UNIVERSIDADE DO ESTADO DE SANTA CATARINA – UDESC
CENTRO DE CIÊNCIA TECNOLÓGICAS – CCT
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO APLICADA – PPGCAP

MARISANGILA ALVES

POSICIONAMENTO DE CACHES EM REDES CELULARES HETEROGÊNEAS CONSIDERANDO O ROTEAMENTO DE REQUISIÇÕES

JOINVILLE
2022

MARISANGILA ALVES

POSICIONAMENTO DE CACHES EM REDES CELULARES HETEROGÊNEAS CONSIDERANDO O ROTEAMENTO DE REQUISIÇÕES

Dissertação apresentada ao Programa de Pós–Graduação em Computação Aplicada do Centro de Ciências Tecnológicas da Universidade do Estado de Santa Catarina, como requisito parcial para a obtenção do grau de Mestre em Computação Aplicada.

Orientador: Dr. Guilherme Piêgas Koslovski

JOINVILLE
2022

Alves, Marisangila
Posicionamento de Caches em Redes Celulares Heterogêneas Considerando o Roteamento de Requisições / Marisangila Alves. - Joinville, 2022.
85 p. : il. ; 30 cm.

Orientador: Dr. Guilherme Piêgas Koslovski.
.
Dissertação (Mestrado) - Universidade do Estado de Santa Catarina, Centro de Ciências Tecnológicas, Programa de Pós-Graduação em Computação Aplicada, Joinville, 2022.

1. Cache. 2. Redes Celulares Heterogêneas. 3. Programação Linear. 4. Posicionamento de Cache. 5. Roteamento de Requisições. I. Koslovski, Dr. Guilherme Piêgas . II. Universidade do Estado de Santa Catarina, Centro de Ciências Tecnológicas, Programa de Pós-Graduação em Computação Aplicada. III. Título: Posicionamento de Caches em Redes Celulares Heterogêneas Considerando o Roteamento de Requisições.

MARISANGILA ALVES

POSICIONAMENTO DE CACHES EM REDES CELULARES HETEROGÊNEAS CONSIDERANDO O ROTEAMENTO DE REQUISIÇÕES

Dissertação apresentada ao Programa de Pós–Graduação em Computação Aplicada do Centro de Ciências Tecnológicas da Universidade do Estado de Santa Catarina, como requisito parcial para a obtenção do grau de Mestre em Computação Aplicada.

Orientador: Dr. Guilherme Piêgas Koslovski

BANCA EXAMINADORA:

Dr. Guilherme Piêgas Koslovski
Universidade do Estado de Santa Catarina

Membros:

Dr. Omir Correia Alves Junior
Universidade do Estado de Santa Catarina

Dr. Luis Carlos Erpen de Bona
Universidade Federal do Paraná

Joinville, 26 de julho de 2022

Dedico este trabalho aos meus pais!

AGRADECIMENTOS

Agradeço ao orientador Guilherme Piêgas Koslovski pela oportunidade, confiança, incentivo, disponibilidade e, sua excelência para uma orientação constante, dedicada, solícita e criteriosa durante o planejamento e desenvolvimento desta pesquisa.

Agradeço, especialmente aos meus pais, minha mãe Elizabet Santana Müller Alves e meu pai Pedro de Oliveira Alves, exemplos de determinação, responsáveis por ensinamentos, os quais foram alicerce em toda minha jornada.

Agradeço à minha namorada Flávia Stocloska por me manter sempre confiante durante esta trajetória, pela sua compreensão, seu valioso apoio e incentivo, os quais tornaram este caminho imensamente mais agradável.

Agradeço a todos os docentes, desde o ensino fundamental até a pós-graduação, responsáveis pela fundamentação desta jornada. Sem eles esta trajetória acadêmica seria imensamente árdua.

Agradeço aos meus amigos, Éwerton de Oliveira Cercal, Mariele de Carvalho dos Santos e, minha prima Fabiana Cristina Müller, grandes responsáveis pelo incentivo e apoio durante todo percurso.

Agradeço a amizade e contribuição técnica de Talita Valéria Araújo da Matta no desenvolvimento deste projeto.

Agradeço ao professor Fabian Quevedo da Rocha, seu trabalho foi fundamental para que eu evoluísse na direção de passos importantes.

Agradeço as agências de fomento à pesquisa e aos demais responsáveis por manter resistente a ciência brasileira e demais instituições e projetos fundamentais no desenvolvimento deste trabalho. (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) e Universidade do Estado de Santa Catarina (UDESC).

Agradeço à comunidade científica, ou não seria capaz de avançar um passo sequer nesta pesquisa. Como dito: "Se eu vi mais longe, foi por estar sobre ombros de gigantes"(Issac Newton, 1675).

Por fim, agradeço à todas as pessoas que corroboraram, de forma direta em indireta para o sucesso deste trabalho.

"*Extraordinary claims require extraordinary evidence.*"(Carl Sagan)

RESUMO

O crescimento da quantidade de dispositivos móveis é realidade nos últimos anos, aliado ao aumento da demanda por tráfego de dados decorrentes, principalmente, da popularização de aplicações multimídia. Ademais, observa-se o surgimento de novas aplicações com rigorosos requisitos relacionados à latência e vazão de dados. Frente a esse cenário, as redes 5G desencadeiam uma necessária e transformadora evolução para às redes móveis. Novas tecnologias vem sendo estudadas para atender tais aplicações. Dentre essas tecnologias, *cache* de conteúdo aliado à técnicas de *Multi-Access Edge Computing* é uma alternativa promissora no que se dispõe. Entretanto, desenvolver políticas de *cache* para redes móveis têm características desafiadoras como capacidade de armazenamento limitado, diferentes padrões de popularidade de conteúdo, comportamento da rede e mobilidade do usuário. Esse trabalho tem como objetivo desenvolver um modelo matemático para uma política de *cache* cooperativa orientada à rede com o fim de reduzir a latência final percebida pelo usuário aplicável em *Heterogeneous Cellular Network*. Um modelo foi desenvolvido através de *Integer Linear Programming* conduzindo conjuntamente os problemas de inserção de conteúdo e roteamento de requisições. Simulações numéricas demonstraram que os princípios de cooperação multissaltos e orientação à rede obtiveram êxito na escolha de caminhos mantendo a latência estável. A política de *cache* cooperativa orientada à rede mostrou-se eficiente na redução da latência. Especificamente, para 99% da amostra a latência foi inferior a 10 milissegundos contra 91% para a abordagem com o modelo multissaltos.

Palavras-chave: Cache. Redes Celulares Heterogêneas. Programação Linear. Posicionamento de Cache. Roteamento de Requisições.

ABSTRACT

In past years there has been an increase in the number of mobile devices and also an increase in data traffic triggered by the popularization of multimedia applications. Moreover, there is a need deployment new applications and services with restricted requirements of delay and throughput. In this sense, the 5G contributes to the evolution of wireless networks and it could support this new environment. New technologies have been researched to support these new applications. The content caching jointly Multi-Access Edge Network is a promising alternative to achieve these challenges. However, there are still challenges concerning the design of caching policy, such as limited storage, different popularity, mobility and network congestion. The objective of this research is to develop a model of network-aware cooperative caching policies to decrease the latency experienced by user-based in Heterogeneous Cellular Network. A model was formulated through Integer Linear Programming and conducts both placement caching problems and request routing problems. Numerical simulations showed that the network dynamics sensitivity feature successfully chose paths between *cache* and content origin for different scenarios, simultaneously decreasing the network latency. The network-aware cooperative cache policy was efficient in decrease de latency. On the one hand, to 99% of the sample, the latency was lower than 10 ms. On the other hand, the model multi-hop was lower than 91%.

Keywords: Caching. Heterogeneous Cellular Networks. Linear Programming. Cache Placement. Request Routing.

LISTA DE ILUSTRAÇÕES

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

3GPP	3rd Generation Partnership Project
5G	Fifth Generation Technology Standard
5GPPP	5G Infrastructure Public Private Partnership
AR	Augmented Reality
BBU	Baseband Unit
BH	Backhaul
BILP	Binary Integer Linear Programming
BS	Base Station
C-RAN	Cloud Radio Access Network
CC	Congestion Control
CCN	Content Centric Networking
CDN	Content Delivery Networks
CoMP	Coordinated Multi-Point
CWND	Congestion Window
D2D	Device-to-Device
DAS	Distributed Antenna System
DNS	Domain Name System
E2E	End-to-End
eMBB	enhanced Mobile Broadband
ETSI	European Telecommunications Standards Institute
FH	Fronthaul
H-CRAN	Heterogeneous Cloud Radio Access Network
HCN	Heterogeneous Cellular Network
ICN	Information-Centric Networking
ILP	Integer Linear Programming
IMT-2020	International Mobile Telecommunications
IoT	Internet of Things
IP	Internet Protocol
ISP	Internet Service Provider
ITU	International Telecommunication Union

KPI	Key Performance Indicators
LFU	Least-Frequently Used
LP	Linear Programming
LRU	Least-Recently Used
M2M	Machine-to-Machine
MBS	Macro Base Station
MCFP	Multi-Commodity Flow Problem
MEC	Multi-Access Edge Computing
MILP	Mixed-Integer Linear Programming
mMTC	massive Machine Type Communications
MNO	Mobile Network Operator
QoE	Quality-of-Experience
QoS	Quality-of-Service
RAN	Radio Access Network
RRH	Remote Radio Head
RTO	Retransmission Time Out
RTT	Round Trip Time
SBS	Small Base Station
SDN	Software Defined Networking
SINR	Signal-to-Interference-Plus-Noise Ratio
SLA	Service Level Agrement
TCP	Transmission Control Protocol
UDN	Ultra-Dense Network
UE	User Equipment
UHD	Ultra-High-Definition
URLLC	Ultra-Reliable and Low-Latency Communications
VM	Virtual Machine
VoD	Video on Demand
VR	Virtual Reality
WWW	World Wide Web

SUMÁRIO

1 INTRODUÇÃO

Os dispositivos móveis estão amplamente presentes no cotidiano das pessoas e têm sido um dos principais meios de acesso a Internet para os usuários finais. Estima-se que o tráfego de dados móveis alcance 77 Exabytes até o ano de 2022 (CISCO, 2019), isto é, pode crescer até 46% em relação ao ano de 2017. Enquanto aproximadamente 5,7 bilhões de assinantes de rede móvel são esperados até 2023, representando 71% da população mundial (CISCO, 2020). Ainda, é possível alcançar 6,9 bilhões de usuários até o ano de 2026, ou seja, 45% mais usuários em relação ao ano de 2020 (ERICSSON, 2020). Estima-se que sejam trafegados até 34 GB de dados por mês através de *smartphones* no ano de 2026 (ERICSSON, 2020), ou seja, um crescimento de 24% a partir do ano de 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 emergentes, motivadas pela popularização ou surgimento de novos serviços e aplicações.

A rede *Fifth Generation Technology Standard (5G)* tem como objetivo possibilitar o suporte a essas novas demandas e, além disso, suportar maior quantidade de usuários, dispositivos finais e tráfego na rede (ITU, 2017). A definição das especificações da rede 5G é resultante de esforços de iniciativas e colaborações governamentais e comerciais tais como: *International Telecommunication Union (ITU)*, *3rd Generation Partnership Project (3GPP)* e *5G Infrastructure Public Private Partnership (5GPPP)*. De tal forma, o *International Mobile Telecommunications (IMT-2020)* definiu requisitos mínimos de desempenho e cenários de uso para as redes 5G (NAVARRO-ORTIZ et al., 2020).

Dentre as aplicações que compõem os cenários definidos, estão aquelas 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)*, indústria 4.0, entre outros. Em síntese, se faz necessário aumentar a largura de banda, reduzir a latência e otimizar o uso da capacidade energética dos dispositivos. Em geral, para ambientes urbanos e ultra densos, o IMT-2020 sugere dois valores indicativos para latência: 4 milissegundos para aplicações como *Video on Demand (VoD)* e 1 milissegundo para aplicações sensíveis a latência. Além disso, a vazão é essencial para transferir uma grande quantidade de dados, coletados a partir de dispositivos IoT ou gerados a partir de conteúdo multimídia (PHAM et al., 2020).

Nesse sentido, evitar que os dados sejam trafegados até o núcleo da rede, por meio de enlaces compartilhados, pode permitir que requisitos mínimos sejam alcançados. Em vista disso, o paradigma *Multi-Access Edge Computing* permite que recursos computacionais estejam presentes na borda da rede, isto é, próximos geograficamente do usuário. O paradigma MEC tem como principal objetivo reduzir o congestionamento, aumentando a capacidade da rede e reduzindo a latência, possibilitar maior privacidade dos dados e ampliar a autonomia energética de dispositivos de borda (ABBAS et al., 2018).

De forma complementar, *Heterogeneous Cellular Network* permitem implantar *Base*

Stations (BSs) com capacidade e recursos heterogêneos dentro da rede móvel (A BS é o meio de conexão entre o dispositivo final e a Internet). Portanto, é possível expandir a capacidade da rede móvel a partir de elementos heterogêneos tais como *Small Base Stations (SBSs)* (DAMNJANOVIC et al., 2011). As SBSs têm menor capacidade, se localizam mais próximas ao usuário e podem abrigar recursos computacionais (KAMEL; HAMOUDA; YOUSSEF, 2016).

Juntamente, MEC e HCN podem ser combinadas e, portanto, permitir que dispositivos com diferentes capacidades computacionais estejam próximos ao usuário (ANDREWS, 2013). Assim, conciliadas, tais tecnologias permitem a implantação de *cache* de conteúdo em redes móveis, através de servidores MEC implantados em BSs. Consequentemente, implantar *cache* de conteúdo dentro da rede móvel, é alternativa promissora. O posicionamento de *cache* próximo ao usuário tem como finalidade contribuir para à redução de latência, originada pela proximidade e redução de transmissão de dados replicados no enlace de *Backhaul (BH)*. O BH é o enlace de conexão entre rede móvel e Internet (PASCHOS et al., 2018) (KABIR et al., 2020) (WU et al., 2021).

Os projetos para *cache* em redes móveis possuem dois problemas enfatizados na literatura. O primeiro problema diz respeito à inserção de conteúdo, é a fase que determina qual, onde e como o conteúdo deve ser armazenado (WU et al., 2021). O segundo problema está relacionado a entrega de conteúdo e à forma como o conteúdo será entregue ao usuário. Além disso, o segundo problema é composto pela fase de roteamento de requisições (DEHGHAN et al., 2017) e associação do usuário à BS (HARUTYUNYAN; BRADAI; RIGGIO, 2018), tais fases consistem em otimizar e escolher o caminho de origem e destino do conteúdo, respectivamente.

Em geral os trabalhos presentes na literatura têm como principal interesse o problema de inserção de conteúdo (SHANMUGAM et al., 2013) (BASTUG; BENNIS; DEBBAH, 2014). Por sua vez, dentre os trabalhos direcionados ao problema de roteamento de requisições, alguns propõem abordagens cooperativas, nas quais a capacidade de armazenamento é 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. No entanto, tais abordagens consideram somente BSs vizinhas (JIANG; FENG; QIN, 2017).

Ainda, há abordagens multissaltos, em que a cooperação é realizada sob perspectiva global. Desse modo, pode ampliar a capacidade de atender solicitações de conteúdo através de *cache* e,consequentemente, evita o tráfego através do enlace de BH para consumir o conteúdo diretamente de sua origem. Dentre as abordagens que consideram a busca cooperativa multissaltos (SONG et al., 2021; LI et al., 2017), a mobilidade não é um fator fundamental a ser considerado no desenvolvimento de políticas de *cache*.

Além dos aspectos e abordagens mencionados, considerar o estado da rede é de extrema importância. O estado da rede é alterado de acordo com a variabilidade das condições (vazão ou latência) originada por possíveis mudanças nas demandas. Tais mudanças podem causar congestionamentos ou sobrecarga na rede (HARUTYUNYAN; BRADAI; RIGGIO, 2018). Diferentemente dos trabalhos mencionados, 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. Dentre as abordagens disponíveis para gerenciamento (*e.g.*, reserva de recursos, marcação de pacotes), é possível estimar a sobrecarga de um enlace e, acompanhar a dinâmica da rede, como amplamente utilizado pelos algoritmos de *Congestion Control (CC)* que fazem parte do *Transmission Control Protocol (TCP)*. Destaca-se que a estimativa de vazão é baseada em premissas consolidadas em redes de computadores (BRAKMO; PETERSON, 1995). No entanto, se propõe que a estimativa de vazão seja atribuída a camada de aplicação, em vista disso, apenas os conceitos de algoritmos de CC do TCP são utilizados na camada de aplicação, os quais conduzem a política de *cache* orientada à rede. Ou seja, a política não atua no TCP, apenas utiliza os conceitos fundamentais sobre estimativa de vazão e latência.

A questão de pesquisa deste trabalho é dada pelos problemas de inserção de conteúdo e roteamento de requisições e, resumidamente, pergunta: A partir de qual estratégia é possível reduzir a latência *End-to-End (E2E)* no que compreende os problemas de inserção de conteúdo e roteamento de requisições para compor uma política de *cache*? Portanto, este trabalho propõe um modelo para uma política de *cache* cooperativa orientada à rede, unindo os problema de inserção de conteúdo e roteamento de requisições multissaltos com objetivo de minimizar a latência em HCN. Ademais, a cooperação multissaltos é dada através de uma perspectiva global, em outras palavras, a busca ou posicionamento do conteúdo é realizada em toda à rede móvel e consiste na cooperação entre as BSs sem hierarquia definida. Um modelo é formulado matematicamente através de *Integer Linear Programming (ILP)*.

Sendo assim, a formulação proposta está sujeita à restrições de capacidade de armazenamento e requisitos de *Quality-of-Service (QoS)* definidos pelo *Service Level Agrement (SLA)* e, ambos os problemas tratados de forma conjunta permitem realizar balanceamento entre a carga dos enlaces e a capacidade de armazenamento dos servidores de borda. A política de *cache* é desenvolvida a partir da perspectiva da rede, portanto, pode ser configurada e administrada pelo *Mobile Network Operator (MNO)*, que é provedor e administrador de infraestrutura móvel.

Para isso, tem-se como hipótese, que a política de *cache* escolha os caminhos com maior vazão e, consequentemente, menor *Round Trip Time (RTT)*, de tal forma pode evitar a diminuição da QoS (CHIU; JAIN, 1989). Além disso, a política de *cache* busca atender as solicitações sempre em *cache*, ou seja, sem trafegar o conteúdo através do enlace de BH. No entanto, a política de *cache* realiza o balanceamento da carga entre consumo através do BH ou a partir da *cache*, desse modo prioriza a minimização da latência. Por fim, a cooperação multissaltos através de uma perspectiva global permite que a busca de conteúdo seja realizada em todas BS, proporcionando maior taxa de *cache hit*.

Simulações numéricas foram realizadas a partir do modelo formulado através de ILP e executado a partir de um administrador de eventos discretos. Tais simulações demonstraram que os princípios cooperação multissaltos e orientação à rede, ou seja, a estimativa de sobrecarga, obteve êxito na escolha de caminhos entre *cache* e conteúdo original em diferentes cenários, sem impacto significativo na latência. Obteve média de *cache hit* $6,3$ vezes maior em relação a

média do modelo de único salto. Destaca-se que o modelo para à política de *cache* cooperativa orientada à rede mostrou-se eficiente na redução da latência. Para 99% da amostra a latência foi inferior a 10 milissegundos contra 91% para o modelo multissaltos. Sobretudo, demonstrou estabilidade na latência, devido distribuição do tráfego e balanceamento de carga entre conteúdo trafegados através de BH e *cache*.

1.1 OBJETIVOS

O objetivo geral do presente trabalho é especificar, desenvolver e analisar um modelo para uma proposta de política de *cache* cooperativa orientada à rede com a finalidade de reduzir a latência em HCN. Essa política, considera os problemas de inserção de conteúdo e roteamento de requisições, respeitando restrições de capacidade de armazenamento e de QoS determinadas pelo SLA.

1.1.1 Objetivos Específicos

De acordo com o objetivo geral, os seguintes objetivos específicos foram definidos e são elencados:

- Realizar revisão bibliográfica, analisar e comparar trabalhos relacionados, nos quais o problema de pesquisa é direcionado para o roteamento de requisições.
- Especificar e desenvolver uma política de *cache* cooperativo orientado à rede baseada em mecanismos presentes no TCP (tais mecanismos podem estimar a vazão atual da rede).
- Definir um modelo generalizável através de ILP para a política de *cache* cooperativa orientado à rede, que consiste em uma função objetivo, variáveis de decisão e suas restrições.
- Implementar o modelo ILP através de um solucionador.
- Executar simulação numérica a partir de parâmetros definidos.
- Coletar e interpretar dados obtidos através da simulação numérica.
- Apresentar e analisar resultados numéricos de acordo com as métricas definidas.

1.2 METODOLOGIA

O presente trabalho pode ser classificado quanto sua natureza, objetivo, lógica, procedimentos e abordagem. Primeiramente, quanto a natureza, classifica-se como ciência aplicada, ou seja, parte de uma solução que pode ser implementada tecnologicamente (PRODANOV; FREITAS, 2013). Além disso, seu objetivo caracteriza-se como descritivo, isto é, consiste em

um estudo, no qual estabelece a análise de relacionamento entre variáveis (GIL, 2010) (WAZLAWICK, 2010) (PRODANOV; FREITAS, 2013). Quanto à abordagem do método científico, designa-se a partir de uma lógica dedutiva. Portanto, parte de uma premissa, isto é, uma regra geral para validar a hipótese (MARCONI; LAKATOS, 2003). Em outras palavras, baseia-se em uma premissa consolidada em redes de computadores, na qual determina que quanto menor o RTT maior a vazão (CHIU; JAIN, 1989), outro sim, quanto maior a vazão melhores são as condições da rede (STALLINGS, 2015). O procedimento técnico é de característica experimental, através da justificativa de se caracterizar como uma análise de relacionamento entre variáveis (*e.g.*, a vazão e o RTT têm impacto no posicionamento e roteamento e, consequentemente, na latência da entrega como um todo). De tal modo, são realizados controle e interferência em tais variáveis (PRODANOV; FREITAS, 2013)(GIL, 2010). Por fim, esse trabalho possui abordagem quantitativa, ou seja, as variáveis são em geral quantificáveis, (*e.g.*, como medida de tempo de RTT ou tamanho de conteúdo em *bytes*). A pesquisa com carácter quantitativo analisa uma grande quantidade de dados e, considera que os dados naturalmente sejam ou, podem ser transformados em dados quantificáveis. Ademais, a abordagem quantitativa resulta em dados amplos, estruturados e conclusões objetivas (GIL, 2010) (PRODANOV; FREITAS, 2013).

1.3 CONTRIBUIÇÕES

As contribuições deste trabalho podem ser sumarizadas como segue:

- Uma proposta de uma política para *cache* cooperativa orientada à rede, a qual se propõe a unir duas questões de pesquisa relevantes relacionadas a *cache* em redes móveis, com objetivo de minimizar a latência. Tais questões são identificadas como problema de inserção de conteúdo e o problema de roteamento de requisições.
- Um modelo generalizável que pode ser empregado em demais aplicações que possam surgir futuramente, além das atuais aplicações de VoD.
- Um modelo proposto que pode ser utilizado em benefício do MNO e simultaneamente beneficiar o provedor de conteúdo ou provedor de serviço, no qual as definições SLA são parâmetros centrais nesta política de *cache*. Ainda, o principal beneficiado é o usuário.
- Um modelo pode ser usado para orientar o desenvolvimento de heurísticas futuramente.
- Uma comparação com outras estratégias para políticas de *cache* existentes na literatura.

1.4 ORGANIZAÇÃO DO TEXTO

Este trabalho está estruturado de tal forma: o Capítulo 2 apresenta os tópicos necessários para contextualização do problema tais como, conceitos gerais relacionados a *cache*, computação de borda e redes móveis e, além disso, destaca trabalhos presentes na literatura. Por sua vez,

o Capítulo 3 define o problema e aborda detalhadamente as premissas e a hipótese central da política de *cache* e apresenta a notação matemática bem como a representação através de estrutura de grafos. O Capítulo 4 detalha o método, um modelo de ILP, enquanto o Capítulo 5 apresenta resultados de simulações numéricas e discussões a partir dos resultados obtidos. O Capítulo 6 destaca as principais considerações a respeito deste trabalho, bem como apresenta os trabalhos futuros.

2 REVISÃO DE LITERATURA

O presente capítulo apresenta a fundamentação teórica necessária para compreensão do contexto, metodologia e proposta posteriormente descrita. Portanto, visita conceitos relacionados a *cache*, redes móveis e computação de borda nas Seções 2.1, 2.2 e 2.3, respectivamente. Expõe alguns dos desafios pertinentes a *cache* de conteúdo em redes móveis na Seção 2.4. A Seção 2.5 descreve sintetiza o algoritmo de CC do TCP *Vegas*. A Seção 2.6 apresenta uma breve compreensão a respeito da natureza da complexidade do problema. Por fim, em síntese, destaca os trabalhos relacionados na Seção 2.7 e considerações desse capítulo na Seção 2.8.

2.1 CACHE EM REDES DE COMPUTADORES

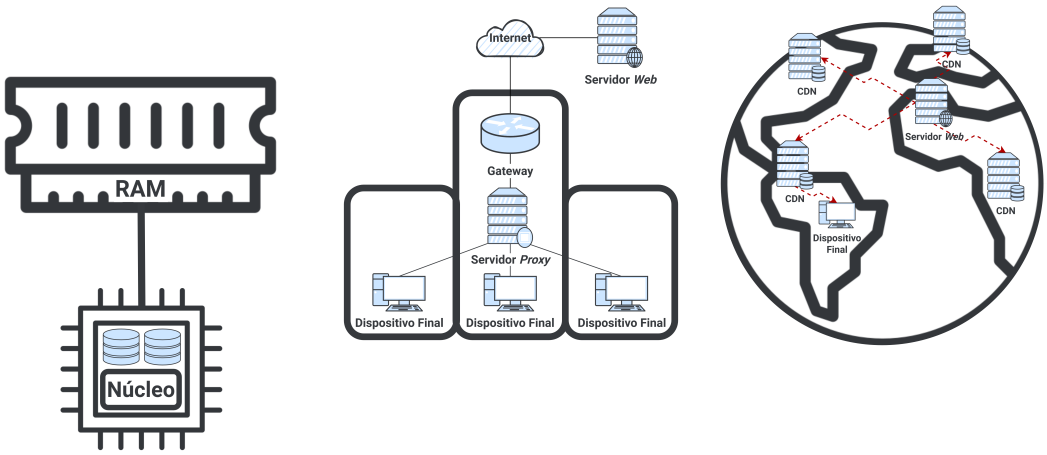
O termo *cache* é de origem francesa que significa armazenar. No contexto computacional, *cache* refere-se ao armazenamento e cópia de informações recentemente acessadas, e que podem ser futuramente recuperadas ou rejeitadas. O *caching* de informação somente obtém vantagem se o seu custo de recuperação é menor em relação ao custo de acesso ao conteúdo original (WESSELS, 2001). Uma das principais métricas de desempenho é a taxa de acertos de *cache*. Dito de outro modo, é a razão entre solicitações atendidas em *cache*, ou seja, *cache hit*, que é dividido pelo total de solicitações recebidas. As solicitações não beneficiadas pelo uso do *cache* e, portanto, atendidas pela entidade de origem do conteúdo são chamadas *cache miss* (KABIR et al., 2020).

Inicialmente, o *cache* foi restrito ao contexto local, ou seja, somente processadores e sistemas operacionais (Figura 1a) (VU et al., 2020). No entanto, com o surgimento da Internet e posteriormente a *World Wide Web (WWW)* por volta da década de 90, esse conceito expandiu-se e a técnica passou a ser incorporada por entidades presentes em redes de computadores, tal como servidores *proxy*. A Figura 1 diferencia os contextos e a evolução do conceito de *cache* ao longo dos anos. Nesse sentido, o tema *caching* de conteúdo é apresentado sob a ótica de serviços *Web*, tais como: *Web Caching* e *Content Delivery Networks (CDN)* nas Subseções 2.1.1 e 2.1.2), respectivamente.

2.1.1 *Web Caching*

Servidores *proxy* são capazes de interceptar solicitações, armazenar o conteúdo solicitado e futuramente responder diretamente ao solicitante. Dessa forma, evita que a busca do conteúdo seja realizada diretamente ao servidor remoto e, além disso, inibe a replicação do conteúdo, como exemplificado da Figura 1b. Consequentemente, tem vantagens tais como: redução de tráfego (e consequentemente consumo de largura de banda), redução de latência e redução de sobrecarga no servidor *Web*. Isso se deve a proximidade geográfica entre *cache* e solicitante e, a redução do tráfego replicado nos principais pontos de compartilhamento de enlaces da rede. Tais benefícios foram observados no advento da *Web* e, além desses: rápido acesso, robustez,

Figura 1 – Sistemas para *cache* de conteúdo.



(a) Cache

(b) Web Caching

(c) CDN

Fonte: Elaborado pela autora (2022).

transparência, escalabilidade, eficiência, adaptatividade, estabilidade, balanceamento de carga, heterogeneidade e simplicidade, eram características previamente desejáveis para um sistema de *Web caching* (WANG, 1999).

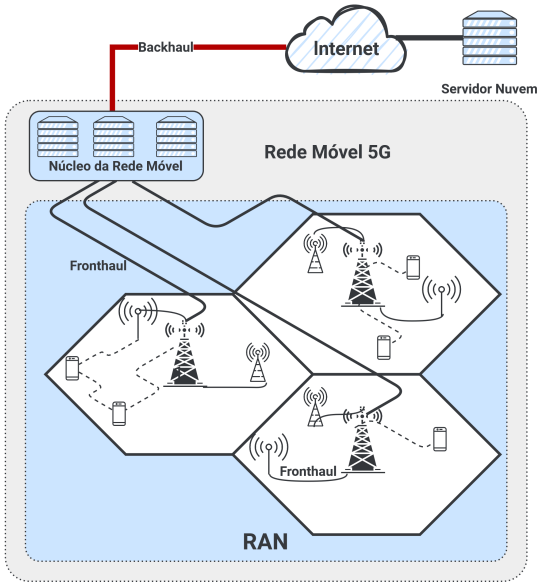
Porém, a utilização de servidores *proxy* possui limitações. Devido a existência de conteúdos dinâmicos, a taxa de *cache hit* limita-se por volta de 40% (WANG, 1999). Tal limitação, compromete a melhoria de escalabilidade, confiabilidade e desempenho. Além disso, a falta de padronização e a demanda de esforços para implementação desmotivam administradores de *Internet Service Provider (ISP)*. No entanto, as CDN, surgiram para superar essas limitações, incentivar e padronizar a implantação de *cache* de conteúdo nas redes de borda, ou seja, em ISPs locais (DILLEY et al., 2002).

2.1.2 *Content Delivery Networks*

Uma CDN é composta por um conjunto de servidores distribuídos hierarquicamente ao longo da Internet com o objetivo de realizar entrega de conteúdo. Dessa forma, réplicas de conteúdo são dispostas em diferentes níveis hierárquicos, o último nível se localiza na borda da rede, em outras palavras, dentro de ISPs locais. Portanto, o conteúdo pode ser consumido diretamente de um CDN que encontra-se geograficamente mais próximo do solicitante, como é possível observar na Figura 1c. Nesse sentido, é possível considerar que a estratégia de CDN é semelhante ao *Web caching* em servidores *proxy*. Entretanto, dispõe de um controle de mapeamento automático baseado em resolução de *Domain Name System (DNS)*, isto é, realiza a seleção entre os servidores pertencentes a hierarquia baseado em critérios, tais como carga do servidor, suporte ao tipo de solicitação, condições da rede, localização do solicitante e disponibilidade de conteúdo (DILLEY et al., 2002). Em resumo, além da entrega conteúdo estático, possibilita a entrega de conteúdo dinâmico e *streaming* de áudio e vídeo, ou seja, CDN

não é responsável apenas pela entrega de conteúdo armazenável em *cache*. Embora, CDN seja uma estratégia promissora e amplamente difundida (PASCHOS et al., 2018) não é capaz de evitar a sobrecarga gerada em enlaces de BH (SHANMUGAM et al., 2013). O BH consiste em um enlace de conexão entre o núcleo da rede móvel e a rede além do domínio do MNO, em outras palavras, a Internet. Por sua vez, o Fronthaul (FH) é o enlace de conexão entre a Radio Access Network (RAN)(toda infraestrutura de rede móvel entre o usuário e o núcleo da rede móvel) e o núcleo da rede móvel. Esses elementos são ilustrados na Figura 2.

Figura 2 – Rede Móvel 5G.



Fonte: Elaborado pela autora (2022).

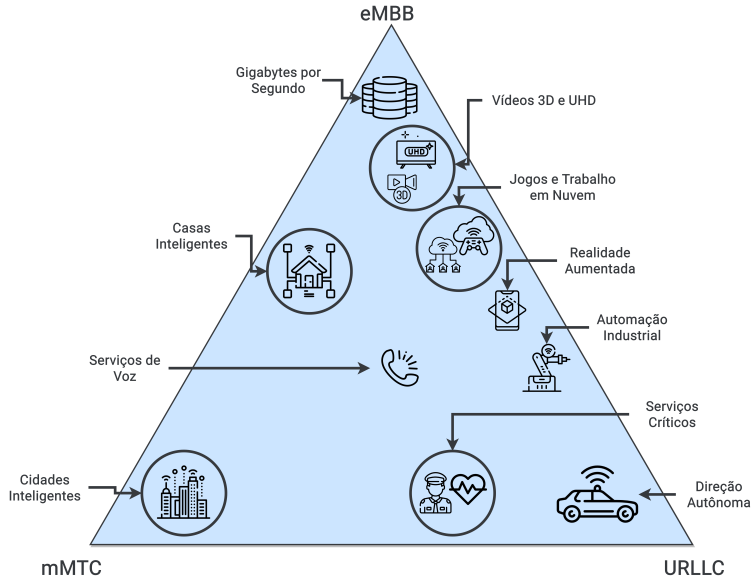
2.2 REDES MÓVEIS

As redes móveis têm evoluído ao longo do anos, com o objetivo suportar novas aplicações e tecnologias. A primeira geração (1G) proporcionou a comunicação por voz, enquanto a segunda geração (2G) permitiu o envio de mensagens de texto. Por sua vez, a terceira geração (3G) passou a transferir dados e, possibilitou a utilização de serviços baseados em *Internet Protocol (IP)*, enquanto a quarta geração (4G) consolidou e ampliou capacidade de envio de dados, permitindo *live streaming* e *on-demand streaming* (PHAM et al., 2020).

Atualmente, a rede denominada 5G está sendo implantada em diversos países, resultado dos esforços de organizações como ITU, *European Telecommunications Standards Institute (ETSI)* e demais parceiros que conduzem pesquisas para desenvolvimento e padronização da rede 5G através de projetos de padronização como 3GPP e 5GPPP. Especificamente, o padrão IMT-2020 formalizou e estabeleceu as especificações mínimas para que a rede 5G fosse implantada até o ano de 2020 (NAVARRO-ORTIZ et al., 2020). Essas especificações indicam que a rede

5G deve prover serviços em três cenários, conforme apresentados na Figura 3: *Ultra-Reliable and Low-Latency Communications (URLLC)*, *enhanced Mobile Broadband (eMBB)* e *massive Machine Type Communications (mMTC)*.

Figura 3 – Cenários IMT-2020.



Fonte: (ITU, 2015)

Aplicações URLLC possuem requisitos rígidos de latência, exigidos por aplicações tais como: direção autônoma, automação industrial ou serviços críticos ligados a segurança e saúde. O IMT-2020 estabelece que latência seja inferior a 1 milissegundo. Para aplicações *eMBB* que incluem a transmissão de vídeos *Ultra-High-Definition (UHD)*, vídeos 3D, AR e VR, espera-se que a taxa de dados mínima experienciada pelo usuário seja 100 Mbps para *download* e 50 Mbps para *upload* e que a latência seja inferior a 4 milissegundos. Como observado na Figura 3, algumas aplicações se encaixam em caminhos intermediários entre esses cenários. Por fim, no cenário mMTC engloba a comunicação *Machine-to-Machine (M2M)*, ou seja, cidades e casas inteligentes e os demais sensores inseridos no contexto de *IoT*, os requisitos se resumem em melhorias da autonomia energética e da capacidade de manter áreas com alta densidade de dispositivos (ITU, 2017).

Os requisitos impostos pela rede 5G criaram diversas oportunidades de pesquisa, que consequentemente resultaram na criação, aprimoramento e utilização combinada de tecnologias gerenciais (BOGALE; LE, 2016; WU et al., 2018; LIU et al., 2018; ANDREWS, 2013; KAMEL; HAMOUDA; YOUSSEF, 2016; BARAKABITZE et al., 2020; PHAM et al., 2020). No escopo do presente trabalho, HCN merece destaque por contribuir para o aumento da qualidade dos serviços e da capacidade da rede móvel, devido a proximidade com o dispositivos finais (LIU et al., 2018).

2.2.1 *Heterogeneous Cellular Network*

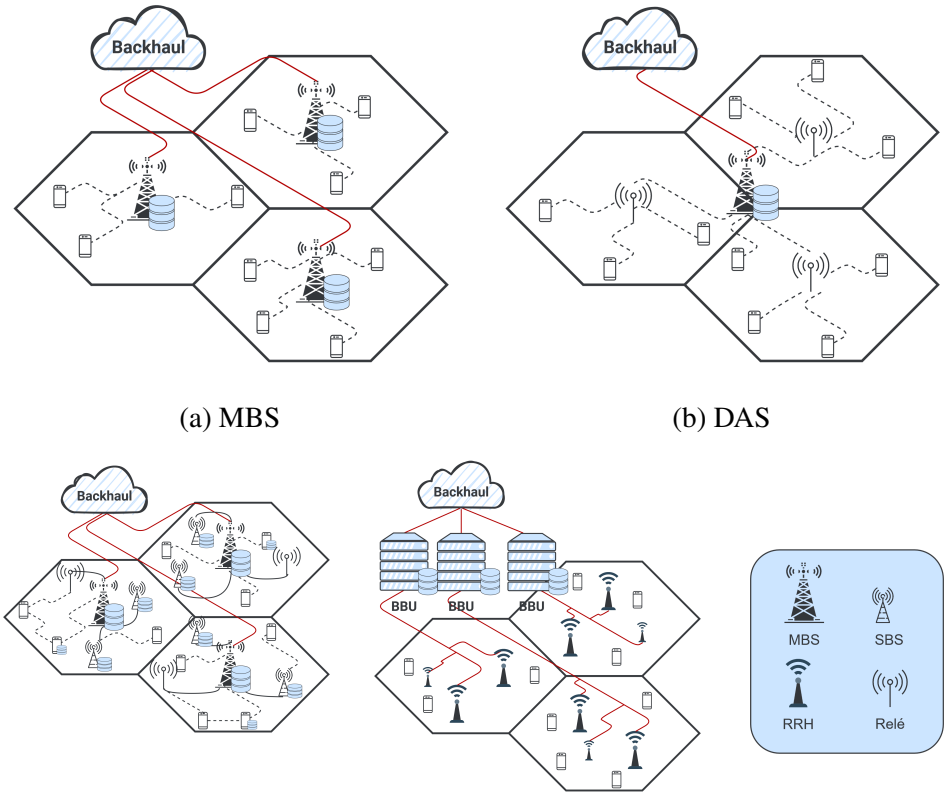
As *HCN* são compostas por BS de alta potência e BS de baixa potência denominadas, *Macro Base Station (MBS)* e SBS, respectivamente. Ao contrário das redes homogêneas, HCNs possuem BSs com diferentes capacidades, potência de sinal, cobertura e frequência, como pode ser observado as diferenças na Figura 4a e 4c. Portanto, HCN tem como principal objetivo melhorar a eficiência espectral e a capacidade da rede (ANDREWS, 2013).

As BS devem ser capazes de atender e comportar pedidos de canais de comunicação solicitado pelo dispositivo final, fornecer um meio de conexão ao núcleo da rede por meio do BH e, para tal, é necessário dispor de uma fonte de energia permanente (ANDREWS, 2013). Por outro lado, relés são extensões de área de cobertura conectados à BS através de meio sem fio (DAMNJANOVIC et al., 2011). Na Figura 4b e 4c, é possível observar a presença de relés em arquiteturas *Distributed Antenna System (DAS)* e HCN.

As SBSs possuem arquitetura distribuída de modo que exercem funções de forma independente, semelhante ao cenário com MBSs, ou seja, são entidades iguais com diferentes capacidades. A arquitetura *Cloud Radio Access Network (C-RAN)* ou *Heterogeneous Cloud Radio Access Network (H-CRAN)* é uma arquitetura de processamento centralizado, em que funções de múltiplas BS são centralizadas em *Baseband Unit (BBU)* e unidades de rádio frequência são instaladas em *Remote Radio Head (RRH)* distribuídas ao longo das células, como pode ser observado na Figura 4d. A arquitetura C-RAN ou H-CRAN é alternativa a arquitetura SBS. Do mesmo modo, SBSs se diferenciam da arquitetura DAS ilustrada na Figura 4b, na qual consiste em uma única BS com unidades de rádio frequência distribuídas para otimizar e estender sua cobertura e potência. A Figura 4 ilustra a diferença entre as arquiteturas mencionadas (KAMEL; HAMOUDA; YOUSSEF, 2016).

A densidade de uma rede móvel *Ultra-Dense Network (UDN)* é mensurada de acordo com a quantidade de células, a qual pode ser definida pela presença de mais células do que usuários ativos na rede com objetivo de ampliar a capacidade de comunicação, como ilustrado na Figura 4c. Dessa forma, é plausível que um dispositivo possa se conectar com múltiplas células. Ademais, densificação pode ser alcançável através de arquitetura distribuída (SBS), ou centralizada (H-CRAN). Ambas arquiteturas possuem vantagens e desvantagens a ser consideradas (KAMEL; HAMOUDA; YOUSSEF, 2016). No entanto, SBSs proporcionam capacidade de armazenamento e recursos computacionais distribuídos, os quais permitem evitar sobrecarga em entidades centralizadoras e, além disso, a autonomia de SBS permite a implantação de servidores MEC que possam realizar tarefas completas tais como: computação *offloading* e *cache* de conteúdo com maior proximidade ao solicitante e, consequentemente é possível reduzir o tráfego no BH (TRAN et al., 2017). Os termos SBS e MBS são definidos como semelhantes neste trabalho, entretanto, SBS possuem menor potência de sinal.

Figura 4 – Arquiteturas.



(a) MBS (b) DAS

(c) HCN/UDN (d) C-RAN/H-CRAN

Fonte: Elaborado pela autora (2022).

2.2.2 Futuro das Redes Móveis

Embora, até o momento, a rede 5G esteja em desenvolvimento e aperfeiçoamento, há pesquisas direcionadas para a próxima evolução em redes móveis, a sexta geração (6G). Espera-se que a rede 6G seja centrada em seres humanos, ao contrário da rede 5G que é centrada em aplicações e dispositivos (DANG et al., 2020). Futuramente, há expectativas, nas quais novas aplicações tenham requisitos que possivelmente não possam ser atendidos pela rede 5G tal como taxa de dados experienciada pelo usuário na ordem de 1 Gbps, 10 vezes maior que o exigido em redes 5G (TARIG et al., 2020).

Intuitivamente espera-se que mesmo com a redução da latência das redes móveis, o uso de *cache* de conteúdo seja difundido e motivado, fundamentalmente porque é necessário que a rede suporte uma grande quantidade de dispositivos, nesse sentido, o *cache* de conteúdo é importe para otimizar a eficiência no uso de enlaces compartilhados no núcleo da rede, de modo que evite tráfego replicado.

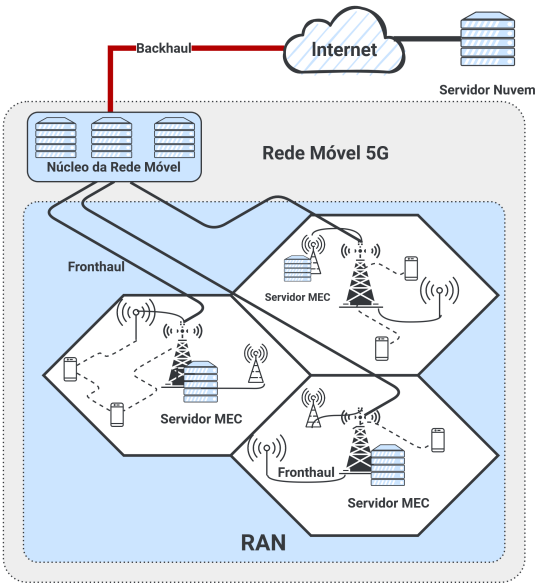
2.3 *Multi-Access Edge Computing*

Devido ao elevado número de dispositivos conectados na borda da rede oriundos da consolidação da IoT, a computação centralizada na nuvem não é suficiente para atender a demanda em termos de qualidade de serviço (*e.g.*, latência, tempo de processamento) e volume de dados. Além de gerar aumento de tráfego em enlaces compartilhados, o aumento do número de dispositivos conectados contribui para o aumento da latência percebida pelo usuário (ou aplicação) final (SHI et al., 2016).

Diante do exposto, o paradigma de *Edge Computing* estabelece que os recursos computacionais estejam presentes também na borda da rede. Quaisquer dispositivos entre o solicitante e a nuvem computacional podem ser considerados um dispositivo de borda, tais como *smartphones* e roteadores. A diversidade de dispositivos de borda é exemplificada na Figura 6. Portanto, *Edge Computing* permite que recursos computacionais estejam presentes em quaisquer entidades na borda da rede, extraindo a vantagem da distância geográfica entre o dispositivo final e a entidade de processamento para redução da latência (SHI et al., 2016).

Especificamente, MEC determina que recursos computacionais, semelhantes aos recursos existentes em nuvens computacionais, estejam próximos aos dispositivos finais (HU et al., 2015). O paradigma MEC permite o posicionamento de recursos computacionais especificamente dentro da *RAN*, que é toda infraestrutura de rede móvel entre o usuário e o *BH* (Meio de conexão entre a rede móvel e a infraestrutura de Internet), como ilustrado na Figura 5 (HU et al., 2015). Além disso, MEC tem como objetivo unir serviços de telecomunicação e computação em nuvem, apresentando-se como evolução da computação em nuvem (KEKKI et al., 2018).

Figura 5 – Rede Móvel 5G e MEC.

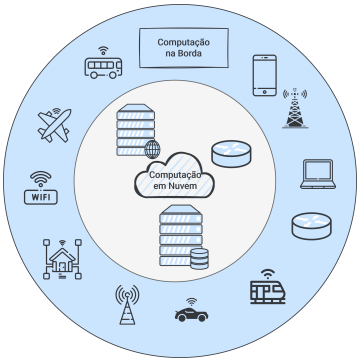


Fonte: Elaborado pela autora (2022).

Frente aos desafios e necessidades de redução da latência e eficiência no consumo de

largura de banda apresentados pela rede 5G, o paradigma MEC é um dos principais pilares para superação desses desafios. Sobretudo, os principais objetivos do MEC são reduzir o congestionamento, aumentando a capacidade da rede e reduzir a latência, possibilitar maior privacidade dos dados e ampliação da autonomia energética de dispositivos de borda.

Figura 6 – Computação em Nuvem e Computação de Borda.



Fonte: Elaborado pela autora (2022).

Sendo assim, o desenvolvimento de tecnologias relacionadas a utilização de *cache* de conteúdo são capazes de possibilitar a diminuição de tráfego de dados replicados em enlaces compartilhados (PHAM et al., 2020).

2.4 *CACHE* NA BORDA DE REDES MÓVEIS

Vantagens da combinação de técnicas MEC e a densificação de redes heterogêneas mostram-se promissoras para a utilização de *cache* de conteúdo na borda de rede móvel. Simultaneamente, a maior capacidade e maior proximidade usufruindo de recursos computacionais, rede, infraestrutura e armazenamento possibilitam evitar congestionamentos no BH devido as solicitações atendidas brevemente na borda da rede. Dessa forma, o *cache* de conteúdo implantado na RAN reduz o tráfego de conteúdo replicado e reduz a distância entre o solicitante e o conteúdo. Portanto, contribui para a redução do tráfego no BH e para diminuição da latência. Nesse sentindo, *cache* de conteúdo na borda da rede é um tema de relevância e uma tecnologia emergente que pode contribuir para que *Key Performance Indicators (KPI)* referentes à rede 5G sejam cumpridos (PARVEZ et al., 2018). No entanto, projetar políticas de *cache* de conteúdo pode ser um problema desafiador (WU et al., 2021). Os termos política de *cache*, abordagem e estratégia de *cache* serão considerados como termos equivalentes neste trabalho.

2.4.1 Desafios no Desenvolvimento de Políticas de *Cache* em Redes Móveis

As políticas de *cache* em redes móveis possuem objetivos distintos e implicam na otimização de diferentes métricas (WU et al., 2021). Dentre as politicas de *cache*, há estratégias cooperativas que buscam otimizar a capacidade de armazenamento, isto é, permitem que BSs

cooperem entre si para buscar o conteúdo entre outras BSs, além da BS em que o solicitante está conectado. A cooperação entre BSs amplia a capacidade de armazenamento de *cache* e evita que o conteúdo seja recuperado diretamente de um servidor remoto, bem como a redundância de conteúdo dentro da RAN (JIANG; FENG; QIN, 2017). Além disso, a cooperação pode ser hierárquica, na qual o conteúdo é buscado e associado de acordo com níveis da topologia da rede (LI et al., 2017).

Uma segunda estratégias considera o *cache* codificado e não codificado, no qual particiona o conteúdo e os armazena em locais diferentes e permite servir múltiplos dispositivos ao mesmo tempo (MADDAH-ALI; NIESEN, 2014). Ademais, em redes HCN há estratégias em que solicitante pode recuperar o conteúdo a partir de um outro dispositivo final através de conexão *Device-to-Device (D2D)*. No entanto, a capacidade de armazenamento é limitada e além disso, pode depender da localização e cooperação entre os usuários (JIANG; FENG; QIN, 2017).

Essas estratégias por sua vez podem ser gerenciadas de forma centralizada ou distribuída. O gerenciamento centralizado permite alcançar a solução ótima, no entanto, é possível que sejam gerados atrasos e sobrecarga no enlace de conexão com a entidade centralizadora. Além disso, as soluções centralizadas, em geral, possuem elevada complexidade computacional. Por outro lado, as políticas de *cache* distribuídas consideram informações locais, que podem comprometer a integridade do resultado final. Por fim, políticas distribuídas podem ser implementadas com baixa complexidade computacional, entretanto sem alcançar a solução ótima (WU et al., 2021).

Dentre os problemas mais desafiadores, apresentam-se as políticas de *cache* proativas ou probabilísticas para inserção e substituição de conteúdo. Em suma, o conteúdo é armazenado antes que seja solicitado, baseando-se em estimativas probabilísticas ou na identificação de padrões de comportamentos do usuário (BASTUG; BENNIS; DEBBAH, 2014). De outro modo, a estratégia reativa armazena o conteúdo no momento em que são solicitados para que seja acessado futuramente caso seja requisitado novamente (GOIAN et al., 2019). Além do mais, de acordo com a variabilidade da capacidade de recursos de armazenamento e capacidade do BH é possível que balanceamento entre esses parâmetros garanta a otimização de diferentes métricas (LI et al., 2017).

Para que sejam projetadas, políticas de *cache* têm dois principais problemas: inserção e substituição de *cache* e o problema da entrega de conteúdo, que podem ser tratados em conjunto (KHREISHAH; CHAKARESKI; GHARAIBEH, 2016) (PU et al., 2018) (DEHGHAN et al., 2017) (SONG et al., 2021) ou separadamente (JIANG; FENG; QIN, 2017) (LI et al., 2017), como fases para inserção/substituição (SHANMUGAM et al., 2013) e posterior de entrega de conteúdo (SHENG et al., 2016).

2.4.1.1 Inserção e Substituição de Conteúdo

Estratégias de inserção ou substituição de conteúdo agregam problemas resumidos pelas seguintes questões: Qual conteúdo deve ser armazenado? Onde o conteúdo deve ser armazenado? e Como o conteúdo deve ser armazenado? As duas primeiras perguntas são as que abarcam

provavelmente o maior desafio de se projetar políticas de *cache* pró-ativas, no qual um desafio é presente na dificuldade de prever o comportamento do usuário. A popularidade do conteúdo pode variar ao longo do tempo, de acordo com a localidade e de usuário para usuário (SHANMUGAM et al., 2013). Algumas estratégias probabilísticas buscam prever o comportamento de usuário e suas preferências. Dentre elas, há estratégias que conectam às relações em redes sociais de usuários e suas preferências (BASTUG; BENNIS; DEBBAH, 2014).

A última questão, como armazenar o conteúdo, se refere à substituição de conteúdo, devido a limitação da capacidade de armazenamento em *cache* e a variabilidade de popularidade do conteúdo. Politicas como: *Least-Frequently Used (LFU)* e *Least-Recently Used (LRU)* são popularmente utilizadas. Os termos de inserção de conteúdo ou posicionamento de *cache*, ambos presentes na literatura serão considerados neste trabalho como termos intercambiáveis.

As políticas de *cache* podem ser elaboradas com diferentes princípios, ou seja, as métricas de otimização podem estar direcionadas ao usuário, ao provedor de conteúdo ou ao operador de rede móvel (WU et al., 2021). O provedor de conteúdo ou serviço, é responsável por disponibilizar o conteúdo bem como a aplicação ou plataforma de acesso para o usuário final. Por sua vez, o provedor de rede, ou seja, MNO considerando um cenário de redes móveis, é responsável por fornecer a infraestrutura de comunicação entre o usuário e o provedor de conteúdo ou serviço. Dito de outro modo, o MNO é responsável por administrar a rede e fornecer a infraestrutura necessária que compõe a RAN tais como: os servidores MEC integrados as BSs, no qual consistem nos servidores de *cache* de conteúdo, bem como os enlace de *FH* e BH. O FH são os enlaces de conexão dentro da RAN, como ilustrado na Figura 2. Além disso, disponibilizam um meio de conexão ao núcleo da rede móvel e, consequentemente, a conexão com a Internet.

2.4.1.2 Entrega de Conteúdo

A entrega de conteúdo, pode ser entendida como uma ampla questão que engloba todo o processo de solicitação do conteúdo pelo usuário até o seu recebimento. É possível resumir a entrega de conteúdo em duas principais questões: roteamento de requisições e associação de usuário. O roteamento de requisições está relacionado a origem do conteúdo, ou seja, a origem do conteúdo consumido pelo usuário. A origem do conteúdo pode ser definida de acordo com a arquitetura da politica de entrega. Há estratégias, tradicionais (DEHGHAN et al., 2017) e cooperativas (PU et al., 2018; LI et al., 2017). Por outro lado, a associação de usuário, está fortemente ligada ao destino do conteúdo, ou seja, onde o usuário deve se conectar para consumir o conteúdo. Em geral, as estratégias costumam basear se na qualidade do canal e recursos de espectro de rádio frequência (HARUTYUNYAN; BRADAI; RIGGIO, 2018).

2.4.1.3 Correlação Entre Inserção de Conteúdo e Roteamento de Requisições

Dentre os problemas descritos nas Subseções 2.4.1.1 e 2.4.1.2, a inserção de conteúdo e o roteamento de requisições possuem forte correlação. Consequentemente, é possível que a

localização do conteúdo impacte diretamente no roteamento de requisições, ou que o roteamento de requisições impacte no posicionamento do conteúdo, embora, a fase de roteamento de requisições seja sustentada pelas decisões da fase de inserção de conteúdo.

Para elucidar essa questão, é possível considerar o seguinte cenário: se um conteúdo é popular, ou seja, possui uma demanda considerável de solicitações de um grupo de usuários, esse conteúdo é capaz de desencadear uma possível sobrecarga sobre um único enlace, no qual impacta diretamente nas decisões do roteamento de requisições e, possivelmente, pode influenciar na latência. Ainda, se o conteúdo for posicionado em um *cache* no qual seu enlace apresenta sinais de sobrecarga, decorrentes de outros conteúdos igualmente populares ou outros tráfegos que compartilham o enlace, certamente, acarreta em impacto na escolha do caminho. Por outro lado, se roteamento pode buscar um caminho, no qual tem menor latência, consequentemente a fase de inserção deve buscar posicionar o conteúdo onde o enlace não apresente sinais de um possível congestionamento na rede (SONG et al., 2021).

2.5 CONTROLE DE CONGESTIONAMENTO DO TCP VEGAS

Os enlaces compartilhados por múltiplos usuários e, consequentemente, diversos protocolos de comunicação, estão sujeitos ao fenômeno de congestionamento. O fenômeno ocorre sempre que a carga oferecida para ser trafegada é superior à capacidade disponível no enlace (mesmo considerando os recursos de *buffer*) (KUROSE; ROSS, 2014). Embora possua uma definição conceitual simples, a identificação da ocorrência de congestionamentos em enlaces é uma tarefa árdua, sobretudo, quando executada diretamente na periferia da rede, sem auxílio dos dispositivos do núcleo. Em suma, os dispositivos finais devem inferir a capacidade do enlace caracterizado como sobrecarregado utilizando apenas informações aproximadas. O envio de dados além do suportado pelo enlace causará congestionamento, enquanto o envio inferior à capacidade do enlace pode resultar na sua subutilização (BRAKMO; PETERSON, 1995).

Para amenizar o impacto sobre as aplicações, alguns algoritmos para controle de congestionamento (ou CC) existentes utilizam mecanismos para prever a capacidade do enlace (a vazão útil) (BRAKMO; PETERSON, 1995) (CARDWELL et al., 2017) (LANGLEY et al., 2017). Especificamente no TCP Vegas, um dos percursores da proposta de predição, é a identificação da janela de congestionamento, que é dada a partir da comparação entre a vazão atual da rede e a vazão esperada. A vazão esperada e a vazão atual são dadas a partir das seguintes Equações 1 e 2, respectivamente.

$$Vazao_{esperada} = \frac{Tamanho_{Janela}}{RTT_{base}} \tag{1}$$

$$Vazao_{atual} = \frac{Tamanho_{Janela}}{RTT_{atual}} \tag{2}$$

Nesse sentido, a vazão atual é obtida através da divisão entre o tamanho atual da janela de congestionamento (*Congestion Window (CWND)*) e o RTT atual e, a vazão esperada é dada pela divisão do tamanho atual da janela de congestionamento e o menor RTT observado na rede, ou seja, o menor RTT possível atingido pela rede quando não há congestionamento presente. Desse modo, se a vazão atual for menor que a vazão esperada é necessário que a janela de congestionamento seja reduzida, caso contrário a janela aumenta. Tal mecanismo pode ser usado basear uma política de *cache* com o princípio de orientação à rede.

2.6 COMPLEXIDADE COMPUTACIONAL DE PROBLEMAS DE *CACHE*

Os problemas de fluxos de redes podem ser resumidos brevemente, como uma forma de transportar entidades entre uma origem e um destino do modo mais eficiente possível. Esses problemas possuem aplicações tais como: sistemas de comunicação, sistemas hidráulicos, sistema mecânicos, circuitos integrados de computador, sistemas de transporte e planejamento gerencial (AHUJA; MAGNANTI; ORLIN, 1993). Além disso, podem ser representados através de modelos de otimização de natureza matemática, que não são iguais a realidade, no entanto, são suficientemente similares de modo que seja possível obter conclusões a partir de sua análise (GOLDBARG; LUNA, 2005). Embora, tais problemas possuam soluções matemáticas, não são computacionalmente triviais. Sobretudo, problemas de fluxos de redes estão sujeitos às múltiplas restrições e variáveis e, além disso, possuem uma quantidade substancialmente numerosa de combinações possíveis, até encontrar a melhor solução para o problema, no qual aumentam no sentindo em que os elementos da rede aumentam (AHUJA; MAGNANTI; ORLIN, 1993).

As topologias de redes podem ser representadas e abstraídas através de uma estrutura de grafos, objetos matemáticos úteis para representar sistemas físicos (AHUJA; MAGNANTI; ORLIN, 1993). Um grafo pode ser definido como estrutura de abstração, ou diagrama, no qual pode ser usado para traduzir elementos e suas conexões, que podem ser denotados matematicamente a partir de um conjunto de vértices que representam os elementos e um conjunto de arestas que representam a relação ou conexão entre tais elementos (BONDY; MURTY, 1976).

Nesse sentido, o problema de roteamento de requisições juntamente ao problema de inserção de conteúdo pode ser uma aplicação intrínseca de um problema de fluxos em redes, no qual será formalizado no Capítulo 4 e, especificamente, pode ser baseado no *Multi-Commodity Flow Problem (MCFP)*, o qual é um problema combinatorial, e portanto, pertence a classe de problemas NP-Completo e não pode ser resolvido em tempo polinomial (EVEN; ITAI; SHAMIR, 1975 apud KARP, 1975). O MCFP consiste em um problema de transmissão de elementos distintos com origem e destinos diferentes. Entretanto, tais elementos compartilham as mesmas restrições e capacidade de uma única rede, com o objetivo de otimizar seu custo geral (AHUJA; MAGNANTI; ORLIN, 1993).

Sobretudo, é possível obter a solução ótima a partir métodos exatos ou modelos, tal qual a

ILP, no qual busca otimizar o problema, ou seja, ou maximizar, ou minimizar uma função a partir de variáveis e restrições bem estabelecidas para o problema (WOLSEY; NEMHAUSER, 1988). Nesse sentido, 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 depender de 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 de uma hipótese antes de investir em uma análise experimental de um problema (META, 2021).

2.7 TRABALHOS RELACIONADOS

Diante dos esforços para impulsionar o futuro da evolução de redes móveis, técnicas de *cache* de conteúdo têm sido uma abordagem promissora para reduzir a latência e o tráfego nos enlaces de BHs. Assim, há um crescente número de trabalhos direcionados para desenvolver políticas de *cache*. Em geral, abordagens de *cache* proativo, probabilístico, baseado em redes sociais e, que utilizam técnicas de aprendizado de máquina. No entanto, os trabalhos citados têm seu foco direcionado especificamente para o problema de roteamento de requisições de usuários e técnicas cooperativas que buscam compartilhar o armazenamento do sistema de *cache* distribuído ao longo da RAN.

2.7.1 Políticas de *Cache* Não Cooperativo

O trabalho de Shanmugam et al. (2013) é precursor ao abordar o problema de inserção de *cache* de conteúdo na borda de redes móveis. Sobretudo, com pretensão de implantar *cache* em *femtocells* (os detalhes arquiteturais foram apresentados na Seção 2.2.1) com restrições de capacidade de armazenamento, topologia da rede e distribuição de popularidade do conteúdo com objetivo de minimizar a latência. Além disso, em suas contribuições, os autores demonstram que o problema é NP-Completo e formulam um modelo ótimo, através de ILP, para *cache* codificado e, um algoritmo "guloso"de baixa complexidade para o problema na versão de *cache* não codificado e demonstram a garantia de aproximação da solução ótima.

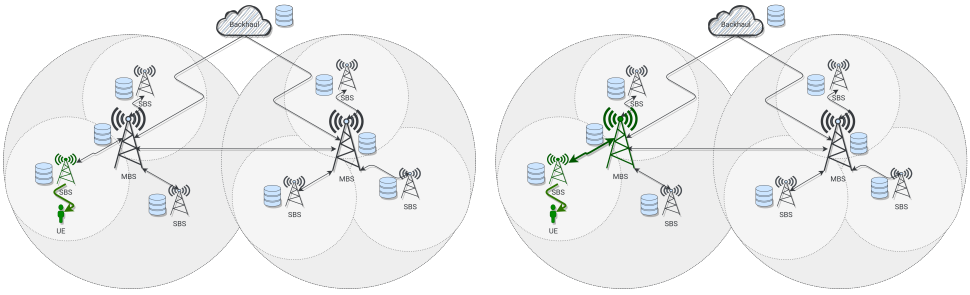
Da mesma forma, Dehghan et al. (2017) formulam a otimização do problema de inserção de conteúdo e roteamento de requisições de usuários de forma conjunta para otimizar latência, no qual consideram restrições de capacidade de armazenamento. Em suma, tal abordagem consiste em obter o conteúdo solicitado diretamente na BS em que usuário está conectado, quando o conteúdo estiver armazenado. Caso contrário, o conteúdo pode ser obtido diretamente do servidor remoto ou pode ser armazenado na BS em que o usuário está conectado. Além disso, duas variantes do problema são formuladas: um modelo sem percepção de congestionamento e um modelo sensível ao congestionamento, no qual a formulação ocorre através de uma fila que representa a carga no servidor remoto. Por fim, os autores demonstram que o problema é

NP-completo em ambos os casos e desenvolvem um algoritmo "ganancioso"com garantia de proximidade da solução ótima e, um segundo algoritmo de baixa complexidade.

Por sua vez, o trabalho de Harutyunyan, Bradai e Riggio (2018) formula através de ILP o problema de inserção de conteúdo e associação de usuário conjuntamente com o objetivo de realizar o balanceamento da utilização de recursos de rádio e utilização do enlace de BH. Complementarmente, recursos de rádio, podem ser entendidos de forma simplificada como meio de conexão entre BSs e usuários. Assim, a otimização atende a restrições de capacidade de armazenamento, capacidade do enlace de BH e capacidade de recursos de rádio. Além da formulação ótima, os autores desenvolvem uma heurística e incorporam a mobilidade no seu modelo de associação de usuário, ou seja, o distanciamento ou deslocamento geográfico do usuário acarreta condições de qualidade inferiores à adequada. Mesmo que o conteúdo solicitado esteja armazenado na BS, o usuário pode se conectar a outra BS na qual esteja mais próxima, entretanto, não possui o conteúdo em *cache* e, portanto, pode ser mais vantajoso consumir o conteúdo diretamente do servidor remoto.

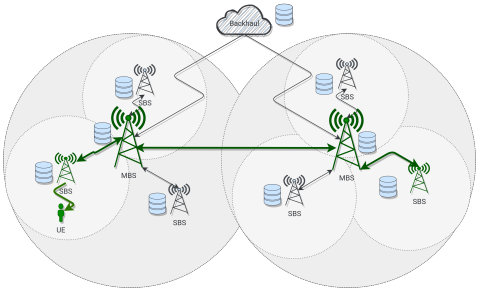
Em resumo, embora os usuários possam se conectar a múltiplos *caches*, caso o conteúdo solicitado não seja encontrado diretamente no *cache* associado ao usuário, o conteúdo é servido através do servidor remoto. Ou seja, não há cooperação entre os *caches* para buscar o conteúdo solicitado como pode ser observado na Figura 7a.

Figura 7 – Diferenças entre abordagens da literatura.



(a) Sem Cooperação - Único Salto.

(b) Cooperação - Único Salto.



(c) Cooperação - Multissaltos.

Fonte: Elaborado pela autora (2022).

Evidentemente, essas políticas estão em desvantagem frente a políticas que utilizam artefatos de colaboração entre múltiplas entidades da rede para amplificar a capacidade de armazenamento. Portanto, é possível que a taxa de *cache hit* seja maior e, consequentemente,

que a utilização do enlace BH seja menor (LI et al., 2017).

2.7.2 Políticas de *Cache* Cooperativo

A política de *cache* proposta por Pu et al. (2018) apresenta uma formulação conjunta entre o problema de inserção de conteúdo, roteamento de requisições e alocação de recursos para uma arquitetura C-RAN. Assim, objetivando a redução de custos, os quais 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. Portanto, VMs e BBU são hospedadas em *central offices* no qual colaboram entre si para disponibilizar ao usuário o conteúdo em *cache*, no entanto, a cooperação é de um único salto (Figura 7b). Além da formulação ótima através de *Mixed-Integer Linear Programming (MILP)*, os autores desenvolvem um algoritmo de aproximação, com desempenho computacional superior ao algoritmo proposto por Dehghan et al. (2017).

No mesmo sentido, uma formulação ótima do problema de inserção e entrega de conteúdo cooperativo para HCN é proposta por Jiang, Feng e Qin (2017) com objetivo de reduzir a latência, considerando restrições de topologia da rede, probabilidade de solicitação do conteúdo, capacidade da armazenamento e capacidade de enlace. Ressalva-se que entrega de conteúdo, nesse contexto, pode ser entendida como o problema de roteamento de requisições, de forma que, nesse trabalho os autores consideram escolha do caminho baseando-se na origem do conteúdo (Como elucidado na Subseção 2.4.1.2). A cooperação ocorre entre SBSs e dispositivos finais por meio de comunicação *D2D* de acordo com o vizinhos diretamente conectados ao usuário.

Uma política de inserção e roteamento de requisições por meio de cooperação hierárquica é proposta por Li et al. (2017) com objetivo de maximizar a taxa de *cache hit*. De tal modo, a formulação possui restrições de capacidade da armazenamento, capacidade do enlace e topologia da rede. Os autores desenvolvem uma formulação ótima por meio de *Binary Integer Linear Programming (BILP)* e uma heurística. Por se tratar de uma abordagem hierárquica, é possível que o usuário se conecte em múltiplos níveis hierárquicos, enquanto o conteúdo pode ser recuperado através de vizinhos no mesmo nível ou níveis superiores da hierarquia.

Uma arquitetura de colaboração entre SBSs e dispositivos finais através de D2D com o objetivo de minimizar a latência é projetada em Sheng et al. (2016). Tal arquitetura é multicamadas, ou seja, pode ser considerada como uma arquitetura hierárquica. De tal modo, inicialmente o usuário faz a busca do conteúdo localmente e posteriormente em outros dispositivos conectados a partir de uma busca multissaltos entre os dispositivos que não estejam conectados diretamente (ou através da SBS). Além disso, os resultados são obtidos através de experimentação, ao contrário da maioria dos trabalhos selecionados.

Do mesmo modo o trabalho de Song et al. (2021) formula o problema de inserção de conteúdo e roteamento de requisições de forma conjunta, como foco em redes não confiáveis. Essa estratégia tem como objetivo minimizar a latência e, além disso, a formulação ILP está sujeita a restrições de capacidade de armazenamento e capacidade de enlace. Além de adotar um

algoritmo de natureza "gananciosa", a proposta considera a cooperação multissaltos, ilustrada na Figura 7c, assim, SBSs podem se comportar como *cache* ou como relé, no qual podem encaminhar o conteúdo até que alcance o solicitante.

Finalmente, o mais recente trabalho de Xie et al. (2022) propôs a inserção de conteúdo e roteamento de requisições com objetivo de minimizar a latência. Além disso, os autores formularam o problema através de *Linear Programming (LP)*, com duas versões, uma com restrição de capacidade de largura de banda e outra sem essa restrição. Por fim, analisaram a complexidade do problema, bem como implementaram um algoritmo executável em tempo polinomial. A política proposta considera o roteamento de requisições para todos os caminhos possíveis, ou seja, é multissaltos.

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 como ilustra a Figura 7.

2.7.3 Discussão de Trabalhos Relacionados

Os trabalhos mencionados nesta Seção 2.7, em geral, são direcionados ao problema de roteamento de requisições. Dentre eles, em sua maioria, partem do princípio de cooperação a partir de uma BS vizinha, ou seja, aquela em que o usuário está associado ou, consideram a cooperação, que pode ser de apenas um único salto, hierárquica ou multissaltos. Abordagens de único salto ou hierárquicas podem limitar o espaço de busca entre os demais *caches* presentes na RAN e, consequentemente, pode reduzir a proporção de *cache hit*, como ilustrado na Figura 7. É possível observar que os trabalhos, normalmente, consideram o roteamento de requisições a partir de um único salto, isto é, único caminho até a *cache* ou até a origem do conteúdo de uma perspectiva E2E. Sobretudo, ignoram as possíveis variações da rede, nas quais podem ocorrer nos caminhos intermediários ou, em outras palavras, não consideram a dinamicidade do estado da rede (DEHGHAN et al., 2017; PU et al., 2018; JIANG; FENG; QIN, 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; XIE et al., 2022).

Além do mais, tais trabalhos usam restrições de largura de banda, portanto, ignoram a existência de outros possíveis fluxos existentes no enlace, isto é, a largura de banda é utilizada como medida da capacidade real do enlace. Tal abordagem pode fornecer uma informação ambígua a respeito do estado do enlace e sua capacidade real (BRAKMO; PETERSON, 1995). Essa premissa será detalhada na Seção 3.1.1.

Ademais, os trabalhos direcionam-se para a minimização da latência, uma métrica de relevância que converge com KPI da rede 5G. Destaca-se que a maioria desses trabalhos abordam o problema de inserção de conteúdo e roteamento de requisições de forma conjunta. Finalmente, tais trabalhos apresentam suas propostas através de métodos exatos(LP), devido a complexidade do problema, no qual é discutida e avaliada em alguns trabalhos (SHANMUGAM et al., 2013;

DEHGHAN et al., 2017; PU et al., 2018; JIANG; FENG; QIN, 2017; SONG et al., 2021; LI et al., 2017). Nesse sentido, são propostos algoritmos "gananciosos", heurísticas ou algoritmos de aproximação. O Quadro 1 sumariza tais comparações entre os trabalhos.

Em síntese, são elencados os itens em que a política de *cache* cooperativa orientada à rede proposta se diferencia dos trabalhos mencionados:

- **Cooperação**: A cooperação é considerada neste trabalho, ou seja, o conteúdo pode ser posicionado em *cache* e a requisição pode ser direcionada para uma BS em que o usuário não está conectado.

- **Cooperação Multissaltos**: Não somente a cooperação, mas a cooperação entre todas as BS que sejam capazes de suportar o *cache* de conteúdo. Assim, os recursos de armazenamento de toda RAN são compartilhados. É possível que o usuário consuma um conteúdo em *cache* posicionado em outro extremo da RAN sem que seja necessário consumir o conteúdo original através do BH. Além disso, o roteamento de requisições multissaltos é realizado de perspectiva global, ou seja, realiza busca completa na RAN sem hierarquia estabelecida.

- **Estimativa de vazão**: Ao contrário de abordagens que restringem a capacidade do enlace e, de tal modo, ignoram outros fluxos que trafegam através da rede, os quais podem causar possíveis congestionamentos. A política de *cache* proposta estima a capacidade real do enlace com base no algoritmo de CC do TCP Vegas.

- **Orientação à rede**: Em geral os trabalhos mencionados não consideram a variabilidade de sobrecarga ou, congestionamento na rede. Para isso, a política de *cache* proposta prioriza o roteamento de requisições para o consumo através de *cache* e, além disso, realiza o balanceamento de carga e pondera se o consumo através de *cache* de fato reduz a latência, caso contrário o conteúdo original pode ser trafegado através do BH.

- **Mobilidade**: A mobilidade é considerada na política de *cache* proposta e, portanto, baseia-se no RTT, o qual pode aumentar ou reduzir em função da distância entre usuário e BS.

- **Múltiplos caminhos**: A política de *cache* proposta considera múltiplos caminhos para o roteamento de requisições, isso é, pode escolher dentre todos os caminhos possíveis dentro da RAN, aquele que resulta em menor latência.

Quadro 1 – Trabalhos Relacionados.

Trabalho	**Cache Cooperativo**	**Multissaltos**	**Sem Restrição de Capacidade do Enlace**	**Múltiplos Caminhos**	**Orientado à Rede**	**Mobilidade**	**Programação Matemática**	**Problema**	**Objetivos**
(SHANMUGAM et al., 2013)	×	×	✓	×	×	×	✓	Inserção de Conteúdo	Minimizar Latência
(DEHGHAN et al., 2017)	×	×	✓	×	parcialmente	×	✓	Inserção de Conteúdo e Roteamento de Requisições	Minimizar Latência
(HARUTYUNYAN; BRADAI; RIGGIO, 2018)	×	×	×	×	×	✓	✓	Inserção de Conteúdo e Associação de Usuário	Balancear o uso de FH e BH
(PU et al., 2018)	✓	×	×	×	×	×	✓	Alocação de Recursos VM, inserção de conteúdo e roteamento de requisições	Minimizar custos (entre eles: latência)
(JIANG; FENG; QIN, 2017)	✓	×	×	×	×	×	✓	Inserção de conteúdo e entrega de conteúdo	Minimizar Latência
(LI et al., 2017)	✓	×	×	×	×	×	✓	Inserção de Conteúdo e Roteamento de Requisições	Maximizar o *cache hit*
(SHENG et al., 2016)	✓	✓	✓	✓	×	×	×	Arquitetura para Roteamento de Requisições	Minimizar Latência
(SONG et al., 2021)	✓	✓	×	✓	×	×	✓	Inserção de Conteúdo e Roteamento de Requisições	Minimizar Latência
(XIE et al., 2022)	✓	✓	ambos	✓	×	×	✓	Inserção de Conteúdo e Roteamento de Requisições	Minimizar Latência
Cooperativa orientada à rede	✓	✓	✓	✓	✓	✓	✓	Inserção de Conteúdo e Roteamento de Requisições	Minimizar Latência

Fonte: Elaborado pela autora (2022).

2.8 CONSIDERAÇÕES PARCIAIS

Esse capítulo apresentou a fundamentação teórica necessária para compreensão do cenário de *cache* em redes móveis. Assim, descreveu brevemente conceitos fundamentais de *cache*, bem como o atual cenário das redes móveis, com ênfase na atual rede 5G. Destacou tecnologias que, se combinadas, proporcionam a evolução das redes móveis. Dentre elas, MEC no qual consiste na implantação de *cache* na borda da rede. Apontou algumas vantagens e desvantagens de arquitetura existentes. Resumiu algumas das tecnologias e alternativas que vem sendo exploradas em projetos de *cache* em redes móveis e, apresentou os desafios existentes, tais como a inserção de conteúdo e roteamento de requisições. Além disso, elucidou a importância e correlação entre os dois problemas. Da mesma forma, a apresentou o algoritmo de CC presente no TCP *Vegas*. Mencionou a necessidade e as vantagens da utilização de formulações matemática e simulações numéricas. Destacou trabalhos relacionados, com ênfase no roteamento de requisições. Finalmente, destacou a relevância de cooperação multissaltos, assim como o planejamento de um roteamento de requisições que considere a dinamicidade rede, tais como sintomas de sobrecarga no enlace e os impactos da mobilidade.

3 FORMULAÇÃO DO PROBLEMA

Inicialmente, este capítulo apresenta a especificação, formulação do problema e o detalhamento das principais premissas da política de *cache* proposta. Em seguida, descreve a infraestrutura na qual a política de *cache* se aplica, dinâmica de configuração e reconfiguração de requisições. Posteriormente, descreve a representação do problema (e seus requisitos) usando estruturas de grafos. Por fim, um cenário de exemplificação é apresentado, demonstrando um potencial caso de uso.

3.1 OBJETIVOS DA POLÍTICA DE *CACHE*

O presente trabalho objetiva especificar, desenvolver e analisar um modelo para uma proposta de política de *cache* de conteúdo cooperativa e orientada à rede com finalidade de reduzir a latência. Em suma, a proposta atua unindo os problemas de inserção de conteúdo e roteamento de requisições com restrições de capacidade de armazenamento e QoS. Os problemas são abordados como partes correlacionadas, em virtude de que a localização do conteúdo influencia diretamente no roteamento da solicitação do usuário, assim como o contrário também é verdadeiro, assim como descrito na Seção 2.4.1.3.

A localização dos dados pode ser afetada pelo padrão da distribuição de popularidade do conteúdo, ou seja, se houver demanda crescente de um mesmo conteúdo direcionada para um subconjunto de usuários, esse comportamento pode gerar sobrecarrega sobre um único enlace (ou caminho). Ainda, se o conteúdo é posicionado em um *cache* no qual o enlace de acesso sofre sobrecarga, ocorre um impacto direto na escolha do caminho, pois a política buscará alternativas para realizar o roteamento (SONG et al., 2021). Por fim, considera-se que os conteúdos possam ser dispostos em quaisquer elementos da RAN (armazenados em *caches*) ou podem ser acessados diretamente de uma nuvem computacional através do enlace de BH. Porém, há restrições que devem ser consideradas para efetivação da política de *cache*:

- **Capacidade de armazenamento dos elementos.** A capacidade de armazenamento refere a quanto e quais conteúdos as MBSs e SBSs podem armazenar em *cache*, dado uma quantidade de *bytes*.
- **Requisitos de QoS** A QoS se refere a capacidade do enlace, ou seja, a possível presença ou ausência de congestionamento ou sobrecarga na rede. A perda de pacotes, pode ser resultante do saturamento de *buffers* dos roteadores ao longo de uma rede, logo é indicativo de congestionamento (KUROSE; ROSS, 2014). Além disso, a QoS pode ser mensurada através do tempo de execução, latência de computação e comunicação (VARGHESE et al., 2021).

Diante do objetivo e das restrições apresentadas, a política de *cache* será baseada em alguns elementos, propriedades e tecnologias consolidadas de redes de computadores, que são

detalhados na sequência.

3.1.1 Orientação à rede

Conforme detalhado na Seção 2.5, o TCP Vegas possui mecanismos de CC baseado na estimativa de *Retransmission Time Out (RTO)* para inferir a ausência ou presença de possíveis congestionamentos na rede. Especificamente, o RTO é obtido com base na variação dos valores de RTT. O presente trabalho utiliza tal conhecimento (cálculo de RTO e RTT) para implementar a política de *cache*, atuando na camada de aplicação. De tal modo, esse mecanismo de CC pode ser usado para inferir o congestionamento da rede móvel, e dessa forma possibilita que a política de *cache* se comporte dinamicamente de acordo com o estado da rede. Sendo assim, a restrição do estado do enlace da rede, segue princípios semelhantes aos do algoritmo de CC do TCP *Vegas*. Portanto, 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 desencadeado pelas retransmissões motivadas pela perda de pacotes pode ser interpretado como um sintoma de congestionamento presente no enlace.

A restrição de QoS considera como referencial a vazão esperada definida pela aplicação. Em outras palavras, um determinado serviço (*e.g.*, VoD) tem a necessidade de que uma vazão mínima seja garantida para que seu serviço seja entregue dentro da qualidade esperada. Essa informação definida pelo SLA deve ser disponibilizada pelo provedor de conteúdo.

Assim, a partir da divisão entre o *buffer*, no qual consiste em uma fração do conteúdo que será enviado ao usuário de acordo com a sua demanda e, o RTT atual do enlace é possível obter a vazão atual (Equação 3). Ressalta-se que o conceito de *buffer*, nesse contexto, consiste na menor parte dos dados que podem ser enviados pela aplicação e, além disso, suas características (*e.g.* tamanho) dependem inteiramente da aplicação.

$$Vazao_{atual}(.) = \frac{Buffer}{RTT_{atual}} \tag{3}$$

De forma semelhante ao TCP *Vegas*, a vazão atual e a vazão esperada, fornecida pela aplicação, serão comparadas. Dito de outro modo, a determinação do estado de um enlace está fortemente conectado aos parâmetros de QoS fornecidos pela aplicação, ou pelo provedor de serviço. A restrição de estado do enlace está atrelada ao problema de roteamento de requisições, no qual determina em quais dos *caches* o usuário consumirá o conteúdo dentro da RAN ou ainda, se irá consumir o conteúdo a partir de uma nuvem computacional, no qual pode se mostrar mais vantajoso do ponto de vista do estado enlace. Ressalta-se que o roteamento de requisições determina a origem do conteúdo assim como os caminhos intermediários entre o usuário e o *cache*, ou entre o usuário e a nuvem computacional.

Diferentemente de outros trabalhos presentes na literatura, 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; BRADAI; RIGGIO, 2018; SHANMUGAM et al., 2013; DEHGHAN et al., 2017; LI et al., 2017; JIANG; FENG; QIN, 2017; PU et al., 2018; SONG et al., 2021). Considerando que 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.

3.1.2 Dinâmica do RTT

A mobilidade é um fator de relevância, sendo um dos principais requisitos da rede 5G. Ou seja, a rede deve ter a competência de suportar dispositivos estacionários ou em movimento tais como pedestres e veículos, com velocidades de aproximadamente até 10 km/h e 120 km/h, respectivamente (ITU, 2017). No entanto, a 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; XU; ANSARI, 2005).

Portanto, é possível deduzir que o RTT está relacionado a distância entre o usuário e a SBS, ou seja, o RTT é linearmente proporcional a distância entre o usuário e a SBS. Partindo da premissa que o meio de conexão é sem fio, esse comportamento é representado pela Equação 4.

$$RTT_{atual}(.) = RTT_{minimo} \times \left(1 + \frac{Distancia_{atual}}{Distancia_{maxima}}\right) \tag{4}$$

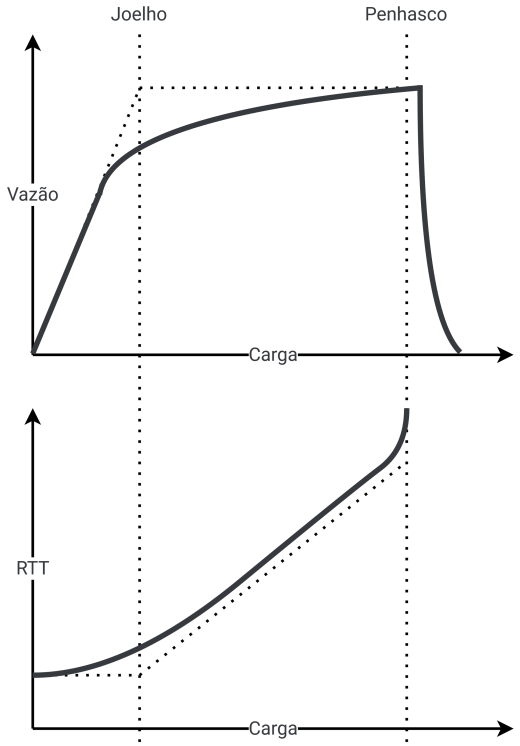
Por fim, é importante destacar que dado a tendência de aumento da tráfego no enlace, o RTT cresce exponencialmente proporcional a carga no enlace (CHIU; JAIN, 1989). Para exemplificar, considera-se que o RTT inicial do enlace é 1 milissegundo e o total trafegado no enlace é multiplicado por uma escala exponencial (*e.g.* $1, 2, 4, 8...n$). Se o tráfego total do enlace é 100 Mb/s, o RTT é multiplicado por 1, se o tráfego no enlace é 200 Mb/s, o RTT é multiplicado por 2, e assim sucessivamente.

3.1.3 Cooperação Multissaltos

A política de *cache* proposta funciona a partir de uma perspectiva global, ou seja, a proposta permite realizar uma busca por conteúdo que seja multissaltos (roteamento de requisições). 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) e (LI et al., 2017).

Dentre os trabalhos encontrados na literatura, há trabalhos que não consideram a cooperação (SHANMUGAM et al., 2013; DEHGHAN et al., 2017; HARUTYUNYAN; BRADAI;

Figura 8 – Desempenho da rede em função da carga.



Fonte: (CHIU; JAIN, 1989).

RIGGIO, 2018) ou, consideraram a busca apenas em BS vizinhas (PU et al., 2018; JIANG; FENG; QIN, 2017; LI et al., 2017). Tal colaboração proposta permite ampliar a capacidade da rede de modo que evite replicações, posicionando os conteúdos sob uma visão global da rede. Portanto, a formulação conjunta dos problemas de inserção e roteamento de requisições permite balancear a carga na rede de acordo com a restrições de capacidade de armazenamento e estado do enlace. Em outras palavras, quanto maior a capacidade de armazenamento presente nas BS, menor é chance do conteúdo ser solicitado à nuvem computacional, sendo assim, evita a sobrecarga do enlace de BH (LI et al., 2017).

3.1.4 Princípios da Política de *Cache*

É possível afirmar que quanto maior a vazão, mais adequado é o enlace de acordo com os requisitos de QoS definidos pelo provedor de serviço. Assim, é deduzível que quanto menor o RTT, maior a vazão (CHIU; JAIN, 1989). Conforme a quantidade de dados enviados através do enlace aumenta, a quantidade de dados presentes no enlace se aproxima da vazão máxima suportada, ilustrada como o *joelho* na Figura 8. Posteriormente, o crescimento da vazão tende a diminuir, até que passe a regredir, esse comportamento pode ser visualizado como *penhasco* na Figura 8. Consequentemente, conforme a vazão entra em declínio, o RTT apresenta um rápido crescimento, conforme é possível observar na Figura 8.

Partindo da seguinte premissa: quanto maior a vazão e menor o RTT, melhor será o desempenho do serviço. É possível deduzir que a escolha de caminhos com maior vazão, além de evitar caminhos possivelmente sobrecarregados, permiti evitar o surgimento ou agravamento de sobrecarga ou congestionamento. Portanto, propicia o benefício a QoS, relacionada, neste caso, diretamente à latência. Assim, 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 multissaltos, no qual, considera as implicações dos estados de enlaces intermediários.

A política de *cache* pretende otimizar as métricas na perspectiva do MNO, ou seja, da rede. Nesse sentindo, a política de *cache* é baseada nos mecanismo de CC do TCP Vegas para estimar a vazão atual do enlace. Assim, a política de *cache* busca o enlace com maior vazão, esperando o desencadeamento de um efeito de espalhamento e distribuição dos fluxos de dados, enquanto que por outro lado busca concentrar o *cache* de conteúdo com menos réplicas possíveis para otimizar a capacidade de armazenamento. Tal consolidação permite que mais conteúdos diferentes sejam armazenados em *cache*. Espera-se que o comportamento da política de *cache* seja preferir caminhos dentro da RAN, e sempre que possível busque trazer 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.

Essencialmente, resumem-se os princípios da política de *cache*:

- O roteamento multissaltos de requisições é baseado na dinamicidade da rede, no qual considera o RTT fortemente correlacionado a carga do enlace óptico e variação do RTT nos enlaces sem fio baseado na degradação da qualidade da conexão ocasionada pela mobilidade do usuário e sobrecarga do enlace.

- A política de *cache* busca enlaces com maior vazão, e portanto, menor RTT para evitar possíveis rotas congestionadas ou desencadear novo congestionamento, consequentemente, miniminizando a latência.

- A busca cooperativa de uma perspectiva global proporciona que mais requisições sejam alocadas em *cache* e, portanto, evita tráfego através do BH, no qual, em geral, possui maior latência e acarreta em custo ao MNO.

- A política busca otimizar e balancear o equilíbrio do uso de recursos de armazenamento, assim como busca respeitar os requisitos de QoS(vazão mínima esperada) definidos pelo servidor de aplicação de acordo com o SLA, distribuindo o tráfego na rede de forma que a latência seja reduzida.

- Estima a vazão real baseada no algoritmo de CC do TCP Vegas. Assim, a política não requer conhecimento prévio sobre a topologia física, permitindo o compartilhamento da infraestrutura com outras aplicações e protocolos.

- Pode beneficiar ambas as perspectivas: MNO, provedor de serviço e usuário. Tais benefícios são propiciados pela melhoras da QoS, redução do uso do BH.

3.2 INFRAESTRUTURA DA REDE

A infraestrutura de rede MEC permite o desenvolvimento de tecnologias que possam contribuir para a redução da latência e a utilização do enlace de BH (KEKKI et al., 2018). Assim, devido a presença de recursos computacionais na borda da rede móvel, é possível que as HCN possibilitem a utilização de tecnologias de *cache* de conteúdo em elementos heterogêneos distribuídos por toda RAN, tais como: MBS, SBS ou dispositivos finais (ANDREWS, 2013). A utilização de *cache* em dispositivos finais compreende alguns desafios tais como: autonomia energética, capacidade de armazenamento restrita (KAMEL; HAMOUDA; YOUSSEF, 2016), deslocamento acentuado (WU et al., 2021) e a dependência do usuário (JIANG; FENG; QIN, 2017). Tais desafios podem dificultar a implementação de políticas de *cache*.

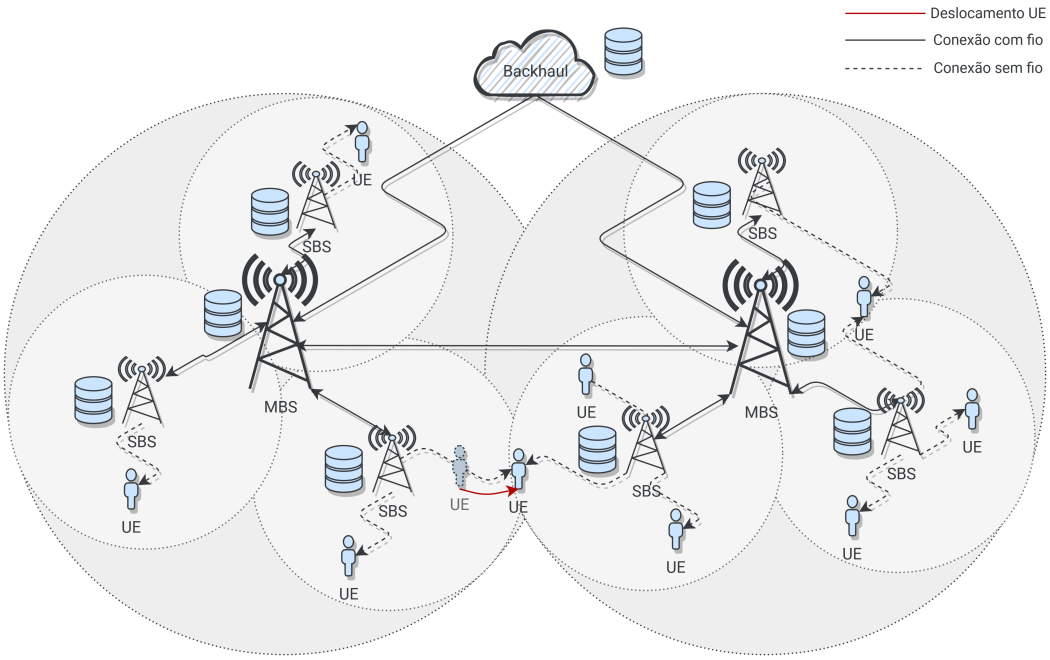
Arquiteturas C-RAN e SBS são consideradas alternativas, possuindo características contrastantes, como detalhando na Subseção 2.2.1. Em outras palavras, devido ao espalhamento das SBS, é possível que sobrecargas em enlaces centralizados sejam evitadas assim como a redução da latência, motivada pela maior proximidade ao usuário (TRAN et al., 2017). Resumidamente, essas tecnologias podem ser combinadas para que a rede seja capaz de alcançar melhor desempenho.

Portanto, a arquitetura assumida nesse trabalho agrega tecnologias de uma HCN, com *caches* implantados em MBS e SBS sob o paradigma MEC, representada através da Figura 9. Assume-se que a RAN, toda a infraestrutura entre o usuário e núcleo da rede móvel, é composta por MBS diretamente conectadas ao núcleo da rede móvel, que por sua vez é responsável por conectar à rede móvel a Internet. 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 uma ou mais nuvens computacionais. Nesse sentido, o consumo do conteúdo diretamente de sua origem será denominado como consumo a partir de uma nuvem computacional.

Conforme demonstrado pela Figura 9, as MBSs podem estar conectadas entre si e as SBSs são conectadas ao núcleo da rede através das MBSs, conforme já definido em outros trabalhos na literatura especializada (LI et al., 2017; SHENG et al., 2016). Através dessa infraestrutura os usuários (UE) se conectam diretamente as SBSs para acessar a Internet ou consumir conteúdos que estejam armazenados em *caches* que são implantados nas MBSs e SBSs, que por sua vez podem ou não suportar determinado servidor de conteúdo ou simplesmente podem não ser capazes de armazenar o *cache* de conteúdo.

Ademais, é necessário que a arquitetura forneça um forma de acesso a camada de dados, que pode obtida através de Software Defined Networking (SDN) ou *Information-Centric*

Figura 9 – Arquitetura da Rede Móvel.



Fonte: Elaborado pela autora (2022).

Networking (ICN). Por fim, a infraestrutura alvo considera a mobilidade dos usuários. É possível que o usuário se desloque ao longo da RAN e alterne sua conexão entre as SBSs e, além disso, os usuários podem se conectar a múltiplas SBSs devido a densidade da rede, conforme exemplificado na Figura 9 que ilustra o deslocamento de um usuário entre SBSs.

3.3 REQUISIÇÕES DE USUÁRIOS

Uma requisição, no contexto da proposta da política de *cache*, é definida como uma solicitação de conteúdo, realizada por um usuário. Em outras palavras, como visto na Figura 9, o *User Equipment (UE)* conectado à SBS solicita conteúdo que pode estar armazenado em uma SBS ou MBS (elementos considerados como similares na arquitetura definida, diferenciados apenas pela capacidade total de armazenamento) que pode fornecer o conteúdo, portanto, a requisição pode ser alocada para servir o conteúdo a partir de um *cache* ou através da origem. Resumidamente, a requisição possui uma origem (conteúdo) e um destino (usuário) e, é composta por três fases: o momento da chegada da requisição, a realocação da requisição e o momento em que a requisição é desalocada.

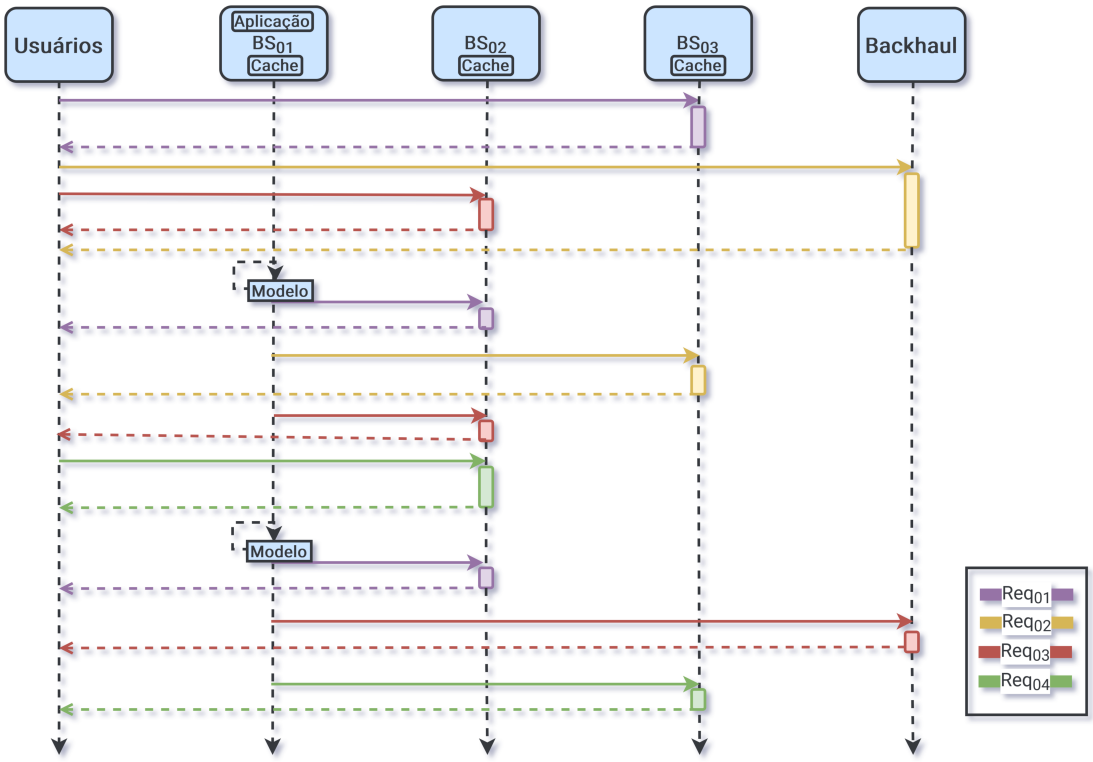
Em um primeiro momento, a solicitação do usuário é atendida e o usuário passa a consumir o conteúdo de um *cache* localizado na RAN ou a partir da nuvem computacional localizada fora da RAN através de um caminho que atenda a exigências de QoS definidas pelo provedor de serviço. No entanto, com a chegada de novas solicitações, é possível que as condições do enlace se degradem, devido a uma alta demanda de um conteúdo específico que

pode concentrar as requisições em um ponto da rede ou, ainda, a degradação pode ser recorrente da mobilidade do usuário. Além disso, pode acarretar na migração (entre *caches* ou entre *cache* e nuvem computacional) desse conteúdo na RAN, de tal modo que, possa distribuir o tráfego.

Em suma, a abordagem dessa política de *cache* proposta é *online*, isto é, não é possível que a política antecipe quais requisições chegarão ou qual será o comportamento do usuário, suas preferências e seu deslocamento, consequentemente, espera-se que as requisições sejam realocadas. Finalmente, as requisições podem ser desalocadas opcionalmente, essa etapa é desencadeada pelo término do consumo determinado pelo usuário ou pela aplicação.

Destaca-se que a otimização do posicionamento do conteúdo, assim como o definição do roteamento de requisições acontece em eventos discretos de tempo configuráveis (que podem ser definidos pelo MNO) (opcionalmente, a política pode ser invocada para atuar a cada chegada de requisição). A Figura 10 exemplifica as três etapas de atuação da política proposta. Inicialmente ocorre a chegada das requisições Req_{01}, Req_{02} e Req_{03}. A política aloca as requisições de acordo com a disponibilidade em *cache* e, portanto, se não houver um *cache* disponível de um conteúdo especifico, o conteúdo original será disponibilizado através do BH.

Figura 10 – Fluxo de Requisições.



Fonte: Elaborado pela autora (2022).

Posteriormente, de acordo com a configuração de tempo realizada pelo MNO o modelo será executado e, consequentemente, otimiza novas requisições. Além disso, realiza migrações se necessário com o objetivo de reduzir a latência. De forma subsequente, uma nova requisição é

recebida Req_{04} e alocada. O modelo é executado novamente, a nova requisição Req_{04} é realocada assim como as demais requisições ativas (Req_{01}, Req_{03}). A Req_{02} é desalocada, devido ao encerramento da solicitação pelo usuário.

3.4 REPRESENTAÇÃO DA PROPOSTA COM GRAFOS

Diante dos fundamentos descritos na Seção 3.1 e da arquitetura da rede definida na Seção 3.2, a presente seção introduz a formalização e notação baseada em grafos.

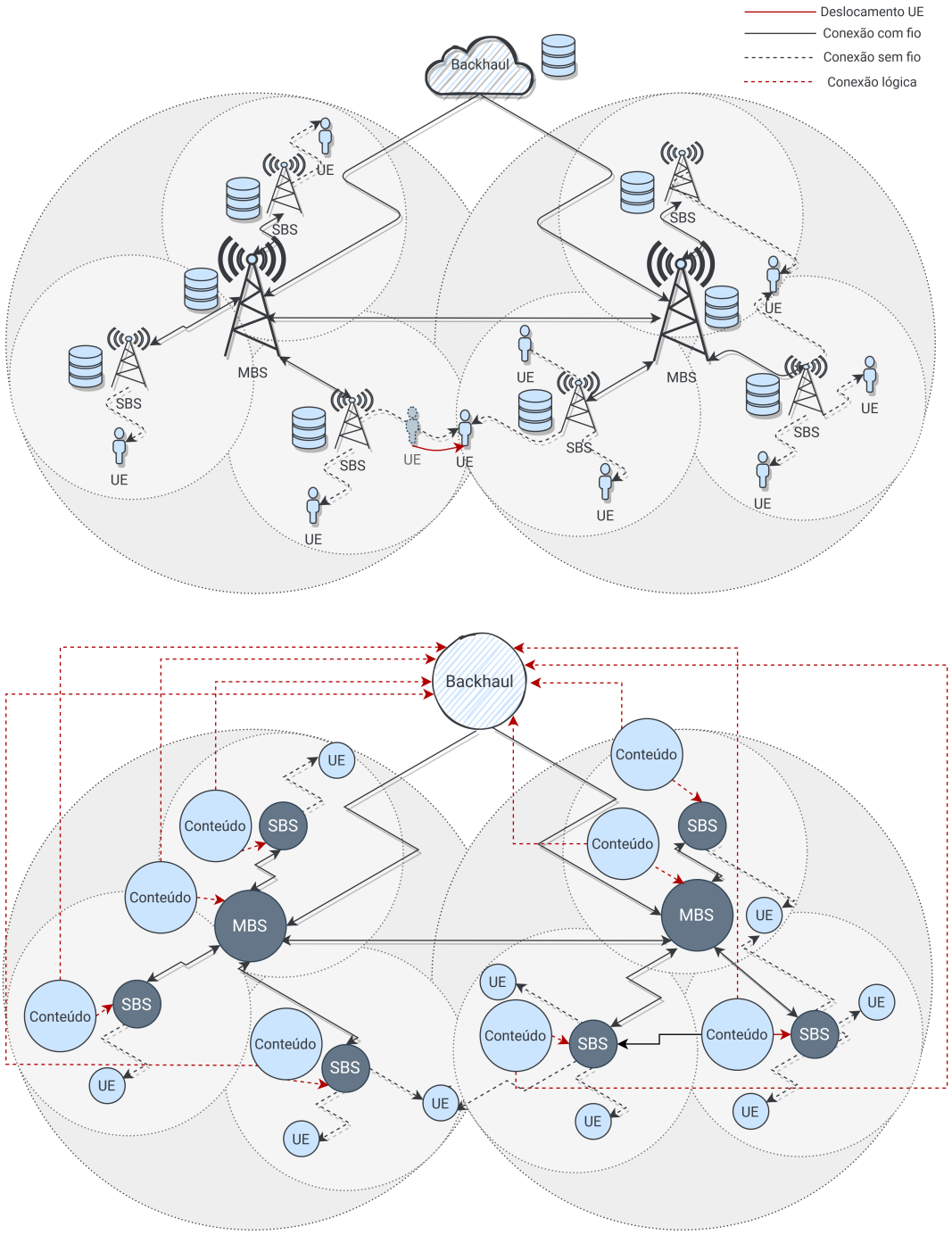
3.4.1 Infraestrutura *HCNs*

Dado um grafo $G(V,E)$, o conjunto V representa os vértices, composto por um ponto de acesso à nuvem (S), conteúdo (C) e a HCN constituída de MBS, SBS (denominados BS) e usuários UE. Ressalva-se que na abstração através de 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 armazenamento e a vazão mínima para garantir QoS, 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 representação da estrutura de grafos pode ser visualizada na Figura 11.

A conexão entre uma nuvem computacional $s \in S$ e, BS $i \in BS$, é dada 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 BS estão contidas no conjunto de arestas $e_{ij} \subseteq E$. É importante observar que as MBS podem ou não possuir conexões entre si. Essas conexões são dadas por arestas bidirecionais. 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 sentido das arestas, fica claro na Figura 11. Igualmente, as conexões sem fio entre SBS e UE, pertencem a $e_{iu} \subseteq E$, entretanto são definidas pela distância entre a UE e a SBS. Nesse caso, cada usuário $u \in UE$ possui as coordenadas (x,y) associadas, no qual podem variar a medida que a UE se desloca sobre o plano, ilustrado na Figura 11.

A função $dis(.) \in \mathscr{R}+$ retorna a distância euclidiana entre UE e BS em plano cartesiano. Para cada $i \in BS$, existe $D \in \mathscr{R}+$, que denota o raio de cobertura de uma BS. Portanto, se $dis(x_i,y_i,x_u,y_u) \leq D$ significa que a $u \in UE$ está dentro do raio de cobertura de $i \in BS$. A aresta de conexão entre SBS e UE possui sentindo unidirecional a partir da SBS. A conexão entre $k \in C$ e $i \in BS$ segue o mesmo raciocínio, ou seja, as arestas pertencem a $e_{ij} \subseteq E$. Dado $\gamma_{ik} \in \{0,1\}$, o valor 1 representa onde é possível que o conteúdo seja armazenado e 0, caso contrário. Os valores de γ podem ser definidos a partir de algoritmos que probabilidade que podem estimar o comportamento do usuário (BASTUG; BENNIS; DEBBAH, 2014). Além disso, o γ pode ser

Figura 11 – Representação do problema usando grafos.



Fonte: Elaborado pela autora (2022).

definido de acordo com a infraestrutura disponibilizada, isto é, quais BS podem suportar o *cache* de uma determinada aplicação. É possível que uma BS não possa hospedar uma *cache*, devido a definições do MNO. Ainda, é possível que o γ seja definido de acordo com a disponibilização de recursos computacionais entre o MNO e provedor de conteúdo, definidos pelo SLA.

Além disso, a conexão é representada por uma aresta unidirecional a partir do conteúdo. A Figura 11 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 $k \in C$.

3.4.2 Orientação à Rede

Cada aresta $ij \in E$ possui um RTT associado, dado por rtt_{ij}, sendo que o RTT é o último observado. Em outras palavras, o tempo de envio do último pacote encaminhado e a confirmação de seu recebimento. Assim, é possível obter o RTO(vazão atual), conforme detalhado na Seção 3.1.1 e introduzido por (BRAKMO; PETERSON, 1995). 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 $i \in BS$. Todos os parâmetros mencionados estão sumarizados no Quadro 2.

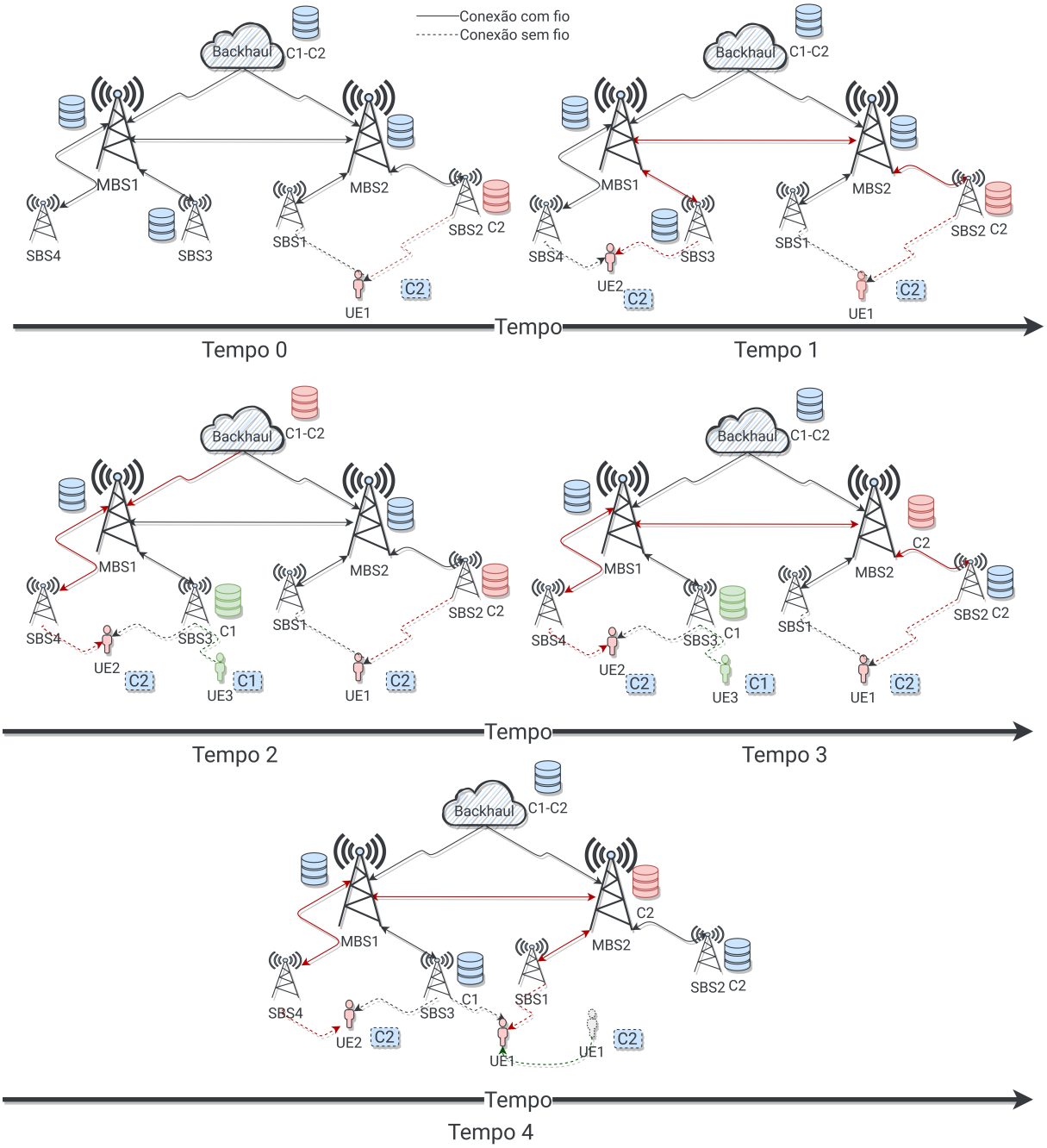
3.4.3 Cenário de Exemplificação

Para fins de exemplificação, um cenário será detalhado, no qual consiste na entrega de conteúdos de uma aplicação VoD. Considerando uma rede móvel, de arquitetura heterogênea, administrada por MNO, no qual o servidor de conteúdo dispõe de uma biblioteca de dois conteúdos, $C(C_1, C_2)$. A rede é composta por duas MBSs, cada qual com duas SBSs agregadas que fornecem infraestrutura de acesso à rede móvel para três usuários, $UE(UE_1, UE_2, UE_3)$. A flecha na Figura 12 ilustra o cenário e indica a passagem de eventos discretos.

Inicialmente, no tempo zero, tem se a primeira requisição de um usuário UE_1, no qual solicita o conteúdo C_2. O usuário está conectado à SBS_1 e SBS_2, no momento em que o usuário solicita o conteúdo, de acordo com as premissas da política de *cache*, o conteúdo é armazenado na SBS_2. Posteriormente, no tempo um, o usuário UE_2 associado às SBS_3 e SBS_4, solicita o mesmo conteúdo, C_2, no entanto, está associado à SBS_3, porém consome o conteúdo a partir da SBS_2. Destaca-se que a busca pelo conteúdo C_2 é realizada, uniformemente dentro de toda a RAN, ou seja, consiste na cooperação entre todas as BSs que permitem o posicionamento de *cache* de forma horizontal e, não apenas em BSs vizinhas, portanto, sem hierarquia definida.

No tempo dois, o usuário UE_3 que está associado à SBS_3, inicia uma solicitação do conteúdo C_1, e assim, o conteúdo é posicionado na SBS_3. De tal modo, outros tráfegos ou mesmo a chegada de mais usuários a SBS_3 sobrecarregam o meio de transmissão e, portanto, o usuário UE_1 passa a consumir o conteúdo através do BH, no qual o caminho entre o usuário UE_1 e o conteúdo C_2 tem uma latência menor atribuída. Ressalva-se que a política de *cache* busca otimizar a latência, ou seja, o caminho E2E, entretanto, analisa cada um dos saltos de forma

Figura 12 – Cenário de Exemplificação.



Fonte: Elaborado pela autora (2022).

Quadro 2 – Notação Matemática.

Parâmetro	**Descrição**
$G(V,E)$	Um grafo representa a RAN (elementos HCN, usuários) e a nuvem computacional.
$V = BS \cup UE \cup C \cup S$	Representa o conjunto de todos os vértices, união dos conjuntos de usuários (UE), *caches* (C), BS (BS) e e a nuvem computacional (S).
$E \subseteq (V \times V)$	Representa a conexão física ou lógica dos enlaces de comunicação entre os componentes da rede.
$c_k^s \in \mathscr{N}+$	Representa os requisitos de armazenamento definidos por SLA, dado um conteúdo $k \in C$.
$c_k^b \in \mathscr{N}+$	Representa o tamanho do *buffer* definido por SLA, dado um conteúdo $k \in C$.
$c_k^{thp} \in \mathscr{N}+$	Representa os requisitos de a vazão mínima definida por SLA, dado um conteúdo $k \in C$.
$bs_i^s \in \mathscr{N}+$	Denota a capacidade total de armazenamento, dada uma BS $i \in BS$.
$\gamma_{ik} \in \{0,1\}$	Indica que um conteúdo $k \in C$ pode ser armazenado, dado uma $i \in BS$.
$r_{uk} \in \{0,1\}$	Indica que um usuário $u \in UE$ solicita um conteúdo $k \in C$.
$rtt_{ij} \in \mathscr{R}+$	Representa o último RTT observado, dado um enlace $ij \in E$.
$thp_{ijk}^{cur} = \frac{c_k^b}{rtt_{ij}} \in \mathscr{N}+$	Representa a última vazão medida, dado um enlace $ij \in E$ baseado em um conteúdo $k \in C$.

Fonte: Elaborado pela autora (2022).

independente.

No tempo três, a política de *cache* redireciona o roteamento e o posicionamento do conteúdo C_2 para a MBS_2, assim o usuário UE_1 e UE_2 passam a consumir o conteúdo sob o novo posicionamento, decorrente da sobrecarga e balanço de vantagens entre os usuários, dado que múltiplos usuários consomem o conteúdo. Ressalta-se que realocar o conteúdo, pode ser mais vantajoso da perspectiva do MNO, que evita tráfego no BH e não replica o conteúdo C_2, portanto, considera o aproveitamento da capacidade de armazenamento do servidores *cache*. Sendo assim, posiciona o conteúdo em um novo local considerando o estado do caminho entre o conteúdo de o usuário. Assim, múltiplos usuários podem se beneficiar de um melhor QoS.

No tempo 4, o usuário UE_3 encerra sua solicitação, ou seja, para de consumir o conteúdo. No mesmo tempo o usuário UE_2 se desloca geograficamente, degradando as condições do meio de transmissão e, nesse sentido, a política de *cache* executa um novo roteamento, para um meio

de transmissão mais adequado, ou seja, o caminho em que o usuário possa usufruir da menor latência. Tal comportamento, consiste, na premissa de que a vazão é um forte indicativo do estado do enlace, a qual deriva-se em parte do RTT (CHIU; JAIN, 1989) (STALLINGS, 2015).

O cenário descrito demonstra as múltiplas possibilidades de migração, realocação de posicionamento do conteúdo, bem como, a realocação de caminhos originada, pelo condução conjunta dos problemas de inserção e roteamento de requisições. Sendo assim, tais escolhas são derivadas do estado do enlace que pode ser afetado pela sobrecarga, a qual pode ser gerada pela demanda de conteúdo com maior popularidade ou a partir dos ruídos gerados pela mobilidade do usuário e, ainda, não deixa de considerar a capacidade de armazenamento como uma unidade em toda RAN, evitando múltiplas réplicas de um mesmo conteúdo.

3.5 CONSIDERAÇÕES PARCIAIS

Esse capítulo descreveu os detalhes do comportamento das requisições dos usuários. Finalmente, delineou e detalhou os princípios da politica de *cache*, tal qual premissas e hipótese dedutiva para uma política de *cache* orientada à rede, orientada pelo algoritmo de CC presente no TCP, especificamente o TCP Vegas. Nesse sentido, a hipótese é dada pelo espalhamento do fluxo na rede, com o intuito de não agravar o congestionamento, partindo da premissa de que quanto menor o RTT maior a vazão e vice e versa. Explicou o impacto da mobilidade no modelo e como ela é encarada, bem como a lógica a respeito do parâmetro RTT.

Destacou, portanto, a principal característica de relevância da política de *cache*, sendo elas: o roteamento de requisições multissaltos no qual considera a QoS em caminhos intermediários e, a cooperação de perspectiva global na qual possui visão total da rede. Assim, a política prioriza o interesse do MNO, de modo que, evita o acesso ao BH otimizando uso de armazenamento e, ao mesmo tempo respeita o SLA definido pelo provedor de conteúdo, no qual estabelece QoS para entregar o conteúdo. Finalmente, detalhou às características da infraestrutura e requisitos estabelecidos pelo escopo do trabalho. Apresentou a abstração do problema através de um estrutura de grafos bem como a notação matemática do problema. Finalmente, ilustrou cenários de exemplo para elucidação do funcionamento da política de *cache*.

4 MODELO PARA UMA POLÍTICA DE *CACHE* COOPERATIVA E ORIENTADA À REDE

O presente capítulo desenvolve uma formulação de ILP para o problema descrito no Capítulo 3. A programação matemática é amplamente adotada em trabalhos semelhantes presentes na literatura (DEHGHAN et al., 2017; JIANG; FENG; QIN, 2017; PU et al., 2018; LI et al., 2017), bem como em soluções comerciais (META, 2021), justificando seu potencial de aplicação para análise de problemas complexos, sem a necessidade de configuração de infraestruturas experimentais específicas.

Conforme os parâmetros apresentados em sua notação formal na Seção 3.4, esse capítulo descreve a formulação de modelo para uma política de *cache* através de ILP. Nesse sentido, 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. Portanto, serão descritas as variáveis de decisão, função objetivo e as restrições do modelo de ILP nas seguintes Seções 4.1 4.2 4.3, respectivamente. A Seção 4.4 detalha a formulação matemática para os modelos existentes na literatura, tais modelo serão utilizados para análise de desempenho do modelo para a política de *cache* proposta.

4.1 VARIÁVEIS DE DECISÃO

A formulação ILP da política de *cache* é baseada no MCFP, um problema NP-Completo (AHUJA; MAGNANTI; ORLIN, 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} é definido de acordo a Equação 5, ou seja y_{ik} é obtido em função de x_{ijk}.

$$y_{ik} = x_{kik}; \forall i \in BS, \forall k \in C \tag{5}$$

4.2 FUNÇÃO OBJETIVO

A função objetivo é descrita pela Equação 6. Para que seja possível atender aos requisitos de latência da rede 5G, o primeiro termo da função objetivo, $\left[\sum_{u \in UE} \sum_{i \in BS} \sum_{k \in C} \frac{(bs_i^s - c_k^s) \times y_{ik} \times r_{uk}}{bs_i^s \times \gamma_{ik} + \delta}\right]$, corresponde à perspectiva de inserção de conteúdo. Sendo assim, há uma tendência em posicionar a maior quantidade possível de conteúdo em *cache*, dito de outro modo, certamente, 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 $\delta \to 0$ 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. Para isso,

considera possíveis congestionamentos presentes na rede decorrentes do aumentos do RTT. Portanto, de acordo com o descrito nas Seções 3.1.1 e 3.1.4 a respeito da relação entre a presença de congestionamento e o aumento do RTT e, consequentemente, a redução da vazão. A partir dessa premissa, é possível inferir que a distribuição dos fluxos direcionada para enlaces que apresentem uma vazão superior seja a melhor escolha (STALLINGS, 2015). Sendo assim, enlaces com vazões menores no qual podem indicar possíveis congestionamentos são evitados, desse modo, o problema não é agravado. Em outras palavras, a equação busca a maior vazão em função de c_k^{thp}. Esse comportamento é dado pelo segundo termo da função objetivo, $\left[\sum_{u\in UE}\sum_{ij\in E}\sum_{k\in C}\frac{c_k^{thp}}{thp_{ijk}^{cur}}\times x_{ijk}\times r_{uk}\right]$, dada pela Equação 6, no qual pode ser atribuído ao problema de roteamento de requisições.

$$min \sum_{u\in UE}\sum_{i\in BS}\sum_{k\in C}\frac{(bs_i^s - c_k^s)\times y_{ik}\times r_{uk}}{bs_i^s\times\gamma_{ik}+\delta} + \sum_{u\in UE}\sum_{ij\in E}\sum_{k\in C}\frac{c_k^{thp}}{thp_{ijk}^{cur}}\times x_{ijk}\times r_{uk} \tag{6}$$

Analisando a Equação 6 sob a perspectiva matemática, o primeiro termo é responsável pelo posicionamento, isto é, o problema de inserção de conteúdo, e sempre prioriza posicionar o conteúdo em *cache* ou o mais próximo possível do usuário. Esse comportamento ocorre porque BSs mais próximas aos usuários normalmente possuem capacidade de armazenamento menor. O contrário ocorre na nuvem computacional, responsável por prover o conteúdo original. Portanto, o resultado da divisão $\frac{(bs_i^s-c_k^s)\times y_{ik}\times r_{uk}}{bs_i^s\times\gamma_{ik}+\delta}$ sempre será menor para BSs com menor capacidade de armazenamento disponível. Isso reflete na função objetivo, que busca minimizar o valor resultante. No entanto, o modelo poderá reorganizar o posicionamento em eventos de tempo para que mais requisições possíveis sejam atendidas em *cache*, assim o uso da capacidade de armazenamento da rede é otimizada como um todo. Por outro lado, o segundo termo da equação prioriza as condições da rede para determinar qual caminho é adequado para trafegar com o conteúdo e consequentemente pode determinar onde o conteúdo é armazenado. Isso é refletido pelo relacionamento entre as variáveis x_{ijk} e y_{ik}. Mesmo que o primeiro termo defina o posicionamento, ele é influenciado pelo segundo termo, que pode determinar a decisão em razão das condições da rede. Nesse contexto, as condições da rede são determinadas pela vazão e RTT, como explicado na Seção 3.1.1. Assim, a vazão atual do enlace ($\frac{c_k^b}{rtt_{ij}}$) reduz se o RTT aumentar em decorrência do distanciamento do usuário ou da sobrecarga dos enlaces. A divisão que coordena o segundo termo, $\frac{c_k^{thp}}{thp_{ijk}^{cur}}$, tende a ser menor para vazões maiores e maior para vazões menores. Assim, espera-se que o modelo escolha os enlaces que resultem em valores menores, ou seja, enlaces com maiores vazões e menores RTT, esse comportamento evita que caminhos que estejam congestionados sejam selecionados. Por fim, a medida que mais requisições são alocadas, é possível que o modelo distribua os caminhos para não sobrecarregar um enlace. Se um enlace estiver sobrecarregado, é possível que o modelo defina que o caminho através do BH é mais vantajoso do ponto de vista da rede, ou seja, possui uma latência total menor.

Destaca-se que as variáveis de decisão estão fortemente relacionadas, ou seja, a inserção de conteúdo é definida em função do roteamento de requisições, assim como o roteamento de requisições é definido em função da inserção de conteúdo. O primeiro termo busca a maximização do uso das *caches*, bem como evita o uso do enlace de BH, respeitando a restrição de capacidade de armazenamento, dada pela Equação 10. Caso contrário, se a restrição for excedida, a política de *cache* tende a buscar o conteúdo diretamente na nuvem computacional.

O segundo termo busca alternar os caminhos intermediários entre o usuário e o conteúdo, de forma que respeite a restrição de QoS definida pelo SLA, dada pela Inequação 11. Com objetivo de buscar os enlaces com maior vazão, com o princípio de evitar enlaces com menor vazão possivelmente congestionados, devido a sobrecarga na rede ou a mobilidade do usuários. Assim, é provável que se um dado caminho até o *cache* presente na RAN estiver congestionado, o melhor caminho escolhido seja até a nuvem computacional, trafegando pelo enlace de BH. Portanto, é plausível que a função objetivo realize balanceamento de carga, a fim atingir o valor mínimo de acordo com os requisitos da otimização.

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 7-9 garantem a conservação do fluxo, isto é, garantem que exista apenas um único caminho entre um usuário $u \in U$ e um conteúdo $k \in C$.

$$\sum_{i \in BS} x_{jik} \times r_{uk} - \sum_{i \in BS} x_{ijk} \times r_{uk} = 0; \forall j \in BS, \forall k \in C, \forall u \in UE \tag{7}$$

$$\sum_{i \in BS} x_{kik} \times r_{uk} - \sum_{i \in BS} x_{ikk} \times r_{uk} = 1; \forall k \in C, \forall u \in UE \tag{8}$$

$$\sum_{i \in BS} x_{uik} \times r_{uk} - \sum_{i \in BS} x_{iuk} \times r_{uk} = -1; \forall k \in C, \forall u \in UE \tag{9}$$

A restrição dada pela Inequação 10, garante que a capacidade de armazenamento de cada BS seja respeitada. Ressalta-se que se não existir solução sob o domínio desta restrição, a política de *cache* segue o comportamento padrão e busca o conteúdo diretamente na origem. Enquanto, a Equação 11 formula a restrição de QoS definida pelo SLA, em outras palavras, garante que os caminhos definidos respeitem a vazão mínima necessária para servir o conteúdo para um aplicação. Por fim a Equação 12 garante que o conteúdo será transmitido apenas por uma única origem (*cache* ou nuvem computacional) a cada requisição,consequentemente, $[\sum_{u \in UE} \sum_{k \in C} \sum_{i \in BS} y_{ik} * r_{uk} = 1]$.

$$\sum_{u \in UE} \sum_{k \in C} c_k^s \times y_{ik} \times r_{uk} \leq b_i^s; \forall i \in BS \tag{10}$$

$$(x_{ijk} \times r_{uk}) \times c_k^{thp} \leq thp_{ijk}^{cur} \times (x_{ijk} \times r_{uk}); \forall ij \in E, \forall k \in C, \forall u \in UE \tag{11}$$

$$\sum_{i \in BS} y_{ik} \times r_{uk} = 1; \forall k \in C, \forall u \in UE \tag{12}$$

4.4 MODELOS DE COMPARAÇÃO

De acordo com os trabalhos discutidos na Seção 2.7, dentre as abordagens propostas para políticas de *cache* existentes na literatura, foram elencadas as seguintes abordagens, com o objetivo de analisar o desempenho do modelo para a política de *cache* orientada à rede:

- **Não Cooperativo - Único Salto:** Tal abordagem considera apenas um único salto, isto é, não implementa a cooperação entre as BSs presentes na RAN.
- **Cooperação Multissaltos:** Tal abordagem considera multissaltos e permite a cooperação entre as BSs, realizando a busca pelo conteúdo em todas as *caches* presentes na RAN.

Para efeito de comparação, os conceitos fundamentais dessas abordagens foram generalizados e implementados a partir de adaptações ao modelo para política de *cache* cooperativa orientada à rede. Nas Subseções 4.4.1 e 4.4.2 serão descritos os modelos para a abordagem de único salto e para abordagem multissaltos, respectivamente.

4.4.1 Não Cooperativo - Único Salto

A abordagem não cooperativa, presente na literatura, será tomada como linha de base para evidenciar o desempenho do modelo para política de *cache* orientada à rede (SHANMUGAM et al., 2013; DEHGHAN et al., 2017; HARUTYUNYAN; BRADAI; RIGGIO, 2018; PU et al., 2018; JIANG; FENG; QIN, 2017; LI et al., 2017). Nesse sentido, é necessário uma nova restrição (elaborada na Equação 13) para limitar o posicionamento. Essa restrição garante que o caminho tenha apenas um único salto entre o usuário e a BS a qual o usuário está conectado diretamente, e um salto lógico que se refere a aresta entre o conteúdo e a BS onde o conteúdo está alocado, ou a aresta que conecta diretamente o usuário ao BH, totalizando dois saltos. Além disso, a função objetivo é representada pela Equação 14, na qual o primeiro termo é exatamente igual ao primeiro termo da Equação 6 relacionado ao modelo para política de *cache* orientada à rede.

Ressalta-se que o modelo único salto usado como linha de base, em geral segue as mesma premissas do modelo para a política de *cache* cooperativa orientada à rede. Portanto, realiza otimização das requisições já alocadas a partir de novas chegadas e busca a otimização baseada em todas a requisições, ou seja, um visão global da rede. Para a aresta entre a origem e

o dispositivo final (*i.e.*, nuvem computacional), o RTT inicial é 10 ms, isto é 10 vezes maior que o RTT inicial usado para as requisições atendidas em cache (LYU et al., 2021).

$$y_{ik} = \sum_{j \in BS} \sum_{i \in BS} \sum_{k \in C} x_{ijk} \times r_{uk} \leq 2 \tag{13}$$

$$min \sum_{u \in UE} \sum_{i \in BS} \sum_{k \in C} \frac{(bs_i^s - c_k^s) \times y_{ik} \times r_{uk}}{bs_i^s \times \gamma_{ik} + \delta} + \sum_{u \in UE} \sum_{ij \in E} \sum_{k \in C} \times x_{ijk} \times r_{uk} \tag{14}$$

4.4.2 Cooperação Multissaltos

A cooperação multissaltos é apresentada por trabalhos na literatura (KHREISHAH; CHAKARESKI; GHARAIBEH, 2016; SHENG et al., 2016). No entanto, tais abordagens não são orientadas à rede, isto é, não são capazes de estimar a vazão atual da rede e além disso, não consideram a mobilidade do usuário. Para fim de comparação, a modelagem segue as mesmas premissas do modelo para a política de *cache* cooperativa orientada à rede, entretanto, não considera à rede em sua otimização. A função objetivo desenvolvida para o cenário multissaltos é representada pela Equação 14, não contendo a razão que determina a orientação à rede e mobilidade. Assim como a linha de base de único salto, multissaltos considera a realocação de requisições, otimização do uso da capacidade total de armazenamento e a inserção de conteúdo o mais próximo possível do usuário (sem considerar a vazão da rede).

4.5 CONSIDERAÇÕES PARCIAIS

Esse capítulo descreveu a metodologia de análise do problema de inserção de conteúdo e roteamento de requisições a partir de ILP. Esse método possibilita a obtenção de uma solução para posterior desenvolvimento de um algoritmo de menor complexidade e solução aproximada. Além disso, permite a variabilidade nos parâmetros, ou seja, é possível obter resultados numéricos, sem que seja necessário a implantação de *hardware* específico. Consequentemente, permite que o cenário de avaliação seja maleável sem custos adicionais, como descrito na Seção 2.6.

Portanto, converge com o objetivo de desenvolvimento de uma formulação para reduzir a latência através de uma política de *cache* cooperativa e orientada à rede, partindo da otimização conjunta dos problemas de inserção de conteúdo e roteamento de requisições, nos quais respeitam às restrições de capacidade de armazenamento e de requisitos de QoS. Nesse sentido, foram descritas as variáveis de decisão, a função objetivo e seu comportamento, assim como às restrições da formulação de ILP que, por sua vez, é baseada no MCFP. Além disso, destacou o balanceamento de carga existente no modelo.

5 SIMULAÇÃO NUMÉRICA

Este capítulo apresenta resultados de simulações numéricas obtidas a partir do modelo proposto. A Seção 5.1 detalha os cenários analisados, métricas e parâmetros utilizados. A Seção 5.2 apresenta as discussões direcionadas aos cenários de análise do modelo para a política de *cache* orientada à rede. A Seção 5.3 avalia o comportamento e eficiência do modelo proposto para à política de *cache* orientada à rede comparado com abordagens existentes na literatura. Finalmente a Seção 5.4 apresenta as considerações do capítulo.

5.1 CENÁRIOS, MÉTRICAS E PARÂMETROS

Esta Seção descreve os detalhes da simulação numérica realizada. Inicialmente, a Subseção 5.1.1 especifica os cenários analisados, enquanto A Subseção 5.1.2 descreve as métricas definidas e suas contribuições para a análise do modelo proposto. A Subseção 5.1.3 detalha todos os parâmetros definidos para a simulação numérica.

5.1.1 Cenários Analisados

Inicialmente, foi realizada uma análise do comportamento da política de *cache* em relação a distribuição de popularidade dos conteúdos e uso do armazenamento (detalhada na Seção 5.2). Destaca-se que devido a natureza matemática do modelo (GOLDBARG; LUNA, 2005), os parâmetros escolhidos tem influência significativa no modelo. A análise de tais parâmetros teve como objetivo evidenciar e analisar o impacto do comportamento dos usuários, a capacidade total de armazenamento sobre o modelo para a política de *cache* frente às métricas definidas.

Após a conclusão do primeiro cenário, a política proposta foi comparada com duas formulações que representam propostas da literatura especializada, conforme apresentado na Seção 5.3. Nesse sentido, a primeira abordagem de comparação não é cooperativa e, portanto, considera apenas um único salto, ou seja, não há cooperação entre os componentes da rede. A segunda abordagem é considera a cooperação multissaltos, ou seja, uma requisição pode utilizar uma *cache* que está disponível através de um caminho na rede. No entanto, não é orientada á rede e não considera a mobilidade do usuário. Tais abordagens serão detalhadas nas Subseções 4.4.1 e 4.4.2.

5.1.2 Métricas de Análise

De acordo com a proposta descrita no Capítulo 3, a formulação matemática definida no Capítulo 4 e, essencialmente, conduzidas pelo objetivo geral descrito no Capítulo 1, as métricas definidas para orientar à análise da política de *cache* cooperativo orientado à rede são as seguintes:

- **Latência**: A latência foi definida como o tempo medido no intervalo entre a solicitação do conteúdo pelo usuário e seu posterior recebimento pelo servidor de *cache*, isto

é, entre origem e destino, ou ainda, BS ou nuvem computacional e UE. A latência de uma requisição é o intervalo de tempo *E2E*, que nas simulações foi baseada nos RTTs contabilizados nos enlaces pertencentes ao caminho realizado pela requisição. Especificamente, o RTT é o tempo medido entre o envio de um pacote e a sua confirmação, no presente caso, os RTTs são contabilizados para cada enlace.

- **Latência Média**: A latência média é obtida através da média das latências das requisições ativas por evento.
- **Latência Total da Rede**: A latência total é obtida através da soma das latências das requisições ativas por evento.
- **Proporção de *cache hit* e *cache miss***: O *cache hit* é a quantidade de requisições atendidas em *cache*. De tal modo, 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.
- **Uso Total do Armazenamento**: O uso de armazenamento considera a capacidade total do sistema de *cache* presente na RAN, ou seja, o somatório das capacidades de cada BS resulta na capacidade total.

Ao contabilizar a latência das requisições atendidas, é possível observar a eficácia do roteamento de requisições na rede. A busca por maiores vazões tende a evitar o surgimento de congestionamento na rede e, consequentemente, estabilizar a latência de forma otimizada. Esse comportamento é representado pelo segundo termo da Equação 6. A métrica de latência total da rede é usada para enfatizar o comportamento ao longo dos eventos discretos.

Por sua vez, a proporção de *cache hit* representa a capacidade da política de *cache* em alocar o maior número de requisições dentro da RAN. Portanto, expressa diretamente o primeiro termo da Equação 6, demonstrando a cooperação multissaltos entre BSs, bem como a priorização de elementos mais à borda da rede móvel. Do mesmo modo, a proporção de *cache miss* contrasta o diferencial entre as proporções. Tal métrica direciona-se para o problema de inserção de conteúdo. As duas métricas são comumente utilizadas na literatura (KABIR et al., 2020).

A análise combinada do uso total do armazenamento com *cache hit/miss* demonstra a relação entre a distribuição de popularidade e as proporções de requisições atendidas na HCN e na nuvem. Ainda, será possível observar que o *cache miss* não é desencadeado unicamente pela capacidade de armazenamento, mas também pela influência da rede.

5.1.3 Parâmetros da Simulação Numérica

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

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

9.1[2]. A simulação foi executada em três computadores: Intel Xeon E312XX com 64 GB com RAM, Intel I7-7700 com 28 GB RAM e, Intel Xeon 4214 com 128 GB RAM.

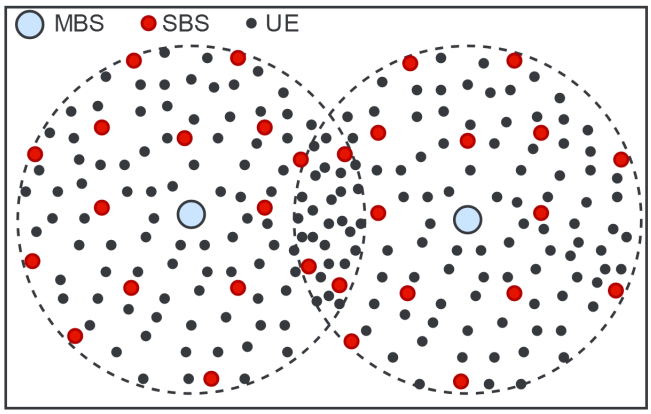
Os parâmetros selecionados foram, sempre que possível, guiados pela literatura especializada, resumidos na Tabela 1. Foram executados 100 eventos discretos (SHENG et al., 2016), ainda, a taxa de chegada de requisições (λ) segue distribuição de Poisson (DEHGHAN et al., 2017). Para compor a HCN, foram consideradas 2 MBSs, na qual cada uma possui 15 SBSs associadas (KHREISHAH; CHAKARESKI; GHARAIBEH, 2016) (JIANG; FENG; QIN, 2017). O raio de cobertura é dado em 70 metros, para SBS (SHANMUGAM et al., 2013), (SHENG et al., 2016), (JIANG; FENG; QIN, 2017) e a capacidade máxima de armazenamento das BS, informada pelo MNO, é 40% da capacidade total para armazenar toda a biblioteca de conteúdo. SBSs e MBSs, possuem 4 GB e 20 GB, respectivamente. Há 200 usuários conectados à rede móvel que somente podem se conectar às SBS, com uma limitação máxima de até 2 SBS ao mesmo tempo (KAMEL; HAMOUDA; YOUSSEF, 2016). A densidade da rede pode ser observada na Figura 13. Os UEs podem se mover de forma randômica (SHANMUGAM et al., 2013) (HARUTYUNYAN; BRADAI; RIGGIO, 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 milissegundo, tanto para meios com fio ou sem fio (ITU, 2017). Requisições permanecem alocadas durante 10 eventos discretos no decorrer da simulação (τ). A desalocação de uma requisição representa a desconexão do usuários, motivada pela conclusão do consumo da aplicação, ou por motivos de preferência ou interesse do usuário. Duas características da aplicação são informadas pelo provedor 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 (SENA et al., 2016), como tamanhos totais entre 2 GB, 4 GB e 8 GB. Ressalta-se que o tamanho do *buffer* depende da aplicação e de suas características, ou seja, podem ser redefinidos para aplicações distintas. Para o parâmetro γ é possível que todas as requisições possam ser alocadas potencialmente em qualquer umas das BSs presente da rede, dado as premissas definidas pelas abordagens (único salto, múltiplos saltos, vazão do enlace ou capacidade de armazenamento).

Além disso, a distribuição de popularidade desses conteúdos é estática e, sendo dada a partir da distribuição Zipf, no qual o parâmetro α é $0,8$. A distribuição de Zipf é usada para representar a distribuição de conteúdos e o comportamento dos usuários em relação as solicitações. Resumidamente, o padrão de solicitação de conteúdo dos usuários segue uma distribuição, definida por $P_k = 1/(k^{\alpha})$. Para valores de α menores, significa que o interesse dos usuários é diversificado entre os conteúdos presentes na biblioteca, enquanto que para valores de α maiores significa que a preferência dos usuários é concentrada nos mesmos conteúdos (BRESLAU et al., 1999).

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

Figura 13 – Densidade da Rede Simulada.



Fonte: Elaborado pela autora (2022).

Tabela 1 – Parâmetros.

Parâmetro	**Valor**
Quantidade de MBS	2
Quantidade de SBS por MBS	15
Tamanho da Biblioteca (C)	100 conteúdos
Quantidade de usuários (UE)	200
Raio de cobertura SBS	70m
Vazão mínima da aplicação (c_k^{thp})	100 Mbps
Capacidade da BS (bs_i^s)	40% da Biblioteca
Tamanho do conteúdo (c_k^s)	2 GB/4 GB/8 GB
Tamanho do *buffer* (c_k^b)	48 Mb
Mobilidade do usuário	10m
Popularidade: Distribuição Zipf(α)	0,8
Taxa de chegada de requisições (λ)	5
Duração de Requisição (τ)	10 eventos
RTT inicial	1ms

Fonte: Elaborado pela autora (2022).

5.2 ANÁLISE DA POLÍTICA ORIENTADA À REDE

Nesta seção são apresentados os resultados e as discussões pertinentes em relação da distribuição de popularidade, capacidade total de armazenamento e, comparação entre os modelos que implementados de acordo com abordagens existentes na literatura, e o modelo para política de *cache* cooperativa orientada à rede.

5.2.1 Distribuição de Popularidade

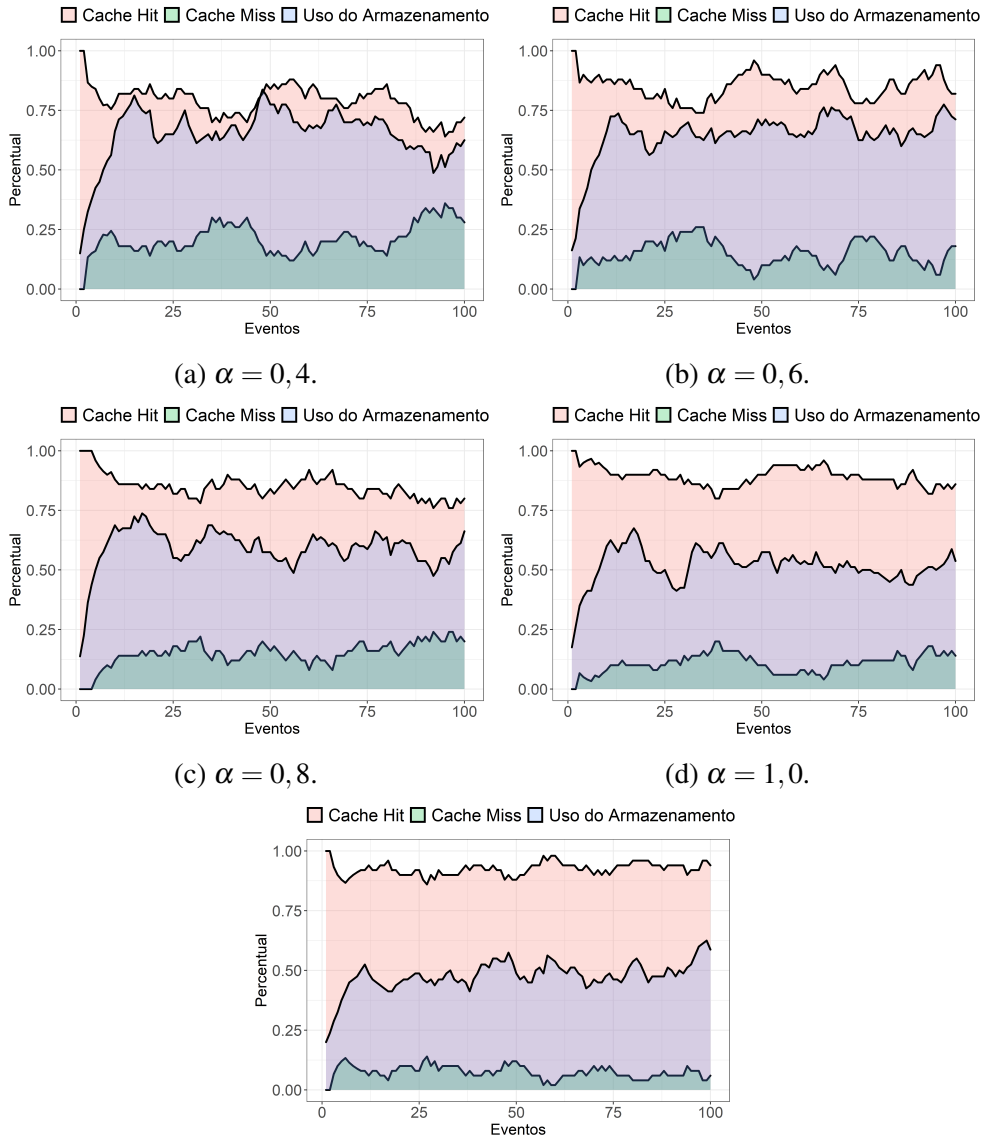
A distribuição de popularidade representa o comportamento dos usuários, isto é, padrões de preferência em relação ao conteúdo. A variação de distribuição de popularidade acarreta em menor variabilidade de popularidade de conteúdos quando o valor de α é maior, ou seja, as solicitações de conteúdo concentram-se em torno de uma pequena parte do conteúdo. Por sua vez, um valor de α menor distribui a popularidade dos conteúdos, desse modo, o padrão de preferência dos usuários é esparso. O parâmetro de distribuição de popularidade de conteúdo abrangeu uma variação no valor α para distribuição Zipf entre $0,4$ e $1,2$, em intervalos de $0,2$.

É possível observar na Figura 14, que 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 conteúdos 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 14a), o uso de armazenamento máximo foi $0,83$ e a proporção máxima de *cache miss* alcançou $0,36$, enquanto o *cache hit* obteve seu valor mínimo de $0,64$.

Em contraste, observa-se que para o valor de $\alpha = 1,2$ (Figura 14e), o máximo do uso total da capacidade de armazenamento foi $0,62$. Enquanto a proporção mínima de *cache hit* foi $0,86$, e $0,14$ para a proporção máxima de *cache miss*. A proporção de *cache hit* aumentou quando o valor de α aumentou, bem como a proporção de *cache miss* reduziu com o aumento do valor de α. Como observado nas proporções de *cache misses*, ainda que, existentes para o valor de $\alpha = 1,2$, a média foi 3 vezes menor em relação a $\alpha = 0,4$, ambos com desvios padrão $0,2$ e $0,06$, respectivamente. Assim, mais solicitações foram atendidas em *cache*, devido a concentração de popularidade dos conteúdos solicitados. Tal comportamento se intensificou à medida que o valor de α aumentou.

Dessa forma, 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 da capacidade 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.

Figura 14 – Cenário de Distribuição de Popularidade.



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

(c) $\alpha = 0,8$. (d) $\alpha = 1,0$.

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

Fonte: Elaborado pela autora (2022).

Finalmente, observa-se na Tabela 2 que a variação nos valores de α não desencadeou impactos significativos na latência. O valor do coeficiente de correlação de Pearson obtido foi $0,06$, indicativo de correlação inexistente. A Figura 15 demonstra graficamente a conformidade entre as diferentes variações de distribuição de popularidade dos conteúdos. Dado os valores divergentes no terceiro e último quartil, possivelmente tal variação é derivada da concentração de popularidade de conteúdo. Conforme o valor de α aumenta, significa que possivelmente um determinado conteúdo tem mais solicitações que outros. Nesse sentido, pode desencadear a sobrecarga de um enlace específico causada pelo crescimento exponencial do RTT em relação aos dados trafegados, como descrito na Subseção 3.1.2. Esse comportamento enfatiza que o modelo para a política de *cache* não criou réplicas e, consequentemente, otimiza o uso total da armazenamento. Ainda assim, houveram requisições que não foram atendidas em *cache*, ou

Tabela 2 – Latência Total em Função da Distribuição de Popularidade dos Conteúdos.

Popularidade	**Quartis**				**Percentil**
	1º	**2º**	**3º**	**4º**	**95º**
α : **0,4**	1,3ms	3ms	5,1ms	11,9ms	7ms
α : **0,6**	1,3ms	3ms	5ms	67ms	7ms
α : **0,8**	1,4ms	3ms	5ms	67,3ms	7ms
α : **1,0**	1,5ms	3,3ms	5,1ms	67,1ms	7ms
α : **1,2**	1,7ms	3,4ms	5ms	35,9ms	7ms

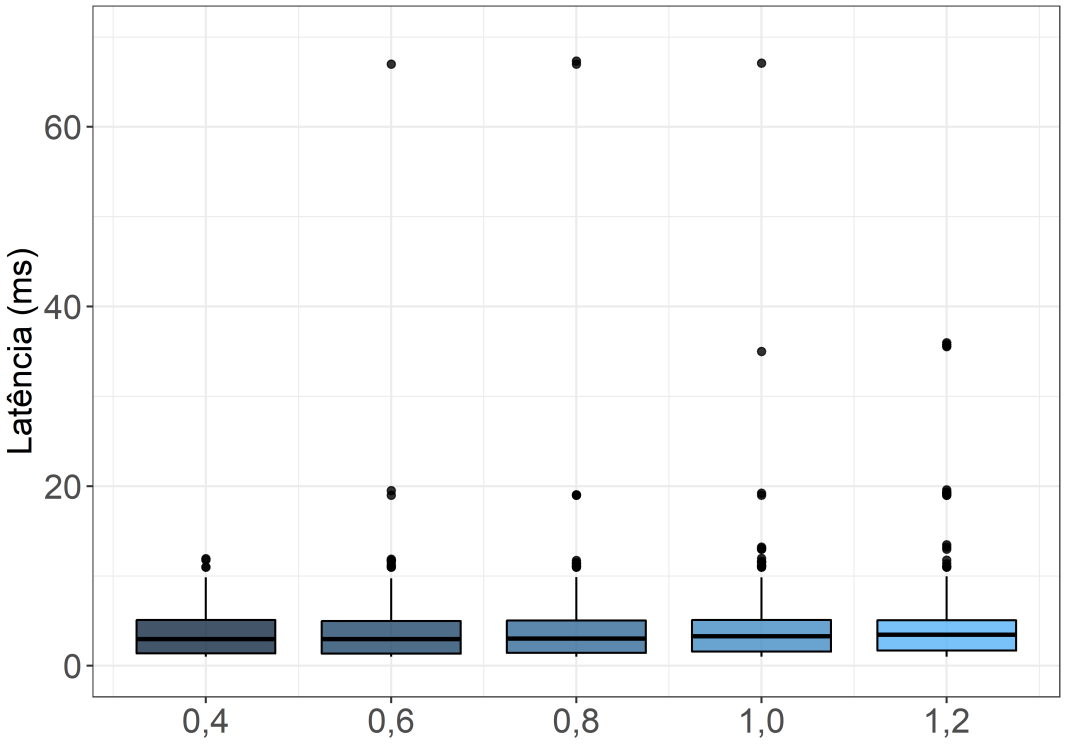
Fonte: Elaborado pela autora (2022).

seja, o caminho até a *cache* de conteúdo possivelmente está sobrecarregado. Assim, a política opta por escolher a transmissão através do BH, que é mais vantajosa na perspectiva do provedor de serviço e do usuário. Desse modo, a distribuição de latência não foi alterada, mesmo com variações de popularidade entre os conteúdos.

A política de *cache* busca otimizar a latência independentemente das variações de popularidade de conteúdo. Para 50% (segundo quartil) da amostra, a latência se manteve inferior a 4 milissegundos, portanto, dentro do limite determinado pelo ITU (2017). Embora no último quartil a latência se manteve inferior a $11,9$ milissegundos no melhor caso e $67,3$ milissegundos no pior caso. Para 95% da amostra se manteve inferior a 7 milissegundos para todos os casos.

Em suma, a partir da perspectiva do MNO, a proporção mínima de *cache hit* foi $0,76$

Figura 15 – Distribuição da Latência em Função da Distribuição de Popularidade.



Fonte: Elaborado pela autora (2022).

no caso médio de distribuição de popularidade, ou seja, para o valor de $\alpha = 0,8$ obteve valores vantajosos. Ademais, Breslau et al. (1999) enfatiza que a distribuição de popularidade se mantém entre $\alpha = 0,64$ e $\alpha = 0,83$, nesse sentido, para o valor de $\alpha = 0,6$ a proporção mínima de *cache hit* foi $0,74$. As proporções máximas de *cache miss* para $\alpha = 0,6$ e $\alpha = 0,8$ foram $0,26$ e $0,24$, respectivamente. 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 6 e demonstra a sensibilidade à dinâmica da rede.

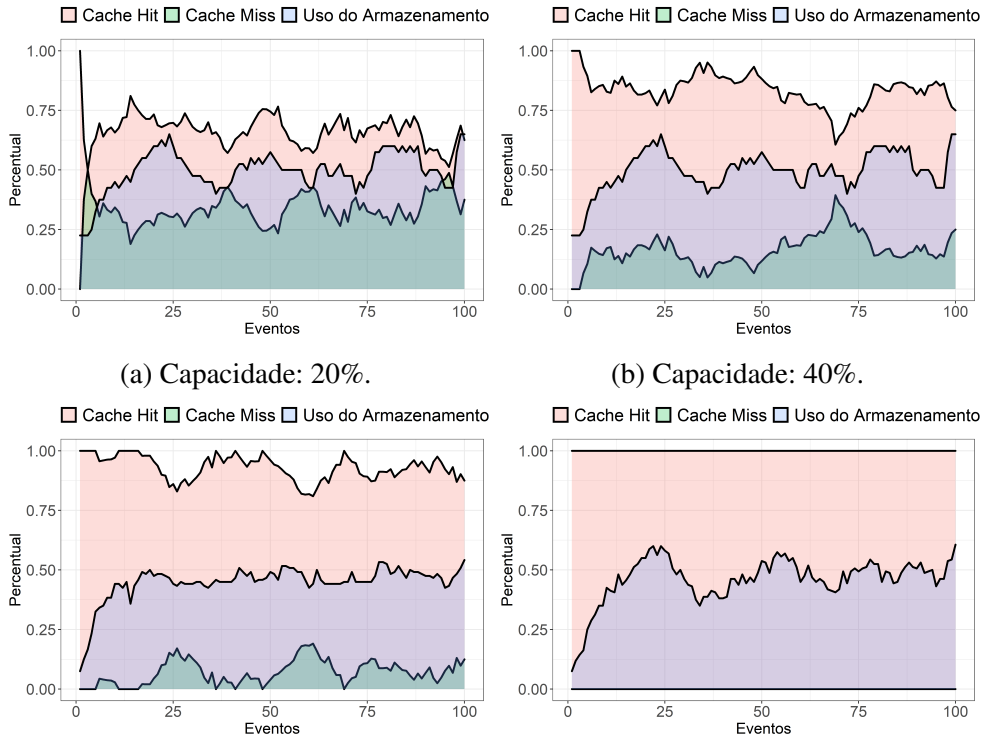
5.2.2 Capacidade Total de Armazenamento

A capacidade total de armazenamento variou entre 20% e 80% do tamanho da biblioteca com intervalos de 20%, representando a proporção de conteúdos que podem ser armazenados em *cache* em relação a capacidade total necessária para armazenar todo o conteúdo em *cache*. A Figura 16 apresenta os resultados obtidos para tais configurações. Inicialmente, é importante ressaltar que com a capacidade de 20% o valor máximo de *cache miss* foi $0,5$, assim como a proporção mínima de *cache hit* foi $0,5$. Ressalta-se que o *cache hit* máximo, ou seja, no valor de 1, foi atingido somente no primeiro evento discreto, no cenário inicial de simulação. Além disso o uso máximo de armazenamento foi $0,65$ relacionado a menor proporção de *cache hit*, em destaque na Figura 16. Enquanto, para capacidade de armazenamento com valor de 80% a uso máximo foi $0,60$. Ademais, há forte correlação estatística entre a proporção de *cache hit* e a variação da capacidade de armazenamento total, corroborado pelo valor do coeficiente de correlação de Pearson de $0,90$. Em suma, é possível observar na Figura 16 que a proporção de *cache hit* e *cache miss* está associada à capacidade total de armazenamento. Portanto, afirma-se que quanto maior a capacidade total de armazenamento, maior será a chance da requisição ser alocada em *cache*.

De modo geral, é possível deduzir que a capacidade de armazenamento influencia diretamente na política de *cache*, que busca posicionar o *cache* em um local da rede que não apresente sobrecarga, devido a maior liberdade na inserção do conteúdo, sem que seja necessário consumir o conteúdo através do BH. Ainda, é possível que a política de *cache* priorize o armazenamento em vez da rede, como ilustrado na Figura 16d, caso o MNO decida configurar o cenário com uma larga capacidade total de armazenamento (em relação ao tamanho da biblioteca). Nesse caso, todas as requisições são alocadas em *cache*, desbalanceando os pesos da política de *cache* conforme reflete o primeiro termo da Equação 6. Esse comportamento ocorre com 80% da capacidade em vez de 100% devido ao valor de $\alpha = 0,8$. Dessa forma, nem todos os conteúdos são solicitados.

A Tabela 3 demonstra as separatrizes em respeito à latência. Não há correlação estatística entre a latência e a variação de capacidade total de armazenamento (o valor do coeficiente de correlação de Pearson foi $-0,22$). Ainda, a Figura 17 demonstra graficamente a conformidade entre as diferentes variações de capacidade de armazenamento. A política de *cache* busca balancear o posicionamento e o roteamento de modo que a latência seja levada em conta. Assim, 50%

Figura 16 – Cenário de Capacidade de Total Armazenamento.



(a) Capacidade: 20%.

(b) Capacidade: 40%.

(c) Capacidade: 60%.

(d) Capacidade: 80%.

Fonte: Elaborado pela autora (2022).

Tabela 3 – Latência Total em Função da Capacidade Total de Armazenamento.

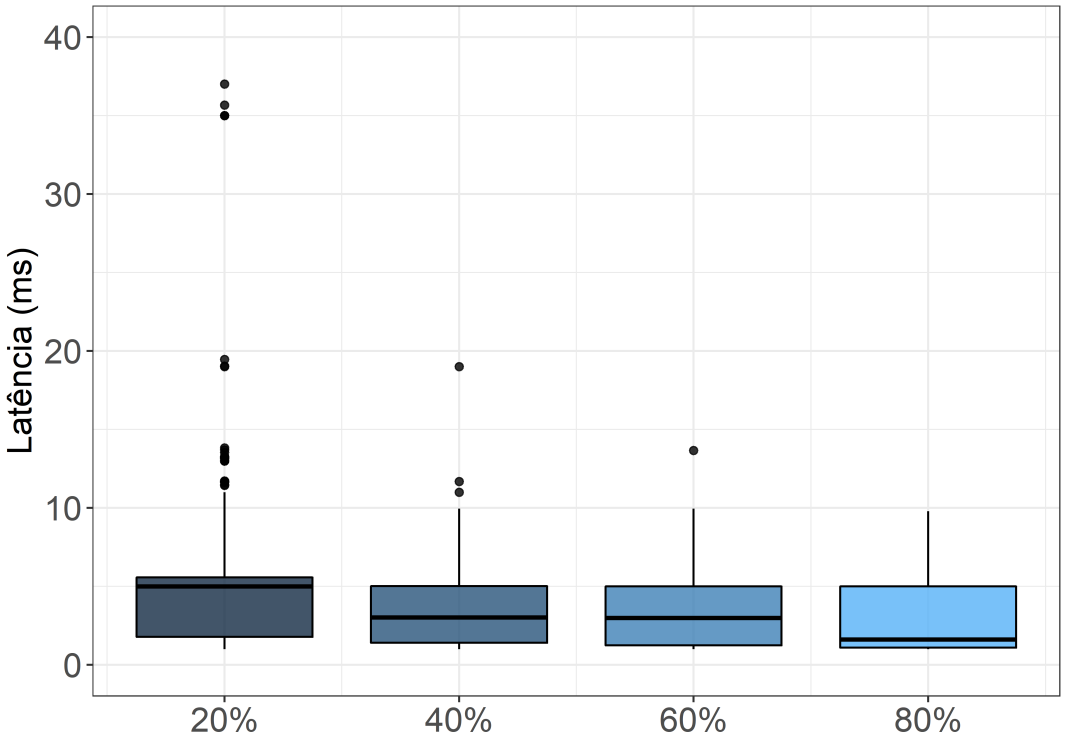
Capacidade	**Quartis**				**Percentil**
	1º	**2º**	**3º**	**4º**	**95º**
20%	1,7ms	5ms	5,5ms	67,6ms	7,4ms
40%	1,4ms	3ms	5ms	19ms	7ms
60%	1,2ms	3ms	5ms	13,6ms	7ms
80%	1,1ms	1,6ms	5ms	9,7ms	6ms

Fonte: Elaborado pela autora (2022).

(segundo quartil) da amostra manteve a latência inferior a 4 milissegundos (requisitos definidos por ITU (2017)), exceto para a capacidade total de armazenamento de 20%. Em 95% da amostra a latência se mostrou inferior a 6 milissegundos no melhor caso e 7,4 milissegundos no pior caso. A redução da capacidade significa também a redução de possibilidade de posicionamento e roteamento. Ou seja, a política de *cache* tem menor liberdade para escolher múltiplos caminhos, assim como menos possibilidades de migração de *caches* ou de ajuste dos caminhos acordo com a chegada de novas requisições.

Sobretudo, diante da variabilidade das capacidades de armazenamento, a política de *cache* demonstrou um comportamento adaptativo à dinamicidade da rede, no qual priorizou otimização da latência.

Figura 17 – Distribuição da Latência em Função da Capacidade Total de Armazenamento.



Fonte: Elaborado pela autora (2022).

5.3 ANÁLISE DE DESEMPENHO DO MODELO PARA POLÍTICA DE *CACHE*

De acordo com os trabalhos discutidos na Seção 2.7, dentre as abordagens propostas para políticas de *cache* existentes na literatura, destacam-se:

- **Não Cooperativa - Único Salto:** Tal abordagem não é cooperativa e considera apenas um único salto, isto é, não implementa a cooperação entre as BSs presentes na RAN.
- **Cooperação Multissaltos:** Tal abordagem permite a cooperação entre as BSs, realizando a busca pelo conteúdo em todas as *caches* presentes na RAN.

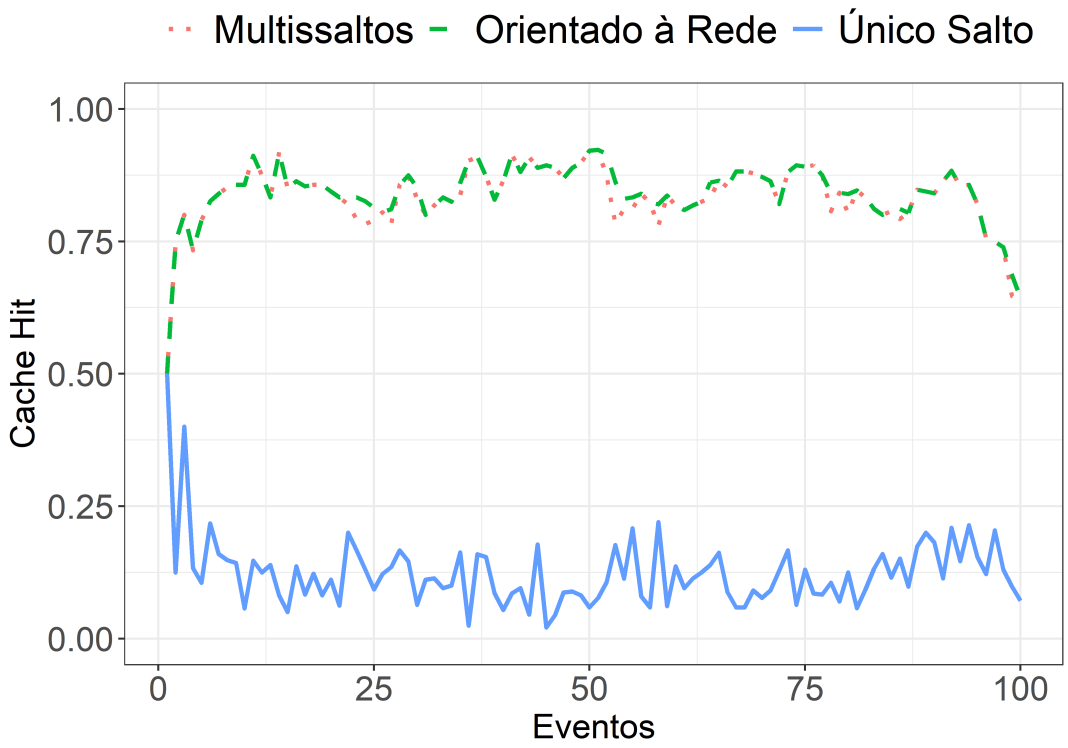
Para efeito de comparação, os conceitos fundamentais dessas abordagens foram generalizados e implementados a partir de adaptações ao modelo para política de *cache* cooperativa orientada à rede. É importante ressaltar que as duas abordagens representativas (único salto e multissalto) implementadas para realizar a análise experimental podem ser consideradas como uma versão otimizada das propostas iniciais. Tal argumentação é decorrente da possibilidade de realizar a reconfiguração das requisições previamente alocadas, bem como buscar que o conteúdo sempre seja posicionado o mais próximo possível do usuário, optimizando assim o uso do armazenamento e a qualidade final do serviço. Tais características são inerentes ao modelo proposto para a política de *cache* cooperativa orientada à rede. Por fim, para analisar o modelo proposto e os modelos de linha de base, o parâmetro γ foi ajustado para limitar o

espaço disponível para o posicionamento inicial das *caches*. Nesse sentido, e espaço de busca para inserção de conteúdo e roteamento de requisições foi limitado, configurando $\gamma = 0,1875$, o que resulta em 6 BSs inicialmente configuradas para receber uma determinada *cache*. Os demais parâmetros previamente apresentados permaneceram inalterados para efetuar a comparação (conforme descritos na Tabela 1). A formulação matemática para o modelo para a abordagem de único salto e para abordagem de multissaltos foi destalhada na Seção 4.4.

5.3.1 Proporção de *Cache Hit*, Uso Total do Armazenamento

Inicialmente, a comparação é realizada considerando os eventos de *cache hit* e o uso de armazenamento na RAN. Conforme a análise da proporção de *cache hit* dos modelos de linha de base e do modelo proposto, é possível observar na Figura 18 que o modelo para a política de *cache* proposta se assemelha ao modelo multissaltos em relação a proporções de *cache hit*. Por esse motivo as linhas no gráfico (Figura 18) praticamente foram sobrepostas. A cooperação, nesse caso, cooperação multissaltos, a qual realiza o posicionamento e o roteamento de requisições em toda a RAN, tem uma proporção de *cache hit* superior para ambos modelos multissaltos em relação ao modelo de único salto. Esse comportamento é esperado e, reafirma o que foi enfatizado por Li et al. (2017).

Figura 18 – Comparação de Proporção de *Cache Hit* .



Fonte: Elaborado pela autora (2022).

A média para o modelo proposto foi $0,84$ e desvio padrão $0,05$, enquanto a média para o modelo que considera um único salto foi $0,12$ e desvio padrão $0,06$. Por sua vez, a média para o

modelo multissaltos foi $0,83$ e desvio padrão $0,06$. A diferença entre as médias dos modelos para único salto e ambos modelos multissaltos foi $\approx 86\%$, ou seja, os modelos multissaltos obtiveram uma proporção de *cache hit* ≈ 7 vezes maior se comparados ao modelo de único salto.

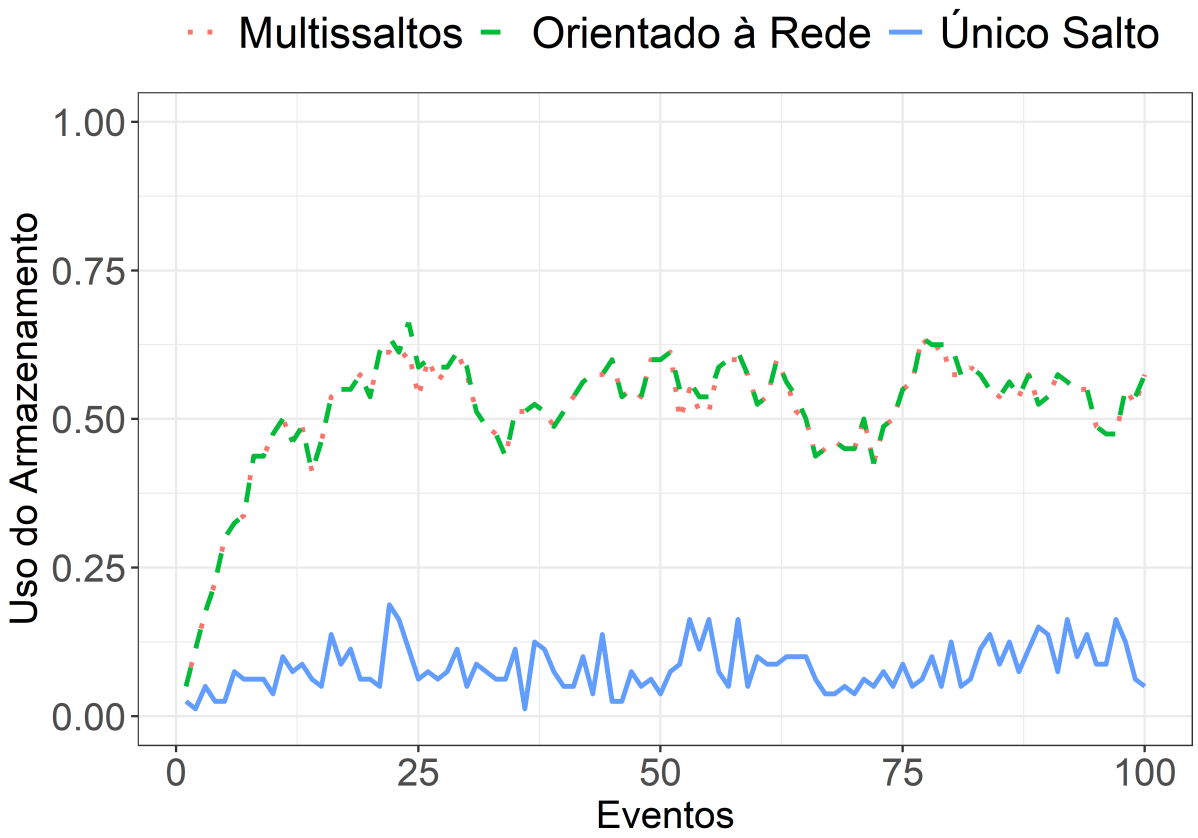
É importante ressaltar que embora o modelo multissaltos considere a cooperação e busca por conteúdo em todas BSs e tenha obtido uma média semelhante ao modelo orientado à rede, a abordagem de apenas posicionar o conteúdo em *cache* sem analisar efetivamente a rede não garante que a latência é reduzida ou que enlaces específicos não foram sobrecarregados. Diferentemente do modelo orientado à rede, o modelo que é apenas multissaltos não considera a dinâmica da rede, logo não tem a capacidade de escolher enlaces com maiores vazões com o objetivo de minimizar a latência.

O uso do armazenamento total reafirma essa conclusão. A média para o modelo de único salto foi $0,08$ e desvio padrão $0,03$. Para os modelos multissaltos e orientado à rede a média foi $0,51$ para ambos, e o desvio padrão foi $0,10$ e $0,07$, respectivamente (conforme apresentado na Figura 19). A média alcançada pelo modelo para único salto é 15% da média alcançada por ambos modelos multissaltos. Em resumo, a média do modelo multissaltos e do modelo proposto é $6,3$ vezes maior em relação a média do modelo de único salto. Ressalta-se que o desvio padrão resultante é causado pelo comportamento inicial, no qual as requisições estão sendo alocadas, visto que considera-se que no primeiro evento não havia nenhum conteúdo consumido originalmente em *cache*. Em resumo, o gráfico na Figura 19 demonstra que a capacidade total de armazenamento existente é subutilizada no modelo de único salto, isto é, para abordagens que não consideram a cooperação multissaltos.

Em decorrência da menor proporção de *cache hit*, as abordagens não cooperativas trafegam mais conteúdo através do BH, o que pode ser menos vantajoso na perspectiva de custo operacional do MNO, além de potencialmente aumentar a latência. Por outro lado, os modelo multissaltos e orientado à rede inserem maior carga sobre os enlaces de FH, ou seja, o trafego é distribuído ao longo do enlaces da RAN para que o conteúdo seja consumido através de *cache*. Em geral, o modelo multissaltos e o modelo orientado à rede prioriza o consumo de conteúdo em *cache*, tal comportamento é orientado pelo primeiro termo da Equação 6 e 14. Assim, o provedor de conteúdo e MNO são beneficiados, de acordo com os requisitos do SLA e, ainda, o MNO não tem sua em rede, o surgimento ou agravamento de possíveis congestionamentos e, portanto, podem oferecer uma melhor QoS para o usuário.

Notadamente, o modelo multissaltos e o modelo para a política de *cache* cooperativa orientada à rede obtiveram destaque em relação ao modelo de único salto. A análise de proporção de *cache hit* e de uso total do armazenamento não evidenciou a característica de balanceamento de carga presente no modelo para a política de *cache* cooperativa orientada à rede. Apenas a cooperação multissaltos não pode garantir a redução na latência. Há dicotomia entre consumir o conteúdo através de uma *cache* ou através do BH para reduzir a latência percebida pelo usuário que não pode ser expressada através da métrica de proporção de *cache hit* e *cache miss*. É necessário realizar análise de desempenho de métricas de latência para o modelo para a política

Figura 19 – Comparação de Proporção do Uso Total de Armazenamento.



Fonte: Elaborado pela autora (2022).

de *cache* cooperativa orientada à rede se comparada as abordagens de linha de base (único salto e multissaltos). Esta análise será detalhada na Subseção 5.3.2.

5.3.2 Latência

A comparação do comportamento da latência entre os modelos de linha de base e o modelo proposto são fundamentais para evidenciar o desempenho do modelo para a política de *cache* cooperativa orientada à rede. Além de priorizar o posicionamento e roteamento de requisições dentro da RAN têm como principal objetivo reduzir a latência da rede. Embora a proporção de *cache hit* seja maior em relação as abordagens de único saltos, apenas o roteamento de requisições e posicionamento multissaltos não garantem a QoS. Consequentemente, é necessário avaliar a eficiência do modelo proposto em relação à redução de latência. Em síntese, não basta que mais requisições consumam o conteúdo através de *cache* e por outro lado, sobrecarreguem um ponto da rede ou direcionem o consumo de conteúdo para um *cache* no qual é necessário que a requisição trafegue por enlaces congestionados e, consequentemente, ocorra piora na QoS.

Nesse sentido, quanto maior a proporção *cache hit* maior é a vantagem do MNO do ponto de vista de custos operacionais, em outras palavras, redução do tráfego através do BH. No entanto, o consumo de conteúdo em *cache* pode gerar congestionamento na rede sob sua

administração e, deste modo pode piorar na QoS. Frente a esse cenário é necessário realizar o balanceamento de carga, ou seja, por um lado priorizar o consumo de conteúdo através da *cache* e, por outro lado, analisar se o consumo através do BH é mais vantajoso a depender da presença de congestionamento em enlaces específicos na rede. Em resumo, sob essa perspectiva, a política proposta tem como um de suas principais características o balanceamento de carga entre *cache* e BH com o objetivo de reduzir a latência. Assim, se propõe a realizar o balanceamento de carga para ponderar o que é mais vantajoso em relação à QoS, principalmente da perspectiva do usuário e do provedor de serviço e, ainda, se possível, pode ser proveitosa a partir da perspectiva do MNO.

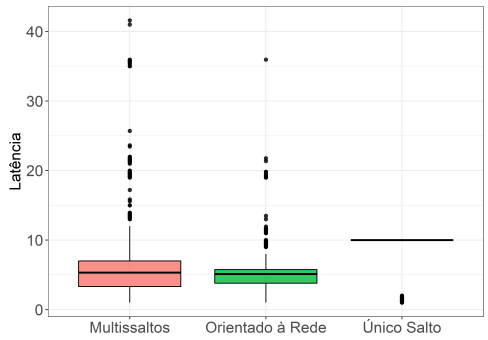
Na Figura 20a é possível observar a distribuição da latência na rede. Para o modelo de único salto, a maioria das requisições foram trafegadas através do BH e, em apenas 11% da amostra a latência foi menor que 10 milissegundos. Tal comportamento, significa que o modelo de único salto tem somente duas escolhas: consumir o conteúdo da BS em que o usuário está conectado, que, em geral, é um caminho menor e, consequentemente com menor latência ou, consumir o conteúdo através do BH, o qual a latência é aproximadamente 10 vezes maior. Esse comportamento é explicado pela a proporção de *cache miss*, que foi em média $0,87$ e desvio padrão $0,06$ para único salto. Para o modelo multissaltos a proporção de *cache miss* foi em média $0,16$ e desvio padrão $0,06$ e $0,15$ e desvio padrão $0,05$ para o modelo orientado à rede. Sendo assim, é natural que maioria das requisições sejam alocadas em nuvem, isto é, sejam transferidas através do enlace de BH e, consequentemente, impactem na latência. Portanto, devido ao espaço de armazenamento limitado e a limitação de conexões diretas com as BS, em geral a maior parte do conteúdo é consumido através do BH.

Por outro lado, o comportamento dos modelos multissaltos e orientado à rede se assemelham conforme enfatizado pela Figura 20a. No entanto, para 75% da amostra a latência foi menor que 7 milissegundos para o modelo multissaltos. Enquanto que para o modelo orientado à rede, 75% da amostra foi menor que 5 milissegundos. Embora 34% das amostras tenham sido inferiores a 4 milissegundos para o modelo multissaltos e 26 milissegundos para o modelo orientado à rede. Esse comportamento pode ser observado na Figura 20b na qual a latência obteve de fato mais valores menores para o modelo multissaltos. Contudo, obteve maior variabilidade na amostra, possivelmente resultado de pontos de sobrecarga na rede, dado que esse modelo não é sensível à rede e não evita o agravamento de congestionamentos (Figura 20a. Sendo assim, para o modelo orientado à rede, 99% da amostra foi inferior a 10 milissegundos contra 91% para o modelo multissaltos.

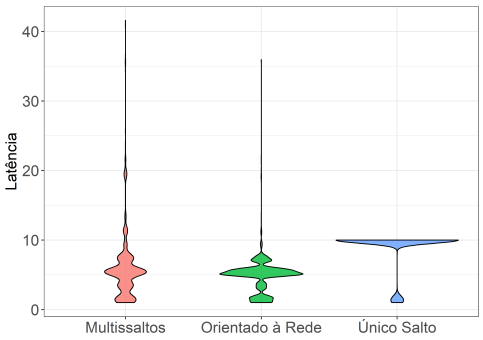
Em síntese é observável que o modelo orientado à rede manteve a latência estável, em razão da menor variabilidade da amostra.

O modelo multissaltos herda a característica de posicionamento mais próximo possível do usuário, comportamento resultante do primeiro termo da Equação 14, semelhante ao primeiro termo da Equação 6 (função objetivo do modelo orientado á rede). No entanto, o modelo multissaltos aloca requisições priorizando a inserção de conteúdo e o roteamento de

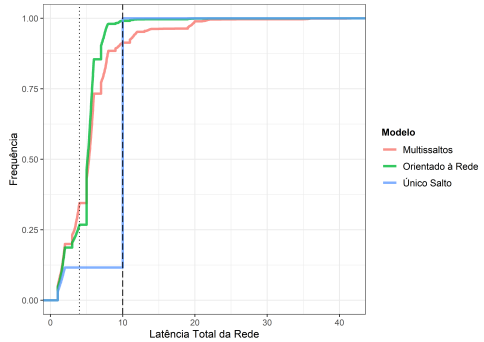
Figura 20 – Latência entre modelo proposto e linhas de base.



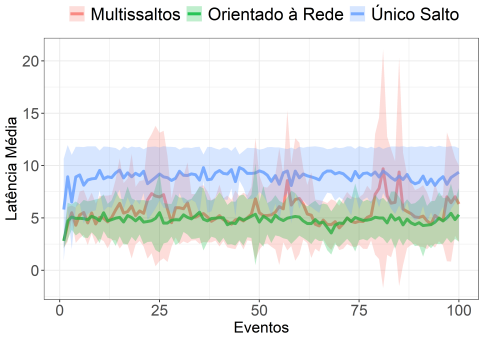
(a) Distribuição da Latência.



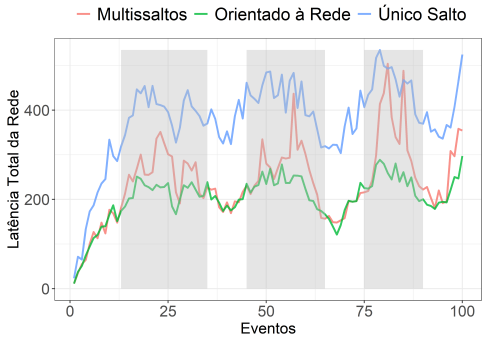
(b) Distribuição da Latência.



(c) Distribuição Acumulada da Latência.



(d) Latência Média.



(e) Latência Total.

Fonte: Elaborado pela autora (2022).

requisições mais próximo a borda da RAN e, posteriormente aloca requisições em níveis mais distante da borda ou trafega o conteúdo através do BH. Desse modo, pode fazer com que o tráfego das requisições sejam agrupados em um único ponto da rede, o qual pode desencadear sobrecarga. Este comportamento pode agravar um congestionamento presente da rede, dado que o modelo não busca por caminhos que apresentam maiores vazões, aumentando a latência resultante, como destacado na Figura 20c. Por sua vez, a linha pontilhada é uma referência para 4 milissegundos (ITU, 2017). É observável que o modelo orientado à rede obteve menores valores em relação a latência.

O gráfico na Figura 20e enfatiza o sucesso do modelo proposto para reduzir a latência em relação aos modelo de linha de base. É possível observar que inicialmente ambos os modelos têm aumento continuo na latência total. Esse comportamento inicial ocorre devido ao primeiro

instante em que as requisições são alocadas, posteriormente há declives na curva. Na sequência, é possível observar três picos na curva destacados na Figura 20e, referentes ao modelo multissaltos que se destacam em relação a curva que representa o modelo orientado à rede e apresentam valores superiores para a latência total. Em contrapartida, esse contraste evidência a estabilidade na redução da latência para o modelo orientado à rede. Ainda, é possível observar na Figura 20d o modelo orientado à rede obteve latência média inferior a 4 milissegundos e, apenas no evento 70 a latência foi superior à 4 milissegundos (ITU, 2017). Além disso, o desvio padrão visualizado na gráfico da Figura 20d evidência a instabilidade na latência da rede. A curva que representa o modelo multissaltos tem picos que enfatizam latências superiores a 4 milissegundos e colaboram para concluir que o modelo multissaltos não orientado à rede não impede que a latência da rede aumente dado a eventual existência de congestionamento.

O modelo orientado à rede, em geral, demonstra maior estabilidade na latência. Isso ocorre devido ao efeito de espalhamento e balanceamento de carga que acontece. Ainda, o modelo busca por enlaces com maiores vazões, buscando sempre menores latências, o que faz com que o modelo evite agravar o congestionamento nos enlaces (enquanto for possível). Por fim, o modelo multissaltos apenas busca alocar as requisições na rede, a medida que a capacidade de armazenamento vai sendo esgotada.

5.4 CONSIDERAÇÕES PARCIAIS

Inicialmente, este capítulo descreveu os cenários, métricas analisadas e detalhou os valores dos parâmetros necessários para execução da simulação numérica. Posteriormente, apresentou e analisou os resultados a partir dos dados obtidos. Nesse sentido, foram executadas simulações em dois cenários, o primeiro cenário avaliou o comportamento do modelo para política de *cache* orientada à rede e, o segundo cenário, por sua vez, avaliou a eficiência do modelo em relação a modelos que seguem abordagens existentes na literatura.

O primeiro cenário teve como objetivo avaliar o comportamento do modelo de acordo com as variações de dois conjuntos de parâmetros: variação da distribuição de popularidade dos conteúdos e a capacidade total de armazenamento disposto na RAN. A variação da distribuição de popularidade dos conteúdos demonstrou a sensibilidade quanto à dinâmica da rede existente na política de *cache*. A proporção de *cache hit* obteve melhores resultados para valores de α maiores. Portanto, é possível deduzir que quanto maior o valor de α, mais concentrada é a distribuição de popularidade dos conteúdo, assim, há uma menor variação entre os conteúdos solicitados e, consequentemente menor uso de armazenamento é necessário. Por outro lado, se α tem um valor menor, a variabilidade do conteúdo aumenta e, consequentemente, requer maior capacidade de armazenamento, provocando aumento da proporção de *cache miss*.

Destaca-se que o uso capacidade de armazenamento total não atingiu o uso máximo, ou seja, ainda que houvesse capacidade suficiente, a política de *cache* optou por disponibilizar o conteúdo a partir do BH. Esse comportamento, possivelmente, demonstra sobrecarga na rede,

causada pela concentração de conteúdos com maior popularidade ou decorrente da mobilidade dos usuários. Tal comportamento é resultante da análise de caminhos intermediários do roteamento de requisições multissaltos, ou seja, o segundo termo da função objetivo (Equação 6).

Por sua vez, a variação da capacidade de armazenamento, demonstrou que a disponibilidade de armazenamento é importante para a cooperação horizontal, o qual impacta diretamente nas proporções de *cache hit* e *cache miss*. A latência não foi impactada, isto é, a política, possivelmente, priorizou a otimização da latência. Em resumo, mesmo com a mudança de comportamento entre os usuários, assim como a capacidade de armazenamento, a política de *cache* obteve êxito no balanceamento e na otimização da latência. Portanto, o modelo para política de *cache* orientada à rede demonstrou sensibilidade quanto à dinâmica da rede.

Finalmente, o segundo cenário teve como objetivo evidenciar a efetividade do modelo para a política de *cache* orientada à rede na redução da latência se comparadas a outras estratégias existentes na literatura. Nesse sentido, foram formulados um modelo de único salto e um modelo multissaltos, tais modelos foram usados como linha de base de comparação. Ressalta-se que ambos modelos de linha de base são adaptações do modelo proposto e, portanto, herdam características importantes, desenvolvidas no presente trabalho. Especificamente, herdam a reorganização e re-otimização das requisições alocadas e a priorização da inserção de conteúdo e roteamento de requisições para a borda da rede.

Em geral, foi possível analisar a eficiência do modelo multissaltos e orientado à rede (também multissaltos), em relação a métrica de proporção de *cache hit*. Ambos modelos obtiveram exito significativo se comparados ao modelo de único salto, reafirmando a eficiência da cooperação de com resultados antecedentes da literatura (LI et al., 2017). Destaca-se, que o modelo para a política de *cache* orientada à rede mantém a latência estável devido ao seu mecanismo de balanceamento de carga e sensibilidade à rede. Tal mecanismo faz com que o modelo proposto busque por caminhos com maiores vazões, distribuindo o tráfego ao longo da rede e além disso, ponderando se o consumo do conteúdo através BH pode ser mais vantajoso para redução da latência em vez do consumo através de um *cache*. Portanto, o modelo para política de *cache* orientada à rede mostrou-se mais eficiente em relação aos modelos de linhas de base.

6 CONCLUSÃO

Frente ao atual cenário de crescimento de tráfego de dados, quantidade de dispositivos móveis, número de usuários e surgimento de novas aplicações com requisitos de menor latência e maior vazão se faz necessário uma evolução nas redes móveis (PARVEZ et al., 2018). Para isso, a rede 5G se apresenta como uma iniciativa perante ao avanço necessário. Especificamente, dentro dos cenários URLLC, eMBB, mMTC determinados pelo ITU, os principais requisitos foram definidos como redução da latência, aumento da capacidade de tráfego de dados e suporte a mobilidade de usuários e alta densidade de dispositivos (ITU, 2017), com objetivo de suportar aplicações IoT, AR, VR, direção autônoma, Indústria 4.0 e entre outras possíveis aplicações ou serviços (PHAM et al., 2020).

Para isso, o *cache* de conteúdo aliado à técnicas de MEC (ABBAS et al., 2018) e HCN (ANDREWS, 2013) é uma alternativa promissora para contribuir para redução da latência e redução de tráfego em enlaces de BH. O *cache* de conteúdo em redes móveis pode evitar tráfego replicado e se apropriar da maior proximidade geográfica com o usuário para reduzir a latência (WU et al., 2021). Desenvolver uma política de *cache* tem características desafiadoras tais como: recursos de armazenamento limitado, mobilidade do usuário, congestionamento da rede, conteúdos com diferentes popularidades, as quais modificam de acordo com a localidade e tempo (GOIAN et al., 2019).

Para responder a questão: A partir de quais mecanismos é possível reduzir a latência *E2E* no que compreende os problemas de inserção de conteúdo e roteamento de requisições. Esse trabalho desenvolveu e especificou 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 através de ILP com restrições de capacidade de armazenamento e requisitos de QoS definidos pelo provedor de serviço.

Ademais, o trabalho apresentou a fundamentação teórica e destacou trabalhos relacionados direcionados ao problema de roteamento de requisições. Posteriormente, a política de *cache* foi detalhada e, assim, enfatizou sua principal premissa. Tal premissa consiste em realizar a predição de um possível congestionamento na rede, baseando-se na seguinte afirmação: Quanto maior a vazão, melhores são as condições do enlace (STALLINGS, 2015). Além disso, baseia-se na afirmativa: Quanto menor o RTT, maior será a vazão do enlace (CHIU; JAIN, 1989). Nesse sentido, tem-se a hipótese que a busca por enlaces com maiores vazões, distribuindo as demandas na rede, pode evitar o agravamento de possíveis congestionamentos. Adicionalmente, considera a cooperação, a partir da busca pelo conteúdo em todas as BSs presentes na RAN e roteamento de requisições multissaltos, no qual analisa os caminhos intermediários entre origem e destino. A política de *cache* se baseou nos mecanismos de CC utilizados no TCP Vegas (BRAKMO; PETERSON, 1995) (CARDWELL et al., 2017) (LANGLEY et al., 2017).

Simulações numéricas realizadas a partir de modelo para política de *cache* orientada

à rede, 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. Por sua vez, a variabilidade na capacidade de armazenamento total não demonstrou impacto significativo na latência.

Em síntese, a variação no comportamento dos usuários e na capacidade total de armazenamento tem impactos na otimização, entretanto, não demonstra impacto sobre à latência. Tais parâmetros foram testados com intuito de gerar estresse na política de *cache* e analisar seu comportamento diante de fenômenos extremos.

Por sua vez, o modelo para à política de *cache* cooperativo orientado à rede mostrou-se eficiente de acordo com as análises realizadas. Em razão da sua característica de cooperação multissaltos a a partir de uma visão global da rede, obteve média $6,3$ vezes maior em relação a média do modelo de único salto, não cooperativo. O modelo orientado à rede, em geral, demonstra maior estabilidade na latência, devido a distribuição do tráfego e balanceamento de carga entre tráfego de conteúdo através de BH e *cache*. O modelo orientado à rede busca por enlaces com maiores vazões, buscando sempre menores latências, consequentemente, evita agravar o congestionamento nos enlaces. Para 99% da amostra foi a latência foi inferior a 10 milissegundos contra 91% para o modelo multissaltos. Em resumo, os resultados das simulações obtidos a partir da especificação, desenvolvimento e análise do modelo para a política de *cache* cooperativa orientada à rede que une os problemas de inserção de conteúdo e roteamento de requisições, indicam confluência em direção ao objetivo geral proposto.

Finalmente, esse trabalho contribuiu concretamente a partir do modelo formulado, que pode ser generalizável, além da aplicação atual existente (VoD), outras aplicações e algoritmos podem ser aplicados futuramente. Assim, é possível que aplicações futuras que necessitem de *cache* possam implementar uma política baseada no modelo proposto. Além disso, a política de *cache* proposta enfatiza princípios importantes da perspectiva do MNO, provedor de serviço ou conteúdo e usuário. Esse trabalho apresentou uma comparação com estratégias para política *cache* existente na literatura.

6.1 TRABALHOS FUTUROS

Esse trabalho avaliou a viabilidade e a efetividade da política de *cache* proposta através modelagem matemática desenvolvida a partir de otimização combinatorial. Em vista da eficiência de uma política de *cache* com a principal estratégia de orientação à rede, futuramente, é necessário desenvolver um algoritmo que seja executável em tempo polinomial e seja capaz de se aproximar dos resultados obtidos a partir da simulação numérica. Assim como, a comparação do custo computacional entre o método exato e o algoritmo que será projetado considerando dados reais

de mobilidade. Por fim novas métricas deveram ser avaliadas como, carga total do FH e BH, número de saltos e migrações.

6.2 PUBLICAÇÕES

A partir deste trabalho foram submetidos e aceitos os seguintes trabalhos:

- Anais da XVIII Escola Regional de Redes de Computadores: Uma Proposta Inicial baseada em Mobile Edge Computing para Orquestrar Caches em Redes 5G (ALVES; KOSLOVSKI, 2020).
- XXVII Workshop de Gerência e Operação de Redes e Serviços: Política Cooperativa Orientada a Rede para Posicionamento de Cache e Roteamento de Requisições em Redes Celulares Heterogêneas (ALVES; KOSLOVSKI, 2022b).
- *27th IEEE Symposium on Computers and Communications*: *Joint Cache Placement and Request Routing Optimization in Heterogeneous Cellular* (ALVES; KOSLOVSKI, 2022a).

REFERÊNCIAS

ABBAS, Nasir et al. Mobile edge computing: A survey. **IEEE Internet of Things Journal**, v. 5, n. 1, p. 450–465, 2018. Disponível em: <https://ieeexplore.ieee.org/abstract/document/8030322>. Acesso em: 8 Set. 2020.

AHUJA, Ravindra K.; MAGNANTI, Thomas L.; ORLIN, James B. **Network Flows**: Theory, algorithms, and applications. United States of America: Prentice Hall, 1993.

ALVES, Marisangila; KOSLOVSKI, Guilherme. Uma proposta inicial baseada em mobile edge computing para orquestrar caches em redes 5g. In: **Anais da XVIII Escola Regional de Redes de Computadores**. Porto Alegre, RS, Brasil: SBC, 2020. p. 66–71. ISSN 0000-0000. Disponível em: <https://sol.sbc.org.br/index.php/errc/article/view/15191>.

ALVES, Marisangila; KOSLOVSKI, Guilherme. Joint cache placement and request routing optimization in heterogeneous cellular networks. In: **27th IEEE Symposium on Computers and Communications**. Rhodes Island, Greece: [s.n.], 2022.

ALVES, Marisangila; KOSLOVSKI, Guilherme. Política cooperativa orientada a rede para posicionamento de cache e roteamento de requisições em redes celulares heterogêneas. In: **XXVII Workshop de Gerência e Operação de Redes e Serviços**. Fortaleza, Brasil: [s.n.], 2022.

ANDREWS, Jeffrey G. Seven ways that hetnets are a cellular paradigm shift. **IEEE Communications Magazine**, v. 51, n. 3, p. 136–144, 2013. Disponível em: <https://ieeexplore.ieee.org/abstract/document/6476878>. Acesso em: 06 Abr. 2021.

BARAKABITZE, Alcardo Alex et al. 5g network slicing using sdn and nfv: A survey of taxonomy, architectures and future challenges. **Computer Networks**, v. 167, p. 106984, 2020. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1389128619304773>. Acesso em: 26 Ago. 2020.

BASTUG, Ejder; BENNIS, Mehdi; DEBBAH, Mérouane. Living on the edge: The role of proactive caching in 5g wireless networks. **IEEE Communications Magazine**, v. 52, n. 8, p. 82–89, 2014. Disponível em: <https://ieeexplore.ieee.org/abstract/document/6871674>. Acesso em: 30 Abr. 2021.

BOGALE, Tadilo Endeshaw; LE, Long Bao. Massive mimo and mmwave for 5g wireless hetnet: Potential benefits and challenges. **IEEE Vehicular Technology Magazine**, v. 11, n. 1, p. 64–75, 2016. Disponível em: <https://ieeexplore.ieee.org/abstract/document/7397887>. Acesso em: 26 Set. 2021.

BONDY, John Adrian; MURTY, Uppaluri Siva Ramachandra. **Graph Theory with Applications**. 1. ed. Great Britain: Elsevier Science Publishing, 1976.

BRAKMO, Lawrence S.; PETERSON, Larry L. Tcp vegas: End to end congestion avoidance on a global internet. **IEEE Journal on Selected Areas in Communications**, v. 13, n. 8, p. 1465–1480, 1995. Disponível em: <https://ieeexplore.ieee.org/abstract/document/464716>. Acesso em: 20 Jan. 2021.

BRESLAU, L. et al. 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. New York, NY, USA: [s.n.], 1999. p. 126–134. Disponível em: <https://ieeexplore.ieee.org/abstract/document/749260>. Acesso em: 26 Ago. 2021.

CARDWELL, Neal et al. Bbr: Congestion-based congestion control. **Communications of the ACM**, v. 60, p. 58–66, 2017. Disponível em: <http://cacm.acm.org/magazines/2017/2/212428-bbr-congestion-based-congestion-control/fulltext>. Acesso em: 28 Jan. 2021.

CHIU, Dah-Ming; JAIN, Raj. Analysis of the increase and decrease algorithms for congestion avoidance in computer networks. **Computer Networks and ISDN Systems**, v. 17, n. 1, p. 1–14, 1989. Disponível em: <https://www.sciencedirect.com/science/article/pii/0169755289900196>. Acesso em: 21 Jul. 2021.

CISCO. **Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2017-2022**. [S.l.], 2019. Disponível em: <https://newsroom.cisco.com/press-release-content?articleId=1955935>. Acesso em: 12 Ago. 2020.

CISCO. **Cisco Annual Internet Report 2018–2023**. [S.l.], 2020. Disponível em: <https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html>. Acesso em: 12 Ago. 2020.

DAMNJANOVIC, Aleksandar et al. A survey on 3gpp heterogeneous networks. **IEEE Wireless Communications**, v. 18, n. 3, p. 10–21, 2011. Disponível em: <https://ieeexplore.ieee.org/abstract/document/5876496>. Acesso em: 20 Set. 2021.

DANG, Shuping et al. What should 6g be? **IEEE Access**, v. 3, p. 20–29, 2020. Disponível em: <https://www.nature.com/articles/s41928-019-0355-6>. Acesso em: 06 Nov. 2020.

DEHGHAN, Mostafa et al. On the complexity of optimal request routing and content caching in heterogeneous cache networks. **IEEE/ACM Transactions on Networking**, v. 25, n. 3, p. 1635–1648, 2017. Disponível em: <https://ieeexplore.ieee.org/abstract/document/7797148>. Acesso em: 28 Mar. 2021.

DILLEY, J. et al. Globally distributed content delivery. **IEEE Internet Computing**, v. 6, n. 5, p. 50–58, 2002. Disponível em: <https://ieeexplore.ieee.org/abstract/document/1036038>. Acesso em: 18 set. 2021.

ERICSSON. **Ericsson Mobility Report**. [S.l.], 2020. Disponível em: <https://www.ericsson.com/4adc87/assets/local/mobility-report/documents/2020/november-2020-ericsson-mobility-report.pdf>. Acesso em: 12 Nov. 2020.

EVEN, S.; ITAI, A.; SHAMIR, A. On the complexity of time table and multi-commodity flow problems. In: **16th Annual Symposium on Foundations of Computer Science (SFCS 1975)**. NW Washington, DC, USA: [s.n.], 1975. p. 184–193. Disponível em: <https://ieeexplore.ieee.org/abstract/document/4567876>. Acesso em: 16 Nov. 2021.

GIL, Antonio Carlos. **Como Elaborar Projetos de Pesquisa**. São Paulo: Atlas, 2010.

GOIAN, Huda S. et al. Popularity-based video caching techniques for cache-enabled networks: A survey. **IEEE Access**, v. 7, p. 27699–27719, 2019. Disponível em: <https://ieeexplore.ieee.org/abstract/document/8658196>. Acesso em: 26 Set. 2021.

GOLDBARG, M.C.; LUNA, H.P.L. **Otimização Combinatória e Programação Linear: Modelos e Algoritmos**. Rio de Janeiro - RJ - Brasil: Elsevier, 2005.

HARUTYUNYAN, Davit; BRADAI, Abbas; RIGGIO, Roberto. Trade-offs in cache-enabled mobile networks. In: **2018 14th International Conference on Network and Service Management (CNSM)**. Rome, Italy: [s.n.], 2018. p. 116–124. Disponível em: <https://ieeexplore.ieee.org/abstract/document/8584988>. Acesso em: 24 Mar. 2020.

HU, Yun Chao et al. **Mobile Edge Computing A Key Technology Towards 5G**. Sophia Antipolis, CEDEX, France, 2015. Disponível em: <https://infotech.report/Resources/Whitepapers/f205849d-0109-4de3-8c47-be52f4e4fb27_etsi_wp11_mec_a_key_technology_towards_5g.pdf>. Acesso em: 9 Set. 2020.

ITU, INTERNATIONAL TELECOMMUNICATION UNION. **IMT Vision – Framework and Overall Objectives of the Future Development of IMT for 2020 and Beyond**. [S.l.], 2015. Disponível em: <https://www.itu.int/rec/R-REC-M.2083>. Acesso em: 25 Set. 2020.

ITU, INTERNATIONAL TELECOMMUNICATION UNION. **Minimum Requirements Related to Technical Performance for IMT-2020 Radio Interface(s)**. [S.l.], 2017. Disponível em: <https://www.itu.int/pub/R-REP-M.2410>. Acesso em: 9 Set. 2021.

JIANG, Wei; FENG, Gang; QIN, Shuang. Optimal cooperative content caching and delivery policy for heterogeneous cellular networks. **IEEE Transactions on Mobile Computing**, v. 16, n. 5, p. 1382–1393, 2017. Disponível em: <https://ieeexplore.ieee.org/abstract/document/7530876>. Acesso em: 01 Abr. 2021.

KABIR, Asif et al. The role of caching in next generation cellular networks: A survey and research outlook. **Trans. Emerg. Telecommun. Technol.**, USA, v. 31, n. 2, p. 1–25, 2020. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/ett.3702>. Acesso em: 25 out. 2021.

KAMEL, Mahmoud; HAMOUDA, Walaa; YOUSSEF, Amr. Ultra-dense networks: A survey. **IEEE Communications Surveys Tutorials**, v. 18, n. 4, p. 2522–2545, 2016. Disponível em: <https://ieeexplore.ieee.org/abstract/document/7476821>. Acesso em: 20 Set. 2021.

KARP, Richard M. On the computational complexity of combinatorial problems. **Networks**, Wiley Online Library, v. 5, n. 1, p. 45–68, 1975.

KEKKI, Sami et al. **MEC in 5G Networks**. Sophia Antipolis, CEDEX, France, 2018. Disponível em: <https://www.etsi.org/images/files/ETSIWhitePapers/etsi_wp28_mec_in_5G_FINAL.pdf>.

KHREISHAH, Abdallah; CHAKARESKI, Jacob; GHARAIBEH, Ammar. Joint caching, routing, and channel assignment for collaborative small-cell cellular networks. **IEEE Journal on Selected Areas in Communications**, v. 34, n. 8, p. 2275–2284, 2016. Disponível em: <https://ieeexplore.ieee.org/abstract/document/7485844>. Acesso em: 29 Mar. 2021.

KUROSE, Jim F.; ROSS, Keith W. **Redes de Computadores e a internet**: Uma abordagem top-down. 6. ed. São Paulo - SP - Brasil: Pearson, 2014.

LANGLEY, Adam et al. The quic transport protocol: Design and internet-scale deployment. In: **Proceedings of the Conference of the ACM Special Interest Group on Data Communication**. New York, NY, USA: Association for Computing Machinery, 2017. p. 183–196. Disponível em: <https://doi.org/10.1145/3098822.3098842>. Acesso em: 17 Out. 2021.

LI, Xiuhua et al. Collaborative multi-tier caching in heterogeneous networks: Modeling, analysis, and design. **IEEE Transactions on Wireless Communications**, v. 16, n. 10, p. 6926–6939, 2017. Disponível em: <https://ieeexplore.ieee.org/abstract/document/8002665>. Acesso em: 21 Set. 2021.

LIU, Chunshan et al. Millimeter-wave small cells: Base station discovery, beam alignment, and system design challenges. **IEEE Wireless Communications**, v. 25, n. 4, p. 40–46, 2018. Disponível em: <https://ieeexplore.ieee.org/abstract/document/8454665>. Acesso em: 26 Set. 2021.

LYU, Xinchen et al. Distributed online learning of cooperative caching in edge cloud. **IEEE Transactions on Mobile Computing**, v. 20, n. 8, p. 2550–2562, 2021. Disponível em: <https://ieeexplore.ieee.org/abstract/document/9050836>. Acesso em: 21 Jul. 2021.

MADDAH-ALI, Mohammad Ali; NIESEN, Urs. Fundamental limits of caching. **IEEE Transactions on Information Theory**, v. 60, n. 5, p. 2856–2867, 2014. Disponível em: <https://ieeexplore.ieee.org/abstract/document/6763007>. Acesso em: 29 Mar. 2021.

MARCONI, Marina de Andrade; LAKATOS, Eva Maria. **Fundamentos de Metodologia Científica**. 5. ed. São Paulo - SP - Brasil: Atlas, 2003.

META, Engineering at. **Network hose: Managing uncertain network demand with model simplicity**. Facebook, 2021. Facebook Engineering. Disponível em: <https://engineering.fb.com/2021/06/15/data-infrastructure/network-hose/>. Acesso em: 22 Nov. 2021.

NAVARRO-ORTIZ, Jorge et al. A survey on 5g usage scenarios and traffic models. **IEEE Communications Surveys Tutorials**, v. 22, n. 2, p. 905–929, 2020. Disponível em: <https://ieeexplore.ieee.org/abstract/document/8985528>. Acesso em: 20 Set. 2021.

PARVEZ, Imtiaz et al. A Survey on Low lLtency Towards 5G: RAN, Core Network and Caching Solutions. **IEEE Communications Surveys and Tutorials**, v. 20, n. 4, p. 3098–3130, 2018. Disponível em: <https://ieeexplore.ieee.org/abstract/document/8367785>. Acesso em: 31 Out. 2020.

PASCHOS, Georgios S. et al. The role of caching in future communication systems and networks. **IEEE Journal on Selected Areas in Communications**, v. 36, n. 6, p. 1111–1125, 2018. Disponível em: <https://ieeexplore.ieee.org/abstract/document/8471001>. Acesso em: 25 Out. 2021.

PHAM, Quoc-Viet et al. A survey of multi-access edge computing in 5g and beyond: Fundamentals, technology integration, and state-of-the-art. **IEEE Access**, v. 8, p. 116974–117017, 2020. Disponível em: <https://ieeexplore.ieee.org/abstract/document/9113305>. Acesso em: 06 Nov. 2020.

PRODANOV, Cleber Cristiano; FREITAS, Ernani Cesar De. **Metodologia do Trabalho Científico: Métodos e Técnicas da Pesquisa e do Trabalho acadêmico-2ª Edição**. 2. ed. Novo Haburgo - RS - Brasil: Editora Feevale, 2013.

PU, Lingjun et al. 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**, v. 36, n. 8, p. 1751–1767, 2018. Disponível em: <https://ieeexplore.ieee.org/abstract/document/8374065>. Acesso em: 27 Mar. 2021.

SENA, Paulo et al. Cache-aware interest routing: Impact analysis on cache decision strategies in content-centric networking. In: **Proceedings of the 9th Latin America Networking Conference**. New York, NY, USA: Association for Computing Machinery, 2016. (LANC '16), p. 39–45. ISBN 9781450345910. Disponível em: <https://doi.org/10.1145/2998373.2998445>. Acesso em: 09 Mai. 2022.

SHANMUGAM, Karthikeyan et al. Femtocaching: Wireless content delivery through distributed caching helpers. **IEEE Transactions on Information Theory**, v. 59, n. 12, p. 8402–8413, 2013. Disponível em: <https://ieeexplore.ieee.org/abstract/document/6600983>. Acesso em: 26 Ago. 2020.

SHENG, Min et al. Enhancement for content delivery with proximity communications in caching enabled wireless networks: architecture and challenges. **IEEE Communications Magazine**, v. 54, n. 8, p. 70–76, 2016. Disponível em: <https://ieeexplore.ieee.org/abstract/document/7537179>. Acesso em: 01 Abr. 2021.

SHI, Weisong et al. Edge computing: Vision and challenges. **IEEE Internet of Things Journal**, v. 3, n. 5, p. 637–646, 2016. Disponível em: <https://ieeexplore.ieee.org/abstract/document/7488250>. Acesso em: 9 Set. 2020.

SONG, Youmei et al. Joint optimization of cache placement and request routing in unreliable networks. **Journal of Parallel and Distributed Computing**, v. 157, p. 168–178, 2021. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0743731521001350>. Acesso em: 21 Jul. 2021.

STALLINGS, William. **Foundations of Modern Networking: SDN, NFV, QoE, IoT, and Cloud**. United States: Addison-Wesley Professional, 2015.

TARIG, Faisal et al. A speculative study on 6g. **IEEE Wireless Communications**, v. 27, n. 4, p. 118–125, 2020. Disponível em: <https://ieeexplore.ieee.org/abstract/document/9170653>. Acesso em: 01 Set. 2021.

TIAN, Ye; XU, Kai; ANSARI, Nirwan. Tcp in wireless environments: Problems and solutions. **IEEE Communications Magazine**, v. 43, n. 3, p. S27–S32, 2005. Disponível em: <https://ieeexplore.ieee.org/abstract/document/1404595>. Acesso em: 21 Nov. 2021.

TRAN, Tuyen X. et al. Collaborative mobile edge computing in 5g networks: New paradigms, scenarios, and challenges. **IEEE Communications Magazine**, v. 55, n. 4, p. 54–61, 2017. Disponível em: <https://ieeexplore.ieee.org/abstract/document/7901477>. Acesso em: 08 Set. 2020.

VARGHESE, Blesson et al. A survey on edge performance benchmarking. **ACM Comput. Surv.**, Association for Computing Machinery, New York, NY, USA, v. 54, n. 3, abr. 2021. ISSN 0360-0300. Disponível em: <https://doi.org/10.1145/3444692>. Acesso em: 27 Abr. 2021.

VU, Thang X. et al. **Wireless Edge Caching**: Modeling, analysis, and optimization. United Kingdom: TJ International Ltd, Padstow Cornwall, 2020.

WANG, Jia. A survey of web caching schemes for the internet. **Computer Communication Review**, New York, NY, USA, v. 29, n. 5, p. 36–46, 1999. Disponível em: <https://dl.acm.org/doi/pdf/10.1145/505696.505701>. Acesso em: 10 set. 2021.

WAZLAWICK, Raul Sidnei. Uma reflexão sobre a pesquisa em ciência da computação à luz da classificação das ciências e do método científico. **Revista de Sistemas de Informação da FSMA**, n. 6, p. 3–10, 2010. Disponível em: <https://www.sisbin.ufop.br/wp-content/uploads/2015/03/Reflexao_pesquisa_ciencia_da_computacao.pdf>. Acesso em: 25 Jul. 2020.

WESSELS, Duane. **Web Caching**. United States of America: O'Reilly Media, 2001. Acesso em: 10 nov. 2020.

WOLSEY, Laurence; NEMHAUSER, George. **Integer and Combinatorial Optimization**. United States: Wiley-Interscience, 1988.

WU, Honghai et al. A comprehensive review on edge caching from the perspective of total process: Placement, policy and delivery. **Sensors**, v. 21, n. 15, 2021. Disponível em: <https://www.mdpi.com/1424-8220/21/15/5033>. Acesso em: 25 Jul. 2021.

WU, Zhanji et al. Comprehensive study and comparison on 5g noma schemes. **IEEE Access**, v. 6, p. 18511–18519, 2018. Disponível em: <https://ieeexplore.ieee.org/abstract/document/8319971>. Acesso em: 26 Set. 2021.

XIE, Tian et al. Joint caching and routing in cache networks with arbitrary topology. In: **ICDCS 2022-42nd IEEE International Conference on Distributed Computing Systems**. Bologna, Italy: [s.n.], 2022. Disponível em: <https://sites.psu.edu/nsrg/files/2022/01/caching_and_routing_report.pdf>. Acesso em: 01 Jun. 2022.