Modelling and Analysis of 5G Networks Based on MEC-NFV for URLLC Services

Caio B. B. de Souza, Marcos R. M. Falcão, Andson M. Balieiro, and Kelvin L. Dias

***Abstract*—Ultra-Reliable and Low Latency Communications (URLLC) is a Fifth Generation (5G) or Beyond(B) 5G of Mobile Networks service that presents strict reliability and latency requirements. Multi-Access Edge Computing (MEC) and Network Function Virtualization (NFV) emerge as essential solutions since they allow network applications/functions to be virtualized and hosted closer to the end-user, improving reliability and latency. However, they may incur delay overheads (e.g., resource setup delay, failure events) to the dynamic resource provisioning that may impair the URLLC, which must be considered in the MEC-NFV node setting. This work addresses the resource provisioning for URLLC services in MEC-NFV-based networks, considering the Virtualized Network Function (VNF) setup time, failure occurrence and resource pre-initialization. A queue-based model is proposed to analyze the MEC-NFV node configuration in terms of average response time, blocking probability, and average number of active resources, under different service arrival and resource setup rates, maximum system capacity, number of resources and pre-initialized ones are carried out. The results show that the resource pre-initialization may mitigate the negative effect of the lower VNF setup rate.**

***Index Terms*—URLLC, MEC, NFV, 5G Network, Queue Theory.**

I. Introdução

A Quinta Geração (5G) de Redes Móveis Sem Fio busca atender demandas que estão além das capacidades dos sistemas atuais, como a densidade alta de conexões da Internet das Coisas e a comunicação com restrição de latência e confiabilidade (ex.veículos autônomos)] [1]. Os serviços 5G diferem nos requisitos de latência, vazão, densidade de conexão, confiabilidade e consumo energético e são categorizados em banda larga móvel melhorada (eMBB), comunicação massiva do tipo máquina (mMTC) e comunicação com confiabilidade muito alta e latência muito baixa (URLLC) [2] [3]. Em consonância com a implantação das redes 5G, a academia e a indústria têm iniciado a concepção da Sexta Geração (6G) de redes móveis sem fio ou *Beyond* 5G (B5G), endereçando questões em aberto da 5ª geração e casos de uso novos (ex. comunicações holográficas 3D e entre veículos aéreos não tripulados) [4]. Dentre as questões em aberto, está o provimento dos serviços URLLC, que demandam requisitos de latência e confiabilidade estritos, da ordem de 1ms para a latência fim-a-fim e 10^{-7} para a probabilidade de perda de pacotes [5], por exemplo. No âmbito da B5G/6G, tais valores tendem a ser mais rigorosos, com a latência na ordem de microssegundos [4]. As restrições dos serviços URLLC introduzem desafios no projeto das redes (B)5G/6G, que requerem melhorias não somente na rede de acesso (RAN) (ex. numerologia, diferentes esquemas de codificação e correção de erro, estrutura do frame variável, uso de ondas milimétricas, comunicação em Terahertz e por luz visível (VLC) [2]), mas também em outros componentes da rede, como a rede de núcleo (CN), geralmente assumida como similar em operação aos dos datacenters comuns.

Nesse sentido, uma alternativa para reduzir a latência é posicionar as funções do núcleo e aplicações na borda da rede, através do uso de Virtualização de Função de Rede (NFV) e Computação de Borda de Acesso Múltiplo (MEC), que possibilita o desacoplamento das funções de rede do hardware dedicado e provê recursos (e.g., contêineres) para que as funções de rede virtualizadas (VNFs) sejam executadas na borda, respectivamente. Entretanto, a transição para MEC pode ser onerosa para o provedor de serviço, especialmente aqueles que endereçam URLLC, pois diferente dos grandes centros de dados atuais, que concentram os recursos em poucas localizações, os nós MEC que processarão serviços URLLC deverão ser distribuídos em vários pontos próximos aos usuários finais para satisfazer aos requisitos estritos de latência e confiabilidade deste tipo de serviço [17]. Em decorrência da distribuição dos recursos e limitações de espaço e custo de cada localidade, é provável que os nós MEC possuam menor potencial de recursos comparados aos centros de dados tradicionais, o que requer o uso eficiente deles.

Trabalhos anteriores têm proposto a ativação e alocação de recursos computacionais sob demanda, que ajuda a minimizar o sub ou sobreprovisionamento de recursos [7]. No entanto, esta não é uma tarefa trivial no contexto de 5G. Especificamente, o custo de operação pode ser reduzido quanto menor for o número de recursos instanciados, mas, por outro lado, o subprovisionamento de recursos pode causar violações do Acordo de Nível de Serviço (SLA), que para serviços URLLC é bastante estrito. Logo, algumas considerações utilizadas no contexto de datacenters comuns para tráfego web como a ausência de falhas no processamento e o tipo host virtual escalonável dinamicamente podem ter um grande impacto sobre o tráfego URLLC.

Este artigo analisa o aprovisionamento de recursos para serviços URLLC executados em nós MEC-NFV considerando eventos adversos típicos dos ambientes virtuais conteinerizados e em contrapartida admitimos o uso da técnica de pré-inicialização de recursos com o intuito de promover o equilíbrio da plataforma, permitindo mais assertividade no dimensionamento de recursos por parte do provedor. Nós consideramos ainda o tempo de inicialização da Função de rede

Os autores são do Centro de Informatica (CIn), Universidade Federal de Pernambuco (UFPE), Av. Jorn. Aníbal Fernandes, Recife, Brasil, e-mail: {cbbs, mrmf, amb4, kld}@cin.ufpe.br.

Este trabalho foi realizado no escopo do projeto "Tecnologias e Mecanismos de Redes Móveis Sem Fio de Quinta Geração (5G) para Suporte a Comunicaçao Ultra Confiavel e com Latência Muito Baixa (URLLC)", financiado pelo CNPQ.

Virtual (VNF) e a possibilidade de falha durante o atendimento de uma requisição que são tipicamente negligenciados em função de não promoverem quebra na qualidade do serviço em *datacenters* tradicionais, mas que para aplicações URLLC, não podem ser ignorados. Para isso, um modelo de fila M/M/C/K com pré-inicialização e falha é proposto e o tempo médio de resposta, a probabilidade de bloqueio e o número médio de recursos ativos são derivados e analisados, sob diferentes cargas e taxas de inicialização (*setup*) de recursos, capacidade máxima do sistema, quantidade de recursos (contêineres) e número de contêineres pré-inicializados. Resultados mostram que a pré-inicialização de contêineres pode mitigar o efeito da taxa de *setup* menor, diminuindo o tempo de espera para atendimento do serviço, e que, a adoção de buffer maior reduz o bloqueio, mas aumenta o tempo de resposta, o que pode causar violação aos requisitos do serviço URLLC.

Este artigo encontra-se assim organizado. A Seção II apresenta trabalhos relacionados. A descrição do modelo proposto e a derivação de métricas são realizadas na Seção III. Validação e análise dos resultados são conduzidas na Seção IV. Seção V conclui este artigo e apresenta direções futuras.

II. Trabalhos Relacionados

O processo de inicialização de uma instância de VNF é um aspecto crucial nos estudos de custos para borda/núcleo da rede 5G, em especial para serviços URLLC, pois violações de nível de serviço (SLA) podem ocorrer caso a VNF não esteja pronta em tempo hábil para iniciar o processamento de fluxos críticos. Alguns trabalhos como [12] ignoram o tempo de inicialização durante o processo de alocação de recursos, ao passo que outros como [7] e [8] o consideram, mas negligenciam outros aspectos como a possibilidade de falha durante o processamento do serviço. Além disso, existem aqueles focados em redes 5G de forma generalista [7] [8], que consideram o núcleo da rede operando e com dimensão e limites de datacenters convencionais (com tráfego web), o que torna a avaliação não representativa para serviços URLLC.

Para fornecer diretrizes teóricas às operadoras de telefonia móvel, os modelos analíticos devem estar alinhados às tendências tecnológicas atuais. Em particular, no contexto 5G, tanto máquinas virtuais (VMs) baseadas em *microkernel* quanto os contêineres baseados em microsserviço tem sido investigados para preencher a lacuna deixada pelas VMs tradicionais que não oferecem um bom compromisso entre desempenho e consumo de recursos físicos. No entanto, os contêineres têm sido amplamente adotados tanto pela academia quanto pela indústria. O trabalho [9] adotou essa suposição na modelagem de desempenho de *datacenter*, mas os autores não trataram de serviços das redes 5G. Eles consideraram três tipos de componentes (Contêineres, Máquinas Virtuais e Máquinas Físicas) escalonáveis. Em termos de alinhamento da tecnologia de virtualização em direção a uma nuvem mais responsiva, acredita-se que o trabalho proposto em [9] seja o mais próximo do nosso, pois fornece uma abordagem sistemática para avaliar a elasticidade da plataforma de microsserviço.

Os autores em [7] adotam VMs como recursos a serem aprovisionados sob demanda. Entretanto, VMs possuem um

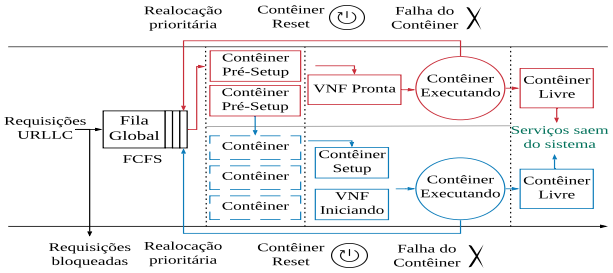


Fig. 1. Sistema URLLC MEC-NFV.

overhead alto de inicialização que inviabiliza os serviços URLLC. Assim, o uso de contêineres baseados em microsservicos para executar as VNFs associado à técnica de pré-inicialização de recurso podem ser uma alternativa para mitigar esse problema. Já em [10], [11] e [13], soluções voltadas para MEC são apresentadas, entretanto somente o último considera tanto o atraso na inicialização dos recursos quanto a possibilidade de falhas durante a operação. Por outro lado, nenhum destes apresenta as métricas indicadas para avaliação de serviços URLLC. O trabalho [12] é um dos poucos que propuseram a avaliação de tráfego URLLC, e, para isso, consideram as métricas de latência e confiabilidade. No entanto, os autores negligenciaram o impacto do tempo de inicialização de recursos.

Embora trabalhos anteriores tenham abordado o aprovisionamento de recursos em redes 5G, nenhum analisa de forma conjunta o impacto da alocação dinâmica considerando a possibilidade de falha durante o processamento do serviço, o atraso de inicialização e pré-inicialização de recurso no atendimento de serviços URLLC em um ambiente MEC-NFV. A Tabela I resume as contribuições dos trabalhos anteriores e as do modelo proposto.

TABELA I
Comparação dos Trabalhos Relacionados

Trabalho	MEC-NFV	URLLC	Falha de VNF	Atraso de Inicialização	Pré-Inicialização de Recursos
[7]	✗	✗	✗	✓	✗
[8]	✗	✗	✗	✓	✗
[9]	✗	✗	✓	✓	✗
[10]	✓	✗	✗	✗	✗
[11]	✓	✗	✗	✗	✗
[12]	✗	✓	✓	✗	✗
[13]	✓	✗	✓	✓	✗
Este Trabalho	✓	✓	✓	✓	✓

III. Modelo do Sistema MEC-NFV

O modelo de sistema é ilustrado na Fig. 1. As solicitações URLLC originadas nos usuários são processadas pela RAN e encaminhadas para o nó MEC-NFV, que executa uma função do núcleo de rede através de uma VNF. Cada VNF executa independentemente em um contêiner e uma unidade de controle decide sobre o bloqueio ou admissão das solicitações. O nó MEC-NFV é composto de $c \in \mathbb{Z}^+$ contêineres, dos quais $n \in \mathbb{Z}^+$ são inicializados previamente ($n < c$), e possui um limite máximo de $k \in \mathbb{Z}^+$ serviços URLLC simultâneos.

A cada chegada de um novo serviço, caso o limite k não tenha sido excedido, a solicitação é admitida no sistema e um contêiner é alocado para o atendimento da demanda (caso haja contêiner ativo e disponível) e um outro contêiner é inicializado (caso haja algum parado) em seguida, visando manter o número de recursos ativos e disponíveis no sistema iguais a n. Caso todos os contêineres já estejam ocupados, os serviços serão colocados em um buffer finito de tamanho $q \in \mathbb{N}$, com $q = k-c$. Durante o atendimento, o contêiner está suscetível a ocorrência de falhas. Nesse caso, ele é reiniciado e o serviço é realocado para um contêiner disponível. Caso não haja, o serviço é posicionado no buffer, tendo maior prioridade no atendimento em relação aos novos serviços. Em ambos os casos, o processamento do serviço é reiniciado.

A ativação da VNF no contêiner compreende a inicialização da imagem do *kernel* e da função de rede especificada, que são englobadas em um único intervalo, denotado como tempo de *setup*, durante o qual recursos e energia são consumidos, mas não há processamento do serviço. Quando a VNF conclui o processamento da solicitação e o número de contêineres pré-inicializados e disponíveis para atendimento é igual a n, o contêiner da VNF que finalizou é parado. O atraso de desligamento de contêiner é abstraído neste trabalho, pois sua magnitude é significativamente menor que o tempo de setup e recuperação de falha [6].

A Fig. 2 apresenta um exemplo do fluxo de operação com $n = 1$, $c = 3$ e $k = 3$. O primeiro evento em t1, é uma solicitação de serviço URLLC, alocada para atendimento no recurso inicializado antecipadamente (CTNR 1). Paralelamente, outro contêiner é inicializado (CTNR 2), visando manter o sistema com um recurso disponível caso outra solicitação chegue. Essa configuração requer um período de espera até que o recurso esteja pronto em t2. Em t3, uma nova solicitação de serviço chega e é alocada para o CTNR2 que já está disponível, enquanto o CTNR3 é inicializado previamente, ficando disponível para atendimento apenas em t4. Em t5, uma falha durante o atendimento força o CTNR2 a reiniciar e mover o serviço atual para outro recurso disponível (CTNR3). Em t6, o CTNR 2 volta a operação e permanece inicializado aguardando que outro serviço chegue para atendimento no sistema. Após terminar o processamento das requisições em t7 e t8, os recursos (CTNR 1 e CTNR 3) são desativados, pois o número de contêineres livres inicializados previamente já foi atingido com o CTNR 2.

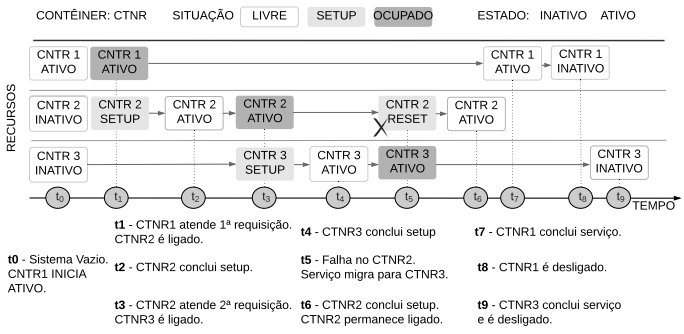


Fig. 2. Exemplo de operação para n = 1, c = 3 e k = 3.

A. Modelo Analítico

O sistema MEC-NFV é modelado através de uma fila M/M/c/k/ com tempo de inicialização (*setup time*), falha de contêineres em atendimento, n contêineres pré-inicializados e disciplina de serviço FCFS regular para os serviço URLLC. A chegada de serviços obedece um processo de Poisson com taxa λ. O atendimento dos serviços é realizado pelos c contêineres idênticos disponíveis no sistema, com tempo de serviço exponencialmente distribuído com taxa μ. Similarmente, o tempo entre falhas e o tempo de inicialização dos contêineres seguem distribuições exponenciais com taxas γ e α, respectivamente. A Fig. 3 apresenta o diagrama de espaço de estados do modelo, com cada estado sendo representado pelo par (i, j), com $i, j \in \mathbb{N}$, onde i denota o número de contêineres inicializados e j o número de serviços no sistema. As transições para a direita (esquerda) indicam a chegada de novos serviços (o término de um serviço sem desligar um contêiner). As transições verticais para baixo (cima) indicam a inicialização de um novo contêiner (a falha do contêiner durante o atendimento). Por fim, as transições na diagonal representam o fim do serviço com o desligamento de um contêiner ocioso conforme a política de pré-configuração utilizada. As transições são compostas por valores fixos e coeficientes gerados dinamicamente, alguns dos quais são definidos pelo operador binário $min(a, b)$, que retorna o menor operando.

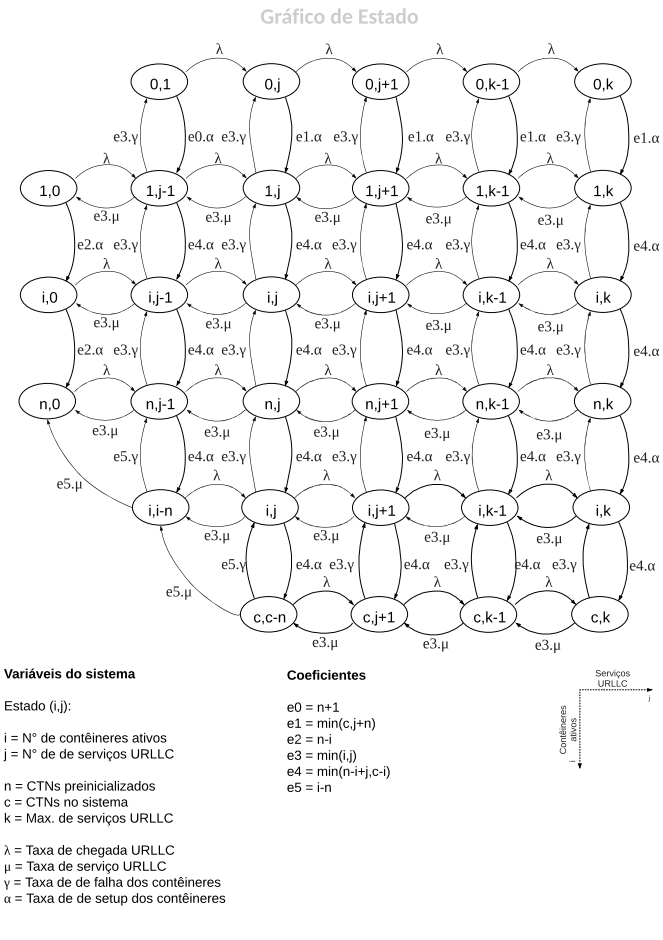


Fig. 3. Diagrama de transição de estado.

TABELA II
DESCRIÇÃO DAS EQUAÇÕES DE BALANÇO

Equação N°	Equação	Estado(s) (i,j)	Condições
(2)	$(\lambda+(n+1)\alpha)\pi_{0,1}=(\gamma)\pi_{1,1}$	$(0,1)$	$(n>0)$.
(3)	$(\lambda+(min(c,j+n)\alpha)\pi_{0,j}=(\lambda)\pi_{0,j-1}+(\gamma)\pi_{1,j}$	$(0,j)$	$(1<j<k)$.
(4)	$(min(c,k+n)\alpha)\pi_{0,k}=(\lambda)\pi_{0,k-1}+(\gamma)\pi_{1,k}$	$(0,k)$	n/a.
(5)	$(\lambda+(n-i)\alpha)\pi_{1,0}=(\mu)\pi_{1,1}$	$(1,0)$	$(n>1)$.
(6)	$(\lambda+(min(i,j)\mu)+(min(n-i+j,c-i)\alpha)+(min(i,j)\gamma))\pi_{i,j}=$ $(\lambda)\pi_{i,j-1}+(min(i,j+1)\mu)\pi_{i,j+1}+(min(n-i+j+1,c-i+1)\alpha)\pi_{i-1,j}+$ $(min(i+1,j)\gamma)\pi_{i+1,j}$	(i,j)	$(0<i<c)$, $(j<k)$ e $(j>max(1,i-n))$.
(7)	$((min(i,k)\mu)+(min(n-i+k,c-i)\alpha)+(min(i,k)\gamma))\pi_{i,k}=$ $(\lambda)\pi_{i,k-1}+(min(n-i+k+1,c-i+1)\alpha)\pi_{i-1,k}+(min(i+1,k)\gamma)\pi_{i+1,k}$	(i,k)	$(0<i<c)$.
(8)	$(\lambda+((n-i)\alpha))\pi_{i,0}=(\mu)\pi_{i,1}+((n-i+1)\alpha)\pi_{i-1,0}$	$(i,0)$	$(n>1)$.
(9)	$(\lambda)\pi_{n,0}=(\mu)\pi_{n,1}+(\mu)\pi_{n+1,1}+(\alpha)\pi_{n-1,0}$	$(n,0)$	$(n>1)$.
(10)	$(\lambda)\pi_{n,0}=(\mu)\pi_{n,1}+(\mu)\pi_{n+1,1}$	$(n,0)$	$(n=1)$.
(11)	$(\lambda+((i-n)\mu)+((i-n)\gamma))\pi_{i,i-n}=$ $((i-n)+min(n,1)\mu)\pi_{i,i-n+1}+((i-n+1)\mu)\pi_{i+1,i-n+1}+(\alpha)\pi_{i-1,i-n}$	$(i,i-n)$	$(n<i<c)$, $(j=i-n)$ e $(c-n>1)$.
(12)	$(\lambda+((c-n)\mu)+((c-n)\gamma))\pi_{c,c-n}=$ $((c-n)+min(n,1)\mu)\pi_{c,c-n+1}+(\alpha)\pi_{c-1,c-n}$	$(c,c-n)$	n/a.
(13)	$(\lambda+(min(c,j)\mu)+(min(c,j)\gamma))\pi_{c,j}=$ $(\lambda)\pi_{c,j-1}+(min(c,j+1)\mu)\pi_{c,j+1}+(\alpha)\pi_{c-1,j}$	(c,j)	$(c-n<j<k)$ e $(k>c-n+1)$.
(14)	$((min(c,k)\mu)+(min(c,k)\gamma))\pi_{c,k}=$ $(\lambda)\pi_{c,k-1}+(\alpha)\pi_{c-1,k}$	(c,k)	$(k>c-n)$.

B. Métricas de Desempenho

O espaço de estados possíveis é dado por $\Omega=\{(i,j|0<i+j,0\leq i\leq c,0\leq j\leq k$, com $i-n\leq j,c\leq k,c\geq 1,k\geq c$ e $n\ \geq 1\}$. Para derivar métricas de desempenho do sistema, a probabilidade dos estados em regime estacionário $\pi(i,j)$ é obtida através da resolução do sistema linear formado pelas as equações de balanço de fluxo (Eqs. 2-14) na Tabela II e a condição de normalização (Eq. 15).

$$\sum_{(i,j)\in\Omega}\pi(i,j)=1 \tag{15}$$

Adotar NFV e MEC para posicionar funções do núcleo da rede e aplicações mais próximas do usuário possibilita reduzir a latência de serviços URLLC. Entretanto, os nós MEC-NFV possuem maior limitação de recursos quando comparados a nuvem central, o que restringe a admissão de serviços/capacidade de atendimento. Quando um serviço não é admitido no nó MEC, ele pode ser direcionado para a nuvem central, experimentando um caminho maior, cruzando nós intermediários, o que pode ocasionar a quebra dos requisitos de latência dos serviços URLLC. Neste aspecto, no dimensionamento/configuração do nó MEC-NFV deve ser analisado o nível de admissão de serviços URLLC. Um bloqueio de serviço acontece quando todos os recursos do nó (contêineres e *buffer*) estão sendo usados durante a chegada da requisição. Assim, a probabilidade de bloqueio (P_B) é obtida através da soma das probabilidade dos estados que representam o sistema cheio, expressa na Eq.16.

$$P_B=\sum_{i=0}^{c}\pi(i,k) \tag{16}$$

Como os serviços URLLC apresentam requisitos estritos de latência de comunicação, a análise do tempo de resposta dos serviços URLLC é crucial na configuração do nó. O tempo médio de resposta (MRT) é dado pela Eq.17, que é a razão entre número de usuários no sistema e a taxa de usuários admitidos. Já o custo de operação é determinado por vários fatores. Um deles é a energia consumida no nó, que é diretamente relacionada à quantidade de recursos ativos. Neste aspecto, no dimensionamento do nó é importante analisar o número de contêineres ativos no sistema (nCTNs), que pode ser obtido através da Eq. 18.

$$MRT=\frac{\sum_{i=0}^{n}\sum_{j=1}^{k}\pi(i,j)j+\sum_{i=n+1}^{c}\sum_{j=i-n}^{k}\pi(i,j)j}{\lambda(1-PB)} \tag{17}$$

$$nCTNs=\sum_{i=1}^{n}\sum_{j=0}^{k}\pi(i,j)i+\sum_{i=n+1}^{c}\sum_{j=i-n}^{k}\pi(i,j)i \tag{18}$$

IV. RESULTADOS

A análise do impacto do escalonamento com pré-inicialização de recursos para os serviços URLLC considerou cenários representativos de um nó MEC-NFV com restrição de recursos atendendo serviços URLLC. Neste aspecto, adotou-se taxas de serviço e chegada dispostos em [14], que descreve um tempo de serviço da rede núcleo de até 1 ms (1 serviço/ms)

e chegadas de serviço com até a 20 solicitações/ms. Destes, consideramos o mesmo valor para a taxa de serviço e a mesma escala para a taxa de chegada em conjunto com uma taxa de falha igual a (γ) 0,001 e de setup (α), dadas em [15]. Variações na taxa de chegada (λ de 5 a 30), número de contêineres pré-inicializados (n de 1 a 4), número total de contêineres (c de 5 a 20), taxa de configuração de contêineres (α de 1 a 4) e capacidade máxima do nó MEC (k de 10 e 25) foram realizadas para analisar seus efeitos na probabilidade de bloqueio do serviço (PB), número médio de contêineres inicializados (nCTNs) e tempo médio de resposta (MRT). Caso não seja especificado o contrário, os seguintes parâmetros são definidos para os modelos (analítico e simulação): $n = 2$, $c = 10$, $k = 15$, $\mu = 1$, $\alpha = 1$, $\gamma = 0,001$, $\lambda = [5, 30]$. No modelo de simulação foram executados aproximadamente 30 milhões de passos. A Tabela III sumariza os parâmetros adotados, organizando-os em cenários, onde em cada um, além da taxa de chegada, outro parâmetro é variado.

TABELA III
PARÂMETROS DE CONFIGURAÇÃO.

Parâmetro	A	B	C	D
CTNs pré-inicializados (n)	2	2	[1, 4]	2
CTNs no sistema (c)	10	[5, 20]	10	10
Capacidade do sistema (k)	15	25	15	[10, 25]
Taxa de serviço URLLC (μ)	1	1	1	1
Taxa de setup dos CTNs (α)	[1, 4]	1	1	1
Taxa de falha dos CTNs (γ)	0,001	0,001	0,001	0,001
Taxa de chegada URLLC (λ)	[5, 30]	[5, 30]	[5, 30]	[5, 30]

As Figuras 4-7 ilustram resultados obtidos através dos modelos analítico e de simulação em termos de probabilidade de bloqueio do serviço, número médio de contêineres inicializados e tempo médio de resposta, onde as linhas e os pontos denotam os resultados analíticos e de simulação, respectivamente. Cada ponto do resultado analítico é o valor médio dos resultados de 15 instâncias de simulação, com 10000 chegadas de serviço cada. O nível de confiança das simulações é de 95% (as barras dos intervalos foram omitidas nos gráficos devido a pequena diferença entre os limites). As seções a seguir apresentam os impactos de λ, α, c, n e k nas métricas PB, nCTNs e MRT.

As Figuras 4a, 5a, 6a e 7a apresentam o impacto de λ na probabilidade de bloqueio dos usuários (PB). Nelas, é possível observar que, inicialmente, a PB é próxima de 0 e cresce gradualmente à medida que λ aumenta. Este comportamento é resultado da crescente ocupação dos contêineres do sistema, acarretando em uma maior quantidade de serviços aguardando processamento. O aumento no tempo de espera faz com que o limite de serviços no sistema seja atingido mais rapidamente.

O impacto de λ no nCTNs é mostrado nas Figuras 4b, 5b, 6b e 7b. Em geral, o comportamento das curvas pode ser dividido em duas fases, subida e estabilização. Na primeira, o número de contêineres do sistema é suficiente para atender a demanda de serviços, sendo inicializados gradualmente conforme o aumento da carga de chegada de usuário, λ. Já na fase de estabilização, com a demanda muito elevada, os contêineres permanecem predominantemente inicializados, atingindo o limite de contêineres inicializados simultaneamente. Os valores escolhidos para λ procuram refletir os casos descritos no documento 3GPP (Rel 16) e podem ser atribuídos a múltiplas classes de aplicações URLLC como a indústria do transporte ou as fábricas inteligentes.

As Figuras 4c, 5c, 6c e 7c apresentam o impacto de λ no MRT. O comportamento das curvas também pode ser dividido em duas fases: subida e estabilização. No início da subida, MRT assume valores próximos de $\frac{1}{\mu}$, mas à medida que λ aumenta, o tempo de configuração de novos contêineres impacta diretamente MRT. A fase de estabilização da curva é atingida quando todos os contêineres já estão inicializados e atendendo aos serviços de forma quase ininterrupta, sem reinicialização devido ao termino de processamento.

A. Impactos da Taxa de Configuração (Cenário A)

A Figura 4a apresenta os resultados de PB sob diferentes valores de taxa de configuração (setup) de contêineres, α. Quanto maior o valor de α, mais rápido os contêineres são configurados, inicializados para processamento. Como se observa, as curvas são próximas entre si, com exceção do ponto em que a taxa de chegada (λ) se aproxima do número total de contêineres no nó MEC (a saber 10), mostrando uma PB ligeiramente maior quando os contêineres demoram mais a ficar operantes. Essa similaridade de desempenho pode indicar que o efeito da taxa de setup menor pode ser mitigado pela pré-inicialização de contêineres, que neste cenário é igual a 2. Quando a demanda é leve (em torno de 5 ou 6), a pré-inicialização consegue manter a probabilidade de bloqueio em níveis similares as configurações com menor tempo de setup do recurso, pois eles conseguem ficar disponíveis (ativos) próximos ao tempo que os usuários chegam. À medida que a demanda aumenta, em torno de λ igual a 8 ou 10, a chegada de usuários se torna mais frequente, gerando ainda a pré-inicialização. Entretanto, o usuário tende a aguardar o setup do contêiner finalizar. Quando a demanda de serviços URLLC (λ) é muito alta, as diferentes taxas de setups não impactam na PB, pois nessa situação os recursos (contêineres) ficam predominantemente ativos, processando serviços ou pré-inicializados, de modo que será pouco provável que eles reinicializem devido a término de atendimento. A reinicialização será predominantemente ocasionada por falhas nos contêineres durante o processamento dos serviços.

A Figura 4b representa os impactos de α no número de contêineres ativos (nCTNs). Observa-se que quanto maior o valor de α, maior é a inclinação das curvas durante a fase anterior a saturação do sistema, pois menor é o tempo que os contêineres levam para ficar prontos para atendimento. Quando a taxa de chegada é alta (valor de λ), a taxa de *setup* não exerce influência no número de contêineres ativos, pois todos os recursos ficam predominantemente ativos para satisfazer a demanda, reinicializando apenas em caso de falha.

A Figura 4c mostra o impacto de α no MRT. Similar a Figura 4b, a influência do valor α é observada na inclinação das curvas. Um valor de α maior diminui o tempo que os contêineres levam para ficar prontos, com isso os serviços URLLC começam a ser processados mais rapidamente e, consequentemente, experimentam um menor MRT. Esse comportamento ocorre quando a taxa de chegada varia até 15,

i.e., quando o processo de inicialização de contêiner acontece com maior frequência. Para taxas de chegadas maiores, não se observa influência dos tempos de setup no MRT, pois os contêineres ficam ativos a maior parte do tempo. Nestes casos, quando um contêiner finaliza o processamento de um serviço URLLC, ele de imediato inicia o processamento de outro, não demandando a sua reinicialização.

B. Impactos do Número de Contêineres (Cenário B)

Na Figura 5a são avaliadas as probabilidades de bloqueio quando nós MEC-NFV com diferentes quantidade de recursos (contêineres, c) são considerados. As quatro curvas aumentam de acordo com λ. Quando λ aumenta, um c maior significa que mais contêineres podem ser usados para lidar com as crescentes solicitações de serviço, diminuindo o número de serviços aguardando processamento na fila. Portanto, PB diminui de acordo com o aumento de c no sistema. Para taxas de chegada (λ) mais baixas (e.g., 5 requisições/ms) e considerando um nó MEC com 5 contêineres (c) a taxa de bloqueio de serviços URLLC é de 4.4%. O incremento de c em 100% acarreta em uma pequena redução (4.4%) na PB, resultando numa relação custo-ganho pouco atrativa para o provedor de serviço. Em contrapartida, esse mesmo incremento resulta em uma redução de 45% na PB quando se tem o dobro de demanda ($\lambda = 10$).

A redução na PB proporcionada pelo incremento de 5 contêineres diminui conforme λ aumenta. Quando a taxa de chegada atinge 30 requisições/ms, o incremento de 5 contêineres ao sistema reduz a PB em aproximadamente 16%. Os eventos de bloqueio de solicitação podem impactar significativamente nas aplicações URLLC, pois quando ocorrem, as alternativas naturais são encaminhar as solicitações bloqueadas para um nó vizinho ou para a nuvem central [16], o que traz incertezas quanto aos níveis de qualidade de serviço a ser alcançado pelas aplicações e podendo ambos incorrer em violações de requisitos dos serviços URLLC (e.g., latência). Logo, o operador deve levar em consideração uma configuração especifica de contêineres disponíveis no nó MEC de acordo com a demanda de usuário tendo em mente a relação custo-ganho.

Os resultados do nCTNs quando nós MEC com diferentes quantidades de contêineres são mostrados na Figura 5b. Nota-se que para todos os valores de c, as curvas crescem a medida que a taxa de chegadas de usuários aumenta e então atingem um limite após o valor numérico de λ ultrapassar a quantidade de contêineres do nó, c. Um c maior proporciona uma maior quantidade de contêineres ativos, que refletirá a demanda de serviços URLLC que chega ao nó MEC.

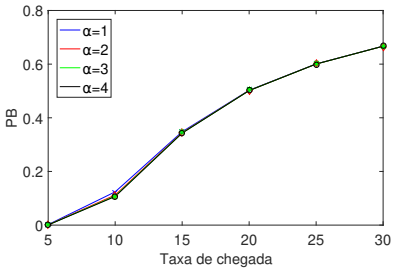
Já em termos de MRT, a Figura 5c ilustra o impacto do número de contêineres no nó MEC sob diferentes cargas de serviços URLLC. Nota-se que quanto maior o valor de c, maior a quantidade de serviços URLLC podem ser processados em paralelo, diminuindo a quantidade de requisições na fila para atendimento e implicando em um menor tempo de resposta. O tempo médio de resposta (MRT) tende a se estabilizar a medida que a taxa de chegada (λ) se aproxima da capacidade do sistema (k), pois MRT será computado apenas pelo tempo de processamento da requisição no contêiner somado do tempo máximo de espera na fila. Considerando a aplicação de automação industrial, cuja a restrição de latência é de 2 ms (3GPP Versão 16 (Rel-16)), nota-se que a configuração com 5 contêineres não atende a este requisito para nenhum valor de taxa de chegada. Dobrando a quantidade de contêineres ($c = 10$), o tempo de resposta consegue ficar abaixo do estipulado quando a demanda de serviços é baixa (λ até 10). Ao passo que adotando as outras configurações ($c = 15$ e $c = 20$) consegue-se suportar a aplicação para todos os valores de carga de serviços analisados.

Em princípio poderia se escolher a configuração com maior número de contêineres para compor o nó MEC e atender o serviço URLLC. Entretanto, essa escolha poderia incorrer em um maior custo de operação ou ociosidade de recurso principalmente em momentos de baixa carga. Desta forma, analisando de uma perspectiva custo e satisfação do requisito da aplicação, para demanda baixa, o operador poderia selecionar a configuração com $c = 10$ e para demandas maiores a configuração com $c = 15$ poderia ser empregada.

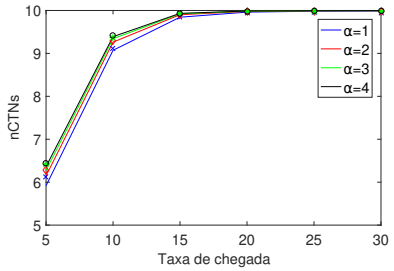
C. Impactos do Número de Contêineres Pré-Inicializados (Cenário C)

A Figura 6a apresentada os resultados da PB sob diferentes números de contêineres pré-inicializados e valores de taxa de chegada. Observa-se que n não tem impacto na PB para taxas de chegadas muito baixas (quando λ não é suficiente para formar fila) ou muito altas (quando o número de contêineres ativos se aproxima ou é igual a quantidade total do nó MEC, c). A diferença de desempenho é percebida quando a taxa de chegada de serviços URRLC fica numericamente em torno de c e k. Um maior n leva a uma PB menor, pois com mais contêineres prontos para atendimento, o processamento dos serviços pode iniciar mais rapidamente e a ocupação da fila tende a diminuir, possibilitando a admissão de novos serviços.

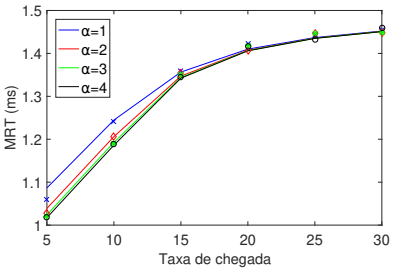
As Figuras 6b e 6c ilustram os impactos de n em nCTNs e MRT, respectivamente. Como já era esperado, na Figura 6b, n mostra maior impacto em nCTNs no ponto inicial, a medida que λ aumenta os valores de nCTNs tendem a convergir para o limite c. Um n maior acarreta em uma saturação de nCTNs mais rápida em relação a λ. Em contrapartida, na Figura 6c os contêineres inicializados previamente impactam diretamente em MRT, diminuindo o tempo de resposta para maiores valores de n no início das curvas, pois novos serviços terão de aguardar menos por um contêiner pronto para atendimento ou até mesmo serem atendidos de imediato. Um nó MEC configurado para manter 2 contêineres inicializados previamente tem uma redução de 0.1 ms (6%) no tempo de resposta para taxa de chegada (λ) igual a 5 em relação a um nó com apenas um contêiner inicializado previamente. À medida que λ aumenta essa redução em MRT tende a diminuir, pois os tempos entre as chegadas passam a ser muito menores em relação ao tempo de ligação dos contêineres ativados com a sua chegada. Por exemplo, quando λ dobra (10) o ganho alcançado com $n = 2$ é de 3,87% em relação ao $n = 1$.



(a) Probabilidade de bloqueio (PB)

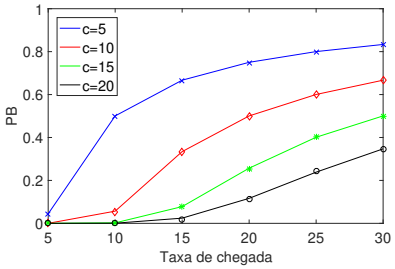


(b) Número médio de contêineres inicializados (nCTNs).

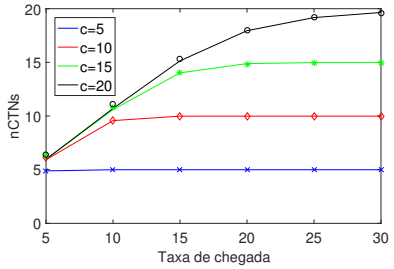


(c) Tempo médio de resposta (MRT).

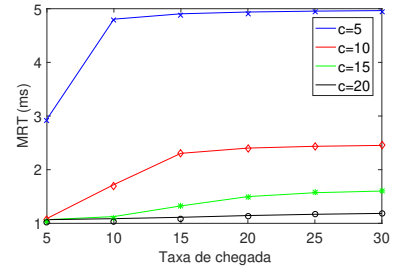
Fig. 4. Impacto da taxa de setup (α).



(a) Probabilidade de bloqueio (PB)

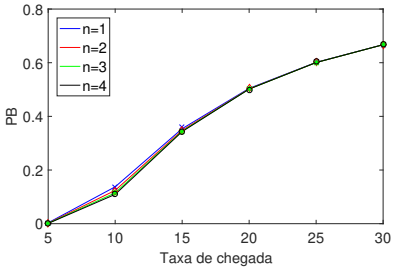


(b) Número médio de contêineres inicializados (nCTNs).

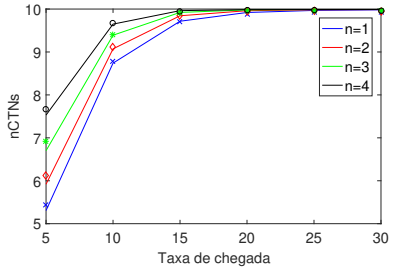


(c) Tempo médio de resposta (MRT).

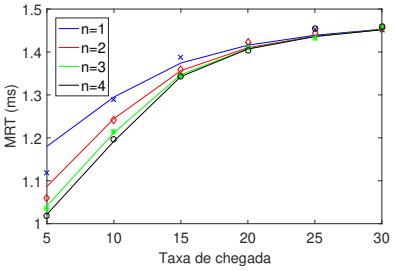
Fig. 5. Impacto do número máximo de contêineres (c).



(a) Probabilidade de bloqueio (PB)

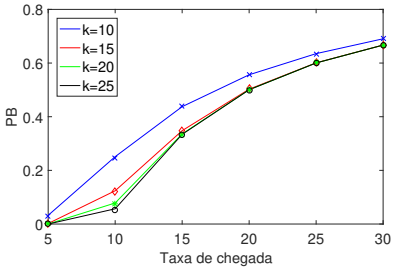


(b) Número médio de contêineres inicializados (nCTNs).

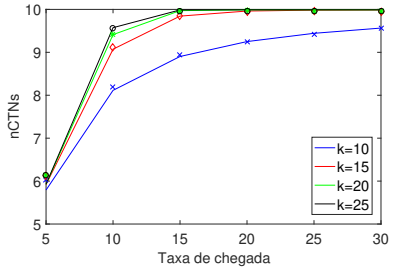


(c) Tempo médio de resposta (MRT).

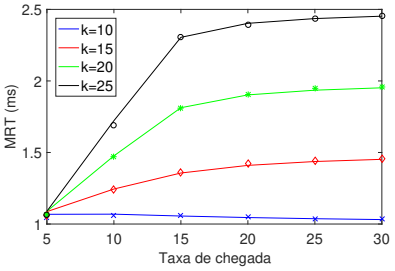
Fig. 6. Impacto do número máximo de contêineres pré-inicializados (n).



(a) Probabilidade de bloqueio (PB)



(b) Número médio de contêineres inicializados (nCTNs).



(c) Tempo médio de resposta (MRT).

Fig. 7. Impacto da capacidade do sistema (k).

D. Impactos da Capacidade do Sistema (Cenário D)

Diferentes configurações da capacidade do sistema (k) são avaliadas em relação a PB na Figura 7a. Inicialmente a PB é próxima de 0, pois os contêineres do sistema são capazes de lidar com a fila formada com uma taxa de chegada de serviços baixa. À medida que λ se aproxima do número de contêineres no sistema (c), o impacto de k é evidenciado, pois, um k maior aumenta o tamanho da fila para novos serviços diminuindo sensivelmente a PB. Quando λ ultrapassa c, nota-se que um k elevado não garante PB baixa, pois com valores de λ mais altos, a admissão no sistema é definida através da vazão dos c contêineres disponíveis, fazendo com que o limite de usuários simultâneos no sistema (k) seja atingido com frequência.

A Figura 7b exibe os impactos de k em nCTNs. A medida que λ aumenta, um k maior possibilita que mais serviços possam aguardar o processamento dos contêineres, fazendo com que nCTNs atinja o limite (c) mais rapidamente em decorrência do uso quase ininterrupto dos contêineres, que, por conta da alta demanda, raramente são parados após o

termino do processamento. No ponto $\lambda = c$, o aumento de 50% em relação ao cenário que permite 10 usuários simultâneos ($k = 10$), acarreta em um aumento de aproximadamente 1,2 contêineres inicializados no sistema (nCTNs).

O gráfico na Figura 7c mostra os impactos de k em MRT. Foi observado que apesar de um maior k diminuir a PB, MRT aumenta drasticamente com a quantidade de serviços aguardando para serem processados, ou seja, o atraso na fila, que é uma componente do tempo de resposta, acaba aumentando a sua contribuição, podendo acarretar em violações de SLA, inviabilizando o serviço URLLC.

V. Conclusão

Neste trabalho analisou-se como o aprovisionamento de recursos para serviços URLLC em redes 5G baseadas na arquitetura MEC-NFV é impactado pelo tempo de inicialização de VNF, por falhas durante o atendimento e a pré-inicialização de recursos. Avaliações de diferentes cenários foram conduzidas e métricas como o tempo médio de resposta, número médio de contêineres ativos e a probabilidade de bloqueio de serviço foram analisadas. Observou-se que o efeito da taxa de *setup* menor pode ser mitigado pela pré-inicialização de contêineres, diminuindo o tempo de espera para atendimento do serviço. Apesar do aumento da capacidade de admissão de serviços do sistema diminuir a probabilidade de bloqueio, ele ocasiona tempo de resposta dos serviços maior. O modelo proposto pode ajudar provedores de serviço no dimensionamento do nó MEC-MFV para suportar serviços URLLC, equilibrando a qualidade de serviço e o custo de operação, dada a carga de serviço e capacidade de atendimento dos recursos utilizados. Como trabalhos futuros, apontam-se a inserção de diferentes tipos de serviços, técnicas de reserva de recurso e priorização de serviço. Outro ponto de estudo é a adoção de ambientes experimentais para a validação dos modelos analíticos e análise de técnicas voltadas para a alocação de recursos.

Agradecimento

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo financiamento do projeto "Tecnologias e Mecanismos de Redes Móveis Sem Fio de Quinta Geração (5G) para Suporte a Comunicação Ultra Confiável e com Latência Muito Baixa (URLLC)", sob nº 433142/2018-9, e pela Bolsa de Produtividade nº 307053/2017-2.

References

[1] I. Parvez, A. Rahmati, I. Guvenc,A.I. Sarwat, and H. Dai, *A survey on low latency towards 5g: Ran, core network and caching solutions*. IEEE Communications Surveys Tutorials, vol. 20, pp. 3098–3130, 2018, doi: 10.1109/COMST.2018.2841349.

[2] G. Pocovi,H. Shariatmadari, G. Beradinelli, K. Pedersen, J. Steiner,and Z. Li, Z. *Achieving Ultra-Reliable Low-Latency Communications: Challenges and Envisioned System Enhancements*, IEEE Network,vol.32, n.2, p. 8-15, 2018, doi: 10.1109/MNET.2018.1700257.

[3] H. Ji, S. Park,J. Yeo, Y. Kim,J. Lee, B. Shim, *Ultra Reliable and Low Latency Communications in 5G Downlink: Physical Layer Aspects*, IEEE Wireless Communications, vol. 25, issue 3, p. 124-130, 2018,doi: 10.1109/MWC.2018.1700294.

[4] M. Giordani,M. Polese, M. Mezzavilla,S. Rangan, and M. Zorzi, *Toward 6G Networks: Use Cases and Technologies*, IEEE Communications Magazine, vol. 58, no. 3, pp. 55-61, 2020, doi: 10.1109/MCOM.001.1900411.

[5] C. She, Z. Chen, C. Yang C., T.Q.S. Quek,Y. LI, and B. Vucetic, *Improving Network Availability of Ultra-Reliable and Low-Latency Communications with Multi-Connectivity*, IEEE Transactions on Communications, 2018, doi:10.1109/TCOMM.2018.2851244.

[6] K. Kaur, T. Dhand, N. Kumar and S. Zeadally, "Container-as-a-Service at the Edge: Trade-off between Energy Efficiency and Service Availabilityat Fog Nano Data Centers,"IEEE WirelessCommunications,vol. 24, pp.48-56, 2017,doi: 10.1109/MWC.2017.1600427.

[7] Y. Ren, T. Phung-Duc, J. Chen and Z. Yu, *Dynamic Auto Scaling Algorithm (DASA) for 5G Mobile Networks,* 2016 IEEE Global Communications Conference (GLOBECOM), pp. 1-6, 2016, doi: 10.1109/GLOCOM.2016.7841759.

[8] Y. Ren, T. Phung-Duc, Y. Liu, J. Chen and Y. Lin, *ASA: Adaptive VNF Scaling Algorithm for 5G Mobile Networks,* 2018 IEEE 7th International Conference on Cloud Networking (CloudNet), 2018, doi: 10.1109/CloudNet.2018.8549542.

[9] H. Khazaei, C. Barna, M. Litoiu, *Performance modeling of microservice-platforms considering the dynamics of the underlying cloud infrastructure*, arXivpreprint arXiv:1902.03387, 2019.

[10] K. Xiong, S. Samuel Rene Adolphe, G. O. Boateng, G. Liu and G. Sun, *Dynamic Resource Provisioning and Resource Customization for Mixed Traffics in Virtualized Radio Access Network,* in IEEE Access, vol. 7, pp. 115440-115453, 2019, doi:10.1109/ACCESS.2019.2935606.

[11] T. Hößler, M. Simsek and G. P. Fettweis, *Mission Reliability for URLLC in Wireless Networks,* in IEEE Communications Letters, vol. 22, no. 11, pp. 2350-2353,2018, doi:10.1109/LCOMM.2018.2868956.

[12] Z. Tong, T. Zhang, Y. Zhu and R. Huang *Communication and Computation Resource Allocation for End-to-End Slicing in Mobile Networks,* in IEEE/CIC International Conference on Communications in China (ICCC), pp. 1286-1291, 2020, doi:10.1109/ICCC49849.2020.9238794.

[13] A. Samanta and J. Tang, *Dyme: Dynamic Microservice Scheduling in Edge Computing Enabled IoT,* in IEEE Internet of Things Journal, vol. 7, no. 7, pp. 6164-6174, 2020, doi:10.1109/JIOT.2020.2981958.

[14] 3GPP, *Release description,* 2020. [Online]. Available: https://www.3gpp.org/release-16/ [Accessed: 27-Dec-2020].

[15] NGMN, *5G Extreme Requirements: EndtoEnd Considerations,* 2018. [Online]. Available: https://www.ngmn.org/publications/5g-extreme-requirements-e2e-considerations/ [Accessed: 27-Dec-2020].

[16] I. Sarrigiannis, K. Ramantas, E. Kartsakli, P. Mekikis, A. Antonopoulos and C. Verikoukis, *Online VNF Lifecycle Management in an MEC-Enabled 5G IoT Architecture,* in IEEE Internet of Things Journal, vol. 7, no. 5, pp. 4183-4194, 2020, doi: 10.1109/JIOT.2019.2944695.

[17] A. Santoyo-González and C. Cervelló-Pastor, *Edge Nodes Infrastructure Placement Parameters for 5G Networks,* in IEEE Conference on Standards for Communications and Networking (CSCN), Paris, 2018, pp. 1-6, doi: 10.1109/CSCN.2018.8581749.

Caio Souza recebeu o título de Bacharel em Ciência da Computação pela Universidade Federal Rural de Pernambuco (UFRPE) em 2020. Atualmente, é mestrando no Centro de Informática (CIn) da Universidade Federal de Pernambuco (UFPE). Seus interesses de pesquisa incluem Wireless Network Virtualization, 5G e B5G Networks e URLLC.

Marcos Falcão é graduado em Engenharia da Computação pela Universidade Federal de Pernambuco (UFPE), Brasil (2013). Ele recebeu o título de Mestre em Ciência da Computação em 2016 pela mesma instituição e atualmente é doutorando também na UFPE. Seus interesses de pesquisa incluem virtualização de redes sem fio, redes de rádio cognitivas, redes 5G e URLLC.

Andson Balieiro possui doutorado em Ciência da Computacão pela Universidade Federal de Pernambuco (UFPE). É professor do Centro de Informática (CIn) da UFPE. Realiza pesquisas em Rádio Cognitivo, URLLC, Virtualização de Funcões de Rede, Multiple Access Edge Computing (MEC) e Redes 5G e B5G/6G.

Kelvin Dias concluiu o doutorado em Ciência da Computação pela Universidade Federal de Pernambuco (UFPE) em 2004. Foi professor adjunto da Universidade da Federal do Pará de 2005 a 2010. Atualmente, é professor associado da UFPE desde 2010. É bolsista de Produtividade do CNPQ (PQ-2). Realiza pesquisas sobre SDN, NFV, MEC, Fog Computing, Rádios Cognitivos e Redes 5G/6G.