Aprovisionamento de Recursos para Serviços URLLC e eMBB em Redes MEC-NFV: Uma Análise Baseada em CTMC

Caio B. Bezerra De Souza[2], Marcos R. de Moraes Falcão [2], Maria G. Lima Damasceno[1,2], Renata K. Gomes Dos Reis[1,2], Andson M. Balieiro[2]

[1]Sidia Instituto de Ciência e Tecnologia
Manaus – AM – Brazil

[2]Centro de Informática (CIn) – Universidade Federal de Pernambuco (UFPE)
Recife – PE – Brazil

{maria.lima, renata.gomes}@sidia.com
{cbbs, amb4, mgld, rkgr}@cin.ufpe.br, mrmfpessoal@gmail.com

Abstract. *Multiple Access Edge Computing (MEC) and Network Function Virtualization (NFV) are key-technologies in the Fifth Generation of Mobile Networks (5G) to support services such as Ultra-Reliable and Low Latency Communication (URLLC) and Enhanced Mobile Broadband (eMBB). However, ensuring the coexistence of these services poses challenges, particularly in dynamic resource allocation within the MEC-NFV domain. This paper presents a Continuous Time Markov Chain (CTMC)- based model to analyze the impact of dynamic resource allocation on both URLLC and eMBB services in an MEC-NFV environment. The analysis considers factors such as virtualization overhead, virtual resource failures, and varying numbers of containers and buffer sizes. The results indicate that availability, response time, and energy consumption are strongly influenced by the number of containers, while buffer size primarily affects response times.*

Resumo. *A Computação de Borda de Acesso Múltiplo (MEC) e a Virtualização de Funções de Rede (NFV) são tecnologias-chave da Quinta Geração de Redes Móveis (5G) para suportar serviços como o de Comunicação Ultra Confiável e com Baixa Latência (URLLC) e a Banda Larga Móvel Melhorada (eMBB). Entretanto, garantir a coexistência desses serviços é desafiador, especialmente na alocação dinâmica de recursos no domínio MEC-NFV. Este artigo apresenta um modelo baseado em Cadeia de Markov de Tempo Contínuo (CTMC) para analisar o impacto da alocação dinâmica de recursos em ambos os serviços em um ambiente MEC-NFV, considerando a sobrecarga de virtualização, falhas nos recursos virtuais e diferentes número de contêineres e tamanho de buffer. Resultados mostram que a disponibilidade, o tempo de resposta e o consumo de energia são fortemente impactados pelo número de contêineres, enquanto o tamanho de buffer afeta principalmente os tempos de resposta.*

1. Introdução

A Computação de Borda de Acesso Multíplo (MEC) e a Virtualização de Funções de Rede (NFV) são tecnologias-chave em Redes da Quinta Geração (5G) para o suporte de serviços como os de Comunicação Ultra Confiável e com Baixa Latência (URLLC) e os

de Banda Larga Móvel Melhorada (eMBB) [Siddiqui et al. 2023]. Enquanto o MEC possibilita a hospedagem de funções de rede, dados e aplicações mais próxima dos usuários finais, reduzindo a latência e aprimorando a confiabilidade geral, a NFV proporciona flexibilidade na disposição das funções de rede, agora virtualizadas (VNFs), e na alocação dinâmica de recursos, alinhando a capacidade da rede com as oscilações da demanda, além de permitir que as VNFs sejam posicionadas na borda da rede, melhorando os tempos de resposta [Setayesh and Bahrami 2022]. Embora permitam diversas aplicações, a operação simultânea dos serviços eMBB e URLLC introduz vários desafios para satisfazer os requisitos contrastantes dessas categorias de serviços [Bairagi et al. 2021], especialmente na alocação dinâmica de recursos no domínio MEC-NFV.

Apesar da literatura endereçar a coexistência de diferentes tipos de serviço em redes 5G, a maioria dos trabalhos trata da alocação de recursos de rádio na rede de acesso de rádio (RAN) [Bairagi et al. 2021] - [Kim and Park 2020], deixando lacunas quanto a fatores relevantes de aprovisionamento de recursos além da RAN. Pesquisas anteriores geralmente pressupõem a rede de núcleo (Core) ou nó MEC (borda) como ambientes em nuvem sem falhas [Li and Jin 2021] ou com tempo instantâneo de aprovisionamento de recursos [Tong et al. 2020], que podem não corresponder a realidade da rede 5G. Além disso, a maioria dos estudos não considera que dentro de cada categoria há aplicações que se diferem amplamente [Liu et al. 2022] e negligenciam fatores como a sobrecarga da virtualização. Por exemplo, na inicialização de uma instância de VNF, a energia é consumida e os recursos são alocados, porém os serviços permanecem sem atendimento, impactando o tempo de resposta e tornando o suporte a serviços críticos desafiador.

Este artigo aborda a combinação de MEC, NFV e alocação dinâmica de recursos virtuais para serviços URLLC e eMBB. Para isso, propõe um modelo baseado em Cadeias de Markov de Tempo Contínuo (CTMC) para caracterizar a alocação dinâmica de recursos virtuais para ambos os serviços e analisar a disponibilidade e tempo de resposta dos servicos bem como o consumo energético do nó MEC. Devido a sua localização descentralizada, os nós MEC possuem uma capacidade de processamento limitada que deve ser capaz de atender aos requisitos de ambos os tipos de serviço. Entretanto, nem sempre é viável optar pela opção de alocação bruta de recursos, pois isso pode incorrer na elevação dos custos de operação e na penalização da qualidade de serviço. Nesse sentido, os impactos da variação conjunta dos recursos de processamento ativo (número de contêineres disponíveis no sistema) e passivo (tamanho do buffer de admissão de cada categoria de serviço) na qualidade do atendimento aos serviços e no consumo de energia do nó MEC-NFV são analisados. Ademais, aspectos práticos que podem impactar no atendimento dos requisitos das aplicações, tais como falhas no recurso de processamento, priorização de serviços e tempos de configuração e reparo, foram incorporados ao modelo. Dessa forma, o modelo proposto pode subsidiar o operador de rede no dimensionamento do nó MEC-NFV para suportar a coexistência dos serviços URLLC e eMBB.

Resultados mostram que cargas mais altas de serviços URLLC diminuem a disponibilidade do sistema e aumentam os tempos de resposta de ambos os tipos de serviços. Por outro lado, esse efeito pode ser mitigado escalonando mais contêineres para atendimento, ao custo de um aumento do consumo energético. Quanto a energia, notou-se que o aumento do número de contêineres não implica necessariamente em um aumento proporcional do consumo energético. Já o aumento no tamanho do buffer melhora le-

vemente a disponibilidade do serviço correspondente, impactando minimamente no consumo energético do sistema, pois as solicitações em buffer não consomem recursos enquanto estão na fila. Por outro lado, o aumento das posições de buffer causa um impacto negativo no tempo de resposta dos serviços. Este artigo encontra-se assim organizado. A Seção 2 apresenta os trabalhos relacionados. O modelo proposto baseado em CTMC para o nó MEC-NFV, considerando as sobrecargas da virtualização, alocação dinâmica de recursos para os serviços URLLC e eMBB e priorização não preemptiva no atendimento, é descrito na Seção 3. A Seção 4 apresenta a validação do modelo e análise de resultados obtidos por extensas simulações de eventos discretos e, finalmente, a Seção 4 conclui este artigo e aponta direções futuras.

2. Trabalhos Relacionados

A literatura tem endereçado diferentes classes de problemas na alocaçao de recursos em sistemas MEC-NFV, incluindo escalonamento, alocação dinâmica e dimensionamento de recursos de rádio e computacional [Li et al. 2021]. Esta seção sintetiza os principais estudos nesse campo, destacando os problemas abordados, os segmentos de rede envolvidos e as categorias de serviço 5G consideradas. Além disso, as ferramentas analíticas utilizadas nos estudos são apontadas.

O compartilhamento de recursos de rádio entre duas categorias de serviço 5G, eMBB e URLLC é abordado em [Bairagi et al. 2021] e [Kim and Park 2020]. O primeiro usa programação combinatorial e considera o compromisso entre latência, confiabilidade e eficiência espectral. O segundo propõe um esquema que realoca parte dos recursos alocados para o tráfego eMBB para serem utilizados pelos pacotes URLLC, resultando em danos ao fluxo eMBB. Os autores estendem o método proposto pela *International Telecommunication Union Radiocommunication* (ITU-R) para refletir a punição do URLLC no eMBB e consideram atrasos adicionais decorrentes de retransmissões, utilizando teoria de filas.

Os trabalhos de [Tong et al. 2020] e [Huang et al. 2021] exploram características fim-a-fim, abrangendo a RAN e o Core. Entretanto, eles consideram apenas uma única categoria de serviço, o URLLC. Em particular, [Tong et al. 2020] desenvolve um algoritmo de alocação dinâmica de recursos que minimiza o atraso fim-a -fim, garantindo uma taxa mínima de serviço e confiabilidade máxima, considerando o mapeamento do VNF na rede de núcleo e de acesso. Da mesma forma, [Huang et al. 2021] propõe um paradigma de rede 5G com NFV para aplicações industriais, garantindo o URLLC por meio de aceleração da cadeia de serviços e compartilhamento dinâmico de recursos de espectro baseado em blockchain entre várias aplicações em execução em equipamentos baseados em NFV.

Já os trabalhos de [Emara et al. 2021] e [Li and Jin 2021] focam nas funções do núcleo de rede, abstraindo as características da RAN, e consideram apenas uma categoria de serviço, não contemplando a coexistência de duas ou mais categorias de serviço 5G. Por exemplo, em [Emara et al. 2021], os autores propõem um modelo baseado em CTMC associado com um problema de otimização para determinar o número ideal de recursos virtuais para maximizar a capacidade de execução de tarefas. O artigo considera a comunicação baseada em contenção para execução remota de tarefas e computação paralela, bem como a ocupação de recursos de processamento no nó MEC propensos a falhas.

Por fim, [Li and Jin 2021] define uma estratégia de execução remota de tarefas nos sistemas MEC para melhorar a qualidade da experiência e aumentar a eficiência energética.

Em termos de virtualização de recursos computacionais para a execução das NFVs e aplicações de rede, o uso de contêiner tem recebido atenção da indústria e da academia e um desafio na sua utilização em infraestruturas MEC-NFV para comunicações móveis está na sua maturidade para esse domínio. Por adotar um kernel compartilhado do sistema operacional, a contêinerização introduz desafios de segurança. Por exemplo, o isolamento de falhas dentro dos contêineres pode não ser trivial e uma falha ser replicada nas outras instâncias. Além de falhas, dois outros fenômenos precisam ser considerados na sua adoção: a instanciação do VNF, que representa o atraso até que uma VNF esteja pronta para processar uma solicitação após ser desligada e um tempo de reparo, que denota a duração tomada para uma VNF para se recuperar de um evento de falha. Embora sejam fatores importantes a serem levados em conta na adoção desta tecnologia em redes 5G para prover os diferentes tipos de serviços, há trabalhos que não os consideram, conforme observado em [Bairagi et al. 2021] e em [Li and Jin 2021]. Negligenciar esses fatores pode ser problemático, pois eles impactam no atendimento de requisitos de confiabilidade e tempo de resposta da aplicação. Por exemplo, se uma estratégia de dimensionamento de recursos não considera a possibilidade de falhas nos recursos, é provável que o robustez computacional do nó resultante seja subestimado. Além disso, mesmo os estudos que consideram eventos de falha, negligenciam o tempo de reparo, como visto em [Liu et al. 2022], [Huang et al. 2021] e em [Zhang et al. 2021]. Essa omissão pode afetar as métricas de desempenho do sistema, como a disponibilidade de recursos e o consumo de energia. Diferentes dos descritos acima, o modelo aqui proposto incorpora tais eventos e considerações apontadas em [Abdelhadi et al. 2022], [Falcao et al. 2022], [Falcao et al. 2023] e [Souza et al. 2021] em um sistema MEC-NFV com coexistência de serviços URLLC e eMBB.

3. Modelo do Sistema

A Figura 1 ilustra o funcionamento do nó MEC considerado nesse trabalho. As solicitações eMBB (fluxo azul) e URLLC (fluxo vermelho) originadas nos Equipamentos do Usuário (UEs) são processadas pela RAN, passadas para o nó MEC e processadas por VNFs hospedadas em contêineres, que são escalonados sob demanda. O trabalho considera as incertezas dos próprios componentes virtuais do nó, como os eventos de configuração, falha e reparo, de forma isolada da RAN, Core e nuvem central. O nó MEC-NFV consiste em uma quantidade finita de contêineres e posições de buffer, que podem ser alocadas para cada tipo, com cada VNF funcionando de forma igual e independente em um único contêiner e onde uma unidade de controle centralizada determina a admissão das solicitações. Uma admissão ocorre se houver recurso suficiente (contêineres disponíveis ou posições no buffer). Nesse caso, cada solicitação pode ser processada ou colocada na fila.

Para lidar com variações de carga, uma estratégia de escalonamento automático dinâmico das VNFs foi incorporada à formulação do sistema. Assim, antes do estágio de processamento, o contêiner com a VNF deve ser inicializado, ficando sujeito a um atraso (tempo de configuração). Além disso, as falhas podem ocorrer durante o processamento do serviço e seu respectivo tempo de reparo também é incorporado. Nesse caso, o contêiner com a VNF é reiniciado e a solicitação é realocada para outro contêiner dis-

ponível ou, se não houver recursos disponíveis, será colocado de volta em sua respectiva fila de serviço, tendo maior prioridade do que as novas solicitações. Nos dois casos, o processamento do serviço é reiniciado.

Além disso, devido aos requisitos críticos de latência dos serviços URLLC, adotou-se a seguinte estratégia de priorização de: (1) Se houver serviços URLLC e eMBB a serem servidos, os serviços URLLC têm maior prioridade; portanto, os contêineres que estão sendo liberados ou ativados são alocados primeiro aos serviços URLLC. (2) No caso em que existe um serviço de URLLC esperando na fila para obter recursos disponíveis e um serviço eMBB foi concluído, o contêiner liberado será reiniciado para ser usado pelo serviço URLLC. No entanto, se houver outros contêineres disponíveis, o atual será alocado para um serviço eMBB sequencial ou desativado se a fila eMBB estiver vazia. (3) A preempção do serviço de prioridade inferior (eMBB) que está sendo processada não é permitida.

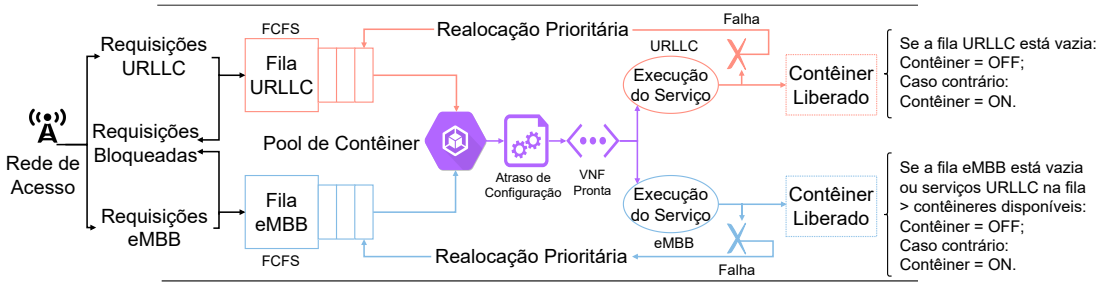

Figura 1. Nó de borda

O sistema é modelado usando uma fila M/N/c/k+K com dois tipos de usuários, priorização, falha, tempo de inicialização, disciplina de serviço Primeiro a Chegar Primeiro a ser Atendido (FCFS) e um buffer limitado para cada tipo de usuário. Os modelos analíticos servem como ferramentas valiosas para avaliar os projetos de infraestrutura de nós MEC distribuídos em larga escala, uma vez que a simulação e os dados de teste, que requerem milhares de nós de borda, nem sempre podem ser viáveis ou estão disponíveis. Os estados modelo são denotados pela tupla (i, j, l, m), onde $i, j, l, m \in N$, com i e j representando o número de serviços URLLC e eMBB e l e m o número de contêineres ativos para cada categoria de usuário, com $l+m$ sendo menor ou igual ao número máximo de contêineres (c).

As chegadas de serviço seguem um processo de Poisson com taxas λ_u e λ_e para os serviços URLLC e eMBB, respectivamente. O processamento de serviço é realizado pelos c contêineres disponíveis, com um tempo de serviço exponencialmente distribuído com taxas μ_u para URLLC e μ_e para eMBB. Da mesma forma, a ocorrência de falha e o tempo de inicialização do contêiner seguem distribuições exponenciais com taxas γ e α, respectivamente. A Fig. 2 resume as transições e estados da CTMC, com seus respectivos parâmetros. Através das probabilidades de estado estacionário (π) do modelo, métricas podem ser derivadas para analisar o desempenho do sistema conforme a seguir.

3.1. Disponibilidade

A combinação MEC e NFV tem sido amplamente reconhecida por seu potencial de reduzir a latência e aumentar a confiabilidade, colocando as funções e aplicações de rede virtualizados mais próximos do UE. No entanto, os recursos limitados dos nós de borda

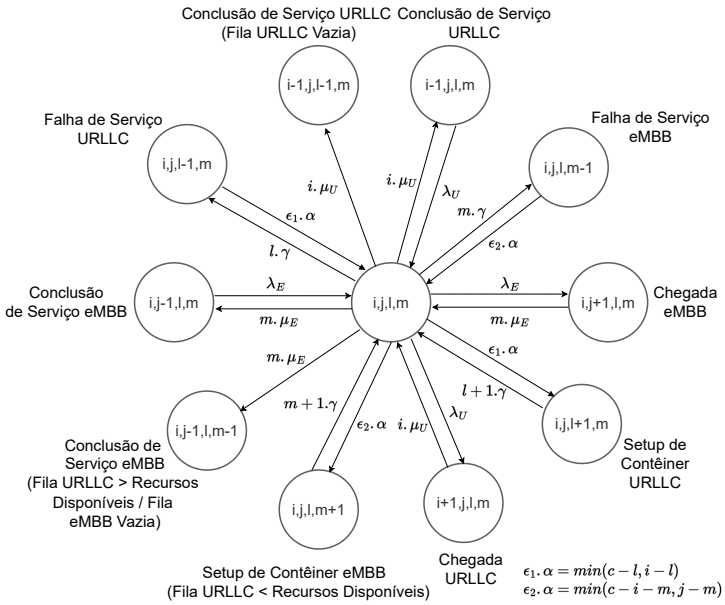

Figura 2. Diagrama de Estados Genérico da CTMC que descreve o nó MEC-NFV

impõem restrições à sua capacidade de serviço, que geralmente é conhecida como disponibilidade. Consequentemente, quando a capacidade máxima é atingida, duas alternativas podem ser utilizadas para o tratamento do fluxo: encaminhar o fluxo excedente para um nó MEC vizinho ou redirecioná-lo para a nuvem central [Sarrigiannis et al. 2020]. Essas alternativas envolvem o estabelecimento de uma nova rota compreendendo vários saltos intermediários, que podem introduzir incerteza significativa na latência e confiabilidade. Assim, a análise da disponibilidade de nós de borda se torna essencial para garantir o atendimento dos requisitos da aplicação. No modelo proposto, a disponibilidade do MEC refere-se à capacidade do sistema de oferecer a quantidade mínima de VNFs funcionais e acessíveis ou posições de buffer para o serviço desejado. Como o uso da priorização do serviço, a disponibilidade do nó MEC é segmentada em termos de cada tipo de serviço, ou seja, URLLC (A_U) e eMBB (A_E), respectivamente, como nas Eqs. 1, que são obtidas somando as probabilidades de todos os estados do modelo, exceto aqueles que representam capacidade total do nó para cada tipo de serviço atingida.

$$A_U = 1 - \sum_{j=0}^{K} \sum_{l=0}^{c} \sum_{m=0}^{min(c-l,j)} \pi_{k,j,l,m}; \quad A_E = 1 - \sum_{i=0}^{k} \sum_{m=0}^{c} \sum_{l=0}^{min(c-m,i)} \pi_{i,K,l,m} \qquad (1)$$

3.2. Tempo de Resposta

O tempo de resposta assume um papel crucial nas aplicações URLLC, embora também seja relevante para as eMBB. Reconhecendo que a sua importância pode variar de acordo com a categoria de serviço, definiu-se o tempo de resposta para a cada categoria. Ele é definido como o intervalo entre a chegada do serviço (no nó MEC-NFV) e sua conclusão, que inclui quaisquer tempos de configuração/reinicialização, caso estes eventos ocorram. As Eqs. 3 denotam os tempos de resposta para os serviços URLLC e eMBB, respectivamemte, os quais são calculados dividindo o número médio de serviços on-line no sistema

para cada categoria (Eqs. 2, com $\overline{U}_U$ e $\overline{U}_E$ para URLLC e eMBB) e suas respectivas taxas de admissão no nó MEC-NFV.

$$\overline{U}_U = \sum_{i=0}^{k}\sum_{j=0}^{K}\sum_{l=0}^{min(c,i)}\sum_{m=0}^{min(c-l,j)} i\pi_{i,j,l,m}; \quad \overline{U}_E = \sum_{i=0}^{k}\sum_{j=0}^{K}\sum_{l=0}^{min(c,i)}\sum_{m=0}^{min(c-l,j)} j\pi_{i,j,l,m} \tag{2}$$

$$T_U = \frac{\overline{U}_U}{\lambda_U A_U}; \quad T_E = \frac{\overline{U}_E}{\lambda_E A_E} \tag{3}$$

3.3. Consumo Energético

O consumo de energia computacional é um componente importante dos custos operacionais do provedor de serviços. Neste modelo, o consumo médio de energia ($\overline{PC}$) é formado a partir da combinação do número médio de recursos virtuais (contêineres) e constantes de consumo de energia para cada estado operacional: em setup (P^{CT}_{setup}) e ocupado ($P^{CT}_{ocupado}$). A Eq. 4 captura a quantidade média de contêineres no estado ocupado, iterando sobre cada estado de carga do sistema e variando a combinação do número de cada tipo de contêiner de 0 até o número de serviços de uma categoria específica ou os recursos máximos disponíveis no sistema. Além disso, a Eq. 5 calcula o número médio de contêineres de VNFs em configuração, iterando nos estados em que o número de serviços no sistema é maior que o número total de recursos ativos para cada categoria de serviço. Finalmente, o consumo total médio de energia ($\overline{PC}$) é dado pela Eq. 6.

$$\overline{CT}_{busy} = \sum_{i=0}^{k}\sum_{j=0}^{K}\sum_{l=0}^{min(c,i)}\sum_{m=0}^{min(c-l,j)} (l+m)\pi_{i,j,l,m} \tag{4}$$

$$\overline{CT}_{setup} = \sum_{i=0}^{k}\sum_{j=0}^{K}\sum_{l=0}^{min(c,i)}\sum_{m=0}^{min(c-l,j)} min((c-l-m),(i+j-l-m))\pi_{i,j,l,m} \tag{5}$$

$$\overline{PC} = P^{CT}_{setup}\overline{CT}_{setup} + P^{CT}_{busy}\ \overline{CT}_{busy} \tag{6}$$

4. Validação e Análise de Resultados

Os resultados analíticos (linhas) foram validados com simulações de eventos discretos (marcadores) (Figuras 3 - 12). Os principais parâmetros foram definidos seguindo a especificação técnica do 3GPP 16 (TR 38.824) [3GPP 2020]. Cada cenário a seguir avalia simultaneamente a influência de um par de parâmetros: o primeiro cenário (Seção 4.1) avalia o impacto do número total de contêineres (c) e o tamanho do buffer para usuários do eMBB (K), que mostra como o aumento da capacidade de processamento paralelo do sistema afeta seu custo energético e a qualidade do serviço assim como as implicações de aumentar a capacidade do sistema de admitir um número maior de serviços eMBB; e segundo cenário (Seção 4.2) analisa o número total de contêineres (c) e o tamanho do buffer para usuários de URLLC (k). Esta seção compartilha um objetivo semelhante a

anterior, mas concentra-se no impacto de expandir a capacidade do sistema para acomodar as solicitações de serviço URLLC. Em todos os cenários, a chegada do serviço URLLC (λ_U) variou de $2,5$ a 25 solicitações/ms para analisar o desempenho do sistema sob diferentes cargas de URLLC. Os valores de base para a taxa de falha (γ) e taxa de configuração (α) foram definidas como $0,001$ e 1 unidade/ms, respectivamente, de acordo com [Kaur et al. 2017]. Os parâmetros restantes podem ser encontrados na tabela 1. Os resultados exibidos representam a média de cada métrica, considerando 10 instâncias de simulaçao, compreendendo 2.700.000 passos de simulação e 2.200.000 atendimentos de serviço cada, com um intervalo de confiança de 95%. Os limites dos intervalos foram muito estreitos, sendo suprimidos dos gráficos para evitar a sobrecarga das imagens.

Tabela 1. Configurações dos Experimentos

Seção	Parâmetros Variados	λ_E	α	γ	μ_U	μ_E	C	K	k
4.1	c, K	10	1	10^{-3}	2	2	4,8,12	16,24	20
4.2	c, k	10	1	10^{-3}	2	2	4,8,12	20	16,24

4.1. Efeitos da variação do número de contêineres (c) e do tamanho do buffer eMBB (K)

Este cenário avalia o impacto da variação do número de contêineres (c) no nó MEC-NFV juntamente com o tamanho do buffer para usuários eMBB (K). Nas Figuras 3 e 4 nota-se que o número de contêineres tem um impacto significativo na disponibilidade para ambas as classes de serviço, mostrando maior disponibilidade para ambientes com maior número de contêineres, representado pelas configurações onde $c = 12$ (linhas vermelhas e laranja), seguido por $c = 8$ (verde e amarelo). Por exemplo, na Figura 3, quando o $\lambda_U = 10$, a disponibilidade do eMBB para as configurações com $c = 8$ é de aproximadamente 20% enquanto para as configurações com $c = 12$ é de cerca de 69%, ou seja, uma diferença absoluta de quase 49%. Por outro lado, as alternativas de buffer testadas tiveram pouco impacto na disponibilidade do eMBB, indicando que exigiriam valores muito maiores que os adotados ($K = 16$ e $K = 24$). No entanto, isto pode ser viável, uma vez que o tamanho do buffer impacta no tempo de resposta, avaliado em seguida. Quanto à disponibilidade do URLLC no sistema (Figura 4), a análise é similar ao eMBB, ou seja, o número de contêineres impacta drasticamente na disponibilidade, enquanto os tamanhos do buffer do URLLC tiveram pouco efeito, resultando em pares de curvas sobrepostas: azul claro/escuro, verde/amarelo e vermelho/laranja.

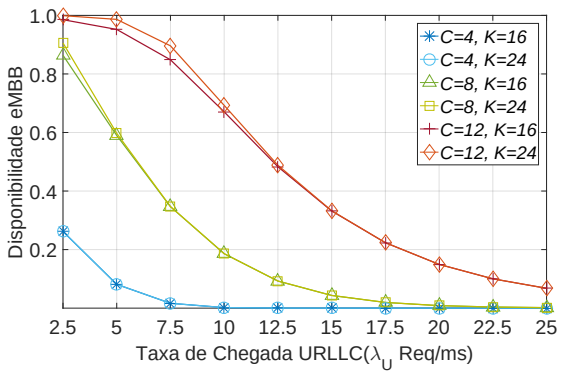

Figura 3. Disponibilidade eMBB

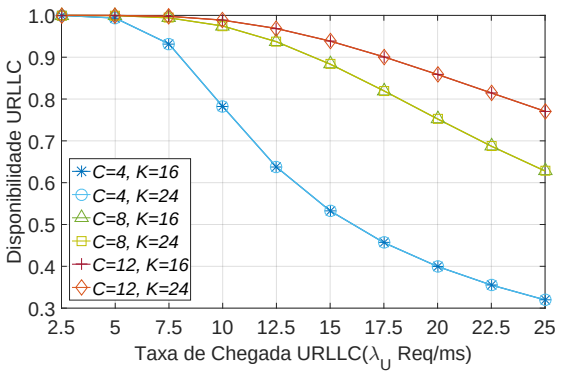

Figura 4. Disponibilidade URLLC

Na Figura 5, um tamanho de buffer maior para a categoria de serviço eMBB também resulta em um aumento no tempo de resposta. Isso se deve ao número de solicitações de serviço enfileiradas antes de cada solicitação eMBB recém-admitida. Por outro lado, um maior número de contêineres disponíveis implica também um menor tempo de fila, reduzindo a contribuição desta componente para o tempo de resposta. Pode-se observar que o subdimensionamento do número de contêineres pode inviabilizar o serviço para usuários de menor prioridade, resultando em grandes tempos de resposta. Por exemplo, as configurações onde $c = 4$ (linhas azuis claras e escuras) mostram-se adequadas para o serviço de Smart Office, que requer latência máxima de 10 ms [Stallings 2021], somente quando $\lambda_U = 2,5$. Em contraste, as configurações restantes podem acomodar esta aplicação com um λ_U tão alto quanto $12,5$.

Similarmente, a variação do tamanho do buffer eMBB também tem pouco impacto no tempo de resposta do URLLC, conforme mostra a Figura 6. Nota-se, portanto, que tempo de resposta é predominantemente impactado pela variação do número de contêineres. Além disso, percebe-se que somente as configurações de sistema com $c = 8$ e $c = 12$ são capazes de atender serviços de Robótica, pois mesmo para $\lambda_U = 2,5$, que é a menor carga de URLLC avaliada nos experimentos, as configurações com $c = 4$ apresentaram tempo de resposta superior a 1 ms. Apesar disso, as configurações com $c = 4$ apresentaram tempo de resposta inferior a 2 ms para todos os λ_U avaliados, mostrando-se capazes de atender o serviços de Sistema de Transporte Inteligente, que permitem latências entre 10 e 100 ms [Siddiqui et al. 2023].

Uma especificidade pode ser encontrada na parte mais à esquerda da Figura 6, onde as curvas com $c = 8$ (linhas verdes e amarelas) e $c = 12$ (linhas vermelhas e laranja) apresentam inicialmente um tempo de resposta decrescente, e, no caso de $c = 8$ ele cresce novamente, atingindo o mesmo valor inicial em $\lambda_U = 25$. Provavelmente, isso se deve ao tempo de configuração do contêiner, que está atendendo às solicitações eMBB ou desligado, considerando a baixa demanda de URLLC de $\lambda_U = 2,5$ até $\lambda_U = 10$. Por outro lado, à medida que a taxa de chegada de URLLC aumenta, pode-se observar uma diminuição no tempo de resposta das solicitações de serviço. Isso ocorre porque mais contêineres ficam disponíveis para atendimento, reduzindo o tempo de espera em relação ao atraso na configuração do contêiner.

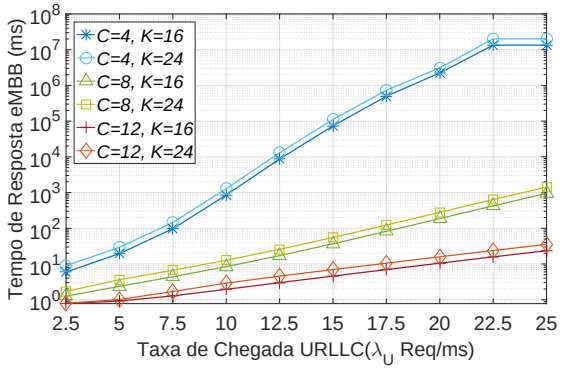

Figura 5. Tempo de Resposta eMBB

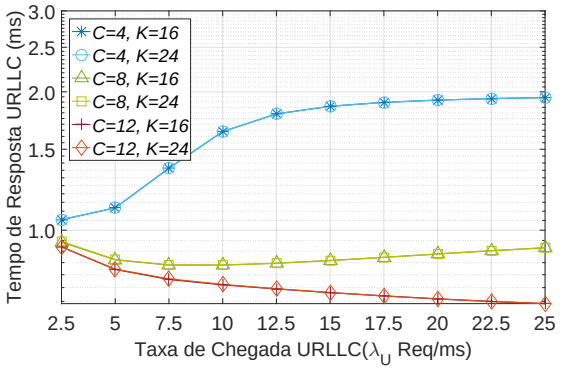

Figura 6. Tempo de Resposta URLLC

Ao contrário do tempo de resposta, uma maior quantidade de contêineres implica inevitavelmente em maior consumo de energia (Figura 7). O consumo de energia não é

exatamente proporcional ao aumento do número de contêineres. Observando o ponto onde o $\lambda_U = 10$ pode-se constatar que entre as curvas azul e verde o número de contêineres dobra de 4 para 8, mas o mesmo não ocorre com o consumo de energia que aumenta cerca de 70%. Isso ocorre porque a quantidade de contêineres em processamento também depende da carga de trabalho que chega ao sistema, ou seja, o consumo de energia só dobraria junto com a quantidade de contêineres se a demanda de atendimento do sistema fosse suficiente para ocupar todos os contêineres disponíveis nas duas configurações do sistema.

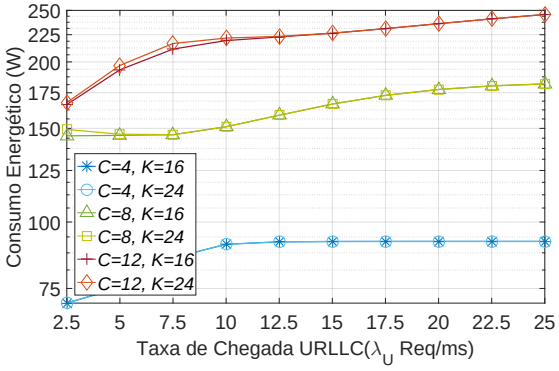

Figura 7. Consumo Energético

Contudo, há pouca diferença no consumo de energia comparando cada par de configurações com as mesmas quantidades de contêineres, isto é, diferentes tamanhos de buffer. Um buffer eMBB maior resulta apenas em um consumo de energia ligeiramente maior porque mais usuários tendem a esperar na fila. Isso evita que o contêiner seja desligado e reiniciado, resultando em menos tempo de configuração e mais tempo de processamento, consumindo mais energia fazendo processamento.

4.2. Efeitos da variação do número de contêineres (c) e do tamanho do buffer URLLC (k)

Esta seção avalia o impacto do número de contêineres (c) no nó MEC-NFV junto com o tamanho do buffer para serviços URLLC (k). No que diz respeito à disponibilidade (Figuras 8-9) tem-se considerações muito semelhantes às realizadas no cenário anterior. Contudo, na Figura 8 é perceptível que há uma inversão na ordem das curvas (do maior tamanho de buffer para o menor), o que é claramente mostrado comparando as curvas com $c = 12$ (vermelho e laranja). Neste caso, a curva vermelha, que possui menos posições de buffer URLLC ($k = 16$) apresenta maior disponibilidade para serviços eMBB do que a laranja ($k = 24$). Isso acontece porque à medida que mais solicitações URLLC são armazenadas, há uma garantia de que elas serão atendidas em vez de descartadas, como acontece com a curva com menos posições de buffer URLLC.

Assim, a carga geral do URLLC aumenta, diminuindo a disponibilidade experimentada pelo serviço eMBB. Por outro lado, na Figura 9, a configuração denotada pela curva laranja ($k = 24$) apresenta uma disponibilidade maior que a vermelha ($k = 16$). Isso era esperado uma vez que a métrica avaliada é a disponibilidade do sistema para serviços URLLC, ou seja, um tamanho de buffer URLLC maior aumenta a disponibilidade para o URLLC, de modo que quando a soma das taxas de chegada para ambas as categorias de serviço se aproxima da capacidade total de processamento, um tamanho de buffer maior implica maior disponibilidade.

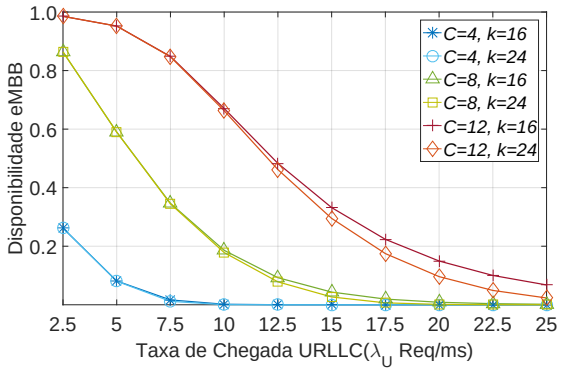

Figura 8. Disponibilidade eMBB

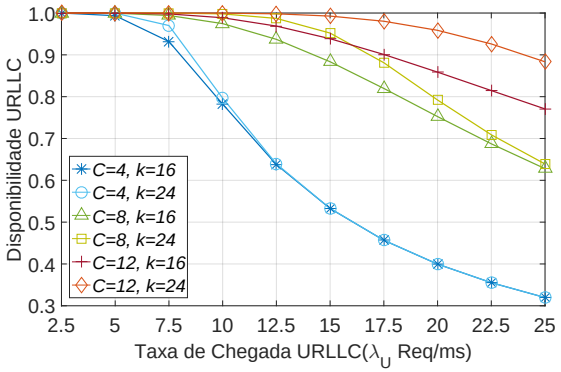

Figura 9. Disponibilidade URLLC

Quanto aos tempos de resposta, representados nas Figuras 10-11, é evidente que o tamanho maior do buffer URLLC exerce um impacto negativo significativo nos tempos de resposta eMBB e URLLC. No entanto, este impacto pode ser atenuado aumentando o número total de contêineres, conforme apresentado nas curvas de ambas as figuras. Na parte mais à esquerda da Figura 10 ($\lambda_U = 2,5$), o tempo de resposta do eMBB permanece abaixo de 10 ms para todas as configurações testadas, embora com taxas de crescimento variáveis. Por exemplo, as curvas correspondentes a $c = 4$ apresentam crescimento exponencial, enquanto aquelas associadas a $c = 12$ apresentam aumento linear. Consequentemente, maiores quantidades de contêineres resultam em melhores tempos de resposta do eMBB, especialmente sob cargas URLLC mais altas, que se aproximam da capacidade do sistema.

Neste cenário, nota-se configurações de sistema com $c = 8$ e $c = 12$ podem efetivamente atender às demandas de serviços de Realidade Virtual e Aumentada, que necessitam de latência de até 8 milissegundos [Raca et al. 2020], para valores de λ_U tão altos quanto 10, enquanto configurações com $c = 4$ só podem acomodar esses serviços para valores λ_U iguais a 2,5, ou seja, baixa carga de tráfego URLLC. Por outro lado, buffers URLLC maiores levam a tempos de resposta eMBB degradados, como evidenciado pelas curvas com $k = 24$ apresentando tempos de resposta mais elevados em comparação com suas respectivas contrapartes com $k = 16$. Além disso, observa-se a partir de $\lambda_U = 17,5$, a curva azul clara (representando $c = 4$ e $k = 24$) assume valores inviáveis (magnitude muito alta). Esta ocorrência é atribuída à pressão intensificada das chegadas de URLLC, juntamente com a adoção de um grande tamanho de buffer, resultando em um tempo de resposta excessivamente elevado do serviço eMBB.

Na Figura 11, o intervalo de possíveis valores de tempo de resposta do URLLC é consideravelmente menor do que o do experimento anterior, devido à maior prioridade concedida às solicitações URLLC. No entanto, há uma variação considerável no comportamento de cada configuração (curva). Algumas curvas apresentam comportamento estritamente ascendente, enquanto outras apresentam fases descendentes e ascendentes. Além disso, uma curva exibe um padrão estritamente descendente. Entretanto, a ordem das curvas em termos de tempo de resposta do URLLC permanece consistente com o experimento anterior (Figura 10). Vale ressaltar que, para um intervalo maior de λ_U, espera-se que as curvas apresentem comportamento semelhante, mas com pequenos deslocamentos.

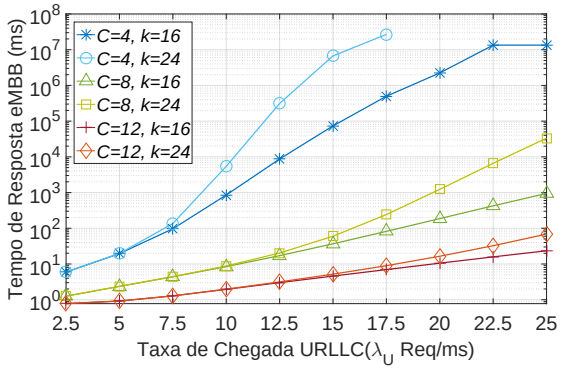

Figura 10. Tempo de Resposta eMBB

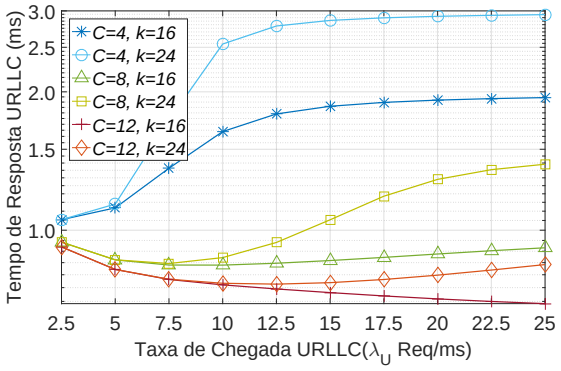

Figura 11. Tempo de Resposta URLLC

Em relação às curvas azul claro e escuro, pode-se inferir que a capacidade do sistema é atingida rapidamente, resultando em tempos de resposta URLLC mais elevados à medida que o buffer se torna mais utilizado. Todavia, mesmo nestes casos, o tempo de resposta do URLLC permanece em um nível aceitável de 3 ms, o que é altamente adequado para a maioria das aplicações URLLC que normalmente requerem tempos de resposta que variam de até 10 ms, como Automação Fabril [Siddiqui et al. 2023]. Quanto à curva estritamente descendente (em vermelho), é provável que o tempo de resposta do URLLC diminua com as novas chegadas de URLLC sendo prontamente processadas por contêineres que estavam anteriormente no modo de configuração, contornando assim o atraso de configuração. Ademais, o tamanho menor do buffer ($k = 16$) leva a menos solicitações na fila de espera, contribuindo assim para um tempo de resposta URLLC geral mais baixo em comparação com configurações com tamanhos de buffer maiores, como $k = 24$ (representado pela linha laranja).

Em termos de consumo de energia, ilustrado na Figura 12, dentro desse cenário, uma observação semelhante pode ser feita em comparação com o experimento anterior. É evidente que um aumento no número de contêineres leva a um maior consumo de energia, alinhando-se com nossas expectativas. Além disso, para a maior parte do quadro de avaliação, o tamanho do buffer URLLC exibe influência mínima nessa métrica de desempenho. Isso é notado a partir do par de curvas sobrepostas, particularmente perceptível para taxas de chegada de URLLC baixas. Esse resultado foi antecipado, pois as solicitações em buffer não consomem recursos enquanto estão na fila. Assim, na maioria dos casos, o tamanho do buffer não afeta significativamente o consumo de energia. No entanto, um ligeiro aumento no consumo de energia é observado quando o sistema se aproxima da capacidade total e utiliza mais posições de buffer. Esse fenômeno ocorre devido aos contêineres ficarem mais tempo em estado de processamento, resultando em períodos reduzidos de estado desligado ou em procedimentos de reinicialização.

5. Conclusão

Este trabalho endereçou alocação dinâmica de recursos virtuais no contexto de sistemas MEC-NFV atendendo serviços URLLC e eMBB. Para isso, um modelo baseado em CTMC englobando fatores práticos, como falhas de recursos, priorização de serviços e tempos de configuração (reparo) foi desenvolvido. Resultados mostraram como as variações fatores como o número de contêineres disponíveis no sistema, o tamanho do buffer para admissão de ambos os serviços e a carga de usuários URLLC podem afetar o

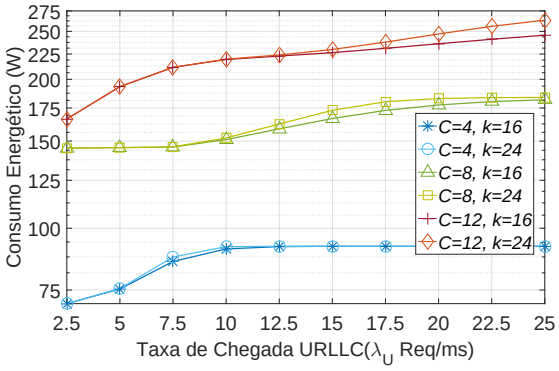

Figura 12. Consumo Energético

atendimento dos serviços de ambos os serviços e o consumo de energia do sistema.

Observou-se que um número de contêineres maior aumenta a disponibilidade e diminuindo os tempos de resposta dos serviços, enquanto o tamanho de buffer afeta principalmente os tempos de resposta. Ademais, o consumo de energia aumenta com o número de contêineres, mas é minimamente afetado por variações no tamanho do buffer. Em suma, o modelo proposto serve como uma ferramenta valiosa para compreender a dinâmica operacional da rede 5G baseada em MEC-NFV no atendimento de diferentes categorias de serviço e subsidia o operador de rede no dimensionamento adequado do sistema MEC-NFV. Como direções futuras, sugere-se a exploração de formulações multiobjetivas para problemas de alocação de recursos dentro do paradigma MEC-NFV, com uma ênfase especial na coexistência dos serviços eMBB e URLLC. Além disso, propõe-se investigar o desenvolvimento de soluções de baixa complexidade para otimizar a alocação de recursos em redes 5G.

Agradecimentos

Este estudo foi financiado pela Coordernaçao de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código 001. Ele contém resultados do Projeto AMAN, executado pelo Sidia Instituto de Ciência e Tecnologia e Samsung Eletrônica da Amazônia LTDA, de acordo com a Lei de Informática n.8387/91 e Art. 39 do Decreto 10.521/2020.

Referências

3GPP (2020). System architecture for the 5g system (5gs). *White Paper*.

Abdelhadi, M., Sorour, S., ElSawy, H., Elsayed, S. A., and Hassanein, H. (2022). Parallel computing at the extreme edge: Spatiotemporal analysis. In *GLOBECOM 2022 - 2022 IEEE Global Communications Conference*, pages 5692–5698.

Bairagi, A. K., Munir, M. S., Alsenwi, M., Tran, N. H., Alshamrani, S. S., Masud, M., Han, Z., and Hong, C. S. (2021). Coexistence mechanism between embb and urllc in 5g wireless networks. *IEEE Transactions on Communications*, 69(3):1736–1749.

Emara, M., ElSawy, H., Filippou, M. C., and Bauch, G. (2021). Spatiotemporal dependable task execution services in mec-enabled wireless systems. *IEEE Wireless Communications Letters*, 10(2):211–215.

Falcao, M., Souza, C., Balieiro, A., and Dias, K. (2022). An analytical framework for urllc in hybrid mec environments. *The Journal of Supercomputing*, 78.

Falcao, M., Souza, C., Balieiro, A., and Dias, K. (2023). Dynamic resource allocation for urllc in uav-enabled multi-access edge computing. *EuCNC 6G Summit*.

Huang, H., Miao, W., Min, G., Tian, J., and Alamri, A. (2021). Nfv and blockchain enabled 5g for ultra-reliable and low-latency communications in industry: Architecture and performance evaluation. *IEEE Transactions on Industrial Informatics*, 17(8):5595–5604.

Kaur, K., Dhand, T., Kumar, N., and Zeadally, S. (2017). Container-as-a-service at the edge: Trade-off between energy efficiency and service availability at fog nano data centers. *IEEE Wireless Communications*, 24(3):48–56.

Kim, Y. and Park, S. (2020). Calculation method of spectrum requirement for imt-2020 embb and urllc with puncturing based on m/g/1 priority queuing model. *IEEE Access*, 8:25027–25040.

Li, C., Cai, Q., Zhang, C., Ma, B., and Luo, Y. (2021). Computation offloading and service allocation in mobile edge computing. *The Journal of Supercomputing*, 77:1–30.

Li, W. and Jin, S. (2021). Performance evaluation and optimization of a task offloading strategy on the mobile edge computing with edge heterogeneity. *The Journal of Supercomputing*, 77(8).

Liu, T., Fang, L., Zhu, Y., Tong, W., and Yang, Y. (2022). A near-optimal approach for online task offloading and resource allocation in edge-cloud orchestrated computing. *IEEE Transactions on Mobile Computing*, 21(8):2687–2700.

Raca, D., Leahy, D., Sreenan, C. J., and Quinlan, J. J. (2020). Beyond throughput, the next generation: A 5g dataset with channel and context metrics. In *Proceedings of the 11th ACM Multimedia Systems Conference*, MMSys '20, page 303–308. Association for Computing Machinery.

Sarrigiannis, I., Ramantas, K., Kartsakli, E., Mekikis, P.-V., Antonopoulos, A., and Verikoukis, C. (2020). Online vnf lifecycle management in an mec-enabled 5g iot architecture. *IEEE Internet of Things Journal*, 7(5):4183–4194.

Setayesh, M. and Bahrami, S. (2022). Resource slicing for embb and urllc services in radio access network using hierarchical deep learning. volume 21.

Siddiqui, M. U. A., Abumarshoud, H., Bariah, L., Muhaidat, S., and Imran, Muhammad, L. (2023). Urllc in beyond 5g and 6g networks: An interference management perspective. *IEEE Access*, 11:54639–54663.

Souza, C., Falcao, M., Balieiro, A., and Dias, K. (2021). Modelling and analysis of 5g networks based on mec-nfv for urllc services. *IEEE Latin America Transactions*, 19(10):1745–1753.

Stallings, W. (2021). *5G Wireless: A Comprehensive Introduction*. Addison-Wesley.

Tong, Z., Zhang, T., Zhu, Y., and Huang, R. (2020). Communication and computation resource allocation for end-to-end slicing in mobile networks. *2020 IEEE/CIC International Conference on Communications in China (ICCC)*, pages 1286–1291.

Zhang, T., Qiu, H., Linguaglossa, L., Cerroni, W., and Giaccone, P. (2021). Nfv platforms: Taxonomy, design choices and future challenges. *IEEE Transactions on Network and Service Management*, 18(1):30–48.