} \quad \sum_{n \in \mathcal{D}_a} d_{f,n,a} = 1 \qquad \forall f \in \mathcal{F}, \forall a \in \mathcal{A}; \tag{2}$$

$$\sum_{f \in \mathcal{F}, a \in \mathcal{A} | n \in \mathcal{D}_a} d_{f,n,a} \cdot D_V(f,n,a) \leq C_V^n \quad \forall n \in \mathcal{N}; \tag{3}$$

$$\sum_{f \in \mathcal{F}, a \in \mathcal{A} | n \in \mathcal{D}_a} d_{f,n,a} \cdot D_B(f,n,a) \leq C_B^n \quad \forall n \in \mathcal{N}; \tag{4}$$

$$l_a \geq d_{f,n,a} \cdot D_L(n,a) \quad \forall f \in \mathcal{F}, a \in \mathcal{A}, n \in \mathcal{D}_a; \tag{5}$$

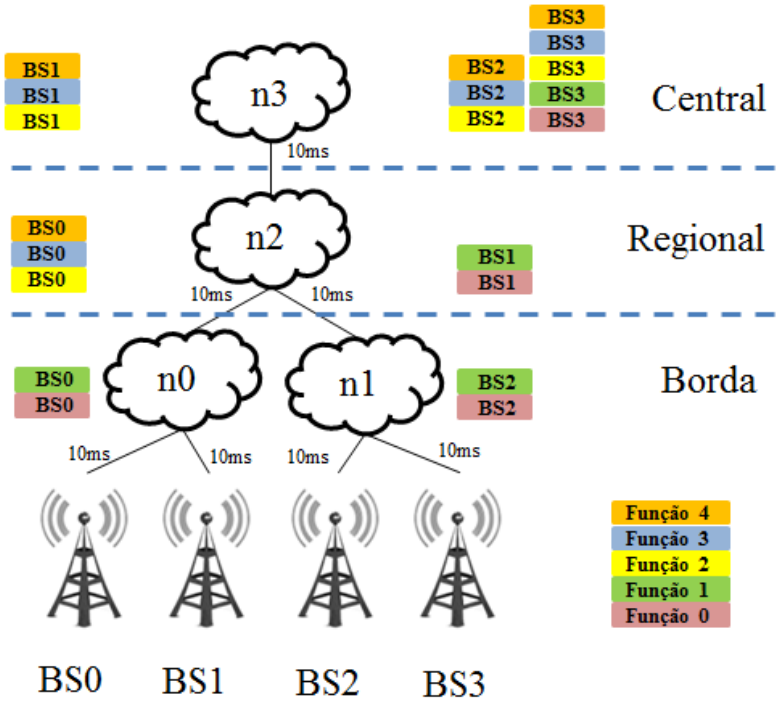
$$l_a \leq l_{max} \quad \forall a \in \mathcal{A}; \tag{6}$$

$$l_a \in \mathbb{R}^+ \quad ; \quad d_{f,n,a} \in \{0,1\}. \tag{7}$$

A Equação 1 é a função objetivo, que consiste em minimizar a latência média, considerando todas as latências l_a de todas as BSs da infraestrutura. Note que $|\mathcal{A}|$ é o número total de BSs da infraestrutura.

A Equação 2 define que cada função f de uma BS a só pode ser posicionada em uma única nuvem n. Note que $\mathcal{D}_a$ é o conjunto das nuvens que podem atender a BS a. Por exemplo, na Figura 9, o $\mathcal{D}_1$ correspondente à BS1 (isto é, $a = 1$) possui as nuvens n0, n4 e n6. A variável de decisão $d_{f,n,a}$ é a saída do problema, que indica se a função f, da

Figura 10 - Exemplo de posicionamento de funções de rádio



BS a, está posicionada na nuvem n.

As Equações 3 e 4 asseguram que o posicionamento de funções respeita, respectivamente, as capacidades de hospedagem de funções e banda de cada nuvem. Neste trabalho, seguindo a mesma nomenclatura de (DALLA-COSTA et al., 2017b), a capacidade de hospedagem de uma nuvem n é medida em número de VDUs (*Virtual Data Units*), dado por C_V^n. Uma VDU é uma fatia indivisível da nuvem, considerando suas capacidades de memória e processamento. Uma função de rede ocupa então um determinado número de VDUs. Assim, o parâmetro $D_V(f, n, a)$ consiste na quantidade de VDUs necessárias para hospedar a função f, da BS a, na nuvem n. Em relação à banda, considera-se como C_B^n a capacidade do enlace de saída da nuvem n para a nuvem hierarquicamente superior. No caso da nuvem central, essa capacidade pode ser considerada como infinita ou mesmo igual à capacidade do enlace de conexão da RAN com outras redes, como o núcleo da rede celular. O parâmetro $D_B(f, n, a)$ é a quantidade de banda que a função f, da BS a, consome no enlace de saída da nuvem n. Por simplicidade, é possível considerar que os parâmetros $D_V(f, n, a)$ e $D_B(f, n, a)$ são independentes de n e a, como realizado mais adiante nesta dissertação.

A Equação 5 calcula a variável l_a para cada BS a, considerando os valores de latência $D_L(n, a)$ entre a nuvem n e a BS a. Note que essa equação força l_a a ser maior ou igual à maior latência entre a BS a e uma nuvem n que hospeda alguma de suas funções.

Entretanto, na solução ótima, l_a será exatamente igual a essa latência mencionada, visto que a função objetivo da Equação 1 minimiza os valores de l_a, atribuindo-a o menor valor possível. A Equação 6 impede o posicionamento de funções em nuvens que possuam uma latência com valor maior do que o máximo permitido, dado por L_{max}. Por fim, a Equação 7 descreve o domínio das variáveis.

Na Figura 10 é possível observar um exemplo de posicionamento ótimo de funções de rádio em uma arquitetura de nuvens para a Topologia 2, ilustrada na Figura 8, descrita mais adiante nesta dissertação. No exemplo dado, a capacidade das nuvens de borda está limitada em 2 VDUs, enquanto a capacidade das nuvens regionais é limitada em 5 VDUs, enquanto a nuvem central possui capacidade infinita de processamento. Este é um caso de rede homogênea, na qual todas as latências dos enlaces tem 10 ms. Nesse exemplo, é possível verificar que a l_a percebida pela BS0 é de 20 ms, dado que foi necessário o posicionamento das funções de rádio da nuvem de borda até a nuvem regional. Já para as BS1, BS2 e BS3, fez-se necessário o posicionamento das funções da nuvem de borda até a nuvem central, logo a l_a percebida por cada uma destas BSs é de 30 ms. Dado o exposto, pode-se notar que a latência média (isto é, o somatório dos l_a dividido pelo número de BSs) para essa topologia, no cenário considerado, é de 27,5 ms.

Por se tratar de um problema com variáveis binárias, a solução por meio de um otimizador MILP pode não escalar com o número de nuvens e BSs. Além disso, o problema deve ser executado periodicamente para agir de acordo com o aumento da demanda das BSs. Consequentemente, sua solução deve ser realizada o mais rápido possível. Para tal, propõem-se duas heurísticas para a solução do problema, descritas a seguir.

4 HEURÍSTICA PARA C-RANS HOMOGÊNEAS

A ideia da heurística proposta é, para cada BS, tentar posicionar todas as suas funções na sua nuvem de borda. Caso não seja possível por falta de capacidade de banda ou de VDU, tenta-se posicionar as funções restantes na nuvem regional. Caso ainda não seja possível, posicionam-se as funções na nuvem central. A cada posicionamento realizado, subtraem-se as demandas $D_V(f, n, a)$ e $D_B(f, n, a)$ dos parâmetros de capacidade C_V^n e C_B^n da nuvem escolhida. Quando todas as funções de uma BS são posicionadas, atende-se uma próxima BS, posicionando suas funções nas nuvens da infraestrutura. Essa heurística considera que a rede é homogênea, ou seja, cada nuvem possui os mesmos valores de latência para todas as BSs que ela pode atender. Nesse caso, o parâmetro $D_L(n, a)$ só depende do nível da nuvem n na hierarquia e não da BS utilizada. Essa consideração permite escolher arbitrariamente a ordem na qual as BSs são atendidas pela heurística, sem priorizar BSs com menores latências. Mais adiante, no Capítulo 5, propõe-se uma heurística que considera redes heterogêneas em relação à latência dos enlaces. A seguir descreve-se formalmente o algoritmo da heurística proposta para redes homogêneas.

4.1 Descrição do Algoritmo

O Algoritmo 1 detalha a heurística para redes homogêneas. A linha 1 itera para todas as BS, enquanto a linha 2 itera para todos os tipos de função. Na linha 3 itera-se entre as nuvens possíveis para a BS a. Por definição, o conjunto $\mathcal{D}_a$ é ordenado da nuvem de menor hierarquia até a nuvem de maior hierarquia. Assim, para três níveis de hierarquia, o algoritmo tenta posicionar a função inicialmente na nuvem de borda, em seguida a regional e finalmente a central. A linha 4 verifica se a função f já foi posicionada em alguma nuvem. Para tal, utiliza-se a variável auxiliar $p_{f,a}$, que indica se a função f da BS a já foi posicionada. Se a função não estiver posicionada, a linha 5 verifica se há VDUs e banda disponíveis na nuvem n e se a latência para essa nuvem é menor ou igual ao maior valor de latência permitido l_{max}. Caso isso seja afirmativo, as linhas 6 e 7 atualizam as variáveis $p_{f,a}$ e $d_{f,n,a}$. Finalmente, nas linhas 8 e 9 são decrementadas da capacidade daquela nuvem a demanda de VDUs e de banda da função posicionada.

Utilizando as variáveis calculadas pelo Algoritmo 1, é possível obter a latência média (isto é, a função objetivo da Equação 1) fazendo:

$$l_{med} = \frac{1}{|\mathcal{A}|} \sum_{\forall a \in \mathcal{A}} l_a = \frac{1}{|\mathcal{A}|} \sum_{\forall a \in \mathcal{A}} \max_{\forall f \in \mathcal{F}, n \in \mathcal{D}_a} (d_{f,n,a} \cdot D_L(n, a)) \tag{8}$$

O Algoritmo 1 possui número de passos proporcional ao número de BSs, dado por

Algoritmo 1: Heurística para Dynamic C-RANs Homogêneas

Entrada: $\mathcal{A}$, $\mathcal{F}$, $\mathcal{D}_a$, C_V^n, C_B^n, $D_V(f,n,a)$, $D_B(f,n,a)$, $D_L(n,a)$, l_{max}
Saída: $d_{f,n,a}$

```
para a ∈ A faça
    para f ∈ F faça
        para n ∈ D_a faça
            se p_{f,a} = 0 então
                se D_V(f,n,a) ≤ C_V^n e D_B(f,n,a) ≤ C_B^n e D_L(n,a) < l_max então
                    p_{f,a} ← 1;
                    d_{f,n,a} ← 1;
                    C_V^n ← C_V^n − D_V(f,n,a);
                    C_B^n ← C_B^n − D_B(f,n,a);
                fim
            fim
        fim
    fim
fim
```

$|\mathcal{A}|$, número de tipos de função, dado por $|\mathcal{F}|$, e número de nuvens que atendem uma BS, dado por $|\mathcal{D}_a|$. Para $|\mathcal{D}_a|$ assume-se que todas as BSs podem ser atendidas pelo mesmo número de nuvens, o que sempre ocorre em topologias em árvore perfeitamente balanceadas, consideradas neste trabalho. Assim, utilizando a notação O para análise de pior caso, tem-se que o algoritmo possui complexidade de tempo $O(|\mathcal{A}||\mathcal{F}||\mathcal{D}_a|)$. Entretanto, o número de tipos de função $|\mathcal{F}|$ é constante para uma determinada tecnologia de comunicação celular, sendo cinco para LTE na abordagem de (WUBBEN et al., 2014). Da mesma forma, $|\mathcal{D}_a|$ é constante para uma determinada arquitetura C-RAN, sendo igual a três nos trabalhos da literatura (DALLA-COSTA et al., 2017b; DALLA-COSTA et al., 2017a). Assim, utilizando as propriedades da notação O, removem-se as constantes da expressão de complexidade, concluindo que a heurística proposta possui complexidade $O(|\mathcal{A}|)$.

4.2 Resultados

Esta seção avalia a heurística proposta em comparação com a solução ótima. Para a solução ótima, executa-se o problema do Capítulo 3 utilizando o GLPK (GNU, 2017) com *scripts* em Python. Para avaliar a heurística, implementa-se o Algoritmo 1 também em Python.

Para realizar os experimentos e solucionar o problema, são utilizadas três topologias com diferentes números de BSs e nuvens dentro da hierarquia. Para todas as topologias, existe apenas uma nuvem central. A Topologia 1, da Figura 7, possui apenas uma nuvem em cada nível da hierarquia e é semelhante à topologia adotada em (DALLA-COSTA et al., 2017b). A Topologia 2, da Figura 8, contém quatro BSs, duas nuvens de borda e uma

Tabela 2 - Parâmetros utilizados na avaliação.

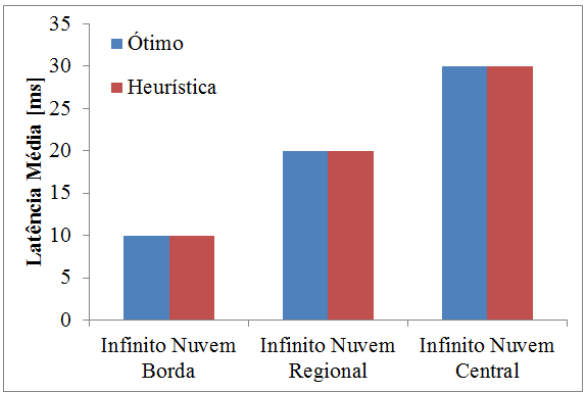
Topologia	Configuração	Capacidade C_V^n (VDUs)			Latência $D_L(n,a)$ (ms)		
		Borda	Regional	Central	Borda	Regional	Central
1	Infinito Nuvem Borda	∞	∞	∞	10	20	30
	Infinito Nuvem Regional	2	∞	∞	10	20	30
	Infinito Nuvem Central	2	2	∞	10	20	30
2	Infinito Nuvem Borda	∞	∞	∞	10	20	30
	Infinito Nuvem Regional	2	∞	∞	10	20	30
	Infinito Nuvem Central	2	2	∞	10	20	30
3	Infinito Nuvem Borda	∞	∞	∞	10	20	30
	Infinito Nuvem Regional	5	∞	∞	10	20	30
	Infinito Nuvem Central	5	5	∞	10	20	30

uma nuvem regional. Já a Topologia 3, da Figura 9, contém oito BSs, quatro nuvens de borda e duas regionais. Em todas as topologias, considera-se que existem cinco tipos de funções (isto é, $|\mathcal{F}| = 5$), como sugerido em (WUBBEN et al., 2014), e cada uma ocupa 1 VDU.

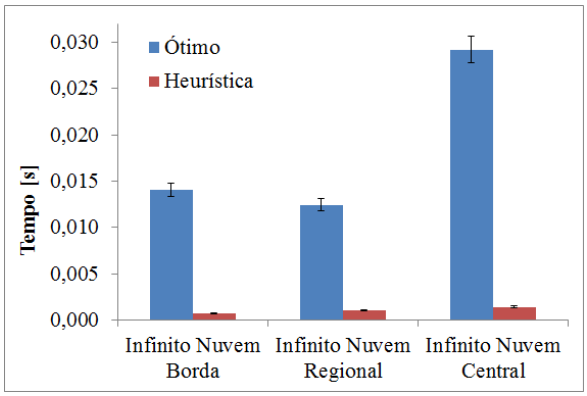
Para cada topologia, são consideradas diferentes configurações dos recursos de banda C_V^n e de VDUs C_B^n de cada nuvem n. Para simplificar a análise, os recursos de banda são mantidos como ilimitados. Os recursos de VDUs são distribuídos, em cada topologia, de acordo com três configurações. Na primeira, denominada Infinito Nuvem de Borda, todas as nuvens possuem capacidade infinita. Na segunda, denominada Infinito Nuvem Regional, as nuvens de borda são limitadas, mas as demais possuem capacidade infinita. Na configuração Infinito Nuvem Central apenas a nuvem central possui capacidade infinita. A Tabela 2 mostra a capacidade em VDUs de cada nuvem utilizada e seus valores de latência para as BSs. Note, pelos valores de latência da Tabela 2, que os enlaces de todas as topologias consideradas possuem 10 ms de latência.

Para a Topologia 1, a Figura 12(a) mostra o valor da função objetivo (isto é, latência média), definida na Equação 8. Os resultados mostram que o valor da função objetivo da solução ótima é igual ao do valor obtido pela heurística. Além disso, é possível observar, como esperado, que a latência aumenta à medida que diminuem-se os recursos das nuvens de menor hierarquia. A Figura 12(b) mostra o tempo de execução da solução ótima e da heurística, representado com nível de confiança de 95% após a coleta de 10 rodadas. É possível notar que, apesar de os tempos serem pequenos para os dois casos, a heurística possui tempos de execução consideravelmente inferiores. É esperado também que, para o caso ótimo, ao reduzirem-se os recursos das nuvens de menor hierarquia, mais possibilidades são acrescentadas ao problema, aumentando o tempo de solução. Por

Figura 11 - Resultados para a Topologia 1 homogênea.

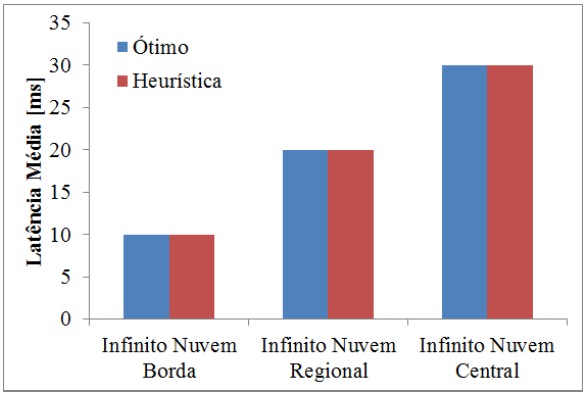


(a) Latência média.

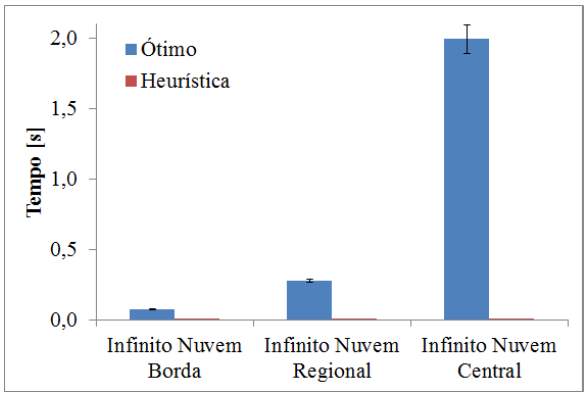


(b) Tempo de execução.

Figura 12 - Resultados para a Topologia 2 homogênea.

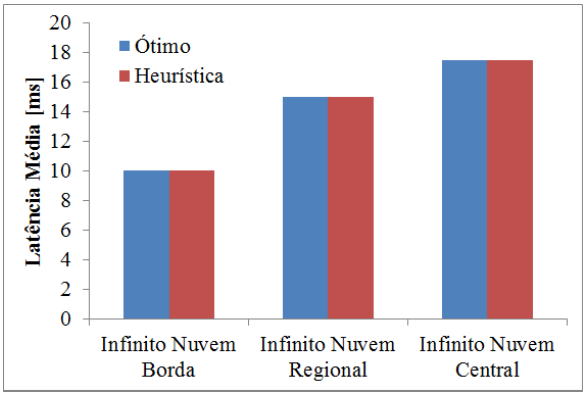


(a) Latência média.

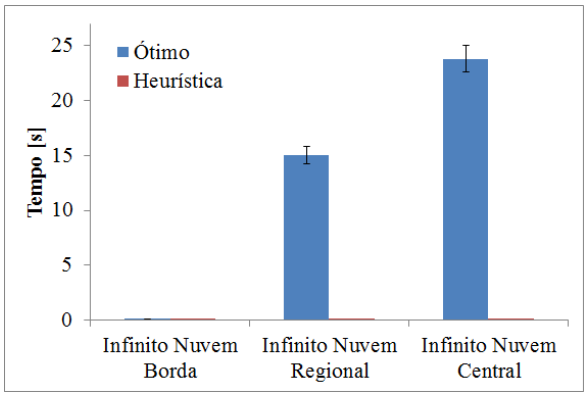


(b) Tempo de execução.

Figura 13 - Resultados para a Topologia 3 homogênea.



(a) Latência média.



(b) Tempo de execução.

exemplo, para o cenário Infinito Nuvem Borda, apenas a nuvem de borda já satisfaz a solução do problema. Entretanto, a Figura 12(b) não mostra esse comportamento entre os cenários Infinito Nuvem Borda e Infinito Nuvem Regional. Isso ocorre devido aos pequenos tempos de execução, que podem acarretar falta de precisão na medição. Esse comportamento está mais evidente nos resultados das Topologias 2 e 3, apresentados a seguir.

As Figuras 12 e 13 apresentam os resultados para as Topologias 2 e 3 respectivamente. Note que, da mesma forma que a Topologia 1, as Figuras 13(a) e 14(a) mostram que a heurística possui a mesma latência média da solução ótima. Nos tempos de execução, apresentados nas Figuras 13(b) e 14(b), é possível observar melhor que limitar os níveis mais baixos da hierarquia aumenta o espaço de solução, aumentando o tempo de execução. Os resultados do tempo de execução também mostram que a heurística obtém a solução de forma significativamente mais rápida, sendo imperceptíveis nos gráficos traçados. Por fim, comparando os tempos obtidos nas Figuras 12(b), 13(b) e 14(b), mostra-se que o tempo de execução da solução ótima aumenta com o número de BSs na topologia.

5 HEURÍSTICA PARA C-RANS HETEROGÊNEAS

Apesar de os resultados da heurística do Capítulo 4 serem iguais aos da solução ótima, isso pode não ocorrer em redes heterogêneas, nas quais uma determinada nuvem possui valores de latência diferentes para as BSs que pode atender (p.ex, na Figura 9 se a nuvem n4 possuir latência de 60 ms com a BS0, mas 20 ms com a BS3). Para efeitos de comparação, foram realizados experimentos para verificar se a ordenação de latências para este tipo de redes interferiria no resultado da solução deste problema.

Dado o exposto, propõe-se neste trabalho uma segunda heurística que aloca as funções em ordem crescente da latência entre uma BS e sua respectiva nuvem. Para tal, ordenam-se todos os possíveis pares (n, a), onde n é uma nuvem e a é uma BS. Após isso, alocam-se as funções na ordem desses pares, privilegiando assim as BSs que possuem menores latências com suas nuvens. A seguir descreve-se o algoritmo da heurística proposta.

5.1 Descrição do Algoritmo

O Algoritmo 2 detalha a heurística para redes heterogêneas. A linha 1 recebe todas as BSs e nuvens e, utilizando os parâmetros $D_L(n, a)$ para cada par (n,a), constrói a lista ordenada. Assim, a linha 2 itera para todos os pares (n,a) na ordem crescente de sua latência. A partir da linha 4, o Algoritmo 2 executa os mesmos passos do Algoritmo 1.

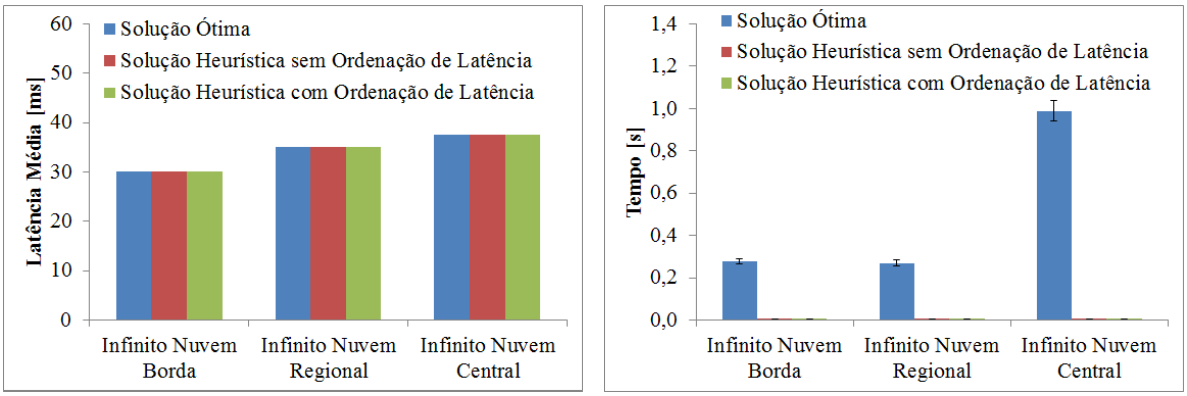
Algoritmo 2: Heurística para Dynamic C-RANs Heterogêneas

Entrada: $\mathcal{A}$, $\mathcal{F}$, $\mathcal{D}_a$, $\mathcal{N}$, C_V^n, C_B^n, $D_V(f,n,a)$, $D_B(f,n,a)$, $D_L(n,a)$, l_{max}
Saída: $d_{f,n,a}$

```
1  listaOrdenada = constroiListaParesOrdenados (A,N);
2  para (n,a) ∈ listaOrdenada faça
3      para f ∈ F faça
4          se p_{f,a} = 0 então
5              se D_V(f,n,a) ≤ C_V^n e D_B(f,n,a) ≤ C_B^n e D_L(n,a) < l_max então
6                  p_{f,a} ← 1;
7                  d_{f,n,a} ← 1;
8                  C_V^n ← C_V^n − D_V(f,n,a);
9                  C_B^n ← C_B^n − D_B(f,n,a);
10             fim
11         fim
12     fim
13 fim
```

A complexidade do conjunto de passos da linha 2 a 13 é a mesma do Algoritmo 1, ou seja, $O(\mathcal{A})$. Entretanto, é necessário considerar também a complexidade da ordenação na linha 1. Considerando um algoritmo simples de ordenação como o *Bubble Sort*, sua

Figura 14 - Resultados para a Topologia 3 heterogênea no Cenário 1.



(a) Latência média.

(b) Tempo de execução.

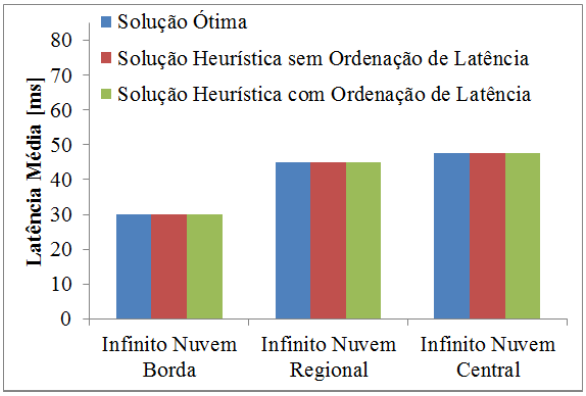
complexidade de pior caso é $O(n^2)$ (SZWARCFITER; MARKENZON, 2013). Assim, como a lista possui $|\mathcal{A}| \cdot |\mathcal{D}_a|$ elementos e $|\mathcal{D}_a|$ é considerado constante, a complexidade do *Bubble Sort* é $O(|\mathcal{A}|^2)$. Consequentemente, a complexidade de todo o Algoritmo 2 é $O(|\mathcal{A}|^2 + |\mathcal{A}|)$. Utilizando as propriedades da notação O, considera-se apenas o termo de maior complexidade. Assim, a complexidade do Algoritmo 2 utilizando *Bubble Sort* é $O(|\mathcal{A}|^2)$.

Apesar do exposto anteriormente, é possível utilizar algoritmos de busca mais sofisticados, como o *Heap Sort*. Esse algoritmo possui complexidade $O(n \log n)$ (SZWARCFITER; MARKENZON, 2013). Desta forma, a complexidade do Algoritmo 2 se torna $O(|\mathcal{A}| \log(|\mathcal{A}|) + |\mathcal{A}|)$. Utilizando as propriedades da notação O, tem-se que Algoritmo 2 com *Heap Sort* possui complexidade de pior caso $O(|\mathcal{A}| \log(|\mathcal{A}|))$.

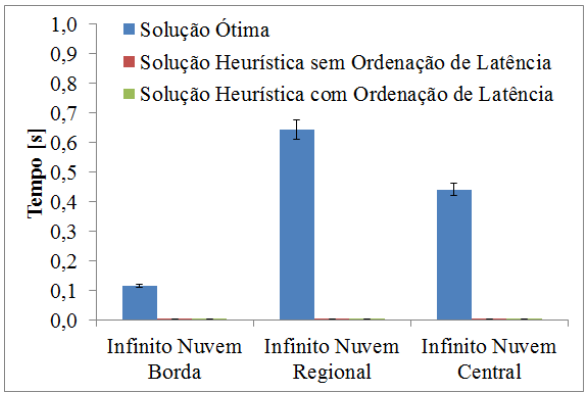
5.2 Resultados

Esta seção analisa os resultados obtidos para a solução ótima, para o Algoritmo 1 e para o Algoritmo 2. Esses resultados são obtidos para a Topologia 3, da Figura 9, com os mesmos parâmetros da Tabela 2. Para essa mesma topologia, são analisados três cenários diferentes em termos de latência. No Cenário 1, todas os enlaces à esquerda de uma nuvem de borda possuem latência de 50 ms e os demais valores de latência são mantidos em 10 ms, como na Tabela 2. Por exemplo, nesse cenário, a nuvem n0 da Figura 9 possui latência de 50 ms com a BS0 e de 10 ms com a BS1. O Cenário 2 engloba os casos do Cenário 1 e considera também que todos os enlaces à esquerda de uma nuvem regional possuem latência de 50 ms. Por exemplo, a latência entre a nuvem n4 e a BS0 é de 100 ms. Já o Cenário 3 engloba o Cenário 2 e considera que o enlace à esquerda da nuvem central possui latência de 50 ms.

Figura 15 - Resultados para a Topologia 3 heterogênea no Cenário 2.

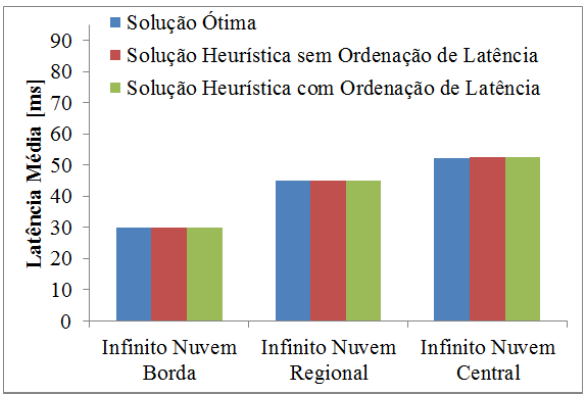


(a) Latência média.

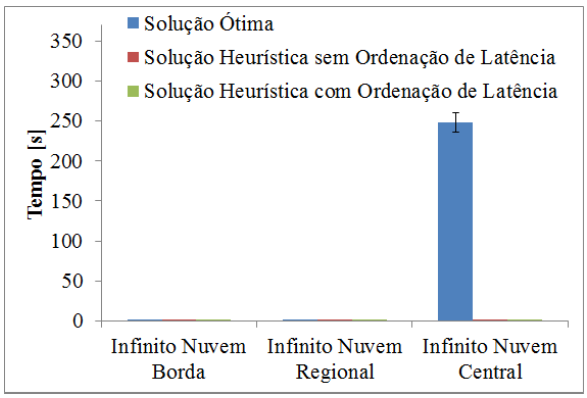


(b) Tempo de execução.

Figura 16 - Resultados para a Topologia 3 heterogênea no Cenário 3.



(a) Latência média.



(b) Tempo de execução.

Tabela 3 - Parâmetros usados na avaliação de Redes Heterogêneas.

Nuvem	**Capacidade C_V^n (VDUs) - Cenário 4**	**Capacidade C_V^n (VDUs) - Cenário 5**
0	2	2
1	2	3
2	3	2
3	3	2
4	4	7
5	7	9
6	∞	∞

Tabela 4 - Latências usadas na avaliação de Redes Heterogêneas para o Cenário 4.

Nuvem / BS	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
0	48	48	-	-	-	-	-	-
1	-	-	1	1	-	-	-	-
2	-	-	-	-	75	75	-	-
3	-	-	-	-	-	-	69	69
4	114	114	67	67	-	-	-	-
5	-	-	-	-	168	168	162	162
6	210	210	163	163	264	264	258	258

A partir da análise dos resultados de todos os cenários, presentes nas Figuras 14, 15 e 16, é possível verificar que a heurística com ordenação (isto é, o Algoritmo 2) possui latência média igual à solução ótima em todos os casos avaliados. No entanto, para os cenários propostos, constatou-se também que a heurística sem ordenação de latências oferece o mesmo resultado do valor ótimo. Para esses cenários avaliados, a ordenação de latências não interferiu na latência média da topologia. Por fim, é possível notar que os tempos de execução da solução ótima são, na maioria dos casos, consideravelmente maiores quando comparados aos tempos de execução das heurísticas. No entanto, se as capacidades das nuvens em VDU e banda também foram heterogêneas, existem casos particulares nos quais a heurística com ordenação possuem latência inferior à da heurística sem ordenação, e há casos nos quais essa situação se inverte. Exemplos desses casos são mostrados a seguir.

5.3 Casos Particulares

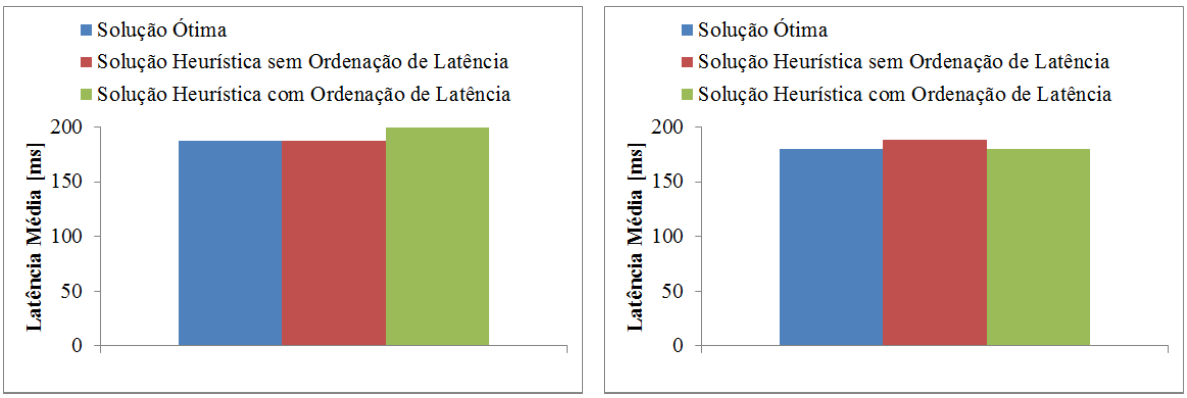
Mesmo que para casos particulares, é possível analisar que a ordenação de latência pode interferir diretamente na latência média da rede. Para esses casos, tem-se também capacidades de processamentos heterogêneas para cada uma das nuvens, e não mais uma quantidade de VDUs igual para as nuvens de mesmo nível. Como indicado na Tabela 3, para essa análise foram propostos dois novos cenários para avaliação dos resultados. Cada um desses cenários para redes heterogêneas considera diferentes latências entre as nuvens e as BSs, conforme indicados nas Tabelas 4 e 5.

A Figura 17 mostra o resultado para os dois cenários. No Cenário 4, para a quan-

Tabela 5 - Latências usadas na avaliação de Redes Heterogêneas para o Cenário 5.

Nuvem / BS	0	1	2	3	4	5	6	7
0	84	84	-	-	-	-	-	-
1	-	-	3	3	-	-	-	-
2	-	-	-	-	49	49	-	-
3	-	-	-	-	-	-	72	72
4	174	174	93	93	-	-	-	-
5	-	-	-	-	146	146	169	169
6	243	243	162	162	215	215	238	238

Figura 17 - Resultados para a Topologia 3 heterogênea nos Cenários 4 e 5.



(a) Latência média para o Cenário 4.

(b) Latência média para o Cenário 5.

tidade de VDUs disponível em cada uma das nuvens, conforme descrito na Tabela 3, foi possível constatar que a ordenação de latência não alcança o resultado ótimo. Já para o Cenário 5, a ordenação de latências foi fundamental para que o resultado alcançasse o valor ótimo. Isso ocorre pois a ordenação de latência não otimiza a utilização de recursos nas nuvens. Assim, ocupar primeiramente uma determinada nuvem que possui menor latência pode prejudicar alocações posteriores, que levariam a uma melhor latência média. Consequentemente, como trabalho futuro isso deve ser investigado para propor uma heurística que considere esses dois fatores.

6 CONCLUSÕES E DIREÇÕES FUTURAS

A preocupação com a latência é um fator importante em uma infraestrutura C-RAN. Assim, é possível utilizar esquemas do tipo *Dynamic* C-RAN para aproximar a nuvem dos usuários e decidir, dinamicamente, a posição das funções da rede. Esse esquema possui nuvens organizadas de forma hierárquica, que são responsáveis por executar funções da C-RAN, como controle de acesso ao meio e correção de erros. Por exemplo, é possível organizar a infraestrutura nos níveis de borda, regional e central. Nesse tipo de infraestrutura, é importante desenvolver algoritmos de otimização que posicionem as funções da rede nas diversas nuvens da hierarquia.

Esta dissertação complementou um modelo de otimização presente na literatura, propondo uma formulação que possui o objetivo de minimizar a latência média na rede. Nenhum trabalho avaliado considera a latência média para a tomada de decisão, sendo assim, avaliou-se a minimização da latência no posicionamento das funções de rádio em *Cloud* RANs, com base no Orquestrador Maestro, proposto na literatura (DALLA-COSTA et al., 2017a). Além disso, este trabalho propôs heurísticas para solucionar o problema formulado. A primeira heurística proposta neste trabalho considera que a infraestrutura é homogênea, ou seja, qualquer nuvem na infraestrutura possui a mesma latência para todas as BSs que podem utilizá-la para posicionar suas funções. Mostra-se neste trabalho que essa consideração permite propor um algoritmo com complexidade $O(n)$ para solucionar o problema. Como medida de comparação, este trabalho propôs uma segunda heurística que considera redes heterogêneas a partir da ordenação das BSs de acordo com a latência. Mostra-se, neste trabalho, que a segunda heurística possui complexidade maior que a primeira, devido à ordenação. Mais especificamente, mostra-se que a complexidade da heurística para redes heterogêneas depende apenas da complexidade do algoritmo de ordenação, que pode ser $O(n^2)$ ou $O(n \log n)$ nos casos mais comuns.

Apesar de mostrar situações nas quais as duas heurísticas propostas alcançam o valor ótimo, este trabalho não prova matematicamente que as heurísticas alcançam sempre o ótimo ou, alternativamente, o quão distante do ótimo o resultado da heurística pode gerar no pior caso. Assim, como trabalhos futuros, pretende-se desenvolver essas provas. Além disso, pretende-se considerar em um trabalho futuro a dinamicidade da rede, já que nesse trabalho os valores de demanda são estáticos; ou seja, o problema deverá ser executado periodicamente em função da mudança nas demandas. Por fim, para tornar a C-RAN uma eficiente solução para a questão de alocação de recursos de processamento, é interessante integrá-la às características da mobilidade celular. A partir do estudo da trajetória de deslocamento dos usuários e sua posição final, é possível realizar uma previsão da demanda das estações e posicionar as funções na infraestrutura de forma mais eficiente. Assim, uma outra direção futura é propor um solução que calcule as demandas

futuras dos usuários, alocando previamente as funções de acordo com essa previsão.

REFERÊNCIAS

ABDELWAHAB, S. et al. Network function virtualization in 5g. *IEEE Communications Magazine*, v. 54, n. 4, p. 84–91, 2016.

ALYAFAWI, I. et al. Critical issues of centralized and cloudified LTE-FDD radio access networks. In: *IEEE International Conference on Communications (ICC)*. [S.l.: s.n.], 2015. p. 5523–5528.

BARTELT, J. et al. Fronthaul and backhaul requirements of flexibly centralized radio access networks. *IEEE Wireless Communications*, v. 22, n. 5, p. 105–111, 2015.

BEYENE, Y. D.; JÄNTTI, R.; RUTTIK, K. Cloud-RAN architecture for indoor DAS. *IEEE Access*, v. 2, p. 1205–1212, 2014.

CHECKO, A. et al. Cloud RAN for mobile networks—a technology overview. *IEEE Communications surveys & tutorials*, v. 17, n. 1, p. 405–426, 2015.

CHIH-LIN, I. et al. Recent progress on C-RAN centralization and cloudification. *IEEE Access*, v. 2, p. 1030–1039, 2014.

CLAYMAN, S. et al. The dynamic placement of virtual network functions. In: *IEEE Network Operations and Management Symposium (NOMS)*. [S.l.: s.n.], 2014. p. 1–9.

COUTINHO, A. A. T. R.; CARNEIRO, E. O.; GREVE, F. G. P. Computação em névoa: Conceitos, aplicações e desafios. In: *Minicursos do XXXIV Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC)*. [S.l.: s.n.], 2016. p. 266–315.

DALLA-COSTA, A. G. et al. Maestro: An NFV orchestrator for wireless environments aware of VNF internal compositions. In: *IEEE International Conference on Advanced Information Networking and Applications (AINA)*. [S.l.: s.n.], 2017. p. 484–491.

DALLA-COSTA, A. G. et al. NFV em redes 5G: Avaliando o desempenho de composição de funções virtualizadas via Maestro. In: *XXXV Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC)*. [S.l.: s.n.], 2017. p. 1–14.

DEMESTICHAS, P. et al. 5g on the horizon: key challenges for the radio-access network. *IEEE Vehicular Technology Magazine*, v. 8, n. 3, p. 47–53, 2013.

ETSI. *ETSI GS NFV 002 V1.1.1: Network Functions Virtualization (NFV); Architectural Framework*. 2013.

GNU. *GLPK (GNU Linear Programming Kit)*. 2017. Https://www.gnu.org/software/glpk/ - Acessado em dezembro de 2017.

HAN, B. et al. Network function virtualization: Challenges and opportunities for innovations. *IEEE Communications Magazine*, v. 53, n. 2, p. 90–97, 2015.

HATOUM, A. et al. Cluster-based resource management in OFDMA femtocell networks with QoS guarantees. *IEEE Transactions on Vehicular Technology*, v. 63, n. 5, p. 2378–2391, 2014.

HEIDEKER, A.; KAMIENSKI, C. Gerenciamento flexível de infraestrutura de acesso público à Internet com NFV. p. 937–950, 2016.

HERRERA, J. G.; BOTERO, J.-F. Resource allocation in NFV: A comprehensive survey. *IEEE Transactions on Network and Service Management*, v. 13, n. 3, p. 518–532, 2016.

LIN, T. et al. Optimal network function virtualization realizing end-to-end requests. In: *IEEE Global Communications Conference (GLOBECOM)*. [S.l.: s.n.], 2015. p. 1–6.

LIU, J. et al. Graph-based framework for flexible baseband function splitting and placement in C-RAN. In: *IEEE International Conference on Communications (ICC)*. [S.l.: s.n.], 2015. p. 1958–1963.

LUIZELLI, M. C. et al. Piecing together the NFV provisioning puzzle: Efficient placement and chaining of virtual network functions. In: *IFIP/IEEE International Symposium on Integrated Network Management (IM)*. [S.l.: s.n.], 2015. p. 98–106.

MAKAYA, C. et al. Policy-based NFV management and orchestration. In: *IEEE Conference on Network Function Virtualization and Software Defined Network (NFV-SDN)*. [S.l.: s.n.], 2015. p. 128–134.

MIJUMBI, R. et al. Network function virtualization: State-of-the-art and research challenges. *IEEE Communications Surveys & Tutorials*, v. 18, n. 1, p. 236–262, 2015.

MIJUMBI, R. et al. Management and orchestration challenges in network functions virtualization. *IEEE Communications Magazine*, v. 54, n. 1, p. 98–105, 2016.

MOHAMMADKHAN, A. et al. Virtual function placement and traffic steering in flexible and dynamic software defined networks. In: *IEEE International Workshop on Local and Metropolitan Area Networks (LANMAN)*. [S.l.: s.n.], 2015. p. 1–6.

MUÑOZ, R. et al. Integrated SDN/NFV management and orchestration architecture for dynamic deployment of virtual sdn control instances for virtual tenant networks. *Journal of Optical Communications and Networking*, Optical Society of America, v. 7, n. 11, p. B62–B70, 2015.

NFV, G. Network functions virtualisation (NFV); architectural framework. *NFV ISG*, 2.

QUEIROZ, G. F. C.; COUTO, R. S.; SZTAJNBERG, A. TRELIS: Posicionamento de funções virtuais de rede com economia de energia e resiliência. In: 16^o *Workshop em Desempenho de Sistemas Computacionais e de Comunicação (WPERFORMANCE)*. [S.l.: s.n.], 2017. p. 1656–1669.

RIGGIO, R. et al. Scheduling wireless virtual networks functions. *IEEE Transactions on Network and Service Management*, v. 13, n. 2, p. 240–252, 2016.

SOARES, J. et al. Cloud4nfv: A platform for virtual network functions. In: *IEEE 3rd International Conference on Cloud Networking (CloudNet)*. [S.l.: s.n.], 2014. p. 288–293.

SUNDARESAN, K. et al. Fluidnet: A flexible cloud-based radio access network for small cells. *IEEE/ACM Transactions on Networking*, v. 24, n. 2, p. 915–928, 2016.

SZWARCFITER, J. L.; MARKENZON, L. *Estruturas de Dados e seus Algoritmos*. 3. ed. [S.l.]: Livros Técnicos e Científicos, 2013.

WANG, K.; ZHAO, M.; ZHOU, W. Traffic-aware graph-based dynamic frequency reuse for heterogeneous cloud-RAN. In: *IEEE Global Communications Conference (GLOBECOM)*. [S.l.: s.n.], 2014. p. 2308–2313.

WUBBEN, D. et al. Benefits and impact of cloud computing on 5G signal processing: Flexible centralization through cloud-RAN. *IEEE signal processing magazine*, v. 31, n. 6, p. 35–44, 2014.

XILOURIS, G. et al. T-nova: A marketplace for virtualized network functions. In: *European Conference on Networks and Communications (EuCNC)*. [S.l.: s.n.], 2014. p. 1–5.