Towards MEC-Servers Deployment for Multimedia Streaming over 5G CRAN-Architecture Networks

Carlo Rodrigues, Vladimir Rocha, and Rodrigo A. C. da Silva

***Abstract*—This article evaluates MEC-servers deployment for multimedia on-demand streaming service over 5G cellular networks. The experiments are based on simulations within a CRAN architecture. Three performance metrics are assessed: operational download, piece delay, and channel utilization. Overall, the experiments reveal substantial effectiveness, enhancing QoS and QoE levels. Among the main findings, we highlight that: (i) for a storage capacity of only 5% of the content providers', MEC-servers deployment reduces data volume on the network by up to 85.7%, whereas (ii) for a storage capacity of 30%, all three metrics above reach their most optimized operating values, regardless of video popularity and network latencies. As its main contribution, this research thus provides valuable insights into the evolving landscape of streaming-service projects in 5G cellular networks equipped with MEC servers. At last, conclusions and future work close this article.**

Link to graphical and video abstracts, and to code: https://latamt.ieeer9.org/index.php/transactions/article/view/9023

***Index Terms*—5G networks, MEC, streaming, multimedia.**

I. Introdução

STREAMING sob demanda de objetos multimídia, e.g., filmes, músicas e séries de TV, constitui um dos principais serviços geradores de tráfego em redes celulares 5G. Diante disso, existe a preocupação de uma possível sobrecarga ou eventual exaurimento da capacidade de transmissão de dados das infraestruturas de comunicação móvel. Hoje já ocorre uma competição explícita entre o tráfego do serviço de *streaming* e os tráfegos isolados de voz e dados [1]–[3].

O paradigma *Multi-access Edge Computing* (MEC) é uma das soluções tecnológicas para otimizar a transmissão de dados por *streaming*. A ideia central é levar as capacidades de armazenamento e de processamento para mais próximo dos usuários finais. Essa abordagem permite a implementação de diferentes estratégias assistidas por rede que melhoram os níveis de Qualidade de Serviço (do inglês, *Quality of Service* - QoS) do sistema e de Qualidade de Experiência (do inglês, *Quality of Experience* - QoE) dos usuários [2], [4]–[6].

Uma das estratégias mais promissoras e simples é usar servidores MEC para armazenar objetos multimídia populares, i.e., frequentemente requisitados, reduzindo o número de acessos direto aos provedores de conteúdo localizados remotamente na nuvem. Nesse sentido, servidores MEC devem naturalmente ter definidos em seus projetos o valor ideal da capacidade de armazenamento [3], [4], [7]. Esse cenário leva consequentemente ao seguinte problema de pesquisa: Qual é a capacidade mínima de armazenamento dos servidores MEC, sem comprometimento dos níveis de QoS e QoE?

Ante o exposto, este artigo avalia o emprego de servidores MEC para o serviço de *streaming* sob demanda de objetos multimídia. Os experimentos são baseados em simulações, considerando uma rede celular 5G de arquitetura *Cloud Radio Access Network* (CRAN) [8]–[10]. São utilizadas três métricas de desempenho: *Download Operacional* (D_O); *Atraso da Peça* (A_P); e *Utilização do Canal* (U_C). Como principal contribuição, esta pesquisa provê importantes subsídios que podem servir de referência e guia para projetos de serviço de *streaming* em redes celulares 5G.

O restante deste artigo é organizado como segue. Fundamentos sobre a arquitetura CRAN estão na Seção II. A Seção III discorre sobre trabalhos relacionados. Na Seção IV, tem-se a explicação da operação de servidores MEC para *streaming*. A descrição e a configuração dos experimentos constituem a Seção V. A Seção VI traz os resultados e as análises realizadas. Finalmente, na Seção VII, encontram-se as conclusões gerais e as sugestões de trabalhos futuros.

II. Arquitetura CRAN

A arquitetura CRAN organizada hierarquicamente em camadas é ilustrada na Fig. 1. [8]–[10]. A *Base Station* (BS) de redes celulares tradicionalmente consiste de dois componentes integrados: *Base-Band Unit* (BBU) e *Radio Head* (RH). Todavia, esses dois componentes são fisicamente separados na CRAN. Essa separação permite obter (i) processamento de alta capacidade por meio de um conjunto centralizado de BBUs, fazendo uso de virtualização e computação em nuvem, e (ii) serviço de rádio colaborativo por meio de RHs remotamente localizados, nomeados como *Remote Radio Heads* (RRHs).

A parte da rede que interliga o conjunto de BBUs à *Core Network* (CN) constitui a infraestrutura de *Backhaul* (BF), enquanto que a parte da rede que interliga o conjunto centralizado de BBUs aos RRHs forma a infraestrutura de *Fronthaul* (FH). Os usuários finais se conectam aos RRHs por meio de seus *User Equipments* (UEs), e.g., *smartphones* e *tablets*. Esse entendimento é mostrado na Fig. 1 [1], [11], [12]. Ademais, para facilidade de leitura, a Tabela I traz uma lista dos principais acrônimos e siglas usados neste artigo.

The associate editor coordinating the review of this manuscript and approving it for publication was Carolina Del-Valle-Soto (*Corresponding author: Carlo Rodrigues*).

Carlo Rodrigues, V. Rocha, and R. Cardoso are with the Center for Mathematics, Computation, and Cognition at the Federal University of ABC, Santo André, Brazil (e-mails: carlo.kleber@ufabc.edu.br, vladimir.rocha@ufabc.edu.br, and cardoso.rodrigo@ufabc.edu.br).

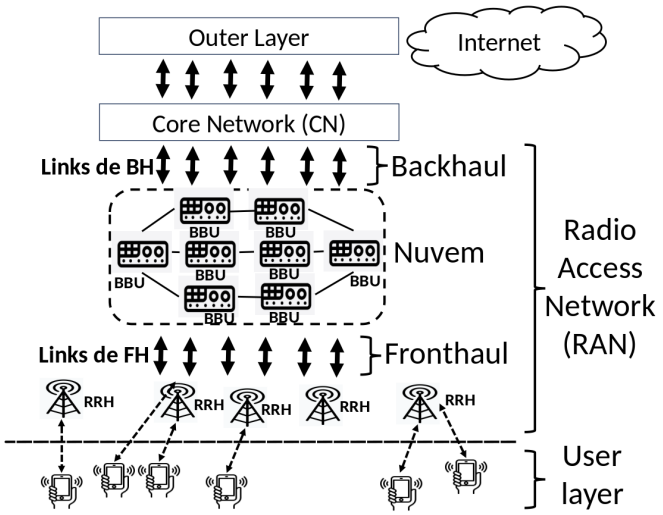


Fig. 1. Arquitetura simplificada da CRAN.

TABELA I
ACRÔNIMOS E SIGLAS

Acrônimo/Sigla	Significado
UE	*User Equipment*
CN	*Core Network*
RAN	*Radio Access Network*
BH	*Backhaul*
FH	*Fronthaul*
BS	*Base station*
QoS	*Quality of Service*
QoE	*Quality of Experience*
MEC	*Multi-access Edge Computing*
CRAN	*Cloud RAN*
BBU	*Base-Band Unit*
RH	*Radio Head*
RRH	*Remote Radio Head*

III. TRABALHOS RELACIONADOS

Trabalhos de pesquisa sobre *streaming* em redes celulares são numerosos e pertencentes a variadas abordagens de mesmo propósito geral: tornar o serviço mais efetivo em termos de QoS e QoE. Ante essa conjuntura, esta seção discorre sobre alguns recentes trabalhos. O objetivo é compartilhar uma sucinta visão geral de algumas recentes abordagens, especialmente evidenciando a diferenciação ou semelhança com este presente trabalho de pesquisa.

Em [13], os autores realizam uma análise comparativa de codecs para transmissão de vídeo em redes 5G. Os experimentos incluem métricas de desempenho relacionadas à taxa de bits e à qualidade da imagem para *streaming* ao vivo. As sequências de teste de vídeo utilizadas nos experimentos são obtidas de diversas fontes, com a característica comum de possuírem definições *High*, *Full* e *Ultra High*. O desempenho dos codecs é testado utilizando bibliotecas de codificação comerciais e *software* de referência dos padrões de codificação. Os resultados experimentais são promissores. É um trabalho voltado eminentemente para técnicas de compressão de vídeo para *streaming* em redes celulares, sendo, portanto, de escopo ortogonal à nossa pesquisa.

Em [14], os autores tratam de *streaming* de vídeo com taxa de bits variável em redes celulares heterogêneas de tecnologia LTE 4G. A questão é abordada como um problema de atribuição e escalonamento, cuja solução reside na teoria de algoritmos gulosos com reconhecimento de QoE. Especificamente, um algoritmo de atribuição guloso vincula os usuários às *femtocells* para baixar os objetos, e o algoritmo de escalonamento guloso aloca intervalos de tempo para os usuários de uma *femtocell*. Os experimentos mostram um desempenho superior, comparativamente a soluções anteriores da literatura. Esse trabalho está no contexto de soluções algorítmicas no nível de aplicação para *streaming*, sendo, como o trabalho em [13], de escopo ortogonal à nossa pesquisa.

Em [15], os autores investigam o posicionamento de conteúdos populares em redes celulares heterogêneas de várias camadas, em que são criados *clusters* de entrega e posicionamento de segmentos de conteúdos para atender solicitações de vídeo correlacionadas. Os autores modelam a decisão de posicionamento de conteúdo usando a formulação do *problema da mochila* e derivam uma estratégia de solução matemática exata para resolvê-lo. O desempenho da estratégia é avaliado usando simulações. É um trabalho, portanto, no contexto de soluções baseadas em servidores MEC para *streaming* em redes celulares. A principal diferenciação em relação à nossa pesquisa é que armazenamos objetos inteiros (e não apenas segmentos) nos servidores MEC e, ainda, discutimos a possível mudança de popularidade desses objetos (vide Seção IV).

Em [16], os autores propõem um esquema de *streaming* de vídeo adaptativo orientado à QoE baseado na teoria de redes do tipo *dueling deep Q-learning*. Esse esquema otimiza o nível de QoE ao considerar conjuntamente a largura de banda de transmissão da camada física e o estado do *buffer* da camada superior. A partir dos resultados numéricos calculados e da técnica de prototipagem, é mostrado que o esquema proposto supera os esquemas existentes, com melhorias médias de QoE de 12.6% a 28.8%. É um trabalho, portanto, no contexto de soluções baseadas em aprendizagem profunda e redes neurais para *streaming* em redes celulares, sendo, como os trabalhos em [13], [14], de escopo ortogonal à nossa pesquisa.

Em [3], os autores abordam o serviço de *streaming* investigando o controle de admissão do usuário e da seleção da taxa de bits de vídeo nas redes de dados móveis de geração 5G e seguintes. É proposto um algoritmo exato para identificar taxas de bits viáveis e seus requisitos mínimos de desempenho quando as BSs, que utilizam servidores MEC isolados, armazenam um subconjunto conhecido de segmentos do vídeo, potencialmente codificados em diferentes resoluções. Os experimentos são baseados em simulações. Como o trabalho em [15], esse trabalho também pertence ao contexto de soluções baseadas em servidores MEC para *streaming*. A principal diferenciação em relação à nossa pesquisa é que armazenamos objetos inteiros (e não apenas segmentos) nos servidores MEC, os quais trabalham de forma cooperativa (vide Seção IV).

Por último, em [4], é proposto um esquema para *streaming* adaptativo com HTTP na borda de redes celulares com uso de servidores MEC sob operação isolada. São analisadas diferentes políticas de pré-busca de segmento, considerando diversas abordagens e técnicas, e.g., solicitações de segmento anteriores, conversão de segmento (i.e., redução da taxa de bits

do segmento), modelo de previsão de Markov e aprendizado de máquina para prever solicitações de segmentos. Como em [3], [15], esse trabalho também se insere no conjunto de soluções baseadas no emprego de servidores MEC. A principal diferenciação em relação à nossa pesquisa é que armazenamos objetos inteiros (e não apenas segmentos) nos servidores MEC, os quais operam cooperativamente, além de discutirmos a mudança da popularidade dos objetos (vide Seção IV).

Ante o discorrido acima, esta presente pesquisa se destaca principalmente por utilizar servidores MEC nos quais objetos inteiros são replicados próximos aos roteadores de encaminhamento de tráfego. A principal vantagem do armazenamento do objeto inteiro é a manutenção da mesma conexão do canal utilizado pelo usuário com o MEC para transmissão de dados, não havendo a necessidade de eventuais reconexões sucessivas para transmissão de segmentos individuais. A principal desvantagem é a necessidade de mais espaço de armazenamento. Comparativamente a soluções anteriores da literatura, que também consideram o armazenamento de objetos inteiros (e.g, CDN, Proxy Cache, entre outros), tem-se que nelas os objetos inteiros são replicados em pontos estratégicos centralizados da rede, e não de forma espalhada sobre a área geográfica atendida, como proposto nesta pesquisa [17].

Para encerrar esta seção, deve-se ainda esclarecer que, salvo melhor juízo, não há trabalhos anteriores na literatura sob MEC que apresentem propostas semelhantes a esta pequisa e, consequentemente, diretamente competidoras no contexto de *streaming* multimídia sob demanda. Análises comparativas com outras propostas da literatura de mesmo propósito geral, bem como avaliação de custo de implementação/operação, são deixadas como trabalhos futuros [18].

IV. Servidores MEC: Operação sob Streaming

Considere a Fig. 1. Os servidores MEC são posicionados na infraestrutura de FH na camada RAN. Esses servidores são equipados com memória secundária para armazenar objetos multimídia populares dos provedores de conteúdos na camada *Outer Layer*. Os servidores têm interface de rede com fio para se conectarem à infraestrutura de FH. Cada servidor atende um conjunto exclusivo de RRHs.

Para ingressar na rede, os usuários, localizados na camada *User Layer*, se conectam aos servidores MEC por meio de RRHs. Assuma que um usuário faz uma requisição para um objeto multimídia. O procedimento a seguir é então executado.

- 1) É verificado se o objeto está armazenado no servidor MEC *local*, ou seja, no servidor MEC que atende o RRH desse usuário. Em caso positivo, o objeto é enviado por *streaming*; senão, tem-se o passo seguinte.
- 2) É verificado se o objeto está armazenado em algum servidor MEC *remoto* da rede, ou seja, um servidor distinto do servidor *local* do usuário requisitante. Em caso positivo, o servidor MEC *remoto* transmite o objeto para o servidor MEC *local*. Em seguida, este servidor envia por *streaming* o objeto para o usuário requisitante. Se houver espaço disponível, o objeto é armazenado neste servidor para atendimento de futuras requisições por esse objeto. Em caso negativo, tem-se o passo seguinte.
- 3) É encaminhada a requisição do objeto para o provedor de conteúdos, por meio do conjunto centralizado de BBUs e da CN. O provedor de conteúdos então transmite o objeto para o servidor MEC *local* do usuário. Por fim, o servidor MEC *local* envia por *streaming* o objeto para o usuário. Se houver espaço disponível, o objeto é armazenado no servidor *local* visando atender futuras requisições por esse mesmo objeto.

A gerência da informação sobre quais objetos multimídia estão armazenados nos servidores MEC é feita da seguinte forma. Cada servidor MEC possui uma lista localmente armazenada. Para cada objeto multimídia disponível no catálogo de títulos do provedor de conteúdos, essa lista informa quais são os servidores MEC constituintes da rede que o possuem.

No início da operação, a lista de cada servidor MEC está vazia, pois nenhum objeto está armazenado. Conforme os usuários ingressem na rede e façam requisições, os objetos vão sendo sendo armazenados localmente nos servidores MEC. Ao ter um objeto armazenando, o servidor MEC atualiza sua lista local e a envia em *broadcast* para que todos os demais servidores MEC da rede atualizem suas respectivas listas.

No estado estacionário, os objetos mais populares (i.e., mais requisitados) devem estar armazenados localmente nos servidores MEC. Quanto maior é a capacidade de armazenamento local, menor é o tráfego de dados entre as camadas *Outer Layer*, CN e RAN (vide Fig. 1), diminuindo a probabilidade de sobrecarga ou exaurimento da capacidade de transmissão de dados da rede celular. Sob um cenário em que as popularidades dos objetos são alteradas dinamicamente e a capacidade de armazenamento é limitada, deve-se empregar uma política de substituição dos objetos armazenados nos servidores MEC. Esse estudo específico não é tratado neste presente trabalho de pesquisa, sendo deixado como trabalhos futuros.

V. Experimentos: Descrição e Configuração

A. Ambiente de Simulação e Modelagem da Rede

O modelo de simulação é desenvolvido em Java como um módulo do ambiente de simulação PeerSim [19], em cuja configuração define-se a quantidade de usuários e considera-se o término da simulação quando todos os usuários recebem o objeto requisitado. Os experimentos são conduzidos no mesmo ambiente utilizando uma única plataforma de *hardware*, constituída por um processador Intel Core i7 (2.6 GHz), 24 GB de RAM e sistema operacional Linux. Os resultados da simulação têm intervalos de confiança de 95% que estão dentro do limite de 5% dos valores médios reportados, tendo sido realizadas 30 execuções (rodadas).

A rede considerada nos experimentos é inspirada nos trabalhos de [20], [21], com dimensionamentos geográfico e demográfico baseados em uma região limitada do bairro Morumbi da cidade de São Paulo, SP, no Brasil. Esse bairro possui uma área total de $11.4\,\text{km}^2$, com uma densidade demográfica de 2832 hab/km^2 [22].

A topologia da rede é definida por um anel de nós r_t, para $t = 1, 2, \ldots, 7$, os quais estão conectados entre si por enlaces bidirecionais de capacidade individual de transmissão $10\,\text{Gbps}$ (em cada direção). Cada nó r_t se constitui em um

roteador de tráfego ao qual se conectam três RRHs via enlaces bidirecionais (um para cada RRH) de capacidade individual de transmissão 1 Gbps (em cada direção). Além disso, cada nó r_i também está conectado a uma única e exclusiva BBU que, por sua vez, se conecta à CN. Cada RRH tem associado um único servidor MEC. Os usuários se conectam individualmente à rede por meio dos RRHs.

Nos experimentos são utilizadas duas latências distintas para a transmissão de dados entre os componentes do sistema (i.e., entre a CN e a BBU, entre a BBU e o RRH, e entre os roteadores adjacentes no anel). Como modelado em [23], as duas latências são: *slow* – 140 ms, correspondendo a conexões lentas e com um alto congestionamento, normalmente associadas a componentes que estão situadas em diferentes áreas geográficas; e *fast* – 40 ms, correspondendo a conexões rápidas e quase sem congestionamento, normalmente associadas a componentes que estão situadas na mesma área geográfica. A utilização de diferentes latências permite, e.g., inferir o nível de otimização que pode ser alcançado devido à evolução da tecnologia 5G, considerando aspectos de *hardware* e *software*.

A Tabela II recapitula os valores numéricos já informados, além de incluir outros parâmetros para uma maior caracterização operacional do sistema. Por fim, a visualização simplificada da topologia é mostrada na Fig. 2. Por simplicidade, as conexões (enlaces) entre roteadores e BBU, bem como entre RRHs e roteadores/usuários/MECs, não aparecem na totalidade, sendo ilustradas em apenas dois roteadores do anel.

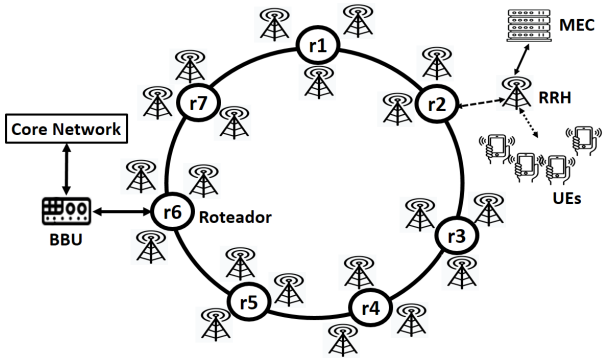


Fig. 2. Topologia simplificada da rede.

TABELA II
CARACTERIZAÇÃO OPERACIONAL DA TOPOLOGIA DE REDE

Definição	Valor
Área de abrangência geográfica da topologia.	4 km^2
Alcance (raio) de cobertura rádio de cada RHH.	500 m
Capacidade de armazenamento do servidor MEC.	Variável
Distância entre RHH e servidor MEC associado.	20 m
Distância entre roteadores adjacentes no anel.	200 m
Distância entre pares de RRHs de um mesmo servidor MEC.	30 m
Capacidade de transmissão do enlace, em cada direção, entre roteadores adjacentes no anel.	10 Gbps
Capacidade de transmissão do enlace, em cada direção, entre RRH e roteador associado.	1 Gbps
Capacidade de transmissão do enlace, em cada direção, entre RRH e UE.	0.5 Gbps

B. Usuários e Provedor de Conteúdos

A caracterização informada a seguir se baseia parcialmente em trabalhos anteriores da literatura, além de informações e estatísticas relacionadas ao serviço de *streaming* e de telefonia móvel comercial (e.g., [24]–[26]).

Para os experimentos são estimados $V = 10195$ usuários distribuídos de forma uniforme entre os RRHs existentes na rede, em consonância com a densidade demográfica e uso de telefonia móvel da região sob análise. O processo de chegada (i.e., ingresso na rede) de usuários segue uma distribuição de Poisson com parâmetro $1/\lambda$ = 1 s. Após seu ingresso, o usuário se conecta a um RRH, por meio de seu UE, e faz a requisição por um objeto multimídia. Em seguida, o usuário passa a se movimentar dentro da área de cobertura do RRH (vide Tabela II) durante todo o tempo de recebimento do objeto por *streaming*. Note que não há, portanto, *handovers* a serem analisados, sendo isso deixado como trabalhos futuros.

A movimentação de cada usuário ocorre de forma independente e aleatória, sendo livre de obstáculos, com velocidade 0–2 m/s e direção 0–360 °. Após o recebimento do objeto requisitado, os usuários deixam a rede. Daí, a simulação termina imediatamente após o último usuário do sistema ter recebido todo o objeto requisitado.

A escolha do objeto a ser requisitado pelo usuário é modelada por uma distribuição Zipf, como explicado a seguir. A probabilidade de escolher um objeto de categoria i é dada por $\frac{1}{i^{1-z}\sum_{j=1}^{K}\frac{1}{j^{1-z}}}$, onde K é o número total de objetos no servidor, e z é o parâmetro *skew factor*. Quanto maior é o valor de z, maior é a probabilidade de escolha dos objetos das primeiras (iniciais) categorias, i.e., passa a haver uma maior concentração de acessos aos objetos das primeiras categorias [26].

A popularidade de um objeto pode, portanto, ser associada à sua categoria i. Os objetos mais populares pertencem às primeiras categorias, enquanto os objetos menos populares pertencem às últimas categorias. Nos experimentos, são então usadas três distribuições Zipf, diferenciadas pelo valor do parâmetro z, para modelar o processo de escolha dos objetos por parte dos usuários, conforme detalhado no que segue.

Para modelar uma *baixa* concentração de acessos, i.e., objetos com frequências semelhantes de acesso, assume-se $z = 0.2$. Para modelar uma *média* concentração de acessos, i.e., os objetos das primeiras categorias com maiores frequências de acesso, assume-se $z = 0.6$. Por fim, para modelar uma *alta* concentração de acessos, i.e., os objetos das primeiras categorias com bem maiores frequências de acesso, assume-se $z = 1.0$. Essa modelagem propicia uma avaliação mais realista da eficiência do emprego de servidores MEC, tendo em vista à possível variação da popularidade dos objetos no provedor de conteúdos [26].

Especificamente sobre o provedor de conteúdos, tem-se o seguinte. Os objetos considerados nos experimentos são exclusivamente do tipo vídeo (filmes), sendo os tipos música e séries de TV deixados como trabalhos futuros. Há um conjunto de 1512 vídeos no total, que corresponde ao catálogo de títulos disponíveis para acesso dos usuários. Cada vídeo possui

1 GB de tamanho, dividido em peças de 1 MB cada. Daí, a capacidade total de armazenamento do provedor é de 1512 GB.

C. Métricas de Desempenho

Nos experimentos, são avaliadas as três métricas a seguir.

- *Download Operacional* (D_O). Estima a taxa média do usuário, em Mbps, para receber as peças do vídeo. A métrica é calculada como a razão entre a quantidade de bytes referente às peças baixadas e o tempo consumido entre o pedido e o recebimento destas. Essa métrica está relacionada à QoS do sistema e permite, e.g., ajustar a taxa de codificação do vídeo. Quanto maior é D_O, mais alto é o nível de QoS. Esta métrica permite estimar indiretamente a redução do *volume de dados* transmitido entre as camadas *Outer Layer*, CN e RAN.
- *Atraso da Peça* (A_P). Estima o tempo médio do usuário, em segundos, para receber as peças do vídeo. A métrica é calculada como a média do tempo consumido entre o pedido da peça e o recebimento desta, para cada uma das peças baixadas. Esta métrica está relacionada ao QoE do sistema e permite, e.g., definir o tamanho de *buffers* locais para mitigar descontinuidades na visualização do vídeo. Quanto menor é A_P, mais alto é o nível de QoE.
- *Utilização do Canal* (U_C). Estima a utilização do enlace RRH-UE, em percentual, para receber as peças de vídeo. A métrica é calculada como a razão entre D_O e a capacidade do enlace do usuário. Esta métrica está relacionada à QoS do sistema e permite, e.g., ajustar a taxa de codificação do objeto transmitido. Quanto maior é U_C, menor é a subutilização do enlace e, portanto, mais alto é o nível de QoS.

VI. Avaliação de Desempenho

A. Organização

Os resultados dos experimentos estão nos gráficos das figuras desta seção, a saber, Fig. 3 (latência *slow*) e Fig. 4 (latência *fast*). O eixo y corresponde à métrica sob análise, enquanto o eixo x informa a capacidade individual de armazenamento dos servidores MEC da rede, denotada por T_M e medida em percentual da capacidade total de armazenamento do provedor de conteúdos. Em cada figura, tem-se resultados para três distintas situações de operação, cujos entendimentos são explicados a seguir.

1) MyRRH-MEC. É a operação em que as requisições de peças do vídeo são atendidas exclusivamente pelo servidor MEC *local* (via RRH) do usuário requisitante. Define o valor do melhor caso (i.e., valor ótimo) de operação do sistema para as três métricas de desempenho.

2) OtherRRH-CN. É a operação em que as requisições de peças são atendidas pelo provedor de conteúdos via CN/BBU ou servidores MEC *remotos*. Define o valor do pior caso para as três métricas de desempenho.

3) *General*. É a operação em que, dependendo de onde o vídeo está armazenado, as peças podem ser obtidas do servidor MEC *local*, do servidor MEC *remoto* ou, ainda, do provedor de conteúdos. São apresentados três conjuntos de resultados diferenciados pela concentração de acessos.

B. Análise dos Resultados

As Figs. 3a e 4a trazem os resultados obtidos para a métrica D_O. Considerando a análise dos conjuntos de valores da operação *General*, temos pontualmente o que segue.

1) Conforme T_M cresce, maior se torna o valor de D_O (i.e., maior QoS), indo do pior valor (operação OtherRRH-CN) para o melhor valor (operação MyRRH-CN). Note que, quanto maior é o valor de T_M, maior é a probabilidade de o vídeo requisitado estar armazenado no servidor MEC *local*.

2) Para que todas as requisições sejam atendidas pelo servidor MEC *local*, o valor mínimo de T_M é função da concentração de acessos. Para *baixa* concentração, $T_M \geq 30\%$; para *média* concentração, $T_M \geq 25\%$; e para *alta* concentração, $T_M \geq 15\%$. Isso ocorre porque, quanto maior é a concentração, menos vídeos diferentes precisam ser armazenados.

3) O valor de D_O sob latência *fast* é cerca de duas vezes maior que sob latência *slow*, chegando mais próximo da capacidade de transmissão do enlace RRH-UE, que é de 0.5 Gbps. Isso ocorre porque, quanto menor é a latência, mais peças são requisitadas e baixadas.

4) O uso de servidores MEC reduz o *volume de dados* (estimado indiretamente por D_O) transmitido entre as camadas *Outer Layer*, CN e RAN. Para $0\% < T_M < 30\%$, as reduções já são bem significativas. Por exemplo, para $T_M = 5\%$ e *alta* concentração de acessos, tem-se uma redução de até 83.9% (latência *slow*) e 85.5% (latência *fast*), como indicado na Tabela III. Por outro lado, para $T_M \geq 30\%$, todas as requisições são atendidas a partir de servidores MEC *locais*.

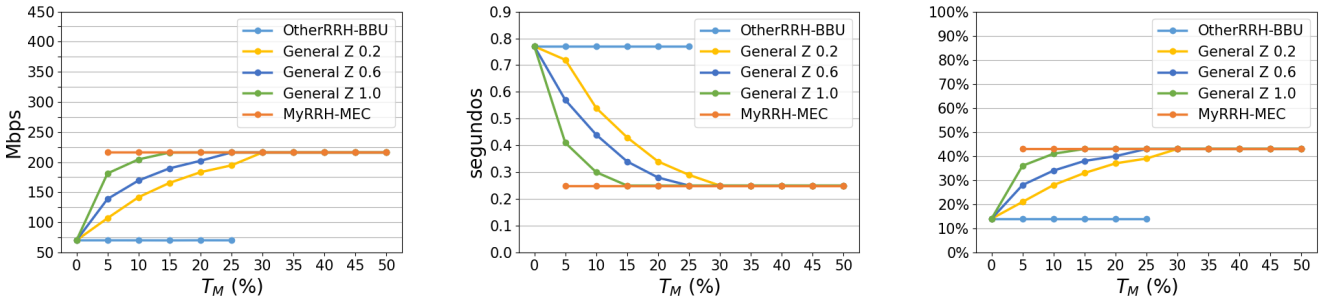
5) O uso de servidores MEC permite atingir adequados níveis de codificação de vídeo em plataformas comercias de *streaming*, mesmo em redes de latências não desprezíveis. Por exemplo, para operação sob uma configuração padrão preconizada pela plataforma YouTube [27] e sob latência *slow*, temos que: para codificações de 4k (i.e, entre 66 e 85 Mbps), então $T_M \geq 5\%$; e para codificações de 8k (i.e., entre 150 e 300 Mbps), então $T_M \geq 15\%$.

Com relação aos resultados obtidos para as métricas A_p (Figs. 3b e 4b) e U_C (Figs. 3c e 4c), chegamos a observações e constatações alinhadas com aquelas já discutidas para a métrica D_O, como destacado brevemente a seguir.

1) O aumento de T_M otimiza os valores de A_P e U_C. A justificativa é a mesma dada na análise de D_O: quanto maior é o valor de T_M, maior é a probabilidade de que o vídeo requisitado já esteja armazenado no servidor MEC *local*, favorecendo a redução de A_P (i.e., maior QoE) e o aumento de U_C (i.e., maior QoS).

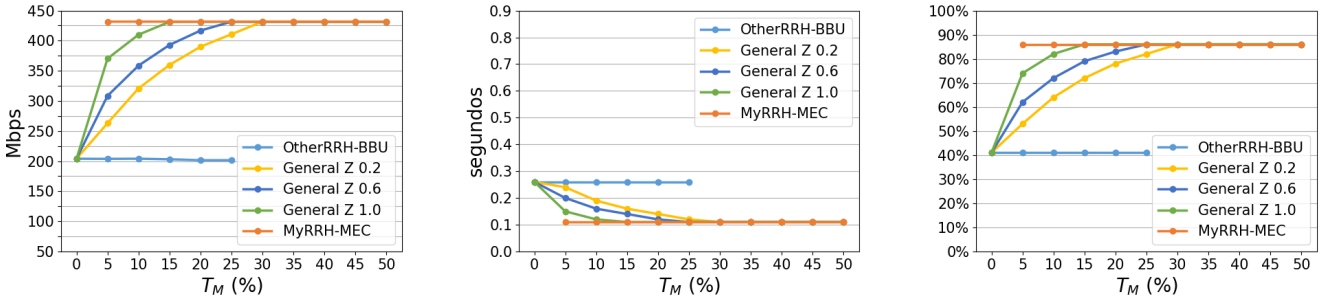
Por exemplo, observando-se A_P, com $T_M = 5\%$ e *alta* concentração de acessos, tem-se uma redução de até 48.1% (latência *slow*) e 46.2% (latência *fast*), como indicado na Tabela III. No caso de U_C, com $T_M = 5\%$ e *alta* concentração de acessos, tem-se um aumento de até 2.85 vezes (latência *slow*) e 1.85 vezes (latência *fast*), como mostrado na Tabela III.

2) Sobre o valor mínimo de T_M para obter os valores mais otimizados possíveis de A_P e U_C, tem-se a mesma constatação da análise anterior de D_O: esse valor é função da concentração de acessos aos vídeos. Ademais, os valores mais otimizados de A_P e U_C são obtidos sob latência *fast*.



(a) *Download Operacional* (D_O). (b) *Atraso da Peça* (A_P) (c) *Utilização do Canal* (U_C)

Fig. 3. Resultados das métricas a) Download Operacional, b) Atraso da peça, e c) Utilização do Canal para latência *slow*.



(a) *Download Operacional* (D_O). (b) *Atraso da Peça* (A_P) (c) *Utilização do Canal* (U_C)

Fig. 4. Resultados das métricas a) Download Operacional, b) Atraso da peça, e c) Utilização do Canal para latência *fast*.

TABELA III
REDUÇÃO DE VOLUME DE DADOS, REDUÇÃO DE A_P, E AUMENTO DE U_C, TODOS PARA $T_M = 5\%$

	VOLUME DE DADOS		A_P		U_C	
Concentração de acessos	*slow* (140 ms)	*fast* (40 ms)	*slow* (140 ms)	*fast* (40 ms)	*slow* (140 ms)	*fast* (40 ms)
Baixa ($z = 0.2$)	49.5%	60.9%	7.8%	11.5%	1.53x	1.31x
Média ($z = 0.6$)	64.3%	71.4%	25.9%	23.1%	2.15x	1.55x
Alta ($z = 1.0$)	**83.9%**	**85.7%**	**48.1%**	**46.2%**	**2.85x**	**1.85x**

VII. CONCLUSÕES E TRABALHOS FUTUROS

Este artigo avaliou o emprego de servidores MEC para o serviço de *streaming* sob demanda de objetos multimídia em redes celulares 5G. Os experimentos foram baseados em simulação, considerando uma arquitetura de rede CRAN.

Os experimentos revelaram uma substancial efetividade sistêmica atingida pelo uso de servidores MEC, incluindo o aumento dos níveis de QoS e de QoE. Em destaque, concluímos que: (i) com uma capacidade de armazenamento de apenas 5% daquela do provedor de conteúdos, os servidores MEC otimizam o volume de dados na rede, o atraso das peças de vídeo e a utilização do canal do usuário em até 85.7%, 48.1% e 2.85 vezes, respectivamente; e (ii) com uma capacidade de armazenamento de 30%, as três métricas citadas já atingem seus valores mais otimizados possíveis de operação, independentemente da popularidade dos vídeos e das latências de transmissão de dados na rede. Neste contexto, a principal contribuição desta pesquisa é o provimento de subsídios para realização de projetos de *streaming* em redes celulares 5G.

Por fim, como trabalhos futuros e conscientes das limitações desta pesquisa, sugerimos: (i) avaliar comportamentos mais complexos de acesso aos objetos multimídia, definindo sessões em que um mesmo usuário faz a requisição de diferentes tipos de objetos multimídia e executa ações de interatividade; (ii) avaliar diferentes políticas de substituições dos objetos armazenados nos servidores MEC, prevendo que as popularidades dos objetos mudam dinamicamente e que o espaço de armazenamento do MEC deve ser utilizado de forma otimizada; (iii) avaliar topologias de redes em áreas mais extensas além de outras propostas da literatura de semelhante propósito, incluindo a ocorrência de *handovers*; por último, (iv) analisar o sistema de *streaming* multimídia sob demanda em termos do custo necessário para implantação e operacionalização de servidores MEC.

REFERÊNCIAS

[1] Y. Li, H. Ma, L. Wang, S. Mao, and G. Wang, "Optimized Content Caching and User Association for Edge Computing in Densely Deployed Heterogeneous Networks," *IEEE Transactions on Mobile Computing*, vol. 21, no. 6, pp. 2130–2142, 2022. doi:10.1109/TMC.2020.3033563.

[2] P. Lin, Z. Ning, Z. Zhang, Y. Liu, F. R. Yu, and V. C. M. Leung, "Joint Optimization of Preference-Aware Caching and Content Migration in Cost-Efficient Mobile Edge Networks," *IEEE Transactions on Wireless Communications*, vol. 23, no. 5, pp. 4918–4931, 2024. doi:10.1109/TWC.2023.3323464.

[3] D. Xenakis, A. Tsiota, and N. Passas, "Admission control and end-to-end slicing for video streaming in MEC-empowered cellular networks," in *GLOBECOM 2022 - 2022 IEEE Global Communications Conference*, pp. 3423–3428, 2022. doi:10.1109/GLOBECOM48099.2022.10001504.

[4] J. Aguilar-Armijo, C. Timmerer, and H. Hellwagner, "SPACE: Segment Prefetching and Caching at the Edge for Adaptive Video Streaming," *IEEE Access*, vol. 11, pp. 21783–21798, 2023. doi:10.1109/ACCESS.2023.3252365.

[5] A. Sarah, G. Nencioni, and M. M. I. Khan, "Resource Allocation in Multi-access Edge Computing for 5G-and-beyond networks," *Computer Networks*, vol. 227, p. 109720, 2023. doi:10.1016/j.comnet.2023.109720.

[6] B. Liang, M. A. Gregory, and S. Li, "Multi-access edge computing fundamentals, services, enablers and challenges: A complete survey," *Journal of Network and Computer Applications*, vol. 199, p. 103308, 2022. doi:10.1016/j.jnca.2021.103308.

[7] Y. Chen, Y. Cai, H. Zheng, J. Hu, and J. Li, "Cooperative caching for scalable video coding using value-decomposed dimensional networks," *China Communications*, vol. 19, no. 9, pp. 146–161, 2022. doi:10.23919/JCC.2022.00.006.

[8] M. Pattaranantakul, C. Vorakulpipat, and T. Takahashi, "Service Function Chaining security survey: Addressing security challenges and threats," *Computer Networks*, vol. 221, p. 109484, 2023. doi:10.1016/j.comnet.2022.109484.

[9] F. Zanferrari Morais, C. André da Costa, A. M. Alberti, C. Bonato Both, and R. da Rosa Righi, "When SDN meets C-RAN: A survey exploring multi-point coordination, interference, and performance," *Journal of Network and Computer Applications*, vol. 162, p. 102655, 2020. doi:10.1016/j.jnca.2020.102655.

[10] R. T. Rodoshi, T. Kim, and W. Choi, "Fuzzy Logic and Accelerated Reinforcement Learning-Based User Association for Dense C-RANs," *IEEE Access*, vol. 9, pp. 117910–117924, 2021. doi:10.1109/ACCESS.2021.3107325.

[11] L. B. Silveira, H. C. de Resende, C. B. Both, J. M. Marquez-Barja, B. Silvestre, and K. V. Cardoso, "Tutorial on communication between access networks and the 5G core," *Computer Networks*, vol. 216, p. 109301, 2022. doi:10.1016/j.comnet.2022.109301.

[12] V. S. Pana, O. P. Babalola, and V. Balyan, "5G radio access networks: A survey," *Array*, vol. 14, p. 100170, 2022. doi:10.1016/j.array.2022.100170.

[13] G. Minopoulos, K. E. Psannis, G. Kokkonis, and Y. Ishibashi, "QoE Assessment of Video Codecs for Video Streaming over 5G Networks," in *2020 3rd World Symposium on Communication Engineering (WSCE)*, pp. 34–38, 2020. doi:10.1109/WSCE51339.2020.9275576.

[14] A. Kulkarni and A. Seetharam, "QoE-aware Video Streaming in Heterogeneous Cellular Networks," in *2021 IEEE 18th Annual Consumer Communications & Networking Conference (CCNC)*, pp. 1–2, 2021. doi:10.1109/CCNC49032.2021.9369556.

[15] T. M. Ayenew, D. Xenakis, N. Passas, and L. Merakos, "Cooperative Content Caching in MEC-Enabled Heterogeneous Cellular Networks," *IEEE Access*, vol. 9, pp. 98883–98903, 2021. doi:10.1109/ACCESS.2021.3095356.

[16] J. Yu, H. Wen, G. Pan, S. Zhang, X. Chen, and S. Xu, "Quality of Experience Oriented Adaptive Video Streaming for Edge Assisted Cellular Networks," *IEEE Wireless Communications Letters*, vol. 11, no. 11, pp. 2305–2309, 2022. doi:10.1109/LWC.2022.3200830.

[17] A. L. S. de Moraes, D. D. de Macedo, and L. Pioli, "Video streaming on fog and edge computing layers: A systematic mapping study," *Internet of Things*, vol. 28, p. 101359, 2024. doi:10.1016/j.iot.2024.101359.

[18] N.-N. Dao, N. H. Tu, T.-D. Hoang, T.-H. Nguyen, L. V. Nguyen, K. Lee, and Laihyuk, "A review on new technologies in 3GPP standards for 5G access and beyond," *Computer Networks*, vol. 245, p. 110370, 2024. doi:10.1016/j.comnet.2024.110370.

[19] A. Montresor and M. Jelasity, "PeerSim: A scalable P2P simulator," in *2009 IEEE Ninth International Conference on Peer-to-Peer Computing*, pp. 99–100, 2009. doi:10.1109/P2P.2009.5284506.

[20] N. Molner, A. de la Oliva, I. Stavrakakis, and A. Azcorra, "Optimization of an integrated fronthaul/backhaul network under path and delay constraints," *Ad Hoc Networks*, vol. 83, pp. 41–54, 2019. doi:10.1016/j.adhoc.2018.08.025.

[21] B. Naudts, M. Kind, S. Verbrugge, D. Colle, and M. Pickavet, "How can a mobile service provider reduce costs with software-defined networking?," *International Journal of Network Management*, vol. 26, no. 1, pp. 56–72, 2016. doi:10.1002/nem.1919.

[22] WIKIPÉDIA, "Morumbi (distrito de São Paulo)," 2023. [ONLINE]. Available at: https://pt.wikipedia.org/wiki/Morumbi_(distrito_de_Sao_Paulo). Accessed on: May 5th, 2024.

[23] A. Narayanan, E. Ramadan, J. Carpenter, Q. Liu, Y. Liu, F. Qian, and Z.-L. Zhang, "A First Look at Commercial 5G Performance on Smartphones," in *Proceedings of The Web Conference 2020*, p. 894–905, 2020. doi:10.1145/3366423.3380169.

[24] J. Stoll, "Countries with most content available on netflix worldwide as of march 2023." White Paper, 2023. [ONLINE]. Available at: https://www.statista.com/statistics/1013571/netflix-library-size-worldwide/. Accessed on: May 5th, 2024.

[25] USP, "Nove em cada dez indivíduos utilizam o celular para acessar a internet em São Paulo." Jornal da USP, 2023. [ONLINE]. Available at: https://jornal.usp.br/radio-usp/nove-em-cada-10-individuos-utilizam-o-celular-para-acessar-a-internet-em-sao-paulo/. Accessed on: May 5th, 2024.

[26] T. Wang, C. Jayasundara, M. Zukerman, A. Nirmalathas, E. Wong, C. Ranaweera, C. Xing, and B. Moran, "Estimating Video Popularity From Past Request Arrival Times in a VoD System," *IEEE Access*, vol. 8, pp. 19934–19947, 2020. doi:10.1109/ACCESS.2020.2966495.

[27] YouTube Help, "Choose live encoder settings, bitrates, and resolutions," 2021. [ONLINE]. Available at: https://support.google.com/youtube/answer/2853702. Accessed on: May 5th, 2024.

Carlo K. da S. Rodrigues received the D.Sc. degree in Systems and Computer Engineering from the Federal University of Rio de Janeiro in 2006, the M.Sc. degree in Systems and Computing from the Military Institute of Engineering in 2000. He is currently a professor in the Center for Mathematics, Computation, and Cognition at the Federal University of ABC, working in the subarea of Computer Networks. http://lattes.cnpq.br/4860474255962383.

Vladimir Rocha holds a Ph.D.'s degree in Computer Engineering and master's degree in Computer Science both at the University of São Paulo. Specialist in Peer-to-Peer and Cloud Computing technologies, works as a professor at the Center for Mathematics, Computation and Cognition department of the Federal University of ABC - Brazil. His research interests include scalability and performance in distributed structures and systems, such as BitTorrent, DHT and Blockchain. http://lattes.cnpq.br/1225632464417175.

Rodrigo A. C. da Silva received the Ph.D. and the M.Sc. degrees in Computer Science from the Institute of Computing, University of Campinas, Brazil, in 2022 and 2015, respectively. He is currently a professor in the Center for Mathematics, Computation, and Cognition at the Federal University of ABC. His current research interests include computer networks, UAV communications, energy efficiency, fog and cloud computing. http://lattes.cnpq.br/8593993681646906