Política Cooperativa Orientada à Rede para Posicionamento de Cache e Roteamento de Requisições em Redes Celulares Heterogêneas

Marisangila Alves[1], **Guilherme Piêgas Koslovski**[1]

[1]Programa de Pós Graduação de Computação Aplicada - PPGCAP
Universidade do Estado de Santa Catarina - UDESC

marisangila.alves@edu.udesc.br, guilherme.koslovski@udesc.br

Abstract. *The 5G networks triggered an evolution aiming at fulfilling the latency and data throughput strict requirements. In this sense, the combination of content cache and Multi-Access Edge Computing are promising alternatives. This work presents a cooperative network-aware cache police to decrease the latency experienced by end-user. The policy formulated through Integer Linear Programming joins the placement content and request routing problems and, moreover, considers challenges in limited storage, content popularity and user mobility. The simulations showed that cache policies are efficient to choose paths in Heterogeneous Cellular Network.*

Resumo. *As redes 5G desencadearam uma evolução para atender novas aplicações com rigorosos requisitos de latência e vazão de dados. Nesse contexto, a combinação de cache de conteúdo e Multi-Access Edge Computing é promissora. Esse trabalho apresenta uma política de cache cooperativa e sensível à rede, para reduzir a latência percebida pelos usuários finais. A política, desenvolvida com Integer Linear Programming, aborda conjuntamente a inserção de conteúdo e o roteamento de requisições, considerando desafios sobre armazenamento limitado, popularidade de conteúdo e mobilidade dos usuários. Simulações demonstraram que a política é eficiente na escolha de caminhos na Heterogeneous Cellular Network.*

1. Introdução

Os dispositivos móveis estão amplamente presentes no cotidiano de usuários e, têm sido um dos principais meios de acesso a Internet. Estima-se que em média, um *smartphone* trafegará até 34 GB de dados por mês no ano de 2026 [Ericsson 2020], o qual representa um crescimento de 24% a partir do ano de 2020. Ainda, o total de usuários até o ano de 2026 deverá ser aproximadamente $6,9$ bilhões de usuários, ou seja, 45% mais usuários em relação ao ano de 2020 [Ericsson 2020]. Nesse contexto, a evolução das redes móveis se torna imprescindível para sustentar o crescente número de usuários e quantidade de tráfego, bem como novas demandas provenientes da popularização ou surgimento de novos serviços e aplicações.

A rede *Fifth Generation Technology Standard (5G)* tem como objetivo sustentar tais demandas, assim como suportar aplicações que necessitam de baixíssima latência ou alta taxa de tráfego de dados tais como: carros autônomos, *Augmented Reality (AR)*, *Virtual Reality (VR)*, *Internet of Things (IoT)*, entre outros. Em síntese, se faz necessário

aumentar a largura de banda, reduzir a latência e otimizar o uso da capacidade energética dos dispositivos. Além disso, a vazão é essencial para transferir dados coletados a partir de dispositivos IoT ou gerados a partir de conteúdo multimídia [Pham et al. 2020].

Para isso, *Multi-Access Edge Computing (MEC)* permite aproximar recursos computacionais da borda da rede, dito de outro modo, os torna mais próximos geograficamente dos usuários, dispondo-os dentro da *Radio Access Network (RAN)* [Hu et al. 2015]. Por sua vez, *Heterogeneous Cellular Network (HCN)* permitem ampliar a capacidade de armazenamento do sistema de *cache* e, ampliar a capacidade da infraestrutura como um todo. HCN são compostas por *Base Station (BS)* de alta potência e BS de baixa potência denominadas, *Macro Base Station (MBS)* e *Small Base Station (SBS)*, respectivamente. Além disso, possuem BSs com diferentes capacidades, potência de sinal, cobertura e frequência, ao contrário das redes homogêneas. Conciliadas, tais tecnologias destacam-se como alternativas promissoras para atender aos requisitos de comunicação supracitados. Nesse sentido, atuam especificamente na redução de transmissão de dados replicados nos enlaces de *Backhaul (BH)*, baseadas no princípio de posicionamento de *caches* próximos aos usuários [Wu et al. 2021].

Os projetos para *cache* em HCN possuem dois problemas enfatizados na literatura. Primeiramente, o problema de inserção de conteúdo, o qual é a fase que determina qual, onde e como o conteúdo deve ser armazenado [Wu et al. 2021]. Por sua vez, o problema de entrega de conteúdo, o qual está relacionado à forma como o conteúdo será entregue ao usuário, o qual abarca os problemas de roteamento de requisições [Dehghan et al. 2017] e associação do usuário à BS [Harutyunyan et al. 2018].

Em geral os trabalhos presentes na literatura tem como principal interesse o problema de inserção de conteúdo [Shanmugam et al. 2013]. Dentre os trabalhos direcionados ao problema de roteamento de requisições, a cooperação entre os recursos da HCN, embora desejada, não é considerada [Dehghan et al. 2017, Harutyunyan et al. 2018]. A cooperação permite que a capacidade de armazenamento seja compartilhada entre as BSs, ou seja, se o conteúdo não é encontrado diretamente na BS que o usuário está conectado é possível realizar uma busca em outras BSs. Por outro lado, alguns trabalhos propõem abordagens cooperativas entre BSs vizinhas [Jiang et al. 2017], ou cooperação multissaltos [Song et al. 2021, Li et al. 2017]. Dentre tais abordagens, mesmo que a partir da cooperação multissaltos, há esquemas hierárquicos que podem limitar o espaço de busca. Por fim, para os trabalhos revisados, a mobilidade não é assumida como um fator fundamental, no desenvolvimento de políticas de *cache*, exceção para [Harutyunyan et al. 2018].

Este trabalho apresenta uma política de *cache* cooperativa e sensível à rede, unindo os problema de inserção de conteúdo e roteamento de requisições multissaltos com o objetivo de minimizar a latência em HCN. A política é formulada matematicamente através de *Integer Linear Programming (ILP)*, respeitando as restrições de capacidade de armazenamento e os requisitos de *Quality-of-Service (QoS)* definidos pelo *Service Level Agrement (SLA)*. Nesse sentido, a modelo matemático apresenta-se como um direcionador para projetar e desenvolver heurísticas ou algoritmos de menor custo computacional. Ademais, a política estima a carga dos enlaces e acompanha a dinâmica da rede com base em algoritmos de *Congestion Control (CC)* que fazem parte do *Transmission Control Protocol (TCP)*, ou seja, premissas consolidadas

em redes de computadores [Brakmo and Peterson 1995, Chiu and Jain 1989]. Em suma, a política busca caminhos com maior vazão e, consequentemente, menor *Round Trip Time (RTT)* [Chiu and Jain 1989]. De tal forma, evita agravar a piora da QoS devido a enlaces sobrecarregados. Simulações numéricas demonstraram que o recurso de sensibilidade à dinamicidade da rede, ou seja, a estimativa de carga, obteve êxito na escolha de caminhos entre *cache* e origem do conteúdo, sem impacto significativo na latência diante da variabilidade no comportamento dos usuários.

O restante deste trabalho é estruturado de tal forma: A Seção 2 destaca os trabalhos relacionados. A Seção 3 detalha as principais premissas da política de *cache* e, define a notação matemática, assim como abstração lógica através de grafos. A Seção 4 detalha as variáveis de decisão, função objetivo e as restrições da formulação matemática. Finalmente, a Seção 5 apresenta a simulação numérica e analisa seus resultados. Posteriormente, na Seção 6 as conclusões são sintetizadas.

2. Trabalhos Relacionados

Quanto à políticas de *cache* não cooperativas, o trabalho precursor ao abordar o problema é focado na inserção de *cache* de conteúdo na borda de redes móveis. Sobretudo, com restrições de capacidade de armazenamento, topologia da rede e distribuição de popularidade do conteúdo com objetivo de minimizar a latência [Shanmugam et al. 2013]. Da mesma forma, um trabalho posterior formulou a otimização do problema de inserção de conteúdo e roteamento de requisições de usuários de forma conjunta para otimizar a latência [Dehghan et al. 2017]. Por sua vez, foram propostas uma formulação ótima através de ILP e uma heurística , para o problema de inserção de conteúdo e associação de usuário conjuntamente. Objetivando realizar o balanceamento da utilização de recursos de rádio e utilização do enlace de BH [Harutyunyan et al. 2018]. Os autores consideraram a mobilidade para a associação de usuário. Em resumo, em políticas não cooperativas embora os usuários possam se conectar a múltiplos *caches*, caso o conteúdo solicitado não seja encontrado diretamente no *cache* associado ao usuário, o conteúdo é servido através do servidor remoto e, portanto, não há cooperação entre BS na busca pelo conteúdo.

Por outro lado, há propostas para uma política de *cache* cooperativa. Assim, um trabalho direcionado para uma arquitetura *Cloud Radio Access Network (C-RAN)* propôs uma formulação conjunta entre o problema de inserção de conteúdo, roteamento de requisições e alocação de recursos. O enfoque do trabalho foi a redução de custos, tais custos são estabelecidos por restrições de capacidade de armazenamento, capacidade de enlace, custo de reconfiguração de *Virtual Machines (VMs)*, custo de migração de consumo da *cache* e restrições de latência [Pu et al. 2018]. No mesmo sentido, uma formulação ótima do problema de inserção e entrega de conteúdo cooperativo para HCN foi proposta. Com objetivo de reduzir a latência, considerando restrições de topologia da rede, probabilidade de solicitação do conteúdo, capacidade da armazenamento e de enlace [Jiang et al. 2017].

Uma política de inserção e roteamento de requisições por meio de cooperação hierárquica foi proposta. Com objetivo de maximizar a taxa de *cache hit*, considerando restrições de capacidade da armazenamento, capacidade de enlace e topologia da rede [Li et al. 2017]. Por sua vez, foi projetada uma arquitetura multicamadas (hierárquica) de colaboração entre SBSs e dispositivos finais através de *Device-to-Device*

(D2D) com o objetivo de minimizar a latência [Sheng et al. 2016]. Do mesmo modo, um trabalho formulou o problema de inserção de conteúdo e roteamento de requisições de forma conjunta, como foco em redes não confiáveis. Essa estratégia teve como objetivo minimizar a latência e, além disso, a formulação ILP continha restrições de capacidade de armazenamento e de enlace [Song et al. 2021]. Além disso, os resultados são obtidos através de experimentação, ao contrário da maioria dos trabalhos selecionados. Em suma, estratégias de *cache* cooperativo (hierárquicas ou multissaltos) são abordagens promissoras se comparadas a abordagens não cooperativas. Tal fato é observado pela melhor utilização dos recursos de armazenamento e de comunicação, considerando que a busca entre os dados é realizada primeiramente entre os dispositivos vizinhos.

Para abordagens com princípio de cooperação a partir de uma BS vizinha, ou seja, aquela em que o usuário está associado ou, aqueles que consideram a cooperação hierárquica. Pode haver limitação no espaço de busca entre os demais *caches* presentes na RAN e, consequentemente, pode reduzir a proporção de *cache hit*, em contra ponto a política de *cache* proposta é cooperativa. Além disso, a política de *cache* proposta considera o roteamento de requisições multissaltos ao contrário dos trabalhos mencionados, os quais possuem um único caminho até o *cache* ou até a origem do conteúdo de uma perspectiva End-to-End (E2E). Sobretudo, a política de *cache* proposta considera as possíveis variações da rede nos caminhos intermediários considerando à dinamicidade da rede [Dehghan et al. 2017, Pu et al. 2018, Jiang et al. 2017, Li et al. 2017]. Os trabalhos que consideram roteamento multissaltos, ignoram o impacto da mobilidade sobre a QoS [Song et al. 2021, Sheng et al. 2016]. Diferentemente de tais trabalhos, a presente proposta de política de *cache* destaca-se por obter a estimativa da capacidade real do enlace e não a capacidade máxima de largura de banda do enlace [Harutyunyan et al. 2018, Shanmugam et al. 2013, Dehghan et al. 2017, Li et al. 2017, Jiang et al. 2017, Pu et al. 2018, Song et al. 2021]. A capacidade máxima de um enlace é uma informação privilegiada, disponível apenas mediante o controle total da rede, a obtenção da estimativa de capacidade real do enlace fornece uma aplicação factível em cenários competitivos, compostos por múltiplos serviços e aplicações.

3. Política de *Cache*

Esta seção detalha as principais premissas da política de *cache* cooperativa orientada à rede e a definição formal matemática, bem como os detalhes da infraestrutura de rede.

3.1. Orientação à Rede e Cooperação

A política de *cache* se beneficia de mecanismos utilizados em algoritmos de controle de congestionamento (ou CC) do TCP. Tais mecanismos, permitem estimar as condições da rede e prever a capacidade do enlace [Brakmo and Peterson 1995, Cardwell et al. 2017]. Sobretudo, a política de *cache* atua na camada de aplicação, embora tais mecanismo sejam originalmente utilizados na camada de transporte. Especificamente o TCP *Vegas*, um dos percursores da proposta de predição, identifica a janela de congestionamento, a partir da comparação entre a vazão atual da rede e a vazão esperada.

O mecanismo de CC pode ser usado para estimar o congestionamento da rede móvel, permitindo que a política de *cache* se comporte dinamicamente de acordo com o estado da rede. A política de *cache* considera como referencial a vazão esperada de-

finida pela aplicação. Em outras palavras, um determinado serviço (e.g. *Video on Demand (VoD)*) tem a necessidade de que uma vazão mínima seja garantida para que seu serviço seja entregue dentro da qualidade esperada. Essa restrição de QoS definida pelo SLA deve ser disponibilizada pelo provedor de conteúdo. Assim, a vazão atual é dada a partir da divisão entre o *buffer*, que consiste em uma fração do conteúdo que será enviado ao usuário de acordo com a sua demanda e, o RTT atual do enlace. Ressalta-se que o conceito de *buffer*, nesse contexto, consiste na menor parte dos dados que podem ser enviados pela aplicação e, além disso, sua características dependem inteiramente da aplicação. Em suma, a política de *cache* parte da seguinte premissa: um forte indicativo de congestionamento é dado pela vazão do enlace, ou seja, quanto menor a vazão, mais forte será o indício de que existe um possível congestionamento no enlace. Nesse sentido, o aumento do RTT pode ser interpretado como um sintoma de congestionamento presente no enlace.

Certamente, mobilidade impõe desafios na garantia de QoS no problema de inserção e roteamento de requisições. É possível que ocorra interferência na qualidade do sinal oriunda de ruídos, decorrentes da mobilidade do usuário, da troca entre SBSs e do distanciamento da SBS. Tais fatos podem acarretar em perdas de pacotes, um fator que impacta diretamente no aumento do RTT [Tian et al. 2005]. Portanto, é possível deduzir que o RTT está relacionado com a distância entre o usuário e a SBS, ou seja, o RTT é linearmente proporcional a distância entre o usuário e a SBS. Além disso, se houver uma tendência de aumento da tráfego no enlace, o RTT cresce exponencialmente proporcional a carga no enlace. Assim, o RTT obtido nos caminhos intermediários, ou seja, por meio de conexão com fio, varia de acordo com a carga do enlace [Chiu and Jain 1989].

Em síntese, enfatiza-se a seguinte premissa: quanto maior a vazão e menor o RTT, melhor será o desempenho do serviço e, mais adequado é o enlace de acordo com os requisitos de QoS definidos pelo provedor de serviço [Chiu and Jain 1989], relacionado, neste caso, diretamente à latência. É possível deduzir que a escolha de caminhos com maior vazão, além de evitar caminhos possivelmente sobrecarregados, permite evitar o surgimento ou agravamento de sobrecarga ou congestionamento. Portanto, converge em sentido ao principal objetivo da política de *cache*, isto é, minimizar a latência. Destaca-se que o roteamento considera o estado do enlace salto a salto, ou seja, consiste em roteamento multissalto, ciente das implicações das condições de enlaces intermediários.

Diferente de outros trabalhos presentes na literatura, a presente proposta destaca-se por obter a estimativa da capacidade real do enlace e não a capacidade máxima de largura de banda do enlace [Harutyunyan et al. 2018, Shanmugam et al. 2013, Dehghan et al. 2017, Li et al. 2017, Jiang et al. 2017, Pu et al. 2018, Song et al. 2021]. Considerando que a capacidade máxima de um enlace é disponível apenas mediante o controle total da rede, a obtenção da estimativa de capacidade real do enlace fornece uma aplicação factível em cenários competitivos, compostos por múltiplos serviços e aplicações.

A política de *cache* pretende otimizar as métricas na perspectiva do *Mobile Network Operator (MNO)*. Assim, a política busca o enlace com maior vazão, esperando o desencadeamento de um efeito de espalhamento e distribuição dos fluxos de dados, ao passo que, por outro lado busca concentrar o *cache* de conteúdo usando o menor número possível de réplicas para otimizar a capacidade de armazenamento. Ademais, a política de

cache deve preferir caminhos dentro da RAN e, sempre que possível recuperar requisições alocadas em uma nuvem computacional para o *cache* mais próximo do usuário. Assim, a política evita que a requisição trafegue pelo BH reduzindo custos do MNO e proporcionando menor latência. Em síntese, uma requisição (determinada como uma solicitação realizada por um usuário a um conteúdo específico) pode ser alocada ou realocada de acordo com as condições das rede. Consequentemente, os caminhos das requisições podem ser redirecionados para atender os requisitos de QoS, ou ainda, a origem de consumo do conteúdo (*cache* ou nuvem) pode ser reconfigurada. Nesse sentido, é possível que o consumo de um conteúdo diretamente de sua origem possa ser momentaneamente proveitoso, a partir de uma perspectiva de minimização da latência.

Finalmente, a política de *cache* funciona a partir de uma perspectiva horizontal e multissaltos. Portanto, quando o usuário envia uma solicitação de conteúdo, a busca é realizada em toda a RAN de forma uniforme sem respeitar hierarquias, ampliando a capacidade de cooperação, quando comparada com abordagens em que o conteúdo é buscado respeitando uma ordem previamente definida de busca (e.g. busca-se primeiro em SBS do mesmo nível, depois busca-se em MBS) [Sheng et al. 2016, Li et al. 2017].

3.2. Infraestrutura HCN

A arquitetura alvo agrega tecnologias de uma rede celular heterogênea ultra-densa, com *caches* implantados em MBS e SBS, de tal modo que o conteúdo (e.g. VoD) pode ser consumido em *cache* ou diretamente em sua origem, ou seja, transferido através de um enlace de BH que possibilita a conexão até o servidor remoto localizado fora do domínio de administração do *MNO*, que poderá estar localizado em nuvens computacionais. Ainda, é necessário que a arquitetura forneça detalhes gerenciais obtidos com *Software Defined Networking (SDN)* ou *Information-Centric Networking (ICN)*. Por fim, os usuários podem se conectar a múltiplas SBSs devido a densidade da rede, sendo possível que o usuário se desloque ao longo da RAN alternando sua conexão entre as SBSs.

Dado um grafo $G(V, E)$, o conjunto V representa os vértices, composto por um ponto de acesso à nuvem (S), servidores de conteúdo (C) e a HCN constituída de MBS, SBS e equipamentos de usuários (*User Equipment (UE)*). Ressalva-se que na abstração com grafos as SBS e MBS consistem em elementos com as mesmas propriedades, assim, de forma simplificada, podem ser considerados BS na notação matemática, isto é, elementos homólogos. Um conteúdo $k \in C$ possui requisitos definidos por SLA, dados por $c_k^s \in \mathcal{N}+$ e $c_k^{thp} \in \mathcal{N}+$, representando os requisitos de QoS para armazenamento e a vazão mínima, respectivamente. Ambas informações são fornecidas pelo provedor de serviço e conhecidas durante a execução da política de *cache*. Além disso, $bs_i^s \in \mathcal{N}+$ denota a capacidade máxima de armazenamento de $i \in BS$ (0 indica que a BS não tem capacidade de armazenamento e, portanto, funciona apenas como um relé).

A conexão entre uma nuvem computacional $s \in S$ e $i \in BS$ é dado por $e_{ij} \subseteq E$, dessa forma a nuvem computacional se conecta com as MBS através de uma aresta unidirecional que representa o BH. As demais conexões cabeadas entre as BSs estão contidas no conjunto de arestas $e_{ij} \subseteq E$. É importante observar que as MBS podem ou não possuir arestas bidirecionais entre si. Do mesmo modo, as SBS se conectam as MBS através de arestas bidirecionais, entretanto, não conectam-se a outras SBS. O relacionamento entre os vértices e direção das arestas é resumido na Figura 1. Igualmente,

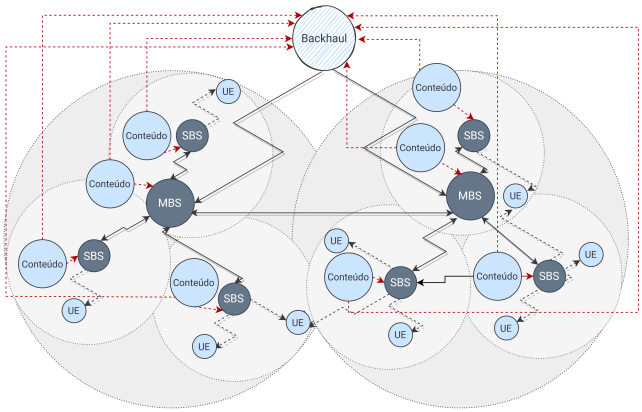


Figura 1. Representação do problema usando grafos.

as conexões sem fio entre SBS e UE pertencem a $e_{iu} \subseteq E$, entretanto são dadas pela distância entre a UE e as SBS. Nesse caso, cada $u \in UE$ possui coordenadas (x, y) que podem variar a medida que a UE se desloca sobre o plano.

A função $dis(.) \in \mathcal{R}+$ retorna a distância euclidiana em plano cartesiano. Para cada $i \in BS$, existe $D \in \mathcal{R}+$ que denota o raio de cobertura de uma BS. Portanto, quando $dis(x_i, y_i, x_u, y_u) \leq D$ a $u \in UE$ está dentro do raio de cobertura de $i \in BS$. A conexão entre $k \in C$ e $i \in BS$ segue o mesmo raciocínio, ou seja, as arestas pertencem a $e_{ij} \subseteq E$. Ainda, dado $\gamma_{ik} \in \{0, 1\}$, o valor 1 representa que o conteúdo k pode ser armazenado em i, enquanto 0 denota o contrário. A Figura 1 ilustra a conexão entre o conteúdo e a BS, destaca-se que todos os conteúdos possuem conexão com o BH. Por fim, o parâmetro $r_{uk} \in \{0, 1\}$, representa se $u \in UE$ solicitou $k \in C$.

Cada aresta $ij \in E$ possui um RTT associado, dado por rtt_{ij}, sendo que o RTT é o último observado no enlace. Assim, é possível obter a vazão atual da rede, conforme detalhado na Seção 3.1. Formalmente, a vazão atual é denotada por $thp^{cur}_{ijk} \in \mathcal{N}+$, sendo $thp^{cur}_{ijk} = \frac{c^b_k}{rtt_{ij}} \in \mathcal{N}+$. Além disso, $rtt_{ij} \to 0$ e $thp^{cur}_{ijk} \to \infty$, em arestas entre $k \in C$ e $i \in BS$.

4. Definição de um Modelo de Programação Linear Inteira

A programação matemática é amplamente adotada em trabalhos semelhantes presentes na literatura [Dehghan et al. 2017, Jiang et al. 2017, Pu et al. 2018, Li et al. 2017], bem como em soluções comerciais [at Meta 2021]. Além disso, a programação matemática pode ser usada para obter a otimização de problemas que não possuem algoritmos viáveis que possam alcançar a resposta em tempo polinomial. Em geral, a programação matemática possui vantagens em obter respostas matemáticas sem desprender recursos adicionais, assim como a possibilidade de variar as entradas, parâmetros e validar múltiplos cenários. Assim, pode ser usada como uma ferramenta estratégica para validação uma hipótese sem a necessidade de configuração de infraestruturas experimentais [at Meta 2021].

Assim, a presente seção descreve a formulação da política de *cache* usando ILP. A formulação busca a otimização conjunta para os problemas de inserção de conteúdo e roteamento de requisições de acordo com restrições de capacidade de armazenamento e

de restrições de QoS com o objetivo de minimizar a latência.

4.1. Variáveis de Decisão

A formulação ILP da política de *cache* é baseada no *Multi-Commodity Flow Problem (MCFP)*, um problema NP-Completo [Ahuja et al. 1993]. Nesse sentido, uma variável binária x_{ijk} indica se existe fluxo do conteúdo $k \in C$ sobre a aresta $e_{ij} \in E$. De tal modo que, se $x_{ijk} = 1$ indica que há fluxo na aresta, caso contrário $x_{ijk} = 0$. Por sua vez, uma variável binária y_{ik} indica se um conteúdo $k \in C$ está armazenado em uma BS $i \in BS$, se $y_{ik} = 1$, ou ausência de armazenamento se $y_{ik} = 0$. Ainda, $y_{ik} = x_{kik}$, ou seja y_{ik} é obtido em função de x_{kik}. Além disso, as Equações 2-4 garantem que exista um único caminho entre um usuário $u \in UE$ e um conteúdo $k \in C$, ou seja, o conteúdo será transmitido apenas por uma única origem (*cache* ou nuvem computacional) a cada requisição.

4.2. Função Objetivo

A função objetivo é descrita pela Equação 1. Para que seja possível atender aos requisitos de latência da rede 5G, o primeiro termo da função objetivo corresponde à perspectiva de inserção de conteúdo. Sendo assim, preferencialmente as requisições devem ser atendidas dentro da RAN sem que seja necessário trafegar pelo enlace de BH até a nuvem computacional. Além disso, o parâmetro $\delta \rightarrow 0$ é inserido para que não ocorra divisão por 0 quando um conteúdo não pode ser hospedado por uma BS.

Por outro lado, a função objetivo busca a otimização do problema através da perspectiva da rede, a medida que preserva a vazão do enlace e respeita sua dinamicidade. Nesse sentido, considera possíveis congestionamentos presentes na rede decorrentes do aumentos do RTT (de acordo com a Seção 3.1). A partir dessa premissa, é possível inferir que a distribuição dos fluxos direcionada para enlaces que apresentem uma vazão superior seja a melhor escolha. Assim, enlaces com vazões menores podem indicar possíveis congestionamentos e são evitados, ou seja, a equação busca a maior vazão em função de c_k^{thp}. Esse comportamento é dado pelo segundo termo da função objetivo, dado pela Equação 1, que em suma representa p problema de roteamento de requisições.

$$min \sum_{u \in UE} \sum_{i \in BS} \sum_{k \in C} \frac{(bs_i^s - c_k^s) \times r_{uk} \times y_{ik}}{bs_i^s \times \gamma_{ik} + \delta} + \sum_{u \in UE} \sum_{ij \in E} \sum_{k \in C} \frac{c_k^{thp}}{thp_{ijk}^{cur}} \times x_{ijk} \times r_{uk} \quad (1)$$

4.3. Restrições

Assim como a otimização conjunta dos problemas de inserção de conteúdo e roteamento de requisições é baseado no MCFP, as restrições dadas pelas Equações 2-4 garantem a conservação do fluxo, isto é, garantem que exista apenas um único caminho entre um usuário $u \in UE$ e um conteúdo $k \in C$.

$$\sum_{i \in BS} x_{jik} \times r_{uk} - \sum_{i \in BS} x_{ijk} \times r_{uk} = 0; \forall j \in BS, \forall k \in C, \forall u \in UE \quad (2)$$

$$\sum_{i \in BS} x_{kik} \times r_{uk} - \sum_{i \in BS} x_{ikk} \times r_{uk} = 1; \forall k \in C, \forall u \in UE \quad (3)$$

$$\sum_{i \in BS} x_{uik} \times r_{uk} - \sum_{i \in BS} x_{iuk} \times r_{uk} = -1; \forall k \in C, \forall u \in UE \quad (4)$$

A restrição dada pela Inequação 5, garante que a capacidade de armazenamento de cada BS seja respeitada, enquanto a Equação 6 formula a restrição de QoS definida pelo SLA, em outras palavras, garante que os caminhos definidos repeitem a vazão mínima necessária para servir o conteúdo para um aplicação. Ainda, ressalta-se que a QoS não pode ser garantida pela política efetivamente, dado que depende das condições da rede. Dado que a capacidade de armazenamento da BS é limitada, mas no entanto, permite posicionar os usuários no mesma BS, o segundo termo da equação tende a evitar que os usuários posicionem-se no mesmo ponto. No entanto, a vazão tende a aumentar no enlace devido ao aumento da sobrecarga derivada da concentração de usuários e, consequentemente, pode extrapolar os requisitos de QoS. Nesse caso, a requisição é atendida de forma convencional através do enlace de BH se a restrição não for atendida.

$$\sum_{k \in C} c_k^s \times y_{ik} \leq bs_i^s; \forall i \in BS \tag{5}$$

$$(x_{ijk} \times r_{uk}) \times c_k^{thp} \leq thp_{ijk}^{cur} \times (x_{ijk} \times r_{uk}); \forall ij \in E, \forall k \in C, \forall u \in UE \tag{6}$$

5. Simulações Numéricas

Esta seção apresenta resultados de simulações numéricas obtidas a partir da implementação da política de *cache*. A simulação numérica foi realizada a partir de um administrador de eventos discretos[1], implementado em linguagem de programação Python 3.9 juntamente com o solucionador Gurobi 9.1[2]. A simulação foi executada em dois servidores: Intel Xeon E312XX com 64 GB RAM, e Intel I7-7700 com 28 GB RAM.

5.1. Parâmetros

A execução foi realizada com 100 eventos discretos [Sheng et al. 2016], com a taxa de chegada de requisições (λ) seguindo a distribuição de Poisson [Dehghan et al. 2017]. Foram consideradas 2 MBSs, na qual cada uma possui 15 SBSs conectadas [Khreishah et al. 2016, Jiang et al. 2017]. O raio de cobertura para SBS é dado em 70 metros [Shanmugam et al. 2013, Sheng et al. 2016, Jiang et al. 2017]. A capacidade máxima de armazenamento das BS, informada pelo MNO, é 40% da capacidade total para armazenar toda à biblioteca de conteúdo. SBSs e MBSs possuem 4 GB e 20 GB, respectivamente. Há 200 usuários conectados à rede móvel e cada usuário pode conectar-se a no máximo 2 SBSs. Os UEs podem se mover de forma randômica [Shanmugam et al. 2013, Harutyunyan et al. 2018, Sheng et al. 2016], em geral com deslocamento de 10 metros de distância, que é a distância euclidiana num plano cartesiano entre UE e SBS e, além disso, as UEs se movem uma vez a cada evento.

Cada enlace possui um valor inicial para o RTT de 1 ms, tanto para meios cabeados ou sem fio [ITU 2017]. As requisições permanecem alocadas durante 10 eventos discretos (τ). Características da aplicação são informadas pelo servidor de conteúdo, são elas: *buffer* e vazão mínima tolerada. Tais valores são 48 Mb e 100 Mbps, respectivamente [ITU 2017]. A biblioteca de conteúdos possui um total de 100 conteúdos distintos,

[1]https://github.com/marischatten/modeling

[2]https://www.gurobi.com/

como tamanhos totais entre 2 GB, 4 GB e 8 GB. Destaca-se que a otimização do posicionamento do conteúdo, assim como o definição do roteamento de requisições acontece em eventos de tempo configuráveis (que podem ser definidos pelo MNO), e não a cada chegada de requisição. Portanto, o modelo matemático é executado a cada intervalo de tempo.

5.2. Métricas e Cenário para Análise

Para análise da proposta, dados sobre quatro métricas, latência, *cache hit* e *cache miss*, uso total do armazenamento e carga da rede (Fronthaul (FH) e BH) foram coletados. A latência de uma requisição é o intervalo de tempo *E2E* que pode ser medido com base no somatório dos RTTs dos enlaces pertencentes ao caminho realizado pela requisição. A proporção de *cache hit* é obtida através da razão entre *cache hit* e total de requisições alocadas por evento discreto. Do mesmo modo, a proporção de *cache miss* é a dada a partir da razão do *cache miss* e total de requisições. O uso total de armazenamento se obtém a partir a divisão do somatório da capacidade utilizada na BS pela capacidade total. A carga total do BH e FH considera o somatório da vazão atual de cada requisição, dada pelo somatório da vazão atual de cada salto da requisição.

Tais métricas demonstram características do comportamento da política de *cache*. A latência objetiva evidenciar que a busca por maiores vazões tende a evitar o surgimento de congestionamento na rede. A proporção de *cache hit* e *cache miss* demonstra a cooperação entre BSs, bem como a priorização de elementos mais à borda da rede móvel. A análise combinada do uso total do armazenamento com *cache hit/miss* demonstra que o *cache miss* não é desencadeado unicamente pela capacidade de armazenamento, mas também pela influência da rede. Por fim, a carga da rede demonstra as vazões utilizadas em função das requisições presentes na rede, assim como o impacto ocasionado pela variação do comportamento dos usuários.

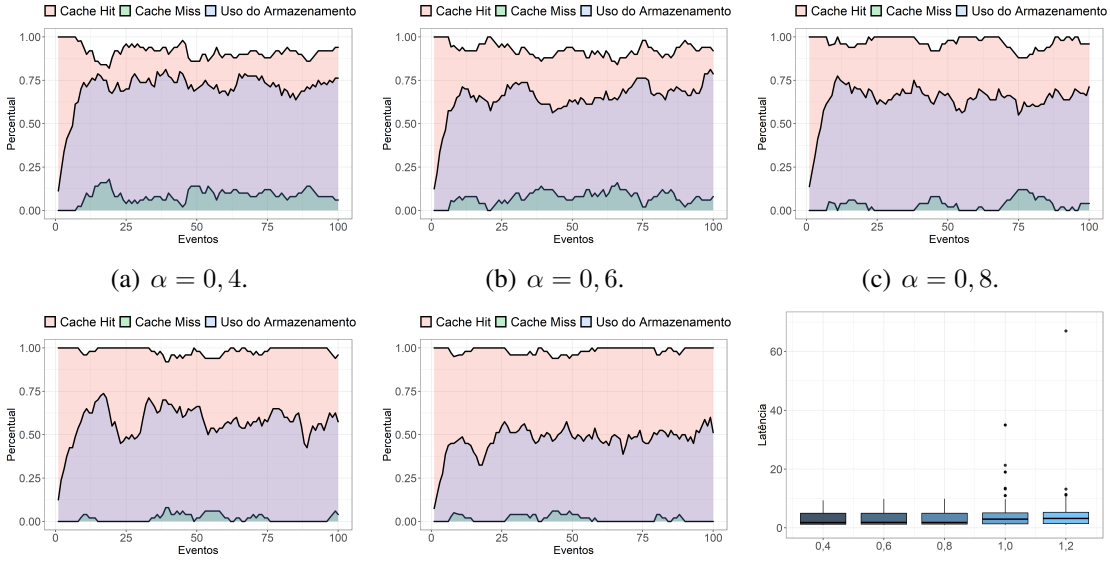
Quanto ao cenário, a simulação considera a variação da distribuição de popularidade dos conteúdos [Song et al. 2021] dada a partir da distribuição Zipf [Breslau et al. 1999]. A variação de distribuição de popularidade acarreta em menor variabilidade de popularidade de conteúdos quando o valor de α é maior, ou seja, as solicitações de conteúdo concentram-se em torno de uma pequena parte do conteúdo. Por sua vez, um valor de α menor distribui a popularidade, desse modo, o padrão de preferência dos usuários é esparso. O parâmetro de distribuição de popularidade abrangeu uma variação no valor α para distribuição Zipf entre $0,4$ e $1,2$, em intervalos de $0,2$.

5.3. Análise dos Resultados

Conforme apresentado na Figura 2, a proporção de *cache hit* reduziu em relação aos valores de α menores. Assim, mais solicitações foram atendidas em *cache*, devido a concentração de popularidade dos conteúdos solicitados. Tal comportamento se intensificou à medida que o o valor de α aumentou. Ainda, quanto menor o valor de α, menor a proporção de *cache hit* e, por outro lado, quanto maior o α menor é a proporção de *cache miss*. Esse comportamento é resultante da possibilidade de alocar uma quantidade menor de conteúdos em *cache* devido a capacidade limitada de armazenamento. Quanto menor o valor de α, mais variados são os conteúdo solicitados e, portanto, é necessário que a capacidade de armazenamento seja superior, para garantir que mais requisições sejam atendidas em *cache*. Para o valor de $\alpha = 0,4$ (Figura 2(a)), o uso de armazenamento

máximo foi $0,81$ e a proporção máxima de *cache miss* alcançou $0,18$, enquanto o *cache hit* obteve ser valor mínimo de $0,82$. Em contraste, para o valor de $\alpha = 1,2$ (Figura 2(e)), o máximo do uso total da capacidade de armazenamento foi $0,60$, enquanto a proporção mínima de *cache hit* foi $0,94$, e $0,06$ para a proporção máxima de *cache miss*.

Tal variação desencadeia efeitos sobre o roteamento de requisições e, portanto, na decisão entre a escolha da origem do conteúdo com objetivo de buscar caminhos menos sobrecarregados, isto é, através da *cache* ou BH. Destaca-se que mesmo com a menor utilização de armazenamento e maior concentração da popularidade dos conteúdos, ainda assim, é possível observar que houve *cache miss*, ou seja, esse comportamento deriva-se do estado da rede, o qual impactou diretamente na decisão da política quanto ao roteamento de requisições e fez com que o conteúdo fosse recuperado a partir do BH.



(a) $\alpha = 0,4$. (b) $\alpha = 0,6$. (c) $\alpha = 0,8$.

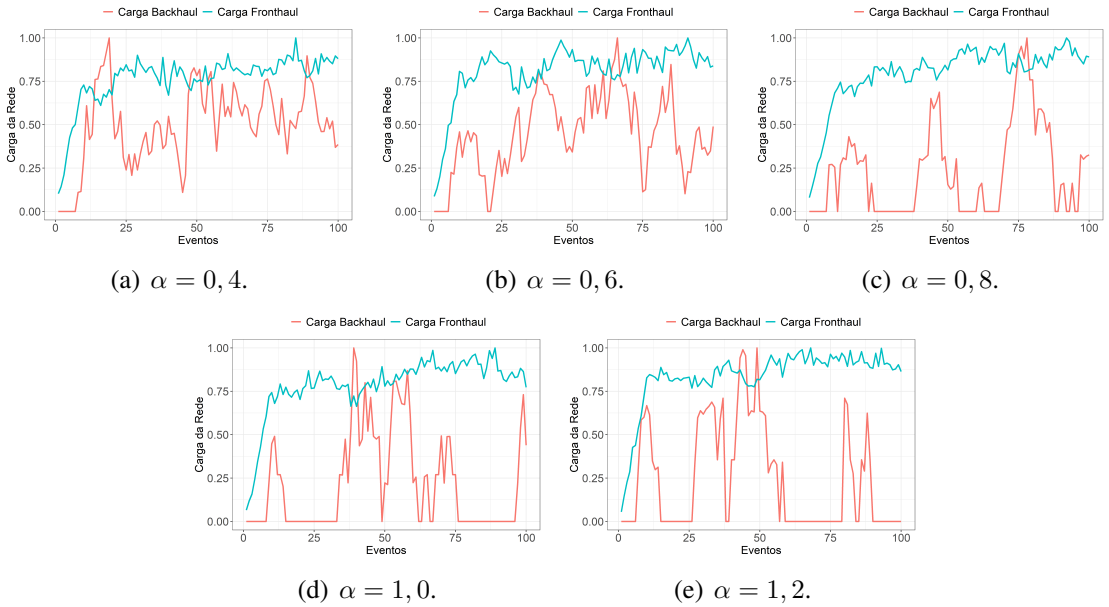
(d) $\alpha = 1,0$. (e) $\alpha = 1,2$. (f) Latência da Rede.

Figura 2. Cache *hit*/*miss*, Armazenamento Total e Latência em Função do Cenário de Distribuição de Popularidade.

Resultados sobre latência da rede são apresentados na Figura 2(f). Observa-se que a variação nos valores de α não desencadeou impactos significativos, com valor para o coeficiente de correlação de Pearson de $0,12$, indicativo de correlação inexistente. Ainda, o gráfico demonstra a conformidade entre as diferentes variações de distribuição de popularidade dos conteúdos. Dado o valor divergente no último quartil, possivelmente tal variação é derivada da concentração de popularidade de conteúdo quando $\alpha = 1,2$. Nesse sentido, pode desencadear a sobrecarga de um enlace específico causada pelo crescimento exponencial do RTT em relação aos dados trafegados, como descrito na Seção 3.1. Ainda, houve requisições não atendidas em *cache*, em outra palavras, é possível notar que o caminho até a *cache* de conteúdo possivelmente está sobrecarregado, e consequentemente a política opta por escolher a transmissão através do BH.

A política de *cache* comporta-se de tal modo que busca otimizar o latência independentemente das variações de popularidade de conteúdo. Para 50% da amostra em ambos os casos, a latência se manteve inferior a 4 ms, portanto, dentro do limite determinado pelo [ITU 2017]. Embora no último quartil a latência se manteve aproximadamente

inferior a $9,3$ ms no melhor caso e 67 ms no pior caso, em 95% da amostra se manteve inferior a $5,9$ ms no melhor caso e $7,2$ ms no pior caso.



(a) $\alpha = 0,4$. (b) $\alpha = 0,6$. (c) $\alpha = 0,8$.

(d) $\alpha = 1,0$. (e) $\alpha = 1,2$.

Figura 3. Carga da Rede em Função do Cenário de Distribuição de Popularidade.

A Figura 3 aprofunda a análise sobre a carga da rede, demonstrando o impacto ocasionado pela variação do comportamento dos usuários. Inicialmente, observa-se comportamento crescente na carga da rede, para ambos (BH e FH), seguida de uma condição de estabilidade. Tal comportamento é motivado pelo estado inicial da rede, ou seja, a rede não possui requisições alocadas e, ao passo que, novas solicitações são acomodas e, posteriormente desalocadas, a rede atinge um nível de estabilidade, constatado visivelmente para requisições alocadas em *cache*. Ainda, é possível notar que há mais requisições alocadas em *cache*, consequentemente, tais requisições ocupam e colocam uma maior carga na rede, ao contrário das requisições consumidas a partir de sua origem. Em geral, a política de *cache* prioriza a colocação de requisições em *cache*, tal comportamento é orientado pelo primeiro termo da Equação 1. Assim, o provedor de conteúdo é beneficiado e, portanto, pode oferecer uma melhor QoS. Da mesma forma, o MNO é beneficiado, atendendo aos requisitos do SLA e, ainda, não comprometendo sua rede com surgimento ou agravamento de possíveis congestionamentos. Essa decisão é orientada pelo segundo termo da Equação 1, ou seja, originado da característica de sensibilidade à rede presente no roteamento de requisições. Por fim, reduz o uso do enlace de BH que pode acarretar em custos para o MNO.

Em síntese, a política demonstrou tendência para alocar requisições em *cache* e, ao mesmo tempo ponderou as condições da rede, para garantir a QoS. Isso reflete o segundo termo da Equação 1 e demonstra a sensibilidade à dinâmica da rede. Sobretudo, tal comportamento demonstra que eventualmente a política de *cache* buscou o enlace de BH para garantir a vazão em momentos de sobrecarga, com o objetivo de manter os serviços alocados e a latência reduzida, mas sob esta condição priorizou a alocação em *cache*, manifestando o primeiro termo da Equação 1.

6. Conclusão

Esse trabalho desenvolveu e descreveu uma política de *cache* cooperativa orientada à rede com objetivo de reduzir a latência em HCN. Além disso, os problemas de inserção de conteúdo e roteamento de requisições compõem a política de *cache*. Sendo assim, foi formulado um modelo ótimo através de ILP com restrições de capacidade de armazenamento e requisitos de QoS definidos pelo provedor de serviço.

Simulações numéricas demonstraram que variações na distribuição de popularidade dos conteúdos e comportamento dos usuários não demonstraram impactos significativos na latência. Destaca-se que, embora, houvesse capacidade de armazenamento suficiente, eventualmente a política de *cache* escolhe recuperar o conteúdo através do enlace de BH. Assim, a política obteve êxito na escolha entre os caminhos, de modo que evitou possíveis caminhos sobrecarregados na RAN, independentemente do comportamento do usuário. Em síntese, a variação no comportamento dos usuários tem impacto na otimização, entretanto, não demonstra impacto sobre à latência.

Futuramente será necessário desenvolver novos cenários para as simulações numéricas. Ainda, é necessário avaliar a dinâmica das migrações de caminhos, migrações e replicações de servidores *cache* e o impacto da desalocações das requisições. Além disso, é necessário incorporar o uso de dados reais de mobilidade urbana.

Agradecimentos: Os autores agradecem o apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Fundação de Amparo à Pesquisa e Inovação (FAPESC), LabP2D e UDESC.

Referências

Ahuja, R. K., Magnanti, T. L., and Orlin, J. B. (1993). *Network Flows*. Prentice Hall, United States of America.

at Meta, E. (2021). Network hose: Managing uncertain network demand with model simplicity. Facebook Engineering.

Brakmo, L. S. and Peterson, L. L. (1995). Tcp vegas: End to end congestion avoidance on a global internet. *IEEE Journal on Selected Areas in Communications*, 13(8):1465–1480.

Breslau, L., Cao, P., Fan, L., Phillips, G., and Shenker, S. (1999). Web caching and zipf-like distributions: evidence and implications. In *IEEE INFOCOM '99. Conference on Computer Communications. Proceedings. Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies. The Future is Now*, pages 126–134, New York, NY, USA.

Cardwell, N., Cheng, Y., Gunn, C. S., Yeganeh, S. H., and Jacobson, V. (2017). Bbr: Congestion-based congestion control. *Communications of the ACM*, 60:58–66.

Chiu, D.-M. and Jain, R. (1989). Analysis of the increase and decrease algorithms for congestion avoidance in computer networks. *Computer Networks and ISDN Systems*, 17(1):1–14.

Dehghan, M., Jiang, B., Seetharam, A., He, T., Salonidis, T., Kurose, J., Towsley, D., and Sitaraman, R. (2017). On the complexity of optimal request routing and content

caching in heterogeneous cache networks. *IEEE/ACM Transactions on Networking*, 25(3):1635–1648.

Ericsson (2020). Ericsson mobility report. Technical report.

Harutyunyan, D., Bradai, A., and Riggio, R. (2018). Trade-offs in cache-enabled mobile networks. In *2018 14th International Conference on Network and Service Management (CNSM)*, pages 116–124, Rome, Italy.

Hu, Y. C., Patel, M., Sabella, D., Sprecher, N., and Young, V. (2015). Mobile edge computing a key technology towards 5g. Technical report, ETSI, Sophia Antipolis, CEDEX, France.

ITU, I. T. U. (2017). Minimum requirements related to technical performance for imt-2020 radio interface(s). Technical report.

Jiang, W., Feng, G., and Qin, S. (2017). Optimal cooperative content caching and delivery policy for heterogeneous cellular networks. *IEEE Transactions on Mobile Computing*, 16(5):1382–1393.

Khreishah, A., Chakareski, J., and Gharaibeh, A. (2016). Joint caching, routing, and channel assignment for collaborative small-cell cellular networks. *IEEE Journal on Selected Areas in Communications*, 34(8):2275–2284.

Li, X., Wang, X., Li, K., Han, Z., and Leung, V. C. M. (2017). Collaborative multi-tier caching in heterogeneous networks: Modeling, analysis, and design. *IEEE Transactions on Wireless Communications*, 16(10):6926–6939.

Pham, Q.-V., Fang, F., Ha, V. N., Piran, M. J., Le, M., Le, L. B., Hwang, W.-J., and Ding, Z. (2020). A survey of multi-access edge computing in 5g and beyond: Fundamentals, technology integration, and state-of-the-art. *IEEE Access*, 8:116974–117017.

Pu, L., Jiao, L., Chen, X., Wang, L., Xie, Q., and Xu, J. (2018). Online resource allocation, content placement and request routing for cost-efficient edge caching in cloud radio access networks. *IEEE Journal on Selected Areas in Communications*, 36(8):1751–1767.

Shanmugam, K., Golrezaei, N., Dimakis, A. G., Molisch, A. F., and Caire, G. (2013). Femtocaching: Wireless content delivery through distributed caching helpers. *IEEE Transactions on Information Theory*, 59(12):8402–8413.

Sheng, M., Xu, C., Liu, J., Song, J., Ma, X., and Li, J. (2016). Enhancement for content delivery with proximity communications in caching enabled wireless networks: architecture and challenges. *IEEE Communications Magazine*, 54(8):70–76.

Song, Y., Wo, T., Yang, R., Shen, Q., and Xu, J. (2021). Joint optimization of cache placement and request routing in unreliable networks. *Journal of Parallel and Distributed Computing*, 157:168–178.

Tian, Y., Xu, K., and Ansari, N. (2005). Tcp in wireless environments: Problems and solutions. *IEEE Communications Magazine*, 43(3):S27–S32.

Wu, H., Fan, Y., Wang, Y., Ma, H., and Xing, L. (2021). A comprehensive review on edge caching from the perspective of total process: Placement, policy and delivery. *Sensors*, 21(15).