Capítulo 1

Gerenciamento e Orquestração de Serviços em O-RAN: Inteligência, Tendências e Desafios

Rodrigo de Souza Couto (UFRJ), Diogo Menezes Ferrazani Mattos (UFF), Igor Monteiro Moraes (UFF), Pedro Henrique Cruz Caminha (UFRJ), Dianne Scherly Varela de Medeiros (UFF), Lucas Airam Castro de Souza (UFRJ), Felipe Gomes Táparo (UFRJ), Miguel Elias Mitre Campista (UFRJ), Luís Henrique Maciel Kosmalski Costa (UFRJ)

Resumo

As redes de acesso via rádio (Radio Access Networks – RANs) atuais adotam soluções monolíticas que implementam toda a pilha de protocolos das comunicações celulares. Essa abordagem monolítica não favorece a integração de infraestruturas de diferentes fornecedores ou o desenvolvimento de ferramentas de controle e gerenciamento agnósticas às interfaces proprietárias. As redes móveis de próxima geração (Next Generation – NextG) são nativamente baseadas em nuvem e construídas sobre arquiteturas desagregadas com hardware de diversos fabricantes e inteligência incorporada. As interfaces de controle padronizadas permitem a definição de laços de controle fechados que garantem a execução de redes autônomas e auto-otimizadas. A O-RAN Alliance é um consórcio formado por membros da indústria e instituições acadêmicas, que visa prover a arquitetura das redes móveis de próxima geração, em que as operadoras de telecomunicações usam interfaces padronizadas e abertas para controlar infraestruturas de diferentes fornecedores e entregar serviços de alto desempenho para os assinantes. O consórcio propõe uma arquitetura inovadora baseada em dois princípios básicos: (i) as funcionalidades da estação rádio-base (Radio Base Station – RBS) são virtualizadas como funções de rede e divididas em vários nós de rede, unidade central (O-RAN Central Unit – O-CU), unidade distribuída (O-RAN Distributed Unit – O-DU) e unidade de rádio (O-RAN Radio Unit – O-RU); e (ii) a existência de um controlador inteligente da rede de acesso via rádio (RAN Intelligent Controller – RIC), que fornece uma abstração centralizada da rede, permitindo que as operadoras implantem funções personalizadas do plano de controle. O consórcio O-RAN prevê laços de controle fechado operando em diferentes escalas de tempo, dependendo das ações de controle, gerenciamento e orquestração a serem implantadas na rede. O objetivo deste capítulo é apresentar a arquitetura O-RAN, com o foco na inteligência para o gerenciamento e a orquestração de serviços.

1.1. Introdução

A rede de acesso via rádio (*Radio Access Network* – RAN) é composta por um conjunto de componentes que interagem entre si para promover a comunicação entre o equipamento de usuário e a rede de núcleo em um sistema de comunicação móvel celular [Arnaz et al., 2022]. Nas primeiras gerações das RANs, uma Estação Rádio-Base (*Radio Base Station* – RBS) é um componente monolítico, acumulando todas as funcionalidades de comunicação via rádio. A RBS conecta-se a uma antena na torre de rádio por meio de cabos elétricos que provocam elevada atenuação e limitam a distância entre a RBS e a antena [Checko et al., 2015]. A terceira geração (3G) separa as funcionalidades de comunicação via rádio em duas partes visando uma melhor flexibilidade. A primeira inclui a transmissão e a recepção, que se tornam responsabilidade da nova RBS, denominada NodeB. A segunda parte inclui o gerenciamento dos recursos de rádio e o processamento relacionado ao usuário, responsabilidades do Controlador da Rede via Rádio (*Radio Network Controller* – RNC) [Arnaz et al., 2022]. A quarta geração (4G) agrega funcionalidades do RNC à RBS, chamada de *Evolved Node B* (eNB), não havendo mais uma entidade de controle separada para a rede de rádio.

Nas redes 3G ou 4G, as funcionalidades da RBS, residentes no NodeB ou no eNB, respectivamente, podem ser desagregadas em Unidade de Rádio Remota (*Remote Radio Head* – RRH), que executa as funções de radio, e na Unidade de Banda Base (*BaseBand Unit* – BBU), responsável pelo processamento de sinal banda base. As funções da RRH consistem, por exemplo, em processamento digital, filtragem de frequência e amplificação de potência. A BBU é responsável pela codificação e aplicação da transformada rápida de Fourier (*Fast Fourier Transform* – FFT) [Checko et al., 2015]. A RRH conecta-se diretamente à antena por meio de um cabo coaxial. A BBU se conecta à RRH por meio de uma rede de transporte denominada *fronthaul*, que pode utilizar fibra óptica ou enlaces de micro-ondas. A desagregação desses dois componentes permite que a BBU esteja fisicamente separada de sua RRH. Assim, a BBU pode ser localizada em um ambiente de mais fácil acesso, podendo estar a até 40 km da RBS [Checko et al., 2015]. A RRH, localizada na RBS, pode assim estar mais próxima da antena, reduzindo a atenuação. Essa arquitetura é denominada RAN distribuída (*Distributed RAN* – D-RAN) [Brik et al., 2022]. Alternativamente, a separação de funcionalidades permite a centralização do serviço da BBU, que pode atender a diversas RRHs, em uma arquitetura denominada RAN centralizada (*Centralized/Cloud RAN* – C-RAN), promovida pela extinta *C-RAN Alliance*. Assim, a C-RAN segue a mesma ideia da computação na nuvem e o dimensionamento de recursos para uma RBS pode ser realizado de acordo com a sua demanda, trazendo eficiência à RAN.

Na quinta geração (5G) de redes celulares, a eNB do 4G evolui para *Next Generation Node B* (gNB). O *3rd Generation Partnership Project* (3GPP) propõe então a desagregação da gNB em três unidades funcionais: a Unidade Central (*Central Unit* – CU), a Unidade Distribuída (*Distributed Unit* – CU) e a Unidade de Rádio (*Radio Unit* – RU) [Polese et al., 2023]. As funcionalidades das camadas física, de enlace e de rede são então divididas entre essas três unidades, como visto mais adiante neste capítulo.

Contrariamente à desagregação dos componentes da RAN, a interação entre os diversos componentes da RAN é feita por interfaces normalmente proprietárias, indepen-

dentemente da geração da rede, forçando a adoção de soluções completas de um único fornecedor por operadora de rede. Os componentes são unidades monolíticas que constituem soluções proprietárias para implantação de RANs, implementando todas as camadas da pilha de protocolos da rede celular [Polese et al., 2023]. Os componentes são produzidos por fornecedores de equipamentos de telecomunicações, que os entregam às operadoras na forma de soluções fechadas. Como consequência, as RANs atuais possuem diversas limitações na capacidade de reconfiguração e refinamento da operação para suportar a diversidade de implementações e diferentes perfis de tráfego. O uso de soluções fechadas também impede a otimização e controle dos componentes da RAN de forma conjunta, dificulta a operação de múltiplas gerações da rede e resulta em bloqueio de fornecedor, limitando as operadoras a implantarem soluções verticais de um único fornecedor. A fim de superar essas limitações, são necessárias soluções abertas para a implementação das RANs. Nesse sentido, o *x-RAN Forum* foi uma iniciativa que visava padronizar a comunicação no *fronthaul* [Polese et al., 2023].

A *O-RAN Alliance*[1] surgiu como fusão do *x-RAN Forum* com a *C-RAN Alliance* [Polese et al., 2023], propondo padronizar uma arquitetura e um conjunto de interfaces para permitirem a realização da RAN aberta [O-RAN Working Group 1, 2023a]. A arquitetura da RAN aberta (*Open RAN* – O-RAN)[2] segue os seguintes princípios fundamentais: desagregação dos componentes da RAN, controle inteligente, virtualização e interfaces abertas [Polese et al., 2023]. A desagregação dos componentes da RAN e sua virtualização permitem a implantação flexível da rede com base em princípios de soluções nativas em nuvem. As interfaces abertas padronizadas abrem o ecossistema da RAN para que empresas menores proponham soluções. As interfaces abertas, juntamente com pilhas de protocolos implantadas em *software*, permitem a integração do controle inteligente. A arquitetura O-RAN visa dividir as funções de rede em componentes de *software* e *hardware*, agnósticos a fornecedores. Assim, a infraestrutura das redes de próxima geração possui a capacidade de fornecer fatias de rede virtual (*slices*) sob demanda e adaptadas a diferentes operadoras de rede virtual, serviços de rede e requisitos de tráfego [Bonati et al., 2021a]. Por fim, a auto-otimização da rede é facilitada por meio da captura e apresentação dos principais indicadores-chave de desempenho (*Key Performance Indicators* – KPIs) e análises da rede por meio das interfaces abertas padronizadas.

O objetivo deste capítulo é apresentar a arquitetura O-RAN, seus princípios de projeto e interfaces. O foco do capítulo é a inteligência para o gerenciamento e a orquestração de serviços. Assim, apresentam-se os princípios do projeto da próxima geração das redes móveis baseada na rede de acesso via rádio aberta; diferenciam-se as principais interfaces, suas funcionalidades e o relacionamento entre módulos da arquitetura O-RAN; discutem-se a principais técnicas e estratégias para realizar o controle inteligente da rede de acesso via rádio; e, por fim, elencam-se os desafios e oportunidades de pesquisa para o desenvolvimento de controles inteligentes em redes móveis de próxima geração.

A Figura 1.1 mostra a organização deste capítulo. A Seção 1.2 apresenta a arquitetura O-RAN. A Seção 1.3 discute o gerenciamento baseado em intenções, contextualiza

[1]Disponível em https://www.o-ran.org/.

[2]O termo *Open RAN* designa iniciativas de RAN aberta em geral, enquanto O-RAN refere-se à arquitetura da *O-RAN Alliance*.

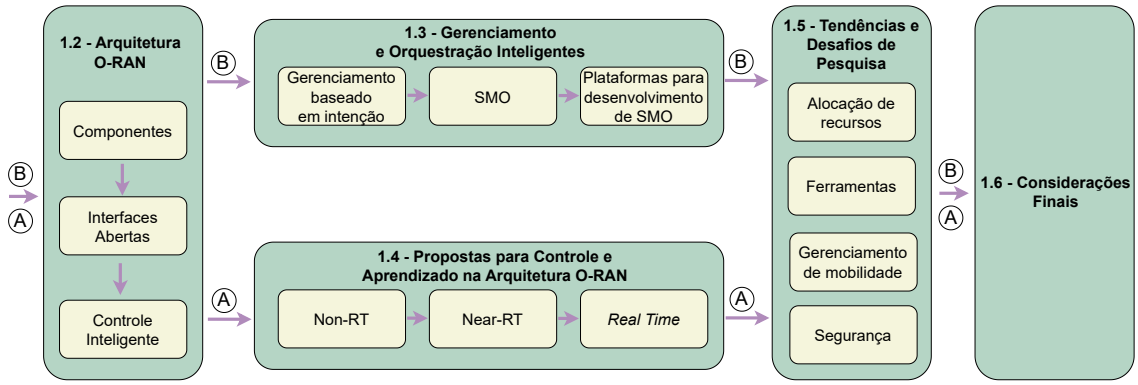

Figura 1.1. Organização deste capítulo e sugestão de dois percursos alternativos. O percurso Ⓐ foca o gerenciamento, orquestração e suas ferramentas, enquanto o Ⓑ foca a abordagem do controle inteligente em artigos científicos.

e exemplifica a implementação do arcabouço de gerenciamento e orquestração de serviços. Propostas para controle da RAN em diferentes escalas de tempo são abordadas na Seção 1.4. As tendências e desafios referentes à pesquisa de interfaces abertas e compreensivas para a RAN são elencados na Seção 1.5. Por fim, as considerações finais são apresentadas na Seção 1.6. Além da leitura linear, seguindo a ordem das seções, a Figura 1.1 mostra dois percursos alternativos. O percurso Ⓐ foca o gerenciamento e orquestração em O-RAN, além das ferramentas associadas. O percurso Ⓑ foca o controle inteligente em O-RAN, revisando a literatura e citando exemplos de aplicações. Em todos os percursos, é possível avançar para as Seções 1.3 ou 1.4 caso o leitor já tenha conhecimentos de O-RAN.

1.2. Arquitetura O-RAN

As redes de acesso atuais são compostas por unidades monolíticas que constituem uma solução *all-in-one*. Essas soluções se caracterizam por implementar todas as camadas da pilha de protocolos da rede celular, sendo fornecidas para as operadoras como "caixas pretas". Esse tipo de solução resulta em reconfigurabilidade limitada, não permitindo ajustes de granularidade fina que suportem a implantação de diferentes perfis de tráfego; coordenação limitada entre os nós da rede, impedindo a otimização e controle conjuntos de componentes da rede de acesso; e dependência de fornecedor, dificultando a utilização pelas operadoras de equipamentos de diferentes fornecedores na rede de acesso. Esses desafios dificultam o gerenciamento otimizado de recursos de rádio e a utilização eficiente do espectro de frequências por meio de adaptação em tempo real [Polese et al., 2023]. A arquitetura definida pela *O-RAN Alliance* para a RAN aberta (Open RAN) tem o objetivo de superar essas limitações, possibilitando a desagregação, virtualização e "softwarização" de componentes, conectando-os através de interfaces abertas padronizadas e permitindo a interoperabilidade entre fornecedores. Dessa forma, há maior flexibilidade na implantação aproveitando-se de princípios das soluções nativas em nuvem e integrando inteligência no controle da rede de acesso [Polese et al., 2023]. Para tanto, as funcionalidades da RBS são desagregadas em três unidades principais (unidade central, unidade distribuída e unidade de rádio). Essas unidades são conectadas a controladores inteligen-

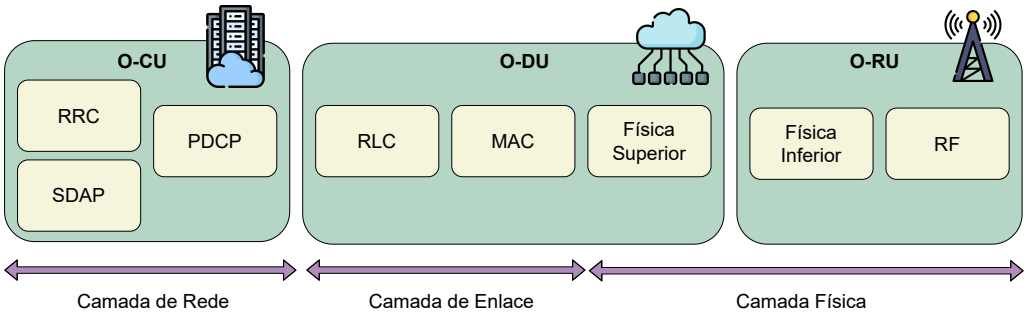

Figura 1.2. Opção de divisão 7.2x e os componentes da O-RAN. A camada física é dividida entre a O-RU e a O-DU. A O-DU também implementa a camada de enlace, enquanto a O-CU é responsável pela camada de rede.

tes através de interfaces abertas. As subseções seguintes detalham a arquitetura O-RAN. Os acrônimos utilizados neste capítulo estão organizados no Apêndice A.

1.2.1. Componentes da Arquitetura O-RAN

A especificação O-RAN para a arquitetura da rede de acesso via rádio aberta tem como base a divisão da estação rádio base em três unidades funcionais, conforme proposto na opção de divisão 7.2x (*Split Option* 7.2x) definida no conjunto de especificações 3GPP *New Radio* (3GPP NR) e apresentada na Figura 1.2. Essa opção de divisão fornece um equilíbrio entre a simplicidade da unidade de rádio e as taxas de dados e latência requeridas entre a unidade de rádio e a unidade distribuída. Assim, a especificação O-RAN [O-RAN Working Group 1, 2023a] cria a Unidade Central O-RAN (O-RAN *Central Unit* – O-CU), Unidade Distribuída O-RAN (O-RAN *Distributed Unit* – O-DU), e a Unidade de Rádio O-RAN (O-RAN *Radio Unit* – O-RU).

A O-RU é um nó lógico que hospeda as funções de camada física inferior (*Low*-PHY) e o processamento de sinais de radiofrequência (RF), incluindo a compensação de fase OFDM (*Orthogonal Frequency Division Multiplexing*) e a transformada rápida de Fourier inversa, estando em acordo com as definições da 3GPP NR 7.2x para a unidade de rádio. A opção de divisão 7.2x também define que a unidade de rádio deve executar operações de adição e remoção de prefixo cíclico. A ideia é tornar a unidade de rádio uma unidade de baixo custo e de fácil implantação.

A O-DU é um nó lógico que hospeda funções de camada física superior (*High*-PHY), a subcamada de controle de acesso ao meio (*Medium Access Control* – MAC) e a subcamada de controle de enlace de rádio (*Radio Link Control* – RLC). As operações realizadas por essas três subcamadas devem ser fortemente sincronizadas, visto que a subcamada MAC gera Blocos de Transporte (*Transport Blocks* – TBs) para serem enviados pela camada física usando dados que são enfileirados pela subcamada RLC. A camada física superior da O-DU deve ser capaz de executar as funções de embaralhamento, modulação, mapeamento de camada e mapeamento de elementos de recursos.

A O-CU implementa as camadas superiores da pilha 3GPP: a camada de Controle de Recursos de Rádio (*Radio Resource Control* – RRC), que gerencia o ciclo de vida das conexões; a camada de Protocolo de Adaptação de Serviços de Dados (*Service Data Adaptation Protocol* – SDAP), que gerencia a qualidade de serviço dos fluxos de tráfego;

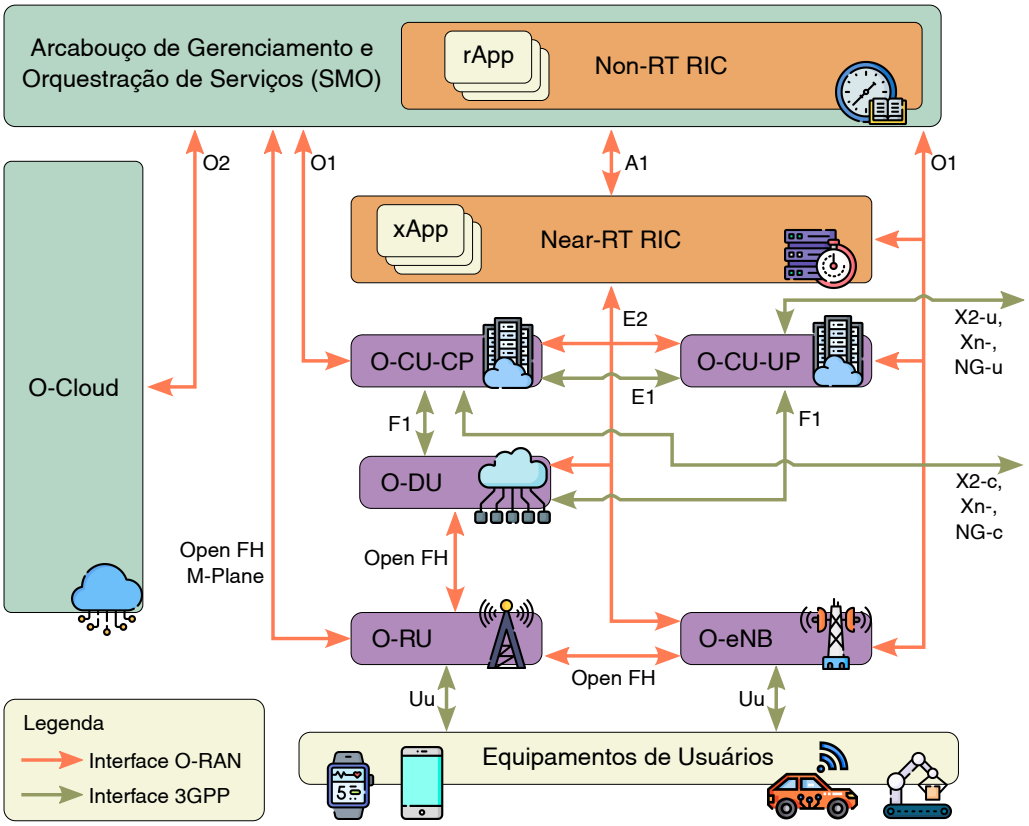

Figura 1.3. Componentes da arquitetura O-RAN. A arquitetura divide as funcionalidades da estação rádio base em três componentes, O-CU, O-DU e O-RU, além de definir o SMO, os RICs e a O-Cloud. Os componentes da arquitetura interagem por meio de interfaces especificadas pela O-RAN Alliance (em laranja) e pelo 3GPP (em verde). Adaptado de [O-RAN Working Group 10, 2023], com ícones de Freepik (flaticon.com).

e a camada de Protocolo de Convergência de Pacotes de Dados (*Packet Data Convergence Protocol* – PDCP), responsável pela reordenação de pacotes, tratamento de pacotes duplicados, criptografia dos dados para a interface aérea, dentre outras funções. A O-CU é responsável por funcionalidades como controle de mobilidade, compartilhamento da RAN, gerenciamento de sessão e transferência de dados do usuário [Arnaz et al., 2022]. A O-CU é subdividida em dois componentes lógicos, um para o plano de controle (O-CU *Control Plane* – O-CU-CP) e outro para o plano de usuário (O-CU *User Plane* – O-CU-UP), a fim de permitir que funcionalidades diferentes possam ser implantadas em diferentes locais da rede e em diferentes plataformas de *hardware*.

A Figura 1.3 mostra a arquitetura O-RAN, com seus diversos componentes e interfaces. Além da desagregação da estação rádio base, a arquitetura O-RAN adota o conceito de componentes programáveis, introduzido através do uso de Controladores Inteligentes da RAN (*RAN Intelligent Controllers* – RICs). Os componentes programáveis são capazes de executar rotinas de otimização com um laço fechado de controle e orquestrar serviços na RAN de forma eficiente, possuindo uma visão abstrata e centralizada da rede. Os RICs introduzidos pela especificação da O-RAN Alliance são o RIC não tempo-real (*Non-Real-Time* RIC – Non-RT RIC) e o RIC quase tempo-real (*Near-Real-Time* RIC –

Near-RT RIC) [Polese et al., 2023], apresentados na Figura 1.3. Os RICs têm acesso a informações de medidas de desempenho e contexto adicionais provenientes de fontes externas à RAN. Esses dados são processados pelos RICs e podem alimentar algoritmos de aprendizado de máquina e inteligência artificial para determinar e aplicar políticas e ações de controle sobre a RAN. Com isso, é possível automatizar procedimentos de otimização da rede com foco no fatiamento dos recursos, balanceamento de carga e mudança de células (*handovers*) [Polese et al., 2023]. A arquitetura O-RAN ainda não define o controle em tempo real, isto é, o controle com tempo de resposta inferior a 10 milissegundos. Portanto, ainda não existe um RIC para tempo real [O-RAN Working Group 1, 2023a].

O Non-RT RIC é um componente do arcabouço de Orquestração e Gerenciamento de Serviços (*Service Management and Orchestration* – SMO), como mostra a Figura 1.3. O Non-RT RIC complementa o Near-RT RIC para fornecer operação inteligente e otimização da RAN em uma escala de tempo maior do que 1 segundo. O SMO é responsável pelo enriquecimento de informação e gerenciamento de modelos de aprendizado de máquina. O Non-RT RIC provê suporte a aplicações de terceiros, as rApps, que fornecem serviços de valor agregado, facilitando a otimização e as operações da RAN [Polese et al., 2023, Arnaz et al., 2022]. O Non-RT RIC pode executar ações de controle a partir do arcabouço do SMO, de forma que, indiretamente, esse RIC pode administrar todos os componentes da arquitetura que estejam conectados ao SMO [Polese et al., 2023].

O Near-RT RIC é implantado na borda da rede e opera laços de controle com periodicidade entre 10 milissegundos e 1 segundo. Esse RIC se comunica com as O-DUs, as O-CUs e as O-eNBs. Uma O-eNB é uma eNB ou *Next-Generation* eNB que suporta a interface E2 [O-RAN Working Group 1, 2023a]. Para a compatibilidade O-RAN, a interface O1 também deve ser suportada. A Figura 1.3 mostra como esses componentes estão interligados. Os equipamentos de usuários (*User Equipments* – UEs) que utilizam serviços da rede 4G/LTE se conectam à O-eNB por meio da interface 3GPP Uu. O provimento de serviços 5G NR é feito por meio da conexão com a O-RU, através da interface Uu. Normalmente, o Near-RT RIC é associado a múltiplos nós da rede de acesso, de forma que o laço de controle fechado de quase tempo real pode afetar a qualidade de serviço de milhares de UEs. O Near-RT RIC é composto por aplicações denominadas xApps e por serviços necessários para a execução das aplicações. A xApp é um microsserviço que pode ser usado para gerenciar recursos de rádio através de interfaces padronizadas e modelos de serviços [Polese et al., 2023, Arnaz et al., 2022]. Para dar suporte às xApps, o Near-RT RIC possui uma base de dados com informações sobre a RAN, como a lista de usuários conectados, que serve como uma camada comum para compartilhamento de dados entre as xApps. O Near-RT RIC também oferece uma infraestrutura de troca de mensagens entre os diferentes componentes, suportando a inscrição de elementos da RAN às xApps. São necessárias terminações para as interfaces abertas e interfaces de programação de aplicação (*Application Programming Interfaces* – APIs) e um mecanismo de resolução de conflitos para orquestrar o controle da mesma função de rede de acesso por múltiplas xApps.

A arquitetura O-RAN prevê a implantação de componentes para gerenciar e otimizar a infraestrutura de rede e as operações, abrangendo desde sistemas de borda até plataformas de virtualização. Nesse sentido, todos os componentes da O-RAN podem

Tabela 1.1. Resumo das funções dos componentes da arquitetura O-RAN.

Componentes	Descrição
SMO	Hospeda o Non-RT RIC e é responsável pelo monitoramento e orquestração da RAN
Non-RT RIC	Suporta rApps, atua em laços de controle maiores que 1 s
Near-RT RIC	Suporta xApps, atua em laços de controle entre 10 ms e 1 s
O-CU	Implementa as camadas superiores da pilha 3GPP, RRC, SDAP, PDCP
O-CU-CP	Componente lógico do plano de controle da O-CU
O-CU-UP	Componente lógico do plano de usuário da O-CU
O-DU	Implementa funções da High-PHY, o MAC e o RLC
O-RU	Implementa funções da Low-PHY e de processamento de sinais de radiofrequência
O-eNB	Estação rádio base 4G/LTE compatível com O-RAN
O-Cloud	Plataforma de computação em nuvem híbrida formada por um conjunto de recursos computacionais e infraestrutura virtualizados reunidos em um ou mais centros de dados

ser implantados em uma plataforma de computação em nuvem híbrida[3], a O-Cloud, que combina nós físicos, componentes de *software* e funcionalidades de gerenciamento e orquestração. Assim, a O-Cloud, mostrada na Figura 1.3, é formada por um conjunto de recursos computacionais e infraestrutura virtualizados reunidos em um ou mais centros de dados [Polese et al., 2023], o que permite o desacoplamento entre componentes de *hardware* e *software*. A O-Cloud permite o compartilhamento de *hardware* entre diferentes inquilinos e automatiza a implantação e instanciação de funcionalidades da RAN, como Funções de Rede Virtuais (*Virtual Network Functions* – VNFs) encontradas na O-CU e as rApps do Non-RT RIC [Arnaz et al., 2022, Polese et al., 2023]. Por meio da padronização de abstrações de aceleração de *hardware*, é possível definir uma API comum entre processadores lógicos baseados em *hardware* dedicados e a infraestrutura O-RAN implantada em *software*. Com essa padronização, a rede de acesso virtualizada passa a suportar os requisitos dos casos de uso do 3GPP NR, como URLLC (*Ultra-Reliable Low-Latency Communication*), usando *hardware* comercial.

1.2.2. Interfaces Abertas

As interfaces definidas pela O-RAN Alliance são específicas para a arquitetura O-RAN, não sendo utilizadas na RAN convencional. O objetivo das interfaces abertas é definir um conjunto de especificações técnicas para padronizar e flexibilizar o acesso aos componentes da RAN, permitindo a conexão entre os diversos componentes da arquitetura [Arnaz et al., 2022], como ilustra a Figura 1.3. Cada interface habilita serviços ao oferecer uma combinação de procedimentos bem definidos [Polese et al., 2023], que envolvem a troca de mensagens nas terminações das interfaces abertas. A arquitetura O-RAN possui interfaces padronizadas tanto pela O-RAN Alliance, como A1, E2, *Open*

[3] A implantação da O-RU em plataforma de computação em nuvem híbrida ainda é objeto de estudo da O-RAN Alliance.

Tabela 1.2. Resumo das interfaces O-RAN e 3GPP presentes na Open RAN.

Interface	Terminação	Tipo
O1	Non-RT RIC, O-eNB Near-RT RIC, O-CU-CP , O-CU-UP, O-DU e O-RU	O-RAN
A1	Non-RT RIC e Near-RT RIC	O-RAN
E2	Near-RT RC e Nós E2	O-RAN
Open FH	O-DU e O-RU	O-RAN
O2	SMO e O-Cloud	O-RAN
E1	O-CU-CP e O-CU-UP	3GPP
F1	O-CU-CP, O-CU-UP e O-RU	3GPP
X2, Xn, NG	O-CU-CP e O-CU-UP	3GPP

FrontHaul (Open FH), O1 e O2, quanto pelo 3GPP, como E1, F1, X2, Xn e NG. A Figura 1.3 mostra as interfaces e os componentes que elas interligam. As interfaces 3GPP possibilitam a desagregação da gNB quando associadas com a interface Open FH. Já as interfaces especificadas pela O-RAN Alliance fornecem dados para os RICs, permitindo a implementação de diversas ações de controle e automação na rede de acesso. Dessa forma, a padronização dessas interfaces auxilia a rede de acesso a não depender de um fornecedor em particular, possibilitando a interoperabilidade entre equipamentos de múltiplos fornecedores. Além disso, essas interfaces permitem selecionar diferentes locais de rede, isto é, nuvem, borda e células, para implantação de diferentes partes de equipamentos. Por exemplo, os RICs podem ser implantados na nuvem, enquanto os O-CUs e O-DUs podem ser implantados na borda e as O-RUs nas células [Polese et al., 2023].

A Tabela 1.2 resume as interfaces da arquitetura O-RAN e suas terminações. O Near-RT RIC é uma das terminações de três interfaces, A1, O1 e E2. O Non-RT RIC também serve como terminação de três interfaces, A1, O1 e O2. A seguir são descritas resumidamente as interfaces padronizadas pela O-RAN Alliance.

Interface O1: serve para comunicação entre o SMO e outros componentes da arquitetura O-RAN. Por exemplo, o SMO usa a interface O1 para comunicação entre o Non-RT RIC e o Near-RT RIC. A interface O1 permite o gerenciamento e orquestração das funcionalidades de rede e segue, sempre que possível, os padrões já existentes estabelecidos pelo 3GPP. Para casos de uso específicos de Open RAN não abordados no padrão 3GPP, o padrão é estendido ou modificado a fim de abordar as necessidades da Open RAN. Pela interface O1 são definidas descrições, requisitos, procedimentos, operações e notificações, a fim de garantir a capacidade de gerenciamento e operação dos componentes da RAN. A interface O1 especifica os Serviços de Gerenciamento (*Management Services* – MnS) suportados na arquitetura O-RAN entre os provedores de MnS e o SMO. Os serviços de gerenciamento incluem o gerenciamento do ciclo de vida dos componentes da O-RAN e a coleta de KPIs.

Dentre os MnS, o de Provisionamento permite que o SMO insira configurações nos nós gerenciados e que os nós gerenciados reportem as atualizações de configurações externas para o SMO. O envio dessas mensagens é feito por meio de uma combinação de APIs REST (*Representational State Transfer)*/HTTPS e NETCONF. Já o MnS de Su-

pervisão de Falha é usado para reportar erros e eventos ao SMO. Nesse caso, os nós da RAN podem reportar os erros usando APIs REST. O MnS de Pulsação (*heartbeat*) permite ao SMO fornecer uma pulsação para os dispositivos gerenciados e gerenciar funções de rede virtuais e físicas. As mensagens de pulsação, isto é, um pacote de dados com informações vitais de gerenciamento, são usadas para monitorar o estado e a disponibilidade dos serviços e nós [Polese et al., 2023]. O MnS de Garantia de Desempenho pode ser usado para transferência em tempo real ou para reportar, via transferência de arquivo, dados de desempenho para o SMO. A ideia é habilitar, por exemplo, a análise e coleta de dados para o fluxo de trabalho de inteligência artificial e aprendizado de máquina. No caso da transferência de arquivos, utiliza-se o protocolo SFTP [Polese et al., 2023]. A transferência de arquivos entre produtores e consumidores é possível graças ao MnS de Gerenciamento de Arquivos [O-RAN Working Group 1, 2021]. Eventos baseados em rastreio podem ser monitorados pelo MnS de Rastreamento, como perfil de chamadas, estabelecimento de conexão da camada de controle de recursos de rádio e falhas de enlace de rádio [Polese et al., 2023]. O MnS de Registro e Inicialização de Funções de Rede Físicas (*Physical Network Functions* – PNF) permite que um nó produtor MnS adquira seus parâmetros da camada de rede via procedimentos estáticos pré-configurados no nó ou via procedimentos dinâmicos durante a inicialização do nó. Durante o processo de aquisição, o nó produtor também adquire o endereço IP do nó consumidor com o qual interage para se registrar. Após o registro, o consumidor MnS pode trocar o estado do produtor para operacional. Por fim, o MnS de *Software* de PNF permite que um nó consumidor solicite a um nó produtor o download, instalação, validação e ativação de novos pacotes de *software*, além de permitir que o produtor reporte suas versões de *software* [O-RAN Working Group 1, 2021]. A Tabela 1.3 resume os MnS. Adicionalmente, existem especificações próprias complementares para funções de gerenciamento específicas do Near-RT RIC, O-CU e O-DU [Polese et al., 2023].

Tabela 1.3. Resumo das funções dos *Management Services* (MnS).

MnS	Descrição
Provisionamento	Permite ao SMO configurar os nós e aos nós relatarem ao SMO as atualizações de configurações externas
Supervisão de Falha	Permite que os nós reportem erros e eventos ao SMO
Pulsação	Permite ao SMO o envio de mensagens de pulsação e o gerenciamento de funções de rede virtuais e físicas
Garantia de Desempenho	Possibilita a transferência de dados de desempenho dos nós ao SMO em tempo real
Gerenciamento de Arquivo	Permite a transferência de arquivos a partir do protocolo SFTP
Rastreamento	Monitora eventos baseados em rastreio
Registro e Inicialização de Funções de Rede Física	Permite a aquisição de parâmetros da camada de rede por um nó produtor
Software de PNF	Permite a solicitação de *download*, instalação, validação e ativação de novos pacotes de *software*

Interface A1: é usada pelo Non-RT RIC para enviar informações ao Near-RT RIC, como dados sobre os casos de uso e enriquecimento de informação. O propósito da interface A1 é permitir que o Non-RT RIC envie orientações baseadas em políticas, gerencie modelos de aprendizado de máquina e envie informações para o Near-RT RIC com o objetivo de otimizar a RAN. Essa comunicação é feita por meio de mecanismos padronizados baseados em uma sintaxe específica que pode expressar intenções de alto nível e políticas. Dessa forma, permite-se a implementação do controle Non-RT (*non-real time*) e de políticas e modelos inteligentes no Near-RT RIC [Polese et al., 2023]. A interface A1 depende do protocolo A1AP (*A1 interface Application Protocol*), que é baseado em um arcabouço 3GPP para implantação de políticas para funções de rede, combinando APIs REST sobre HTTP para transferência de objetos (*JavaScript Object Notation* – JSON) [Polese et al., 2023].

Os serviços suportados pela interface A1 incluem o serviço de gerenciamento de políticas, o serviço de enriquecimento de informação e o serviço de gerenciamento de modelos de aprendizado de máquina. O Serviço de Gerenciamento de Políticas A1 (A1-P) é usado pelo Non-RT RIC para conduzir as funcionalidades do Near-RT RIC de forma a alcançar a intenção de alto nível para a RAN. O Non-RT RIC define as políticas para o Near-RT RIC a partir da observação de eventos e das intenções do sistema. O Near-RT RIC então envia um *feedback* pela interface A1 para o Non-RT RIC, que avalia os impactos das políticas a partir desse *feedback* e das informações sobre a rede obtidas através da interface O1. A partir dessas informações, o Non-RT RIC pode decidir atualizar ou modificar as políticas A1. O Serviço de Enriquecimento de Informação A1 (A1-EI) tem o objetivo de aprimorar o desempenho da RAN fornecendo informação que normalmente não está disponível para a RAN, como previsão de capacidade. Como o Non-RT RIC e o SMO têm uma perspectiva global da rede e acesso a fontes externas de informação, estes podem encaminhar essas informações às xApps no Near-RT RIC usando o serviço A1-EI. Em conjunto com informações já disponíveis, as informações enriquecidas aumentam o desempenho do sistema. Com base nesses dados, o Non-RT RIC pode inferir informações que beneficiem tanto as funções do Non-RT RIC quanto as funções do Near-RT RIC. A interface A1 é utilizada para a descoberta, requisição e entrega de informações enriquecidas, além de descoberta de informações de enriquecimento externas [Polese et al., 2023]. O Serviço de Gerenciamento de Modelos de Aprendizado de Máquina (A1-ML) auxilia no gerenciamento dos modelos de aprendizado de máquina da RAN, permitindo o *download* e distribuição, e o *upload* e agregação em aprendizado federado [O-RAN Working Group 2, 2021a]. Os modelos de aprendizado de máquina podem ser treinados e executados em diferentes locais da arquitetura O-RAN, incluindo o Non-RT RIC e o Near-RT RIC. No caso do Near-RT RIC, o modelo é treinado no SMO e implantado no Near-RT RIC através da interface O1 para otimização da RAN. Para dar suporte aos modelos do Near-RT RIC, o Non-RT RIC pode prover informações enriquecidas via interface A1. Já no caso do Non-RT RIC, os modelos são treinados no SMO e usados pelo Non-RT RIC para aprimorar o monitoramento e a orientação da RAN com base na observabilidade da interface O1. O treinamento e a implantação do modelo são feitos pelo SMO [O-RAN Working Group 2, 2023].

Interface E2: é por meio dessa interface que ocorre a comunicação do Near-RT RIC com os elementos gerenciados, isto é, todos os componentes lógicos da RAN que estão

conectados ao Near-RT RIC pela interface E2. Esses elementos são denominados nó E2, como O-CU, O-DU e O-eNB. A interface E2 possibilita os laços de controle de quase tempo real por meio da transmissão de dados de telemetria da RAN e da resposta de controle do Near-RT RIC. Dessa forma, o Near-RT RIC consegue coletar dados sobre os nós E2. Para isso, a O-RAN Alliance utiliza um conjunto de identificadores únicos baseados nas especificações do 3GPP para a gNB, fatias de rede e classes de qualidade de serviço [Polese et al., 2023]. Para os equipamentos de usuário, a O-RAN Alliance define um identificador comum de usuário (*UE Identifier* – UE-ID) em suas especificações, possibilitando a identificação do mesmo usuário em diferentes nós E2. Assim, existe uma identidade de usuário uniforme e consistente em todo o sistema sem expor informações sensíveis relacionadas ao usuário.

As aplicações que operam sobre a interface E2 usam os *E2 Service Models* (E2SMs) e a comunicação é regida pelo *E2 Application Protocol* (E2AP). O E2AP coordena a comunicação entre o Near-RT RIC e os nós E2. Este protocolo trata do gerenciamento de interface, configuração e conexão dos nós E2 ao Near-RT RIC. A conexão é estabelecida por meio do *Stream Control Transmission Protocol* (SCTP) e, após a conexão, o E2AP provê os serviços RIC, que podem ser combinados de maneiras diferentes para a implementação dos E2SMs [Polese et al., 2023]. Por exemplo, o nó E2 pode transmitir uma solicitação de configuração E2 na qual lista as funções RAN e configurações suportadas juntamente com os identificadores do nó. Ao processar a informação, o Near-RT RIC responde com uma mensagem de configuração de resposta E2. Um E2SM descreve as funções do nó E2 que podem ser controladas pelo Near-RT RIC e os procedimentos relacionados. Assim, um E2SM define uma divisão de gerenciamento de recurso de rádio (*Radio Resource Management* – RRM) específica de função entre o nó E2 e o Near-RT RIC. O Near-RT RIC pode monitorar, suspender, parar, sobrescrever ou controlar o comportamento do nó E2 por meio de políticas, através das funções expostas no E2SM. Assim, os E2SMs definem protocolos específicos de função que são implementados sobre a especificação E2AP [O-RAN Working Group 3, 2023b]. A comunicação é feita sobre SCTP. Atualmente os E2SMs definidos pelas especificações O-RAN são E2SM *Network Interface* (E2SM-NI), E2SM *Key Performance Measurement Monitor* (E2SM-KPM), E2SM RAN *Control* (E2SM-RC) e E2SM *Cell Configuration and Control* (E2SM-CCC) [O-RAN Working Group 3, 2023c].

Cada nó E2 expõe algumas funções RAN, que definem os serviços e capacidades suportados pelos nós. Assim, é possível separar de forma clara as capacidades de cada nó e definir como as xApps devem interagir com a RAN. Após o estabelecimento da conexão entre o Near-RT RIC e o nó E2, o E2AP provê quatro serviços: REPORT, INSERT, CONTROL, POLICY e QUERY [O-RAN Working Group 3, 2023b]. Existem ainda funções de suporte RIC que envolvem procedimentos de gerenciamento de interface e procedimentos de serviços de funções da RAN. Essas funções são E2 SETUP, E2 RESET, RIC SERVICE UPDATE, E2 NODE CONFIGURATION UPDATE e E2 REMOVAL. A combinação desses serviços cria um modelo de serviço, cuja mensagem é inserida como carga útil de uma mensagem E2AP. O conteúdo é codificado utilizando a notação *Abstract Syntax Notation One* (ASN.1)[4].

[4]Padrões ITU-T X.680 a X.699, disponíveis em https://www.itu.int/ITU-T/recommendations/rec.aspx?rec=x.680.

Interface Open FH: permite a interação entre as O-RUs e as O-DUs, conectando a O-DU a uma ou mais O-RUs dentro da mesma gNB. A partir dessa interface, é possível distribuir as funcionalidades da camada física entre a O-DU e a O-RU. Além disso, a interface permite controlar as operações da O-RU a partir da O-DU. A interface Open FH suporta comunicação confiável e de baixa latência entre O-DUs e O-RUs com temporização adequada aos requisitos de fluxos URLLC. O protocolo O-RAN FH inclui quatro planos distintos: Plano de Controle (*Control Plane* – C-Plane), Plano de Usuário (*User Plane* – U-Plane), Plano de Sincronização (*Synchronization Plane* – S-Plane) e Plano de Gerenciamento (*Management Plane* – M-Plane). O C-Plane trata da transferência de comandos entre a camada física superior da O-DU e a camada física inferior da O-RU. Os comandos estão relacionados, por exemplo, às configurações de escalonamento e alinhamento de feixe (*beamforming*) e controle de compartilhamento de espectro. As mensagens do C-Plane são encapsuladas pelos protocolos *evolved Common Public Radio Interface* (eCPRI) ou IEEE 1914.3 com cabeçalhos e comandos específicos para diferentes procedimentos de controle. O U-Plane tem como principal função transferir amostras de sinais em fase e quadratura (I/Q) no domínio da frequência entre a O-RU e a O-DU. O U-Plane também é responsável pela temporização da transmissão de mensagens de forma que sejam recebidas na O-RU com tempo suficiente para processamento antes da transmissão. Esse plano também especifica o ganho digital das amostras e pode comprimi-las para melhorar a eficiência da transmissão dos dados. O S-Plane é responsável por sincronizar o tempo, frequência e fase entre o relógio da O-DU e das O-RUs. Dessa forma, o S-Plane fornece uma referência de relógio compartilhada que permite que a O-DU e a O-RU estejam adequadamente sincronizadas no tempo e na frequência para transmissão e recepção dos sinais. Existem diferentes perfis de de sincronização nas especificações, baseados em diferentes protocolos, como o *Physical Layer Frequency Signals* (PLFS) e o PTP (*Precision Time Protocol*) que podem alcançar uma precisão temporal de submicrossegundo. O M-Plane é um plano que funciona em paralelo aos outros e permite a inicialização e gerenciamento da conexão entre O-DU e O-RU, além da configuração da O-RU. As terminações do M-Plane na O-DU e na O-RU são dedicadas e estabelecem um túnel IPv4 ou IPv6. As especificações preveem duas opções de implantação do M-Plane. Na opção hierárquica, o SMO gerencia a O-DU e a O-DU gerencia a O-RU. Na opção híbrida, o SMO também pode interagir diretamente com a O-RU. As mensagens do M-Plane são criptografadas fim-a-fim por SSH e/ou TLS [Polese et al., 2023].

Interface Open O2: permite a comunicação entre o SMO e a O-Cloud. Com isso, é possível suportar funcionalidades que executam na nuvem. A interface O2 permite definir um inventário dos recursos controlados pela O-Cloud, monitoramento, provisionamento, tolerância a falhas e atualizações. A O-RAN Alliance considera adotar para a interface O2 padrões e soluções abertas, como os padrões da *European Telecommunications Standards Institute* (ETSI) para *Network Function Virtualization* (NFV), interfaces baseadas em serviços do 3GPP, e os projetos Kubernets, OpenStack e ONAP/OSM [Polese et al., 2023]. Existem duas classes de funções oferecidas pela interface O2 que residem na O-Cloud: funções que gerenciam a infraestrutura e funções que gerenciam implantações na infraestrutura. Os Serviços de Gerenciamento de Infraestrutura (*Infrastructure Management Services* – IMS) incluem as funções da interface O2 responsáveis pela implementação e gerenciamento da infraestrutura em nuvem. Os Serviços de Gerenci-

amento de Implantação (*Deployment Management Services* – DMS) incluem as funções relacionadas ao gerenciamento de funções virtualizadas na infraestrutura em nuvem [O-RAN Working Group 6, 2023].

Interfaces E1, F1, X2, Xn, NG, Uu: alguns componentes herdados de outras gerações da RAN usam as mesmas interfaces usadas nas arquiteturas dessas gerações. A interface E1 é um exemplo, sendo responsável por realizar a conexão entre o plano de controle e o plano de usuário presente na O-CU. A interface F1 conecta elementos da O-DU e O-CU para troca de informação sobre o compartilhamento de recursos de rádio e sobre outros estados da rede. As interfaces X2 e Xn ajudam com a interoperabilidade entre nós de diferentes gerações e a interface NG conecta nós 5G à rede de núcleo quando esta opera no modo *standalone*, ou seja, 5G puro. A interface Uu permite a conexão dos UEs à rede.

1.2.3. Controle Inteligente da RAN e Aplicações

As especificações da O-RAN descrevem requisitos e funcionalidades de diferentes componentes dos RICs (*RAN Intelligent Controllers*), de forma que implementações em conformidade com o padrão forneçam os mesmos conjuntos de serviços. Apesar de essas especificações não definirem requisitos de implementação, a Comunidade de Software da O-RAN (*O-RAN Software Community* – OSC) fornece referências de implementação de um Near-RT RIC que segue as especificações O-RAN e podem ser usadas para desenvolvimento de protótipos de soluções O-RAN. A referência de implementação tem como base o uso de múltiplos componentes executados como microsserviços em um *cluster* Kubernetes [Polese et al., 2023]. É importante destacar que o controle inteligente da RAN é efetuado por meio de xApps, que executam no Near-RT RIC, e de rApps, que executam no Non-RT RIC.

Near-RT RIC: os principais componentes de um Near-RT RIC são a infraestrutura de mensagens internas, o componente de mitigação de conflitos, o gerenciador de assinaturas, o subsistema de segurança, o banco de dados da Base de Informações de Rede (*Network Information Base* – NIB), a API de camada de compartilhamento de dados, e o gerenciador de xApp [Polese et al., 2023]. A Figura 1.4 mostra esses componentes. A infraestrutura de mensagens interna interconecta xApps, plataformas de serviços e terminações de interfaces. É necessário suporte ao registro, descoberta e exclusão de terminações e o fornecimento de APIs para envio e recebimento de mensagens, seja por mecanismos de comunicação ponto-a-ponto ou publicador/assinante (*publisher/subscriber*). O componente para mitigação de conflitos deve lidar com os possíveis conflitos entre diferentes xApps, que podem surgir quando xApps distintas requerem configurações conflitantes ao tentar alcançar os objetivos de otimização individuais. Esses conflitos podem resultar em degradação do desempenho geral da rede. As especificações O-RAN destacam três classes de conflitos: diretos, indiretos e implícitos. Os conflitos diretos podem ser detectados pelo componente interno de mitigação de conflitos, por exemplo, quando xApps solicitam mais recursos do que o disponível. Já os indiretos e implícitos não são observados diretamente e podem ser dependentes da relação entre diferentes xApps, por exemplo, configurações que otimizam o desempenho de uma classe de usuários podem degradar o desempenho de outros usuários de forma inesperada. Os conflitos diretos podem ser resolvidos por meio de resoluções pré-ação, por exemplo, limitando o escopo de uma

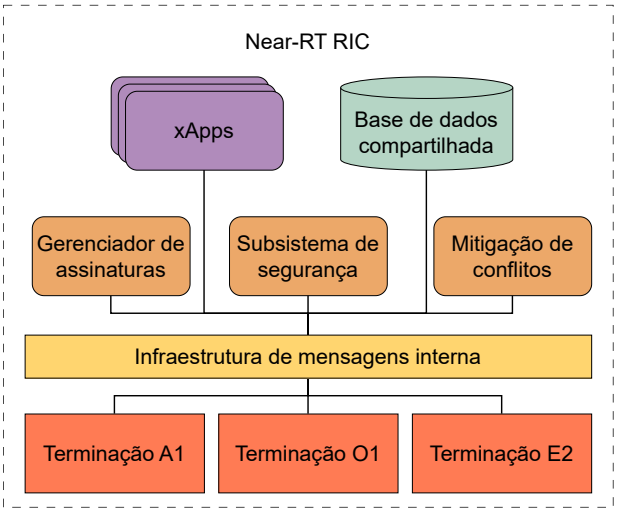

Figura 1.4. O Near-RT RIC é um componente programável da Open RAN que implementa funcionalidades que executam em uma escala de tempo entre 10 ms e 1 s. Os diversos componentes desse RIC oferecem suporte para a execução das xApps. A comunicação com o restante dos componentes da arquitetura é feita pelas interfaces A1, O1 e E2.

ação de controle. Já os indiretos e implícitos são solucionados por verificações pós-ação, ou seja, monitorando o desempenho do sistema após a aplicação de determinada política de controle. O subsistema de segurança previne o vazamento de dados da RAN por xApps maliciosas. Além disso, previne que essas xApps afetem o desempenho da RAN negativamente. A NIB armazena informações sobre os nós E2 e sobre os equipamentos dos usuários e suas identidades. Já a API de camada de compartilhamento de dados serve para que os componentes da plataforma RIC, incluindo as xApps, possam solicitar dados à NIB. O gerenciador de xApp fornece serviços e APIs para o gerenciamento automatizado do ciclo de vida das xApps, incluindo a implantação e terminação, registros de falha, configuração, contabilização, desempenho e segurança (*Fault, Configuration, Accounting, Performance, Security* – FCAPS).

As xApps executadas no Near-RT RIC são componentes *plug-and-play* que implementam uma lógica personalizada. Essas aplicações podem ser usadas, por exemplo, para controle e análise de dados da RAN. As xApps têm acesso a informações de telemetria da RAN e podem enviar ações de controle para serem executadas por elementos da RAN através da interface E2. Um sistema de assinatura permite que xApps se conectem a funções expostas pela interface E2, controlando também o acesso individual das xApps às mensagens nessa interface. Assim, a interface E2 possibilita a associação direta entre a xApp e a funcionalidade da RAN. As informações usadas pelas xApps são obtidas dos modelos de serviço da interface E2 (E2SMs) associados a elas, utilizando as APIs do Near-RT RIC [O-RAN Working Group 3, 2023a]. Isso é possível porque a associação da xApp com um E2SM deve garantir a possibilidade de interação entre a xApp e qualquer nó E2 suportado pelo E2SM associado. As xApps devem ser capazes de fornecer informações sobre registros de coleta, rastreio e métricas para o Near-RT RIC [O-RAN Working Group 3, 2023a].

A xApp é definida por uma imagem de *software* e por um descritor, que inclui informações sobre parâmetros necessários para gerenciar a aplicação e que pode descrever os tipos de dados consumidos e gerados pela xApp e as capacidades de controle [Polese et al., 2023]. O descritor também deve incluir informações de configuração da xApp e uma lista com as métricas fornecidas pela xApp [O-RAN Working Group 3, 2023a]. Essas aplicações podem ser compostas por um ou mais microsserviços, sendo independentes do Near-RT RIC e podendo ser fornecidas por terceiros. No Near-RT RIC da implementação de referência da OSC, a xApp é definida por uma imagem Docker que pode ser implantada em uma infraestrutura Kubernetes por meio da aplicação de um esquema descritor, isto é, um arquivo que especifica os atributos do contêiner [Polese et al., 2023].

Non-RT RIC: é parte do arcabouço SMO e implementa um subconjunto de funcionalidades desse arcabouço. A Figura 1.5 mostra essas funcionalidades. O principal objetivo do Non-RT RIC é realizar a otimização inteligente da RAN por meio de orientação baseada em políticas, gerenciamento de modelos de aprendizado de máquina e enriquecimento de informação para o Near-RT RIC. Dessa forma, o Non-RT RIC é responsável pelos procedimentos de orquestração, gerenciamento e automação para monitorar e controlar os componentes da RAN. O Non-RT RIC e o SMO são terminações lógicas da interface A1. Através dessa interface, o Non-RT RIC pode acessar funcionalidades do arcabouço SMO que não estão implementadas no RIC, influenciando, por exemplo, o que é transferido pelas interfaces O1 e O2. A interação entre as funcionalidades do Non-RT RIC e do SMO é feita por meio de uma infraestrutura interna de mensagens. O Non-RT RIC é composto pelo arcabouço Non-RT RIC e pelas aplicações Non-RT RIC (rApps), como mostra a Figura 1.5. O arcabouço Non-RT RIC oferece serviços para as rApps através da interface R1 dessas aplicações [O-RAN Working Group 1, 2023a]. As rApps, por sua vez, são aplicações modulares que aproveitam as funcionalidades oferecidas pelo arcabouço Non-RT RIC para oferecer serviços de valor agregado a fim de suportar e facilitar a otimização e operação da RAN, além de executar outras funções.

Algumas funcionalidades podem residir tanto no SMO quanto no Non-RT RIC. A infraestrutura de mensagens interna é composta por diversas funções do SMO que permitem que todos os componentes do SMO, inclusive os que fazem parte do Non-RT RIC, acessem e utilizem interfaces, dados e funcionalidades oferecidos tanto pelo SMO quanto pelo Non-RT RIC. Por exemplo, políticas criadas por rApps podem alcançar o Non-RT RIC por meio da terminação R1 e eventualmente alcançar o Near-RT RIC pela interface A1. Para isso, todas as terminações de interfaces são ligadas a funções específicas de interface incluídas na infraestrutura de mensagens interna que facilita a troca de mensagens entre as terminações. O serviço de exposição e gerenciamento de dados também reside em ambos os componentes, SMO e Non-RT RIC. As rApps podem consumir dados produzidos por componentes do SMO ou do Non-RT RIC [Polese et al., 2023]. Outra funcionalidade que reside tanto no SMO quanto no Non-RT RIC é o fluxo de trabalho de aprendizado de máquina e inteligência artificial [Polese et al., 2023].

As especificações da O-RAN Alliance definem alguns requisitos para as rApps. Por exemplo, essas aplicações devem ser capazes de se comunicar por meio da interface R1, fornecendo informações relacionadas aos tipos de dados e à periodicidade com

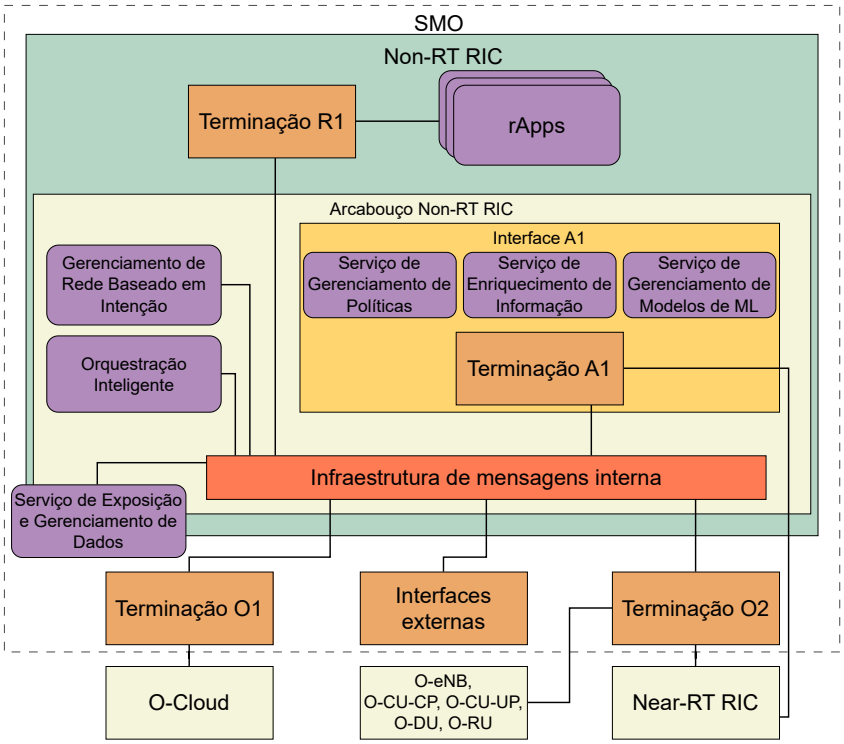

Figura 1.5. O Non-RT RIC faz parte do SMO e é um componente programável da Open RAN que implementa funcionalidades que executam em uma escala de tempo igual ou maior do que 1 s. Os diversos componentes desse RIC oferecem suporte para a execução das rApps. A comunicação com o restante dos componentes da arquitetura é feita pelas interfaces A1, O1 e O2.

a qual a rApp consome e produz dados. Os serviços oferecidos às rApps através da interface R1 possibilitam a obtenção de acesso aos serviços de exposição e gerenciamento de dados, funcionalidades de aprendizado de máquina e inteligência artificial, bem como às interfaces A1, O1 e O2 por meio da infraestrutura de mensagens interna [O-RAN Working Group 1, 2023a]. Dessa forma, as rApps oferecem serviços de orientação de políticas, enriquecimento de informação, gerenciamento de configuração e análise de dados [Polese et al., 2023].

As rApps realizam diversas tarefas de automação e gerenciamento, com laços de controle em uma escala de tempo igual ou maior que um segundo. Durante interações em que dados precisam ser registrados, se não houver uma fonte de dados correspondente para um determinado tipo de dado consumido, a rApp deve ser capaz de determinar se pode ou não continuar a execução sem aquele tipo de dado. Se as necessidades de consumo de dados da rApp não puderem ser cumpridas para alguns tipos de dados, a rApp deve ser capaz de interromper as interações de registro. Quando as necessidades de consumo de dados de uma rApp são modificadas, a rApp é responsável por determinar como acomodar essa mudança [O-RAN Working Group 2, 2021b]. Apesar das rApps poderem suportar as mesmas funcionalidades de controle fornecidas pelas xApps, como direcionamento de tráfego, controle de escalonamento, e gerenciamento de *handover*, em uma escala de tempo maior, as rApps são padronizadas para derivar políticas de controle que operam em alto nível e que podem afetar uma maior quantidade de usuários e nós da rede.

As rApps podem interagir entre si através de interfaces padronizadas para construir funções de automação de rede mais complexas. Alguns exemplos de rApps para aplicações de controle Non-RT da RAN são o gerenciamento de frequências e interferência, compartilhamento da RAN, diagnóstico de desempenho, garantia de Acordo de Nível de Serviço (*Service Level Agreement* – SLA) fim-a-fim e fatiamento da rede [Polese et al., 2023]. No mercado, já existem fabricantes desenvolvendo rApps. A Ericsson, por exemplo, define quatro categorias principais de rApps: as rApps para evolução da rede, implantação da rede, otimização da rede e cura da rede[5].

O Non-RT RIC oferece dois serviços de gerenciamento e orquestração de alto nível. Essa oferta permite que a arquitetura seja suficientemente flexível para que o comportamento de cada componente da rede e funcionalidade possa ser ajustado em tempo real, atendendo aos objetivos e intenções das operadoras. O primeiro serviço é o gerenciamento de rede baseado em intenção, que permite às operadoras especificar intenções utilizando uma linguagem de alto nível, por exemplo linguagem natural, por meio de uma interface homem-máquina. A intenção é automaticamente analisada pelo Non-RT RIC que determina a política e o conjunto de rApps e xApps que devem ser implantadas e executadas para satisfazer as políticas. O segundo serviço é a orquestração inteligente, que permite coordenar e orquestrar as diferentes xApps e rApps que executam em diferentes RICs e locais da rede. O Non-RT RIC é responsável pela orquestração da inteligência da rede para garantir que as aplicações selecionadas sejam adequadas para satisfazer as intenções da operadora e atender aos requisitos impostos. Além disso, o Non-RT RIC deve garantir que as aplicações sejam instanciadas no local apropriado para garantir o controle sobre os elementos da RAN especificados na intenção, sejam alimentadas com dados relevantes, e sejam robustas o suficiente para não gerarem conflitos por condição de corrida entre as aplicações [Polese et al., 2023].

1.3. Gerenciamento e Orquestração Inteligentes

A orquestração e o gerenciamento de serviços (*Service Management and Orchestration* – SMO) na arquitetura O-RAN extrapolam o gerenciamento da rede de acesso via rádio, como definido pelo 3GPP *(NG-)core* e o gerenciamento de fatias de rede de ponta a ponta [Lopez et al., 2022]. Na O-RAN, os módulos SMO são responsáveis por interfaces de gerenciamento FCAPS para funções de rede O-RAN, pela otimização da rede de acesso via rádio em larga escala e pelo gerenciamento e orquestração da O-Cloud por meio da interface O2, incluindo descoberta de recursos, dimensionamento, FCAPS, gerenciamento de *software* e criação, leitura, atualização e exclusão (*create, read, update, delete* – CRUD) de recursos na nuvem. Assim, a orquestração e o gerenciamento são operações que afetam todo o ciclo de vida das funções de rede, desde o seu projeto, criação, otimização, operação, até o inventário de recursos e a extinção da função de rede O-RAN.

1.3.1. Gerenciamento baseado em intenção

O objetivo do gerenciamento baseado em intenção é tornar o gerenciamento e a operação da rede mais simples, exigindo mínima intervenção externa [Clemm et al., 2022]. Para tanto, a intenção é definida como um conjunto de

[5]https://www.ericsson.com/en/ran/intelligent-ran-automation/intelligent-automation-platform/rapps.

objetivos operacionais que a rede deve alcançar e resultados que a rede deve entregar, especificados de uma maneira declarativa, porém sem indicação explícita de como alcançá-los ou implementá-los [Clemm et al., 2022]. A intenção, geralmente, é definida em linguagem natural. A seguir são apresentados exemplos de intenção retirados da RFC 9315 [Clemm et al., 2022]:

1. Desvie o tráfego de rede originário de pontos finais (*endpoints*) que pertencem a uma Região Geográfica A de uma Região Geográfica B, a menos que o destino do tráfego esteja na Região Geográfica B;

2. Evite encaminhar tráfego de rede originário de um conjunto de pontos finais ou tráfego de rede associado a um dado cliente através de equipamentos de um vendedor específico, mesmo que isso custe uma redução dos níveis de serviço;

3. Maximize o uso da rede mesmo se isso significar uma redução dos níveis de serviço, como aumento da latência ou a perda de pacotes, a menos que os níveis de serviço tenham se deteriorado em 20% ou mais em relação à sua média histórica;

4. Garanta que os serviços de redes privadas virtuais (*Virtual Private Networks* - VPNs) tenham proteção de caminho em todos os momentos para todos caminhos.

Os Exemplos 1 e 2 definem os objetivos a serem alcançados, mas não como alcançá-los. No Exemplo 2, ainda são dadas informações adicionais de compromisso entre diferentes objetivos para serem usadas, se necessário. Os Exemplos 3 e 4 definem um resultado desejado que a rede deve entregar sem especificar como alcançá-lo, sem nenhum detalhe de implementação.

O princípio de funcionamento do gerenciamento baseado em intenção é mais do que simplesmente definir mecanismos que permitam a interação do operador com a rede usando abstrações de alto nível. O objetivo é fazer com que o foco dos operadores seja nos resultados desejados, deixando para a rede os detalhes sobre como alcançar tais resultados. O foco nos resultados leva a um aumento da eficiência operacional e da flexibilidade, em escalas de tempo menores e com menos dependência de intervenções humanas e, portanto, com menos possibilidade de erros. Por conta do foco no resultado, o gerenciamento baseado em intenção é um candidato para aplicação de técnicas de inteligência artificial [Clemm et al., 2020].

A orquestração da RAN depende da implantação de políticas complexas. Contudo, nas redes de comunicação móveis atuais, isso é um desafio para as operadoras, pois as políticas normalmente descrevem objetivos de alto nível ou intenções de negócios. Os objetivos de alto nível são representados por KPIs, índices que permitem que gestores acompanhem a evolução das operações, abstraindo especificidades de gerenciamento e operação das redes. As operadoras executam, então, o trabalho complexo, e sujeito a erros, de dividir cada política em ações de baixo nível a serem implantadas nos dispositivos físicos ou virtuais relevantes [Jacobs et al., 2021].

A ideia do gerenciamento baseado em intenção para a RAN é transformar a configuração da RAN de um ajuste de parâmetros técnicos como, por exemplo, os limiares

de *handover*, para definições de alto nível, no caso, a intenção. Dessa forma, as operadoras podem especificar o serviço de conectividade propriamente dito e, por exemplo, definir níveis de prioridade diferentes entre usuários e serviços baseados em intenções de negócio [Westerberg e Fiorani, 2020].

Em uma RAN, tipicamente, existem milhões de decisões tomadas a cada segundo sobre qual usuário atender pela interface de rádio e como atender a esse usuário. Cada uma dessas decisões contribui para a qualidade do serviço e a priorização entre usuários e serviços em caso de conflitos. Tradicionalmente, essas decisões são determinadas por uma combinação de opções de projeto do fornecedor e definições de parâmetros de configuração de rede feitas pela operadora. Nos sistemas 2G, relativamente simples, o efeito de uma mudança de configuração era quase sempre possível de se entender. Nas redes de nova geração multisserviço, é praticamente impossível, de maneira econômica, prever o efeito que um determinado conjunto de alterações de configuração terá nos serviços do usuário final. No entanto, a intenção da RAN continua a mesma de oferecer conectividade aos clientes das operadoras de forma rentável e com qualidade de serviço.

Uma intenção é normalmente definida em linguagem natural. Sendo assim, para que sejam usadas como entrada em sistemas de gerenciamento baseados em intenção, é necessário que sejam processadas para extração inteligente de fatos e indicadores. Somente após esse procedimento, as ações necessárias para atingir os objetivos de gerenciamento são inferidas. O Processamento de Linguagem Natural (PLN), também conhecido como linguística computacional, consolida-se como um campo de pesquisa que envolve modelos e processos computacionais para a solução de problemas práticos de compreensão e manipulação de linguagens humanas [Otter et al., 2020]. Independentemente de sua forma de manifestação, textual ou fala, a linguagem natural é entendida como qualquer forma de comunicação diária entre humanos. Tal definição exclui linguagens de programação e notações matemáticas, consideradas linguagens artificiais. As linguagens naturais estão em constante mudança, dificultando o estabelecimento de regras explícitas para computadores [Otter et al., 2020].

Expressar intenções diretamente em linguagem natural possibilita abstrair as interfaces de gerenciamento de diferentes equipamentos. Assim, é possível reduzir a probabilidade de erros humanos ao dividir manualmente as políticas em comandos de configuração de equipamentos. Contudo, a linguagem natural é sujeita a ambiguidade e, assim, dificulta o sistema capturar a intenção do operador de maneira inequívoca e precisa. Os sistemas de gerenciamento baseados em intenção não garantem a sustentabilidade da rede, pois não contemplam todas as possíveis situações que possam surgir. Como contraponto, foram propostos sistemas de gestão do conhecimento que facilitam o processo de tomada de decisão [Leivadeas e Falkner, 2022].

A arquitetura O-RAN, em particular, define que uma intenção da RAN é uma expressão de alto nível que define objetivos operacionais ou de negócios a serem alcançados pela rede de acesso via rádio, permitindo que um operador especifique os acordos de nível de serviço desejados para a RAN cumprir para todos ou para uma classe de usuários em um dada área em um período de tempo [O-RAN Working Group 2, 2023]. Para efeito de comparação, a O-RAN define que uma política é um conjunto de regras que governa as escolhas de comportamento de um sistema.

O serviço do gerenciamento de rede baseado em intenção é de responsabilidade do Non-RT RIC, que deve permitir a injeção de intenções por fontes externas, como ilustra a Figura 1.6. Por isso, a definição do formato das intenções da RAN está fora do escopo da especificação da O-RAN. Uma intenção recebida é automaticamente analisada pelo Non-RT RIC para extrair os objetivos de alto nível contidos na intenção. Baseado nesses objetivos, em eventos e em contadores fornecidos pela interface O1, o Non-RT RIC determina políticas e o conjunto de rApps que devem ser implantadas e executadas para satisfazer tais políticas. Em seguida, o Non-RT RIC usa o serviço A1-P, citado na Seção 1.2.2, para fornecer as políticas para o Near-RT RIC através da interface A1. Por isso, tais políticas são chamadas de políticas A1. O objetivo das políticas A1 é ajustar o desempenho da RAN para que o objetivo geral expresso na intenção da RAN seja alcançado. As políticas A1 são políticas declarativas que contêm declarações sobre objetivos de política e recursos de política aplicáveis a equipamentos de usuários e células [O-RAN Working Group 2, 2023].

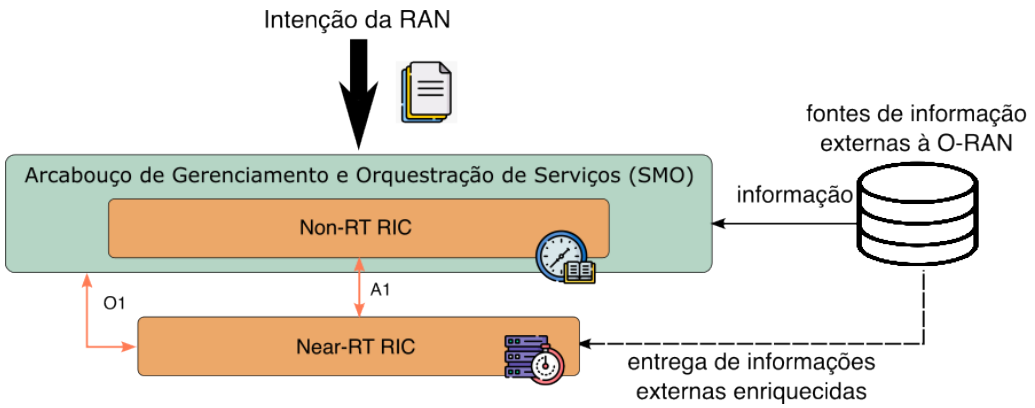

Figura 1.6. Elementos da arquitetura O-RAN envolvidos no gerenciamento baseado em intenção, adaptado [O-RAN Working Group 2, 2023]. A intenção é injetada por fontes externas no Non-RT RIC, que automaticamente extrai o objetivo geral da intenção e, baseado nesse objetivo e em eventos e contadores fornecidos pela interface O1, determina um conjunto de políticas que são enviadas ao Near-RT RIC pela interface A1. O Non-RT RIC também pode fornecer informações enriquecidas para auxiliar a aplicação das políticas no Near-RT RIC.

O Non-RT RIC usa a realimentação das políticas A1 e informações do estado da rede fornecidas pela interface O1 para avaliar continuamente o impacto das políticas A1 no cumprimento da intenção. A partir daí, o Non-RT RIC pode decidir por atualizar os objetivos definidos nas políticas A1 ou até mesmo remover políticas. Por exemplo, se o Non-RT RIC avaliar que os recursos de rede disponíveis em uma determinada área não são suficientes para atender o acordo de nível de serviço definido pela intenção para todos os usuários de uma fatia de rede, o Non-RT RIC pode decidir por, temporariamente, alterar os níveis de qualidade de serviço de alguns usuários pertencentes à mesma fatia. Para isso, alteraria o conjunto de políticas A1. O Non-RT RIC também pode fornecer informações enriquecidas para auxiliar a aplicação das políticas no Near-RT RIC através da interface A1. O Non-RT RIC é parte do SMO que é detalhado na próxima seção.

1.3.2. Arcabouço de gerenciamento e orquestração de serviços (SMO)

A arquitetura de Operação e Gerenciamento (*Operation And Management* - OAM) da O-RAN identifica serviços de gerenciamento, funções gerenciadas e elementos gerenciados suportados pela O-RAN, incluindo a interoperabilidade entre gerenciamento e orquestração de serviços e outros componentes da O-RAN, tal como o gerenciamento da infraestrutura. A arquitetura identifica as interfaces entre o SMO e os elementos gerenciados (*Managed Elements* – ME) para diferentes modelos e exemplos de implantação.

A arquitetura O-RAN OAM de referência é desenvolvida pela OSC[6] e detém os seguintes requisitos [O-RAN Working Group 10, 2023]. A arquitetura deve suportar a interação entre o SMO e a O-Cloud através da interface O2 para executar a orquestração de recursos virtualizados. Para tanto, o SMO deve consumir o serviço de gerenciamento de provisionamento exposto por cada elemento gerenciado O-RAN, implementado como uma PNF ou uma VNF por meio da interface O1. A arquitetura deve oferecer suporte à criação, modificação e encerramento de VNFs em uma rede O-RAN através do SMO e deve suportar registro e inventário de VNFs e PNFs, assim como suportar a configuração de VNFs e PNFs. A arquitetura O-RAN OAM deve suportar também o gerenciamento de dados de desempenho, tais como coleta, armazenamento, consulta e relatórios estatísticos de dados dos componentes O-RAN. A arquitetura O-RAN OAM deve oferecer suporte ao gerenciamento hierárquico e híbrido dos componentes O-DU e O-RU [O-RAN Working Group 4, 2023]. A arquitetura e as interfaces O-RAN OAM devem suportar o fatiamento de rede, em que uma instância da função gerenciada O-RAN pode ser associada a uma ou mais fatias. A arquitetura O-RAN OAM deve suportar a interface O1 para todos os elementos gerenciados, com exceção da O-RU que suporta a interface *Open Fronthaul M-Plane*. A arquitetura O-RAN OAM deve propiciar que o SMO seja capaz de descobrir os recursos de gerenciamento relacionados a falha, configuração, contabilização, desempenho e segurança (FCAPS) da função de rede O-RAN na qual há a terminação da interface O1. O SMO deve ainda descobrir os recursos de gerenciamento relacionados a FCAPS da função de rede O-RAN em que está a terminação a interface *Open Fronthaul M-Plane* e da infraestrutura O-Cloud.

O SMO, mostrado na Figura 1.7, supervisiona o gerenciamento do ciclo de vida das funções de rede e da nuvem (O-Cloud). Na arquitetura O-RAN OAM, o lado das funções gerenciadas (*Managed Functions*) da rede de acesso de rádio inclui Near-RT RIC, O-CU-CP, O-CU-UP, O-DU e O-RU, enquanto o lado do gerenciamento é composto pelo SMO, o qual engloba o Non-RT RIC. No ambiente NFV, os elementos de rede O-RAN também podem ser implementados de forma virtualizada e, portanto, incluem uma camada adicional de infraestrutura baseada em O-Cloud. As especificações da O-RAN definem que um arcabouço SMO inclui: (i) um ambiente de projeto para desenvolvimento rápido de aplicativos; (ii) uma plataforma comum de coleta de dados para gerenciamento da RAN; (iii) o suporte para licenciamento, controle de acesso e gerenciamento do ciclo de vida de funções de inteligência artificial, juntamente com as interfaces *northbound* herdadas; (iv) as funções de operação e suporte existentes, como orquestração de serviços, inventário, topologia e controle de políticas; e (v) a interface R1 para permitir portabilidade e gerenciamento do ciclo de vida de rApps. O SMO inclui um controlador

[6]Disponível em https://o-ran-sc.org/.

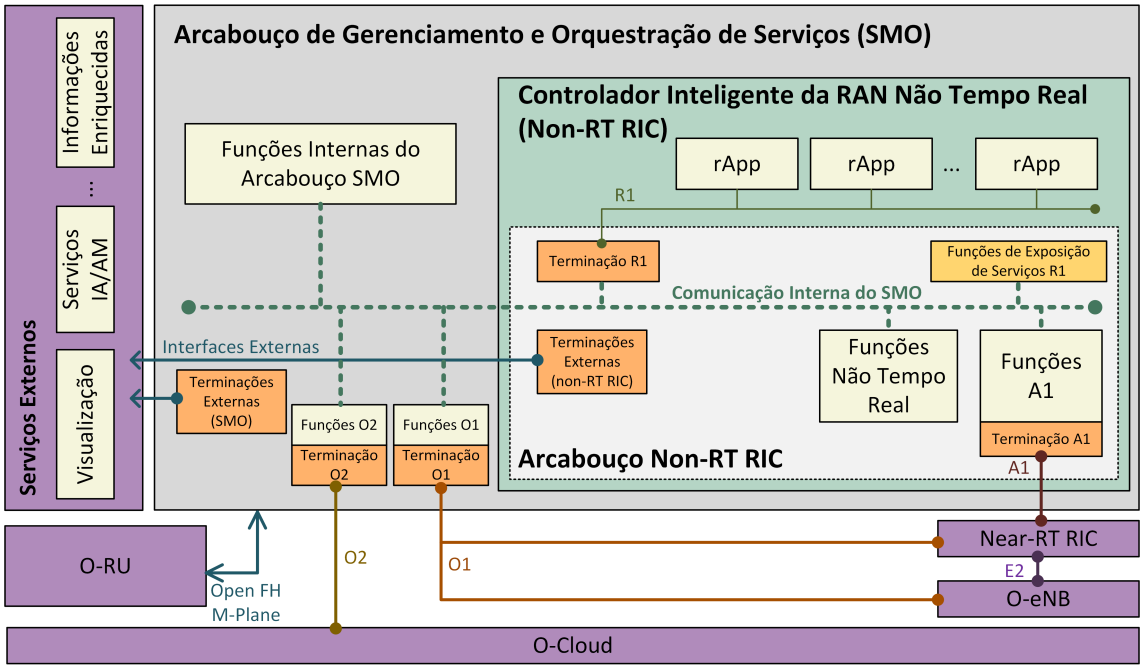

Figura 1.7. Arcabouço de Gerenciamento e Orquestração de Serviços (*Service Management and Orchestration* - SMO). A interface O1 ocorre entre o SMO e os elementos gerenciáveis. A interface O2 permite que o SMO exerça o gerenciamento de recursos na nuvem O-Cloud. A interface R1 permite que as rApps se comuniquem com o arcabouço do Non-RT RIC. A interface A1 permite o Non-RT RIC fornecer informações enriquecidas ao Near-RT RIC para a otimização da RAN. A interface Open Fronthaul M-Plane é alternativa à O1 para a comunicação SMO e O-RU.

inteligente de rádio não tempo-real (Non-RT RIC) e define interfaces entre o arcabouço SMO e as funções de rede na RAN (A1 e O1) e entre o SMO e a O-Cloud (O2). As interfaces permitem que o SMO gerencie redes O-RAN de vários fornecedores. O SMO possui ainda as interfaces *Open FrontHaul* e R1, que permitem portabilidade entre fornecedores. A interface R1 foi projetada para oferecer suporte à portabilidade de rApps de vários fornecedores. A interface é uma coleção de serviços, incluindo serviços de registro e descoberta de outros serviços, serviços de autenticação e autorização, serviços de *workflow* de aprendizado de máquina e serviços relacionados às interfaces A1, O1 e O2. O arcabouço SMO suporta também a interface Open FrontHaul M-Plane baseada em NETCONF/YANG como uma alternativa à interface O1 para suportar integração de unidades de rádio de vários fornecedores. O Open FrontHaul M-plane oferece suporte aos recursos de gerenciamento, incluindo inicialização, gerenciamento de *software*, gerenciamento de configuração, gerenciamento de desempenho, gerenciamento de falhas e gerenciamento de arquivos.

O protocolo NETCONF é um protocolo de gerenciamento de rede bem estabelecido que permite que um sistema de gerenciamento de rede (*Network Management System* – NMS) forneça, modifique e exclua configurações de dispositivos de rede [Enns et al., 2011]. O protocolo emprega codificação de dados baseada em XML para os dados de configuração e mensagens. As operações NETCONF são realizadas como chamadas de procedimento remoto (*Remote Procedure Calls* – RPCs). O protocolo

NETCONF facilita a automação e a orquestração da rede, pois fornece uma interface consistente para diferentes tipos de dispositivos e fornecedores, reduzindo a complexidade e o custo do gerenciamento de rede. A manipulação de todos os dados de configuração de um dispositivo garante precisão e integridade. Ao passo que a alterações simultâneas em vários dispositivos com atomicidade e confiabilidade evita configurações parciais ou inconsistentes que podem causar problemas na rede. O NETCONF suporta configuração dinâmica e baseada em modelo, permitindo que a rede se adapte a requisitos e condições em constante mudança. Por sua vez, a YAML é uma linguagem de serialização de dados amigável ao ser humano, passível de uso por diferentes linguagens de programação. Portanto, a YAML é mais legível por humanos e mais fácil de entender do que outros formatos de dados, como o XML, já que tem uma sintaxe simples e compacta que usa recuo e dois pontos para indicar estrutura e pares chave-valor para representar os dados. A YAML é um superconjunto do JSON e, assim, também pode usar a sintaxe JSON. A YAML suporta vários tipos de dados, como escalares, listas, mapas, conjuntos e pares, e permite comentários e auto-referências.

A implementação de referência da OAM define que todos os MEs, incluindo o *near RT-RIC*, O-CU, O-DU e O-RU, implementam a interface O1 [O-RAN Working Group 1, 2023a]. A especificação da interface O1 define um servidor NETCONF para gerenciamento de configuração (*Configuration Management* - CM) e um cliente HTTP para gerenciamento de falhas (*Fault Management* – FM), gerenciamento de desempenho (*Performance Management* – PM), além de outros eventos em cada Provedor de Serviços de Gerenciamento (*Management Service Provider* - MnS-Provider) em execução nos elementos gerenciados. Cada MnS-Provider e cada ME implementa uma interface (TLS/)NETCONF para gerenciamento de configuração e consome mensagens TLS/HTTP-POST com um corpo JSON no formato de mensagem *Virtual Event Streaming* (VES). O VES é um coletor RESTful para processamento de mensagens JSON. O coletor verifica a origem e valida os eventos no esquema VES antes de distribuir aos tópicos. O método de assinatura/cancelamento do VES deve ser realizado via NETCONF, pois o VES não disponibiliza tal função. O MnS-Consumer usa a interface NETCONF para tal operação. A interface O2 permite o gerenciamento de infraestruturas O-Cloud e o gerenciamento do ciclo de vida de implantação de funções de rede O-RAN nativas da nuvem que executam na O-Cloud. A interface A1 permite que a função Non-RT RIC forneça orientação, gerenciamento de modelo de aprendizado de máquina e informações de enriquecimento para a função Near-RT RIC para a otimização da RAN. De forma simplificada, o SMO recebe as intenções de gerenciamento através de interfaces externas (interfaces gráficas ou API), processa através de funções internas do SMO ou através de rApps no Non-RT RIC e, então, as converte em políticas e informações de enriquecimento que são expressas através da interface A1 para o Near-RT RIC. O Near-RT RIC toma as ações de otimização da RAN em laços fechados de controle da ordem de 10ms a 1s, baseado nas políticas definidas pelo SMO/Non-RT RIC.

O Open FrontHaul da O-RAN é uma interface lógica, consistindo na divisão da camada inferior (*Lower-Layer Split* - LLS) em plano de controle (LLS-CP) e plano de usuário (LLS-UP), plano de sincronização e plano de gerenciamento (M-Plane). O Open FrontHaul O-RAN especifica uma nova interface de transporte cooperativo (*Cooperative Transport Interface* - CTI) que destina-se a apoiar a cooperação em tempo real e em

tempo não real entre o eNB/gNB e a rede de transporte baseada na alocação de recursos. Quando a rede de transporte (*fronthaul*) consiste em um sistema baseado em pacotes, interconectando vários O-DUs para vários O-RUs, o CTI é usado para identificar cada fluxo de transporte e acionar decisões de agendamento apropriadas pelos nós de transporte para que os requisitos de rede sejam atendidos [Garcia-Saavedra e Costa-Pérez, 2021].

O arcabouço SMO é capaz de fornecer suporte a redes não virtualizadas e a redes virtualizadas. Para elementos não virtualizados, o arcabouço SMO suporta a implantação de elementos de rede física nos recursos físicos dedicados de destino que atendam aos requisitos de cobertura do operador de rede, com gerenciamento por meio da interface O1. Para elementos de rede virtualizados, o SMO interage com a O-Cloud para executar o gerenciamento do ciclo de vida do elemento de rede por meio da interface O2. O SMO consome o serviço de gerenciamento de provisionamento através da interface O1 para gerenciar a configuração dos elementos de rede [O-RAN Working Group 10, 2023]. O SMO age com a O-Cloud para realizar a implantação e provisionamento dos elementos de rede O-RAN virtualizados, criando uma rede O-RAN para fornecer serviço aos consumidores.

Por sua vez, o Non-RT RIC age como o centro de gerenciamento inteligente localizado no SMO, determinando quais os dados de medição de desempenho são necessários e, então, interage com as funções do SMO para coletar dados de medição da rede para treinamento, inferência e análise de modelos de inteligência artificial ou de aprendizado de máquina. A partir dos modelos de aprendizado, operações de otimização são executadas para melhorar a experiência de serviço do usuário de ponta a ponta e o desempenho da rede. Para atender às necessidades de dados do Non-RT RIC, o SMO deve gerar *jobs* de gerenciamento de desempenho (*Performance Management* - PM) e executar as operações de controle do PM, além de suportar o consumo de dados de medição pelo Non-RT RIC.

O Non-RT RIC é integrado ao SMO e opera em uma escala de tempo maior que 1s [Gramaglia et al., 2022]. Seu principal objetivo é apoiar otimização da RAN inteligente, fornecendo orientação baseada em políticas, gerenciamento de modelo aprendizado de máquina e informações de enriquecimento para a função Near-RT RIC. O Non-RT RIC pode também executar a função gerenciamento de recursos de rádio inteligente em tempo não real. Em contraponto, o Near-RT RIC, situado fora do SMO, é uma função lógica que permite o controle e otimização quase em tempo real da RAN e seus recursos por meio de coleta de dados refinada e ações em interfaces abertas e com laços de controle na ordem de subsegundos. O Near-RT RIC hospeda um ou mais xApps, aplicativos projetados para coletarem informações quase em tempo real e fornecerem controle sobre a RAN. O controle é conduzido através das políticas e de informações enriquecidas fornecidos pelo Non-RT RIC. As rApps fornecem serviços de valor agregado para apoiar e executar otimização e operações de RAN.

1.3.3. Plataformas para desenvolvimento de SMO

Open Source Management and Orchestration (OSM)[7], *Open Network Automation Platform* (ONAP)[8] e *Open Network Management System* (OpenNMS)[9] são as principais plataformas de gerenciamento e orquestração de código aberto e disponíveis publicamente, que estão sendo integradas à arquitetura O-RAN. Entretanto, a ONAP e a OSM são as mais utilizadas atualmente e, portanto, mais detalhadas neste capítulo. Ambas são plataformas abrangentes que permitem automação e orquestração em redes virtualizadas e baseadas em *software* [Polese et al., 2023].

A plataforma Open Source MANO[10] (OSM) foca a orquestração de infraestruturas híbridas e hiperconvergentes, que consolidam todos os elementos de um centro de dados tradicional, sendo as infraestruturas compostas por contêineres e máquinas virtuais e diferentes tecnologias coexistentes. A Open Source MANO visa implantar uma única camada de orquestração e gerenciamento para a infraestrutura complexa das redes. Além disso, visa aplicações futuras no contexto de um sistema de integração e entrega contínuas (CI/CD) DevOps para a virtualização de funções de rede (*Network Function Virtualization* – NFV). A OSM suporta a interface com vários tipos de gerenciadores de infraestruturas virtualizadas (*Virtualized Infrastructure Manager* – VIM), tais como nuvens privadas usando OpenStack e nuvens públicas na Amazon Web Services (AWS) ou Microsoft Azure. A integração de diferentes VIMs permite que o consumo de recursos nesses diferentes ambientes seja transparente para o usuário através do OSM. O OSM também suporta o estabelecimento de sobreposição de conectividade interna e entre centros de dados explorando um sistema baseado em redes definidas por *software* (SDN). Vários controladores SDN são suportados, como ONOS, Juniper Contrail e Arista. O OSM suporta nativamente a interação com infraestruturas nativas da nuvem, tal como Kubernetes, e explora diferentes gerenciadores de aplicativos, tais como Helm[11] e Juju[12]. Assim, o OSM age como uma interface única de gerenciamento e orquestração da rede facilitando a implantação de políticas e estratégias de otimização de forma independente da realização das políticas sobre recursos físicos ou virtuais.

A Figura 1.8 exibe a arquitetura da plataforma OSM. A OSM é mantida pela ETSI e possui uma arquitetura leve e simples, com poucos módulos. O objetivo da plataforma OSM é o desenvolvimento de um orquestrador de serviços de rede fim-a-fim para serviços de telecomunicações. A plataforma OSM provê a Interface Norte Unificada (*Northbound Interface* – NBI), baseada na especificação ETSI GS NFV-SOL 005, que permite a configuração e o controle do ciclo de vida de serviços de rede e fatias de rede. Além disso, o módulo de Gerenciamento de Ciclo de Vida (*LifeCycle Management* - LCM) é responsável por gerenciar funções virtuais e fatias de rede, além de permitir a operação de controle de laço fechado em conjunto com o Módulo de Políticas (*Policy Module* – POL). A comunicação entre o LCM e o Orquestrador de Recursos (*Resource Orchestrator* –

[7]Disponível em https://osm.etsi.org/.
[8]Disponível em https://www.onap.org/.
[9]Disponível em https://www.opennms.com/.
[10]*Management ANd Orchestration (MANO).*
[11]Disponível em https://helm.sh/.
[12]Disponível em https://juju.is/.

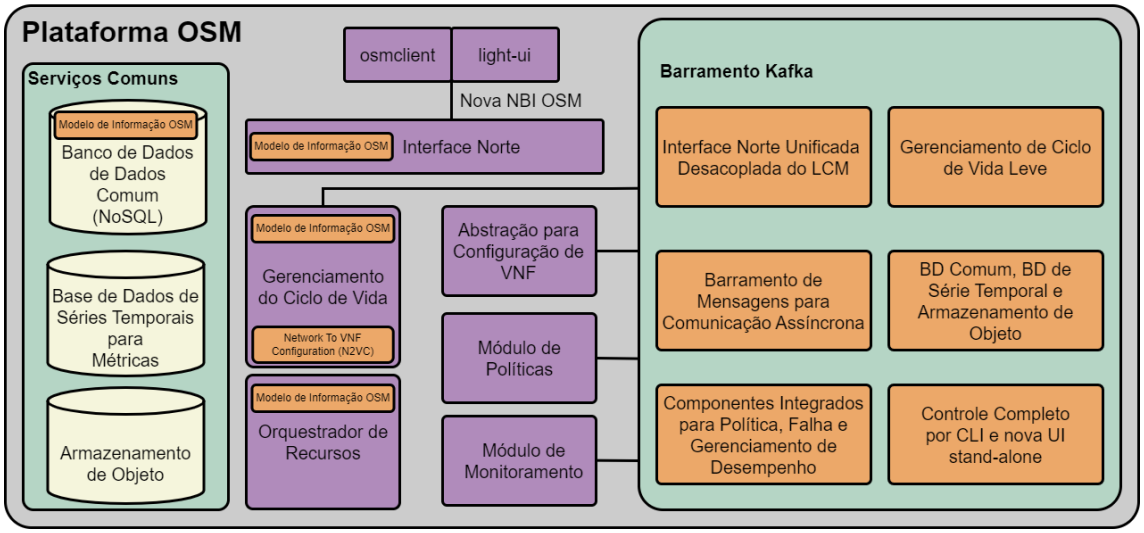

Figura 1.8. Principais módulos e componentes da plataforma OSM. A interface O1 é realizada pelos módulos LCM e POL. O barramento de comunicação interno é realizado pelo Apache Kafka. O módulo Interface Norte realiza a interface R1 da O-RAN.

RO) é realizada através do barramento de mensagens que utiliza o Apache Kafka[13]. Por fim, o Módulo de Monitoramento (*Monitoring Module* - MON) coleta métricas relacionadas ao desempenho do sistema e as armazena na Base de Dados de Séries Temporais (*Time-Series Data Base* – TSDB).

As funcionalidades da interface O1 são implementadas na plataforma OSM através dos módulos LCM e POL, utilizando o Apache Kafka como o barramento de mensagens. A base de dados para o armazenamento de métricas de séries temporais é o Prometheus[14], enquanto os demais dados são armazenados na base de dados SQL MongoDB[15]. O Grafana[16] provê a interface gráfica para visualização dos dados.

A plataforma ONAP permite a orquestração e a automação em tempo real, orientadas por políticas de funções de rede físicas, virtuais e nativas da nuvem. Assim, a plataforma habilita a automação rápida de novos serviços e o gerenciamento completo dos ciclos de vida correspondentes. A ONAP provê agilidade ao oferecer suporte a modelos de dados para implantação rápida de serviços e recursos e ao fornecer um conjunto comum de APIs REST *northbound*, abertas e interoperáveis, além de oferecer suporte a interfaces orientadas a modelos para as redes. A plataforma ONAP agrega recursos independentes de serviços para projeto, criação e gerenciamento do ciclo de vida. Além disso, combina a velocidade de abordagens DevOps/NetOps com os modelos e processos formais que as operadoras de telecomunicações exigem para introduzir novos serviços e tecnologias. A plataforma utiliza tecnologias nativas da nuvem, incluindo Kubernetes, para gerenciar e implantar rapidamente seus componentes.

[13]Disponível em https://kafka.apache.org/.
[14]Disponível em https://prometheus.io/.
[15]Disponível em https://www.mongodb.com/.
[16]Disponível em https://grafana.com/.

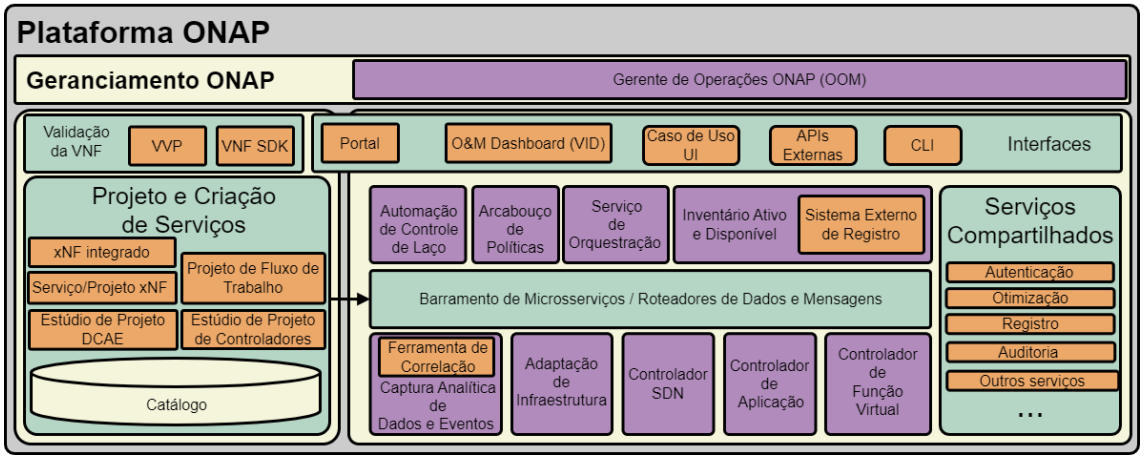

Figura 1.9. Principais módulos e componentes da plataforma ONAP. A plataforma é complexa e apresenta módulos especializados alinhados com as definições da O-RAN Software Community para plataformas de Gerenciamento e Orquestração de Serviços (SMO).

A Figura 1.9 exibe a arquitetura da plataforma ONAP. A plataforma ONAP provê integração nativa com projetos tais como Kubernetes, Akraino, Acumos e OpenDaylight, por serem todos projetos mantidos pela Linux Foundation. Acima da arquitetura, há a NBI para que os operadores configurem os módulos existentes na plataforma. O Gerente de Operações ONAP (*ONAP Operations Manager* – OOM) é o módulo responsável pela orquestração fim-a-fim, gerenciamento e monitoramento do ciclo de vida dos componentes existentes na plataforma ONAP. Esse módulo realiza um papel similar ao módulo LCM da plataforma OSM. Suas mensagens são enviadas através do barramento de microsserviços. O Inventário Ativo e Disponível (*Active and Available Inventory* – AAI) provê a visualização dos recursos do sistema e serviços em tempo real. Outros módulos como o Controlador SDN (*SND Controller* – SDNC), Automação de Controle de Laço (*Control Loop Automation* – CLAMP), Serviço de Orquestração (*Service Orchestration* – SO), Controlador de Aplicação (*Application Controller* – APPC) e Controlador de Função Virtual (*Virtual Function Controller* – VFC) permitem a automação do controle de laço fechado na plataforma.

O microsserviço *Common Controller Software Development Kit* (CCSDK) é responsável por implementar as funcionalidades de políticas de serviço e do adaptador da interface A1 na plataforma ONAP. O módulo SDNC também faz parte do adaptador da interface, sendo parte necessária para a comunicação do SMO com as APIs A1 do Near-RT RIC [Bonneau e Keeney, 2022]. As mensagens da interface são enviadas e recebidas por meio dos Roteadores de Dados e Mensagens (*Message & Data Routers* – DMaaP) e interfaces API REST para a configuração de políticas.

A comparação entre as arquiteturas OSM e ONAP permite observar que a plataforma ONAP é mais complexa, com diversos módulos e componentes que inexistem na plataforma OSM. Os módulos na plataforma ONAP realizam serviços específicos, enquanto os módulos da plataforma OSM atendem diversas necessidades. A plataforma OSM é leve e menos complexa do que a plataforma ONAP, porém seu desenvolvimento para oferecer funcionalidades de SMO ainda está em um estágio inicial. Por essa razão a

plataforma ONAP é adotada como a solução de SMO pela OSC [OSC, 2022], uma colaboração entre a Linux Foundation e a O-RAN Alliance [Polese et al., 2023]. Como pode ser destacado nas duas arquiteturas, atualmente a documentação da OSC não padroniza a integração de módulos de ambas as plataformas com relação à Interface O2 que deve ser oferecida por plataformas SMO.

A plataforma OpenNMS permite a visualização e o monitoramento em tempo real de redes locais e remotas. Para isso, a plataforma utiliza ferramentas similares as adotadas pela plataforma OSM, como Apache Kafka, ElasticSearch, Grafana e Kibana. Entretanto, a principal atividade foca o monitoramento das atividades de rede para detecção de intrusão, diferentemente das plataformas OSM e ONAP que oferecem serviços mais amplos relacionados a orquestração do ambiente. A plataforma possui integração com o OpenDaylight para o suporte a SDN e não possui suporte à NFV. Além disso, a plataforma fornece a Arquitetura para Habilitação de Aprendizado e Correlação (*Architecture for Learning Enabled Correlation* – ALEC), que integra funcionalidades de aprendizado de máquina ao ambiente, a fim de automatizar a análise de eventos de rede. Os dados coletados são armazenados em uma base de dados de série temporal Cassandra[17] e os dados do núcleo da plataforma são armazenados na base de dados PostgreSQL[18]. Atualmente a plataforma disponibiliza duas opções de instalação: Horizon, uma versão gratuita e com as funcionalidades mais novas, e a Meridian, a versão estável que possui suporte por meio de uma assinatura anual.

A Tabela 1.4 exibe uma comparação entre os componentes utilizados pelas principais plataformas de SMO para oferecer os serviços das interfaces [Skorupski e Brakle, 2020].

1.4. Propostas para Controle e Aprendizado na Arquitetura O-RAN

O controle e aprendizado em uma RAN pode atuar nas escalas de tempo Non-RT, Near-RT e tempo real (RT), como visto na Seção 1.2. A orquestração de serviços e recursos é executada por rApps, que são aplicações do Non-RT RIC. Essas rApps visam realizar ações que impactam uma grande quantidade de dispositivos e usuários, complementando e configurando as xApps [Polese et al., 2023]. As xApps, por sua vez, configuram O-CUs, O-DUs e O-RUs. Um exemplo disso é a alocação de fatias de rede. Uma fatia de rede consiste em um conjunto isolado de recursos para atender os requisitos de um determinado serviço [Popovski et al., 2018]. Esses recursos podem ser computacionais, como processamento, armazenamento e memória utilizados pelas O-CUs e O-DUs, além de comunicação, como alocação de PRBs (*Physical Resource Blocks*) nas RBSs. Um PRB é a menor unidade de alocação do enlace de rádio de uma rede celular, composto por subportadoras alocadas em uma determinada frequência e durante um intervalo de tempo [Chiarello et al., 2021]. A alocação de recursos para cada fatia é uma atribuição das xApps. Entretanto, as rApps, por terem uma visão global da rede, influenciam as decisões das xApps.

Devido aos mecanismos de isolamento inerentes, o fatiamento de redes permite que serviços com diferentes requisitos de qualidade de serviço (*Quality of Service* – QoS)

[17]Disponível em: https://cassandra.apache.org/
[18]Disponível em: https://www.postgresql.org/

Tabela 1.4. Mapeamento das interfaces do SMO em plataformas de gerenciamento e orquestração.

Componente SMO	Protocolo	ONAP	OSM	OpenNMS	Outras
Terminação O1 NetConf/YANG	Cliente NetConf/YANG	ODL / CCSDK / SDNC			OpenDaylight Apache Karaf
Terminação O1 VES	Servidor VES	Coletor VES Coletor HV-VES			
Painel O1	Aplicação Web	ODLUX			
Barramento de Mensagens		DMaaP	Apache Kafka	Apache Kafka	Apache Kafka
Base de Dados Persistente	SQL e Não-SQL	ElasticSearch	MongoDB ou SQL	ElasticSearch MariaDB	ElasticSearch MariaDB
Provisionamento de Serviços		SO			
Otimização		OOF			
Política		Política			
Análise de Dados		DCAE			Acumos
Inventário	REST	A&AI			ElasticSearch
Servidor de Certificação		AAF	Keystore		
Registro		Elastic	ElasticSearch		ElasticSearch Kibana
Painel de Registro	Aplicação Web	Kibana			

possam coexistir na RAN [Popovski et al., 2018]. A ITU (*International Telecommunication Union*) define três classes de serviços do 5G que podem, por exemplo, estabelecer os requisitos de uma determinada fatia, que são eMBB (*enhanced Mobile Broadband*), URLLC e mMTC (*massive Machine Type Communication*) [Popovski et al., 2018]. A classe eMBB suporta aplicações que necessitam de uma comunicação estável com alta vazão, como *streaming* de vídeo e jogos imersivos. A URLLC suporta aplicações que geram pacotes pequenos, mas que necessitam de uma latência de transmissão muito baixa e com alta confiabilidade, como veículos autônomos e cirurgia remota. Por fim, a mMTC suporta uma grande quantidade de dispositivos que enviam de forma esporádica pequenos pacotes à RBS, como em aplicações de Internet das Coisas [Motalleb et al., 2023]. O controle e aprendizado da arquitetura O-RAN pode atuar, por meio de rApps e xApps, para garantir os requisitos de QoS dessas diferentes classes de serviços.

Apesar de a arquitetura O-RAN ser preparada para suportar controle em todas as escalas de tempo, no momento da escrita deste capítulo ainda não há especificação da O-RAN para aplicações RT. Há, porém, iniciativas para padronizá-las por meio das dApps [D'Oro et al., 2022]. Esta seção tem como objetivo descrever propostas em cada uma das três escalas de tempo, exemplificando o uso da arquitetura O-RAN. Demais tendências e desafios de pesquisa são apresentados na Seção 1.5.

1.4.1. *Non Real Time* (Non-RT)

Diversos trabalhos da literatura propõem rApps e xApps para orquestrar serviços e recursos na arquitetura O-RAN. D'Oro *et al.* propõem a rApp OrchestRAN para orquestrar a inteligência em uma infraestrutura O-RAN [D'Oro et al., 2022]. A orquestração da

inteligência relaciona-se a tarefas como treinar e escolher os modelos de ML/AI utilizados e definir quais locais da infraestrutura instalá-los, de forma a atender as requisições da operadora. A Figura 1.10 mostra uma visão geral da OrchestRAN. Essa rApp recebe requisições da operadora, que podem ser relacionadas à implantação de funcionalidades como fatias de rede e escalonamento do enlace. A atuação para satisfazer as requisições pode ser diretamente nas O-CUs, O-DUs ou O-RUs[19], quando se trata de operações que influenciam o controle em tempo real, como modelos que realizam gerenciamento de feixe [Polese et al., 2021]. Em operações que influenciam o Near-RT RIC, como no gerenciamento das fatias de redes [Motalleb et al., 2023], a OrchestRAN interage com as xApps na interface A1. A OrchestRAN também pode interagir, por exemplo, para obter informações dos componentes da infraestrutura. Como mostrado na Figura 1.10, a OrchestRAN possui quatro módulos principais, descritos a seguir.

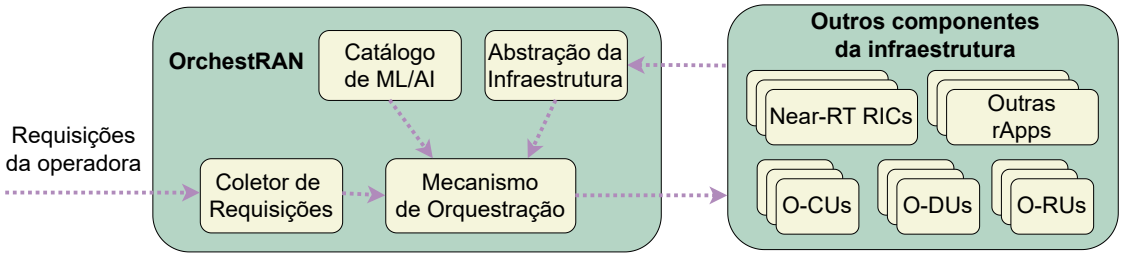

Figura 1.10. Visão geral da OrchestRAN. A rApp recebe requisições da operadora e utiliza seu mecanismo de orquestração para escolher quais modelos serão instanciados em quais localidades. Esses modelos são utilizados para implementar as funcionalidades solicitadas. Adaptada de [D'Oro et al., 2022].

- **Abstração da infraestrutura.** Este módulo coleta informações sobre a infraestrutura e cria uma abstração de alto nível a ser usada pelo Mecanismo de Orquestração. Para tal, cria-se uma árvore na qual a raiz é o Non-RT RIC que executa a OrchestRAN e os demais nós são os Near-RT RICs, as O-CUs, as O-DUs e as O-RUs. Os enlaces desse grafo representam as conexões entre os componentes e o grafo pode ser usado para modelar a alcançabilidade entre um componente e outro. Por exemplo, é possível verificar se informações do sinal de rádio de uma O-RU específica podem ser obtidas por um determinado nó que executa um Near-RT RIC.

- **Catálogo de ML/AI.** Neste módulo estão descritos os modelos de ML/AI pré-treinados, que realizam tarefas de inferência. Assim, para cada modelo, há informações sobre as funcionalidades as quais estão associados, como escalonamento de enlace e *handover*; quais entradas recebem, por exemplo, medidas de vazão e de tamanho de buffer; seus indicadores de desempenho, como acurácia em uma determinada funcionalidade; e aos recursos que necessitam para executar, como a quantidade de núcleos de CPU. A entrada necessária para um modelo pode ser obtida diretamente no nó no qual está instanciado ou é possível recebê-la remotamente utilizando as interfaces da O-RAN. Entretanto, o envio remoto de dados de entrada pode inserir uma sobrecarga na rede. Dessa forma, o Catálogo de ML/AI

[19] Apesar de haver propostas na literatura, como as dApps [D'Oro et al., 2022], que atuam em tempo real, a arquitetura O-RAN ainda não especifica esse tipo de aplicação. Assim, assume-se que a OrchestRAN conecta-se diretamente às O-CUs, O-DUs ou O-RUs, e não utilizando interfaces com dApps.

também possui, para cada combinação de modelo e funcionalidade, um indicador de adequação dessa combinação para um determinado nó. Por exemplo, modelos para gerenciar o feixe precisam de informações facilmente obtidas na O-RU, mas que podem sobrecarregar a rede caso sejam enviados para um Near-RT RIC.

- **Coletor de Requisições.** Este módulo recebe as requisições da operadora, que especificam quais funcionalidades devem ser instaladas em quais nós. Além disso, para cada combinação funcionalidade-nó, deve-se informar qual é o desempenho esperado, quais são os requisitos de tempo de resposta e qual é o conjunto de nós-fonte que podem fornecer os dados de entrada do modelo.
- **Mecanismo de Orquestração.** Este módulo resolve o problema de orquestração a partir das informações recebidas pelos demais módulos. Esse problema deve escolher, para cada combinação funcionalidade-nó de cada requisição, qual nó e qual modelo instanciado nesse nó serão utilizados para a combinação. É importante notar que o nó em que o modelo é instanciado pode ser diferente do nó em que a funcionalidade é oferecida. Além disso, um determinado modelo em um nó pode ser utilizado para oferecer funcionalidades em diferentes nós, tornando mais eficiente o uso de recursos. Cada requisição possui um valor associado. Assim, o problema de orquestração é modelado como um ILP (*Integer Linear Programming*) binário que possui o objetivo de maximizar o valor total oferecido pelo atendimento das requisições. Caso todas as requisições possuam o mesmo valor, esse objetivo pode ser entendido como maximizar o número de requisições atendidas. As restrições que o problema lida considera fatores como requisitos de desempenho e recursos disponíveis na infraestrutura.

O trabalho da OrchestRAN mostra que o problema de orquestração é NP-difícil e propõe algoritmos para solucioná-lo de forma eficiente. Além disso, implementa diversos mecanismos para instanciar contêineres para executar funcionalidades e modelos. A proposta é validada experimentalmente em um ambiente com 7 RBSs e 42 UEs na plataforma Colosseum [Bonati et al., 2021b].

Bonati *et al.* propõem o arcabouço NeutRAN para um cenário de infraestrutura hospedeira neutra [Bonati et al., 2023a]. Esse tipo de infraestrutura consiste em espectro e componentes da RAN oferecidos por um provedor para diversas operadoras. Assim, as operadoras alugam recursos do provedor. Esses recursos são então compartilhados por diversas operadoras, reduzindo o custo da infraestrutura. O compartilhamento é realizado por meio de virtualização de O-CUs, O-DUs e O-RUs, além de configuração de fatias de rede. O objetivo do NeutRAN é então automatizar o compartilhamento da infraestrutura entre múltiplos inquilinos, permitindo rápida implantação e gerenciamento de RANs que atendam suas necessidades. A Figura 1.11 apresenta uma visão geral do NeutRAN, que é composto por rApps, xApps, centros de dados de borda e outros componentes auxiliares. O provisionamento de uma RAN segue os seguintes passos identificados na figura:

1. **Submissão de requisições.** Nesta etapa, as operadoras inquilinas submetem suas requisições. Nessas requisições, especificam-se requisitos como as áreas geográficas que devem ser cobertas, a quantidade de espectro necessária, a duração da alocação e o nível de tolerância a falhas exigido.

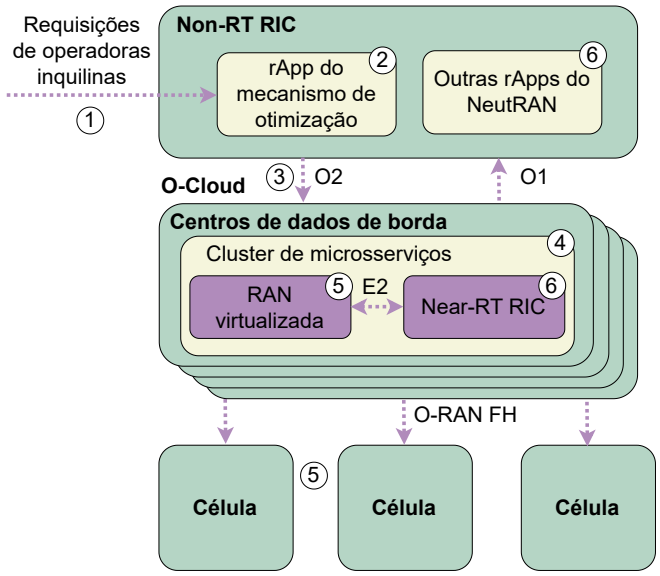

Figura 1.11. Visão Geral do NeutRAN. As operadoras inquilinas solicitam o provisionamento de uma RAN em uma infraestrutura hospedeira neutra. O Non-RT RIC, por sua vez, executa o mecanismo de otimização para provisionar a RAN virtualizada. A partir das decisões desse mecanismo, o Near-RT RIC realiza a configuração e inicialização de serviços para completar o provisionamento. Adaptada de [Bonati et al., 2023a].

2. **Execução do mecanismo de otimização.** A rApp do mecanismo de otimização do NeutRAN recebe as requisições e define as políticas de alocação a serem enviadas aos centros de dados de borda. A otimização executa o problema do hospedeiro neutro, modelado como um QCQP (*Quadratically Constrained Quadratic Program*). Esse problema utiliza os requisitos dos inquilinos e a disponibilidade da infraestrutura e do espectro para, da mesma forma que a OrchestRAN, atender o maior número possível de requisições, considerando o valor associado a cada uma. Como o problema é NP-difícil, o trabalho usa técnicas para reduzir o número de variáveis e transformar expressões quadráticas em lineares. As requisições recebidas pela rApp são armazenadas em um *buffer*. A cada intervalo de Δ segundos, o problema tenta atender a todas as requisições já armazenadas. As requisições não atendidas permanecem armazenadas para o próximo intervalo ou são removidas pelas suas operadoras correspondentes.

3. **Envio de políticas.** Após a finalização de um intervalo de otimização, a rApp envia as políticas para os centros de dados de borda pela interface O2. Essas políticas incluem, para cada requisição, especificações sobre a frequência e banda do espectro das O-RUs, recursos computacionais das O-CUs e O-DUs virtualizados, além de quais células são utilizadas.

4. **Inicialização de serviço na O-Cloud.** Nesta etapa, utiliza-se um orquestrador de contêiner, o OpenShift[20], para alocar componentes da O-Cloud de acordo com as políticas recebidas. Esses componentes são a RAN virtualizada (isto é, O-CUs e O-DUs), Near-RT RICs, a rede de núcleo e xApps específicas dos inquilinos.

[20]Disponível em https://www.redhat.com/en/technologies/cloud-computing/openshift

Para cada requisição instancia-se um *cluster* de microsserviços com os diversos componentes.

5. **Provisionamento completo.** Nesta etapa os serviços requisitados pelos inquilinos são completamente provisionados. Por exemplo, configuram-se fatias de redes para inquilinos diferentes que compartilham o mesmo espectro.

6. **Execução de serviços de monitoramento.** Para fornecer a automação do NeutRAN, é necessário provisionar serviços de monitoramento. Assim, esta etapa visa instanciar xApps e rApps para verificar frequentemente o estado dos *clusters* de microsserviços e das células, além de se recuperar automaticamente de falhas.

O NeutRAN é avaliado experimentalmente em um protótipo com 4 RBSs e 10 UEs de três diferentes inquilinos. Os resultados mostram, por exemplo, que é possível instanciar uma infraestrutura para um inquilino em aproximadamente dez segundos. Isso, por exemplo, facilita a criação de RANs virtualizada paras suprir demandas temporárias. Além disso, mostram-se ganhos de vazão para os usuários da rede devido ao uso eficiente da infraestrutura. Essa eficiência ocorre graças ao uso de virtualização, automação do provisionamento, otimização de recursos da RAN e compartilhamento do espectro.

Uma rApp do Non-RT RIC também pode atuar para enriquecer informações para o Near-RT RIC. Por exemplo, um Non-RT RIC pode ser o ponto central de uma estratégia de aprendizado federado (*Federated Learning* – FL). No FL, é possível treinar um modelo global a partir da interação com modelos distribuídos, sem a necessidade de coletar os dados locais [Neto et al., 2020, Ramos et al., 2021]. Os diversos trabalhos de FL em O-RAN diferem em relação aos objetivos e componentes desenvolvidos, mas, em geral, utilizam uma arquitetura similar à da Figura 1.12. Em uma infraestrutura O-RAN, um Non-RT RIC pode executar uma camada de federação que recebe vetores de parâmetros de diversos modelos locais, instanciados em Near-RT RICs distribuídos pela infraestrutura. A partir desses modelos, essa camada executa uma estratégia de federação, como realizar a média dos parâmetros recebidos, gerando um modelo global. O Non-RT RIC pode, então, estar em um servidor na nuvem central, enquanto os Near-RT RICs estão em nuvens regionais [Bonati et al., 2021a]. Cada Near-RT RIC pode ser responsável pelo controle da RAN de uma determinada localidade por meio da interação com O-CUs ou O-DUs pela interface E2 [Singh e Khoa Nguyen, 2022]. A camada de federação, após aplicar a estratégia de federação com os vetores recebidos, envia o modelo global para os Near-RT RICs via interface A1. Os Near-RT RICs, por sua vez, utilizam o vetor de parâmetros do modelo global como seus pesos iniciais de uma nova etapa de treino, que podem utilizar novos dados de entrada vindos da interface E2. Esse treino gera um novo vetor de parâmetros em cada Near-RT RIC, que pode ser enviado novamente para o Non-RT RIC pela interface A1. Esse processo repete-se até atingir a convergência do modelo centralizado. O FL permite então que um modelo seja treinado com informações de diversos nós locais sem que esses enviem seus dados brutos para o ponto central. Assim, Near-RT RICs de diferentes provedores podem colaborar sem violação de privacidade. Além disso, o envio apenas de vetores de parâmetros diminui o tráfego na rede em comparação ao envio de dados brutos, podendo acelerar o processo de convergência e diminuir a sobrecarga de controle.

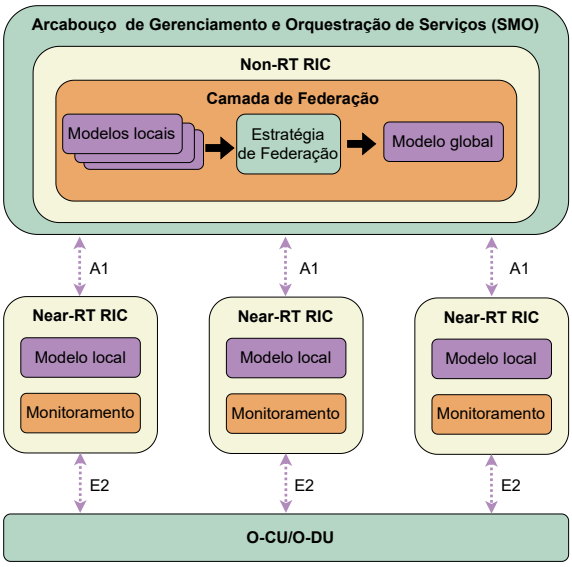

Figura 1.12. Exemplo de arquitetura de FL. Os Near-RT RICs treinam modelos locais e enviam seus parâmetros ao Non-RT RIC. A camada de federação do Non-RT RIC executa uma estratégia de federação, agregando os parâmetros recebidos e enviando aos Near-RT RICs. A partir desses parâmetros, os Near-RT RICs treinam um novo modelo local.Adaptada de [Rezazadeh et al., 2023].

Um exemplo de uso de FL é o trabalho de Singh e Nguyen, que visa treinar modelos para orquestrar recursos das fatias de rede [Singh e Khoa Nguyen, 2022]. No exemplo apresentado no trabalho, utiliza-se um modelo para predizer o tráfego nas fatias de rede, podendo ser usado para dimensionar os recursos para cada uma e definir políticas apropriadas. No cenário considerado por Singh e Nguyen, os O-CU-CPs enviam, por meio da interface E2, dados de telemetria das fatias para seus Near-RT RICs associados. Cada Near-RT RIC está em uma infraestrutura de borda e treina localmente um modelo com base nesses dados. Para realizar a estratégia de FL, os Near-RT RICs interagem com o Non-RT RIC utilizando a interface A1. Como uma iteração entre Near-RT RICs e o Non-RT RIC consome recursos de processamento e rede, é necessário escolher uma fração dos possíveis nós que participarão do treinamento. Assim, o trabalho de Singh e Nguyen lida com o desafio de escolher quais Near-RT RICs participarão de uma iteração de treinamento. Para tal, propõem o O-RANFed que, além de escolher os nós que participarão de uma iteração, aloca os recursos de rede e processamento necessários para o treinamento. A escolha de nós e alocação de recursos são realizadas por problemas de otimização formulados e resolvidos no trabalho. Os autores também propõem, em outro trabalho, o MCORANFed (*Momentum Compressed* O-RANFed) [Singh e Nguyen, 2022]. A proposta é baseada no O-RANFed, mas realiza compressão de dados para tornar mais eficiente o envio de parâmetros dos Near-RT RICs para o Non-RT RIC. O objetivo nesse caso é reduzir o tempo de treinamento sem a necessidade de alocar mais recursos de rede.

Rezazadeh *et al.* utilizam uma estratégia de aprendizado por reforço profundo federado (*Federated Deep Reinforcement Learning* – FDRL) para escolher a alocação de PRBs em cada uma das RBSs utilizadas por uma fatia de rede [Rezazadeh et al., 2023]. Uma estratégia centralizada, localizada no Non-RT RIC, pode gerar alto tráfego na infraestrutura (por exemplo, nas interfaces A1), pois exige uma visão completa da rede. Além disso, o cálculo centralizado pode levar a problemas de escalabilidade do algoritmo de

alocação. Assim, para uma determinada fatia, Rezazadeh *et al.* propõem o uso de um agente associado a cada uma das RBSs. Em uma fatia, cada agente executa no Near-RT RIC relacionado à sua RBS e toma decisões locais. Cada fatia possui a própria camada de federação e agentes que colaboram por meio dessa camada, instanciada no Non-RT RIC. Assim, não há comunicação entre os agentes de diferentes fatias nem entre suas camadas de federação. Dentro de uma fatia, o objetivo é escolher a alocação de PRBs em cada RBS de forma a atender os requisitos de vazão e de latência. A vazão é considerada como a demanda agregada de todos os usuários da fatia. A latência é o tempo médio que o tráfego de uma fatia precisa esperar em uma fila antes de ser atribuído a um PRB. Essa espera é necessária visto que o espectro é compartilhado por diferentes fatias em uma RBS.

Considerando o uso de uma arquitetura como a da Figura 1.12 em cada fatia, cada Near-RT RIC da fatia possui um agente que colabora com os outros agentes por meio de uma camada de federação localizada no Non-RT RIC. O agente executa uma estratégia de aprendizado por reforço para escolher, no seu RBS, o número de PRBs alocados em um determinado instante para a fatia. No aprendizado por reforço, um agente realiza ações em um ambiente que levam a mudanças de estado. Essas ações recebem como retorno recompensas ou punições [Santos Filho et al., 2020]. Na proposta de Rezazadeh *et al.*, a ação de uma agente consiste em quantos PRBs serão alocados para a fatia para o PRB associado no intervalo de tempo de decisão [Rezazadeh et al., 2023]. Os estados são as medidas realizadas na RBS nesse intervalo, que consiste em uma tupla com a relação sinal-ruído (*Signal-to-Noise Ratio* – SNR) média observada pelos usuários da fatia, o volume de tráfego e a latência desses usuários. A recompensa utiliza esses estados para punir ações que levem a uma subutilização ou sobrecarga do enlace de rádio. Dessa forma, tenta-se calcular a correta alocação de PRBs necessária para os usuários da fatia.

Os algoritmos de aprendizado por reforço estimam a recompensa de uma ação a partir de um estado e escolhem as ações a serem tomadas de forma a maximizar a recompensa ao longo do tempo [Filho et al., 2022]. Diversos algoritmos podem ser utilizados para estimar essa recompensa. Rezazadeh *et al.* utilizam o mecanismo *Double Deep Q-Network* (DDQN), que possui duas redes neurais para realizar as estimativas [Van Hasselt et al., 2016]. Cada agente possui suas próprias redes neurais. Os parâmetros dessas redes são periodicamente enviados para a camada de federação pela interface A1. Essa camada, por sua vez, utiliza uma estratégia de federação, como realizar a média dos parâmetros recebidos, para gerar o modelo global. Em seguida, os agentes são atualizados com esse modelo. Entretanto, alguns agentes podem possuir demandas de tráfego e padrões de mobilidade de usuários muito diferentes entre si, não devendo colaborar na estratégia de FDRL. Assim, o trabalho também propõe um algoritmo de clusterização para escolher dinamicamente quais grupos de agentes irão colaborar entre si pela camada de federação. Esse algoritmo executa no Non-RT RIC que, utilizando a camada de federação, calcula um modelo global para cada *cluster* da fatia de rede. O trabalho é validado por meio de simulações que mostram que uma estratégia que não considera a clusterização, isto é, usando o modelo de todos os agentes da fatia, possui problemas de convergência. Logo, dada a heterogeneidade das demandas das RBS, não há um modelo único que satisfaça todos os agentes da fatia. Assim, mostra-se que a clusterização torna os agentes de um mesmo *cluster* mais especializados em um determinado padrão de tráfego, facilitando a convergência dos modelos.

1.4.2. *Near Real Time* (Near-RT)

O controle da RAN na escala de tempo "próxima ao tempo real" (Near-RT) diz respeito a operações que devem ocorrer na ordem de dezenas de milissegundos até 1 segundo. As operações acima de 1 segundo são consideradas Non-RT. O controle nesta escala de tempo é implementado por xApps do Near-RT RIC que interage com dois componentes das gNBs, a O-CU e a O-DU. Um controlador Near-RT pode estar associado com diversas gNBs. Assim, as decisões tomadas podem afetar milhares de usuários. Diferentes estratégias de aprendizado de máquina são propostas na literatura para a inteligência do Near-RT RIC, tais como redes neurais profundas e redes neurais de grafos (*Graph Neural Networks* – GNNs).

Bonati *et al.* [Bonati et al., 2021a] demonstram o funcionamento do controle Near-RT realizando experimentos no *testbed* Colosseum, uma implementação experimental que contém diversos elementos da O-RAN para emulação de cenários próximos aos reais. O controle é realizado através da implementação de xApps executando em um Near-RT RIC. O objetivo do controlador implementado é a otimização das políticas de escalonamento utilizadas em diferentes fatias de rede. Agentes de aprendizado profundo (*Deep Reinforcement Learning* – DRL) executando nas xApps são responsáveis por selecionar a melhor política de escalonamento para cada fatia de rede. Os agentes DRL são treinados com dados sobre diferentes métricas de desempenho da rede, como vazão e taxa de erro de bit; informações sobre o estado dos elementos de rede, como tamanho de filas de transmissão, SINR, informação de qualidade do canal (*Channel Quality Indicator* – CQI), e estratégias de alocação de recursos (fatiamento e escalonamento) [Bonati et al., 2021a].

Na configuração experimental, há 4 RBSs e 40 UEs distribuídos em um cenário urbano (Roma, Itália). As localizações das RBSs são extraídas a partir da localização de células em operação. Cada UE é associado de forma estática a uma fatia de rede, podendo requisitar três tipos de serviço: eMBB, URLLC ou mMTC. As RBSs proveem os serviços nas fatias de rede utilizando políticas de escalonamento, podendo ser estas *proportionally fair* (PF), *waterfilling* (WF) e *round robin* (RR). O número de PRBs alocados para cada fatia de rede também pode variar ao longo do tempo. Cada agente é responsável pelo controle de uma fatia de rede em uma RBS, sendo assim, 12 agentes DRL são executados como xApps no Near-RT RIC. Por meio da interface O-RAN E2, os agentes recebem métricas de desempenho relativas à fatia de rede sob seu controle. O agente utiliza uma rede neural com 5 camadas e 30 neurônios para determinar a melhor política de escalonamento (PF, WF ou RR) ao longo do tempo. Essa política é informada à RBS correspondente, através do envio de mensagens de controle utilizando a interface E2. A recompensa dos agentes depende do serviço considerado: agentes eMBB e mMTC são treinados para maximizar a vazão obtida pelos UEs, enquanto agentes URLLC são treinados para minimizar a latência experimentada, por exemplo alocando PRBs o mais rapidamente possível.

Orhan *et al.* propõem uma estratégia de gerenciamento de conexão inteligente para atribuir usuários a células considerando a vazão da rede, a cobertura da célula e o balanceamento de carga [Orhan et al., 2021]. O principal diferencial da proposta de acordo com os autores é que normalmente a estratégia de conexão e reconexão a células, quando o UE se movimenta, é realizada pelo próprio UE. Uma técnica bastante usada é o UE medir a potência de referência do sinal recebido (*Received Signal Reference Power* – RSRP).

À medida que o UE se movimenta, a RSRP da célula à qual está conectado diminui. Outra célula (RBS) vizinha é selecionada pelo UE de acordo com a RSRP observada. A ideia é que a escolha da próxima célula leve em consideração também a capacidade da infraestrutura de rede, não somente a visão local dos UEs que pode levar a RBSs sobrecarregadas. Assim, os autores propõem uma estratégia de gerenciamento de conexões que executa no Near-RT RIC ciente da carga da rede.

Existem propostas que utilizam GNNs [Capanema et al., 2022] e aprendizado por reforço para escolher a próxima célula. A modelagem é feita considerando-se uma abstração da O-RAN onde RBSs e UEs são nós do grafo e a qualidade dos enlaces sem-fio é representada pelos pesos dos enlaces do grafo. Para levar em conta a carga na infraestrutura de rede, etiquetas são associadas a nós, e enlaces refletem condições de carga, qualidade do canal, taxa média dos UEs, entre outros. Utilizando o aprendizado por reforço e GNNs são realizadas as decisões de *handover* dos UEs. Os autores propõem uma estratégia de Q-learning profundo, onde a função Q é aprendida a partir das instâncias de UEs e células implantadas e da recompensa obtida de cada configuração de rede. O objetivo nesse modelo é obter a melhor função Q, capturada através da GNN. A cada passo da busca pela solução ótima, o estado atual corresponde ao grafo de conexão entre UEs e células atual, a ação tomada é conectar ou desconectar um UE de uma célula, e a recompensa é definida como a função de utilização da rede após a tomada desta ação. A proposta é avaliada através de simulações, com redes entre 3 e 9 células, e 20 a 60 UEs, demonstrando o ganho em termos de vazão, balanceamento de carga e cobertura comparados à utilização de uma estratégia gulosa onde a tomada de decisão de conexão é realizada de forma completamente distribuída pelos UEs.

Como visto na Seção 1.4.1, o Non-RT RIC pode ser utilizado para aplicar estratégias de FL em modelos dos Near-RT RICs. De forma similar, um Near-RT RIC pode ser utilizado para aplicar FL em modelos que atuam em tempo real. Um dos exemplos é o uso de aprendizado federado para realizar controle de acesso de UEs [Cao et al., 2022]. O controle de acesso, ou associação, de UEs consiste em selecionar a qual RBS um UE se conectará em um determinado instante. Tradicionalmente, um UE toma decisões de associação escolhendo a RBS que possui o RSS (*Received Signal Strength*) mais forte. Entretanto, essa estratégia pode levar a *handovers* frequentes, visto que o sinal pode ter alta variação. Além disso a estratégia pode levar à sobrecarga de uma RBS, visto que um alto RSS pode torná-la atrativa para muitos UEs [Cao et al., 2022]. Há diversas propostas na literatura para solucionar esse problema, que consideram a tomada de decisões pela própria RBS ou pela interação entre as diversas RBSs de uma região. Entretanto, Cao *et al.* [Cao et al., 2022] consideram que, no caso da O-RAN, a decisão deva ser tomada pelo UE. Isso justifica-se pois, dada a desagregação e flexibilidade da O-RAN, muitas RBSs podem estar disponíveis em uma região. Assim, a interação entre RBSs pode ser custosa em termos de sinalização e complexidade dos algoritmos de decisão.

O objetivo do trabalho de Cao *et al.* é propor um esquema para o UE escolher, em um determinado intervalo de tempo, qual RBS e quais PRBs serão utilizados [Cao et al., 2022]. Essa escolha deve ser feita de forma a maximizar sua vazão de *downlink* e minimizar a frequência de *handovers*. Para tal utiliza-se uma estratégia de FDRL. As ações da estratégia especificam, em cada intervalo de tempo, qual RBS deve ser usada e quais PRBs são alocados no *downlink*. A partir da definição dessas ações, o

UE solicita a alocação de PRBs à RBS selecionada. Essas ações levam a um estado que possui cinco componentes. Os dois primeiros indicam a utilização, em termos de UEs associados, das RBSs e dos PRBs. Há também dois componentes indicando o nível de RSS das RBSs e dos PRBs. O último componente indica a vazão do UE. A recompensa visa aumentar a vazão dos UEs e punir os *handovers*. O nível de punição depende de um parâmetro que pesa a importância da frequência baixa de *handovers* em relação à vazão.

Assim como o trabalho de Rezazadeh *et al.*, abordado na Seção 1.4.1, Cao *et al.* utilizam a estratégia DDQN de aprendizado por reforço, na qual duas redes neurais profundas são utilizadas por cada UE. O FDRL é então utilizado para que os UEs colaborem, via Near-RT RIC, na melhoria do modelo. O uso de aprendizado por reforço é justificável pois, em uma região controlada pelo mesmo Near-RT RIC, as decisões dos UEs interferem entre si. Apesar de cada UE tomar decisões locais, o FDRL permite que o sistema seja capaz de convergir para uma melhoria global da vazão dos UEs. Na estratégia de FDRL de [Cao et al., 2022], os UEs enviam parâmetros para o Near-RT RIC que, por sua vez, constrói um modelo global. Assim como diversas propostas de FL em O-RAN, Cao *et al.* propõem um mecanismo para escolher um subconjunto de UEs para enviar parâmetros em uma determinada rodada de aprendizado. Esse mecanismo toma decisões utilizando aprendizado por reforço por meio do *Upper Confidence Bound* (UCB) [Auer et al., 2002]. Em linhas gerais, da mesma forma que uma estratégia gulosa, o UCB visa escolher um subconjunto de UEs que possuam os melhores modelo locais. Entretanto, o UCB também explora eventualmente outros UEs, que não necessariamente possuam os melhores valores de vazão e de frequência de *handovers*. Apesar do pior desempenho em um determinado instante, os UEs contribuem para um melhor modelo global no longo prazo.

Outra contribuição de Cao *et al.* é que apenas partes dos parâmetros da DDQN de cada UE são usadas no aprendizado federado [Cao et al., 2022]. Isso permite que as amostras enviadas pelos UEs possuam independência entre si e reflitam o efeito de ações locais de cada UE. A proposta de Cao *et al.* é avaliada por meio de simulações e comparada com diferentes soluções, como o método tradicional de escolha por RSS e estratégias mais simples de aprendizado por reforço e otimização estocástica. Os resultados mostram ganhos de vazão agregada e redução da frequência de *handovers*.

1.4.3. *Real Time*

A especificação O-RAN ainda não possui definições a respeito de um laço de controle em tempo real [Polese et al., 2023]. Porém, há serviços e tarefas desempenhadas pelos elementos da arquitetura O-RAN que necessitam de respostas em tempo real. Os exemplos mais imediatos são tarefas relacionadas à camada MAC e à camada física, como alinhamento de feixe (*beamforming*), modulação e codificação. Naturalmente, é possível estabelecer no Non-RT RIC ou no Near-RT RIC políticas de alto nível capazes de influenciar no gerenciamento dessas tarefas. Porém, as tarefas em si não serão executadas nos laços de controle Non-RT ou Near RT, impossibilitando o gerenciamento de aspectos de mais baixo nível. Assim, encontram-se na literatura trabalhos que buscam oferecer soluções para alguns dos desafios relacionados ao controle em tempo real. Além de oferecer soluções para desafios individualmente, esses trabalhos podem ajudar a estabelecer direções para as especificações O-RAN.

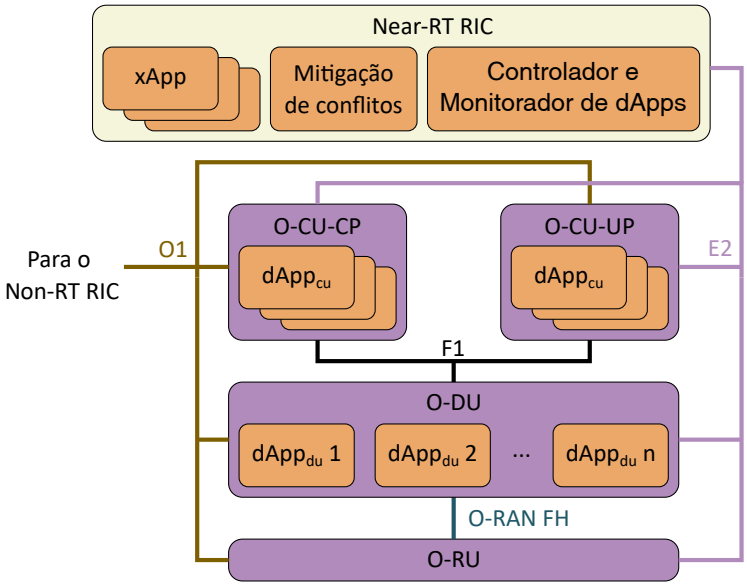

Figura 1.13. Arquitetura O-RAN com alterações para dar suporte às dApps. Os dApps são executados pela O-CU e pela O-DU. Adaptada de [D'Oro et al., 2022].

ChARM é um arcabouço que altera os parâmetros da O-DU e da O-RU de acordo com interferências identificadas nos sinais recebidos [Baldesi et al., 2022]. O arcabouço age em tempo real sobre os dados recebidos, mas é projetado para ser implantado como uma xApp, a fim de garantir compatibilidade com o padrão O-RAN existente. O ChARM é executado pelo Near-RT RIC e recebe amostras I/Q, decidindo se é necessário alterar algum dos parâmetros da comunicação. Se for necessário, as novas configurações são enviadas para a O-DU através da interface E2.

D'Oro *et al.* propõem que o controle em tempo real seja realizado através de dApps [D'Oro et al., 2022]. As dApps são análogas às rApps e xApps, mas são executados dentro das O-CUs e O-DUs. Assim, a informação que as dApps recebem diretamente das O-CUs, O-DUs e O-RUs possui menor latência do que a informação recebida pelas xApps ou rApps, possibilitando que as dApps efetuem inferência e controle mais rápido.

A utilização das dApps requer que O-CUs e O-DUs suportem a execução de aplicações na forma de contêineres. A Figura 1.13 ilustra a proposta de D'Oro *et al.* para mudanças na arquitetura O-RAN, a fim de dar suporte às dApps. A orquestração é realizada pelo Controlador e Monitorador de dApps (*dApps Controller and Monitor*), executado no Near-RT RIC. O Controlador e Monitorador verifica o desempenho das dApps, observando se as métricas estabelecidas pela operadora estão sendo alcançadas. A adição de dApps aumenta as chances de algum conflito como os conflitos mencionados na Seção 1.2, de forma que as medidas para mitigação de conflitos levem em consideração as dApps e suas possíveis interações diretas e indiretas com rApps e xApps.

O arcabouço *DeepBeam* utiliza uma rede neural convolucional para o gerenciamento de feixes de ondas milimétricas (*millimetre waves – mmWave*) [Polese et al., 2021]. O *DeepBeam* utiliza informações referentes ao ângulo de chegada de feixes e da transmissão de ondas para alimentar uma rede neural convolucional e tomar decisões sobre o controle de feixes. Tanto a coleta de informações quanto o controle dos feixes devem ser realizados em tempo real. É possível que o *DeepBeam* seja implementado como uma ou mais dApps, incluindo seu gerenciamento de feixes no laço de controle em tempo real.

Alguns tipos de recursos devem ser alocados e liberados em tempo real. Assim, os procedimentos para a alocação desses recursos também devem ocorrer em tempo real. Um exemplo possível é a alocação de PRBs para cada fatia da RAN. Motalleb *et al.* modelam o enlace de rádio para resolver um problema de otimização para alocar recursos a diferentes fatias da RAN [Motalleb et al., 2023]. Na proposta, cada fatia atende aos requisitos de um dos serviços eMBB, URLLC e mMTC. Outro exemplo de trabalho que foca na alocação de recursos em tempo real é o de Sharara *et al.* [Sharara et al., 2022]. Os autores utilizam aprendizado por reforço para alocar recursos computacionais para o processamento de quadros de usuários dos serviços de eMBB e URLLC. O principal desafio reside em lidar com um grande número de usuários, com condições restritas de atraso. Originalmente, Sharara *et al.* propõem a execução do algoritmo proposto em alguma infraestrutura de computação de borda. Porém, tanto o trabalho de Sharara *et al.* quanto o de Motaleb *et al.* podem ser implantados como dApps, fazendo parte de laços de controle em tempo real.

1.5. Tendências e Desafios de Pesquisa

Esta seção discute as tendência e desafios de pesquisa no gerenciamento e orquestração de serviços em O-RAN. Assim, destaca os principais desafios das soluções apresentadas na Seção 1.4 e complementa com outros trabalhos da literatura. Para tal, focam-se quatro aspectos: alocação de recursos, ferramentas de desenvolvimento e testes de xApps e rApps, gerenciamento de mobilidade e segurança.

1.5.1. Alocação de recursos

A alocação de recursos é um dos desafios mais importantes para garantir redução de custos e melhoria na qualidade da experiência dos usuários. Nesse aspecto, há desafios de pesquisa nas diversas escalas de tempo presentes em uma arquitetura O-RAN. Em geral, é necessário propor algoritmos escaláveis que possam orquestrar fatias de rede e serviços. Como visto nos trabalhos [D'Oro et al., 2022, Bonati et al., 2023a] esse problema é particularmente importante no Non-RT RIC, pois suas ações consideram informações de uma grande quantidade de elementos de rede e usuário. Assim, é necessário propor algoritmos de otimização eficientes para o Non-RT RIC, visando a orquestração de um número crescente de fatias de rede, com diversos requisitos. O FL também pode ser uma alternativa viável, possibilitando o treinamento distribuído de modelos para tomadas de decisões locais nos Near-RT RICs, utilizando uma camada de federação no Non-RT RIC. Entretanto, é necessário levar em conta o tráfego de controle gerado pelo treinamento distribuído. Dessa forma, como visto nos trabalhos abordados anteriormente [Singh e Khoa Nguyen, 2022, Singh e Nguyen, 2022, Rezazadeh et al., 2023], é necessário escolher, dentre os diversos Near-RT RICs, quais participarão de uma rodada de FL além de comprimir os dados enviados. Um Near-RT RIC também pode atuar como camada de federação para modelos de tempo real (RT), possuindo desafios semelhantes ao Non-RT, de escolha de nós para participar do treinamento e tráfego de controle gerado [Cao et al., 2022].

Outro problema relacionado à alocação de recursos é orquestração da inteligência, como a realizada pela OrchestRAN [D'Oro et al., 2022]. Essa orquestração possui como desafios lidar com as diferentes escalas de tempo dos modelos e escolher de onde coletar

os dados. Por exemplo, modelos que atuam em escalas de tempo maiores, como os executados por rApps no SMO, podem necessitar de dados dos O-RUs. O envio de dados dos O-RUs para o SMO pode ser custoso e adicionar latência na comunicação. Por outro lado, executar um modelo em um local mais próximo da O-RU, como em um Near-RT RIC, pode tirar sua visão global da infraestrutura. Outro desafio da orquestração da inteligência é garantir que diferentes modelos não tenham conflito entre si. Dada a flexibilidade e modularidade da arquitetura O-RAN, diferentes modelos podem coexistir. Entretanto, um mesmo parâmetro ou funcionalidade não pode ser controlado por mais de um modelo, de forma a evitar decisões conflitantes. A orquestração da inteligência também é abordada em [Martin-Perez et al., 2022]. Esse trabalho indica que, em muitas situações é importante escolher quais dados utilizados para treinar um modelo ao invés de usar todos as amostras disponíveis, dada a heterogeneidade dos componentes da infraestrutura.

Além das propostas apresentadas na Seção 1.4.1, outros trabalhos propõem mecanismos para alocação de recursos pelo Non-RT RIC ou ferramentas de apoio às suas decisões. Em [Saraiva Jr et al., 2022] propõe-se um classificador de tráfego para verificar o tráfego oriundo de UEs é eMBB, URLLC ou mMTC. Essa classificação pode ser utilizada, por exemplo, para dimensionar corretamente as fatias relacionadas a cada uma das classes. Em [Almeida et al., 2023], considera-se a desagregação de um Near-RT RIC, de forma que parte de seus componentes possam ser instalados em diferentes locais na infraestrutura: nó próprio nó E2, em uma infraestrutura de borda ou em uma nuvem de alta capacidade. Para isso, o trabalho propõe um mecanismo de orquestração, que executa no Non-RT RIC, para posicionar os componentes dos Near-RT RICs na infraestrutura. Esse mecanismo é construído com base em um problema de otimização que visa reduzir o custo da instanciação dos componentes, mas respeita requisitos de latência e capacidade de processamento, memória e armazenamento dos nós da infraestrutura. Outro trabalho relacionado ao Non-RT RIC é o SEM-O-RAN [Puligheddu et al., 2023], que realiza orquestração de fatias de rede baseada na semântica aplicação. Ao invés de estabelecer requisitos das fatias de forma pré-definida e com medidas genéricas, como as propostas da Seção 1.4.1, a orquestração do SEM-O-RAN visa requisitos específicos de tarefas de reconhecimento de objetos, como acurácia e latência. A ideia geral é usar características específicas da aplicação, como o nível de compressão de imagem tolerado, para realizar configurações de mais baixo nível da RAN, como configuração de recursos computacionais e PRBs. O SEM-O-RAN se baseia no fato de que, dependendo da aplicação e da disponibilidade de recursos, é possível aplicar compressão de imagens e ainda manter níveis aceitável de acurácia. Além disso, diferentes formas de combinar recursos de rede e computação podem levar a um mesmo desempenho da aplicação.

A alocação de recursos é também importante no Near-RT RIC. Em geral, os recursos são alocados considerado a mobilidade dos UEs, como visto na Seção 1.4.2 e descrito mais adiante na Seção 1.5.3. Como visto anteriormente no trabalho [Cao et al., 2022], os xApps podem atuar em tornar mais eficiente o controle de acesso de UEs. O trabalho de Tang *et al.* [Tang et al., 2023] é um outro exemplo, mas que considera um cenário no qual há uma presença maciça de UEs esparsas, ou seja, apenas uma pequena parte está ativa em um determinado momento. Esse cenário é típico de serviços mMTC e o Near-RT RIC deve alocar recursos a partir da detecção de quais UEs estão ativas, de forma a realizar um compartilhamento eficiente do meio. Para tal, Tang *et al.* propõe uma estratégia de

aprendizado por reforço para detectar UEs ativas, executada por xApps em uma infraestrutura O-RAN. Há também propostas que auxiliam a tomada de decisões de alocação de recursos no Near-RT RIC. Em [Rego et al., 2022] propõem-se xApps para realizar sensoriamento de espectro, que pode ser utilizado para fornecer informações para alocação e escalonamento dos recursos da O-RU.

Apesar de não fazer parte das especificações atuais, é esperado o suporte futuro à alocação de recursos em tempo real [Polese et al., 2023, D'Oro et al., 2022]. Uma iniciativa para facilitar a padronização é a de D'Oro *et al.*, que propõe dApps. Os dApps são estruturas similares aos xApps, porém gerenciados pelo nearRT RIC e executados no O-DU [D'Oro et al., 2022]. Os laços de controle em tempo real podem beneficiar tarefas como o alinhamento de feixes ou alocação de recursos de rádio [Polese et al., 2021, Motalleb et al., 2023]. A Seção 1.4.3 descreve propostas para o controle em tempo real.

1.5.2. Ferramentas de desenvolvimento e testes

Ferramentas de desenvolvimento e testes que suportem cenários de larga escala são importantes para que novas soluções para O-RAN sejam criadas e adotadas [Polese et al., 2022]. Para que sejam úteis, as plataformas de testes devem atender a uma gama de casos de uso, com diferentes requisitos [Khatib et al., 2023]. A *O-RAN Alliance* possui uma verificação de conformidade para Centros Abertos de Integração e Testes (*Open Testing and Integration Centres*), que certifica que um centro é capaz de executar testes seguindo os padrões O-RAN. Porém, até o momento de escrita deste texto, apenas 11 centros estão certificados[21]. Um exemplo de plataforma de testes é o Colosseum [Bonati et al., 2021b], utilizado em diversos experimentos para validar diferentes propostas [Baldesi et al., 2022, Bonati et al., 2021a, D'Oro et al., 2022, Polese et al., 2022]. O Colosseum possui 256 rádios definidos por *software*, capazes de modelar diferentes condições de sinal e de interferência. Outra plataforma é a POWDER [Johnson et al., 2022], que está em fase de implantação, mas já possui algumas funcionalidades para testes e experimentos com O-RAN. Ela possui 64 estações de rádio, entre outros equipamentos disponíveis. Além disso, é importante a disponibilização de plataformas de código aberto, facilitando e em algum nível a padronização o desenvolvimento e os testes de rApps e xApps. Do ponto de vista de testes, o ColO-RAN é um arcabouço de testes de larga escala para xApps, usando a infraestrutura de rádio do Colosseum. Em termos de desenvolvimento de aplicações, o OpenRAN Gym é um conjunto de ferramentas para o desenvolvimento de xApps com aprendizado de máquina [Bonati et al., 2023b]. O OpenRAN Gym possui um fluxo de coleta de dados, de aprendizado de máquina e de implantação em RANs.

1.5.3. Gerenciamento da mobilidade

As especificações da O-RAN estabelecem diversos casos de uso como gerenciamento da mobilidade, como o gerenciamento de troca de células (*handover*), alocação de recursos baseada em trajetórias, gerenciamento de feixes e outros [O-RAN Working Group 1, 2023b]. É esperado que os RICs armazenem informa-

[21] `https://www.o-ran.org/testing-integration`

ções sobre a troca de células, sobre a trajetória e sobre a velocidade de dispositivos, para inferir as melhores estratégias de gerenciamento de mobilidade. Zhang *et al.* propõem um método de mitigação de conflitos de xApps através de uma técnica batizada pelos autores de aprendizado em equipe (*team learning*). As simulações de Zhang *et al.* mostram que a velocidade dos usuários pode alterar a estratégia ótima de alocação de recursos, mesmo quando não há troca de células [Zhang et al., 2022]. O gerenciamento de mobilidade é um desafio fortemente relacionado à alocação de recursos, uma vez que a alocação ótima depende do padrão de mobilidade dos dispositivos. Assim, a alocação de recursos e o fatiamento da rede devem levar a mobilidade em consideração. Filali *et al.* desenvolvem um método baseado em aprendizado por reforço profundo que considera a velocidade dos dispositivos para alocar recursos de rádio [Filali et al., 2023]. Coronado *et al.* propõem o Roadrunner [Coronado et al., 2022], um arcabouço compatível com o O-RAN para a seleção de célula para *handover*. A estratégia utilizada pelo Roadrunner privilegia altas taxas de dados, ao contrário das estratégias tradicionais, que privilegiam a qualidade do sinal. O gerenciamento da mobilidade é especialmente importante para as aplicações veiculares [Arnaz et al., 2022]. O principal desafio relaciona-se com o desenvolvimento de aplicações de gerenciamento de procedimentos de *handover*, com previsão de mobilidade e disponibilidade de recursos.

1.5.4. Segurança

A desagregação promovida pela O-RAN aumenta a superfície de ataque, dada a existência de múltiplas interfaces e a existência de diversas aplicações de gerenciamento [Polese et al., 2023]. Os ataques podem ser direcionados à infraestrutura celular, à arquitetura aberta, à virtualização, ao aprendizado de máquina ou à arquitetura 5G na qual o O-RAN se insere [Mimran et al., 2022]. Adicionalmente, o desenvolvimento da O-RAN envolve muito mais partes interessadas, o que também aumenta a probabilidade de haver desafios de segurança [Liyanage et al., 2023]. Haas *et al.* defendem que para atingir a latência exigida pelas aplicações do 5G, é necessário que a O-RAN seja executado sobre uma plataforma de hardware confiável e elabora uma arquitetura para aumentar a confiabilidade do hardware. Outra proposta para aumentar a segurança da O-RAN é uma arquitetura para a execução da O-RAN em uma infraestrutura pouco confiável [Ramezanpour e Jagannath, 2022]. Na proposta, cada pedido de cada usuário é avaliado com auxílio de aprendizado de máquina, para definir autorizações e comportamentos suspeitos. A implantação plena da O-RAN deve aumentar o incentivo para ataques ao mesmo tempo em que aumenta as oportunidades para ataques. Assim sendo, as pesquisas em segurança devem reduzir a superfície de ataques, assim como aumentar o custo e reduzir o benefício de ataques de sucesso.

1.6. Considerações Finais

Este capítulo explorou a tendência de migração das redes de acesso via rádio, monolíticas e com arquitetura proprietária, para redes de acesso via rádio modulares com arquitetura aberta. A RAN aberta (Open RAN), promovida pela O-RAN Alliance, reproduz o movimento de desagregação do hardware e da função de rede através de tecnologias

como as redes definidas por software (*Software Defined Networks* – SDN) e a virtualização das funções de rede (*Network Function Virtualization* – NFV). A Open RAN permite a desagregação, virtualização e "softwarização" de componentes conectados através de interfaces abertas padronizadas. Para isso, as funcionalidades da estação rádio base são desagregadas em três unidades principais – unidade central, unidade distribuída e unidade de rádio – que são conectadas a controladores inteligentes através de interfaces abertas. Essa mudança arquitetural permite maior competitividade entre fornecedores, tendo em vista o potencial de interoperabilidade e programabilidade; e integração de inteligência no controle da rede de acesso, tendo em vista o uso de Controladores Inteligentes da RAN (*RAN Intelligent Controllers* - RICs).

As iniciativas para prover redes de acesso via rádio aberta são, então, cada vez mais presentes e permitem o desenvolvimento de controladores inteligentes em diferentes escalas de tempo. Este capítulo introduziu os diferentes tipos de RIC, Non-RT RIC, Near-RT RIC e RT RIC, sendo este último ainda não especificado. Assim, as oportunidades de pesquisa são diversas. Novos controles em tempo real para otimização e orquestração de funções da rede de acesso via rádio são tendências de pesquisa atuais, alinhados com mecanismos de aprendizado de máquina e inteligência artificial distribuídos e federados. O capítulo mostrou quatro tipos de desafio de pesquisa em O-RAN, alocação de recursos, ferramentas de desenvolvimento e testes de xApps e rApps, gerenciamento de mobilidade e segurança. A alocação de recursos é um dos desafios mais importantes para garantir redução de custos e melhoria na qualidade da experiência dos usuários. Apesar de não fazer parte das especificações atuais da O-RAN, é esperado o suporte futuro à alocação de recursos em tempo real para que, assim, seja possível a criação de laços de controle em tempo real para tarefas como o alinhamento de feixes ou alocação de recursos de rádio. Ferramentas de desenvolvimento e testes que suportem cenários de larga escala são importantes para que novas propostas para O-RAN sejam criadas e validadas. A disponibilização de plataformas de código aberto é importante para garantir a interoperabilidade e continuidade de desenvolvimento da RAN aberta. O gerenciamento da mobilidade é especialmente importante para as aplicações veiculares. O principal desafio relaciona-se com o desenvolvimento de aplicações de gerenciamento de procedimentos de *handover*, com previsão de mobilidade e disponibilidade de recursos. No caso da segurança, a desagregação aumenta a superfície de ataque dada a existência de múltiplas interfaces e existência de diversas aplicações de gerenciamento. Por fim, também é tendência de pesquisa o desenvolvimento de mecanismos de otimização da operação da rede que realizam instrumentação da rede para a geração de dados de desempenho. Os dados de desempenho da rede alimentam *workflows* de aprendizado de máquina e políticas de gerenciamento de alto nível focadas no KPIs das operadoras de rede ao invés de métricas estritas de gerenciamento e operação da rede.

Agradecimentos

Este capítulo foi realizado com recursos do CNPq, CAPES, FAPERJ, RNP, PR2/UFRJ e PGC/UFF.

A. Acrônimos

Os acrônimos utilizados neste texto estão organizados na Tabela 1.5.

Tabela 1.5. Acrônimos utilizados no capítulo.

Acrônimo	Descrição	Acrônimo	Descrição
3GPP	*3rd Generation Partnership Project*	O-RAN	*Open RAN*
A1-EI	Serviço de Enriquecimento de Informação A1	O-CU	*O-RAN Central Unit*
A1-ML	Serviço de Gerenciamento de Modelos de Aprendizado de Máquina	O-CU-CP	*O-CU Control Plane*
A1-P	Serviço de Gerenciamento de Políticas A1	O-CU-UP	*O-CU User Plane*
A1AP	*A1 interface Application Protocol*	O-DU	*O-RAN Distributed Unit*
AAI	*Active and Available Inventory*	O-RU	*O-RAN Radio Unit*
ALEC	*Architecture for Learning Enabled Correlation*	OAM	*Operation And Management*
API	*Application Programming Interface*	Open FH	*Open FrontHaul*
APPC	*Application Controller*	OSC	*O-RAN Software Community*
BBU	*BaseBand Unit*	OSM	*Open Source Management and Orchestration*
C-Plane	*Control Plane*	ONAP	*Open Network Automation Platform*
CCSDK	*Common Controller Software Development Kit*	OOM	*ONAP Operations Manager*
CLAMP	*Control Loop Automation*	OpenNMS	*Open Network Management System*
CM	*Configuration Management*	PDCP	*Packet Data Convergence Protocol*
CQI	*Channel Quality Indicator*	PF	*proportionally fair*
CRUD	*create, read, update, delete*	PLFS	*Physical Layer Frequency Signals*
CTI	*Cooperative Transport Interface*	PLN	Processamento de Linguagem Natural
DDQN	*Double Deep Q-network*	PM	*Performance Management*
DMaaP	*Message & Data Routers*	PNF	*Physical Network Function*
DMS	*Deployment Management Services*	POL	*Policy Module*
DRL	*Deep Reinforcement Learning*	PRB	*Physical Resource Blocks*
E2AP	*E2 Application Protocol*	PTP	*Precision Time Protocol*
E2SM	*E2 Service Model*	QCQP	*Quadratically Constrained Quadratic Program*
E2SM-CCC	E2SM *Cell Configuration and Control*	QoS	*Quality of Service*
E2SM-KPM	E2SM *Key Performance Metrics*	RAN	*Radio Access Network*
E2SM-NI	E2SM *Network Interface*	RBS	*Radio Base Station*
E2SM-RC	E2SM *RAN Control*	RIC	*RAN Intelligent Controller*
eCPRI	*evolved Common Public Radio Interface*	RLC	*Radio Link Control*
eMBB	*enhanced Mobile Broadband*	RNC	*Radio Network Controller*
eNB	*Evolved Node B*	RO	*Resource Orchestrator*
ETSI	*European Telecommunications Standards Institute*	RPC	*Remote Procedure Calls*
FCAPS	*Fault, Configuration, Accounting, Performance, Security*	RR	*Round Robin*
FDRL	*Federated Deep Reinforcement Learning*	RRC	*Radio Resource Control*
FL	*Federated Learning*	RRH	*Remote Radio Head*
FM	*Fault Management*	RRM	*Radio Resource Management*
gNB	*Next Generation Node B*	RSRP	*Received Signal Reference Power*
GNNs	*Graph Neural Networks*	RSS	*Received Signal Strength*
ILP	*Integer Linear Programming*	S-Plane	*Synchronization Plane*
IMS	*Infrastructure Management Services*	SCTP	*Stream Control Transmission Protocol*
ITU	*International Telecommunication Union*	SDAP	*Service Data Adaptation Protocol*
KPI	*Key Performance Indicator*	SDNC	*SND Controller*
LCM	*Life Cycle Management*	SLA	*Service Level Agreement*
LLS	*Lower Layer Split*	SMO	*Service Management and Orchestration*
M-Plane	*Management Plane*	SNR	*Signal-to-Noise Ratio*
MAC	*Medium Access Control*	SO	*Service Orchestration*
ME	*Managed Elements*	TB	*Transport Block*
mMTC	*massive Machine Type Communication*	TSDB	*Time-Series Data Base*
mmWave	*millimetre waves*	U-Plane	*User Plane*
MnS	*Management Services*	UE	*User Equipment*
MnS-Provider	*Management Services Provider*	UE-ID	*UE Identifier*
MON	*Monitoring Module*	UCB	*Upper Confidence Bound*
Near-RT RIC	*Near-Real-Time* RIC	URLLC	*Ultra-Reliable Low-Latency Communication*
NFV	*Network Function Virtualization*	VES	*Virtual Event Streaming*
NIB	*Network Information Base*	VFC	*Virtual Function Controller*
NBI	*Northbound Interface*	VIM	*Virtualized Infrastructure Manager*
NMS	*Network Management System*	VNF	*Virtual Network Function*
Non-RT RIC	*Non-Real-Time* RIC	VPN	*Virtual Private Network*
NR	*New Radio*	WF	*waterfilling*

Referências

[Almeida et al., 2023] Almeida, G. M., Bruno, G. Z., Huff, A., Hiltunen, M., Duarte Jr, E. P., Both, C. B. e Cardoso, K. V. (2023). RIC-O: Efficient Placement of a Disaggregated and Distributed RAN Intelligent Controller with Dynamic Clustering of Radio Nodes. *arXiv preprint arXiv:2301.02760*.

[Arnaz et al., 2022] Arnaz, A., Lipman, J., Abolhasan, M. e Hiltunen, M. (2022). Toward Integrating Intelligence and Programmability in Open Radio Access Networks: A Comprehensive Survey. *IEEE Access*, 10:67747–67770.

[Auer et al., 2002] Auer, P., Cesa-Bianchi, N. e Fischer, P. (2002). Finite-Time Analysis of the Multiarmed Bandit Problem. *Machine learning*, 47:235–256.

[Baldesi et al., 2022] Baldesi, L., Restuccia, F. e Melodia, T. (2022). ChARM: NextG Spectrum Sharing through Data-Driven Real-Time O-RAN Dynamic Control. Em *IEEE Conference on Computer Communications (INFOCOM)*, p. 240–249.

[Bonati et al., 2021a] Bonati, L., D'Oro, S., Polese, M., Basagni, S. e Melodia, T. (2021a). Intelligence and Learning in O-RAN for Data-Driven NextG Cellular Networks. *IEEE Communications Magazine*, 59(10):21–27.

[Bonati et al., 2021b] Bonati, L., Johari, P., Polese, M., D'Oro, S., Mohanti, S., Tehrani-Moayyed, M., Villa, D., Shrivastava, S., Tassie, C., Yoder, K. et al. (2021b). Colosseum: Large-Scale Wireless Experimentation through Hardware-in-the-Loop Network Emulation. Em *IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN)*, p. 105–113.

[Bonati et al., 2023a] Bonati, L., Polese, M., D'Oro, S., Basagni, S. e Melodia, T. (2023a). NeutRAN: An Open RAN Neutral Host Architecture for Zero-Touch RAN and Spectrum Sharing. *arXiv preprint arXiv:2301.07653*.

[Bonati et al., 2023b] Bonati, L., Polese, M., D'Oro, S., Basagni, S. e Melodia, T. (2023b). OpenRAN Gym: AI/ML development, data Collection, and Testing for O-RAN on PAWR Platforms. *Computer Networks*, 220:109502.

[Bonneau e Keeney, 2022] Bonneau, M. e Keeney, J. (2022). O-RAN A1 Policies in ONAP. Relatório técnico. Disponível em `https://wiki.onap.org/pages/viewpage.action?pageId=84672221`.

[Brik et al., 2022] Brik, B., Boutiba, K. e Ksentini, A. (2022). Deep Learning for B5G Open Radio Access Network: Evolution, Survey, Case Studies, and Challenges. *IEEE Open Journal of the Communications Society*, 3:228–250.

[Cao et al., 2022] Cao, Y., Lien, S.-Y., Liang, Y.-C., Chen, K.-C. e Shen, X. (2022). User Access Control in Open Radio Access Networks: A Federated Deep Reinforcement Learning Approach. *IEEE Transactions on Wireless Communications*, 21(6).

[Capanema et al., 2022] Capanema, C. G. S., Silva, F. A. e Loureiro, A. A. F. (2022). Redes Neurais de Grafos no Contexto das Cidades Inteligentes. Em *Minicursos do XL Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC)*.

[Checko et al., 2015] Checko, A., Christiansen, H. L., Yan, Y., Scolari, L., Kardaras, G., Berger, M. S. e Dittmann, L. (2015). Cloud RAN for mobile networks—a technology overview. *IEEE Communications Surveys & Tutorials*, 17(1):405–426.

[Chiarello et al., 2021] Chiarello, L., Baracca, P., Upadhya, K., Khosravirad, S. R. e Wild, T. (2021). Jamming Detection with Subcarrier Blanking for 5G and Beyond in Industry 4.0 Scenarios. Em *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, p. 758–764.

[Clemm et al., 2022] Clemm, A., Ciavaglia, L., Granville, L. Z. e Tantsura, J. (2022). Intent-Based Networking - Concepts and Definitions. RFC 9315.

[Clemm et al., 2020] Clemm, A., Faten Zhan, M. e Boutaba, R. (2020). Network Management 2030: Operations and Control of Network 2030 Services. *Journal of Network and Systems Management*, 28(4).

[Coronado et al., 2022] Coronado, E., Siddiqui, S. e Riggio, R. (2022). Roadrunner: O-RAN-based Cell Selection in Beyond 5G Networks. Em *IEEE/IFIP Network Operations and Management Symposium (NOMS)*, p. 1–7.

[D'Oro et al., 2022] D'Oro, S., Polese, M., Bonati, L., Cheng, H. e Melodia, T. (2022). dApps: Distributed Applications for Real-Time Inference and Control in O-RAN. *IEEE Communications Magazine*, 60(11):52–58.

[D'Oro et al., 2022] D'Oro, S., Bonati, L., Polese, M. e Melodia, T. (2022). OrchestRAN: Network Automation through Orchestrated Intelligence in the Open RAN. Em *IEEE Conference on Computer Communications (INFOCOM)*, p. 270–279.

[Enns et al., 2011] Enns, R., Björklund, M., Bierman, A. e Schönwälder, J. (2011). Network Configuration Protocol (NETCONF). RFC 6241.

[Filali et al., 2023] Filali, A., Nour, B., Cherkaoui, S. e Kobbane, A. (2023). Communication and Computation O-RAN Resource Slicing for URLLC Services using Deep Reinforcement Learning. *IEEE Communications Standards Magazine*, 7(1):66–73.

[Filho et al., 2022] Filho, R. H. S., Ferreira, T. N., Mattos, D. M. F. e Medeiros, D. S. V. (2022). An Efficient and Decentralized Fuzzy Reinforcement Learning Bandwidth Controller for Multitenant Data Centers. *Journal of Network and Systems Management*, 30(4).

[Garcia-Saavedra e Costa-Pérez, 2021] Garcia-Saavedra, A. e Costa-Pérez, X. (2021). O-RAN: Disrupting the Virtualized RAN Ecosystem. *IEEE Communications Standards Magazine*, 5(4):96–103.

[Gramaglia et al., 2022] Gramaglia, M., Camelo, M., Fuentes, L., Ballesteros, J., Baldoni, G., Cominardi, L., Garcia-Saavedra, A. e Fiore, M. (2022). Network Intelligence for Virtualized RAN Orchestration: The DAEMON Approach. Em *Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit)*, p. 482–487.

[Jacobs et al., 2021] Jacobs, A. S., Pfitscher, R. J., Ribeiro, R. H., Ferreira, R. A., Granville, L. Z., Willinger, W. e Rao, S. G. (2021). Hey, Lumi! Using Natural Language for Intent-Based Network Management. Em *USENIX Annual Technical Conference*, p. 625–639.

[Johnson et al., 2022] Johnson, D., Maas, D. e Van Der Merwe, J. (2022). NexRAN: Closed-Loop RAN Slicing in POWDER-A Top-to-Bottom Open-Source Open-RAN use Case. Em *ACM Workshop on Wireless Network Testbeds, Experimental evaluation & CHaracterization (WiNTECH)*, p. 17–23.

[Khatib et al., 2023] Khatib, E. J., Álvarez-Merino, C. S., Luo-Chen, H. Q. e Moreno, R. B. (2023). Designing a 6G Testbed for Location: Use Cases, Challenges, Enablers and Requirements. *IEEE Access*, 11:10053–10091.

[Leivadeas e Falkner, 2022] Leivadeas, A. e Falkner, M. (2022). A Survey on Intent Based Networking. *IEEE Communications Surveys & Tutorials*, p. 1–32.

[Liyanage et al., 2023] Liyanage, M., Braeken, A., Shahabuddin, S. e Ranaweera, P. (2023). Open RAN Security: Challenges and Opportunities. *Journal of Network and Computer Applications*, 214:103621.

[Lopez et al., 2022] Lopez, M. A., Barbosa, G. N. N. e Mattos, D. M. F. (2022). New Barriers on 6G Networking: An Exploratory Study on the Security, Privacy and Opportunities for Aerial Networks. Em *International Conference on 6G Networking (6GNet)*, p. 1–6.

[Martin-Perez et al., 2022] Martin-Perez, J., Molner, N., Malandrino, F., Bernardos, C. J., Oliva, A. d. l. e Gomez-Barquero, D. (2022). Choose, not Hoard: Information-to-Model Matching for Artificial Intelligence in O-RAN. *IEEE Communications Magazine*, p. 1–7. Aceito para publicação.

[Mimran et al., 2022] Mimran, D., Bitton, R., Kfir, Y., Klevansky, E., Brodt, O., Lehmann, H., Elovici, Y. e Shabtai, A. (2022). Security of Open Radio Access Networks. *Computers & Security*, 122:102890.

[Motalleb et al., 2023] Motalleb, M. K., Shah-Mansouri, V., Parsaeefard, S. e López, O. L. A. (2023). Resource Allocation in an Open RAN System Using Network Slicing. *IEEE Transactions on Network and Service Management*, 20(1):471–485.

[Neto et al., 2020] Neto, H. N. C., Mattos, D. M. F. e Fernandes, N. C. (2020). Privacidade do Usuário em Aprendizado Colaborativo: Federated Learning, da Teoria à Prática. Em *Minicursos do XX Simpósio Brasileiro de Segurança da Informação e de Sistemas Computacionais (SBSeg)*, p. 1–55.

[O-RAN Working Group 1, 2021] O-RAN Working Group 1 (2021). O-RAN Operations and Maintenance Interface Specification. Especificação Técnica v04.00, O-RAN Alliance. Disponível em `https://orandownloadsweb.azurewebsites.net/specifications`.

[O-RAN Working Group 1, 2023a] O-RAN Working Group 1 (2023a). O-RAN architecture description 8.0. Especificação Técnica v08.00, O-RAN Alliance. Disponível em https://orandownloadsweb.azurewebsites.net/specifications.

[O-RAN Working Group 1, 2023b] O-RAN Working Group 1 (2023b). Use cases detailed specification. Especificação Técnica v10.00, O-RAN Alliance. Disponível em https://orandownloadsweb.azurewebsites.net/download?id=376.

[O-RAN Working Group 10, 2023] O-RAN Working Group 10 (2023). O-RAN architecture description 8.0. Especificação Técnica v08.00, O-RAN Alliance. Disponível em https://orandownloadsweb.azurewebsites.net/specifications.

[O-RAN Working Group 2, 2021a] O-RAN Working Group 2 (2021a). AI/ML workflow description and requirements. Especificação Técnica v01.03, O-RAN Alliance. Disponível em https://orandownloadsweb.azurewebsites.net/specifications.

[O-RAN Working Group 2, 2021b] O-RAN Working Group 2 (2021b). Non-RT RIC: Functional Architecture. Relatório Técnico v01.01, O-RAN Alliance. Disponível em https://orandownloadsweb.azurewebsites.net/specifications.

[O-RAN Working Group 2, 2023] O-RAN Working Group 2 (2023). A1 interface: General aspects and principles. Especificação Técnica v03.01, O-RAN Alliance. Disponível em https://orandownloadsweb.azurewebsites.net/specifications.

[O-RAN Working Group 3, 2023a] O-RAN Working Group 3 (2023a). Near-rt ric architecture. Especificação Técnica v04.00, O-RAN Alliance. Disponível em https://orandownloadsweb.azurewebsites.net/specifications.

[O-RAN Working Group 3, 2023b] O-RAN Working Group 3 (2023b). O-RAN E2 General Aspects and Principles (E2GAP). Especificação Técnica v03.00, O-RAN Alliance. Disponível em https://orandownloadsweb.azurewebsites.net/specifications.

[O-RAN Working Group 3, 2023c] O-RAN Working Group 3 (2023c). O-RAN e2 service model (e2sm). Especificação Técnica v03.00, O-RAN Alliance. Disponível em https://orandownloadsweb.azurewebsites.net/specifications.

[O-RAN Working Group 4, 2023] O-RAN Working Group 4 (2023). O-RAN Management Plane Specification 11.0. Especificação Técnica v11.00, O-RAN Alliance. Disponível em https://orandownloadsweb.azurewebsites.net/specifications.

[O-RAN Working Group 6, 2023] O-RAN Working Group 6 (2023). O2 Interface General Aspects and Principles. Especificação Técnica v03.00, O-RAN Alliance. Disponível em https://orandownloadsweb.azurewebsites.net/specifications.

[Orhan et al., 2021] Orhan, O., Swamy, V. N., Tetzlaff, T., Nassar, M., Nikopour, H. e Talwar, S. (2021). Connection Management xAPP for O-RAN RIC: A Graph Neural Network and Reinforcement Learning Approach. Em *IEEE International Conference on Machine Learning and Applications (ICMLA)*, p. 936–941.

[OSC, 2022] OSC (2022). The O-RAN Software Community (SC) Documentation. Relatório técnico. Disponível em `https://docs.o-ran-sc.org/en/latest/index.html`.

[Otter et al., 2020] Otter, D. W., Medina, J. R. e Kalita, J. K. (2020). A Survey of the Usages of Deep Learning for Natural Language Processing. *IEEE Transactions on Neural Networks and Learning Systems*.

[Polese et al., 2022] Polese, M., Bonati, L., D'Oro, S., Basagni, S. e Melodia, T. (2022). ColO-RAN: Developing Machine Learning-Based xApps for Open RAN Closed-Loop Control on Programmable Experimental Platforms. *IEEE Transactions on Mobile Computing*.

[Polese et al., 2023] Polese, M., Bonati, L., D'Oro, S., Basagni, S. e Melodia, T. (2023). Understanding O-RAN: Architecture, Interfaces, Algorithms, Security, and Research Challenges. *IEEE Communications Surveys & Tutorials*, p. 1–1.

[Polese et al., 2021] Polese, M., Restuccia, F. e Melodia, T. (2021). DeepBeam: Deep Waveform Learning for Coordination-Free Beam Management in mmWave Networks. Em *International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing (MobiHoc)*, p. 61–70.

[Popovski et al., 2018] Popovski, P., Trillingsgaard, K. F., Simeone, O. e Durisi, G. (2018). 5G Wireless Network Slicing for eMBB, URLLC, and mMTC: A Communication-Theoretic View. *IEEE Access*, 6:55765–55779.

[Puligheddu et al., 2023] Puligheddu, C., Ashdown, J., Chiasserini, C. F. e Restuccia, F. (2023). SEM-O-RAN: Semantic and Flexible O-RAN Slicing for NextG Edge-Assisted Mobile Systems. Em *IEEE Conference on Computer Communications (INFOCOM)*.

[Ramezanpour e Jagannath, 2022] Ramezanpour, K. e Jagannath, J. (2022). Intelligent Zero Trust Architecture for 5G/6G Networks: Principles, Challenges, and the Role of Machine Learning in the Context of O-RAN. *Computer Networks*, p. 109358.

[Ramos et al., 2021] Ramos, H. S., Maia, G., Papa, G. L., Alvim, M. S., Loureiro, A. A. F., Cardoso-Pereira, I., Campos, D. H. C., Filipakis, G., Riquetti, G., Chagas, E. T. C., Barros, P. H., Gomes, G. N. e Cid-Allende, H. (2021). Aprendizado Federado Aplicado à Internet das Coisas. Em *Minicursos do XXXIX Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC)*.

[Rego et al., 2022] Rego, I., Medeiros, L., Alves, P., Goldbarg, M., Lopes, V., Flor, D., Barros, W., Sousa, V., Aranha, E., Martins, A. et al. (2022). Prototyping Near-Real Time RIC O-RAN xApps for Flexible ML-based Spectrum Sensing. Em *IEEE Conference on Network Function Virtualization and Software Defined Networks (NFV-SDN)*.

[Rezazadeh et al., 2023] Rezazadeh, F., Zanzi, L., Devoti, F., Chergui, H., Costa-Pérez, X. e Verikoukis, C. (2023). On the Specialization of FDRL Agents for Scalable and Distributed 6G RAN Slicing Orchestration. *IEEE Transactions on Vehicular Technology*, 72(3):3473–3487.

[Santos Filho et al., 2020] Santos Filho, R. H., Mattos, D. M. F. e Medeiros, D. S. V. (2020). Agentes Inteligentes baseados em Aprendizado por Reforço para Alocaçao Dinâmica de Tráfego em Nuvens. Em *XXXVIII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*, p. 141–154.

[Saraiva Jr et al., 2022] Saraiva Jr, R. G., Oliveira, K. K. L. e Nascimento, F. A. O. (2022). Classificação de Tráfego em Redes Móveis Inteligentes Usando Abordagem de Aprendizado de Máquina. Em *XL Simpósio Brasileiro de Telecomunicações e Processamento de Sinais (SBrT 2022)*, p. 1–5.

[Sharara et al., 2022] Sharara, M., Pamuklu, T., Hoteit, S., Vèque, V. e Erol-Kantarci, M. (2022). Policy-Gradient-Based Reinforcement Learning for Computing Resources Allocation in O-RAN. Em *IEEE International Conference on Cloud Networking (CloudNet)*, p. 229–236.

[Singh e Khoa Nguyen, 2022] Singh, A. K. e Khoa Nguyen, K. (2022). Joint Selection of Local Trainers and Resource Allocation for Federated Learning in Open RAN Intelligent Controllers. Em *IEEE Wireless Communications and Networking Conference (WCNC)*, p. 1874–1879.

[Singh e Nguyen, 2022] Singh, A. K. e Nguyen, K. K. (2022). MCORANFed: Communication Efficient Federated Learning in Open RAN. Em *IFIP Wireless and Mobile Networking Conference (WMNC)*, p. 15–22.

[Skorupski e Brakle, 2020] Skorupski, M. e Brakle, T. V. (2020). SMO - Service Management and Orchestration. Relatório técnico. Disponível em `https://wiki.o-ran-sc.org/display/OAM/SMO+-+Service+Management+and+Orchestration`.

[Tang et al., 2023] Tang, X., Liu, S., Du, X. e Guizani, M. (2023). Sparsity-Aware Intelligent Massive Random Access Control in Open RAN: A Reinforcement Learning Based Approach. *arXiv preprint arXiv:2303.02657*.

[Van Hasselt et al., 2016] Van Hasselt, H., Guez, A. e Silver, D. (2016). Deep Reinforcement Learning with Double Q-learning. Em *AAAI conference on artificial intelligence*, p. 2094–2100.

[Westerberg e Fiorani, 2020] Westerberg, E. e Fiorani, M. (2020). The Innovation Potential of Non Real-time RAN Intelligent Controller. Relatório técnico, Ericsson. Disponível em `https://www.ericsson.com/en/blog/2020/10/innovation-potential-of-non-real-time-ran-intelligent-controller`.

[Zhang et al., 2022] Zhang, H., Zhou, H. e Erol-Kantarci, M. (2022). Team Learning-based Resource Allocation for Open Radio Access Network (O-RAN). Em *IEEE International Conference on Communications (ICC)*, p. 4938–4943.