Uma Proposta Inicial baseada em *Mobile Edge Computing* para Orquestrar *Caches* em Redes 5G

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

[1]Programa de Pós Graduação de Computação Aplicada - PPGCA
Universidade do Estado de Santa Catarina - UDESC - Joinville - SC - Brasil

marisangila.alves@gmail.com, guilherme.koslovski@udesc.br

Resumo. *A quinta geração de redes móveis - 5G - é uma tecnologia em evolução e, portanto, existem questões abertas e um número substancial de propostas que visam atender seus requisitos. Dessa forma, pesquisas sugerem a dissipação de servidores cache na periferia da Internet com a arquitetura distribuída proposta pelo paradigma de Mobile Edge Computing. Esse artigo descreve uma proposta preliminar para otimizar o uso servidores cache e diminuir a latência, requisito da 5G. Com base na ponderação de latência e recursos do servidor, o algoritmo decide entre buscar o conteúdo na sua origem ou redirecionar o usuário para outro servidor cache presente na rede.*

1. Introdução

Serviços e aplicações hospedados e fornecidos por computação em nuvem têm gerado um aumento significativo de tráfego na rede. Além disso, o número de dispositivos móveis está em crescimento, com isso se tornam a principal forma de acesso a Internet para usuários finais [Rocha et al. 2016]. Provedores de serviços disponibilizam seu conteúdo replicado através de múltiplos servidores para reduzir a latência percebida por seus usuários e, o congestionamento no núcleo da rede. Esses servidores podem ser geograficamente espalhados em diversas localizações pelo mundo, e.g. servidores de *Content Delivery Network* (CDN), ou genericamente, servidores de *cache*.

Ainda, nos últimos anos obteve-se o aumento de dispositivos móveis com elevada capacidade de processamento, consequentemente aumentando os requisitos de *Quality-of-Service* (QoS) e *Quality-of-Experience* (QoE) [Iamnitchi et al. 2015]. Em outras palavras, as redes móveis se popularizam. Atualmente a quinta geração – 5G, resultado dos esforços de trabalho do *International Telecommunication Union* (ITU) e dos principais grupos de pesquisa envolvidos (3GPP[1] e 5GPPP[2]), está sendo implantada em alguns países [Parvez et al. 2018] e, consequentemente acarreta em novos desafios.

Soluções para redução de latência em 5G, sugerem o paradigma de *Mobile Edge Computing* (MEC) [Abbas et al. 2018]. Esse paradigma resume-se em aproximar o poder computacional a borda da rede, especificamente na *Radio Access Network* (RAN), ou seja, próximo ao usuário [Hu et al. 2015]. A rede de acesso via rádio pode ser definida como o meio responsável pela conexão entre usuários móveis e o centro da rede [Abbas et al. 2018].

[1]https://www.3gpp.org/
[2]https://5g-ppp.eu/

Diante do exposto, nesse artigo é abordado o problema da minimização de latência *End-to-End* (E2E) em redes 5G. Portanto, o objetivo deste trabalho é apresentar uma proposta preliminar de um algoritmo de gerenciamento de *cache* distribuído dentro da RAN com infraestrutura MEC para contribuir com a diminuição da latência entre os usuários finais e servidores de conteúdo, bem como otimizar o uso de recursos nos servidores. Baseando-se na vazão mínima tolerada pela aplicação para orquestrar servidores *cache*, derivada do *Round Trip Time* (RTT) e tamanho do conteúdo. A principal contribuição da proposta é melhorar QoS e QoE a partir do redimensionamento de *cache* baseado na sensibilidade ao congestionamento da rede .

Assim, o artigo apresenta a fundamentação teórica na Seção 2. Os trabalhos relacionados são apresentados na Seção 3 enquanto a proposta inicial é descrita na Seção 4. A Seção 5 apresenta as conclusões.

2. Fundamentação Teórica

Essa seção apresenta uma breve revisão dos conceitos necessários para compreensão do problema e a abordagem adotada na proposta preliminar de solução.

2.1. Computação na Borda da Internet

[Hu et al. 2015] define que MEC fornece infraestrutura de serviços de Tecnologia da Informação e capacidade de computação em nuvem na borda da rede móvel, dentro da RAN, ou seja, nas proximidades de assinantes móveis. RAN é a infraestrutura de acesso que permite a comunicação entre dispositivos móveis e a rede principal, nesse contexto a comunicação é realizada via propagação de ondas de rádio [Abbas et al. 2018]. Em resumo, o paradigma MEC objetiva utilizar o poder computacional ou capacidade de armazenamento presente na computação em nuvem na borda de redes móveis, ou seja, dentro de sua infraestrutura RAN. Em suma os conceitos que aproximam os serviços de computação dos usuários tem como objetivo reduzir a latência e aprimorar a experiência final [Abbas et al. 2018].

2.2. *Cache* em Redes 5G

Segundo [Sengupta et al. 2017] projetos de *cache* em MEC resumidamente possuem dois principais problemas, sendo eles: qual conteúdo disponibilizar nos servidores *cache* disponíveis na borda e, como entregar o conteúdo. Normalmente a literatura classifica o primeiro como problema de posicionamento de conteúdo, em geral pesquisadores propõem formas de otimizar política de *caching* baseada em popularidade do conteúdo. O segundo problema origina dois sub problemas: o primeiro direcionado ao uso de recursos e como entregar o conteúdo, enfatizando otimização do uso de recursos de rádio, otimização do uso de enlaces *Backhaul* (BH) e *Fronthaul* (FH), otimização do uso de armazenamento, *Random Access Memory* (RAM) e processamento do servidor responsável por fornecer o serviço de *cache*. Por fim, existe um sub-problema relacionado a associação de usuário, de maneira que seja priorizado uma associação que proporcione a menor latência percebida pelo usuário ou maior taxa de vazão.

Os problemas de posicionamento do conteúdo definidos pelas políticas de *cache* e o problema de associação de usuário estão diretamente conectados. A definição do escopo da proposta apresentada será direcionada somente ao problema de associação de usuário.

2.3. Definição do Problema

Dentre os requisitos que devem ser alcançados na rede 5G, a ultra baixa latência merece destaque. Dessa forma, a 5G abre portas para a popularização de serviços que exigem *Ultra-Reliable Low-Latency Communication* (URLLC) como realidade virtual, realidade aumentada, Indústria 4.0, direção autônoma e IoT em geral. Essas aplicações possuem requisito de latência rígido, em sua maioria, menores que 1 milissegundo [Parvez et al. 2018]. Além do aumento de dispositivos na borda da Internet a centralização se consolida em *data centers* e serviços de armazenamento, através da computação em nuvem. A combinação desses fatores contribui para possíveis congestionamentos em enlaces compartilhados em razão de existir mais dispositivos finais conectados diretamente ao núcleo da Internet, esse fator pode causar aumento de latência [Abbas et al. 2018] e, consequentemente piora nos indicadores de QoS e QoE.

Existem elementos que dificultam o escalonamento de serviços de *caches* como as condições de estabilidade de uma conexão. Por exemplo, um usuário que está em deslocamento geográfico pode perceber aumento de latência e degradação na QoS e QoE, ou até mesmo ter sua conexão interrompida e perder os dados escalonados pelo servidor. Espera-se que estabilidade desse usuário seja totalmente diferente de um usuário fixo, que está utilizando os serviços através de computador ou dispositivo móvel em rede local, em sua casa, trabalho, *shopping* ou universidade. A heterogeneidade entre aplicações faz com que a alocação de recursos no servidor de *cache* tenha necessidades diferentes para cada tipo de aplicação, e.g. uma aplicação que fornece compartilhamento de serviços de mídias requer maior alocação de recursos no servidor de *cache*, enquanto uma aplicação de uma página Web requer menos recursos alocados.

Sabe-se que a demanda de enlaces compartilhados cresce com aumento do número de dispositivos e, consequentemente cria sobrecarga na rede, que é um elemento agravante para piora de QoS e QoE. O paradigma MEC pode ajudar na redução de tráfego em enlaces compartilhados no núcleo da Internet fazendo com que o dispositivo final acesse diretamente um servidor *cache* que provisiona uma réplica do serviço solicitado, localizada mais próximo geograficamente (ou em termos de latência de acesso) do dispositivo final [Abbas et al. 2018]. Com isso entende-se que para garantir QoS e QoE para o usuário final é interessante obter informações de suas conexões para proporcionar escalonamento adequado de serviços de *cache*.

3. Trabalhos Relacionados

O problema de associação de usuário é considerado como um problema da classe NP-difícil. Assim, autores propõem heurísticas como [Lei et al. 2017] que propõe um algoritmo *Deep Neural Network* (DNN), que consiste na otimização do posicionamento de conteúdo, entrega de conteúdo e associação de usuário, além de otimizar a eficiência energética, o algoritmo DNN alcançou resultado 90% ótimo. Por sua vez, [Tang et al. 2018] implementa uma solução com algoritmo genético alcançando resultados quase ótimos, considerando *caching* hierárquico e cooperativo entre *Edge Nodes* (ENs). De maneira semelhante, os autores [Zhou et al. 2018] adotam a abordagem de *caching* cooperativo. O algoritmo *Most Popular Content* (MPC) é escolhido para definir a popularidade do conteúdo e, com base na popularidade um modelo *Integer Linear Program* (ILP), permite que conteúdos mais populares sejam armazenados em todos ENs

e conteúdos menos populares sejam distribuídos localmente. Em [Chen et al. 2020] é proposto o protocolo *Software-Defined Networking (SDN) Based Adaptive Transmission Protocol* (SDATP) sobre a camada de transporte implantado na RAN com objetivo de reduzir a retransmissão de pacotes baseando-se na popularidade probabilística e em informações a respeito do congestionamento no enlace. A retransmissão dinâmica é realizada com base em SDN.

Os artigos em geral apresentam soluções sub-ótimas, heurísticas, DNN ou algoritmos genéticos ou formulações usando ILP, em razão da complexidade do problema e dificuldade de simulação.

Em geral, os autores apresentam propostas proativas e estáticas, ou seja, a ação de armazenamento de conteúdo no *cache* é normalmente realizada antes que usuário solicite o conteúdo, sem alterações em tempo de execução. Entretanto se o estado da rede muda e, ocasiona piora da QoS e QoE, não é possível realocar os recursos de forma inteligente e redimensionar o usuário. Sendo assim, a proposta preliminar desse artigo explora essa limitação, destaca a necessidade de obter informações de congestionamento e, além disso, considera a mobilidade dos usuários, de forma que suas contribuições futuras diferenciam-se dos trabalhos mencionados, sendo uma alternativa reativa e dinâmica.

4. Proposta Preliminar

Em resumo, o objetivo é espelhar o algoritmo de orquestração de *caching* de RAN em algoritmos de congestionamento que são parte do protocolo *Transmission Control Protocol* (TCP) da camada de transporte, baseando-se no RTT e a vazão mínima tolerada pela aplicação. Esse algoritmo pode ser implementado como uma heurística e, executado em um simulador como o NS-3[3] que possui o módulo desenvolvido pelo projeto 5G-LENA[4] que implementa a arquitetura de redes 5G ou, de forma numérica através de *Mixed-integer Linear Programming* (MILP) , pelo motivo de se tratar de um problema de grafos de uma especialização do problema genérico de *Maximum Flow Problems* (MFP) da classe de complexidade NP-Difícil.

A Figura 1 ilustra um modelo lógico representado por um grafo para exemplificar o cenário do problema. O grafo representa a infraestrutura *intra-AS*. O vértice *núcleo* representa a conexão da rede *intra-AS* com redes externas, referindo-se ao *Core* da rede. Os vértices estação de base (EB) representam as *base stations* (BS) que podem ser primárias ou secundárias, chamadas *macro base stations* (MBS) ou *small base stations* (SBS), respectivamente. Dessa forma, elementos *Distributed Antenna System* (DAS) podem ser estruturados dentro da RAN respeitando uma topologia hierárquica. Em suma, do ponto de vista lógico, esses elementos não possuem diferença, por esse motivo ambos recebem nome genérico. Tais elementos podem ou não ter capacidade de armazenamento de *cache* disponível. Os vértices equipamento de usuário (EU) representam dispositivos finais que podem ser de *interface* homem para máquina ou máquina para máquina. As arestas representam os enlaces de FH que conectam as MBS, SBS e EU e enlaces de BH que são responsáveis por conectar MBS ao *Core*. Além disso, arestas possuem custos atribuídos que se referem a uma representação ordinal a respeito do congestionamento da rede e disponibilidade de recursos de armazenamento disponíveis na *cache*.

[3]https://www.nsnam.org/

[4]https://5g-lena.cttc.es/

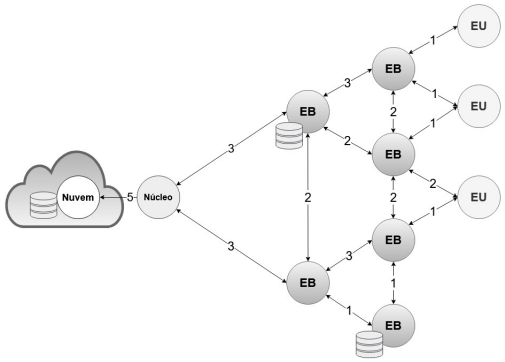

Figura 1. Modelo de perspectiva intra-AS.

No momento em que existe uma conexão entre um dispositivo final e um servidor de borda e, esse servidor armazena uma réplica do conteúdo solicitado, pode ocorrer instabilidade na RAN. Sendo assim, a QoS e QoE de provisionamento do serviço podem ser comprometidas devido aos fatores de congestionamento em um enlace, sobrecarga ou mobilidade do usuário. Para proporcionar melhor QoS e QoE enquanto o usuário estiver consumindo a *caching*, será proposto um algoritmo distribuído que orquestra a alocação da *cache* durante o provisionamento de conteúdo. Com o objetivo de que o servidor *cache* não armazene recursos desnecessários que não serão utilizados pelo dispositivo final, em razão do usuário estar em deslocamento geográfico ou de que o enlace esteja congestionado, isso evita que o provedor desperdice *hardware* e largura de banda com conexões ambíguas. O algoritmo pode considerar o custo de enlace, que nesse caso é definido como o custo de acesso ao conteúdo externo, ou seja, largura de banda de um enlace fornecido por outro AS e, além disso, objetiva otimizar o consumo de armazenamento no servidor.

A proposta inicial pretende monitorar o estado da conexão de usuário e, a partir dessa informação detectar congestionamento ou conexões de baixa qualidade de transmissão de dados baseando-se no RTT medido, tamanho do conteúdo da *caching* e na vazão mínima tolerada pela aplicação. Partindo do princípio que a vazão mínima tolerada seja informada pela aplicação e, a vazão real, obtida pela divisão do RTT atual e tamanho do conteúdo. A partir desses dados, comparando a vazão mínima e vazão atual, é possível identificar formação de congestionamento no enlace e, em seguida o servidor de *cache* que executa o algoritmo distribuído pode redirecionar a conexão para outro servidor de *cache* com condições estáveis para alocar a conexão. Isso evita que o servidor de *cache* faça conexões desnecessárias com o provedor de nuvem para obter menor latência.

5. Conclusão

Com o aumento do uso de dispositivos móveis, surgem novos desafios para garantir a melhor QoS e QoE possível, com avanço das redes móveis, a 5G está visivelmente no centro das discussões tecnológicas, além possibilitar o uso de aplicações com requisitos URLLC. Para isso, implantar o paradigma MEC combinado ao uso de servidores *cache* tem sido uma alternativa para minimizar a latência da rede. Em resumo, para evitar aumento da latência ou o uso de recursos desnecessários em um servidor de borda, é possível aproveitar conteúdo alocado em um servidor *cache* dentro da RAN evitando com que o conteúdo seja buscado diretamente em sua origem. Esse artigo apresentou uma alternativa a associação de usuários a *cache* e justificou como servidores *cache* podem contribuir

para a evolução das redes móveis e o problema de minimização de latência.

A proposta preliminar sugere um algoritmo heurístico distribuído baseando-se no RTT, tamanho do conteúdo e vazão mínima tolerada pela aplicação para minimizar a latência E2E, assim como, otimizar o uso de recursos de RAM nos servidores *cache*, com o objetivo de melhorar a QoS e QoE percebida pelo usuário final. Futuramente, será realizado o desenvolvimento de um ambiente de simulação para o cenário detalhado, além disso, será realizada a definição de detalhes do algoritmo e, por fim realização de testes em ambiente controlado, para posterior avaliação dos resultados encontrados.

Agradecimentos: UDESC, LabP2D, FAPESC e CAPES.

Referências

Abbas, N., Zhang, Y., Taherkordi, A., and Skeie, T. (2018). Mobile edge computing: A survey. *IEEE Internet of Things Journal*, 5(1):450–465.

Chen, J., Ye, Q., Quan, W., Yan, S., Do, P. T., Yang, P., Zhuang, W., Shen, X., Li, X., and Rao, J. (2020). Sdatp: An sdn-based traffic-adaptive and service-oriented transmission protocol. *IEEE Transactions on Cognitive Communications and Networking*, 6(2):756–770.

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

Iamnitchi, A., Datta, A., Montresor, A., Higashino, T., Barcellos, M., Garcia Lopez, P., Epema, D., Riviere, E., and Felber, P. (2015). Edge-centric Computing: Vision and Challenges. *ACM SIGCOMM Computer Communication Review*, 45(5):37–42.

Lei, L., You, L., Dai, G., Vu, T. X., Yuan, D., and Chatzinotas, S. (2017). A deep learning approach for optimizing content delivering in cache-enabled HetNet. In *Proceedings of the International Symposium on Wireless Communication Systems*, pages 449–453.

Parvez, I., Rahmati, A., Guvenc, I., Sarwat, A. I., and Dai, H. (2018). A survey on low latency towards 5G: RAN, core network and caching solutions. *IEEE Communications Surveys and Tutorials*, 20(4):3098–3130.

Rocha, A. A. A., Sampaio, L. N., Vieira, A. B., Wehmuth, K., and Ziviani, A. (2016). Revisitando metrologia de redes – do passado às novas tendências. In *Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos, SBRC 2016*, pages 151–209.

Sengupta, A., Tandon, R., and Simeone, O. (2017). Fog-Aided Wireless Networks for Content Delivery: Fundamental Latency Tradeoffs. *IEEE Transactions on Information Theory*, 63(10):6650–6678.

Tang, Q., Xie, R., Huang, T., and Liu, Y. (2018). Hierarchical collaborative caching in 5G networks. *IET Communications*, 12(18):2357–2365.

Zhou, F., Fan, L., Wang, N., Luo, G., Tang, J., and Chen, W. (2018). A Cache-Aided Communication Scheme for Downlink Coordinated Multipoint Transmission. *IEEE Access*, 6:1416–1427.