Orquestração dinâmica total de fatiamento de rede no núcleo 5G sobre plataforma nativa de computação em nuvem

Felipe Hauschild Grings[1], Lucas Baleeiro Dominato Silveira[2], Nathan Junges Muller[1], Lúcio Prade[1], Kleber Vieira Cardoso[2], Sand Luz Correa[2] e Cristiano Bonato Both[1]

[1] Universidade do Vale do Rio dos Sinos (UNISINOS), São Leopoldo, RS – Brasil
[2]Universidade Federal de Goiás (UFG), Goiânia, GO – Brasil
{luciorp, cbboth}@unisinos.br; {felipehg, natanjm}@edu.unisinos.br
{lucassilveira,kleber,sand}@inf.ufg.br

Abstract. *Technological advances in the fifth-generation (5G) mobile networks are based on native cloud computing platforms and Kubernetes has emerged as the orchestration system for virtualized infrastructure. However, these platforms were not designed to natively support 5G services. To illustrate, Kubernetes is designed to be agnostic to the services which orchestrates and is not able to dynamically reconfigure the 5G core according to existing network resources, i.e., it provides a partial dynamic orchestration to perform network slicing. This paper proposes a solution integrated with Kubernetes to allow full dynamic orchestration of network slicing at runtime, adjusting the 5G core. This integration is accomplished through a Kubernetes-integrated controller and a proxy for control plane. The controller adjusts the 5G core and adapts the virtualized infrastructure, while the proxy creates an abstraction for the control communication between access and transport networks with the core. The experimental results showed a reconfiguration total dynamic orchestration without interruption of the services, reducing the total reconfiguration requests number by network slices by 47.5%.*

Resumo. *Os avanços tecnológicos na quinta geração (5G) de redes móveis estão alicerçados em plataformas nativas de computação em nuvem e Kubernetes tem se destacado como o sistema de orquestração de infraestrutura virtualizada. Entretanto, essas plataformas não foram projetadas para suportar nativamente os serviços em 5G. Para ilustrar, Kubernetes foi planejado para ser agnóstico aos serviços que orquestra e não é capaz de reconfigurar o núcleo 5G dinamicamente de acordo com os recursos de rede existentes, i.e., provê uma orquestração dinâmica parcial para realizar um fatiamento da rede. Este artigo propõe uma solução integrada ao Kubernetes para permitir a orquestração dinâmica total de fatiamento de rede em tempo de execução, ajustando núcleo 5G. Essa integração é realizada através de um controlador integrado ao Kubernetes e de um proxy de plano de controle. O controlador é capaz de ajustar o núcleo 5G e adaptar a infraestrutura virtualizada, enquanto o proxy cria uma abstração para a comunicação de controle entre as redes de acesso e transporte com o núcleo. Os resultados experimentais apresentaram uma reconfiguração na orquestração dinâmica total sem interrupção dos serviços, reduzindo em 47,5% o número total de requisições de reconfiguração por fatias de rede.*

1. Introdução

Redes 5G são cruciais para operadoras de telecomunicações participarem do novo mercado da indústria vertical [Ordonez-Lucena et al. 2021]. Entretanto, esse mercado não é

mais exclusivo dessas operadoras, pois gigantes como Google, Amazon e Oracle entraram definitivamente nesse mercado [Cortesão et al. 2021]. Considerando esse contexto, a área de telecomunicações está gradualmente incorporando plataformas nativas de computação em nuvem. Por exemplo, o fatiamento da rede baseado em conceitos de computação em nuvem permite, de forma mais flexível e eficiente, compartilhar a infraestrutura subjacente com diferentes serviços que são instanciados sobre redes virtuais isoladas. Essas plataformas de computação em nuvem são vistas hoje como fundamentais para prover serviços em ambientes densos e heterogêneos, como estádios, estações de metrô e situações de emergência suportando diversas demandas com requisitos específicos dos usuários finais, tais como baixa latência, grande quantidade de dispositivos, etc. [Zhou et al. 2020].

As tradicionais infraestruturas de telecomunicações, baseadas em hardware e software especializados, estão sendo migradas para infraestruturas virtualizadas, utilizando tecnologias de orquestração voltadas para computação em nuvem, tais como OpenStack, VMware vCloud Director, Amazon Web Services (AWS) e *clusters* baseados em *Container Network Function* (CNF) do Kubernetes [Guan et al. 2021]. Uma das grandes vantagens de incorporar os conceitos de computação em nuvem na área de telecomunicações é permitir gerenciar e orquestrar dinamicamente o funcionamento dos recursos virtualizados para atender demandas específicas. Por exemplo, o ciclo de vida de uma fatia de rede é formado pelos estados de criação, execução e encerramento [Abbas et al. 2021] que devem ser orquestrados eficientemente. Na criação, são definidos os requisitos do serviço que serão prestados, seguindo um modelo (*template*), por exemplo, usando padrões definidos pela GSM Alliance (GSMA) [Ordonez-Lucena et al. 2021]. No estado de execução, uma das principais ações a serem realizadas é a modificação da fatia de rede criada. Essa ação de modificação do fatiamento refere-se à capacidade de alocação e desalocação dos recursos de rede em tempo real, respeitando os acordos de nível de serviço estabelecidos. Neste contexto, a indústria e academia têm adotado amplamente o Kubernetes para orquestração de infraestrutura virtualizada em sistemas 5G.

O Kubernetes é nativamente projetado para atender as demandas de serviços com flutuações dos recursos computacionais [Arouk and Nikaein 2020]. Nesse contexto, existem duas abordagens de adaptação de infraestrutura virtualizada largamente utilizadas em computação em nuvem: (i) elasticidade vertical e (ii) elasticidade horizontal. A elasticidade vertical lida com a flutuação da demanda aumentando ou reduzindo os recursos da máquina utilizada pelo serviço. A elasticidade horizontal propõe a criação de uma ou mais instâncias virtuais com as mesmas configurações da máquina original, distribuindo a demanda para essas instâncias virtuais que executam o serviço [Arteaga et al. 2020]. Entretanto, ferramentas de orquestração de computação em nuvem, incluindo Kubernetes, não foram projetadas para suportar nativamente os serviços de telecomunicações. Além disso, para ser capaz de lidar com qualquer tipo de serviço, Kubernetes foi planejado para manter um determinado nível de abstração dos serviços, não atendendo algumas necessidades do sistema 5G. Por exemplo, a versão nativa do Kubernetes não é capaz de reconfigurar as métricas e variáveis do núcleo 5G dinamicamente de acordo com os recursos da rede, garantindo o cumprimento de acordos de nível de serviço. Essa reconfiguração dinâmica é essencial para suportar funcionalidades avançadas de fatiamento de rede, pois cada alteração deve identificar, modificar e atualizar várias funções de rede que envolvem uma fatia (de rede), incluindo componentes do núcleo 5G. Isso é essencial para usar eficientemente os recursos da infraestrutura e garantir os acordos de nível de serviço.

Os trabalhos relacionados sobre fatiamento de rede usando conceitos de computação em nuvem mostram que, em sua maioria, mecanismos de gerenciamento

de recursos de fatiamento de rede apresentam restrições para modificá-los em tempo de execução. Utilizando técnicas de elasticidade horizontal e vertical disponíveis em ferramentas de orquestração e *Virtual Infrastructure Managers* (VIMs), os trabalhos replicam exatamente as mesmas configurações de rede pré-estabelecidas adicionando auto-escala dos recursos computacionais (i.e., no estado de criação), reduzindo assim possibilidades de adaptação dos serviços de rede sob demanda e realizando uma orquestração dinâmica parcial de fatiamento de rede. As possibilidades de adaptação de serviços, juntamente com auto-escala dos recursos computacionais, são apresentadas como trabalhos futuros em diversos artigos [Chahbar et al. 2021, Chochliouros et al. 2020, Breitgand et al. 2021]. Dessa forma, observa-se a necessidade de avançar as funcionalidades das ferramentas de orquestração para modificar e adaptar as fatias de rede de forma dinâmica. Mais especificamente, existe a necessidade de propor técnicas de gerenciamento e orquestração que permitam a reconfiguração de fatias de rede dinamicamente para melhorar o desempenho do serviço prestado, com alto nível de automação da rede [Chochliouros et al. 2020]. Além disso, a literatura apresenta trabalhos com ênfase no controle de modificação de fatias de rede a partir de agentes externos ao núcleo da rede 4G [Baranda et al. 2020]. Por exemplo, observam-se estudos que usam ferramentas para alocação de novos recursos com base nas informações disponibilizadas sobre o estado atual da rede. Entretanto, tais trabalhos apresentam limitações, principalmente não considerando os padrões do 3GPP [3GPP 2018], responsáveis pelas mensagens de sinalização para o fatiamento da rede localizada no núcleo 5G, i.e., através da função *Network Slice Selection Function* (NSSF). Finalmente, existem propostas que abordam a otimização da alocação dos recursos em tempo de criação, utilizando redes neurais e algoritmos de predição. Entretanto, tais trabalhos não suportam a alteração após a criação da fatia, sendo necessário a desativação para realizar tal alteração [Zhou et al. 2020]. Em resumo, os trabalhos não abordam soluções para reconfiguração de fatias de rede após sua inicialização, i.e., as soluções não são capazes de reconfigurar o núcleo 5G dinamicamente de forma integrada com os recursos da infraestrutura virtualizada para garantir o cumprimento dos acordos de nível serviço, realizando uma orquestração dinâmica total.

Para superar as limitações apresentadas na literatura, este artigo introduz uma solução integrada à ferramenta Kubernetes, permitindo uma orquestração dinâmica total de fatiamento de rede em tempo de execução. Essa integração é realizada através de um controlador que possui interfaces para ajustar os serviços especializados do núcleo 5G (e.g., NSSF) e também adaptar o ambiente de virtualização e computação em nuvem, i.e., permitindo elasticidade vertical e horizontal dos recursos computacionais no núcleo da rede 5G. A integração também faz uso de um *proxy* de plano de controle que é responsável pela abstração da comunicação (de controle) entre rede de acesso e transporte com o núcleo 5G. A abstração torna transparente a comunicação com diferentes instâncias de componentes do núcleo, além de oferecer balanceamento de carga e monitoramento de disponibilidade das instâncias. A solução proposta permite reconfigurar o fatiamento de rede em tempo de execução para a rede 5G suportar diversas demandas dos usuários finais. Dessa forma, as principais contribuições do presente artigo são: (i) adaptação dinâmica total de fatias de rede 5G sob demanda dos serviços, (ii) um controlador acoplado ao Kubernetes para permitir a integração entre as funções do núcleo 5G e o ambiente de computação em nuvem, (iii) um *proxy* para comunicação de controle com o núcleo 5G e (iv) uma avaliação experimental destacando a reconfiguração baseada na orquestração dinâmica total sem interrupção dos serviços prestados, reduzindo em 47,5% o número total de requisições de reconfiguração por fatias de rede.

O restante do artigo está organizado conforme descrito a seguir. A Seção 2 apresenta uma visão geral sobre trabalhos que exploram propostas similares à aqui apresentada. A Seção 3 descreve a orquestação dinâmica total de fatiamento de rede em tempo de execução proposta. A Seção 4 apresenta o protótipo desenvolvido e na Seção 5 discute a avaliação de desempenho da orquestração dinâmica total de fatiamento de rede comparada com a orquestração dinâmica parcial. Por fim, a Seção 6 conclui o artigo com considerações finais e direções para trabalhos futuros.

2. Trabalhos Relacionados

Esta seção apresenta os trabalhos relacionados sobre soluções de fatiamento de rede em 5G que utilizam infraestruturas virtualizadas baseadas em conceitos de computação em nuvem. Todos os trabalhos descritos a seguir possuem algum nível de orquestração de fatiamento de rede. Por exemplo, no cenário mais básico, a configuração estática replica exatamente os mesmos parâmetros da rede pré-estabelecidos no estado inicial do ciclo de vida de uma fatia de rede. Entretanto, essa revisão da literatura foi direcionada para trabalhos que apresentam as funções do núcleo 5G com suporte a arquitetura baseada em serviços (do inglês, *Architecture-based Service* - SBA), os quais possuem nativamente a função NSSF, bem como componentes para o gerenciamento de fatias de rede, como *Network Slice Subnet Management Function* (NSSMF) e *Network Slice Management Function* (NSMF). Além disso, discute-se soluções que apresentam projetos de criação de modelos (*templates*) de fatias de redes que permitem a modificação no estado de execução dessas fatias. Por fim, a principal análise dos trabalhos da literatura refere-se a orquestração dinâmica das fatias de rede. Para tornar mais claro e organizado o grupo de trabalhos encontrados na literatura considerando esse tópico, introduziu-se o conceito de orquestração dinâmica de forma *Parcial* e *Total*. A orquestração *Parcial* refere-se às funcionalidades disponíveis nas ferramentas de computação em nuvem para auto-escala de recursos computacionais para compartilhar a infraestrutura virtualizada com diferentes serviços ou configurações especializadas dentro do núcleo 5G para suportar fatiamento de rede. A orquestração *Total* diz respeito às soluções que permitem reconfigurar o núcleo 5G dinamicamente de forma integrada com os recursos da infraestrutura virtualizada.

As principais iniciativas de código aberto em computação em nuvem utilizadas pelas operadoras de telecomunições são: *Open Network Automation Platform* (ONAP) [Foundation 2022] e Open Source *Management and Orchestration* (MANO) (OSM) [ETSI 2022]. ONAP é uma plataforma abrangente da *The Linux Foundation*, composta por módulos para orquestração, gerenciamento e automação de serviços em tempo real. Essa plataforma é orientada a políticas de funções de rede físicas e virtuais, permitindo uma rápida automação de novos serviços e o gerenciamento do ciclo de vida dessas funções para redes 5G. OSM é uma iniciativa da *European Telecommunications Standards Institute* (ETSI) que visa alinhar as atividades de desenvolvimento com a evolução do padrão ETSI *Network Function Virtualization* (NFV), permitindo que operadoras e fornecedores tenham um ecossistema baseado na arquitetura NFV MANO. As duas iniciativas possuem diversas funcionalidades considerando o escopo de fatiamento de rede, tais como suporte a núcleo 5G, possibilidade de utilização de modelos para a criação de fatias de redes e suporte a auto-escala de recursos computacionais utilizando funcionalidades de computação em nuvem baseadas no Kuberneters, por exemplo. Entretanto, apesar do ONAP e OSM implementarem *Virtual Network Functions* (VNFs) isoladas, essas plataformas não apresentam um controle adaptativo sobre essas funções, ou seja, não permitem a reconfiguração das funções do núcleo 5G dinamicamente e integradas com os

recursos que gerenciam a infraestrutura virtualizada. Por exemplo, para aplicar uma nova configuração em uma fatia de rede, é necessário encerar o ciclo de vido dessa fatia e iniciar uma nova fatia de rede, interrompendo o serviço prestado. Baseada nessa característica, a orquestração de fatias de rede dessas iniciativas é classificada como dinâmica *Parcial*, como pode ser observado na Tabela 1. O projeto Europeu 5G-TOURS [Garcia-Aviles et al. 2020] possui as mesmas características das iniciativas ONAP e OSM.

Tabela 1. Resumo dos Trabalhos relacionados.

Trabalhos	Características		
	Núcleo 5G SBA	Projeto Modelo	Orquestração* Dinâmica
ONAP [Foundation 2022]	●	●	◒
OSM [ETSI 2022]	●	●	◒
5G-TOURS [Garcia-Aviles et al. 2020]	●	●	◒
5GZORRO [Breitgand et al. 2021]	○	●	◒
5Growth [Baranda et al. 2020]	○	●	◒
5G-COMPLETE [Gkatzios et al. 2020]	○	●	◒
5G-CLARITY [Ordonez-Lucena et al. 2021]	○	○	◒
DYSOLVE [Kukkalli et al. 2020]	○	○	◒
OpenSlice [Tranoris 2021]	○	●	●
Este trabalho	●	●	●

*Orquestração dinâmica: Ausente (○) *Parcial* (◒) *Total* (●)

Trabalhos como 5GZORRO [Breitgand et al. 2021] e 5Growth [Baranda et al. 2020] apresentam modelos de orquestração dinâmica *Parcial*, com elasticidade vertical e horizontal para o fatiamento de rede, utilizando a arquitetura MANO. 5GZORRO e 5Growth exploram o controle do ciclo de vida do fatiamento de rede considerando os três domínios, i.e., rede de acesso, transporte e núcleo, seguindo os padrões propostos pelas entidades 3GPP e ETSI. Entretanto, esses trabalhos abordam o núcleo 5G de forma única e imutável, limitando o gerenciamento dos recursos disponibilizados pela arquitetura SBA proposta pelo 3GPP. Neste contexto, a ferramenta 5G-COMPLETE [Gkatzios et al. 2020] também considera o núcleo 5G de forma única. Tais trabalhos não conseguem cumprir determinados acordos de nível de serviços, afetando a qualidade do serviço prestado.

A literatura sobre fatiamento de rede também apresenta trabalhos com uma perspectiva ampla. Por exemplo, o projeto Europeu 5G-CLARITY [Ordonez-Lucena et al. 2021] visa desenvolver um novo plano de gerenciamento baseado nos princípios de *Software-Defined Networking* (SDN), NFV e no uso de algoritmos de inteligência artificial para habilitar o fatiamento de rede em equipamentos neutros. Adicionalmente, DYSOLVE [Kukkalli et al. 2020] é uma proposta de alocação dinâmica de recursos para fatiamento dinâmico de rede 5G para um cenário de emergência veicular. Essa solução tem como objetivo alocar recursos de rádio e da rede de transporte de operadoras de forma cooperativa para otimizar o custo da fatia de rede, garantindo a disponibilidade do serviço. Nesses trabalhos, a orquestração da infraestrutura virtualizada está presente, i.e., auto-escala dos recursos computacionais caracterizando uma orquestração dinâmica *Parcial* do fatiamento de rede. Entretanto, devido à perspectiva ampliada, esses trabalhos não consideram as funcionalidades de fatiamento de rede do núcleo 5G, nem os modelos de fatias definidos pela GSMA para auxiliar no gerenciamento dessas fatias de redes.

Uma iniciativa que se destaca na orquestração dinâmica de fatias de rede é o projeto OpenSlice [Tranoris 2021]. Essa solução de código aberto é baseada no OSM e permite que usuário de diferentes fatias de rede explorem especificações de serviço oferecidas para infraestrutura de computação em nuvem. Além disso, OpenSlice permite que desenvolvedores de NFV integrem e gerenciem artefatos de VNF e serviços de rede. Desta forma, o projeto OpenSlice é classificado com uma orquestração dinâmica *Total* para fatiamento de redes, como observado na Tabela 1. Entretanto, OpenSlice ainda não suporta características de funções especializadas do núcleo 5G, restringindo sua aplicabilidade. Analisando a literatura de fatiamento de rede, percebe-se que não existem soluções que suportem a reconfiguração do núcleo 5G dinamicamente após a inicialização de uma fatia de rede para garantir o cumprimento dos acordos de nível serviço em uma rede 5G. A solução proposta neste artigo suporta essas características, detalhada na próxima seção.

3. Orquestração Dinâmica Total de Fatiamento de Rede

Esta seção apresenta a arquitetura integrada para a orquestração dinâmica total de fatiamento de redes em tempo de execução. Além disso, um diagrama de sequência detalhando a comunicação entre os principais componentes da arquitetura é discutida.

3.1. Arquitetura Integrada para a Orquestração Dinâmica Total

A visão geral da arquitetura integrada para o gerenciamento dinâmico total de fatias de redes é estruturado em duas camadas, como pode ser observada na Figura 1. Na camada inferior da arquitetura encontra-se a infraestrutura composta pelas redes de acesso, de transporte e de núcleo. Nessa camada pode existir diversas fatias de rede virtuais sobre uma infraestrutura física, composta por equipamentos de rádio, de encaminhamento, servidores, entre outros recursos computacionais. A comunicação entre essas redes e seus equipamentos é realizada através de planos de controle e de dados. Já, na camada superior é definido o Sistemas de Apoio à Operação (do inglês, *Operation Support Systems* - OSS) que gerencia todos os recursos usados para o provisionamento e operação dos fatiamentos de rede em tempo de execução. A camada OSS segue o padrão ETSI NFV, baseada na arquitetura NFV MANO e possui componentes definidos pelo padrão 3GPP.

A Figura 1 apresenta um exemplo com três diferentes fatias de rede, com a mesma finalidade de entregar serviços *Ultra-Reliable Low-Latency Communication* (URLLC), i.e., prover redundância e controle isolado para cada fatia específica. É importante destacar que cada fatia de rede apresenta características específicas como: distribuição geográfica, recursos computacionais, tempos de respostas, limite máximo de usuários conectados simultaneamente, etc. Essas características são consideradas na arquitetura integrada para orquestração dinâmica total, oferecendo um provisionamento das fatias, bem como uma adaptação desses serviços durante o ciclo de vida de cada fatia. Além disso, as funções do núcleo 5G possuem mecanismos para atuar sobre as redes de acesso e transporte para o provisionamento e adaptação das configurações em tempo de execução, através do plano de controle. As mensagens de sinalização necessárias para a orquestração das diferentes fatias de rede são implementadas pela função NSSF no núcleo 5G. Essa função trabalha integrada com outras funções do núcleo, tais como *Access and Mobility Management Function* (AMF), *Session Management Function* (SMF), *User Plane Function* (UPF), entre outras. O objetivo das funções do núcleo é prover a sinalização para o controle dos fatiamentos de rede, i.e., o núcleo 5G não tem responsabilidade em orquestrar o ciclo de vida das fatias em tempo de execução, com a camada de infraestrutura.

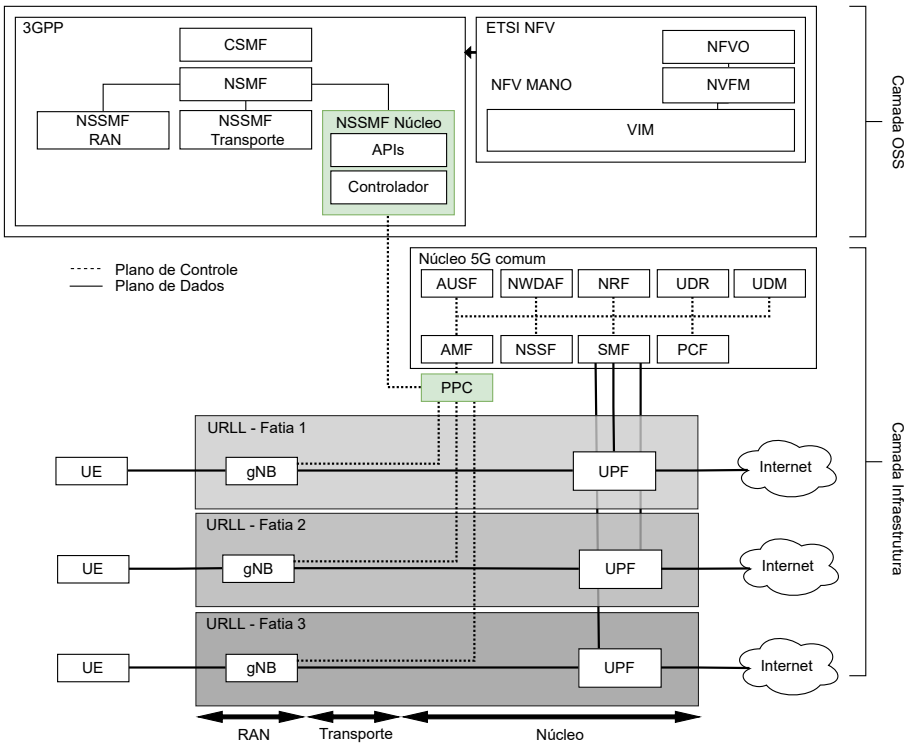

Figura 1. Arquitetura de controle em tempo de execução de fatias de rede.

A camada OSS é responsável por toda operação, administração e manutenção (do inglês, *Operations, Administration, and Maintenance* - OAM) das funções modulares encontradas na camada de infraestrutura, permitindo que os operadores gerenciem as diferentes fatias de rede ao longo de todo seu ciclo de vida. As definições dos padrões 3GPP e ETSI NFV tem grande impacto na camada OSS. Por exemplo, as funções NSMF e NSSMF definidas pelo 3GPP devem estar fortemente integradas com toda a arquitetura de orquestração dinâmica total, baseada no NFV MANO para o fatiamento de rede. Dessa forma, cada NSSMF possui instâncias para a rede de acesso (NSSMF RAN), de transporte (NSSMF Transporte) e de núcleo (NSSMF Núcleo), coordenada por uma função NSMF. Neste trabalho, um controlador é proposto na instância NSSMF Núcleo para suportar a orquestração dinâmica total integrada com a camada de infraestrutura, utilizando contêineres Kubernetes. Assim, a integração é realizada por meio de APIs HTTP REST para o gerenciando o ciclo de vida das *Network Slice Instances* (NSIs). Além disso, a NSSMF Núcleo foi projetada com APIs para suporte a criação, remoção e modificação de fatiamento de rede em tempo de execução. A modificação de uma fatia de rede é realizada através do controlador, responsável pela comunicação para disponibilizar as NSIs, cumprindo os acordos de nível de serviço. Por meio de comunicações HTTP REST, o controlador envia as requisições para o Kubernetes criar, alterar e remover as NFs.

Para que essa comunicação ocorra entre as camadas OSS e de infraestrutura, bem como com núcleo 5G, foi proposto um *Proxy* de Plano de Controle (PPC), baseado em conceitos de um *proxy* reverso. O PPC é responsável pela abstração da comunicação entre a RAN e rede de transporte com o núcleo, através no plano de controle. A abstração é realizada apresentando apenas o IP de conexão do PPC e escondendo todas as possibilidades de conexões com AMFs. O PPC realiza a conexão entre uma estação de rádio base, chamada gNB, e AMF utilizando *Stream Control Transmission Protocol* (SCTP) e

o algoritmo Round Robin para distribuição dos pedidos de registro dos *User Equipments* (UEs). A lista de AMFs armazenada no PPC contém o estado de cada AMF, com métricas de atualização dos recursos computacionais. As AMFs com configurações de fatias e conexões atualizadas são consideradas ”saudáveis”, enquanto AMFs desatualizadas devem ter suas conexões finalizadas o mais breve possível. As novas conexões são encaminhadas para quaisquer AMFs que estejam com o estado ”saudável”, enquanto AMFs desatualizadas só podem receber mensagens de comunicações em andamento antes da sua troca de estado. Assim, encerrando quaisquer comunicação com AMFs desatualizadas e garantindo a atualização de configuração da rede, sem interrupção no serviços e no processamento de requisições. O controlador, através do PPC, garante uma adaptação contínua e transparente para as conexões de UEs, realizando alterações em tempo de execução sem interrupção de serviço.

3.2. Diagrama de Sequência do Funcionamento da Orquestração Dinâmica Total

Para um melhor detalhamento da comunicação entre os componentes propostos na arquitetura integrada para a orquestração dinâmica total, essa subseção apresenta um diagrama de sequência do funcionamento dessa orquestração de fatiamento de rede no núcleo 5G sobre a plataforma nativa de computação em nuvem, Kubernets. Inicialmente, como pode ser observado na Figura 2, um UE envia uma mensagem de solicitação de registro no sistema 5G (1), utilizando o protocolo *Next Generation Application Protocol* (NGAP) para uma AMF registrada. Essa mensagem é recebida pelo PPC que (2) redireciona para uma AMF ”saudável”, através de uma conexão SCTP. Esta mensagem ao chegar na AMF escolhida (#1), processa o pedido de registro, com as informações do UE, por exemplo, endereço IP, *Service Slice Type* (SST) e *Slice Differentiator* (SD). Após, a AMF (#1) interage com várias outras Funções de Rede do núcleo 5G (3) (do inglês, *Network Functions* - NFs), até o estabelecimento da sessão de dados realizada pelo gNB e UPF (4).

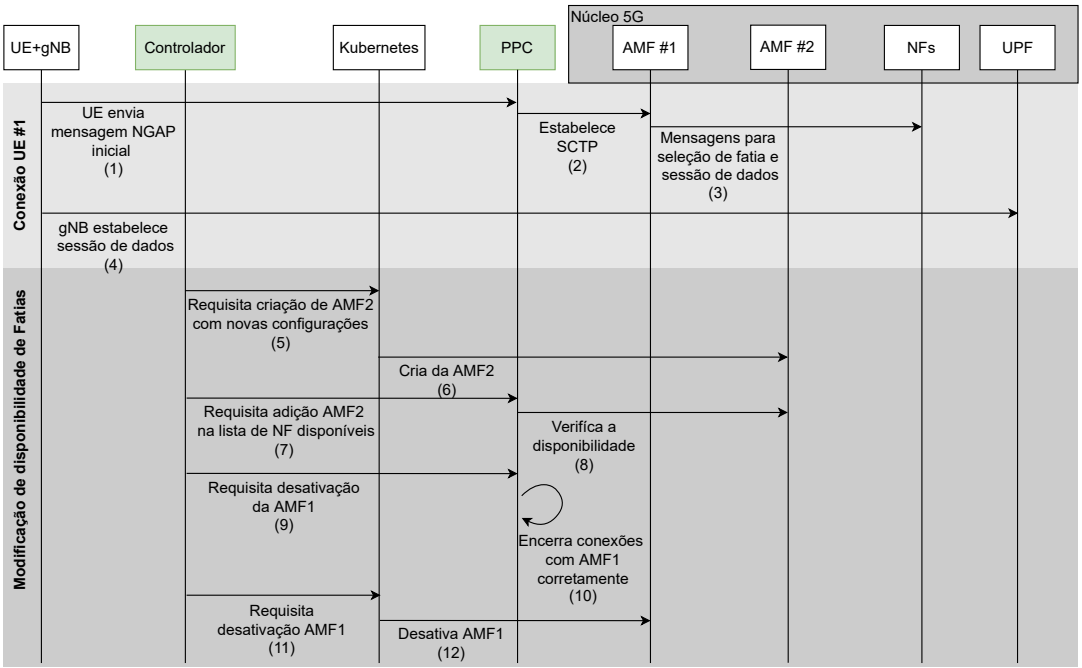

Figura 2. Fluxo de trabalho para modificação de AMF pelo controlador.

A Figura 2 apresenta a sequência de integrações para o controle das NSIs em tempo de execução, após a conexão do UE. O controlador requisita a configuração de

uma nova AMF (#2) utilizando as novas configurações recebidas pelo o Kubernetes (5), através da API disponibilizada na NSSMF Núcleo. Após, o Kubernetes se encarrega da disponibilização e validação da NF construída (6). O controlador realiza a requisição para cadastro de uma nova AMF (#2) disponível para requisições ao PPC (7), com o objetivo de adicioná-la na lista de NF disponíveis. O PPC verifica a disponibilidade das NFs (8) e o controlador requisita a desativação da AMF antiga (#1) com configurações desatualizadas (9). O PPC encerra de forma correta todas as conexões restantes (10), impedindo novas conexões com a AMF (#1) em desligamento, finalizando as requisições em andamento. Finalmente, o controlador requisita a desativação da AMF (#1) para o Kubernetes (11), que desativa a AMF (#1) liberando os recursos computacionais utilizados (12).

4. Protótipo

Esta seção apresenta o protótipo da arquitetura integrada entre o controlador, PPC, núcleo 5G e Kubernetes, provendo uma orquestração dinâmica total de fatiamento de rede. A ferramenta Terraform[1] foi utilizada para a padronização do ambiente, provisionamento e replicação das máquinas virtuais (do inglês, *Virtual Machine* - VM). Já, o gerenciamento de pacotes e atualização dos módulos de softwares utilizados foram realizados pela ferramenta Ansible[2], responsável, por exemplo, pelas configurações de rede, Kubernetes, *Global System for Mobile Communications Tunnelling Protocol* (GTP) para 5G (GTP5G), etc. O sistema operacional utilizado foi o Ubuntu 18.04.1 LTS com kernel 5.0.0-23-genericm com suporte ao módulo *Low-Latency*. Além disso, o Kubernetes foi configurado com um *Master* e dois *Nodes*, como pode ser observado na Figura 3. O protótipo suporta *Container Network Interface* (CNI) extra para gerenciar redes e subredes necessárias para a comunicação interna nos protocolos NGAP e *Non-Access Stratum* (NAS) com o núcleo 5G, utilizando o projeto free5GC[3]. Essa integração foi realizada utilizando Multus CNI[4]. Por fim, um emulador de UEs e RAN para a geração de pedidos de conexão com o núcleo 5G e tráfego de rede foi integrado no protótipo. O testador 5G chamado my5G-RAN-tester[5] [Dominato et al. 2021] foi escolhido para essa emulação.

É importante destacar que a NSSMF Núcleo e PPC são baseados em conteineres e, desta forma, implementadas como componentes internos ao Kubernetes. Além disso, a NSSMF Núcleo e PPC foram desenvolvidos utilizando a linguagem de programação Typescript e Go. Dessa forma, o prototipo segue a arquitetura *Service Oriented Architecture* (SOA), utilizando protocolo HTTP REST para o interface com componentes externos. Finalmente, o PPC utiliza o protocolo SCTP e o algoritmo Round Robin para distribuição de novas conexões no plano de controle com o núcleo 5G. O protótipo desse artigo está disponível no GitHub em https://github.com/fhgrings/ODT-5gc.

5. Avaliação de Desempenho

Essa seção apresenta inicialmente a metodologia de avaliação de desempenho realizada para avaliar o protótipo desenvolvido. Posteriormente, os resultados são discutidos comparando a orquestração dinâmica total e parcial, encontradas na literatura.

[1] https://www.terraform.io/ [2] https://www.ansible.com/ [3] https://www.free5gc.org/ [4] https://01.org/kubernetes/projects/multus-cni [5] https://github.com/my5G/my5G-RANTester

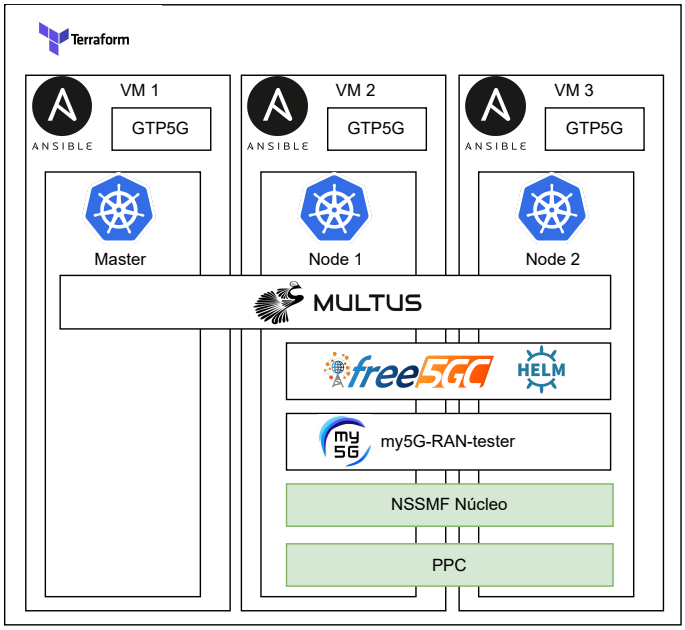

Figura 3. Protótipo.

5.1. Metodologia

O protótipo foi avaliado em um servidor com processador Intel(R) Xeon(R) CPU E5-2407@2.20GHz, utilizando três VMs, com as seguintes configurações: VM1 (*Master*) com 4 vCPUs, 8 GB RAM, 100 GB HD; VM2 e VM3 (*Nodes*) com 2vCPUS, 4 GB RAM, 100 GB HD. O experimento considera o ambiente distribuído entre os dois Kubernetes Nodes, com interfaces de redes Flannel e Multus. A distribuição dos IPs entre as sub-redes foi realizada de forma estática, sendo necessário a seleção de um novo IP para cada implementação de componente na rede.

Para a geração de requisições de registros de UEs no núcleo 5G utilizou-se o my5G-RAN-tester, selecionando fatias de rede específicas e estabelecendo conexões entre gNB e núcleo 5G. Entre as rajadas de solicitações de registros foram realizadas alterações nas configurações das fatias de rede suportadas pelo núcleo 5G, através da orquestração dinâmica parcial e total. As mensagens foram registradas em formatos de logs, sendo posteriormente analisadas nos resultados apresentados, i.e., os resultados representam uma média de seis execuções sem apresentar variação significativa entre as execuções.

5.2. Resultados

A primeira análise refere-se a disponibilidade (0 ou 1) de uma fatia de rede durante o seu processo de reconfiguração em tempo de execução. Esse processo diz respeito a solicitação, processamento e resposta de requisições de reconfiguração das funções de redes providas pelo núcleo 5G, bem como as alterações na infraestrutura virtualizada. A Figura 4 mostra o início e o fim desse processo de reconfiguração ao longo de uma janela de tempo. Além disso, na parte superior da Figura 4, pode-se observar que a orquestração dinâmica total não presenta nenhuma interrupção na prestação de serviço durante o processo de reconfiguração e o tempo de reconfiguração é de 9 segundos. Esse comportamento ocorre graças a camada de abstração criada pelo PPC sobre a comunicação gNB e núcleo 5G. O PPC trabalha utilizando um único IP informado para todas gNBs, bem como possui todas as reconfigurações realizada em tempo de execução, i.e., encerrando e inicializando novas conexões SCTP com as AMFs. Dessa forma, o tempo de reconfiguração

é resultado da alocação dos AMFs e da reconfiguração do núcleo, sendo diretamente relacionado a fila de requisições processada no momento de alteração das AMFs.

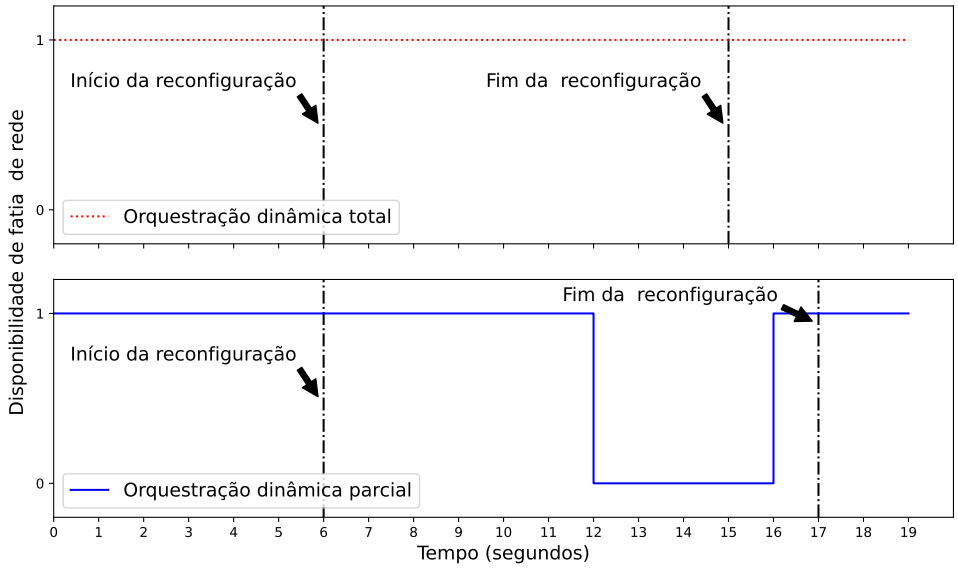

Figura 4. Disponibilidade das fatias de rede.

Na orquestração dinâmica parcial, representada na parte inferior da Figura 4, observa-se uma indisponibilidade entre os segundos 12 e 16, além do tempo de reconfiguração ocorrer em 11 segundos. As requisições em processamento são encerradas corretamente, porém é observado essa indisponibilidade durante o período de destruição, configuração da nova fatia de rede e instanciação das funções de rede do núcleo 5G. Esse comportamento ocorre, pois a destruição da AMF, para atualização de suas configurações, não é possível, devido a uma nova instância requisitar o mesmo IP da sub-rede de protocolo NGAP, gerenciada dentro do Kubernetes. Dessa forma, é importante destacar que o controlador proposto na NSSMF Núcleo e o PPC tem impacto direto sobre a qualidade de serviço durante etapas de reconfiguração, pois garante o serviço disponível no processo de reconfiguração do fatiamento de rede.

A quantidade de mensagens de requisições para realizar reconfigurações de fatias de rede também foi analisada. A Tabela 2 apresenta o número de mensagens de requisições, considerando três etapas do processo de reconfiguração de fatias de rede, i.e., criação, alteração e remoção. A criação de uma nova fatia de rede é dada pela instanciação das funções de rede do núcleo 5G e PPC. Neste trabalho, foram consideradas as funções básicas do núcleo, i.e., AMF, UPF, SMF. O PPC só é utilizado na orquestração dinâmica total. Os resultados representam a média de requisições nas interfaces CNI do Kubernetes. Neste caso, observa-se um aumento de 61 requisições da orquestração dinâmica parcial para 99 na orquestração dinâmica total, i.e., um aumento de 61,61% nas requisições totais. Esse comportamento é resultado das requisições de configuração do PPC.

Tabela 2. Número de requisições.

	Criação	Alteração	Remoção
Orquestração dinâmica parcial	61	101	40
Orquestração dinâmica total	99	53	47

Uma segunda observação refere-se a média de requisições para a alteração das

configurações de disponibilidade de uma fatia de rede. Neste caso, a orquestração dinâmica parcial precisou de 101 requisições, enquanto a orquestração dinâmica total utilizou 53 mensagens de requisição, i.e., uma redução de 47,5% nas requisições deste tipo. Esse comportamento é consequência da possibilidade de alteração de uma fatia de rede em tempo de execução utilizando funções de redes independentes. Como a orquestração dinâmica parcial não implementa totalmente tal funcionalidade, percebe-se um número maior de requisições, pois é necessária a desativação da fatia de rede para posteriormente realizar a reconfiguração de uma nova fatia de rede. Finalmente, investigou-se a média de requisições realizadas para remoção de uma fatia de rede. A orquestração dinâmica parcial apresentou 40 mensagens de requisição, enquanto, a orquestração dinâmica total necessitou de 47 mensagens. É importante destacar que o número de mensagens extras é insignificante e permite o serviço permanecer disponível, sem interrupções.

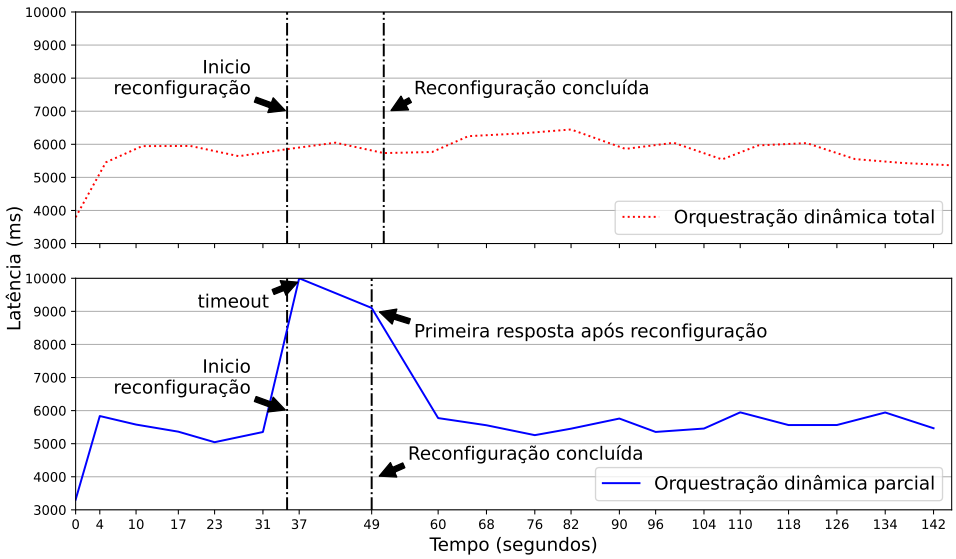

Figura 5. Latência antes, durante e após o processo de reconfiguração.

Uma análise importante refere-se a latência apresentada para o registro de UEs, antes, durante e após o processo de reconfiguração de uma fatia de rede. A parte superior da Figura 5 mostra a latência do processo de reconfiguração, considerando a orquestração dinâmica total. Neste caso, pode-se observar que a latência não sofreu variações significativas antes, durante e após a reconfiguração do fatiamento de rede, caracterizando qualquer interrupção de serviço. A latência observada ficou estável em torno de 5500ms. Essa latência elevada é resultado do ambiente de baixo recurso computacional disponível e da comunicação entre múltiplas redes e sub-redes utilizadas na cenário de avaliação.

A parte inferior da Figura 5 apresenta a latência para a orquestração dinâmica parcial. Após a inicialização do processo de reconfiguração, no segundo 35, pode-se percebe que existe um aumento significativo na latência. Além disso, a partir do segundo 37, pode se observar um período, em média de 12 segundos, sem respostas das requisições de registros, i.e., as mensagens forma perdidas, devido ao *timeout* e o serviço ficou indisponível. Após a reconfiguração pode-se perceber que a latência estabilizou em torno de 5500ms.

A última análise refere-se a adaptabilidade das fatias de redes utilizadas na orquestração dinâmica total. A parte superior da Figura 6 apresenta uma fatia de rede configurada para aceitar até sete UEs simultaneamente, com o objetivo de prover alta qualidade de serviço para essas UEs conectadas. Quando a quantidade de UEs é alcançada

na fatia de rede 1, o controlador da NSSMF Núcleo aciona a reconfiguração da rede, interrompendo a disponibilização da fatia de rede 1 para novas requisições de registros de UE e inicia o processo de disponibilidade na fatia de rede 2. Através do controlador proposto nesse trabalho é possível garantir o cumprimento dos acordos de nível de serviços estabelecidos em tempo de execução, sem interferir em outros serviço disponíveis.

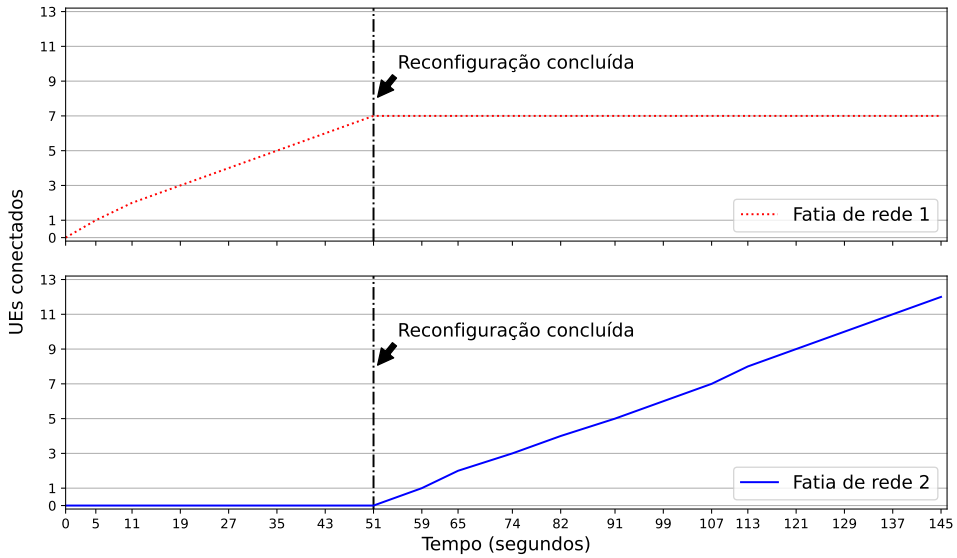

Figura 6. Adaptabilidade das fatias de rede.

6. Considerações Finais

Este artigo apresentou uma solução integrada à ferramenta Kubernetes, utilizando uma orquestração dinâmica total de fatiamento de rede em tempo de execução. Essa integração foi realizada através da implementação de um controlador que possui interfaces para ajustar os serviços especializados do núcleo 5G e adaptar o ambiente de virtualização em computação em nuvem, permitindo elasticidade vertical e horizontal dos recursos computacionais no núcleo da rede 5G. Avaliações experimentais foram realizadas comparando a orquestração dinâmica parcial com a orquestração dinâmica total. Os resultados demonstram a eficiência da orquestração dinâmica total, através da reconfiguração de fatiamento de redes. Destaca-se nessa comparação, que a orquestração dinâmica total realiza a reconfiguração do fatiamento de rede sem interrupção dos serviços e reduz a quantidade de mensagens de requisições de registros, na solicitação de alteração de requisições. Como trabalhos futuros, pretende-se integrar aprendizagem de máquina na adaptabilidade das fatias de redes de forma proativa, aumentando o nível de automatização da solução.

Agradecimentos

Os autores agradecem à Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) por meio dos Projetos PORVIR-5G: Programabilidade, ORquestração e VIRtualização de Redes em 5G, concessão 2020/05182-3, e SAMURAI: núcleo 5G inteligente e integração de múltiplas redes de acesso, concessão 20/05127-2.

Referências

3GPP (2018). System Architecture for the 5G (Release 15). Technical Recommendation (TR) 23.501, 3^{rd} Generation Partnership Project (3GPP).

Abbas, K. et al. (2021). Network Slice Lifecycle Management for 5G Mobile Networks: An Intent-Based Networking Approach. *IEEE Access*, 9:80128–80146.

Arouk, O. and Nikaein, N. (2020). 5G Cloud-Native: Network Management amp; Automation. In *IEEE/IFIP Network Operations and Management Symposium*, pages 1–2.

Arteaga, C. H. T. et al. (2020). A scaling mechanism for an evolved packet core based on network functions virtualization. *IEEE Transactions on Network and Service Management*, 17(2):779–792.

Baranda, J. et al. (2020). Scaling Composite NFV-Network Services. In *Proceedings of the Twenty-First International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing*, page 307–308.

Breitgand, D. et al. (2021). Dynamic Slice Scaling Mechanisms for 5G Multi-domain Environments. In *IEEE International Conference on Network Softwarization*, pages 56–62.

Chahbar, M. et al. (2021). A Comprehensive Survey on the E2E 5G Network Slicing Model. *IEEE Transactions on Network and Service Management*, 18:49–62.

Chochliouros, I. P. et al. (2020). Dynamic network slicing: Challenges and opportunities. *IFIP Advances in Information and Communication Technology*, 585 IFIP:47–60.

Cortesão, R. et al. (2021). Cloud-Based Implementation of a SON Radio Resources Planning System for Mobile Networks and Integration in SaaS Metric. *IEEE Access*, 9:86331–86345.

Dominato, L. B. et al. (2021). Tutorial on communication between access networks and the 5G core. *arXiv, disponível em https://arxiv.org/pdf/2112.04257.pdf*.

ETSI (2022). Open Source Management and Orchestration. https://osm.etsi.org/.

Foundation, T. L. (2022). Open Network Automation Platform. https://www.onap.org/.

Garcia-Aviles, G. et al. (2020). Experimenting with open source tools to deploy a multi-service and multi-slice mobile network. *Computer Communications*, 150:1–12.

Gkatzios, N. et al. (2020). Optimized placement of virtualized resources for 5G services exploiting live migration. *Photonic Network Communications*, 40:233–244.

Guan, W., Zhang, H., and Leung, V. C. M. (2021). Customized Slicing for 6G: Enforcing Artificial Intelligence on Resource Management. *IEEE Network*, 35(5):264–271.

Kukkalli, H. et al. (2020). Evaluation of Multi-operator dynamic 5G Network Slicing for Vehicular Emergency Scenarios. In *IFIP Networking Conference*, pages 761–766.

Ordonez-Lucena et al. (2021). On the Rollout of Network Slicing in Carrier Networks: A Technology Radar. *Sensors*, 21(23).

Tranoris, C. (2021). Openslice: An opensource OSS for Delivering Network Slice as a Service. *arXiv, disponível em https://arxiv.org/pdf/2102.03290.pdf*.

Zhou, J., Zhao, W., and Chen, S. (2020). Dynamic Network Slice Scaling Assisted by Prediction in 5G Network. *IEEE Access*, 8:133700–133712.