# Alocação e posicionamento de recursos para redes de acesso virtualizadas com diferentes níveis de centralização

**Phelipe A. de Souza**[1], **Elivelton F. Bueno**[2], **Kleber V. Cardoso**[1]

[1]Instituto de Informática – Universidade Federal de Goiás (UFG)
Goiânia – GO – Brasil

[2]Escola de Matemática Aplicada (EMAp) – Fundação Getulio Vargas
Rio de Janeiro – RJ – Brasil

phelipedesouza@inf.ufg.br, bueno@impa.br, kleber@inf.ufg.br

***Abstract.*** *There are great expectations on technologies such as network functions virtualization (NFV) and centralized or cloud radio access network (CRAN) and how they can accelerate deployment of new network services and at the same time decrease some costs of the network operators. In this context, there is a relevant problem that involves three main concerns: 1) which cell sites to be updated; 2) how to update the selected cell sites, i.e., change to fully virtual or not; and 3) where to serve the visualized cell sites. Those issues are influenced by the centralization level employed on a certain radio access network (RAN). We propose an optimization model that allows the decision maker to define the weight (or reward) of the centralization level and evaluate the impact on metrics such as the necessary investment and the achieved level of centralization. The model shows how the investment should be allocated depending on the level of centralization and the relative cost among the different resources. Our heuristic presents performance similar to the deterministic approach but it is able to obtain solutions much faster and to deal with large networks (i.e., cities and metropolitan regions).*

***Resumo.*** *Existem grandes expectativas nas tecnologias de centralização (CRAN) e de virtualização de rede (NFV); e como elas podem acelerar a implantação de novos serviços e, ao mesmo tempo, diminuir os custos das operadoras de redes. Vários trabalhos discutiram os benefícios de implantar uma nova infraestrutura de rede, mas apenas alguns investigaram como deve ser a transição de uma rede legada. Nesse contexto, existe um problema relevante que envolve três questões principais: 1) quais locais da rede devem ser atualizados; 2) como atualizar o local selecionado, i.e., para totalmente virtualizado ou não; e 3) quem deve atender aos locais virtualizados. Essas questões são influenciadas pelo nível de centralização empregado em uma determinada rede de acesso (RAN). Aqui, propomos um modelo de otimização e uma heurística que permite ao tomador de decisão definir o nível de centralização desejado e avaliar seu impacto em algumas métricas, tais como o investimento necessário e o nível de centralização efetivamente alcançado. O modelo mostra como o investimento deve ser aplicado de acordo com o nível de centralização e o custo relativo entre os diferentes recursos. Nossa heurística apresenta desempenho semelhante à abordagem determinística, mas pode obter soluções mais rapidamente e lidar com redes maiores (i.e., cidades e regiões metropolitanas).*

## 1. Introdução

A computação em nuvem tem mostrado que a virtualização de recursos é essencial para alcançar metas como escalabilidade, flexibilidade e robustez. A rede definida por *software* (SDN, Software Defined Networking) e a virtualização das funções de rede (NFV, Network Functions Virtualization) procuram alcançar um sucesso semelhante no contexto das redes de comunicação através da definição e adoção de um conjunto de padrões que utilizam equipamentos de propósito geral capaz de executar diversas funções convencionais de rede (i.e., DHCP, DNS, Firewall, DPI) e rede sem fio, e.g., controle de interferência (eICIC, Enhanced Inter-cell Interference Coordination) e processamento de banda [Nunes et al. 2014]. Vários benefícios podem ser obtidos com a utilização de SDN/NFV, como a redução dos custos de investimento (CAPEX) e operacionais (OPEX), e melhor compartilhamento dos recursos de *hardware*. No entanto, esses benefícios dependem fortemente da alocação eficiente dos recursos da rede.

A rede de acesso de rádio (RAN, Radio Access Network) tradicional apresenta uma arquitetura distribuída, onde cada local/célula (CS, Cell Site) contém uma estação/torre rádio base (BS, Base Station) que é divida em duas partes: 1) a unidade de banda base (BBU, baseband unit), que fica na base da torre, é responsável por executar o processamento digital (modulação e demodulação) dos sinais de banda base, realizar a conexão com o núcleo da rede e executar funções de controle e monitoramento; 2) e a unidade de rádio remoto (RRH, Remote Radio Head) ou equipamento de rádio (RF, Radio Equipament), que recebe e transmite sinais analógicos ao equipamento de usuário (UE, User Equipament) e é instalado no topo da torre. Esses dois equipamentos se comunicam através de uma interface, e.g., a interface comum de rádio pública (CPRI, Common Public Radio Interface), óptica ou coaxial. As BSs compõem a rede de acesso que é chamada de *fronthaul*. Além disso, todos os *links* e equipamentos de comunicação (e.g., roteadores e switches) entre a RAN e o núcleo da rede determinam o *backhaul*.

Desse modo, a RAN centralizada ou em nuvem (CRAN, Centralized-RAN ou Cloud-RAN) aparece como uma abordagem promissora para reduzir o CAPEX e o OPEX da rede. Ao implantar a CRAN, um operador pode reduzir o consumo energético, melhorar a acurácia dos serviços de controle de interferência e facilitar as atualizações da rede [Checko et al. 2015]. Naturalmente, o uso conjunto de SDN/NFV e CRAN oferece importantes vantagens, como a padronização de *hardware* e *software* para virtualização, e a alta flexibilidade na manutenção e operação da rede. Os benefícios são ainda maiores em um contexto de CRAN combinado com a tradicional RAN distribuída, o que tende a ser um cenário comum para a evolução da rede de acesso [Asensio et al. 2016]. Neste contexto, identificar a quantidade e a localização dos recursos necessários para atender a essa demanda tornou-se um problema relevante a ser abordado. Esse problema possui características específicas quando as demandas estão relacionadas às funções da rede sem fio [Musumeci et al. 2016]. Por exemplo, a função *coordinated multi-point* (CoMP) permite a coordenação dinâmica de transmissão e recepção de dados pelo usuário com uma variedade de BSs diferentes, no entanto, a troca de informações de controle entre as BSs e o núcleo da rede apresenta restrições muito rigorosas em relação ao baixo nível de latência exigido pelo CoMP.

Neste trabalho, apresentamos um modelo de otimização de alocação de recursos que minimiza o custo da atualização de uma infraestrutura RAN com diferentes níveis de

centralização, que pode ser composta por diferentes gerações de *hardware* e *software*. A solução do modelo indica I) os locais a serem atualizados, II) quais deles devem ser centralizados e III) onde a centralização deve ocorrer. Nosso modelo considera os custos envolvendo os *links* e os locais, e também a capacidade máxima dos recursos de rede. Apresentamos um algoritmo heurístico eficiente, capaz de resolver o modelo proposto para redes mais realistas (i.e., cidades e regiões metropolitanas) e que assegura uma solução satisfatória.

Esse artigo está organizado da seguinte forma: Na Seção (2) apresentamos os trabalhos relacionados, na Seção (3) introduzimos o problema de alocação de recursos sob diferentes níveis de centralização. Na Seção (4) propomos a formulação matemática do problema como um modelo de programação linear inteira e, também, uma abordagem heurística de solução. Na Seção (5) realizamos as avaliações do modelo e da heurística. Por fim, apresentamos as considerações finais e trabalhos futuros na Seção (6).

## 2. Trabalhos Relacionados

A modelagem de alocação de recursos e virtualização de funções de rede tem sido investigada em vários artigos a partir de diferentes perspectivas [Bouet et al. 2015, Baumgartner et al. 2015, Basta et al. 2014]. Por exemplo, uma alocação de função de rede é apresentada em [Bouet et al. 2015], que se concentra na função de inspeção de pacotes e tem como objetivo instalar o menor número de locais de alta centralidade capazes de lidar com toda a demanda de tráfego do *backhaul* da rede. Os autores de [Baumgartner et al. 2015] abordam a virtualização com NFV de funções específicas do núcleo da rede, como SGW, PGW, MME e HSS. O objetivo é minimizar o custo dos recursos necessários por essas funções, tanto dos *links* de comunicação, quanto de computação de cada vértice da rede. [Basta et al. 2014] utilizam SDN e NFV para realizar a virtualização e alocação dos gateways do núcleo da rede (i.e., SGW e PGW) em data center. O objetivo é minimizar o atraso e a carga do plano de dados, além da sobrecarga do plano de controle em uma rede SDN. Naturalmente, fica claro o caminho em se virtualizar as funções do *backhaul* e do núcleo da rede (EPC, Envolved Packet Core). No entanto, há também a necessidade de se virtualizar as funções da RAN para que junto com as funções do *backhaul* e do núcleo da rede seja possível obter benefícios no gerenciamento e na utilização dos recursos disponíveis na rede.

Visando a rede de acesso, [Musumeci et al. 2016] apresenta um modelo de otimização para a implantação de CRANs através do posicionamento de BBUs e o uso de redes de acesso/agregação de baixa latência, com alta vazão e que utilizem fibra óptica que operem com multiplexação por divisão de comprimento de onda (WDM, Wavelength-division Multiplex). Diferente do nosso trabalho, eles empregam uma abordagem comum ao assumir uma solução CRAN completa na rede, i.e., onde todos os locais da rede são compostos apenas por RRHs e devem ser atendidos por um conjunto de BBUs centralizadas. Além disso, [Musumeci et al. 2016] não considera o custo de investimento na atualização dos CSs e dos escritórios centrais (CO, Central Office).

Embora vários artigos tenham discutido a implantação de CRANs como uma solução disruptiva de evoluir as infraestruturas atuais, a combinação de CRAN com as RANs tradicionais pretende melhor atender as expectativas de investimento dos provedores. Neste contexto, [Asensio et al. 2016] apresenta um modelo que minimiza o custo

de CAPEX (e.g., equipamentos de rede e servidores de virtualização) para equipar os locais que hospedam CRANs. Os autores avaliam diferentes níveis de centralização da rede e como isso pode afetar na redução de custos. No entanto, em [Asensio et al. 2016], os autores assumem que toda a infraestrutura de rede foi desenvolvida para suportar a virtualização, ou seja, todos os locais da rede já possuem RRH e podem conter BBU.

No modelo a ser apresentado, assumimos que a infraestrutura de rede pode ser composta por dispositivos de diferentes gerações, e.g., 2G, 3G, 4G/LTE (Long Term Evolution), e as novas versões de 4G conhecidas como LTE-A (Long Term Evolution Advanced). Desse modo, nosso modelo é mais genérico que o modelo apresentado por [Asensio et al. 2016], e pode efetivamente representar a evolução gradual de uma infraestrutura legada para uma mais moderna. Além disso, nosso modelo considera a demanda em cada local da rede e o quanto ela afeta a disponibilidade dos recursos de rede.

## 3. Modelo do sistema

Nesta seção, introduzimos o problema de alocação de recursos em diferentes níveis de centralização, ao qual referenciamos por DCL-RAN. Por conta dos equipamentos e do modo de operação da rede atual, o custo de se construir e atualizar a RAN está ficando cada vez mais caro, enquanto a receita não cresce na mesma taxa para o operador da rede. Desse modo, substituindo os equipamentos proprietários (que o operador não pode modificar) por plataformas de propósito geral (i.e., PC x86) capazes de dar suporte às técnicas de virtualização da rede deve ser possível fornecer um ambiente mais flexível para o operador, reduzir a complexidade da rede e os custos de investimento.

Assumimos que o operador precisa atualizar sua infraestrutura para atender a uma crescente demanda dos terminais conectados à rede. Para isso, ele deseja aplicar o mínimo de investimento necessário. Uma vez que o operador sabe que a demanda está aumentando e que ela flutua muito em diferentes partes da rede (i.e., devido à mobilidade dos usuários), ele escolhe investir em tecnologias (i.e., CRAN e SDN/NFV) que oferecem mais flexibilidade e que podem ser facilmente reconfiguradas e expandidas.

Para implementar CRAN com funções virtualizadas de rede (VNF, Virtualized Network Functions), os locais a serem atualizados devem ser identificados. O custo desses locais pode variar de um para o outro dependendo do *hardware*, *software* e das interfaces de comunicação. Quando um local é atualizado, ele recebe um conjunto de RRHs e se torna um *virtual eNodeB* (veNodeB). Além disso, um conjunto específico de locais podem ser escolhidos para hospedar, de maneira centralizada, as BBUs (i.e., um hotel de BBUs). De acordo com a ETSI (*European Telecommunications Standards Institute*), em [ETSI 2014], um conjunto de componentes de *hardware* e *software* capazes de fornecer um ambiente de execução para as VNFs é chamado de NFVI-Node (*Network Functions Virtualization Infrastructure Node*).

A figura (1) ilustra uma RAN híbrida composta de elementos centralizados e distribuídos. Os elementos centralizados são os NFVI-Nodes que concentram as funções de rede, processamento e controle dos vértices atualizados como veNodeB. Os locais distribuídos, são os elementos convencionais da rede que não foram atualizados, sendo denominados de *legacy nodes* (ou vértices/itens legados). Embora a RAN parcialmente centralizada possa não ser a melhor escolha por conta do seu desempenho, essa topologia é razoável ou até mesmo preferível financeiramente para descrever uma rede em evolução.

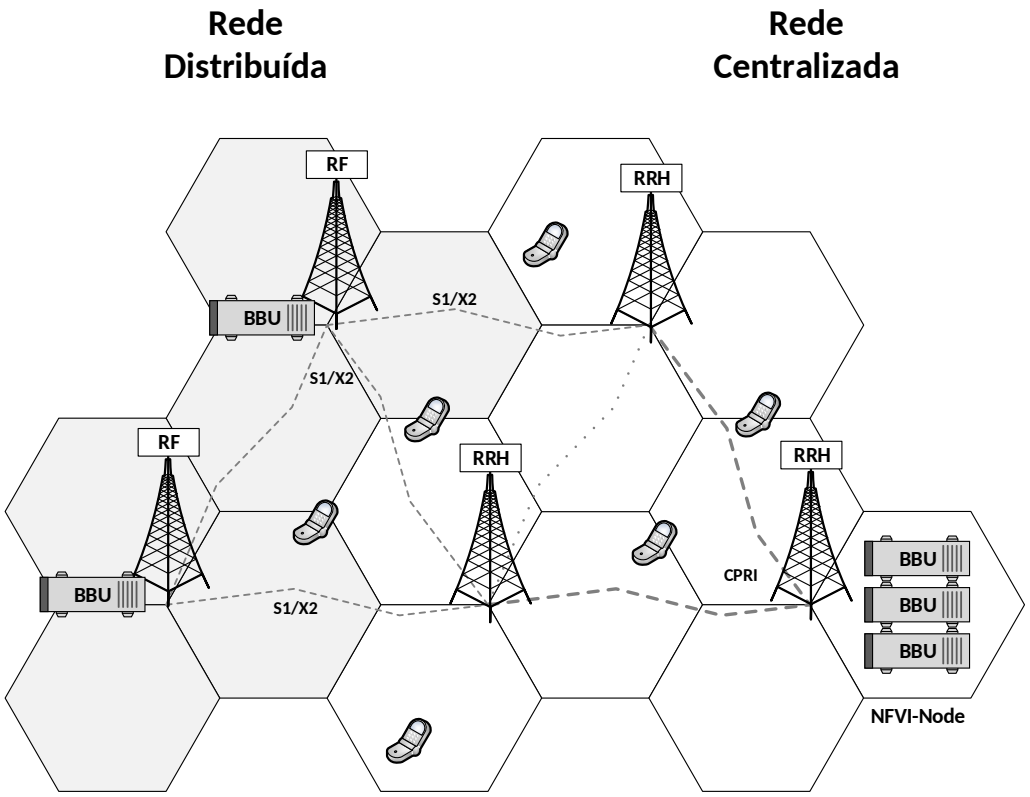



**Figura 1. Rede de acesso de rádio distribuída e centralizada.**

A topologia da RAN é representada como um grafo conectado e não-direcionado $G = (V, E)$, no qual o vértice $v \in V$ é um elemento da rede celular e a aresta $(i, j) \in E$ é um link de comunicação. Em nosso modelo, um vértice é um CO com grande número de recursos de computação e *links* de comunicação, i.e., arestas. Assim, um CO aparece como um vértice convencional no grafo. Além disso, cada aresta $(i, j) \in E$ tem uma capacidade de $u_{i,j}$ bps.

Se um vértice $s \in V$ for selecionado para ser atualizado como um veNodeB (i.e., para conter apenas RRHs), então um custo $c_s^1$ deve ser pago. Ao longo deste trabalho, essa decisão é representada pela atribuição do valor 1 para a variável binária $y_s^1 \in \{0, 1\}$. Da mesma forma, se um vértice $t \in V$ for selecionado para ser atualizado como um NFVI-Node (i.e., para conter RRHs e BBUs), então um custo $c_t^2$ deve ser pago e denotamos tal decisão como $y_t^2 = 1$. O custo efetivo de atualização pode variar de um site para outro, uma vez que cada um apresenta uma quantidade diferente de recursos e esses recursos podem estar em diferentes níveis de atualização. Cada vértice $s \in V$ tem uma demanda $d_s$ bps e pode servir a um número máximo de $b_s$ bps devido aos seus recursos limitados de computação, memória e das interfaces de comunicação sem fio. Desse modo, em um NFVI-Node, a escassez de recursos afeta o número de VNFs e a quantidade de veNodeBs que podem ser atendidos.

Denominamos $P$ o conjunto de todos os caminhos candidatos entre cada par de vértices no grafo $G$, com $P_{s,t} \subset P$ representando o conjunto de todos os caminhos entre os vértices $s \in V$ e $t \in V$. Logo, os caminhos candidatos são todos os caminhos que satisfazem às restrições de latência e de capacidade máxima das arestas se o vértice $s$ for

servido por $t$. Associado ao conjunto $P_{s,t}$, definimos a variável de decisão $x^p_{s,t} \in \{0,1\}$, que deve ser igual a 1 apenas se $t$ for efetivamente escolhido para servir $s$ através do caminho $p \in P_{s,t}$. Isso implica diminuir em $d_{\boldsymbol{s}(p)}$ bps a capacidade de cada aresta ao longo do caminho $p$ onde $d_{\boldsymbol{s}(p)}$ é a demanda do primeiro vértice de $p$.

Normalmente, para ser devidamente atendido por uma BBU remota, um RRH apresenta uma rigorosa restrição de latência na comunicação com a BBU [Checko et al. 2015]. No entanto, a NGFI (*Next Generation Fronthaul Interface*) propõe uma nova interface para o *fronthaul* que facilita a implantação da CRAN/NFV, compatível com as plataformas de propósito geral e suficientemente escalável para suportar a evolução da rede [Chih-Lin et al. 2015]. Ambas as abordagens podem ser representadas em nosso modelo como o atraso máximo permitido pelo vértice $s$ ao longo do caminho $p \in P_{s,t}$ até chegar ao vértice $t$.

A tabela 1 resume a descrição dos parâmetros de entrada empregados no nosso modelo do sistema e também na formulação do problema, que será apresentada na Seção (4) a seguir.

**Tabela 1. Resumo das notações**

| **Parâmetro** | **Definição** |
|---|---|
| $V$ | Conjunto de todos os vértices. |
| $E$ | Conjunto de todas as arestas. |
| $P$ | Conjunto de todos os caminhos candidatos (viáveis) entre os vértices. |
| $P_{s,t}$ | Conjunto de todos os caminhos candidatos de $s \in V$ até $t \in V$. |
| $P(i,j)$ | Todos os caminhos candidatos que utilizam a aresta $(i,j) \in E$. |
| $\boldsymbol{s}(p)$ | Primeiro vértice de um caminho $p \in P$. |
| $\boldsymbol{t}(p)$ | Último vértice de um caminho $p \in P$. |
| $u_{i,j}$ | Capacidade da aresta $(i,j) \in E$. |
| $d_s$ | Demanda do vértice $s \in V$. |
| $b_t$ | Capacidade no vértice $t \in V$. |
| $c^1_s$ | Custo de se atualizar o vértice $s \in V$ para uma *veNodeB*. |
| $c^2_t$ | Custo de se atualizar o vértice $t \in V$ para um *NFVI-Node*. |
| $\alpha$ | Peso (recompensa) da centralização. |
| **Variável** | **Definição para quaisquer vértices** $s,t \in V$ **e caminho** $p \in P_{s,t}$ |
| $x^p_{s,t} \in \{0,1\}$ | Valor 1 se $s$ usa $p$ para ser servido por $t$; valor 0, caso contrário. |
| $y^1_s \in \{0,1\}$ | Valor 1 se $s$ torna uma veNodeB; valor 0, caso contrário. |
| $y^2_t \in \{0,1\}$ | Valor 1 se $t$ torna um NFVI-Node; valor 0, caso contrário. |

## 4. Formulação do problema e solução

Nesta seção, iremos apresentar a formulação matemática do problema deste estudo, DCL-RAN, como um modelo de programação linear inteira (ILP) e, também, propor uma

solução heurística para ele.

### 4.1. Modelo ILP do problema DCL-RAN

Usando o cenário proposto e as notações apresentadas na Tabela 1, propomos o seguinte modelo ILP para o problema DCL-RAN:

$$\textit{minimizar} \quad \sum_{s \in V} c_s^1 y_s^1 + \sum_{t \in V} c_t^2 y_t^2 - \alpha \sum_{s \in V} y_s^1 \tag{1}$$

*sujeito a*:

$$y_t^1 + y_t^2 \leq 1, \quad \forall t \in V \tag{2}$$

$$x_{s,t}^p + y_t^1 \leq 1, \quad \forall s, t \in V, p \in P_{s,t} \tag{3}$$

$$x_{s,t}^p - y_t^2 \leq 0, \quad \forall s, t \in V, s \neq t, p \in P_{s,t} \tag{4}$$

$$\sum_{t \in V} \sum_{p \in P_{s,t}} x_{s,t}^p = 1, \quad \forall s \in V \tag{5}$$

$$\sum_{s \in V} \sum_{p \in P_{s,t}} x_{s,t}^p d_s \leq b_t, \quad \forall t \in V \tag{6}$$

$$\sum_{p \in P(i,j)} d_{\boldsymbol{s}(p)} x_{\boldsymbol{s}(p),\boldsymbol{t}(p)}^p \leq u_{i,j}, \forall (i,j) \in E \tag{7}$$

$$x_{s,t}^p \in \{0,1\}, \quad \forall s, t \in V, p \in P_{s,t} \tag{8}$$

$$y_s^1 \in \{0,1\}, \quad \forall s \in V \tag{9}$$

$$y_t^2 \in \{0,1\}, \quad \forall t \in V \tag{10}$$

A restrição (2) garante que nenhum vértice pode ser configurado para se tornar, ao mesmo tempo, um veNodeB e um NFVI-Node. Além disso, a restrição (3) impede que um vértice puramente virtual (i.e., veNodeB) seja atribuído para servir a si mesmo ou qualquer outro vértice da rede, pois, um veNodeB não contém recursos de processamento e é composto apenas por um conjunto de RRHs. Por outro lado, a restrição (4) garante que apenas os vértices atualizados para serem NFVI-Node podem servir outros vértices (i.e., veNodeBs) além de si mesmos.

Para definir a alocação dos vértices, a restrição (5) garante que todo vértice seja servido por pelo menos um vértice da rede sempre através de um único caminho. Desse modo, um vértice pode atender a si mesmo, se mantendo como um eNodeB convencional (ou *legacy node*), ou ser atendido por um vértice remoto (i.e., NFVI-Node). Além disso, para definir o limite de utilização dos recursos, a restrição (6) impede que um vértice seja atribuído para atender a uma demanda que exceda a sua capacidade. Da mesma forma, a restrição (7) garante que a carga total transportada por uma aresta não exceda a sua capacidade. As restrições (8), (9) e (10) garantem a integralidade e a condição de não negatividade para as variáveis de decisão de acordo com suas respectivas definições.

E, finalmente, a função objetivo (1) é composta por duas partes: os dois primeiros somatórios priorizam uma configuração de rede que minimize o custo total de atualização de seus elementos (i.e., para veNodeB ou para NFVI-Node); e o último somatório, ponderado por $\alpha$, estimula a virtualização pura (para veNodeB) dos elementos da rede. Em particular, esse estímulo à virtualização pura tende a priorizar os elementos com menor

custo $c_s^1$ de atualização para veNodeB e, também, os de menor demanda $d_s$ (pois, em geral, um veNodeB com $d_s$ grande violaria mais restrições de capacidade de arestas em (7)), assim como os de menor capacidade $b_s$ (pois, em geral, elementos com $b_s$ maiores suportam mais designações na restrição correspondente em (6), o que os tornariam melhores candidatos à NFVI-Node).

### 4.2. Heurística de Alocação de Recursos

Nossa formulação se assemelha ao problema clássico de otimização conhecido como um problema de *location-allocation* (LA) [Azarmand and Neishabouri 2009]. No problema de LA, o objetivo é minimizar o custo para localizar um conjunto de novas instalações, considerando o custo de transporte das instalações até os clientes e o número de instalações a serem alocadas para satisfazer a demanda do cliente. Esse problema é NP-hard. Portanto, nenhum algoritmo de otimização em tempo polinomial é conhecido por resolver o problema de LA. Além do nosso modelo ser semelhante ao problema de LA, as evidências obtidas com o uso do CPLEX, que descrevemos na Seção 5, sugerem que estamos lidando de fato com um problema NP-hard.

Como apresentado na Tabela 2, o grande número de antenas instaladas nas principais cidades e regiões metropolitanas do Brasil, sugere a necessidade de se desenvolver métodos aproximativos ou heurísticas que consigam resolver em tempo polinomial o problema DCL-RAN. De fato, como veremos na Seção 5 a implementação do nosso modelo ILP com um dos mais eficientes pacotes de otimização exata da atualidade ainda não nos permite resolver instâncias de relevância prática para o problema. Desse modo, desenvolvemos uma heurística que consegue resolver de forma aproximada, mas em tempo polinomial, esse problema de alocação de recursos para cenários mais realistas.

**Tabela 2. Número de BS nas maiores capitais brasileiras [Telebrasil 2017].**

| *Cidades* | **CLARO** | **OI** | **TIM** | **VIVO** | *Total* |
|---|---|---|---|---|---|
| São Paulo | 991 | 1590 | 1427 | 1230 | 5238 |
| Rio de Janeiro | 839 | 813 | 1522 | 925 | 4099 |
| Brasília | 409 | 401 | 576 | 337 | 1723 |
| Belo Horizonte | 300 | 275 | 402 | 306 | 1283 |
| Salvador | 286 | 260 | 357 | 185 | 1088 |
| Fortaleza | 141 | 253 | 321 | 190 | 915 |
| Goiânia | 179 | 188 | 220 | 103 | 690 |

O Algoritmo 1 resume a nossa abordagem heurística, à qual nos referimos HDCL-NFV. O algoritmo começa ($\ell.$ 2) ordenando, em ordem não-decrescente, todos os vértices de $V$ em relação à razão $\frac{c_t^2}{b_t}$ entre o custo de atualização e a capacidade de cada vértice $t \in V$. Sobre esse conjunto ($N$), define uma janela $w$ de deslocamento ($\ell.$ 3), que representa um certo nível de aleatoriedade que $\alpha$ gera na seleção dos elementos que se deseja virtualizar na rede. Para determinar o valor de $w$, o algoritmo utiliza o número de vértices, $|N|$, o maior custo de atualização da rede, $\max\{c_t^2 | t \in V\}$, e o nível de centralização, $\alpha$. Quanto menor o valor de $\alpha$ maior deve ser o tamanho da janela $w$ e, a aleatoriedade na

escolha dos elementos. O tamanho da janela, $w$, é definido pela equação:

$$w = \max\left\{1, |N| \cdot \left(1 - \frac{\alpha}{\max\limits_{t \in V}\{c_t^2\}}\right)\right\}$$

A cada iteração, um vértice $t$ é selecionado ($\ell$. 5) aleatoriamente do subconjunto $[0, w] \subset N$. Logo, $t$ é designado a processar sua própria demanda e posteriormente a sua capacidade residual, `cap[t]`, é atualizada. Para representar o conjunto $N_{l:r} \subset N$ de elementos da janela $w$, foi definido os índices $r$ e $l$, onde $r$ representa a posição final e o índice $l$ a posição inicial da janela $w$. Enquanto houver elementos para serem alocados, $r > 0$, e o vértice $t$ possuir capacidade de atender o vértice de menor demanda da rede, $\min_{k \in N_{0:r}}\{d_k\}$, um vértice $s$ poderá ser alocado em $t$. Desse modo, para $s$ ser alocado em $t$, ($\ell$. 12), utiliza-se a mesma estratégia aplicada na seleção de $t$, no entanto, priorizando os elementos que apresentam a maior relação custo/capacidade presente no final do conjunto $N$. Assim, inicia-se selecionando uma posição aleatória do subconjunto $N_{l:r} \subset N$, onde o valor $s$ dessa posição é trocado com o valor da posição $r$, de forma que o elemento $s$ passe a ser selecionado pela posição $r$. Logo, realiza-se o deslocamento da janela $w$, atualizando-se os valores de $r$ e $l$, permanecendo o valor de $s$ no final da janela para o caso de $s$ não ser alocado em $t$. Assim, o vértice $s$ só vai ser atendido por $t$ quando o vértice $t$ possuir capacidade suficiente para atender a demanda de $s$ e algum caminho $p \in P_{s,t}$ dispor de capacidade residual para transportar os dados de $s$.

No final, para cada vértice $i \in V$ da rede, o algoritmo atualiza $i$ para um NFVI-Node se ele for escolhido para atender demandas diferentes da sua, ou atualiza $i$ para veNodeB se sua demanda for processada por algum outro vértice, ou mantém $i$ como um vértice legado se ele simplesmente for designado para processar sua própria demanda.

## 5. Avaliação

Nesta seção, avaliamos nosso modelo ILP e mostramos como os parâmetros de entrada afetam o custo da atualização da rede. Analisamos nossa heurística e resolvemos o modelo ILP (1)–(10) com diferentes instâncias de RANs. Utilizamos Python para gerar os cenários da rede e resolver nossa heurística e CPLEX para resolver o problema de otimização.

Semelhante ao que foi apresentado em [Bouet et al. 2015], empregamos o modelo Barabási-Albert para criar topologias de RAN de forma aleatória, com diferentes tamanhos e com no máximo 200 vértices. Os caminhos candidatos são pré-computados e fornecidos como dados de entrada. Na fase de pré-processamento, verificamos, para cada par de vértices ($s$, $t$), cada um dos caminhos elementares entre eles e formamos um conjunto apenas com aqueles que são, a priori, viáveis em relação a três critérios: a capacidade do vértice de destino ($t$) é suficiente para processar toda a demanda do vértice de origem ($s$); a latência entre os dois vértices não é maior que a latência máxima permitida; e a capacidade da aresta que faz parte do caminho é o suficiente para transportar a demanda adicional do vértice de origem. Assim, eliminamos apenas caminhos que são de fato inviáveis e que não afetam na obtenção de solução ótima por parte do solver exato, i.e., CPLEX.

**Algorithm 1** Abordagem heurística.

**Input:** Todas as entradas listadas na Tabela 1.
**Output:** Possível $\hat{x}^p_{s,t}$, $\hat{y}^1_t$, e $\hat{y}^2_t$, $\forall s, t \in V,\ p \in P_{s,t}$.

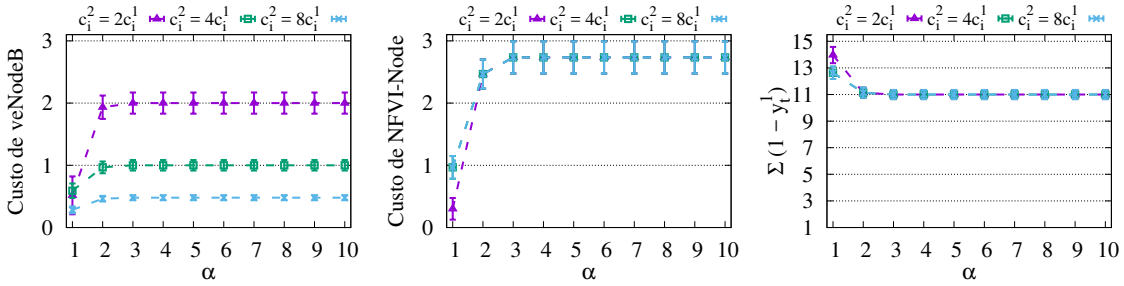
$\hat{x}^p_{s,t} \leftarrow 0$, $\hat{y}^1_t \leftarrow 0$, $\hat{y}^2_t \leftarrow 0$, $\forall s, t \in V,\ p \in P_{s,t}$
$N \leftarrow$ ORDENA_CUSTO2_CAPACIDADE($V$)
$w \leftarrow$ JANELA_DE_ALEATORIEDADE($|N|, \alpha$)
**while** $N \neq \emptyset$ **do**
    $t \leftarrow$ OBTÉM_POSSÍVEL_NFVI($N, w$)
    $\hat{x}^p_{t,t} \leftarrow 1$ ▷ $p$ identifica o caminho de um ponto $(t, t)$
    $\texttt{cap}[t] \leftarrow b_t - d_t$
    $N \leftarrow N - \{t\}$
    $r \leftarrow |N|$
    $l \leftarrow \max\{0, r - w\}$
    **while** $r > 0$ **and** $\texttt{cap}[t] \geq \min_{k \in N_{0:r}}\{d_k\}$ **do**
        $s \leftarrow$ OBTÉM_POSSÍVEL_VENODEB($N_{l:r}$) ▷ $N_{l:r} = \{N_i : l \leq i < r\}$
        $p \leftarrow$ SELECIONA_CAMINHO_VIÁVEL($P_{s,t}$)
        $r \leftarrow r - 1$
        $l \leftarrow \max\{0, r - w\}$
        **if** $d_s < \texttt{cap}[t]$ **and** $p \neq \texttt{nil}$ **then**
            $\hat{x}^p_{s,t} \leftarrow 1$, $\hat{y}^1_s \leftarrow 1$, $\hat{y}^2_t \leftarrow 1$
            $\texttt{cap}[t] \leftarrow \texttt{cap}[t] - d_s$
            $N \leftarrow N - \{s\}$

Avaliamos os cenários com vértices homogêneos e heterogêneos em termos de recursos ($b_i$) e custos de atualização ($c^1_i, c^2_i$). No caso homogêneo, a capacidade computacional máxima da topologia se baseia no valor médio obtido do modelo de demanda [Lee et al. 2014]. Devido ao tempo de execução em cenários com dados homogêneos o modelo de otimização é avaliado em topologias com até 15 vértices. No caso heterogêneo, esse valor difere na maioria dos vértices com a execução do modelo de demanda. Desse modo, se a demanda do vértice $s \in V$ for atendida por um vértice $t \in V$, então a demanda adicionada em $t$ é $min\{d_s, |b_t - d_t|\}$. Nossos resultados são calculados com uma média de 30 execuções por simulação, cada uma correspondente a uma configuração de demanda diferente ($d_i$). Os valores médios são mostrados com um intervalo de confiança de 95%.

A demanda é baseada no modelo proposto por [Lee et al. 2014]. Esse modelo representa a distribuição espacial do tráfego da rede móvel como uma distribuição log-normal. A área de avaliação é dividida em pixels quadrados e para cada pixel é atribuída uma demanda. Nossa área de avaliação tem $5km$ x $5km$ e as células foram distribuídos empregando o modelo *hardcore point process* (HCPP) [ElSawy et al. 2013] com uma distância mínima de 500 metros entre cada célula. Assumimos que a área está totalmente coberta e utilizamos *Voronoi Tessellation* para descrever a cobertura efetiva de cada célula. Desse modo, cada célula cobre um conjunto diferente de pixels, o que significa uma quantidade diferente de demanda ($d_i$) dos usuários.

As figuras 2(a), 2(b), e 2(c) apresentam o investimento necessário e o número de vértices que não são puramente virtuais em função de $\alpha$. O número de vértices que não são

puramente virtuais (i.e., não são veNodeB) representa o nível de centralização. Quanto menor o número de vértices, maior deveria ser o nível de centralização. Os vértices são homogêneos em $c_i^1, c_i^2$ e $b_i$. A homogeneidade de custo limita o impacto de $\alpha$, que se torna estável após a mudança inicial. A capacidade dos vértices (igual à demanda média) e a grande variação da demanda limita o número de vértices que não são puramente virtuais. Além disso, essa configuração motiva o investimento em NFVI-Nodes ser maior que o investimento em veNodeBs. Observamos tendências semelhantes quando os vértices são homogêneos em $c_i^1$ e $c_i^2$, mas heterogêneos em $b_i$. No entanto, existem diferenças, a heterogeneidade da capacidade permite uma melhor centralização e um aumento no investimento em veNodeB.



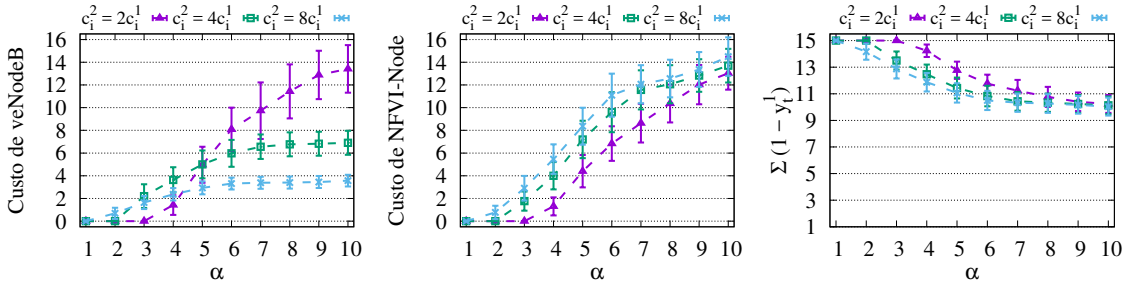

(a) Investimento em veNodeB. (b) Investimento em NFVI-Node. (c) Vértices que não são puramente virtuais.

**Figura 2. Investimento necessário e nível de centralização com $c_i^1, c_i^2$, e $b_i$ homogêneos.**

As figuras 3(a), 3(b), and 3(c) também apresentam o investimento necessário e o número de vértices que não são puramente virtuais em função de $\alpha$, mas agora os vértices são heterogêneos em $c_i^1, c_i^2$ e $b_i$. Esta configuração representa cenários em que os CS têm diferentes recursos, em geral, para atender demandas diferentes. A heterogeneidade do custo é comum quando diferentes gerações de tecnologia de comunicação são empregadas, por exemplo, 2G, 3G e 4G. A escolha de $\alpha$ tem um grande impacto no custo e no nível de centralização. Geralmente, o investimento em NFVI-Nodes é ainda maior do que o investimento em veNodeB, exceto quando $c_i^2 = 2c_i^1$. Como anteriormente, o nível de centralização eleva com o aumento do $\alpha$ para todas as relações entre $c_i^1$ e $c_i^2$. No entanto, as curvas ilustram como $\alpha$ afeta cada relação entre $c_i^1$ and $c_i^2$. Observamos tendências semelhantes quando os vértices são heterogêneos em $c_i^1$ e $c_i^2$, mas homogêneos em $b_i$. Da mesma forma que os resultados anteriores, a homogeneidade da capacidade afeta a centralização e o investimento em veNodeB.

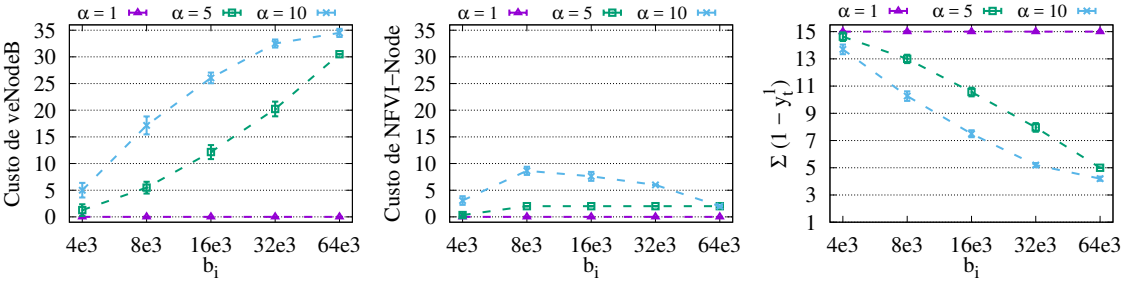
As figuras 4(a), 4(b), e 4(c) também apresentam o investimento necessário e o número de vértices que não são puramente virtuais, mas em função da capacidade ($b_i$). Os vértices são heterogêneos em $c_i^1$, $c_i^2$, e $c_i^2 = 2c_i^1$. Com base nos resultados anteriores, espera-se que $\alpha = 1$ não implique em investimento e nem na centralização dos elementos da rede. Por outro lado, $\alpha \geq 5$ motiva um grande investimento, principalmente em veNodeBs. Naturalmente, isso acontece devido à oportunidade de melhorar em grande parte o nível de centralização. Conforme ilustrado pela figura 4(c), o nível de centralização é fortemente afetado pela capacidade dos vértices ($b_i$).

Finalmente, comparamos o desempenho da heurística com a abordagem de




(a) Investimento em veNodeB. (b) Investimento em NFVI-Node. (c) Vértices que não são puramente virtuais.

**Figura 3. Investimento necessário e nível de centralização com $c_i^1, c_i^2$, e $b_i$ heterogêneos.**



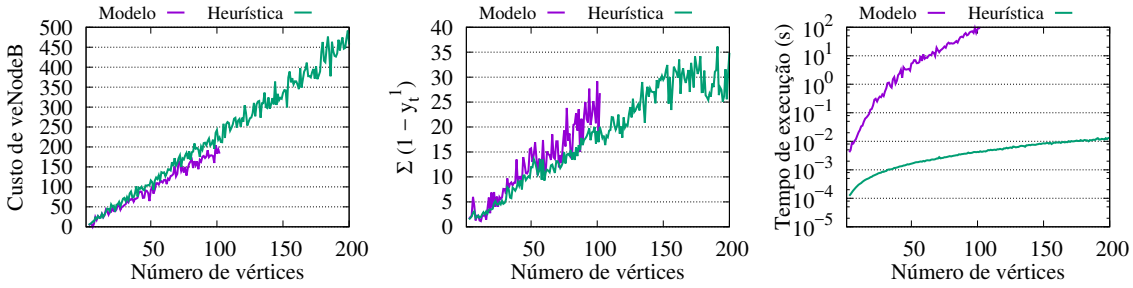

(a) Investimento em veNodeB. (b) Investimento em NFVI-Node. (c) Vértices que não são puramente virtuais.

**Figura 4. Investimento necessário e nível de centralização com $c_i^1$ e $c_i^2$ heterogêneos.**

otimização exata, através do nosso modelo ILP. Isto está ilustrado nas figuras 5(a), 5(b), e 5(c). Estes resultados correspondem à seguinte configuração: área de $12km$ x $12km$, $\alpha = 5$, custo heterogêneo $c_i^2 = 2c_i^1$, e capacidade heterogênea. A heurística apresenta valores próximos à abordagem exata em todas as métricas, e.g., o investimento em veNodeB na figura 5(a) e número de vértices que não são puramente virtuais na figura 5(b). No entanto, o tempo de computação consumido pela heurística é notavelmente menor do que o apresentado através do modelo, o que permite encontrar solução para redes de até 200 vértices, capazes de representar o cenário de uma operadora em Goiânia, segundo a Tabela 2.

## 6. Conclusão

Uma evolução disruptiva da rede de acesso de rádio para uma abordagem centralizada e virtualizada pode ser uma solução atraente para gerenciar, expandir e criar serviços através da rede. No entanto, isso implicaria em um aumento excessivo nos custos operacionais e de infraestrutura. Neste artigo, propomos um modelo de otimização de alocação de recursos que minimiza o custo de atualização de uma infraestrutura RAN, que pode ser composta por diferentes gerações de *hardware* e *software*, e.g., 2G, 3G e 4G. Nosso modelo permite que o tomador de decisão indique o nível de centralização e, como consequência, o quanto ele/ela quer gastar na atualização da rede. Investigamos como vários parâmetros afetam o investimento necessário e o nível de centralização alcançado. Observamos que a




(a) Investimento em veNodeB. (b) Vértices que não são puramente virtuais. (c) Tempo de execução.

**Figura 5. Comparação entre o modelo e a heurística com $c_i^1, c_i^2$, e $b_i$ heterogêneos.**

capacidade dos vértices tem forte impacto no nível de centralização alcançado. Uma vez que uma abordagem de otimização exata através do nosso modelo está limitado a encontrar soluções para redes de menos de 15 vértices, em um cenário homogêneo, desenvolvemos e avaliamos uma heurística para o problema. Nossa heurística alcançou desempenho semelhante à abordagem de otimização inteira e conseguiu encontrar a solução para uma rede de 200 vértices em milissegundos.

Como trabalho futuro, pretendemos lidar com o impacto dos usuários não atendidos, já que nosso modelo atual pressupõe que nenhuma ação é possível quando o vértice está saturado, ou seja, toda a sua capacidade foi consumida. Uma vez que os recursos adicionais poderiam ser adquiridos, o nível do serviço não atendidos poderia ser diminuído. Também estamos interessados em evoluir a formulação e a solução do nosso problema para uma versão estocástica, onde a demanda seria a principal fonte de incerteza. No entanto, os recursos também podem apresentar incertezas, por exemplo, a melhoria promovida pelo eICIC ou o CoMP não são conhecidos de forma determinística. Por fim, temos interesse em trabalhar com a virtualização completa da RAN (vRAN, Virtualized Radio Access Networks), onde todas as funções, de qualquer camada, podem ser virtualizadas e localizadas em diferentes pontos da rede utilizando NFV e Edge Computing [Garcia-Saavedra et al. 2018].

## Agradecimentos

Este trabalho foi parcialmente financiado pela Fundação de Amparo à Pesquisa do Estado de Goiás (FAPEG) na chamada No. 05/2016, processo No. 201610267001193. O trabalho também foi parcialmente financiado pela Rede Nacional de Ensino e Pesquisa (RNP).

## Referências

Asensio, A., Saengudomlert, P., Ruiz, M., and Velasco, L. (2016). Study of the Centralization Level of Optical Network-Supported Cloud RAN. In *2016 International Conference on Optical Network Design and Modeling (ONDM)*, pages 1–6.

Azarmand, Z. and Neishabouri, E. (2009). *Location Allocation Problem*. Physica-Verlag HD.

Basta, A., Kellerer, W., Hoffmann, M., Morper, H. J., and Hoffmann, K. (2014). Applying NFV and SDN to LTE Mobile Core Gateways, the Functions Placement Problem. In

*Proceedings of the 4th Workshop on All Things Cellular: Operations, Applications, and; Challenges*, pages 33–38.

Baumgartner, A., Reddy, V. S., and Bauschert, T. (2015). Mobile core network virtualization: A model for combined virtual core network function placement and topology optimization. In *Proceedings of the 2015 1st IEEE Conference on Network Softwarization (NetSoft)*, pages 1–9.

Bouet, M., Leguay, J., Combe, T., and Conan, V. (2015). Cost-based placement of vDPI functions in NFV infrastructures. *International Journal of Network Management*, 25(6):490–506.

Checko, A., Christiansen, H. L., Yan, Y., Scolari, L., Kardaras, G., Berger, M. S., and Dittmann, L. (2015). Cloud RAN for Mobile Networks - A Technology Overview. *IEEE Communications Surveys Tutorials*, 17(1):405–426.

Chih-Lin, Yuan, Y., Huang, J., Ma, S., Cui, C., and Duan, R. (2015). Rethink fronthaul for soft RAN. *IEEE Communications Magazine*, 53(9):82–88.

ElSawy, H., Hossain, E., and Haenggi, M. (2013). Stochastic Geometry for Modeling, Analysis, and Design of Multi-Tier and Cognitive Cellular Wireless Networks: A Survey. *IEEE Communications Surveys & Tutorials*, 15(3):996–1019.

ETSI (2014). Network Functions Virtualisation (NFV); Terminology for Main Concepts in NFV. `http://www.etsi.org/deliver/etsi_gs/NFV/001_099/003/01.02.01_60/gs_NFV003v010201p.pdf`. [Último acesso: 28-03-2017].

Garcia-Saavedra, A., Iosifidis, G., Costa-Perez, X., and J.Leith, D. (2018). FluidRAN: Optimized vRAN/MEC Orchestration. In *INFOCOM 2018*.

Lee, D., Zhou, S., Zhong, X., Niu, Z., Zhou, X., and Zhang, H. (2014). Spatial modeling of the traffic density in cellular networks. *Wireless Communications, IEEE*, 21(1):80–88.

Musumeci, F., Bellanzon, C., Carapellese, N., Tornatore, M., Pattavina, A., and Gosselin, S. (2016). Optimal BBU Placement for 5G C-RAN Deployment Over WDM Aggregation Networks. *Journal of Lightwave Technology*, 34(8):1963–1970.

Nunes, B., Mendonca, M., Nguyen, X. N., Obraczka, K., and Turletti, T. (2014). A Survey of Software-Defined Networking: Past, Present, and Future of Programmable Networks. *IEEE Communications Surveys Tutorials*, 16(3):1617–1634.

Telebrasil (2017). Telebrasil. `http://www.telebrasil.org.br/panorama-do-setor/mapa-de-erbs-antenas/`.