Modelo e Avaliação da Recuperação de Conteúdos Através de Funções de Rede Virtuais na Arquitetura de Computação na Borda em Redes Móveis

Ian Vilar Bastos[1,2], **Igor Monteiro Moraes**[1,2], **Nguyen Thi-Mai-Trang**[2], **Guy Pujolle**[2]

[1]Instituto de Computação – Laboratório Mídiacom – Universidade Federal Fluminense (UFF)
Niterói – RJ – Brasil

ianvilar@id.uff.br, igor@ic.uff.br

[2]Laboratoire d'Informatique de Paris 6 – Sorbonne Université
Paris – France

{Thi-Mai-Trang.Nguyen, Guy.Pujolle}@lip6.fr

Abstract. *The mobile traffic grows every year. Resource requirements of mobile applications, as processing power and storage capacity, transformed the architecture of cellular networks into a centralized infrastructure, the C-RANs. The centralized nature of C-RAN allows an efficient resource management through virtualization techniques and opens the horizon of ubiquitous computing for the Internet of Things. On the other hand, providing resources close to base stations, as the Mobile Edge Computing suggests, allows an immediate processing for delay sensitive applications. In this work, we formulate as a mixed integer linear programming the retrieval of media contents through caches acting as virtual network functions. The problem has as a solution the minimization of the retrieval cost as well as the minimization of virtual network functions.*

Resumo. *O aumento no consumo de tráfego móvel pelos usuários cresce a cada ano. A exigência por recursos pelas aplicações de dispositivos móveis, como poder de processamento e capacidade de armazenamento, fez com que a arquitetura das redes móveis se redesenhasse para uma estrutura centralizada, as C-RANs. A natureza centralizada da C-RAN possibilita gerenciar os recursos mais eficientemente através de técnicas de virtualização e abre o horizonte da computação ubíqua para processar os dados do paradigma da Internet das Coisas. Por outro lado, prover recursos computacionais próximos às estações base, como sugere a arquitetura de computação na borda, permite que aplicações sensíveis ao atraso possuam um processamento mais próximo do usuário. Neste trabalho, formula-se um problema de programação linear inteira mista para o posicionamento dos caches como funções de rede virtuais o local de recuperação dos conteúdos por parte dos usuários. O problema possui como solução a minimização do custo de recuperação por parte dos usuários, assim como a minimização da implantação de funções de rede virtuais.*

1. Introdução

O tráfego gerado por dispositivos móveis cresce ano após ano. Espera-se que o tráfego provindo das redes móveis alcance 49 exabytes por mês em 2021 [CISCO 2017].

Um enorme esforço é feito para evoluir o padrão e a tecnologia *Long Term Evolution* (LTE) com o objetivo de acomodar os futuros requisitos dos usuários móveis. Diversas propostas são estudadas na camada física para aprimorar a eficiência espectral do padrão LTE, como o *Multiple Input Multiple Output* (MIMO) multi-usuário, o MIMO maciço e o uso de pequenas células heterogêneas (*Heterogeneous and Small Cell Networks* - HetSNets). Embora as propostas para a camada física aumentem a capacidade da rede, a interferência entre as células continua um problema e o aumento no número de células resulta em um alto custo de implantação [Checko et al. 2015]. Para superar as limitações na capacidade das redes celulares fora do espectro físico, duas arquitetura foram propostas, a *Cloud Radio Access Network* (C-RAN) e a *Mobile Edge Computing* (MEC).

A ideia por trás da arquitetura C-RAN está em centralizar todas as unidades de processamento de banda base (*Baseband Units* - BBU) em um ambiente virtualizado e compartilhar esses BBUs entre as diferentes redes celulares de acesso por demanda [Lin et al. 2010]. Ao desacoplar os recursos de processamento da BBU e as funcionalidades de rádio, diferentes estações base (*evolved NodeB* - eNB) pertencentes à mesma operadora de rede podem compartilhar recursos e informações sobre as aplicações em uso. A centralização da RAN permite integrar um alto poder de processamento e capacidade de armazenamento no provedor da nuvem. Dessa forma, torna-se mais fácil implantar a computação ubíqua ao interconectar os vários dispositivos com os serviços compartilhados [Tran et al. 2017]. Por outro lado, enquanto centralizar os recursos abre novos horizontes para o paradigma da Internet das Coisas (*Internet of Things* - IoT), a troca de informações entre as antenas das eNBs e o ambiente virtualizado no qual se encontram as BBUs necessita de enlaces de alta vazão e com baixa latência.

Muitas aplicações sensíveis ao atraso, como as de *streaming* de vídeo, computação visual e inteligência artificial demandam respostas quase em tempo-real. Portanto, implantar serviços e recursos próximos aos usuários móveis pode preservar os requisitos de aplicações sensíveis ao atraso. O paradigma da arquitetura MEC utiliza nuvens locais e/ou regionais fisicamente mais próximas das eNBs para complementar a arquitetura C-RAN. Na arquitetura MEC, nós de borda atuam como um poder computacional adicional para satisfazer de uma forma mais imediata as requisições dos usuários ou para pré-processá-las antes de enviá-las para a nuvem centralizada, economizando os enlaces que conectam internamente a rede celular [Hu et al. 2015].

Em 2016, 60% do tráfego móvel foi derivado do tráfego de vídeo móvel [CISCO 2017]. Atualmente, todo tipo de mídia transferido pela Internet é recuperado através dos protocolos *Hypertext Transmission Protocol* (HTTP) e *Transmission Control Protocol* (TCP). A capacidade da transmissão fim-a-fim possui extremas variações em poucos segundos devido às condições do canal de rádio, operações de *handover* e do uso interno dos enlaces na rede celular. Como resultado dessas variações, o TCP não consegue estimar precisamente a capacidade de transmissão fim-a-fim e, logo, não só a qualidade experienciada (*Quality of Experience* - QoE) é afetada como os recursos são sub-utilizados.

Nesse trabalho, formula-se como um problema de programação inteira linear mista a recuperação de mídias através da arquitetura MEC na qual as mídias são armazenadas em funções de rede virtuais atuando como *caches*. As topologias consideradas

nas avaliações são derivadas de um conjunto de dados disponibilizados pela prefeitura de Paris, e as requisições pelas mídias são modeladas através de um conjunto de dados disponibilizados pela Netflix. Os resultados de simulação evidenciam um compromisso entre o custo de recuperação das mídias e o custo associado à instanciação dos *caches* em redes de altos diâmetros. Observa-se também um compromisso entre a capacidade de armazenamento dos *caches* e o número de cópias e seus locais de armazenamento na rede para as mídias mais requisitadas entre as diferentes eNBs.

O restante deste trabalho está organizado da seguinte forma. A Seção 2 apresenta uma visão geral da arquitetura das redes celulares e do paradigma de computação na borda em redes móveis. A Seção 3 apresenta as notações matemáticas utilizadas e a formulação do problema. A Seção 4 apresenta e discute os resultados obtidos através das simulações. A Seção 5 apresenta os trabalhos relacionados e a Seção 6 conclui este trabalho.

2. Visão Geral da Arquitetura de Computação na Borda em Redes Móveis

A rede celular conecta os usuários móveis, ou os equipamentos de usuários (*User Equipment* - UE), ao núcleo da rede (*Evolved Packet Core* - EPC). A conexão entre os UEs e a EPC é feita normalmente por meio da RAN via um ponto de acesso sem fio, a eNB. A EPC é essencialmente composta de um *serving gateway* (S-GW), um *packet data network gateway* (P-GW), um *home subscriber* (HSS), um *mobility management entity* (MME) e um *policy and charging rules function* (PCRF). O S-GW é responsável pelo encaminhamento de pacotes dentro da EPC e atua como um gerenciador de mobilidade quando os UEs se movem de uma eNB para outra, realizando todo o processo de *handover*. O P-GW provê conectividade entre os UEs e qualquer tipo de rede externa, atuando como o nó de saída para a Internet. O HSS é a base de dados central da rede celular, na qual as informações sobre assinaturas de todos os usuários da operadora de rede estão armazenados. É responsável pelo estabelecimento de ligações e sessões, autenticação de usuários e a autorização de acesso. O MME é responsável por rastrear a localização dos UEs e seu estado dentro da rede. Quando os usuários são autenticados pelo HSS, o MME gerencia o estabelecimento das portadoras para a transmissão de dados ao selecionar os S-GW e P-GW apropriados. Todos os componentes são conectados uns aos outros através da conexão interna da rede celular, chamada de *backhaul*.

O paradigma da arquitetura MEC possui como foco prover um ambiente de nuvem na borda das redes celulares, na qual recursos de virtualização estão localizados dentro da RAN e próximos aos UEs [Boccardi et al. 2014]. A ideia está em reduzir o uso dos enlaces do *backhaul* e em prover baixa latência aos serviços requisitados pelos UEs. A arquitetura MEC é implantada sobre um ambiente virtualizado, o qual opera em conjunto com o paradigma da virtualização das funções de rede (*Network Function Virtualization* - NFV). Sua principal diferença para o paradigma empregado pela arquitetura MEC está em seu foco. Enquanto a MEC foca na virtualização de aplicações móveis o NFV foca na virtualização das funções de rede [Hu et al. 2015]. O NFV instancia funções como *gateways*, *firewalls*, sistemas de detecção de intrusão, *web caches*, balanceadores de carga, entre outros. A arquitetura MEC pode se beneficiar desse ambiente virtualizado para as funções de rede já existente para implantar o paradigma das aplicações móveis virtualizadas. Além disso, pode também utilizar essas funções de rede para atuar em conjunto com as aplicações instanciadas na rede celular.

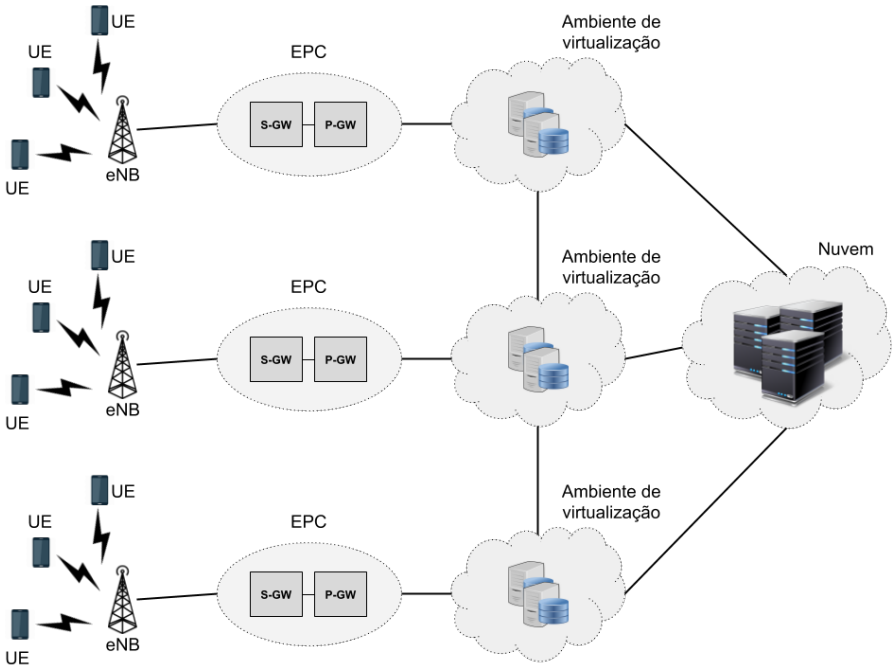


Figura 1. A arquitetura *Mobile Edge Computing*.

A Figura 1 ilustra um exemplo da arquitetura MEC. A rede de acesso para os usuários móveis pode ser qualquer tipo de infraestrutura sem fio, como WiFi, celular ou ambos, que dispõem de um *backhaul* cabeado para interconectar os diferentes ambientes virtualizados. O ambiente virtualizado pode ser disposto mais próximo dos usuários móveis ou da saída para a Internet dependendo dos requisitos que a operadora da rede deseja garantir. Independentemente da topologia de rede adotada pela operadora de rede, em árvore, anel ou em malha, a característica mais importante está na capacidade de criar enlaces virtuais com o objetivo de interconectar os pares de nós pertencentes ao mesmo nível hierárquico [Ceselli et al. 2017]. Com o advento das VNFs, até mesmo os nós que compõem a EPC podem ser máquinas virtuais e interconectados por enlaces virtuais.

3. Formulação do Problema

Neste trabalho, formula-se a obtenção de mídias pelos UEs ao recuperá-las de VNFs atuando como *caches* em um problema de programação linear inteira mista. O problema pode ser interpretado como a atribuição dos *caches* aos ambientes virtuais associados às eNBs e a atribuição das mídias aos *caches* disponíveis para satisfazer uma ou mais requisições dos UEs.

A rede é modelada como um grafo não-direcionado $G(N, E)$, no qual N é o conjunto de eNBs e E é o conjunto de enlaces que interconectam as eNBs. A função $D_{i,j}$, tal que $i, j \in N$, define a função de custo para que a eNB i alcance a eNB j. A distribuição de requisições de cada eNB j é representada por λ_j. φ representa o conjunto de todas as mídias disponíveis que podem ser requisitadas pelos UEs e $\mu_\nu[bytes]$ é o tamanho de cada mídia $\nu \in \varphi$. O conjunto ψ representa o conjunto de todas as *cache* VNFs disponíveis para serem alocadas às eNBs. Considera-se que $U \subseteq N$ como o conjunto de eNBs que hospedam um *cache* VNF. A capacidade de armazenamento de cada *cache*

VNF instanciada $u \in U$ é representada por θ_u. Com o objetivo de representar o esforço associado em instanciar um *cache* VNF $k \in \psi$ em uma eNB $j \in N$, assume-se um custo de configuração p_k^j.

A formulação também possui três variáveis de solução. A variável binária t_k^j indica se a eNB j hospeda ou não um *cache* VNF k. A obtenção da mídia ν pertencente a λ_j através da eNB u é representada pela variável binária $X_{j,u}^{\nu}$. Para definir que uma mídia ν está armazenada em uma eNB u, utiliza-se uma variável binária auxiliar δ_u^{ν}. Essa variável auxiliar indica que a mídia ν somente pode estar disponibilizada na eNB u caso a mídia ν tenha sido previamente requisitada por um UE e consequentemente armazenada na eNB u, como mostra a Equação 1.

$$\delta_u^{\nu} = \begin{cases} 1, & \sum_{j \in N} X_{j,u}^{\nu} > 0 \\ 0, & \text{caso contrário} \end{cases} \tag{1}$$

A formulação desenvolvida é apresentada a seguir e todas as notações utilizadas nesta formulação estão resumidas na Tabela 1.

$$\text{minimizar} \sum_{j \in N} \sum_{\nu \in \lambda_j} \sum_{u \in U} D_{j,u} X_{j,u}^{\nu} + \sum_{u \in U} \sum_{k \in \psi} t_k^u p_k^u \tag{2}$$

$$\text{sujeito a} \sum_{k \in \psi} t_k^j \leq 1 \qquad ,\forall j \in N \tag{3}$$

$$\sum_{j \in N} X_{j,u}^{\nu} \leq W.\delta_u^{\nu} \qquad , \forall u \in U, W \geq |N| \tag{4}$$

$$\sum_{j \in N} X_{j,u}^{\nu} \leq \sum_{k \in \psi} t_k^u \qquad , \forall u \in U \tag{5}$$

$$\sum_{\nu \in V} \mu_{\nu} \delta_u^{\nu} \leq \theta_u \qquad , \forall u \in U \tag{6}$$

A função objetivo da formulação apresentada é definida na Equação 2. A função objetivo define que o objetivo está em minimizar o custo da distância entre a eNB j, na qual a mídia ν é requisitada, e a eNB u, na qual a mídia ν está armazenada em seu *cache* VNF. A função objetivo também define que a quantidade de *cache* VNFs instanciados na rede deve ser minimizada. Nesta formulação, assume-se que há recursos disponíveis em todas as eNBs para instanciar um *cache* VNF e a Equação 3 define que cada eNB j pode instanciar no máximo um *cache* VNF. A Equação 4 define que para a eNB j recuperar a mídia ν da eNB u, a eNB u deve possuir a mídia ν armazenada em seu *cache* VNF. Da mesma forma, a Equação 5 complementa a Equação 4 ao definir que as mídias só podem estar armazenadas na eNB u caso a eNB u possua um *cache* VNF. A última restrição apresentada pela Equação 6 define que a capacidade de armazenamento dos *cache* VNFs instanciados não seja ultrapassada.

A formulação apresentada nesta seção é derivada do problema de atribuição generalizado (*Generalized Assignment Problem* - GAP). O problema tem como objetivo alocar n tarefas em l máquinas com a restrição de que cada tarefa $j = 1, ..., n$ deve ser atribuída

exatamente a uma única máquina. Cada máquina m possui espaço para acomodar um determinado número de tarefas e cada tarefa ocupa um certo espaço da máquina ao ser atribuído a ela. O processo de atribuir uma tarefa l a uma máquina m incorre em um custo de atribuição. O objetivo do problema está em encontrar a atribuição de todas as tarefas com o menor custo de atribuição possível. Como o GAP é um problema NP-difícil e a formulação deste trabalho é derivada de sua formulação, a complexidade computacional da formulação apresentada também é não-polinomial e cresce com a combinações formadas pelo número de eNBs na rede, com o número de eNBs requisitantes e o número de mídias requisitadas.

Tabela 1. Notações utilizadas na formulação.

Notação	Descrição	Tipo
N	eNBs	Conjunto
E	Enlaces que interconectam as eNBs	Conjunto
φ	Mídias	Conjunto
λ_j	Distribuição das requições na eNB j	Conjunto
U	eNB que contém um *cache* VNF	Conjunto
μ_ν	Tamanho de cada mídia ν em bytes	Parâmetro
θ_u	Capacidade de cada *cache* VNF u instanciado	Parâmetro
p_k^j	Custo de configuração do *cache* VNF k na eNB j	Parâmetro
t_k^j	eNB j que hospeda um *cache* VNF k	Variável
$X_{j,u}^\nu$	Requisição na eNB j para a mídia ν satisfeita pela eNB u	Variável
δ_u^ν	Mídia ν armazenada na eNB u	Variável

4. Resultados

Para avaliar como as mídias se distribuem na rede, assim como são instanciados os *cache* VNFs, em relação à capacidade de armazenamento e do custo de configuração de VNFs, implementa-se um simulador na linguagem de programação Python. O simulador gera as topologias utilizadas, as requisições dos UEs e o modelo de otimização. As topologias geradas são baseadas em um conjunto de dados sobre os *hotsposts* WiFi disponibilizado pela prefeitura de Paris [Paris 2018]. As três topologias utilizadas nas avaliações são construídas ao utilizar os identificadores e as coordenadas de latitude/longitude dos *hotspots* WiFi.

$$d_{x,y} = 2\ R\ \arcsin\left(\sqrt{\sin^2\left(\frac{lat_y - lat_x}{2}\right) + \cos(lat_x)\cos(lat_y)\sin^2\left(\frac{lon_y - lon_x}{2}\right)}\right) \tag{7}$$

Para gerar a topologia, assume-se que dois *hotsposts* separados por um determinado limiar de distância estão de alguma forma conectados. As três topologias são geradas ao utilizar como limiares 100 m, 300 m e 500 m. Cada limiar gera múltiplas componentes conexas e a maior componente conexa é selecionada como a topologia para o respectivo limiar. O custo dos enlaces é representado pelo inverso da distância entre dois nós. A distância entre dois nós $d_{x,y}$ é a distância de Haversine calculada através da Equação 7, na

qual $R = 6371$ Km é o raio da Terra e as variáveis lat e lon são as respectivas posições de latitude e longitude das eNBs vizinhas x e y. O custo total entre quaisquer pares de eNBs i e j é definido como a soma do inverso das distâncias entre todas as eNBs que compõem o menor caminho entre as eNBs i e j, como mostra a Equação 8. As características de cada topologia estão resumidas na Tabela 2.

$$D_{i,j} = \sum_{(x,y) \in SP(i,j)} \frac{1}{d_{x,y}} \tag{8}$$

Tabela 2. Características das topologias geradas.

Limiar	# de nós	# de enlaces	Densidade	Diâmetro
100 m	9	29	0.805556	2
300 m	17	84	0.617647	5
500 m	59	249	0.145529	11

Para a requisição de mídias, considera-se um conjunto de dados disponibilizado pela Netflix [Bennett et al. 2007] como modelo para gerar as requisições. O conjunto de dados é composto por 100 milhões de avaliações em uma escala de 1 a 5 que estão distribuídas sobre aproximadamente 18 mil filmes, como mostra a Figura 2. Assume-se que as mídias possuem diferentes popularidades e que a probabilidade de uma mídia ser requisitada é proporcional ao número de avaliações que ela recebeu pelos usuários no conjunto de dados. A probabilidade é gerada ao normalizar a quantidade de avaliações recebidas pelo soma das avaliações de todos os filmes.

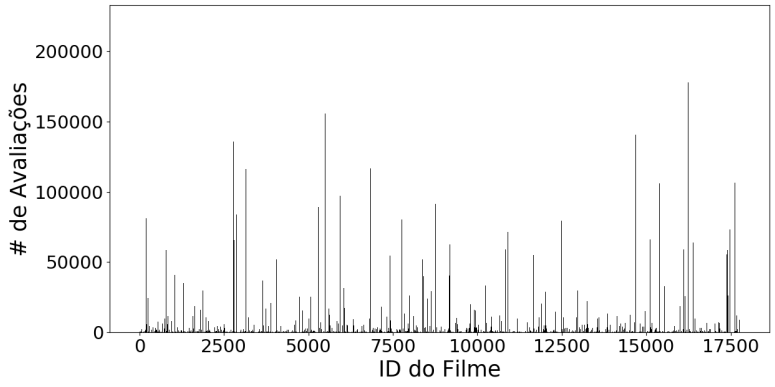


Figura 2. Distribuição das avaliações dos usuários sobre os diferentes filmes no conjunto de dados da Netflix.

Os parâmetros do simulador são o número de eNBs requisitantes, a quantidade de mídias a serem requisitadas, o tamanho das mídias, a capacidade de armazenamento dos *cache* VNFs e o custo de configuração dos *cache* VNFs. Neste trabalho, assume-se que todas as mídias possuem o mesmo tamanho, todas as eNBs requisitam a mesma quantidade de mídias e os *cache* VNFs possuem a mesma capacidade de armazenamento. As eNBs requisitantes são selecionadas aleatoriamente seguindo uma distribuição uniforme desconsiderando-se o nó de controle (NC). Representa-se o custo de configuração dos

cache VNFs p_k^j como a distância $D_{nc,j}$ para que o NC alcance a eNB j multiplicado por uma constante de penalidade α, isto é, $p_k^j = \alpha \times D_{nc,j}$. O nó escolhido na topologia para representar o NC é o nó que possui o maior valor de intermediação. A intermediação é uma medida de centralidade na rede cujo valor é calculado pela quantidade de menores caminhos de todos os nós para quaisquer outros nós da rede que passam pelo nó avaliado. A parametrização das avaliações está resumida na Tabela 3 e o modelo de otimização é solucionado através do IBM ILOG CPLEX Optimization Studio 12.8.0.

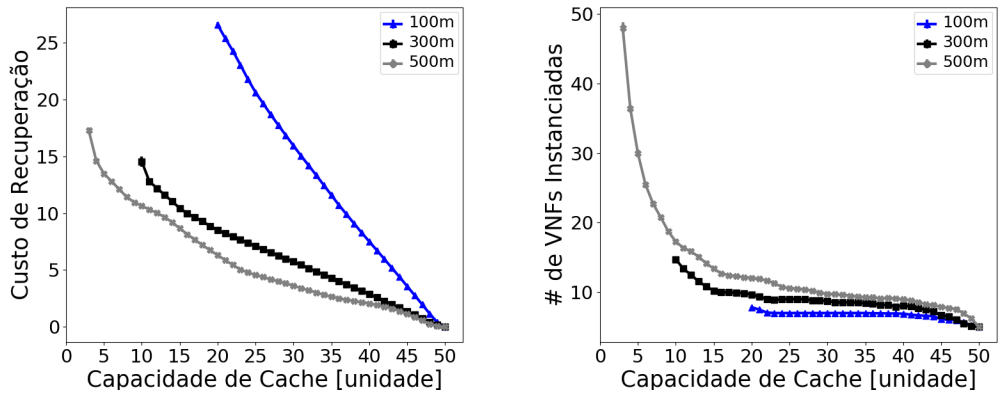
Tabela 3. Parametrização

Parâmetro	**Valor**
# de eNBs requisitantes	5 eNBs
# de mídia requisitadas por eNB	50 mídias
Tamanho de *cache*	[20, 50] mídias
Fator de penalidade α	[0,2]

A Figura 3 apresenta os resultados para as três topologias quando não há uma penalidade para instanciar as VNFs e a otimização busca somente minimizar o custo de recuperação das mídias. Pela Figura 3(b) é possível observar que conforme a rede celular comporta mais instâncias, menor é o tamanho de *cache* necessário para que haja soluções viáveis. Como a quantidade de VNFs instanciadas não impacta a solução final, independentemente do tamanho do *cache*, a topologia de 500 m mantém um número de instâncias superior às demais topologias. Ao possuir uma diversidade maior de enlaces que podem ser utilizados para a recuperação das mídias, o custo para recuperá-las é reduzido, como mostra a Figura 3(a). Outra característica importante de ser observada é que, quando a rede possui recursos o suficiente para instanciar VNFs com tamanho de *cache* capaz de armazenar todas as mídias solicitadas, a diversidade de nós e enlaces não é um fator predominante. Tamanhos de *cache* próximos a 50 unidades já são capazes de armazenar todas as mídias requisitadas na simulação, portanto é suficiente instanciar as VNFs nas eNBs nas quais os UEs estão solicitando as mídias e armazenar uma cópia de cada mídia nas VNFs instanciadas sem a necessidade de percorrer os enlaces de *backhaul* da rede celular.

Visto que todas as três topologias apresentam soluções viáveis a partir de uma capacidade de *cache* em 20 unidades, as Figuras 4-6 mostram os resultados com as capacidades de *cache* variando entre 20 e 50 unidades. A Figura 4 apresenta os resultados do custo de recuperação das mídias em função das configurações de capacidade em *cache* e fator de penalidade. É possível observar que, com exceção da topologia de 500 m com fator de penalidade $\alpha = 2$, existe uma tendência de reduzir o custo de recuperação das mídias conforme o tamanho dos *caches* aumente e com um leve aumento ao incrementar o fator de penalidade. Uma vez que os *cache* VNFs comportem um maior número de mídias, é natural que seja necessário um menor número de instâncias, como mostra a Figura 5, e que estas instâncias estejam localizadas mais próximas das eNBs nas quais as mídias estão sendo requisitadas.

A pequena variação no custo de recuperação das mídias para as topologias de 100 m e 300 m se deve ao diâmetro da rede. Visto que o diâmetro de ambas as rede é curto e o NC localiza-se em um posição de centralidade da rede, o custo associado aos poucos enlaces a serem percorridos para as instanciações dos *cache* VNFs é desprezível



(a) Custo de recuperação das mídias em função do tamanho de cache.

(b) VNFs instanciadas em função do tamanho de cache.

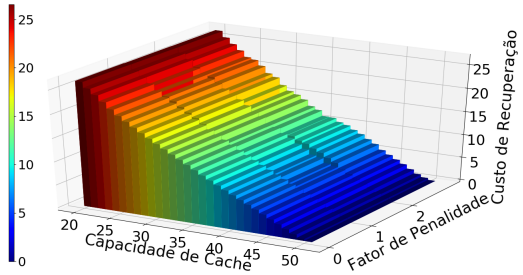
Figura 3. Avaliações do custo de recuperação de mídias e da quantidade de VNFs instanciadas com o fator de penalidade $\alpha = 0$.

comparado ao de recuperar inúmeras mídias com custos cada vez menores. Entretanto, o mesmo comportamento não corre com fator de penalidade $\alpha = 2$ para a topologia de 500 m. O valor $\alpha = 2$ indica que os enlaces percorridos para instanciar um *cache* VNF possuem o dobro do custo comparado ao de percorrer os enlaces para a recuperação das mídias. Dessa forma, em um rede com um alto diâmetro, existe um compromisso entre recuperar as mídias próximas às eNBs requisitantes e o NC necessitar percorrer mais enlaces para aproximá-las. Enquanto a capacidade do *cache* não for grande o suficiente para produzir soluções com uma redução significativa na quantidade de VNFs instanciadas, a melhor solução encontra-se em manter as VNFs próximas às eNBs requisitantes. Quando o número necessário de VNFs é reduzido, a tendência é que esse número se mantenha o mesmo até o momento no qual a capacidade do *cache* permite que as requisições sejam todas satisfeitas pela eNB na qual as requisições são recebidas.

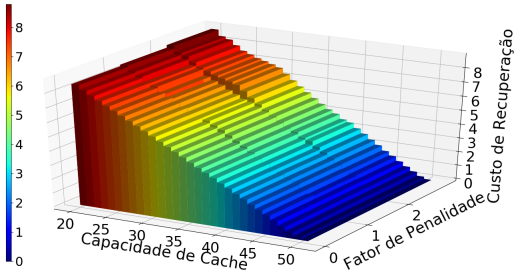
A Figura 6 torna este comportamento mais evidente. Os gráficos apresentam os resultados em função da popularidade das mídias para todas as configurações de capacidade de *cache* com fator de penalidade $\alpha = 2$, nos quais os índices mais baixos representam as mídias mais requisitadas. Observa-se que em todas as topologias, as mídias mais populares possuem a tendência de estarem armazenadas em poucas instâncias e consequentemente nas instâncias mais centrais. Dessa forma, todas as eNBs que estejam requisitando essa mesma mídia conseguem recuperá-la com um custo equivalente, enquanto as mídias menos populares ocupam a maior parte dos demais *caches*, que por sua vez são instanciados mais próximos somente das eNBs que as estão requisitando. Quando o tamanho do *cache* é capaz de armazenar todas as mídias requisitadas, a tendência é que haja uma cópia de cada mídia popular próxima às eNBs que as requisitam.

5. Trabalhos Relacionados

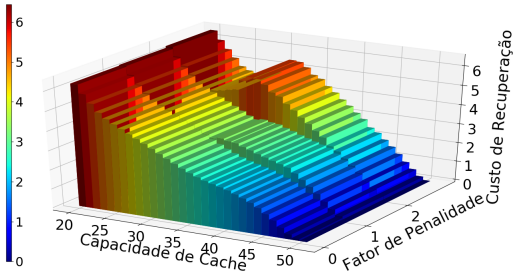
Baştuğ *et al.* [Baştuğ et al. 2014] discutem que as RANs devem mudar seu paradigma reativo para um paradigma proativo. Os autores argumentam que é mais viável somente tomar uma decisão após a chegada de fluxos de dados na RAN com o aumento exacerbado de UEs, *streaming* de vídeo, aplicações web e das redes sociais. As redes



(a) Custo de recuperação das mídias em todos os cenários para a topologia de 100 m.

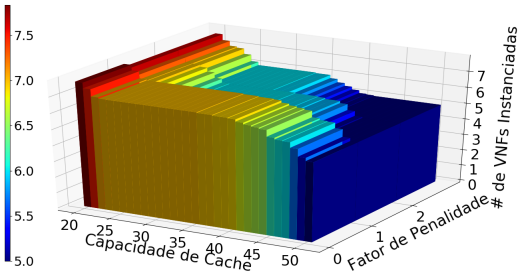


(b) Custo de recuperação das mídias em todos os cenários para a topologia de 300 m.

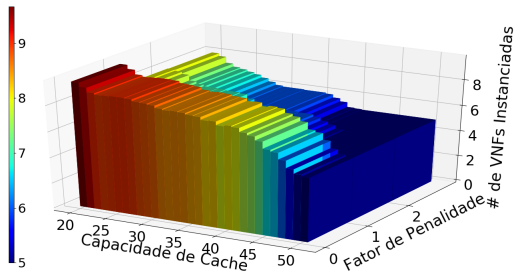


(c) Custo de recuperação das mídias em todos os cenários para a topologia de 500 m.

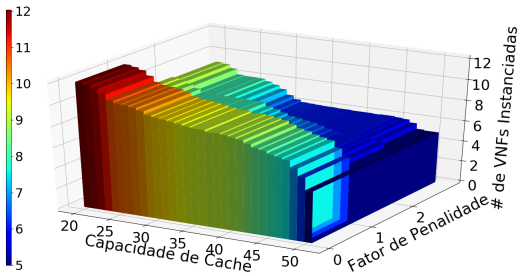
Figura 4. Avaliações do custo de recuperação de mídias para todos os fatores de penalidade em todas as topologias.



(a) Número de VNFs instanciadas em todos os cenários para a topologia de 100 m.



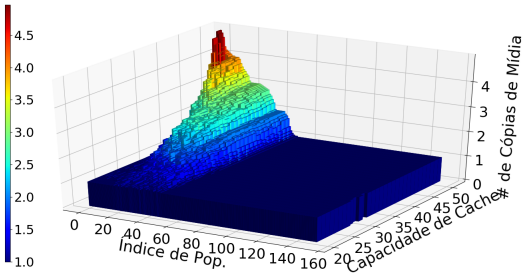
(b) Número de VNFs instanciadas em todos os cenários para a topologia de 300 m.



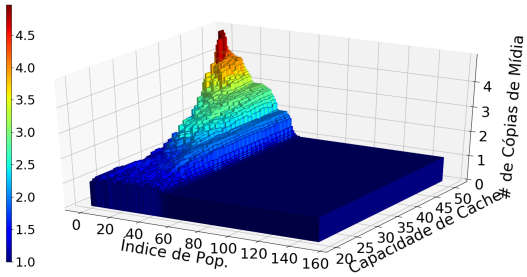
(c) Número de VNFs instanciadas em todos os cenários para a topologia de 500 m.

Figura 5. Avaliações do número de VNFs instanciadas para todos os fatores de penalidade em todas as topologias.

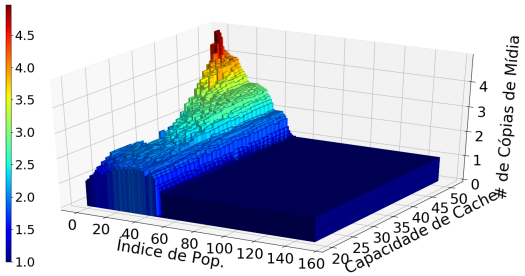
celulares devem proativamente explorar as informações de contexto dos usuários para antecipar futuras requisições. Afirma-se no trabalho que usuários móveis são previsíveis



(a) Relação entre o número de cópias e a popularidade da mídia na topologia de 100 m.



(b) Relação entre o número de cópias e a popularidade da mídia na topologia de 300 m.



(c) Relação entre o número de cópias e a popularidade da mídia na topologia de 500 m.

Figura 6. Relação entre a popularidade das mídias e o número de cópias armazenadas em diferentes *cache* VNFs para o fator de penalidade $\alpha = 2$ em todas as topologias.

até um certo grau e a rede celular pode pré-armazenar conteúdos que um conjunto de usuários possua alta probabilidade de requisitá-los. Como a capacidade de armazenamento de usuários móveis está cada vez maior, os autores defendem que os conteúdos sejam proativamente armazenados nos próprios UEs para serem acessados localmente, sem a necessidade de se comunicar com a RAN.

Paschos *et al.* [Paschos et al. 2016] argumentam que a infraestrutura celular evoluiu para acompanhar o crescimento de dispositivos móveis que estão acessando a Internet. No entanto, a explosão de RANs e de UEs tornam os enlaces de *backhaul* congestionados caso nenhuma outra solução seja proposta para reduzir sua carga. Nesse sentido, armazenar conteúdos na borda da rede celular possui o potencial em aliviar a infraestrutura cabeada das redes celulares. Entretanto, técnicas de *caching* normalmente empregadas em redes de distribuição de conteúdos não são apropriadas para as redes celulares, visto que o número de conteúdos alcançáveis é ordens de magnitude maior e sua popularidade varia bruscamente com o passar do tempo. Políticas de retirada de *cache* baseadas em modelos de popularidade estáticos, como a *Least Recently Used* (LRU), não produzem ganhos satisfatórios. As políticas devem se encaixar no período que os usuários acessam os conteúdos como pulsos, e a amplitude desses pulsos serem vistas como popularidades instantâneas para produzir melhores estimativas de popularidade e, portanto, de *caching* [Leconte et al. 2016]. Os autores também afirmam que algoritmos de aprendizado de máquina e *big data* serão primordiais para rastrear as variações de popularidade.

Li *et al.* [Li et al. 2015, Li et al. 2017] propõem o uso de *caching* como serviço (*Cachin as a Service* - CaaS) nas redes celulares. Com a tendência em virtua-

lizar estações base em conjuntos de BBUs, assim como os elementos do EPC, as redes 5G devem incorporar o *caching* como parte do ambiente virtual para aprimorar a entrega de serviços multimídia. Os autores discutem os diferentes prós e contras relacionados ao local de implantação dos recursos de *caching*. Recursos físicos mais próximos das eNBs são naturalmente mais escassos, o que torna mais difícil produzir um grande número de acertos em *cache*. Ao mover o ambiente virtual mais adentro da EPC os recursos virtuais das *cache* VMs aumentarão, assim como os acertos em *cache*. Por outro lado, a complexidade de roteamento dentro da EPC aumentará, assim como a carga nos enlaces de *backhaul*. Além disso, o controlador centralizado do CaaS terá um papel fundamental com as *cache* VMs livres para serem instanciadas em qualquer ambiente da rede celular com o objetivo de atender os requisitos de QoS e dos usuários móveis. O controlador será responsável por aumentar e diminuir a escala de *cache* VMs instanciadas, na decisão do local de instanciação das *cache* VMs, em configurar rotas eficientes e em pré-armazenar conteúdos baseando-se nas aplicações utilizadas pelo usuários móveis.

Moon *et al.* [Moon et al. 2017] propõem um sistema de virtualização de *cache* baseado em NFV e em um modelo hierárquico de *caching*. Os autores projetam o plano de dados e o plano de controle do sistema. O plano de dados é responsável por receber as requisições dos usuários móveis através de um protocolo conhecido como *Common Public Radio Interface* (CPRI). O roteador de borda verifica se o conteúdo requisitado já está armazenado em um *cache* de borda ao analisar as informações do cabeçalho HTTP através da técnica de inspeção profunda de pacotes (*Deep Packet Inspection* - DPI). A análise do fluxo recebido é utilizada para consultar a tabela de encaminhamento (*Forwarding Information Base* - FIB). Caso o conteúdo requisitado não esteja armazenado em um *cache* virtual na borda da rede, o roteador de borda encaminha o fluxo para o S-GW através do protocolo de tunelamento (*General Packet Radio Service Tunneling Protocol* - GTP). O S-GW consulta sua FIB para determinar se o conteúdo requisitado está armazenado no EPG. Se não for o caso, a requisição é encaminhada para o P-GW para ser recuperado pela Internet. O conteúdo requisitado segue o caminho inverso dentro da rede celular para ser entregue ao UE, e o orquestrador de *caches* virtuais- (*Virtual Cache Orchestrator* - vCO) decide o *cache* virtual no qual o conteúdo será armazenado. No plano de controle, o vCO é responsável em calcular o roteamento ótimo e em obter informações sobre os *caches* virtuais.

6. Conclusão

O paradigma da computação na borda em redes móveis é mais um esforço para a próxima geração das redes celulares em fornecer conexões de alta vazão e latências baixas ao ponto de haver respostas em 1 ms. A evolução do paradigma da computação nas redes de computadores permite criar um balanço no qual as informações são armazenadas e processadas no local que torna mais conveniente a sua transferência ao usuário final. Em uma rede na qual a transferência de dados cresce, aproximar as informações e o processamento de informações dos usuários móveis é uma tendência na implantação das redes 5G.

Neste trabalho, formulou-se um modelo de otimização para estudar como as mídias requisitadas em diferentes estações base que possuam um certo grau de redundância são armazenadas. Para armazená-las, assume-se um ambiente virtual próximo às estações base no qual são instanciados *caches* como funções de rede virtuais. Através

da metodologia utilizada, o estudo aponta vantagens em armazenar poucas cópias de mídias populares e em nós mais centrais quando os recursos são mais escassos ou quando o processo de configuração de funções virtuais é custoso para a rede. Além disso, redes com diâmetros grandes apresentam um compromisso em instanciar os *caches* mais próximos às estações base quando o custo associado à percorrer os enlaces para instanciá-los é maior que o custo associado ao da recuperação das mídias pelos usuários móveis.

Como trabalhos futuros, pretende-se desenvolver heurísticas para a alocação dos *caches* e do armazenamento das mídias ao levar em consideração informações como popularidade das mídias, dimensões da rede, políticas de atribuição e descarte nos *caches* e o processo de configuração das funções virtuais. Pretende-se também estudar a estimação da popularidade das mídias requisitadas pelos usuários móveis através da correlação de atributos presentes em metadados das mídias utilizando algoritmos de aprendizado de máquina.

Agradecimentos

Os autores agradecem ao CNPq, CAPES, FAPERJ, Proppi/UFF, TBE/ANEEL, CELESC/NeoDomino/ANEEL and FAPESP pelo suporte financeiro para o desenvolvimento desta pesquisa.

Referências

Baştuğ, E., Bennis, M., and Debbah, M. (2014). Living on the edge: The role of proactive caching in 5g wireless networks. *IEEE Communications Magazine*, 52(8):82–89.

Bennett, J., Lanning, S., et al. (2007). The netflix prize. In *Proceedings of KDD cup and workshop*, volume 2007, page 35. New York, NY, USA.

Boccardi, F., Heath, R. W., Lozano, A., Marzetta, T. L., and Popovski, P. (2014). Five disruptive technology directions for 5g. *IEEE Communications Magazine*, 52(2):74–80.

Ceselli, A., Premoli, M., and Secci, S. (2017). Mobile edge cloud network design optimization. *IEEE/ACM Transactions on Networking (TON)*, 25(3):1818–1831.

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.

CISCO (2017). Cisco visual networking index: Global mobile data traffic forecast update, 2016–2021. Technical report. `https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html`.

Hu, Y. C., Patel, M., Sabella, D., Sprecher, N., and Young, V. (2015). Mobile edge computing—a key technology towards 5g. *ETSI white paper*, 11(11):1–16.

Leconte, M., Paschos, G., Gkatzikis, L., Draief, M., Vassilaras, S., and Chouvardas, S. (2016). Placing dynamic content in caches with small population. In *INFOCOM 2016-The 35th Annual IEEE International Conference on Computer Communications, IEEE*, pages 1–9. IEEE.

Li, X., Wang, X., Li, K., and Leung, V. C. (2017). Caas: Caching as a service for 5g networks. *IEEE Access*, 5:5982–5993.

Li, X., Wang, X., Zhu, C., Cai, W., and Leung, V. C. (2015). Caching-as-a-service: Virtual caching framework in the cloud-based mobile networks. In *INFOCOM Workshops*, pages 372–377.

Lin, Y., Shao, L., Zhu, Z., Wang, Q., and Sabhikhi, R. K. (2010). Wireless network cloud: Architecture and system requirements. *IBM Journal of Research and Development*, 54(1):4–1.

Moon, S., Shin, Y., Chung, S., and Kim, S. (2017). Design and analysis of virtualized caching service on cellular infrastructure. In *Proceedings of the 15th ACM International Symposium on Mobility Management and Wireless Access*, pages 45–54. ACM.

Paris (2018). Liste des sites municipaux équipés d'un point d'accès WiFi - Mairie de Paris / Direction des Systèmes et Technologies de l'Information, 30/05/2018, sous license ODbL. Technical report. `https://opendata.paris.fr/explore/dataset/liste_des_sites_des_hotspots_paris_wifi/information/`.

Paschos, G., Bastug, E., Land, I., Caire, G., and Debbah, M. (2016). Wireless caching: Technical misconceptions and business barriers. *IEEE Communications Magazine*, 54(8):16–22.

Tran, T. X., Hajisami, A., Pandey, P., and Pompili, D. (2017). Collaborative mobile edge computing in 5g networks: New paradigms, scenarios, and challenges. *IEEE Communications Magazine*, 55(4):54–61.