Avaliação de desempenho de uma arquitetura de vídeo sob demanda usando rede de filas fechada

Leonardo Cristian, Rubenilson de Sousa e Francisco Airton Silva

[1] Laboratório de Pesquisas Aplicadas a Sistemas Distribuídos (PASID)
Universidade Federal do Piauí (UFPI)
Picos, PI, Brasil

{leonardocristian, rubenilson, faps}@ufpi.edu.br

Abstract. *The streaming service has been receiving great visibility lately, getting more and more users. This notoriety is due to the current context of the world, where the pandemic caused by COVID-19 led the population to adhere more to the entertainment offered by video-on-demand platforms. However, to support this high demand for its services, it is feasible to have an assessment regarding the performance in this type of architecture, in order to avoid bottlenecks that can spoil the users' experience. This article proposes a closed queue model for performance evaluation. This model allows us to calculate some metrics such as mean response time (MRT) and component utilization, as well as perform simulations in different scenarios. The proposed model allows the designer to previously assess the behavior of this type of infrastructure without the need for prior expenses, enabling an accurate simulation before its implementation. And with the simulation results, it was possible to identify some factors that affect the model's performance, as well as its best configuration.*

Resumo. *O serviço de streaming vem recebendo uma grande visibilidade ultimamente, obtendo cada vez mais usuários. Essa notoriedade é devido ao contexto atual do mundo, onde com a pandemia ocasionada pelo COVID-19, levou a população a aderir mais pelo entretenimento oferecido pelas plataformas de vídeo sob demanda. No entanto, para suportar essa alta demanda pelos seus serviços, é viável ter uma avaliação a respeito do desempenho neste tipo de arquitetura, a fim de evitar gargalos que podem estragar a experiência dos usuários. Este artigo propõe um modelo de filas fechadas para avaliação de desempenho. Tal modelo nos permite calcular algumas métricas como, tempo médio de resposta (MRT) e utilização dos componentes, bem como realizar simulações em diferentes cenários. O modelo proposto permite que o projetista avalie previamente o comportamento deste tipo de infraestrutura sem a necessidade de gastos prévios, possibilitando uma simulação precisa antes da sua implantação. E com os resultados das simulações, foi possível identificar alguns fatores que prejudicam o desempenho do modelo, bem como sua melhor configuração.*

1. Introdução

Nos últimos vinte anos, o mundo presenciou a mudança relativamente rápida dos modos de entretenimento através da tecnologia. Os meios de se consumir músicas, vídeos e

filmes principalmente, mudaram com o passar desses anos, e o que antes era algo de difícil acesso, hoje pode estar a apenas a um ”click”de distância. Com o avanço dos dispositivos tecnológicos, é possível acessar qualquer tipo de informação em qualquer lugar, possibilitando o indivíduo a ter um maior acesso a recursos de multimídia. Um deles que ganha maior visibilidade a cada dia é o serviço de streaming de vídeo. De acordo com a pesquisa da Kantar IBOPE Media, em um período de 12 meses, 58% dos usuários de internet disseram que viram mais vídeo e TV online em streaming pago durante os períodos de isolamento. O tempo em frente a televisão aumentou 37 minutos diários, e cada indivíduo passou cerca de 1h49 por dia assistindo a conteúdos em plataformas de streaming (Silva 2021). Essa preferência a tal serviço, é resultado da versatilidade de usar as plataformas em dispositivos variados, além da opção de escolha do que o usuário deseja assistir, diferenciando de outros canais de comunicação, como a TV aberta, onde a programação já é definida.

É perceptível que as pessoas estão cada vez mais interessadas em assistir conteúdos disponibilizados na internet, visto que, a Netflix - uma das maiores empresas de streaming de vídeo sob demanda da atualidade - conseguiu fechar o primeiro trimestre de 2020 com 15,77 milhões de novos assinantes (Alecrim 2020), número consideravelmente alto comparado com os anos anteriores. Mas ela não é a única no ramo, empresas como Youtube, Amazon Prime Video, HBO GO, a recém chegada Disney+, entre várias outras, englobam todo esse contexto multimilionário que só tende a crescer no mercado. Portanto, a avaliação de desempenho se torna útil para ajudar na implementação prévia da arquitetura, analisando métricas de desempenho como tempo de resposta nível de utilização dos componentes, a fim de analisar configurações eficientes no sistema servindo para mitigar problemas que podem ocorrer. A interação que ocorre entre o usuário e o serviço multimídia, é feita pela solicitação de pacotes (Wu et al. 2001), onde a pessoa faz determinada ação como pausar, continuar, pular, voltar, aumentar ou diminuir a qualidade do vídeo, dentre outras, que geram requisições. Essas requisições serão enviadas para uma fila de processamento que ao chegar na nuvem, será processada retornando determinada ação para a Internet e assim atender a solicitação do usuário (Dantas et al. 2016). A avaliação de desempenho se torna útil para ajudar na implementação prévia da arquitetura, servindo assim para mitigar problemas que podem ocorrer.

Este artigo propõe elaborar um modelo de filas fechadas, a fim de analisar o desempenho de uma arquitetura de vídeo sob demanda avaliando algumas métricas como: utilização dos componentes, tempo médio de resposta e número de requisições do sistema, todas variando a quantidade de núcleos das máquinas utilizadas, tal avaliação nos permite uma análise detalhada do comportamento da arquitetura antes mesmo da sua implantação. O restante do documento está organizado da seguinte forma. A Seção 2 apresenta os trabalhos relacionados. A Seção 3 descreve o cenário avaliado para projetar o modelo das filas. a seção 4 apresenta o modelo de filas proposto e uma explicação do seu funcionamento. A seção 5 exibe os resultados das análises. E por fim, temos a Seção 6 que mostra a conclusão do trabalho e considerações finais.

2. Trabalhos relacionados

Nesta seção temos os trabalhos relacionados, exibindo suas características e evidenciando o diferencial do presente artigo. Embora muito já se tenha feito a respeito da transmissão de vídeo em tempo real, não existe um modelo geral que define uma arquitetura escalável

e flexível para análise de desempenho (Hoque and Miranskyy 2018). Portanto, todos os títulos aqui citados têm as suas métricas para avaliação do serviço, sendo eles para vídeo em específico ou não. Na Tabela 1, é possível visualizar melhor as comparações, onde temos as colunas divididas por título dos artigos, contexto da aplicação do trabalho, método utilizado pelos autores e se foi utilizado uma avaliação de capacidade ou não. Para contextualizar, foi definido a principal área em que cada artigo trabalha, já que o contexto streaming pode ser amplo. Dos trabalhos vistos, a grande maioria teve como foco o *Streaming* de vídeo, por enquadrar melhor no contesto desse artigo. Porém, o (Hoque and Miranskyy 2018) foi o único que optou por não focar na questão de vídeo, mas sim no serviço streaming em geral.

Apesar da maioria utilizar métricas de desempenho para avaliação, houve uma grande variação dos métodos utilizados. Pois, como já mencionado, não há a existência de um padrão para essa análise. Diferente de todos os artigos vistos, esse apresenta as métricas de MRT, Utilização e Número de Requisições no sistema. Tais métricas são importantes para melhor compreensão de como os componentes estão sendo sobrecarregados no modelo. O Zhao et al. 2019 foi o que apresentou características mais semelhantes, pois também utilizava redes de filas como base para a sua avaliação de desempenho. Sobre Avaliação de Capacidade, apenas Niu et al. 2011 e Zhao et al. 2019 a utilizaram em seus trabalhos. Porém, nosso diferencial consiste na avaliação de capacidade do modelo atribuindo diferentes prioridades para cada tipo de usuário. Além das métricas para um melhor monitoramento, como o nível de utilização dos componentes, tempo médio de resposta e quantidade de requisições.

Tabela 1. Trabalhos Relacionados

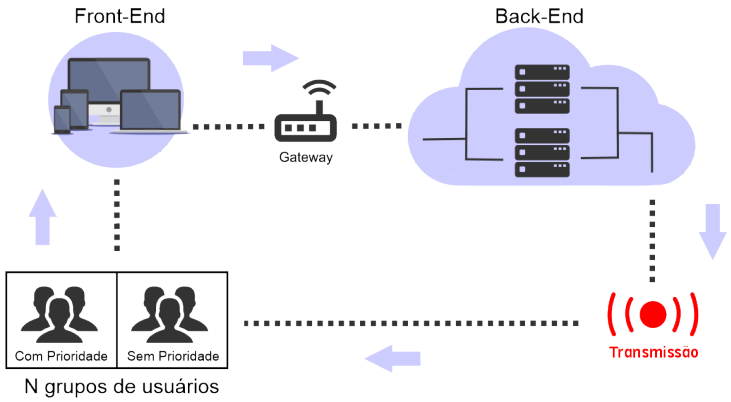
Título	Métricas Utilizadas	Avaliação de Capacidade
(Niu et al. 2011)	População, upload de pares e a demanda de largura de banda	Sim
(De Cicco and Mascolo 2013)	CDF, tempos transientes e fator de segurança	Não
(Juluri et al. 2015)	Qualidade de experiência	Não
(Bagci et al. 2016)	Taxas de bits	Não
(Hoque and Miranskyy 2018)	Um diagrama da arquitetura de 7 camadas.	Não
(Duanmu et al. 2018)	Qualidade de experiência	Não
(Zhao et al. 2019)	Taxa de rejeição, utilização, taxa de chegada, N° de requisições e QLRA	Sim
(Irawan and Surantha 2020)	Algoritmos de enfileiramento (DropTail, BLUE e CoDel)	Não
Este trabalho	MRT, utilização e número de requisições no sistema	Sim

Fonte: Autor.

3. Cenário Avaliado

Esta seção apresenta uma visão geral acerca de uma arquitetura que fornece serviço de streaming de vídeo sob demanda. A Figura 1 nos apresenta o cenário em que temos os componentes principais para o funcionamento da plataforma. O fluxo de dados ocorre da esquerda para a direita, partindo do usuário, em que, utilizando dispositivos como *PC's, tablets e smartphones*, irá interagir com todo esse sistema em questão de milésimos, muitas vezes sem ao menos ter ciência disso.

Figura 1. Arquitetura avaliada de uma plataforma de serviço de vídeo sob demanda.



Fonte: Autor.

Os grupos de usuários, que estarão utilizando o serviço de *stream*, irão executar ações diretamente com o Front-End da aplicação, aqui representados por variados tipos de dispositivos que tenham suporte a tal plataforma. Essas ações, serão as solicitações da pessoa ao sistema, que depende do que a aplicação dispõe a seus clientes, mas em geral podemos citar as mais comuns, que são o ato de pausar, pular, voltar ou avançar um vídeo, dentre outras. Essas requisições irão para o gateway, a fim de estabelecer uma conexão com a Nuvem, um provedor de internet que estará fornecendo servidores de processamento. No Back-End, as requisições dos usuários serão devidamente processadas, para que assim os dados possam ser transmitidos novamente aos usuários em forma de resposta a ação que eles pretendiam realizar no sistema.

Como o Back-End é o componente responsável pelo processamento das requisições, o mesmo pode variar de capacidade, podendo atender a uma maior demanda de usuários. No cenário avaliado, existem N grupos de M usuários em que podem ter prioridades diferentes no sistema, ex: o grupo A tem uma maior condição de contratar um serviço de internet mais eficiente, portanto, possui uma prioridade maior em relação ao grupo B. Os grupos de usuários irão utilizar os mesmos recursos do Back-End, podendo ter diferentes resultados dependendo da prioridade de cada grupo. Tal arquitetura pode ser evoluída considerando múltiplos provedores de nuvem.

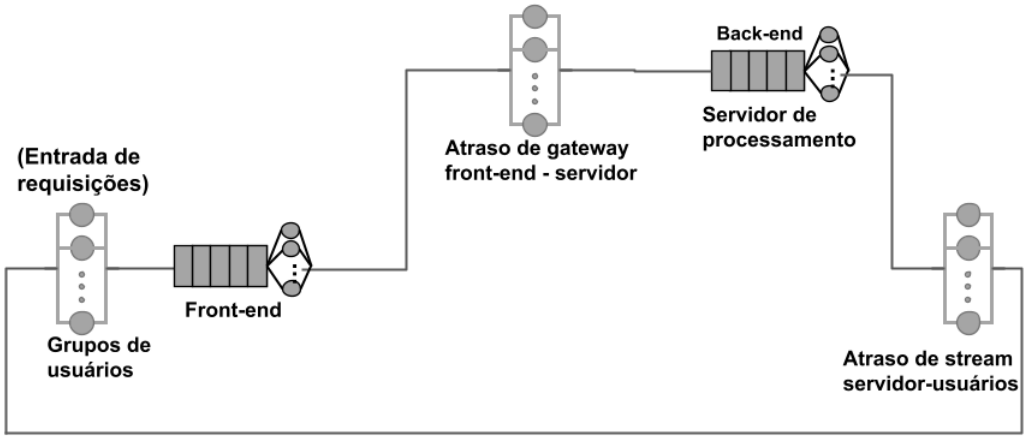
4. Modelo de Filas proposto

A teoria das filas foi desenvolvida para prover modelos que retratem previamente o comportamento de um sistema que forneça serviços que possuam demandas que aumentem aleatoriamente(Dantas et al. 2016). Estas filas, são ocasionadas quando a capacidade de processamento não consegue suportar as requisições que chegam no sistema. Para modelar um sistema usando filas, é necessário ter em mente que devemos definir alguns parâmetros como: taxa de chegada das requisições, taxa de atendimento do servidor, política de filas a ser utilizada e capacidade do sistema. As máquinas Front-End e Back-End foram construídas usando o modelo de fila M/M/C/K seguindo o padrão de serviço e chegada de clientes exponenciais, e cada estação C com uma capacidade limitada de K nós (núcleos). A fim de simular diferentes grupos de pessoas, utilizamos o conceito de prioridades de classes, que consiste em atribuir um valor de prioridade para cada classe,

onde, quanto maior for o valor, maior é sua prioridade no sistema.

Esta seção apresenta o modelo proposto para simular o comportamento da arquitetura mostrada na seção 3. A Figura 2 representa um modelo de filas fechadas onde os N grupos de usuários da plataforma enviam requisições para o Front-End, representado aqui como uma fila de tamanho fixo responsável por armazenar cada requisição dos usuários, que em seguida as encaminha para serem processadas na fila do Back-End, este, representado como uma fila que recebe os dados transmitidos pelo gateway vindos do Front-End, simbolizado por um atraso que faz a ponte entre o Front-End e o Back-End.

Figura 2. Modelo proposto para avaliar uma arquitetura de vídeo sob demanda utilizando filas.



Fonte: Autor.

Cada fila possui um tamanho máximo para armazenar essas requisições, que serão enviadas durante um curto intervalo de tempo fixo. O Back-End é o componente responsável por realizar o processamento dessas requisições antes de enviá-las de volta aos usuários, que geram novas, formando um ciclo. Nele, as requisições são recebidas utilizando a política de filas *First Come First Serverd (FCFS)*. Os parâmetros utilizados para avaliar os componentes do modelo são: quantidade de núcleos (cores) de processamento, representados como servidores internos da fila, tamanho da fila, tempo de atendimento ou serviço e taxa de queda.

Com base no cenário descrito na seção anterior, iremos realizar algumas simulações, variando a quantidade de núcleos de processamento, a fim de calcular algumas métricas. Consideramos que existem dois grupos de usuários, representados no modelo como Classe 1 e Classe 2, ambas com prioridades diferentes, sendo a Classe 2 mais prioritária do que a Classe 1. Buscamos aqui avaliar a utilização das filas, tempo médio de resposta para as Classes, e número de requisições. O desempenho dos outros componentes não serão avaliados, visto que, os mesmos não terão seus parâmetros alterados durante a simulação.

5. Simulações

Esta seção apresenta os parâmetros utilizados para avaliar o desempenho do modelo proposto, calculando a utilização, tempo médio de resposta e número de requisições no sis-

tema. Todas as simulações, além da construção do modelo foram feitas utilizando a plataforma *Java Modeling Tool (JMT)*. Os parâmetros para cada componente do modelo são: para as máquinas Front-End e Back-End, com o tempo de 25ms e 33.3ms, com o tamanho de 200 e 500 respectivamente. Onde o tempo (ms) do componente da fila (estação), corresponde ao período que leva pra requisição ser processada ou seu tempo de atendimento. Além do tempo dos atrasos, sendo eles o Servidor-usuários, com 2000ms, e o *Gateway* com 3.3ms.

Após a configuração de todos os componentes, foram calculadas as métricas de utilização, tempo médio de resposta e número de requisições. Para avaliar essas métricas foi realizada diferentes simulações, variando a quantidade de núcleos de processamento das filas, primeiro com 8, em seguida com 12 núcleos. Tais simulações foram feitas utilizando uma funcionalidade da plataforma, em que realiza uma séria de simulações variando alguns parâmetros do modelo pré estabelecidos a fim de obter resultados precisos. O parâmetro escolhido para as simulações foi o número de requisições.

6. Resultados

Esta seção apresenta os resultados obtidos com as simulações do modelo proposto. No modelo apresentado na Figura 2, simulamos uma situação em que dois grupos de 100 usuários, totalizando 200 pessoas na rede, estão utilizando o serviço da plataforma de streaming, as configurações podem ser alteradas para N grupos de M usuários. As requisições serão enviadas e processadas em um servidor na nuvem. Para obter os resultados, foram realizadas 10 simulações, aumentando o número de requisições por simulação e variando a capacidade de processamento do Back-End, criando assim duas configurações, onde a Configuração A, tem o Front-End e o Back-End com 8 núcleos. Já a Configuração B, altera o Back-End para a quantidade de 12 núcleos.

Para realizar as simulações, foi alterada a capacidade de processamento apenas do Back-End, visto que, o mesmo é responsável por realizar todo o trabalho pesado para atender as requisições dos diferentes grupos de usuários. Com isso, podemos observar na Figura 3 o nível de utilização do Front-End e Back-End em ambas as configurações. Note aqui que o Back-End possui um nível maior de utilização na configuração A, visto que, com poucos núcleos de processamento, a máquina precisa trabalhar mais para atender a demanda de requisições dos usuários. No Front-End, podemos notar que o nível de utilização sobe quando o Back-End possui mais núcleos, isso se dá pelo fato do Back-End suportar mais requisições por segundo que o Front-End, isso faz com que o dispositivo Front-End trabalhe mais para enviar as requisições para o Back-End.

Na Figura 4 é exibido o número de requisições do sistema. Diferente da simulação do MRT e da Utilização, nesse caso foi analisado todas as Classes em uma única vez, variando apenas os Núcleos presentes no Back-End, entre 8 e 12. Como visto no gráfico foi possível observar que houve uma estabilidade em ambos os casos, porém, com 8 Núcleos, o número de requisições foi maior. A Figura 5 representa o tempo médio de resposta em milissegundos que leva para as requisições de cada grupo de usuários serem processadas no Back-End e retornadas novamente aos mesmos, levando em consideração diferentes prioridades para cada grupo de usuários como mencionado anteriormente. Podemos observar que, com 8 núcleos no Back-End, a classe 1 possui um MRT maior que a classe 2, isso ocorre porque a prioridade da classe 2 é maior em relação a classe 1,

Figura 3. Utilização

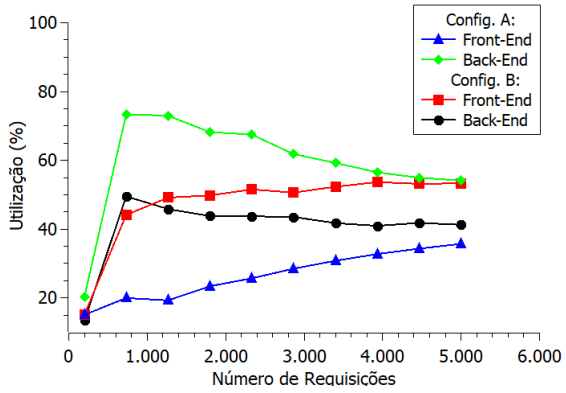


Figura 4. Número de requisições do sistema

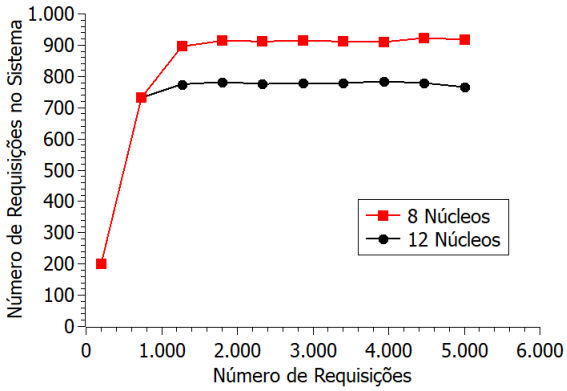
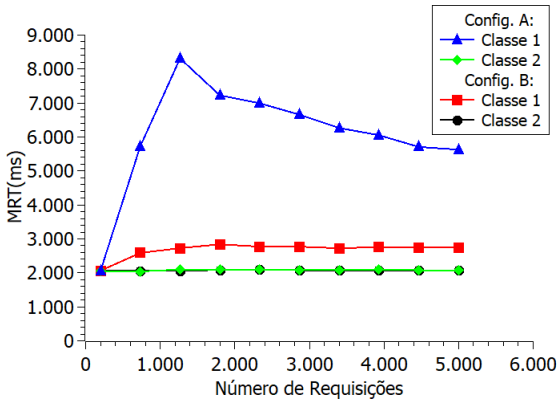


Figura 5. tempo médio de resposta



Fonte: Autor.

pois, quanto maior o valor, maior sua prioridade e suas requisições são processadas com mais eficiência. E como a classe 1 possui uma prioridade menor, o sistema demora mais para processar suas requisições. E quando aumentamos a capacidade de processamento do Back-End, há uma grande diminuição do MRT da classe 1, visto que, o mesmo possui mais capacidade para processar todas as requisições de ambas as classes com eficiência.

7. Conclusão

No desenvolvimento deste trabalho, foi realizada uma avaliação de desempenho de uma arquitetura de vídeo sob demanda utilizando filas fechadas. Propomos um modelo que permite simular o comportamento dessa infraestrutura, a fim de analisar alguns fatores que podem causar impacto no sistema. Concluímos que, ao alterar a quantidade de núcleos de uma estação do modelo, conseguimos diminuir o nível de utilização do mesmo, tornando assim mais eficiente e menos sobrecarregado. Vimos que, com diferentes grupos de usuários utilizando o mesmo sistema, este processa as informações das classes de maior prioridade de forma mais eficiente e com menor tempo de resposta. Com o modelo proposto, conseguimos realizar uma análise do seu comportamento, analisando o impacto causado por variações em alguns componentes.

Referências

Alecrim, E. (2020). Netflix tem crescimento recorde e vai a 183 milhões de assinantes.

Bagci, K. T., Sahin, K. E., and Tekalp, A. M. (2016). Queue-allocation optimization for adaptive video streaming over software defined networks with multiple service-levels. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 1519–1523. IEEE.

Dantas, J., Matos, R., Araujo, J., Oliveira, D., Oliveira, A., and Maciel, P. (2016). Hierarchical model and sensitivity analysis for a cloud-based vod streaming service. In *2016 46th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshop (DSN-W)*, pages 10–16. IEEE.

De Cicco, L. and Mascolo, S. (2013). An adaptive video streaming control system: Modeling, validation, and performance evaluation. *IEEE/ACM Transactions on Networking*, 22(2):526–539.

Duanmu, Z., Rehman, A., and Wang, Z. (2018). A quality-of-experience database for adaptive video streaming. *IEEE Transactions on Broadcasting*, 64(2):474–487.

Hoque, S. and Miranskyy, A. (2018). Architecture for analysis of streaming data. In *2018 IEEE International Conference on Cloud Engineering (IC2E)*, pages 263–269. IEEE.

Irawan, Y. and Surantha, N. (2020). Performance evaluation of queue algorithms for video-on-demand application. In *2020 International Conference on Information Management and Technology (ICIMTech)*, pages 966–971. IEEE.

Juluri, P., Tamarapalli, V., and Medhi, D. (2015). Measurement of quality of experience of video-on-demand services: A survey. *IEEE Communications Surveys & Tutorials*, 18(1):401–418.

Niu, D., Liu, Z., Li, B., and Zhao, S. (2011). Demand forecast and performance prediction in peer-assisted on-demand streaming systems. In *2011 Proceedings IEEE INFOCOM*, pages 421–425. IEEE.

Silva, R. (2021). Um ano depois do início da pandemia, plataformas de streaming contabilizam ganhos.

Wu, D., Hou, Y. T., Zhu, W., Zhang, Y.-Q., and Peha, J. M. (2001). Streaming video over the internet: approaches and directions. *IEEE Transactions on circuits and systems for video technology*, 11(3):282–300.

Zhao, H., Wang, J., Wang, Q., and Liu, F. (2019). Queue-based and learning-based dynamic resources allocation for virtual streaming media server cluster of multi-version vod system. *Multimedia Tools and Applications*, 78(15):21827–21852.