Explorando o Redirecionamento de *Streaming* de Vídeo em Infraestruturas de Borda-Nuvem para Usuários Móveis

Eduardo Cristaldo Panizzon
Federal University of Santa Catarina - UFSC
dudu123du@gmail.com

Alison R. Panisson
Federal University of Santa Catarina - UFSC
alison.panisson@ufsc.br

Roberto Rodrigues-Filho
Federal University of Santa Catarina - UFSC
roberto.filho@ufsc.br

Resumo

Video streaming has become one of the most popular applications in recent years. With the exponential increase in demand for content, the development of adaptive solutions that balance the efficient use of computational resources while maintaining a high quality of experience for users, especially mobile users, has become essential. In this context, the integration between edge and cloud infrastructures emerges as a promising approach for delivering high-quality video as the user moves. This integration encompasses a wide range of devices, from mobile equipment to servers in data centers, including intermediary servers based on fog computing, many of which are deployed near 4G/5G base stations. This work proposes to investigate rule-based autonomic computing strategies for the dynamic adaptation of streaming video services to better utilize computational resources in edge and cloud infrastructures. Specifically, it explores the content-steering architecture, implemented by the Dynamic Adaptive Streaming over HTTP (DASH) protocol, as a solution to optimize streaming video, focusing on enhancing the quality of experience for mobile users.

Keywords

Infraestruturas de Borda-Nuvem, Streaming Adaptativo de Vídeo, Usuários Móveis

1 Introdução

As aplicações de *streaming* de vídeo se tornaram protagonistas na Internet, com plataformas como *YouTube*, *Netflix*, *Twitch* e *Amazon Prime* dominando o mercado, sustentadas por gigantes como *Google*, *Netflix* e *Amazon*. Redes sociais como *Facebook*, *Instagram*, *TikTok* e *X* (antigo *Twitter*) também intensificam a divulgação de conteúdos em vídeo, tanto sob demanda quanto em transmissões ao vivo [1]. Esse cenário evidencia a crescente demanda por soluções que garantam uma experiência de alta qualidade para os usuários, o que é essencial para o sucesso dessas plataformas.

Para atender a essa demanda, tecnologias como infraestrutura de borda-nuvem [2], arquiteturas de *content-steering* [3] e *streaming* de vídeo adaptativo [4] têm sido amplamente utilizadas [5]. A infraestrutura de borda-nuvem, ou contínuo computacional (*computing continuum*), integra dispositivos heterogêneos desde usuários finais até servidores em *data centers*, passando por servidores intermediários na *fog* [6]. Essa hierarquia permite explorar recursos computacionais distintos em cada camada, permitindo a movimentação de serviços e dados para dispositivos mais próximos dos usuários, com o objetivo de reduzir a latência, evitar congestionamentos e melhorar a qualidade de experiência.

No contexto de *streaming* de vídeo, o contínuo computacional é utilizado em conjunto com a arquitetura de *content-steering* para redirecionar usuários a servidores com condições ideais de atendimento. Este trabalho investiga o uso dessa arquitetura, implementada pelo protocolo *Dynamic Adaptive Streaming over HTTP* (DASH), com o objetivo de otimizar o redirecionamento de requisições dos usuários. A abordagem direciona as requisições para servidores mais próximos do usuário, à medida que este se movimenta. Essa estratégia busca aprimorar a qualidade do serviço, garantindo que as requisições sejam processadas por servidores adequados à localização dos usuários, especialmente em cenários de alta mobilidade.

2 Fundamentação Teórica

2.1 Infraestruturas de Borda-Nuvem

As infraestruturas de borda-nuvem, ilustrada na Fig. 1, são infraestruturas hierárquicas com dispositivos com diferentes capacidades de recursos computacionais. Na camada inferior, a borda (*edge)*, é onde está localizado os dispositivos próximos aos usuários. Nesta camada se encontra, por exemplo, computadores (laptops), sensores, drones, carros autonomous, *etc.*

Na camada da borda, os dispositivos possuem as seguintes características: baixa disponibilidade, ou seja, estão sempre entrando e saindo da infraestrutura, baixa latência de rede, devida a sua localização próxima ao usuário final, e baixa capacidade de recursos computacionais (*ex.,* CPU, memória, armazenamento, *etc.*).

A segunda camada é composta por servidores intermediários. Esses servidores, geralmente possuem recursos computacionais com mais capacidade que os dispositivos da borda, mas com menos capacidade dos dispositivos de nuvem. A latência entre esses servidores e os usuários é maior que os dispositivos de borda, mas menor se comparada aos dispositivos de nuvem. O mesmo pode ser dito para disponibilidade, já que são servidores que possuem uma maior disponibilidade do que os dispositivos de borda, mas tendem a ter uma menor disponibilidade que os recursos na nuvem.

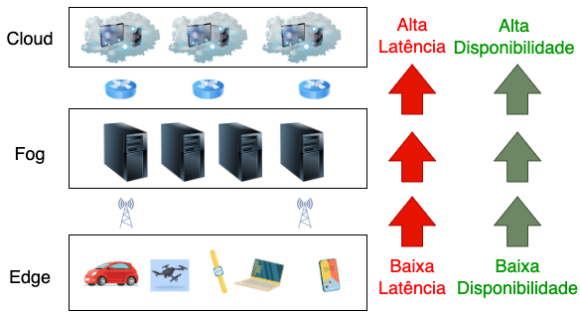

Figura 1: Infraestruturas de Borda-Nuvem

Por fim, no topo da hierarquia, se encontra as plataformas de computação em nuvem. Os serviços que executam nessas plataformas possuem a maior latência de rede até atingirem o usuário final. Em contrapartida, os recursos computacionais disponíveis na nuvem possuem maior capacidade e maior disponibilidade. Aplicações contemporâneas exploram os recursos computacionais dessa infraestrutura ao movimentar serviços entre as camadas, explorando os *trade-offs* entre capacidade de recursos computacionais e latência.

2.2 *Dynamic Adaptive Streaming over HTTP* (DASH)

O *Dynamic Adaptive Streaming over HTTP* (DASH) [7] é um protocolo amplamente utilizado para *streaming* de vídeo adaptativo que opera sobre HTTP, sendo particularmente eficiente e bastante adotado por aplicações de grande escala.

O funcionamento do DASH envolve a preparação do conteúdo de vídeo em diferentes níveis de qualidade, que é segmentado em trechos menores. Cada segmento corresponde a uma parte do vídeo que será reproduzida pelo *player* do usuário. Quando o usuário solicita um vídeo, o servidor envia os segmentos de forma contínua por meio de pacotes de resposta HTTP. À medida que esses segmentos chegam, o *player* os reproduz, proporcionando uma experiência contínua de *streaming* para o usuário.

O diferencial do DASH é sua capacidade de adaptação em tempo real, baseada em métricas de rede, como latência e largura de banda disponível. Para isso, o vídeo deve estar disponível em várias versões, com diferentes níveis de qualidade, como mencionado anteriormente. Segmentos de qualidade inferior possuem tamanhos menores, consumindo menos recursos da rede, enquanto os de qualidade superior exigem maior largura de banda para transmissão. O *player*, ao detectar condições adversas de rede, como alta latência ou congestionamento, pode solicitar segmentos de qualidade inferior para evitar interrupções na reprodução. Por outro lado, à medida que as condições de rede melhoram, o *player* pode retomar a solicitação de segmentos de maior qualidade, garantindo uma experiência equilibrada entre fluidez e qualidade de vídeo.

2.3 Arquitetura *Content-Steering*

O desenvolvimento de aplicações de *streaming* de vídeo exige um planejamento rigoroso e uma arquitetura bem estruturada para garantir que o conteúdo chegue ao usuário final com qualidade. Essas aplicações geralmente utilizam *Content Delivery Networks* (CDNs), redes de distribuição de conteúdo estrategicamente posicionadas próximas aos usuários para melhorar sua experiência.

Para atingir esse objetivo, as aplicações replicam seus conteúdos em diversas CDNs e dependem de mecanismos eficientes para redirecionar as requisições dos usuários a servidores localizados em regiões estratégicas. Nesse contexto, as comunidades responsáveis pela padronização do *streaming* de vídeo introduziram a arquitetura de *content-steering*. Essa arquitetura permite que o *player* do usuário faça requisições periódicas a uma entidade gerenciadora, que responde com informações sobre os servidores mais adequados para fornecer os conteúdos de vídeo.

A entidade gerenciadora, conhecida como servidor de *content-steering*, é responsável por direcionar os usuários a servidores estrategicamente localizados e com condições de rede adequadas para atender às demandas. A interação entre o *player* do usuário, o servidor de *content-steering* e os servidores de borda que armazenam os vídeos está ilustrada na Fig. 2.

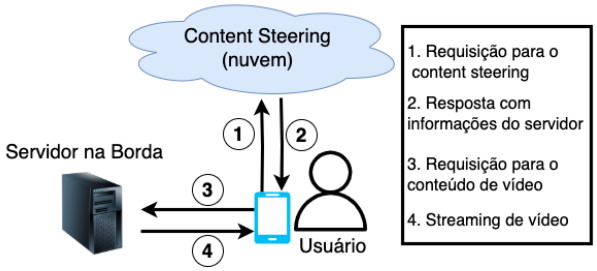

Figura 2: Arquitetura *Content-Steering*

3 Redirecionamento de *Streaming* de Video

3.1 Arquitetura e Simulador

Neste trabalho, utilizamos a arquitetura de *content-steering* do protocolo DASH e servidores cache na borda para redirecionar requisições de *streaming* de vídeo, acomodando usuários móveis. A solução foi validada com a máquina virtual de Rodrigues-Filho et al. [8] e um simulador desenvolvido para representar o movimento de um usuário móvel.

A arquitetura consiste em três elementos principais: um orquestrador (*content-steering*), servidores cache na borda e um usuário móvel. À medida que o usuário se desloca, o *player* consulta periodicamente o orquestrador, hospedado na nuvem, que determina o servidor de borda mais adequado com base na localização geográfica do usuário e na disponibilidade dos servidores. Essa abordagem assegura que o vídeo seja obtido do servidor mais eficiente para atender às condições do usuário em movimento.

O simulador foi implementado em uma máquina virtual baseada no VirtualBox (Ubuntu 22.04), contendo três contêineres Docker que executam servidores cache de borda com o mesmo vídeo armazenado. O orquestrador *content-steering* direciona o *player* do cliente ao servidor mais apropriado. A máquina virtual e o código foram disponibilizados publicamente[1].

3.2 Estudo de Caso

Neste trabalho, modelamos um dispositivo móvel que requisita a transmissão de vídeo a partir do servidor *content-steering* na nuvem enquanto percorre um trajeto que passa por diversos servidores de borda (Fig. 3). O servidor na nuvem monitora a distância do cliente em relação aos servidores de borda e determina, em tempo real, o servidor mais próximo com o conteúdo solicitado.

O *player* do usuário realiza consultas periódicas ao servidor de *content-steering*. Quando o *content-steering* detecta que a distância do usuário ao servidor de borda atual ultrapassa um limite definido, ele seleciona um novo servidor de borda mais próximo para dar continuidade à transmissão. Na consulta subsequente, o *content-steering* informa ao *player* o endereço do novo servidor. O *player* então redireciona suas requisições para o novo servidor, assegurando a continuidade do *streaming* sem interrupções, perdas de conexão ou variações bruscas de latência.

[1] https://github.com/robertovrf/content-steering-tutorial.

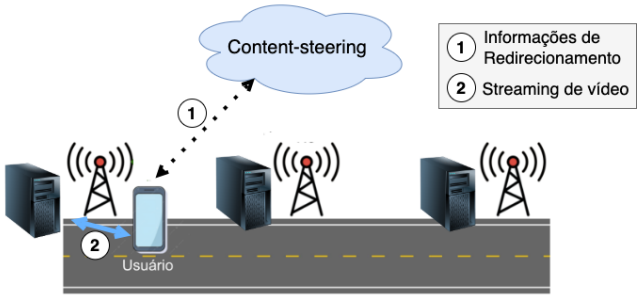

Figura 3: Usuário se movendo em linha reta (para direita) e trocando de um servidor de borda para outro, de acordo com as informações passadas pelo servidor *content-steering*.

A simulação da mobilidade do usuário funciona atribuindo um valor inicial de distância entre o usuário e cada servidor cache de borda. A cada solicitação ao servidor de *content-steering*, a distância entre o usuário e o servidor de borda transmissor é aumentada em 2 unidades, enquanto as distâncias dos demais servidores diminuem progressivamente. Quando a distância do servidor transmissor atinge o limite predefinido para troca, o *content-steering* seleciona um novo servidor de borda para continuar a transmissão.

Essa abordagem simula um cenário em que o usuário se desloca ao longo de uma estrada, inicialmente próximo a um servidor de borda, mas se afastando em direção a outro. Os experimentos mostram que, utilizando a arquitetura de *content-steering*, o redirecionamento do *streaming* de vídeo ocorre de forma antecipada, evitando que o aumento da distância comprometa a qualidade da experiência do usuário. Além disso, o processo garante a continuidade do *streaming* sem interrupções perceptíveis.

4 Trabalhos Relacionados

Diversos estudos têm investigado estratégias de cache para *streaming* de vídeo na borda. Trabalhos como [9] e [10] são exemplos representativos dessa abordagem, enquanto outros, como [11] e [12], se concentram em alocação de recursos e balanceamento de carga entre servidores em CDNs e dispositivos de borda.

Gama *et al.* [4] e Rodrigues-Filho *et al.* [8], por sua vez, exploram arquiteturas de *content-steering* voltadas para *streaming* de vídeo em infraestruturas de borda-nuvem. Esses trabalhos destacam o potencial do redirecionamento de requisições de usuários como uma estratégia adaptativa para otimizar o uso dos recursos computacionais em infraestruturas de borda-nuvem, melhorando a eficiência e a qualidade do serviço.

Este trabalho, por sua vez, representa um esforço inicial para aprofundar a utilização de recursos em infraestruturas de borda-nuvem para *streaming* de vídeo, com o diferencial de focar no redirecionamento de requisições por meio do *content-steering*, no contexto de usuários móveis.

5 Conclusão

Este trabalho investiga o uso de arquiteturas de *content-steering* para o redirecionamento de requisições de *streaming* de vídeo em infraestruturas de borda-nuvem. A abordagem proposta explora o direcionamento dinâmico de requisições para servidores estrategicamente localizados na borda, permitindo que, conforme o usuário se desloca, as requisições sejam redirecionadas para o servidor mais próximo à sua nova localização.

Demonstramos que é possível implementar políticas adaptativas flexíveis, aproveitando a arquitetura de *content-steering* para otimizar o atendimento de requisições de *streaming* de vídeo. Essa estratégia maximiza a eficiência no uso de recursos das infraestruturas de borda-nuvem, proporcionando uma melhor experiência para usuários móveis. O código desenvolvido durante este projeto foi disponibilizado em um repositório público no Github[2]. Em trabalhos futuros, pretende-se explorar o uso de técnicas de aprendizado de máquina, possivelmente utilizando abordagens de IA distribuída como em [13], prevendo a movimentação do usuário e antecipando necessidade de mudança de servidor de borda.

Referências

[1] Xiangbo Li, Mahmoud Darwich, Mohsen Amini Salehi, and Magdy Bayoumi. Chapter four - a survey on cloud-based video streaming services. volume 123 of *Advances in Computers*, pages 193–244. Elsevier, 2021. doi: https://doi.org/10.1016/bs.adcom.2021.01.003. URL https://www.sciencedirect.com/science/article/pii/S0065245821000280.

[2] Schahram Dustdar, Victor Casamayor Pujol, and Praveen Kumar Donta. On distributed computing continuum systems. *IEEE Transactions on Knowledge and Data Engineering*, 35(4):4092–4105, 2023. doi: 10.1109/TKDE.2022.3142856.

[3] Daniel Silhavy, Will Law, Stefan Pham, Ali C. Begen, Alex Giladi, and Alex Balk. Dynamic cdn switching - dash-if content steering in dash.js. In *Proceedings of the 2nd Mile-High Video Conference*, MHV '23, page 130–131, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701603. doi: 10.1145/3588444.3591027. URL https://doi.org/10.1145/3588444.3591027.

[4] Eduardo S. Gama, Roberto Rodrigues-Filho, Edmundo R. M. Madeira, Roger Immich, and Luiz F. Bittencourt. Enabling adaptive video streaming via content steering on the edge-cloud continuum. In *2024 IEEE 8th International Conference on Fog and Edge Computing (ICFEC)*, pages 35–42, 2024. doi: 10.1109/ICFEC61590.2024.00018.

[5] Haoyu Zhan, Lisheng Fan, Chao Li, Xianfu Lei, and Feng Li. Cloud–edge learning for adaptive video streaming in b5g internet of things systems. *IEEE Internet of Things Journal*, 11(24):40140–40148, 2024. doi: 10.1109/JIOT.2024.3450477.

[6] Jagdeep Singh, Parminder Singh, and Sukhpal Singh Gill. Fog computing: A taxonomy, systematic review, current trends and research challenges. *Journal of Parallel and Distributed Computing*, 157:56–85, 2021. ISSN 0743-7315. doi: https://doi.org/10.1016/j.jpdc.2021.06.005. URL https://www.sciencedirect.com/science/article/pii/S0743731521001349.

[7] Iraj Sodagar. The mpeg-dash standard for multimedia streaming over the internet. *IEEE MultiMedia*, 18(4):62–67, 2011. doi: 10.1109/MMUL.2011.71.

[8] Roberto Rodrigues-Filho, Eduardo S Gama, Marcio Miranda Assis, Roger Immich, and Edmundo Madeira. Content steering: Leveraging the computing con-tinuum to support adaptive video streaming. *Minicursos do XLII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos. XLIed.: SBC*, pages 1–39, 2024.

[9] Jesús Aguilar-Armijo, Christian Timmerer, and Hermann Hellwagner. Space: Segment prefetching and caching at the edge for adaptive video streaming. *IEEE Access*, 11:21783–21798, 2023. doi: 10.1109/ACCESS.2023.3252365.

[10] Yinxin Li, Haiyan Tu, Guorong Zhou, Ting Li, Yunfeng Wang, Kai Liang, Zhigang Wang, and Liqiang Zhao. Design and implementation of adaptive-bitrate-streaming-based edge caching. In *2022 IEEE 95th Vehicular Technology Conference: (VTC2022-Spring)*, pages 1–5, 2022. doi: 10.1109/VTC2022-Spring54318.2022.9860428.

[11] Haolin Liu, Xiaoling Long, Zhetao Li, Saiqin Long, Rong Ran, and Hui-Ming Wang. Joint optimization of request assignment and computing resource allocation in multi-access edge computing. *IEEE Transactions on Services Computing*, 16(2):1254–1267, 2023. doi: 10.1109/TSC.2022.3180105.

[12] Kai Xu, Xiang Li, Sanjay Kumar Bose, and Gangxiang Shen. Joint replica server placement, content caching, and request load assignment in content delivery networks. *IEEE Access*, 6:17968–17981, 2018. doi: 10.1109/ACCESS.2018.2817646.

[13] Bernardo Pandolfi Costa, Heitor Henrique da Silva, Analucia S. Morales, Luiz F. Bittencourt, Alison R. Panisson, and Roberto Rodrigues-Filho. A multi-agent approach to self-distributing systems. In *International Conference on Advanced Information Networking and Applications (AINA)*, 2025.

[2] https://github.com/robertovrf/content-steering-tutorial.