BRASÍLIA - DF

SBRC

Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos

22 A 26 DE MAIO DE 2023

Minicursos da 41ª edição do Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos

Organizadores:

Geraldo Pereira Rocha Filho (UESB)

Marcelo Antonio Marotta (UnB)

Marcos Fagundes Caetano (UnB)

Rafael Lopes Gomes (UECE)

Organizadores:
Geraldo Pereira Rocha Filho (UESB)
Marcelo Antonio Marotta (UnB)
Marcos Fagundes Caetano (UnB)
Rafael Lopes Gomes (UECE)

Minicursos da 41ª edição do Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos

Porto Alegre
Sociedade Brasileira de Computação – SBC
2023

Dados Internacionais de Catalogação na Publicação (CIP)

S612 Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (41. : 22 – 26 maio 2023 : Porto Alegre)

Minicursos do SBRC 2023 [recurso eletrônico] / organização: Geraldo Pereira Rocha Filho... [et al.]. – Dados eletrônicos. – Porto Alegre: Sociedade Brasileira de Computação, 2023.

215 p. : il. : PDF.

Modo de acesso: World Wide Web.

Inclui bibliografia

ISBN 978-85-7669-543-1 (e-book)

1. Computação – Brasil – Evento. 2. Redes de computadores. 3. Sistemas distribuídos. 4. Gerenciamento de serviços. 5. Internet das Coisas. 6. Aplicações críticas. I. Rocha Filho, Geraldo Pereira. II. Marotta, Marcelo Antonio. III. Caetano, Marcos Fagundes. IV. Gomes, Rafael Lopes. V. Sociedade Brasileira de Computação. VI. Título.

CDU 004(063)

Ficha catalográfica elaborada por Annie Casali – CRB-10/2339

Biblioteca Digital da SBC – SBC OpenLib

Índices para catálogo sistemático:

1. Ciência e tecnologia dos computadores : Informática – Publicação de conferências,congressos e simpósios etc. ... 004(063)

Sumário

Mensagem dos Coordenadores dos Minicursos do SBRC 2023

É com grande satisfação que apresentamos os Minicursos do 41º Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC), realizado de 22 a 26 de Maio de 2023 em Brasília. Os Minicursos do SBRC produzem os capítulos do Livro de Minicursos que, tradicionalmente, trazem material de alta qualidade, abordando temas de pesquisa e desenvolvimento de interesse da comunidade. Apesar das apresentações terem sido realizadas virtualmente, o alto nível técnico e o engajamento dos participantes demonstram a importância deste evento, já tradicional ao longo das edições do SBRC. O livro, assim como as apresentações, tem sempre um viés didático, capaz de motivar alunos de graduação, pós-graduação e profissionais da área a mergulharem nos temas propostos. A ideia é abrir horizontes e complementar a formação dos participantes em temas recentes que, possivelmente, ainda não são cobertos nas grades curriculares das instituições tradicionais de ensino e pesquisa.

Neste ano, 8 submissões de propostas foram recebidas, levando em conta o prazo final, que ocorreu após dois adiamentos. Os adiamentos foram realizados para que mais submissões pudessem surgir e, de fato, surgiram. Todas as propostas foram atribuídas a, no mínimo, quatro revisores, onde propostas com notas próximas tiveram até seis revisores. Os membros do comitê técnico aceitaram e revisaram no prazo. A organização dos Minicursos do SBRC 2023 agradece aos 24 membros do comitê de 2023, que se dedicaram à revisão das propostas de forma sempre atenciosa.

O montante de submissões foi providencial, uma vez que deixou a organização confortável para selecionar as 4 melhores propostas, número este que vem sendo praticado ao longo dos últimos anos. As 4 propostas selecionadas representaram certamente ótimas oportunidades a atrair público ao evento, já que abordam temas como Softwarização de redes, Aplicações Inteligentes, Segurança em redes e em sistemas distribuídos e Redes 5G. É válido mencionar que, entre as propostas não contempladas, algumas teriam plenas condições de serem aceitas, o que infelizmente não foi possível, desta vez.

Agradeço aos organizadores gerais do SBRC 2023, Geraldo Pereira Rocha Filho (DCET-UESB), Marcelo Antonio Marotta (CIC-UnB) e Marcos Fagundes Caetano (CIC-UnB), que, com muita dedicação e empenho, organizaram o SBRC de 2023. Agradeço, especialmente, pela confiança em mim depositada para a organização dos Minicursos, pelo trato sempre muito gentil e pela infinita paciência. Agradeço, também, aos autores dos minicursos, pois sem a dedicação e esmero de todos na escrita e na entrega pontual do material, nada disso seria viável. Temos certeza que todas as apresentações abrilhantaram todo o trabalho e possibilitaram plena absorção técnica da audiência. O desfecho dos minicursos foi suave graças à colaboração e ao profissionalismo de todos. Agora, desejamos que o árduo trabalho do evento gere frutíferas discussões em direção ao avanço da ciência nacional.

Rafael Lopes Gomes (UECE)

Coordenador de Minicursos do SBRC 2022

Comitê de Programa dos Minicursos do SBRC 2023

1. Alberto Egon Schaeffer-Filho – UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
2. Aldri Luiz dos Santos – UNIVERSIDADE FEDERAL DE MINAS GERAIS
3. Alex Borges Vieira – UNIVERSIDADE FEDERAL DE JUIZ DE FORA
4. Alfredo Goldman – UNIVERSIDADE DE SÃO PAULO
5. Christian Esteve Rothenberg – UNIVERSIDADE ESTADUAL DE CAMPINAS
6. Daniel de Oliveira Cunha – UNIVERSIDADE FEDERAL DO RIO DE JANEIRO
7. Daniel Fernandes Macedo – UNIVERSIDADE FEDERAL DE MINAS GERAIS
8. Daniel Ludovico Guidoni – UNIVERSIDADE FEDERAL DE OURO PRETO
9. Daniel Menasché – UNIVERSIDADE FEDERAL DO RIO DE JANEIRO
10. Denis Rosário – Federal University of Para
11. Dianne de Medeiros – UNIVERSIDADE FEDERAL FLUMINENSE
12. Edmundo Madeira – UNIVERSIDADE ESTADUAL DE CAMPINAS
13. Eduardo Coelho Cerqueira – UNIVERSIDADE FEDERAL DO PARÁ
14. Helder Oliveira – UNIVERSIDADE FEDERAL DO ABC
15. Igor Monteiro Moraes – UNIVERSIDADE FEDERAL FLUMINENSE
16. Juliano Fischer Naves – Instituto Federal de Educação
17. Kelvin Dias – UNIVERSIDADE FEDERAL DE PERNAMBUCO
18. Leobino Nascimento Sampaio – UNIVERSIDADE FEDERAL DA BAHIA
19. Luis Carlos Erpen De Bona – Federal University of Parana
20. Marcelo Dias de Amorim – CNRS
21. Mauro Fonseca – UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
22. Rodrigo de Souza Couto – UNIVERSIDADE FEDERAL DO RIO DE JANEIRO
23. Ronaldo Alves Ferreira – Federal University of Mato Grosso do Sul
24. Weverton Luis da Costa Cordeiro – UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

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*.

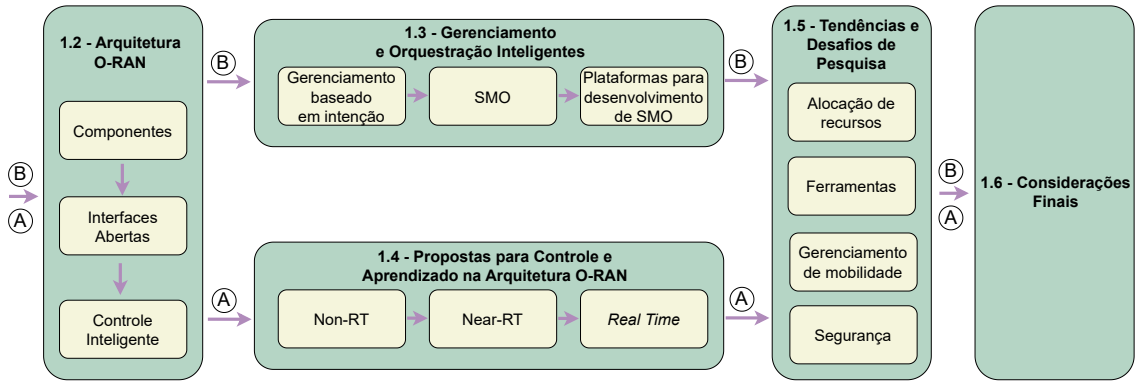

Figura 1.1. Organização deste capítulo e sugestão de dois percursos alternativos. O percurso Ⓐ foca o gerenciamento, orquestração e suas ferramentas, enquanto o Ⓑ foca a abordagem do controle inteligente em artigos científicos.

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-

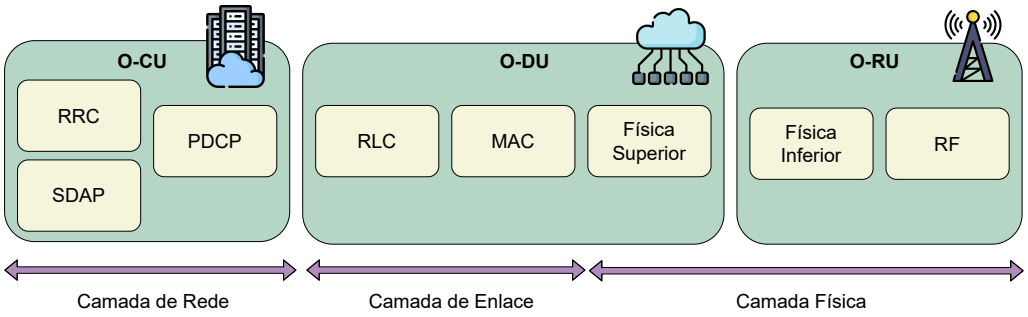

Figura 1.2. Opção de divisão 7.2x e os componentes da O-RAN. A camada física é dividida entre a O-RU e a O-DU. A O-DU também implementa a camada de enlace, enquanto a O-CU é responsável pela camada de rede.

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;

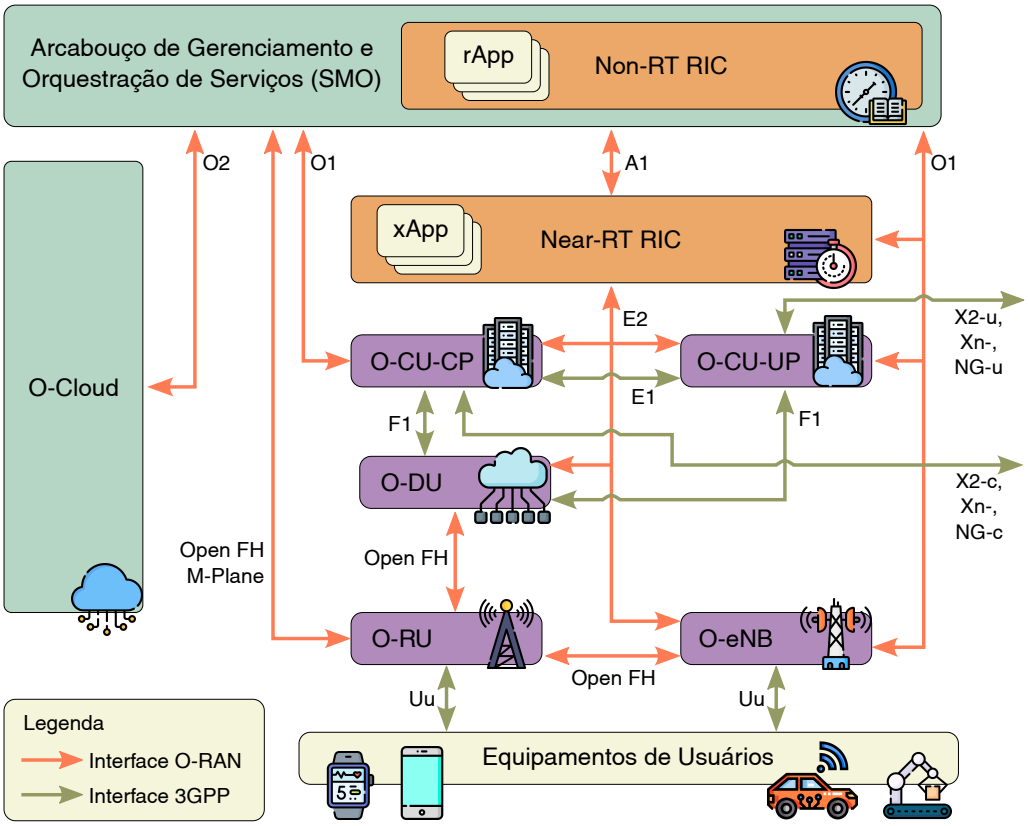

Figura 1.3. Componentes da arquitetura O-RAN. A arquitetura divide as funcionalidades da estação rádio base em três componentes, O-CU, O-DU e O-RU, além de definir o SMO, os RICs e a O-Cloud. Os componentes da arquitetura interagem por meio de interfaces especificadas pela O-RAN Alliance (em laranja) 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

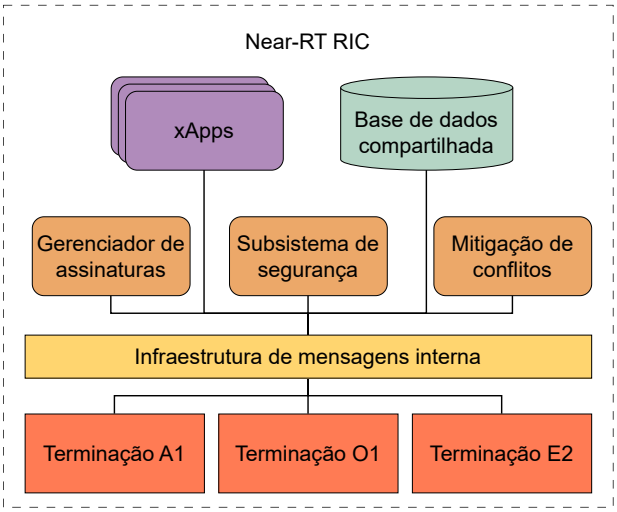

Figura 1.4. O Near-RT RIC é um componente programável da Open RAN que implementa funcionalidades que executam em uma escala de tempo entre 10 ms e 1 s. Os diversos componentes desse RIC oferecem suporte para a execução das xApps. A comunicação com o restante dos componentes da arquitetura é feita pelas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

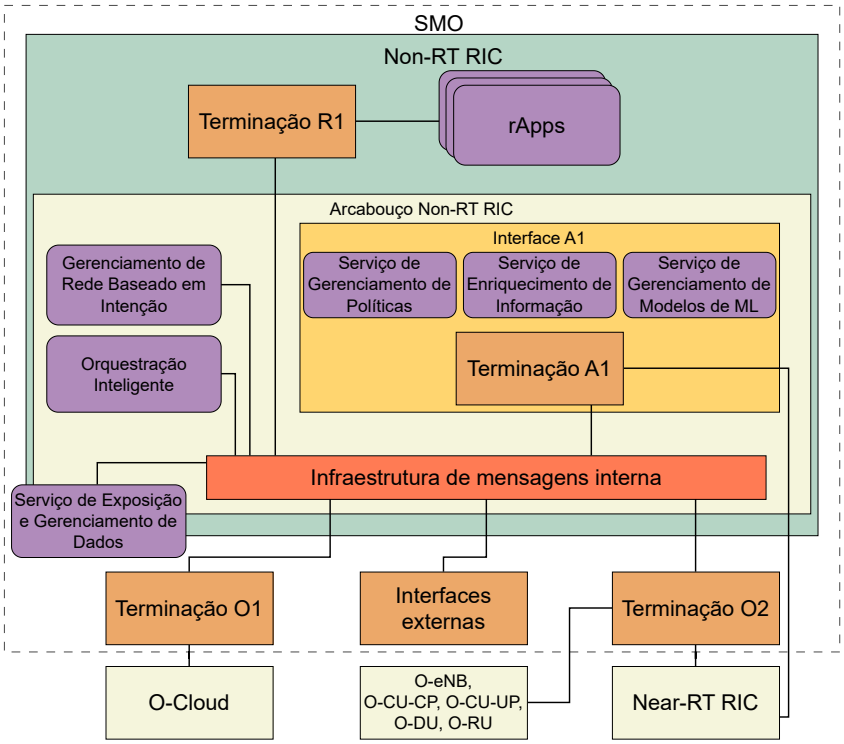

Figura 1.5. O Non-RT RIC faz parte do SMO e é um componente programável da Open RAN que implementa funcionalidades que executam em uma escala de tempo igual ou maior do que 1 s. Os diversos componentes desse RIC oferecem suporte para a execução das rApps. A comunicação com o restante dos componentes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

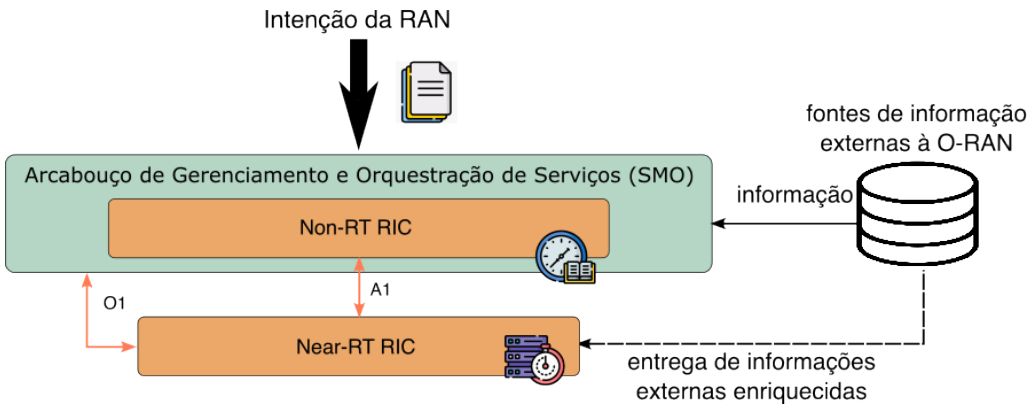

Figura 1.6. Elementos da arquitetura O-RAN envolvidos no gerenciamento baseado em intenção, adaptado [O-RAN Working Group 2, 2023]. A intenção é injetada por fontes externas no Non-RT RIC, que automaticamente extrai o objetivo geral da intenção e, baseado nesse objetivo e em eventos e contadores fornecidos 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/.

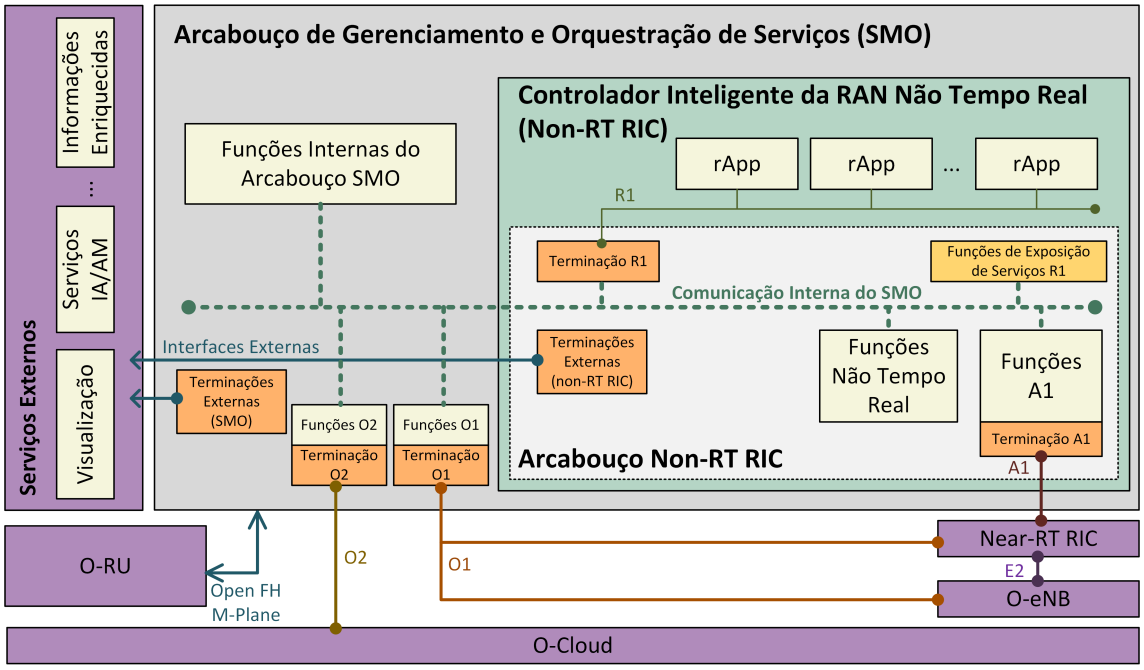

Figura 1.7. Arcabouço de Gerenciamento e Orquestração de Serviços (*Service Management and Orchestration* - SMO). A interface O1 ocorre entre o SMO e os elementos gerenciáveis. A interface O2 permite que o SMO exerça o gerenciamento de recursos na nuvem O-Cloud. A interface R1 permite que as rApps se 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/.

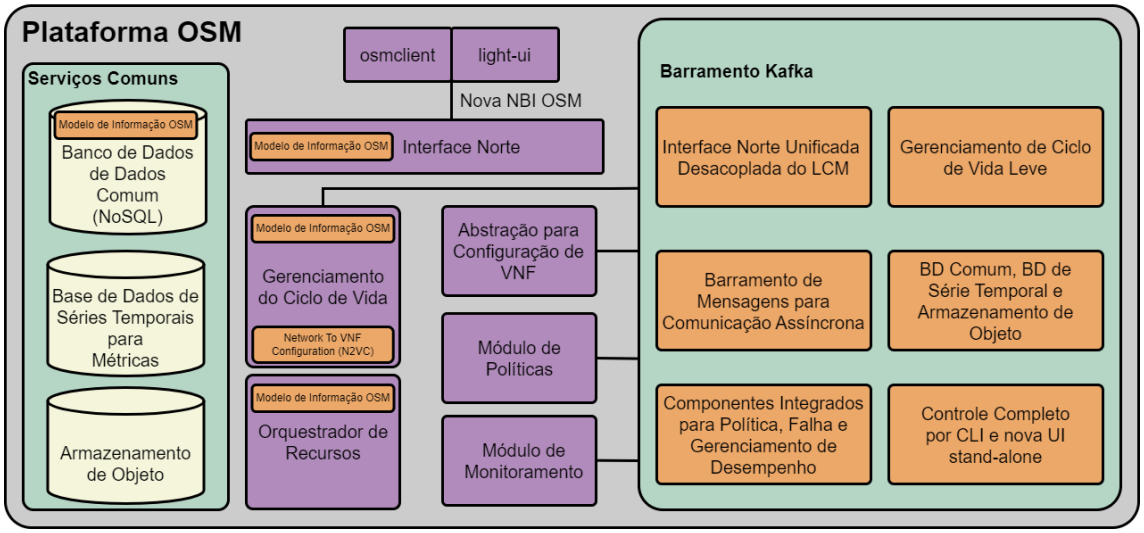

Figura 1.8. Principais módulos e componentes da plataforma OSM. A interface O1 é realizada pelos módulos LCM e POL. O barramento de comunicação interno é realizado pelo Apache Kafka. O módulo Interface Norte realiza a interface R1 da O-RAN.

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/.

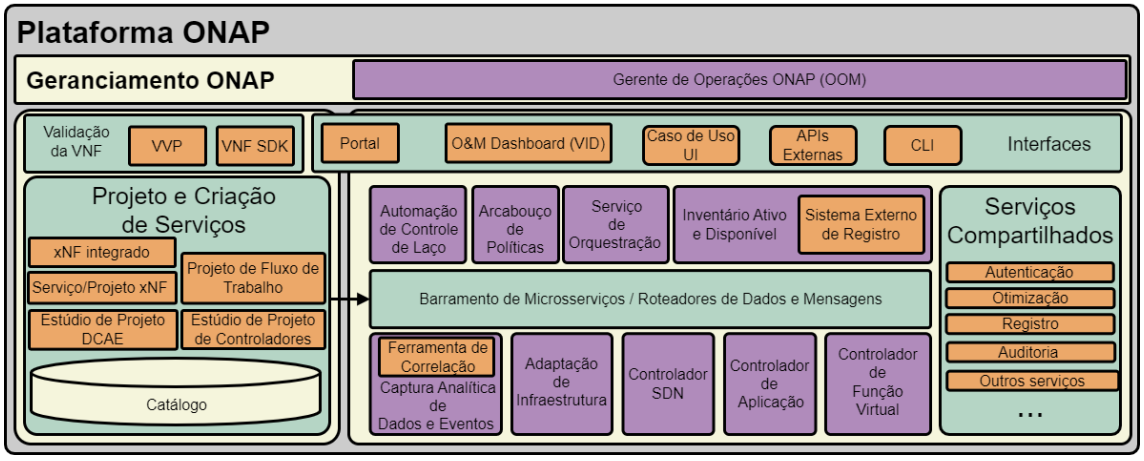

Figura 1.9. Principais módulos e componentes da plataforma ONAP. A plataforma é complexa e apresenta módulos especializados alinhados com as definições da O-RAN Software Community para plataformas de Gerenciamento e Orquestração de Serviços (SMO).

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.

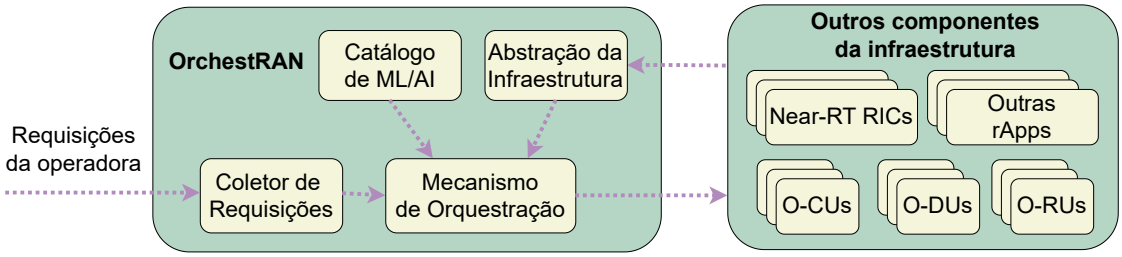

Figura 1.10. Visão geral da OrchestRAN. A rApp recebe requisições da operadora e utiliza seu mecanismo de orquestração para escolher quais modelos serão instanciados em quais localidades. Esses modelos são utilizados para implementar as funcionalidades solicitadas. Adaptada de [D'Oro et al., 2022].

- **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.

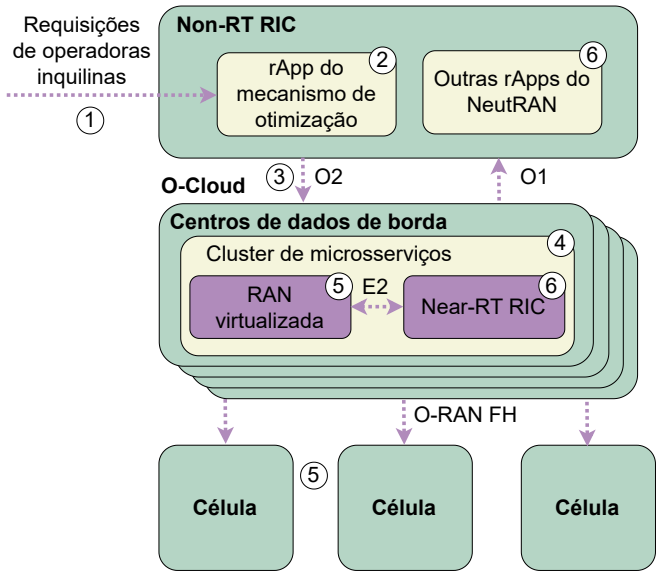

Figura 1.11. Visão Geral do NeutRAN. As operadoras inquilinas solicitam o provisionamento de uma RAN em uma infraestrutura hospedeira neutra. O Non-RT RIC, por sua vez, executa o mecanismo de otimização para provisionar a RAN virtualizada. A partir das decisões desse mecanismo, o Near-RT RIC realiza 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.

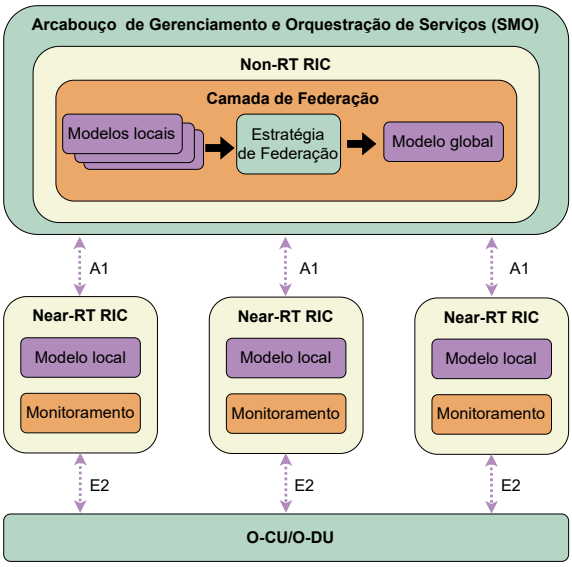

Figura 1.12. Exemplo de arquitetura de FL. Os Near-RT RICs treinam modelos locais e enviam seus parâmetros ao Non-RT RIC. A camada de federação do Non-RT RIC executa uma estratégia de federação, agregando os parâmetros recebidos e enviando aos Near-RT RICs. A partir desses parâmetros, os Near-RT RICs 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.

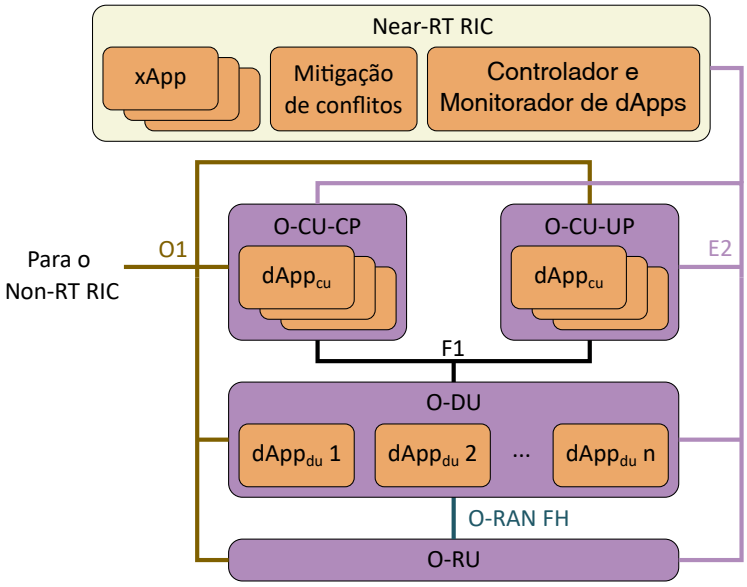

Figura 1.13. Arquitetura O-RAN com alterações para dar suporte às dApps. Os dApps são executados pela O-CU e pela O-DU. Adaptada de [D'Oro et al., 2022].

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.

Capítulo 2

Um Panorama dos Serviços de Saúde Avançados: Conectividade e Segurança em Sistemas de Vida Assistida

Adriana V. Ribeiro (UFBA), Fernando Nakayama (UFPR), Michele Nogueira (UFMG), Leobino N. Sampaio (UFBA)

Abstract

An ambient assisted living is an advanced health service that includes smart space applications and location-independent individual monitoring. The development of assisted living applications relies on several technologies, such as the Internet of Things, short and long-range communication protocols, middlewares, cloud computing, and artificial intelligence. The heterogeneity of the architectural components and communication protocols arises challenges in providing network and security requirements. This chapter overviewes this service and identifies the main communication and security requirements and the challenges to address them.

Resumo

O ambiente de vida assistida é um serviço avançado de saúde que inclui aplicações em espaços inteligentes e o monitoramento da saúde de indivíduos, independente de sua localização. O desenvolvimento dessas aplicações segue diferentes tecnologias, como a Internet das Coisas, protocolos de comunicação de curto e longo alcance, middlewares, computação em nuvem e inteligência artificial. A heterogeneidade dos componentes que fazem parte dessas arquiteturas e dos próprios protocolos de comunicação utilizados geram desafios para alcançar os requisitos de rede e segurança das aplicações. Este capítulo apresenta uma visão geral desse serviço, identifica os principais requisitos de comunicação e segurança e os principais desafios para endereçá-los.

2.1. Introdução

Em 2017, o Ministério da Saúde do Brasil informou que as hospitalizações de pessoas idosas no Sistema Único de Saúde (SUS) custam aproximadamente 30% a mais

quando comparadas às de adultos entre 25 e 59 anos [Heemann and Hermsdorf 2017]. Paralelo a isso, o Instituto Brasileiro de Geografia e Estatística (IBGE) estima que o número de pessoas com idade superior a 65 anos no Brasil será equivalente a 25.5% da população até 2060 [IBGE 2018]. Embora esses sejam dados brasileiros, essa mudança demográfica e o aumento dos custos na área de saúde têm sido observados ao redor do mundo [United Nations 2020], intensificando a necessidade de pesquisas que contribuam com o desenvolvimento de soluções economicamente viáveis [Maskeliūnas et al. 2019, Tun et al. 2021]. Além da questão econômica, também é importante ressaltar que a mudança demográfica mundial requer uma maior quantidade de profissionais de saúde para atender à população. Portanto, a tecnologia é vista como aliada para oferecer saúde de qualidade e a um menor custo.

Com o objetivo de produzir soluções mais baratas e eficazes, a Internet das Coisas da Saúde (do inglês, *Internet of Health Things* – IoHT) se baseia no uso de dispositivos e aplicações de Internet das Coisas (do inglês, *Internet of Things* – IoT) e em arquiteturas de rede para dar suporte às aplicações e serviços de saúde [Rodrigues et al. 2018]. A IoHT auxilia na diminuição dos custos da saúde pública e privada, através de mecanismos de monitoramento e análise de dados que possibilitem a prevenção de eventos adversos e, consequentemente, a redução de hospitalizações [Tun et al. 2021]. Como as aplicações da IoHT podem estar relacionadas a diversos públicos e funcionalidades, existem vários critérios que podem ser utilizados para classificá-las. Em [Rodrigues et al. 2018], os autores as classificam em quatro áreas principais: monitoramento de saúde remoto, soluções de saúde baseadas em *smartphones*, dispositivos vestíveis e Ambiente de Vida Assistida (do inglês, *Ambient Assisted Living* – AAL).

A AAL é um serviço que faz uso de sistemas e aplicações IoT para promover mais independência e bem estar às pessoas idosas ou com deficiência [Rodrigues et al. 2018], bem como para avaliar e identificar padrões em populações [Wang et al. 2022]. Os principais tipos de aplicações AAL incluem o monitoramento do espaço físico (e.g., segurança do local, temperatura e umidade) e do indivíduo (e.g., frequência cardíaca, pressão sanguínea e temperatura corporal). No entanto, as soluções de AAL comumente estão restritas a um espaço e cobrem apenas uma área previamente estabelecida. Assim, o monitoramento é interrompido caso o usuário saia do espaço monitorado. Para que haja um acompanhamento efetivo da saúde do indivíduo, as atividades de monitoramento devem ocorrer independente de sua localização [Nakayama et al. 2022]. Portanto, o uso de dispositivos móveis na coleta e encaminhamento dos dados de forma contínua é imprescindível para o funcionamento de diversas aplicações.

As aplicações de saúde mais simples geralmente têm um objetivo específico e são baseadas no monitoramento de poucas características. Por exemplo, uma aplicação simples para controle de diabetes pode envolver a utilização de sensores de glicose para monitorar o nível de açúcar no sangue e emitir algum alerta para o indivíduo, um familiar ou membro da equipe médica em caso de problemas. No entanto, algumas soluções de saúde atuais têm caráter mais complexo pois buscam ofertar serviços de saúde holísticos e mais inteligentes. O aumento na complexidade das aplicações introduz a necessidade de outras tecnologias para dar suporte aos serviços de saúde avançados. Em [Philip et al. 2021], os autores destacam cinco tecnologias principais para o desenvolvimento de soluções de saúde: aplicações IoT, computação em nuvem, utilização de *middlewares*, comunicações

de rede de curto alcance e sensores. O uso dessas tecnologias deve considerar os requisitos do usuário e da aplicação para que haja uma adoção efetiva das soluções.

Alguns requisitos gerais que os sistemas de vida assistida devem atender incluem a utilidade do serviço, a facilidade de uso e a curva de aprendizado. Além disso, a adoção de soluções na área de saúde é influenciada diretamente por aspectos como idade, gênero e escolaridade dos indivíduos [Maskeliūnas et al. 2019]. Tendo em vista que, em muitos casos, a população idosa tem menos familiaridade com o manuseio de equipamentos tecnológicos, é preciso considerar o conforto do dispositivo, sua capacidade de gerenciamento e facilidade de uso para que haja sucesso na adoção dessas soluções. No entanto, em um sistema como esse, envolvendo uma arquitetura com diversos componentes distribuídos, também é fundamental pensar em questões relacionadas à conectividade e à segurança dos dados e sistema. A conectividade é essencial para que haja comunicação entre os componentes do sistema. Enquanto a segurança gera a proteção do sistema e da aplicação contra ações de modificação dos dados e recursos, acesso indevido aos mesmos e indisponibilidade do serviço.

Os requisitos de comunicação e de segurança relacionados às aplicações de saúde são discutidos ao longo deste capítulo de livro considerando os diferentes processos que acontecem para que uma aplicação IoHT funcione adequadamente: coleta de dados, transmissão local e encaminhamento de dados para serviços remotos. Além disso, também serão tratados os requisitos gerais envolvendo conectividade, mobilidade, Qualidade de Serviço (do inglês, *Quality of Service* – QoS) e Qualidade de Experiência (do inglês, *Quality of Experience* – QoE). Do ponto de vista de segurança, é preciso observar os requisitos voltados aos princípios básicos, como disponibilidade, integridade, confidencialidade, privacidade e controle de acesso.

Como a saúde é a área que mais tem desenvolvido soluções baseadas em IoT [Rodrigues et al. 2018], é imprescindível que os protocolos e os dispositivos criados consigam ser utilizados de forma efetiva para proporcionar soluções nessa área, pois isso também influencia o desenvolvimento da própria IoT. O desenvolvimento de soluções reais depende do uso de mecanismos que garantam ou deem suporte à segurança e à privacidade dos dados [Rodrigues et al. 2018]. Essa temática tornou-se popular nos últimos anos, mas o uso de protocolos de comunicação e serviços eletrônicos e a preocupação com questões de privacidade e segurança são tópicos tratados há um tempo. O potencial para uso desses serviços tem aumentado e a área de saúde tem liderado o *ranking* de pesquisas envolvendo IoT. Uma característica interessante nas aplicações atuais é a necessidade do monitoramento holístico do indivíduo para o desenvolvimento de serviços inteligentes. Considerando isso e a relevância dessas aplicações para a saúde dos indivíduos, este minicurso discute o processo de evolução do uso da tecnologia na saúde de idosos e as características das aplicações atuais. Ele também apresenta os conceitos relacionados à IoHT, os requisitos de conectividade e segurança das aplicações e como novos modelos e arquiteturas de rede suportam essas soluções. Por fim, serão apresentados dois estudos de casos e ambientes de experimentação associados à área.

2.2. Evolução do uso da tecnologia na saúde de idosos

Nos últimos 20 anos foram desenvolvidas as principais tecnologias utilizadas nas soluções de saúde atuais, incluindo dispositivos e aplicações IoT, *smartphones*, equipamentos vestíveis e computação em nuvem. Na Figura 2.1, são listadas as características e os aspectos marcantes dessa evolução entre os anos 2000 e 2020. Uma pesquisa publicada em 2004 indicou aplicações passíveis de beneficiar os cuidados com a saúde de idosos: auxílio em situações de emergência, prevenção de queda, monitoramento de temperatura e de parâmetros fisiológicos, sistemas de segurança para a casa, monitoramento de medicação, entre outros [Demiris et al. 2004]. Apesar disso, as soluções da época estavam mais voltadas à telemedicina, à criação de sistemas eletrônicos para a saúde e à utilização de e-mail, SMS e fax para facilitar a comunicação entre o indivíduo e a equipe médica. Havia iniciado o uso de câmeras para fazer monitoramento dos indivíduos, mas a privacidade era uma preocupação dos idosos já nesse período. Outras preocupações incluíam a substituição da assistência humana pela tecnologia, o uso de dispositivos amigáveis e a necessidade de treinamentos específicos. Além disso, algumas características das redes sem fio, como alto custo e dificuldade de interoperabilidade, limitaram o desenvolvimento de soluções móveis [Istepanian and Lacal 2003].

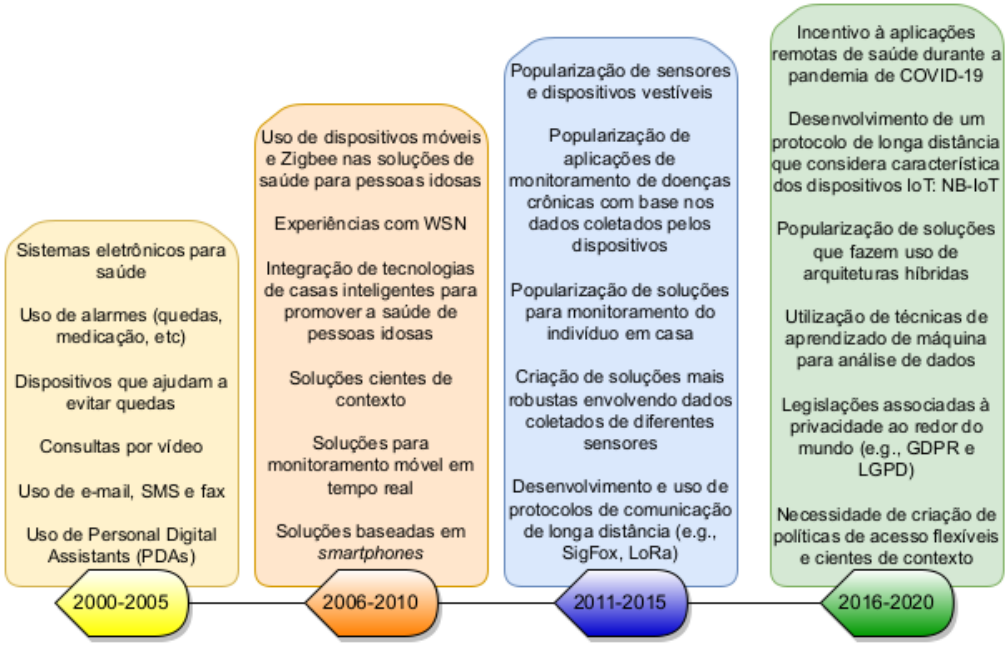

Figura 2.1. Marcos importantes para o desenvolvimento e uso de aplicações de saúde para pessoas idosas.

Entre 2006 e 2010, o desenvolvimento e a popularização dos *smartphones* e a primeira geração de serviços de nuvem influenciaram o crescimento de novas soluções de saúde. Uma pesquisa publicada por [Steele et al. 2009] em 2009 demonstrou que a independência é um dos aspectos mais valorizados por pessoas idosas e que há uma abertura maior ao uso de sistemas e serviços que possam prolongar esse sentimento. Nesta mesma pesquisa, os autores indicaram que as preocupações associadas ao uso de redes de sensores sem fio (do inglês, *Wireless sensors networks* – WSN) estavam relacionadas a fatores como custo, impacto social do uso de dispositivos, possíveis efeitos colaterais causados

pelas ondas de radiofrequência emitidas pela WSN, ansiedade, privacidade, confidencialidade e confiabilidade do sistema. Os participantes destacaram ainda suas preferências em relação ao uso de sensores embutidos em relógios e/ou anéis, para que houvesse mais privacidade em relação ao uso de um dispositivo de monitoramento de saúde. Nesse período, as soluções híbridas começaram a ser pensadas e houve um aumento na diversidade de aplicações, que variavam desde o monitoramento da casa ao monitoramento de sinais vitais e doenças crônicas [Arcelus et al. 2007, Jones et al. 2010].

De 2011 até 2015, ocorreu um cenário mais propício para o desenvolvimento dos atuais serviços de saúde avançados. Os dispositivos vestíveis começaram a se tornar itens mais populares entre as pessoas, independente da idade. Esse fator é importante ao pensar em soluções de saúde para pessoas idosas, uma vez que elas relatam preocupação em relação ao desconforto e impacto social de ter que utilizar sensores de monitoramento de saúde. Além da popularização dos vestíveis, tem-se a popularização de aplicações de monitoramento de doenças crônicas e do ambiente do indivíduo. Em [Rashidi and Mihailidis 2012], os autores destacam as aplicações e as tecnologias associadas ao uso de AAL. Esse trabalho trouxe uma lista dos principais tipos de sensores de ambiente e dispositivos vestíveis usados na época e foi possível observar o uso de algoritmos de reconhecimento de imagens e aplicações voltadas ao monitoramento de atividades cardíacas, cerebrais e musculares, glicose, pressão sanguínea e outras. A integração dessas soluções possibilitou propostas de aplicações mais robustas. Em [Kim et al. 2014], os autores apresentaram soluções de cuidados da saúde com diferentes arquiteturas e envolvendo o acesso à informação por várias partes, como equipe médica e indivíduo. As comunicações entre os dispositivos dentro do ambiente eram baseadas em protocolos de curto alcance, como Zigbee, Bluetooth, RFID e Wi-Fi. Enquanto as comunicações de longa distância se baseavam, principalmente, nos sistemas de telecomunicações móveis [Amiribesheli et al. 2015]. Apesar do grande salto no desenvolvimento de dispositivos e dos aspectos positivos possibilitados através das soluções na época, algumas pessoas idosas relataram não utilizar as aplicações de saúde, pois achavam que não melhorariam sua qualidade de vida de forma significativa [Heart and Kalderon 2013].

A evolução dos serviços remotos de saúde continuou e a evidência de sua necessidade foi ainda mais acentuada durante a pandemia de COVID-19. No primeiro ano da pandemia, muitas consultas e cirurgias eletivas foram suspensas devido à necessidade de concentrar os recursos de saúde no enfrentamento à COVID-19. Além disso, como os idosos faziam parte do grupo de risco, houve receio em ir aos médicos para suas consultas de rotina. A necessidade de alternativas remotas fez com que o governo brasileiro aprovasse a Lei 13.989/2020, que permitia aos profissionais de saúde a realização de atendimentos online. Embora essa lei tenha sido revogada, o senado aprovou o PL998/2020 que garantia que qualquer profissional de saúde pudesse realizar atendimento online. Outras medidas legislativas também foram adotadas no Brasil nesse período e influenciam diretamente as aplicações de saúde. Uma dessas medidas foi a Lei Geral de Proteção de Dados (LGPD) que busca garantir o direito à privacidade dos usuários através da regulamentação do tratamento dos dados pessoais.

Durante os últimos 10 anos, houve mudanças legislativas e tecnológicas que impulsionaram o desenvolvimento e a preocupação com aplicações de saúde. Houve o desenvolvimento de novos dispositivos e tecnologias de suporte, além da adoção de arqui-

teturas híbridas, uso de técnicas e algoritmos de inteligência artificial (IA) e a criação de protocolos de comunicação de longa distância específicos para IoT. A junção dessas soluções tem possibilitado a proposta e o desenvolvimento de aplicações cada vez mais robustas e inteligentes que relacionam dados do indivíduo para possibilitar tratamentos personalizados e, ao mesmo tempo, permitem avaliar políticas de saúde pública e auxiliar na análise e na correlação de fatores associados a diversas doenças. A expectativa – que tem se tornado realidade – é que a IoHT ajude a endereçar diversos problemas associados a pessoas idosas, como limitações em suas atividades diárias, risco de queda, monitoramento de doenças crônicas, demência, desordens mentais e gerenciamento de medicamentos [Maskeliūnas et al. 2019]. A seguir, serão descritas algumas dessas aplicações atuais e suas características gerais.

2.2.1. Cenários de aplicação

Algumas das doenças mais comuns em pessoas idosas incluem problemas cardíacos, diabetes, hipertensão, câncer e Alzheimer. O monitoramento do ambiente e da saúde do indivíduo é essencial para manter essas doenças sob controle, principalmente nos casos das chamadas "doenças silenciosas", como a hipertensão. A Tabela 2.1 apresenta uma relação dos domínios de aplicações de saúde associados às pessoas idosas. Esses domínios têm pontos de intersecção nas aplicações de saúde: por exemplo, o monitoramento de uma doença crônica pode ser realizado com vestíveis, através de aplicações móveis ou mesmo com uso de AAL. Além disso, o monitoramento dos parâmetros de saúde auxilia na prevenção e no diagnóstico de novas doenças. A Tabela 2.2 resume as aplicações de saúde para pessoas idosas.

Tabela 2.1. Domínios de monitoramento associados à saúde de pessoas idosas.

Domínio	Contribuições
Vestíveis e sensores	Executam funções como a detecção de quedas, o monitoramento de padrões de sono, condições cardíacas, níveis de oxigênio no sangue, pressão arterial, temperatura corporal e comportamentos sedentários. Permitem o monitoramento contínuo do indivíduo e podem emitir alertas para familiares e/ou equipe médica em casos de emergência.
AAL	Coleta variáveis do ambiente, como qualidade do ar e intensidade da luz para avaliar suas implicações na saúde do indivíduo. Além disso, é possível integrar aplicações de AAL com robôs, equipamentos vestíveis e tecnologias móveis. Essa integração viabiliza a criação de aplicações mais robustas.
Telemedicina	Definida como o uso da tecnologia para realizar diagnósticos e tratamentos de forma remota. Bastante popular durante a pandemia de COVID-19 e cada vez mais utilizada. É uma alternativa para proporcionar acesso a profissionais de saúde para pessoas em áreas remotas.
Aplicações móveis	Usam serviços de nuvem para armazenar os dados e diferentes pessoas (e.g., equipe médica e familiares) acessam as informações através de aplicativos. Auxilia no monitoramento de inúmeras condições e doenças, bem como possibilita *feedbacks* aos indivíduos com tratamento personalizado.

A relevância no desenvolvimento dessas aplicações é demonstrada através de dados estatísticos: dados do Sistema de Mortalidade do Brasil (SIM), disponibilizados pelo

Tabela 2.2. Aplicações de saúde para pessoas idosas.

Aplicação	Características
Monitoramento de parâmetros de saúde	Medição de parâmetros como nível de oxigênio no sangue, pressão arterial, pulsação, níveis de açúcar, temperatura corporal, ritmo de caminhada, equilíbrio, perfil lipídico, entre outros. Geralmente são utilizados para realizar o monitoramento de doenças crônicas e detecção de quedas.
Monitoramento de doenças crônicas	Faz uso do monitoramento de parâmetros de saúde e podem apresentar requisitos de baixa latência e monitoramento em tempo real. Pode incluir soluções de monitoramento para doenças como Alzheimer e Parkinson.
Rastreamento de atividades	Faz uso de dispositivos como sensores, acelerômetros e GPS para mapear e incentivar a prática de atividade física por pessoas idosas. Os dados coletados podem ser utilizados para aplicações de detecção de queda ou mesmo como insumo para o monitoramento de doenças mentais.
Gerenciamento de medicação	São aplicações que alertam o indivíduo para tomar suas medicações. Além disso, as informações coletadas por aplicações de monitoramento de parâmetros de saúde podem ser usadas como insumo para realizar ajustes medicamentosos pela equipe médica.
Monitoramento da saúde mental e de doenças cognitivas	As doenças mais comuns associadas à saúde mental de idosos incluem Alzheimer, Parkinson, demência, depressão e esquizofrenia. O uso de informações de rastreamento de atividades diárias possibilita avaliar a progressão dessas doenças a partir da análise de padrões e mudanças na rotina. Embora esses aplicativos não substituam o acompanhamento humano, podem auxiliar no bem estar desses indivíduos.
Serviços de emergência	Em emergências, esses serviços são utilizados para enviar alertas e possibilitar acesso ao histórico médico por familiares e/ou equipe médica.

DataSUS[1], mostram que entre o período de 2015 e 2020 foram registrados aproximadamente 4,8 milhões de óbitos de pessoas com idade igual ou superior a 65 anos. Mais de 50% dessas mortes foram ocasionadas por três causas principais: 31,8% às doenças do aparelho circulatório, 16,1% a tumores e 15% às doenças do aparelho respiratório. A maioria dessas doenças pode se beneficiar de soluções IoHT: uma aplicação de monitoramento de diabetes demonstrou uma melhora do controle glicêmico em 0,8% para diabetes do tipo 2 e 0,3% para diabetes do tipo 1 [Kitsiou et al. 2017]. Em outro estudo, foi demonstrado que o monitoramento de pacientes com problemas cardíacos reduziu as taxas de hospitalizações e mortalidade [Bui and Fonarow 2012].

2.3. Principais Conceitos dos Sistemas de Vida Assistida

Os serviços de saúde avançados que têm como foco o atendimento às pessoas idosas, são suportados por cinco tecnologias principais: dispositivos IoT e *mHealth*, protocolos de comunicação de curto e longo alcance, arquiteturas baseadas em computação em nuvem e em névoa, *middlewares* e uso de técnicas de aprendizado de máquina. Cada uma dessas tecnologias tem papel fundamental no desenvolvimento de soluções mais inteligentes e serão discutidas nesta seção.

[1] http://tabnet.datasus.gov.br/cgi/deftohtm.exe?sim/cnv/obt10uf.def

2.3.1. IoHT e *mHealth*

Os dispositivos IoT possibilitam funções de coleta e armazenamento de dados e a execução de ações autônomas com o uso de atuadores [Rodrigues et al. 2018]. As características desses dispositivos variam de acordo com sua aplicação, protocolos de comunicação, mecanismos de coleta, mobilidade, custo, entre outros. Essa heterogeneidade de dispositivos possibilita o desenvolvimento de diferentes aplicações, ao mesmo tempo que as demandas por mais serviços de saúde requerem melhorias e novas funcionalidades nos dispositivos. A arquitetura IoHT segue um modelo definido em três camadas: percepção, rede e aplicação. A Figura 2.2 ilustra essa arquitetura e o papel de cada uma dessas camadas é discutido ao longo desta seção. Existem variações desta arquitetura que incluem um modelo estendido em mais camadas com o objetivo de explicitar algumas funcionalidades e tecnologias, como o uso de *middlewares*.

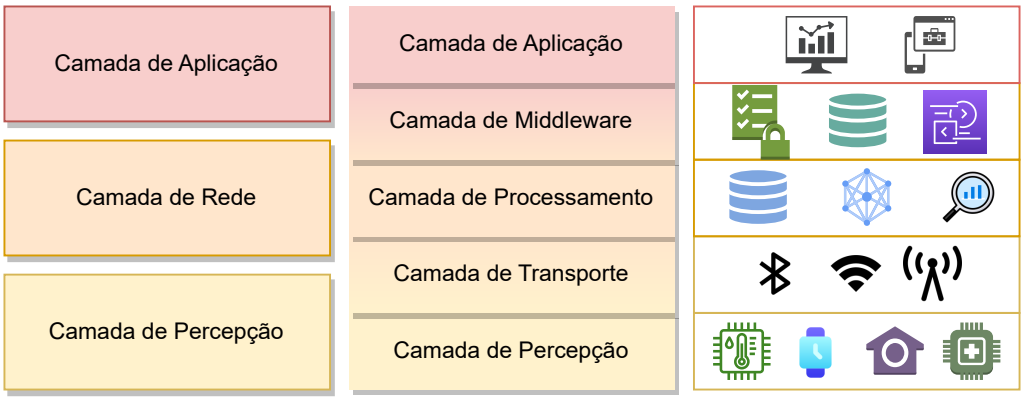

Figura 2.2. Representação em camadas da arquitetura IoT e seus componentes.

Na **camada de percepção**, estão localizados os sensores e atuadores. Os sensores associados à área de saúde podem coletar dados como sinais vitais, pressão sanguínea e temperatura corporal. Em fase de prototipagem, esses sensores comumente são associados às placas de Arduíno ou Raspberry Pi. Em cenários industriais, destacam-se os *smartwatches* como exemplo de equipamento vestível que possui sensores e que tem se tornado popular [Rodrigues et al. 2018]. Os sensores e dispositivos IoT podem variar em relação a inúmeros aspectos, como capacidade de processamento, protocolos de comunicação e papel que executa na arquitetura. Alguns desses dispositivos são considerados inteligentes e, além da coleta de dados, também executam algum tipo de processamento local. Em geral, os dispositivos comunicam-se entre si e/ou com um *gateway* para troca de informações e envio de dados para processamento remoto. O *gateway* é um dispositivo com mais recursos e autonomia energética que centraliza ações de coordenação e gerenciamento dos dispositivos, além de garantir a comunicação com equipamentos e serviços remotos. Nesse sentido, além de serem equipados com sensores, os dispositivos também possuem um ou mais mecanismos de comunicação.

Considerando a heterogeneidade dos dispositivos, foi estabelecida uma taxonomia baseada em recursos que os classifica em três categorias: classe 0, classe 1 e classe 2 [Bansal and Kumar 2020]. A classe 0 é composta pelos sensores e atuadores, que são dispositivos que geralmente têm poucos recursos e muitas limitações de segurança. Já

a classe 1, engloba elementos que podem atuar como *gateways* básicos, suportam múltiplos protocolos de comunicação e têm recursos superiores aos dispositivos de classe 0. Por fim, os dispositivos associados à classe 2 são mais robustos, possibilitam o uso de técnicas de IA e aprendizado de máquina (do inglês, *Machine Learning* – ML), dão suporte a vários protocolos de comunicação (e.g., Gigabit Ethernet, Wi-Fi e Bluetooth) e podem usar mecanismos de segurança mais avançados. Na Tabela 2.3 há um comparativo entre as classes de dispositivos com base na taxonomia de [Bansal and Kumar 2020].

Tabela 2.3. Classificação dos dispositivos IoT.

Classificação	Comunicação	Dispositivos	Características gerais
Classe 0 **Sensores** **Atuadores**	Protocolos *lightweight* (e.g., IEEE 802.15.4)	Telos Open mote Waspmote Tmote Sky	Muito vulneráveis às ameaças de segurança
Classe 1 ***Gateways***	Múltiplos protocolos (e.g., Wi-Fi)	Arduíno Netduino Tessel 2 ESP8266	Suporte à criptografia São *gateways* básicos em arquiteturas IoT
Classe 2 **Controladores**	Praticamente todos os tipos de protocolos	Raspberry Pi Orange Pi Banana Pi Cubieboard Beagleboard	Tem mais recursos de memória, processamento e armazenamento Suportam aplicações de IA, ML e processamento de linguagem São *gateways* mais complexos

Além da capacidade de recursos, os dispositivos IoT podem ser classificados e tratados com base em outros critérios. Por exemplo, quando um equipamento é classificado como dispositivo médico nos Estados Unidos, ele precisa ser aprovado pela FDA (*Food and Drug Administration*). Essas ações são essenciais para tentar minimizar a liberação de dispositivos que possam causar riscos à saúde do indivíduo ou desconforto devido a questões como tamanho, frequências de comunicação e aquecimento [Philip et al. 2021, Cornet et al. 2022]. Os dispositivos médicos têm buscado possibilitar a autonomia do indivíduo e a acurácia dos dados coletados. Em paralelo, os sensores têm sido utilizados para fornecer diferentes mecanismos de monitoramento: ECG, acelerômetro, SPo2, frequência cardíaca, etc [Philip et al. 2021]. Esses sensores estão embutidos em acessórios cotidianos, como pulseiras e relógios, para promover mais conforto físico e social ao indivíduo. Além disso, há a preocupação de que suas baterias sejam recarregáveis e fáceis de substituir, pois em aplicações de monitoramento contínuo, a inatividade do dispositivo e a inacurácia dos dados pode gerar diagnósticos incorretos [Cornet et al. 2022]. Tanto dispositivos médicos quanto equipamentos vestíveis e *smartphones* têm sido utilizados para acompanhamento de doenças crônicas através do monitoramento contínuo dos parâmetros de saúde do indivíduo [Perez et al. 2022].

O monitoramento contínuo depende de dispositivos que garantam a transmissão dos dados independente da localização do indivíduo. A *mHealth* refere-se ao uso de dispositivos móveis aplicados à área da saúde. Nessa área, é importante que os dispositivos suportem protocolos de comunicação de longa distância para encaminhar os dados em tempo real para a aplicação. Além disso, a importância do suporte à mobilidade também é importante pois a coleta de dados de mobilidade é um dos aspectos que mais contribui

para detecção na variação da saúde de pessoas idosas [Tun et al. 2021].

Para conseguir transmitir os dados coletados, os dispositivos precisam comunicar-se e trocar informações entre si. Portanto, é comum que dispositivos de classe 0 e classe 1 comuniquem-se e enviem dados para armazenamento e processamento. A conectividade entre os dispositivos ou entre a rede local e serviços remotos é feita com protocolos de comunicação de curta e longa distância. Esses protocolos são utilizados na transmissão dos dados e estão associados às funcionalidades da **camada de rede** na arquitetura representada na Figura 2.2. A escolha do protocolo de comunicação vai depender dos requisitos da aplicação e das características dos protocolos, que podem variar em banda de frequência, taxa de transmissão, latência e consumo energético. Além disso, também deve-se considerar aspectos como mobilidade, interoperabilidade e métricas de QoS. Informações sobre os protocolos que atuam na camada de rede e os principais requisitos associados à conectividade são discutidos nas Seções 2.3.2 e 2.4, respectivamente.

A camada de rede possibilita a troca de mensagens entre as camadas de percepção e a aplicação. A **camada de aplicação** é responsável por prover serviços para usuários finais (e.g., indivíduos, familiares e equipe médica). Como foi visto na Seção 2.2.1, as aplicações podem incluir monitoramento de parâmetros de saúde, gerenciamento de medicação, serviços de emergência, entre outras funcionalidades e podem ser realizadas a partir de diferentes domínios, como uso de vestíveis ou AAL. Os requisitos da camada de aplicação servirão como guia na definição de quais dispositivos e protocolos de comunicação devem ser utilizados. Além disso, aplicações mais robustas incluem serviços de computação em nuvem, desenvolvimento de *middlewares* e uso de técnicas de ML. Alguns desses tópicos serão abordados nas seções seguintes.

2.3.2. Protocolos de comunicação

Os dispositivos IoT e sensores utilizados em sistemas de vida assistida apresentam propriedades de comunicação variadas. Essa variação tem relação direta com a função executada pelo dispositivo e com as características dos dados coletados (e.g., volume dos dados e frequência de coleta). Na Tabela 2.4 estão listadas as propriedades de algumas aplicações de saúde em relação às características da comunicação, como frequência de banda, taxa de transmissão, latência e consumo energético. Uma das consequências dessas variações de requisitos é a utilização de diferentes protocolos de comunicação.

Tabela 2.4. Relação entre aplicações de saúde e características do meio de comunicação. Adaptado de [Cornet et al. 2022].

Aplicação	Bandas de frequência	Taxa de transmissão	Latência	Uso de energia
Voz	–	100kbps	< 250ms	–
Sensor de movimento	30-100Hz	4.8-35kbps	< 250ms	–
Streaming de vídeo	–	< 10Mbps	< 250ms	Alto
Pressão sanguínea	< 100Hz	< 10kbps	< 150ms	Alto
ECG	< 500Hz	3kbps	< 150ms	Alto
Streaming de áudio	–	1Mbps	< 250ms	Alto
Fluxo sanguíneo	< 40Hz	480kbps	< 150ms	Baixo
Temperatura	< 1Hz	120bps	< 150ms	Baixo
Nível de glicose	< 50Hz	< 1kbps	< 150ms	Muito baixo

Os protocolos de comunicação utilizados em sistemas de vida assistida geralmente são classificados em protocolos de curta ou longa distância. Os protocolos de curta distância estão associados, principalmente, a dois tipos de redes: redes de área corporal sem fio (do inglês, *Wireless Body Area Networks* – WBAN) e WSN. Os protocolos de longa distância geralmente são utilizados para possibilitar a comunicação entre os dispositivos locais e a nuvem, através da Internet. Os protocolos de curto e longo alcance variam em cobertura e frequência do sinal, consumo energético, largura de banda, mecanismos de comunicação, entre outros. Na Tabela 2.5 é realizada uma comparação entre esses diferentes protocolos, considerando suas principais características. Fazendo uma relação dessa tabela com a anterior, observa-se que para sensores com baixa frequência de sinal (e.g., temperatura), utilizar uma tecnologia como o Bluetooth Low Energy (BLE) é suficiente. Enquanto para sensores com frequência de sinal mais alta e que requerem mais largura de banda (e.g., *streaming* de áudio), é possível utilizar Wi-Fi [Philip et al. 2021]. Além disso, no caso de aplicações que são baseadas em monitoramento de vídeo ou voz, também é importante considerar métricas de rede, como latência.

Tabela 2.5. Tecnologias de comunicação sem fio usadas em sistemas de vida assistida. Fonte: [Nogueira et al. 2021].

	Tecnologia	Bandas de frequência	Alcance	Taxa de transmissão	Uso de energia
Curta distância	RFID	125 - 134 kHz, 13.56 MHz, 860 - 960 MHz	Até 100m	Depende da frequência	Muito baixo
	NFC	13.56 MHz	<0.2 m	Até 424 kbps	Muito baixo
	BLE (802.15.1)	2.4 - 2.48 GHz	Até 100m	Até 24 Mbps	Baixo
	Zigbee (802.15.4)	868 - 868.6 MHz, 902 - 928 MHz, 2.4 - 2.49 GHz	Até 100m	Depende da frequência	Muito baixo
	Wi-Fi (802.11a/b/g/n)	2.4 - 2.48 GHz, 4.9 - 5.8 GHz	20-250 m	2-600 Mbps	Médio
	Wi-Fi 5 (802.11ac)	4.9 - 5.8 GHz	Até 70m	Até 3.5 Gbps	Alto
	Wi-Fi 6 (802.11ax)	1 - 6 GHz	Até 120m	Até 9.6 Gbps	Alto
Longa distância	NB-IoT	Frequências da LTE	Até 15Km	Até 250 kbps	Baixo
	LTE-M	Frequências da LTE	Até 10Km	Até 1 Mbps	Baixo
	LoRa	867 - 869 MHz	Até 25Km	50 kbps	Muito baixo
	Sigfox	868–878.6 MHz	Até 40Km	100 bps	Muito baixo

Dentre os protocolos apresentados na Tabela 2.5, os protocolos NB-IoT, Lora e Sigfox foram criados especificamente para dispositivos IoT. A importância de criar protocolos específicos para IoT se dá, principalmente, porque esses protocolos consideram as características e limitações dos dispositivos. Nesse sentido, o padrão 3GPP (5G IoT) tem sido bem avaliado como alternativa para fornecer conexões celulares de baixa potência, baixa taxa de dados e ampla cobertura de sinal [Philip et al. 2021]. Além disso, o desenvolvimento de protocolos de longa distância que consumam baixa energia é essencial para o monitoramento contínuo de idosos. A proposta de sistemas de vida portátil [Nakayama et al. 2022] destaca a necessidade de utilização de mecanismos de comunicação que garantam a continuidade do serviço à medida que o indivíduo se movimente e ocupe diferentes espaços. Na proposta, os autores sugerem e avaliam o uso de protocolos de múltiplos caminhos como uma alternativa para prover resiliência na comunicação.

Ao considerar as comunicações de curta distância, há a predominância de três protocolos: BLE, Zigbee e Wi-Fi. Esses protocolos atuam, em geral, em uma frequên-

cia de banda de 2.4GHz, que é considerada o padrão para WBANs. No entanto, tem-se observado que essa frequência será insuficiente para atender os requisitos de comunicação de alta velocidade e tempo real exigidos por algumas aplicações [Cornet et al. 2022]. Além disso, como muitos serviços existentes utilizam também a frequência de banda de 2.4GHz, isso pode causar prejuízos na confiabilidade do funcionamento da rede de comunicação. Em decorrência disso, as frequências de onda milimétricas (do inglês, *milimeter wave* – mmWave) têm sido considerada uma alternativa para proporcionar altas velocidades a um custo baixo de processamento. Uma das principais limitações no uso de altas frequências é que estão mais propensas a sofrer com ruídos, reflexão e difração. No entanto, esse problema pode ser menos crítico para as comunicações de curto alcance. Além disso, há a possibilidade de usar métodos propostos para o funcionamento adequado das redes 5G, como o *beamforming*, que envia o sinal em apenas uma direção (a direção do destinatário) em vez de ser enviado em todas as direções.

A comunicação fim a fim em sistemas de vida assistida é caracterizada por múltiplos protocolos de comunicação sem fio ou cabeado, de curta e longa distância, cujas taxas de transmissão de dados e latência fim a fim são determinadas pelas taxas de transmissão e latências intermediárias. Assim, é preciso identificar e evitar possíveis gargalos na rede sem perder de vista que altas taxas de envio de dados e retransmissões entre os dispositivos podem significar um maior consumo energético. Os autores em [Rodrigues et al. 2018] ressaltam, ainda, a dificuldade em criar soluções de segurança completas que considerem as características de diferentes protocolos de comunicação, cabeados e sem fio. No entanto, apesar de a segurança dos canais de comunicação ser algo essencial para as aplicações de IoHT, é preciso assegurar que as soluções de privacidade não causarão sobrecarga nos canais de comunicação, principalmente aqueles com baixa capacidade, visto que isso pode ocasionar falhas ou erros no funcionamento da aplicação [Mukherjee et al. 2018]. Além disso, as soluções de segurança devem considerar os dispositivos e canais envolvidos na coleta (segurança de sensores e dispositivos em geral), transmissão (protocolos de comunicação), armazenamento e processamento de dados (computação em nuvem).

2.3.3. Computação em nuvem

A computação em nuvem tem sido utilizada no contexto de IoHT com duas funções principais: armazenamento e processamento de dados. O armazenamento pode ser realizado em nuvem pública, pessoal, privada, híbrida ou comunitária [Yang et al. 2020]. Existem vantagens e desvantagens no uso de cada tipo de armazenamento. Comparando, por exemplo, a nuvem pública com a privada, percebe-se que para manter uma nuvem privada é necessário investimento em infraestrutura física e profissionais para administrar os equipamentos e serviços. Isso gera um custo elevado de gerenciamento, mas aumenta as garantias relacionadas à confidencialidade e privacidade dos dados. Já com a utilização de nuvens públicas, os usuários pagam proporcionalmente ao serviço contratado e o gerenciamento é mais simples. No entanto, a principal preocupação no uso de nuvens públicas em IoHT está associada à segurança dos dados. Dentre os principais aspectos de segurança que devem ser observados ao armazenar dados em nuvem, é importante citar:

- **Confidencialidade, integridade e disponibilidade dos dados**: é preciso garantir que os dados não sejam alterados ou acessados por pessoas indevidas, e que os

dados enviados são exatamente os mesmos dados armazenados na nuvem. Além disso, os dados devem estar disponíveis sempre que necessário.

- **Controle de acesso granular e compartilhamento seguro de dados em grupos dinâmicos**: é preciso garantir políticas de controle de acesso com diferentes níveis de granularidade e flexibilidade. Assim, será possível compartilhar dados com diferentes partes de acordo com o contexto e fazer revogações de acesso, se necessário.
- **Privacidade e remoção completa dos dados**: a privacidade dos dados tem que ser definida de tal forma que profissionais que trabalham no provedor de nuvem não tenham acesso aos dados. E, caso a contratação do serviço de nuvem seja encerrada, os dados devem ser completamente excluídos.

Para garantir alguns requisitos de segurança, como confidencialidade e privacidade dos dados, geralmente são utilizados mecanismos criptográficos. Na Tabela 2.6, há um resumo dos mecanismos citados em [Yang et al. 2020]. A escolha do tipo de mecanismo utilizado vai ser determinada por diversos fatores e determina o uso de diferentes funções (e.g., busca em dado criptografado) e níveis de segurança.

Tabela 2.6. Características de protocolos criptográficos usados em ambientes de nuvem.

Características	Classificação dos algoritmos	Revogação de Acesso
Identity-Based Encryption		
-Mecanismo tradicional de Infraestrutura de Chave Pública -Confirma a identidade através de Autoridade Certificadora confiável -Utiliza a chave pública do usuário para criptografar os dados	-N/A	-Existem alternativas na literatura [Boneh and Franklin 2001, Li et al. 2013] -Em [Boneh and Franklin 2001], a chave pública é definida utilizando ID + período de validade
Attribute-Based Encryption		
-Utiliza um conjunto de atributos para criptografar os dados -Apenas usuários com os atributos conseguem acessar os dados -Permite controle de acesso granular	-Key Policy Attribute-Based Encryption -Ciphertext-Policy Attribute-Based Encryption	-Oferta esquemas de revogação indireta ou direta [Xu et al. 2019, Attrapadung and Imai 2009, Shi et al. 2015]
Homomorphic Encryption		
-Possibilita a realização de operações algébricas nos dados criptografados -A acurácia do resultado deve ser avaliada	-Partial Homomorphic Encryption -Somewhat Homomorphic Encryption -Full Homomorphic Encryption	-N/A
Searchable Encryption		
-Busca em dados criptografados -Adequada para cenários com quantidade limitada de palavras-chave e volume de dados reduzido	-Searchable Symmetric Encryption -Public Key Encryption with Keyword Search	-N/A

Os mecanismos criptográficos garantem requisitos de privacidade e confidencialidade no armazenamento dos dados. No entanto, em alguns casos os dados precisam ser descriptografados na nuvem para processamento. Como esse processo pode deixá-los suscetíveis a acessos indevidos, tem-se observado a utilização de ambientes confiáveis de execução (do inglês, *Trusted Execution Environment* – TEE), cujo objetivo é proteger os dados no momento do processamento através da garantia de requisitos de confidencialidade e integridade. No entanto, há alguns desafios no uso de TEE, pois esses ambientes têm acesso limitado à memória e os aceleradores de renderização de gráficos (e.g., GPU) e de tarefas de aprendizado profundo (e.g.,TPU) ainda não proveem esse serviço. Portanto,

o uso de TEE para processamento de dados provê um mecanismo de processamento mais seguro, mas aumenta o tempo de análise. Em [Narra et al. 2019], os autores propõem uma solução denominada Origami para endereçar os problemas de eficiência ao utilizar TEE para processamento de dados sensíveis. Na solução proposta, os dados criptografados são enviados para o TEE, onde são descriptografados e particionados. Em seguida, o Origami aplica uma técnica de blindagem criptográfica para ofuscar os dados e encaminhá-los para processamento em um ambiente não seguro, como uma GPU.

Além dos processos de armazenamento e processamento dos dados, as aplicações de saúde necessitam também de políticas de acesso que possibilitem o compartilhamento seguro das informações geradas. A possibilidade de utilizar serviços de nuvem para compartilhar informações de saúde do indivíduo entre profissionais de saúde, cuidadores e os próprios pacientes minimiza os riscos de perda de registros, mas pode adicionar riscos associados à segurança e privacidade dos dados [Dang et al. 2019]. Nesses casos, os processos criptográficos e de autenticação devem considerar aspectos como liberação e revogação de acesso a diferentes usuários, mecanismos cientes de contexto para lidar com situações atípicas e/ou de emergência e soluções dinâmicas e flexíveis para controle de acesso [Yang et al. 2020]. Por exemplo, um indivíduo pode consentir que seus dados sejam acessados por equipes médicas para obter um tratamento personalizado, mas pode restringir o uso dos dados para usos secundários, como pesquisas governamentais sobre uma determinada doença [Philip et al. 2021]. É importante ressaltar que o múltiplo acesso às informações pode ocorrer entre diferentes provedores de nuvem. Então, é preciso atentar-se à interoperabilidade de estratégias de armazenamento de dados e segurança entre provedores distintos. Todos esses pontos relacionados à segurança são fundamentais para viabilizar aplicações em IoHT. No entanto, as características da comunicação entre os dispositivos e a nuvem também impactam o desenvolvimento dessas soluções.

A comunicação entre a borda da rede (sensores e dispositivos IoT) e a nuvem pode apresentar diferentes latência, largura de banda e níveis de segurança. Como algumas aplicações na área de saúde tem requisitos mais rigorosos em relação à conectividade e segurança, a computação em névoa tem sido usada como solução em algumas arquiteturas. Em [Aazam et al. 2020], os autores sugerem o uso de computação em névoa como uma espécie de *middleware* entre a borda e a nuvem. Os autores argumentam que a distância entre os servidores da nuvem e os equipamentos de borda limita o desenvolvimento de soluções de tempo real em IoHT. Além disso, os autores também argumentam que com o uso de computação em névoa, é possível prover um monitoramento mais individualizado dos dispositivos IoT, bem como estabelecer mecanismos de segurança e privacidade mais adequados às aplicações. É importante ressaltar, entretanto, que os dispositivos localizados na névoa não têm a mesma capacidade de armazenamento e processamento da computação em nuvem. Além disso, como os dispositivos estão mais próximos dos indivíduos ou instituições, em caso de desastres, é possível que ambos (borda e névoa) fiquem indisponíveis. Assim, fica clara a necessidade de usar uma arquitetura híbrida para possibilitar o desenvolvimento de aplicações que tenham requisitos de QoS mais estritos e para garantir premissas de segurança, como armazenamento de backup em locais distantes. Para realizar a integração e segurança entre os inúmeros componentes das aplicações baseadas em arquiteturas híbridas envolvendo borda, névoa e nuvem, comumente são utilizados *middlewares*.

2.3.4. *Middlewares*

Middlewares são *softwares* que atuam entre a aplicação e o sistema operacional e/ou rede. No contexto de IoHT, os *middlewares* são utilizados para endereçar requisitos da infraestrutura e da aplicação [Zgheib et al. 2019]. Em relação aos requisitos de infraestrutura, destaca-se a interoperabilidade entre os dispositivos, escalabilidade (quantidade de dispositivos), interações entre os equipamentos e diversidade da infraestrutura de comunicação. Já em relação aos desafios associados à aplicação, o foco é na disponibilidade e confiabilidade do serviço, tolerância a falhas, monitoramento em tempo real e aspectos de segurança e privacidade. Além disso, há uma necessidade de desenvolvimento de *middlewares* semânticos, visto que o uso de semântica pode auxiliar na redução de número de sensores utilizados e no desenvolvimento de aplicações mais inteligentes.

Os *middlewares* associados à saúde podem ser classificados em sete categorias: baseado em névoa, em modelo *publish/subscribe*, na Web das Coisas, em arquitetura orientada a serviço, orientado a eventos e orientado a mensagens [Fersi 2020]. A partir dessa classificação, os autores consideram a névoa como um *middleware* entre a borda e a nuvem, que possibilita alcançar requisitos de aplicação, como baixa latência. De acordo com essa classificação, não há um único tipo de *middleware* adequado para todas as aplicações de IoHT, portanto é possível combiná-los para tentar endereçar os requisitos da aplicação. Além disso, os *middlewares* de segurança devem ter constantes atualizações, para garantir eficácia à medida que novos mecanismos de ataque são identificados.

Pesquisadores da Universidade da California - Irvine (UCI) desenvolveram um *middleware* denominado Tippers[2] para gerenciamento de dados de sensores em espaços inteligentes. O Tippers consegue reduzir a complexidade do desenvolvimento de aplicações ao funcionar como um concentrador de dados de fontes variadas. A solução é considerada escalável e de fácil adaptação, uma vez que permite que novos sensores e tipos de sensores sejam adicionados sem que haja mudança no código da aplicação. Além disso, o Tippers também considera a implantação de estratégias de gerenciamento de dados que possibilita a integração de técnicas que garantam a privacidade dos usuários. Esse *middleware* tem sido utilizado no desenvolvimento de várias soluções que consideram questões como conectividade e privacidade dos usuários [Lin et al. 2020, Mehrotra et al. 2020, Ghayyur et al. 2020]. Atualmente, o Tippers tem sido usado no contexto do projeto CareDEX[3], que se trata de uma plataforma inteligente que tem como objetivo melhorar o tratamento de pessoas idosas, em cenários de desastre, através da troca segura de informações entre equipes de primeiros socorros, cuidadores e idosos em instituições de repouso. O Tippers também tem sido utilizado em outros campi para apoiar serviços de localização. O principal requisito endereçado por esse *middleware* é a interoperabilidade entre os dispositivos e transparência para a aplicação.

Em [Madureira et al. 2019] os autores desenvolveram um *middleware*, denominado My-AHA, com o objetivo de integrar soluções de saúde para pessoas idosas de forma segura. Os autores sinalizam que a maioria dos *middlewares* desenvolvidos são focados na interoperabilidade de dispositivos e eles expandem essa interoperabilidade para as soluções. O My-AHA é composto por um conjunto de conectores responsáveis por

[2] https://tippers.ics.uci.edu/
[3] https://sites.uci.edu/caredex/

coletar os dados disponibilizados pelas plataformas externas. Essa função é executada rotineiramente para que a aplicação tenha uma versão atualizada dos dados. A arquitetura da solução também é composta por um mecanismo de autenticação e autorização. Os usuários devem conceder acesso ao My-AHA, para que os conectores consigam consultar novos dados. Os dados advindos de outras plataformas são armazenados de forma anonimizada em uma base de dados do My-AHA. Os usuários podem interromper o compartilhamento de dados e solicitar que os dados enviados anteriormente sejam removidos da plataforma. Por fim, o My-AHA utiliza um mecanismo *publish/subscribe* para garantir o acesso dos dados coletados para diferentes partes. Esse *middleware* endereça requisitos como segurança e interoperabilidade de dispositivos e plataformas.

Em [Mukherjee et al. 2018], os autores propõem um *middleware* para prover segurança de forma flexível para aplicações de saúde. A proposta é promover um mecanismo de segurança flexível, que atenda os requisitos de segurança de cada aplicação, baseado nos recursos dos dispositivos na borda e na nuvem. Para identificar a melhor abordagem de segurança que deve ser utilizada, os autores consideram a intermitência da rede e restrições dos dispositivos (e.g., energia, computação, armazenamento, etc). De acordo com essas informações, é definido um esquema de segurança que inclui mecanismos de autenticação, criptografia dos dados e códigos de autenticação de mensagem. Caso haja problemas de intermitência na comunicação de rede, a solução faz uso de algoritmos de retomada de sessão, que reutilizam sessões criptografadas anteriormente por um mesmo dispositivo para retomar a conexão que foi interrompida. Além disso, o *middleware* também é responsável por determinar qual o melhor esquema de segurança para um cenário, considerando os requisitos da aplicação e as características dos dispositivos.

2.3.5. Aprendizado de máquina

O uso de técnicas de ML é comum nas aplicações e serviços de saúde para correlacionar e analisar os dados coletados pelos sensores, tornando possível a identificação de padrões, predição de eventos e sugestão de melhorias no tratamento dos indivíduos [Maskeliūnas et al. 2019, Wang et al. 2022]. Como os dispositivos IoT possuem restrições energéticas e baixa capacidade de armazenamento e processamento, os dados são enviados para serem armazenados e processados em outro local. Assim, as técnicas de ML podem ter funções variadas no contexto de saúde, como a análise de dados de conectividade que contribuem com o monitoramento do indivíduo ou com o desenvolvimento de novas técnicas de aprendizado que considerem requisitos de privacidade.

Ao executar as funções básicas de conectividade, os dispositivos de rede geram informações que podem beneficiar as aplicações. Em [Lin et al. 2020], os autores propõem a utilização de dados de conectividade Wi-Fi para localizar indivíduos em áreas internas de um espaço físico, como salas dentro de um prédio. Eles aplicam técnicas de pré processamento para melhorar a qualidade dos dados e aplicam uma técnica de aprendizado semi supervisionado e método probabilístico para realizar uma identificação semântica da localização do indivíduo. O uso desse sistema tem sido explorado no projeto CareDEX para identificar a localização de residentes dentro de instituições de repouso. Com a localização semântica dos residentes, é possível usar a aplicação para auxiliar na evacuação de idosos em cenários de emergência, bem como para identificar mudanças na rotina dos indivíduos. Nessa solução, os autores utilizaram um mapeamento entre identificador do

usuário, endereço MAC do dispositivo vestível, conexões dos dispositivos com os pontos de acesso e localização dos pontos de acesso na construção da solução. Além das informações utilizadas pelos autores, a análise de dados disponibilizados pelos equipamentos de rede e coletados por meio de protocolos como o *Simple Network Management Protocol* (SNMP) podem auxiliar as aplicações. Por exemplo, a análise automatizada de informações como frequência de *reboots* e aumento da temperatura do equipamento, quando associadas às informações de contexto, como localização dos dispositivos e horário, podem auxiliar na identificação e predição de problemas no local.

Do ponto de vista de segurança, embora o uso de TEE possa garantir algum nível de confidencialidade e privacidade dos dados, há casos em que a centralização de dados para processamento em um único local não é viável ou recomendada. Os dados dos indivíduos e de diferentes instituições podem ser utilizados, juntos, para avaliar e prover políticas de saúde para a população. No entanto, o compartilhamento desses dados é regulamentado por diferentes legislações e/ou termos de privacidade e o envio dos dados para processamento centralizado pode violar premissas de privacidade. Portanto, tem-se observado o uso de aprendizado federado, que permite que diferentes instituições compartilhem um modelo de aprendizado global enquanto mantém seus dados localmente, preservando requisitos de privacidade e confidencialidade. A arquitetura proposta para uso do aprendizado federado por ser vista na Figura 2.3. As instituições médicas atuam como nós de borda que executam localmente um modelo de aprendizado e deve encaminhá-lo periodicamente para o nó agregador. Esse nó agregador recebe os modelos de cada instituição e cria um modelo de treinamento global. O modelo de treinamento global é enviado novamente para as instituições. Durante esse processo, os dados são mantidos nas instituições e apenas os parâmetros e pesos utilizados no modelo são encaminhados entre os nós [Konečnỳ et al. 2016]. Esse método traz mais garantias de privacidade mas os resultados têm menos acurácia. Além disso, apesar de a solução não encaminhar dados médicos entre os nós, é importante avaliar se os parâmetros e pesos encaminhados entre os nós podem gerar vazamentos em relação à privacidade.

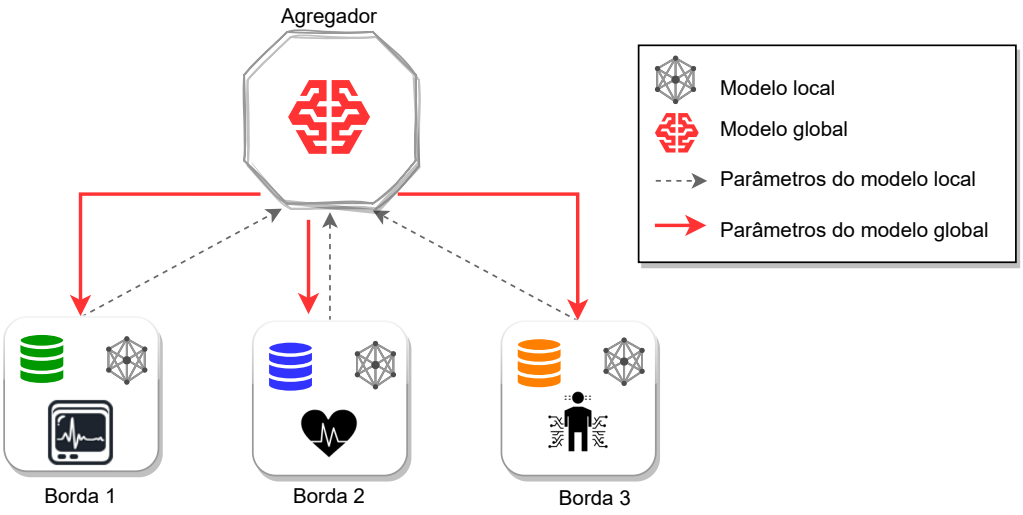

Figura 2.3. Representação do funcionamento da técnica de aprendizado federado.

Os autores em [Du et al. 2020] fazem uma análise de diferentes técnicas de aprendizado de máquina em relação a aspectos como distribuição de dados, acurácia, comunicação e privacidade. O aprendizado centralizado é aquele que possui maior acurácia, pois todos os dados estão no mesmo local. No entanto, isso significa mais riscos à privacidade e maior consumo de banda, visto que todos os dados devem ser enviados para um nó centralizador. O aprendizado federado é o que menos consome banda e mais preserva a privacidade, mas tem menor acurácia. Assim, a solução de aprendizado par-a-par pode ser considerada um meio termo entre essas soluções, pois também preserva a privacidade do usuário e tem uma acurácia moderada. O aspecto negativo é um consumo de banda superior às técnicas de aprendizado federado. Em suma, é preciso avaliar quais tipos de técnicas de ML são adequadas a cada aplicação considerando seus requisitos.

2.4. Requisitos de Comunicação e Segurança da Informação

A adoção de soluções de sistema de vida assistida dependem diretamente do tratamento dos requisitos dos usuários e das aplicações. Como a comunicação e a segurança são aspectos essenciais na viabilidade desses sistemas, a forma de tratamento desses requisitos é descrita nesta seção. Inicialmente, serão discutidos os requisitos gerais, e, em seguida, os principais requisitos em termos de conectividade, mobilidade, QoS e QoE. Em seguida, serão abordados os aspectos relacionados à segurança, considerando princípios como disponibilidade, integridade e confidencialidade.

2.4.1. Requisitos Gerais e de Comunicação

Em geral, os requisitos dos sistemas de vida assistida estão associados a três etapas: *(i)* coleta dos dados e transmissão até um ponto de acesso, *(ii)* transmissão dos dados através das redes de acesso até a Internet e *(iii)* a integração das informações coletadas e da transmissão de dados via comunicação fim-a-fim. A Figura 2.4 ilustra essas etapas da comunicação em sistemas apoiados pela IoHT e as principais entidades envolvidas.

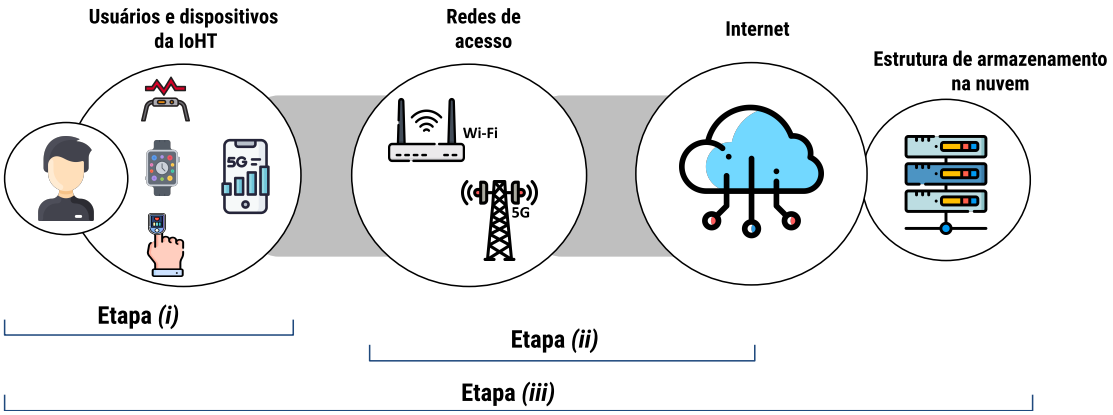

Figura 2.4. Etapas da comunicação em sistemas baseados na IoHT.

Considerando inicialmente a **comunicação entre um dispositivo de coleta de dados e um ponto de acesso**, os principais requisitos estão relacionados com os dispositivos e com os protocolos de comunicação de curto alcance. Em relação aos dispositivos, tem-se os seguintes requisitos: os dispositivos de coleta de dados devem funcionar de maneira

ininterrupta para aplicações críticas (e.g., monitoramento da frequência cardíaca), deve haver compatibilidade entre os dispositivos, os dispositivos devem alcançar um ciclo de vida de bateria longo e utilizar baterias que sejam facilmente recarregadas ou substituídas. Além disso, os protocolos de comunicação devem considerar a eficiência energética e os dispositivos devem ter recursos de armazenamento para evitar perda de dados nos casos de instabilidade da rede ou desconexões. Por fim, o sistema deve gerenciar a entrada e saída de nós de forma automática e seus componentes devem ser facilmente substituídos e atualizados, favorecendo requisitos de gerenciamento, manutenibilidade e escalabilidade.

A segunda etapa de comunicação está relacionada com **o acesso da rede interna à Internet**. A comunicação entre os pontos de acesso e a Internet pode ocorrer de maneiras distintas nos sistemas de vida assistida, dependendo do nível de mobilidade envolvido. Os sistemas mais tradicionais usam dispositivos estáticos e pontos de acesso fixos atrelados a redes estruturadas. No entanto, esse mecanismo de conexão tem cobertura de sinal limitada e restringe a mobilidade do usuário. Assim, é comum o uso de redes móveis, como as redes 4G/5G, que permitam a comunicação de dispositivos portáteis ou vestíveis à Internet mesmo quando o usuário está em movimento. Independente do tipo de rede utilizada, o principal requisito dessa etapa é que os dados coletados cheguem até a infraestrutura que irá armazená-los. Para garantir essa premissa, é importante que haja redundância nos canais ou tecnologias de comunicação e redundância de dispositivos de rede, como os pontos de acesso. Embora essas garantias de redundância possam ser difíceis de serem alcançadas em cenários domésticos devido às restrições de custo ou mesmo por falta de conhecimento técnico, essas alternativas devem ser consideradas imprescindíveis em instituições de saúde como hospitais ou casas de repouso para idosos.

A etapa da **comunicação fim-a-fim** compreende todo o caminho percorrido pelos dados. Os requisitos desta etapa incorporam as duas etapas anteriores e todos os elementos envolvidos. No entanto, mesmo atendendo os requisitos citados anteriormente, ainda há problemas que podem impactar as aplicações. Por exemplo, mesmo que haja redundância e que os problemas de compatibilidade sejam sanados, ainda pode haver variações no desempenho da comunicação. Neste caso, torna-se necessária a redundância ou variação nas tecnologias de comunicação que garantam a transmissão das informações em tempo hábil e de forma íntegra. Alguns protocolos, como o BFD (*Bidirectional Forwarding Detection*), são utilizados para detecção de falhas e mudanças de rota de forma proativa. Esses protocolos geralmente funcionam adequadamente em casos de desconexões, realizando a migração para o caminho redundante, mas podem apresentar comportamento indesejado em casos de instabilidade de enlaces: perdas de pacotes no enlace principal podem gerar mudanças frequentes entre as rotas principal e secundária. Por fim, as mesmas garantias de disponibilidade sinalizadas anteriormente devem ser ofertadas considerando o destino final da informação, seja o armazenamento em um dispositivo remoto ou na nuvem. O não cumprimento desses requisitos pode implicar em dados corrompidos, ausentes ou entregues com alto atraso, inviabilizando o uso de algumas aplicações.

Outros requisitos relacionados à infraestrutura, mas complexos de serem alcançados incluem: redundância de dispositivos de coleta de dados, multiplicidade de dispositivos com mesma funcionalidade e tecnologias de comunicação distintas, pontos de acesso que suportam múltiplas tecnologias de comunicação de curto alcance e atendam diversos dispositivos, mecanismos de proteção contra a interferência cruzada causada pela vari-

edade de tecnologias de comunicação, interoperabilidade de dispositivos proprietários e sua integração transparente ao ambiente. A dificuldade em alcançar esses objetivos ocorre por duas razões principais: falta de integração entre as etapas da comunicação e excesso de padrões existentes e em desenvolvimento para cada uma delas. A variedade de padrões dificulta a comunicação em cenários que diferentes dispositivos se comunicam entre si em busca de oferecer serviços mais robustos aos usuários, como é o caso de sistemas de vida assistida. Alguns dispositivos IoT suportam protocolos M2M, como o CoAP, MQTT e AMQP, mas é preciso ainda garantir a interoperabilidade entre esses protocolos e dispositivos médicos que são baseados em outros protocolos de troca de mensagem.

Alguns requisitos dos sistemas de vida assistida estão relacionados à adoção dos dispositivos e do sistema pelos usuários. De uma forma geral, os usuários buscam requisitos como simplicidade de uso, adaptabilidade, uso de contexto, intuitividade, utilidade e confiabilidade. Além disso, muitos dispositivos são posicionados no corpo do indivíduo e devem ser anatômicos e não causar desconforto ao usuário. A não garantia desses requisitos pode gerar fracasso na adoção das soluções. Assim, dispositivos portáteis e vestíveis com interfaces de toque na tela devem ter operação intuitiva e facilitada para que pessoas de diferentes faixas etárias consigam utilizá-las. As funcionalidades e os dispositivos do sistema devem ser expostos ao usuário para que o mesmo não tenha dúvidas sobre como e com qual finalidade os dados serão utilizados. Em relação aos dados coletados, para garantir suporte aos sistemas de vida assistida, é importante endereçar requisitos como veracidade, integração, privacidade, confiabilidade e segurança. Uma discussão mais aprofundadas desses tópicos será realizada na Seção 2.4.4.

2.4.2. Mobilidade

A mobilidade nos sistemas de vida assistida pode ocorrer de duas formas: de maneira mais tradicional como nas aplicações AAL, cujo objetivo é fornecer mobilidade em um espaço delimitado e restrito por um tecnologia sem fio de curto alcance, como o Wi-Fi; ou em uma vertente mais atual, com uso de sistemas de vida assistida portáteis que visam oferecer uma mobilidade maior ao usuário, empregando tecnologias de comunicação de longo alcance e permitindo uma maior área de cobertura e comodidade para o usuário. Os sistemas de vida assistida portáteis promovem uma maior mobilidade e qualidade de vida para o usuário ao criar um ambiente mais imersivo e integrado em termos de dispositivos e tecnologias de comunicação. Isso é possível pois os dispositivos da IoHT são dotados de tecnologias de comunicação cada vez mais eficientes e versáteis. Os *smartphones* incorporam processadores de alta capacidade e complexidade computacional e oferecem as tecnologias de comunicação mais recentes. Em geral, a configuração inicial dos sistemas de vida assistida portáteis usa um *smartphone* como coordenador dos dispositivos de coleta de dados. Esses dispositivos confiam no coordenador para executar as funções que demandam mais processamento e memória, e para garantir o acesso à Internet. Essa configuração é cômoda para o usuário uma vez que grande parte da população possui um *smartphone* para executar tarefas básicas e se conectarem à Internet quando estão em movimento. Além disso, esses equipamentos contam com tecnologias de curto alcance como Bluetooth e NFC para a sincronização de dispositivos.

Considerando a miniaturização dos componentes e a inclusão de novas tecnologias de comunicação em equipamentos vestíveis, uma nova configuração para os dispositivos

da IoHT começa a surgir: os dispositivos estão sendo disponibilizados nativamente com conexões de longa distância, como a LTE. Isso modifica a arquitetura tradicional baseada em um dispositivo coordenador e traz mais liberdade para os usuários pois sensores e equipamentos vestíveis com tecnologias de longo alcance permitem a conexão constante com a Internet e dispensam o uso de um coordenador ou um ponto de acesso fixo.

2.4.3. Qualidade de Serviço e Qualidade de Experiência

Sistemas baseados em IoT conectam, além de pessoas, vários dispositivos que, de maneira ativa ou passiva, compõem o sistema. Para integrar esses componentes, empregam-se diferentes protocolos e tecnologias de comunicação com capacidades distintas de oferecer QoS. Entretanto, para garantir que as aplicações apoiadas nesses sistemas funcionem conforme o esperado torna-se necessário enfrentar os desafios em termos de variação de desempenho e heterogeneidade de tecnologias. Na Tabela 2.7 há uma lista de dispositivos empregados na IoHT, as tecnologias de comunicação disponíveis e os respectivos requisitos para comunicação relacionados com algumas métricas de QoS e QoE. Nota-se que dispositivos com mais recursos apoiam mais aplicações e possuem requisitos mistos, enquanto dispositivos de transmissão de áudio e vídeo têm requisitos estritos em termos de latência e capacidade, e sensores específicos para aplicações de saúde requerem altos índices de confiabilidade e baixa latência devido à criticidade das informações.

Tabela 2.7. Requisitos de comunicação para os dispositivos da IoHT e suas tecnologias. Adaptado de [Devi et al. 2023].

Dispositivo	Tecnologias	Requisitos para Comunicação		
		Latência	Capacidade	Confiabilidade
Smartphone	LTE, Bluetooth, mmWave Celular, WLAN	Regulares	Regulares/Altos	Regulares
Relógio ou Óculos Inteligente	LTE, Bluetooth	Regulares	Regulares	Baixos
Disp. de Realidade Virtual e Realidade Aumentada	mmWave Celular, WLAN	Altos	Altos	Baixos
Tablet	LTE, Bluetooth, mmWave Celular, WLAN	Regulares	Regulares/Altos	Regulares
Roupa ou Tênis Inteligente	Zigbee, Bluetooth	Baixos	Baixos	Baixos
Sensores Médicos	LTE, Bluetooth	Altos	Baixos	Altos

Garantir QoS em sistemas de vida assistida é um grande desafio devido às variações de desempenho das tecnologias de comunicação empregadas. Essas variações ocorrem devido às especificações heterogêneas das tecnologias de comunicação, aos múltiplos enlaces que podem ocasionar gargalos na comunicação e à multiplicidade de aplicações concorrentes que utilizam o sistema. A falha ao atender os requisitos de QoS tem impactos negativos na percepção do usuário ao utilizar as aplicações que requerem interação humana, ocasionando baixa QoE. A ligação entre QoS e QoE é evidente, mas existe um desafio ao aferir a QoE para as aplicações nos sistemas de vida assistida que não exigem participação específica de um usuário. Embora os cumprimentos dos requisitos de QoS sejam um bom parâmetro, eles não garantem uma boa experiência para o usuário final.

Existem diversos trabalhos na literatura que exploram o levantamento de requisitos de QoS para aplicações apoiadas na IoT tradicional [Adil et al. 2022]. Entretanto,

os sistemas de vida assistida demandam alterações na identificação dos requisitos e nas implementações em nível operacional para o cumprimento dos mesmos. A Tabela 2.8 ilustra algumas diferenças fundamentais entre a IoT e a IoHT. As características da IoHT em termos de criticidade das aplicações de saúde, necessidade de mobilidade, resiliência das informações e eficiência energética são os principais pontos que diferenciam a IoHT e a IoT e que demandam requisitos mais estritos. Apesar dessas diferenças serem significativas, poucos trabalhos disponíveis na literatura abordam a QoS e QoE em IoHT.

Tabela 2.8. Particularidades operacionais da IoT e IoHT. Adaptado de [Adil et al. 2022].

Particularidades da IoT tradicional	Particularidades da IoHT
Aplicações para diversas finalidades	Aplicações relacionadas à saúde
Aplicações com requisitos variados	Aplicações com requisitos estritos
Aplicações não demandam alta precisão nos resultados	Requer resultados precisos e específicos
Facilidade de implementação e configuração	Precisão depende da correta configuração
Falhas não impactam o sistema de forma crítica	Falhas colocam em risco a saúde do usuário
Tráfego de dados intermitente e volumoso	Tráfego crítico e eventual
Eficiência energética impacta o custo do sistema	Eficiência energética impacta a qualidade de vida

Considerando as particularidades da IoHT e observando os requisitos de comunicação e segurança nos sistemas de vida assistida, encontramos os seguintes desafios ao propor níveis de QoS que satisfaçam a necessidade das aplicações: desafios relacionados à coleta dos dados, às arquiteturas de rede, à segurança, à interoperabilidade e à escalabilidade. A coleta dos dados requer rapidez e acurácia por parte de sensores e dispositivos, para que isso aconteça uma série de outros requisitos devem ser cumpridos sob risco da incidência de atrasos e informações faltantes. Sendo assim, os desafios na etapa da coleta de dados acabam se estendendo para a correta instalação e uso dos dispositivos por parte dos usuários ou prestadores de serviço, pois o atraso por falhas ou uso inadequado de um dispositivo irá se propagar para os serviços dependentes da informação. A necessidade da informação sempre à disposição impacta ainda os dispositivos alimentados por baterias e a capacidade dos canais de comunicação estarem sempre disponíveis.

Uma das formas de lidar com os problemas relacionados à garantia da QoS em ambientes IoHT envolve o estudo de protocolos de roteamento mais eficientes. Entretanto, a segmentação das redes em conjunto com o uso de tecnologias proprietárias, principalmente nas redes dos dispositivos de coleta de dados, causam problemas de interoperabilidade e dificultam o cumprimento dos requisitos de QoS. Outros problemas que afetam a entrega de QoS incluem: mecanismos proprietários de criptografia dos dados e autenticação dos usuários, formatos específicos de compressão dos dados que impactam no cálculo do atraso, dificuldade em estimar os tempos de gravação e recuperação dos dados em ambientes distintos como nuvem, névoa, bordas ou em um dispositivo específico.

A avaliação de QoE depende da interpretação do usuário, do desempenho do sistema e do contexto de utilização. Geralmente, aplicações com um fluxo constante de dados tem maior impacto na percepção do usuário. Aplicações que envolvam áudio e vídeo permitem que os usuários detectem falhas mais facilmente e estão presentes na telemedicina e no monitoramento de pacientes em tempo real. No entanto, a interconexão exclusivamente entre dispositivos é uma das características da IoHT, incluindo a ausência completa de interação humana em diversas etapas do ciclo da informação. Isso traz difi-

culdade para mensurar a QoE uma vez que os dispositivos não têm a mesma percepção de um usuário e não é possível confiar somente nos requisitos da QoS devido aos problemas relacionados à heterogeneidade do ambiente. Nesse sentido, algumas propostas para prover a QoS de forma autônoma a partir de diversos parâmetros de configuração e requisitos, começam a surgir na literatura. No trabalho proposto em [Bardalai et al. 2022], os autores empregam modelos de aprendizado de máquina para auxiliar na tomada de decisão, fortalecendo o provisionamento tanto de QoS quanto de QoE.

Existe uma diversidade de pontos em aberto envolvendo QoS e QoE em ambientes complexos, como a IoHT. Os esforços estão direcionados para resolver os problemas em cada etapa da comunicação e os desafios são numerosos. Ainda assim, mesmo que um segmento específico consiga cumprir todos os requisitos e garantir os requisitos de QoS isoladamente, ainda existe o problema relacionado à segmentação dos serviços, como o atraso e as falhas em cascata, onde um problema original causa transtornos nos serviços subsequentes. Uma solução seria a possível orquestração dos serviços dependentes, mas essa integração é complexa uma vez que em geral não existe controle sobre os enlaces mais distantes do usuário como as redes de acesso e a própria nuvem e seus canais de comunicação. Nesse cenário, a QoE torna-se um parâmetro avaliativo útil porém sem uma indicação específica que contribua para redefinir os requisitos de QoS.

2.4.4. Segurança da Informação

Os princípios básicos de segurança aplicados aos sistemas tradicionais como: disponibilidade, integridade, confidencialidade e privacidade, são ainda mais importantes em sistemas de vida assistida devido à sensibilidade das informações coletadas e distribuídas. Em geral, os dados existentes nesses sistemas representam informações pessoais e privadas. Sendo assim, sua disseminação não autorizada pode ter consequências graves para os usuários do sistema, prestadores de serviço e fabricantes de dispositivos. Adicionalmente, as informações armazenadas e transmitidas devem estar imunes de manipulação por entidades não autorizadas e disponíveis sempre que uma entidade autorizada necessitar acessá-las. Finalmente, não deve ser possível identificar um usuário do sistema através de dados coletados nos canais de comunicação ou armazenados em dispositivos e bases de dados, sem a devida autorização. A Figura 2.5 ilustra os princípios básicos de segurança nos sistemas de vida assistida e as principais ameaças a cada um deles.

A segurança de um sistema, como um todo, é proporcional às garantias de segurança do elo mais fraco e os princípios de segurança apresentam relações entre si. A partir da Figura 2.5, é possível observar relações de interdependência entre os princípios de segurança e a diversidade de alvos em relação aos tipos de ataque. Por exemplo, ataques que afetam a autenticação e autorização de dispositivos podem comprometer a integridade do sistema como um todo. Neste sentido, tem-se o desafio de desenvolver e usar soluções de segurança que sejam integráveis ou interoperáveis para prover uma alternativa de segurança que contemple todos os componentes do sistema. Durante esse processo, deve-se avaliar os recursos dos componentes e as alternativas que podem ser utilizadas. Em paralelo, esses mecanismos de segurança devem ainda atender às regulações de privacidade – que podem variar em diferentes países e estados – e proporcionar políticas de controle de acesso que podem variar em diferentes níveis (e.g., borda, névoa e nuvem) e contextos (e.g., emergências e desastres). Alguns desses pontos serão discutidos nas próximas

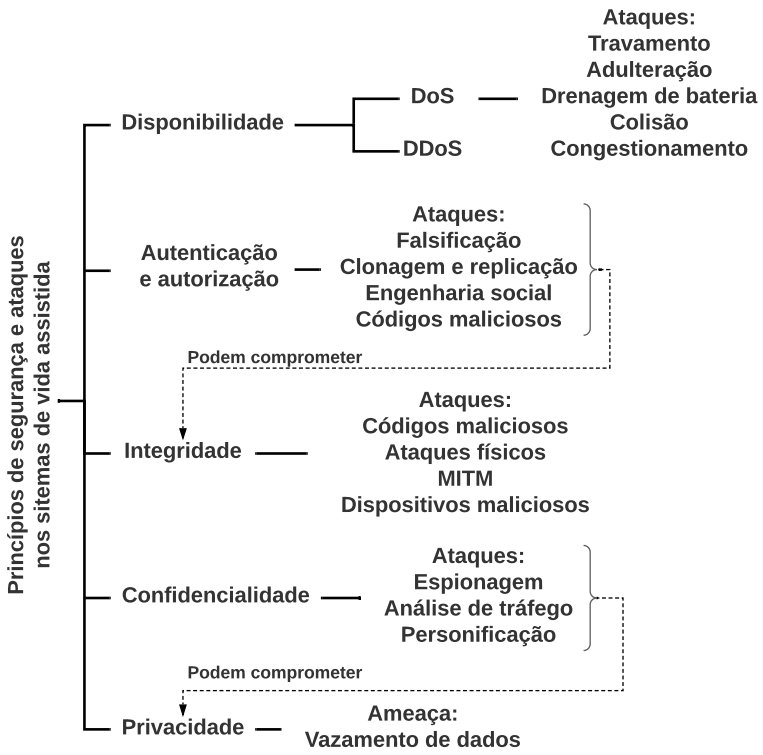

Figura 2.5. Princípios básicos e ameaças de segurança em sistemas de vida assistida.

seções, onde serão explorados os requisitos associados à segurança em IoHT.

2.4.4.1. Disponibilidade

O princípio da disponibilidade está associado com o acesso às informações e visa assegurar que os dados e serviços estejam disponíveis, quando necessário, e que os serviços prestados não sejam negados a nenhum usuário ou subsistema autorizado. Nos sistemas de vida assistida, a disponibilidade das informações e serviços está diretamente relacionada com a confiabilidade do sistema pelos indivíduos, familiares e equipe médica. Isso se dá, principalmente, porque interrupções na oferta de serviços ou a indisponibilidade dos dados requisitados pode ocasionar situações que coloquem o usuário em perigo, incluindo casos de risco à vida. Nestes cenários, os riscos associados à disponibilidade e as estratégias para mitigá-los devem considerar as características dos elementos do sistema: sensores e dispositivos IoT, canais de comunicação e serviços de nuvem.

Todas as soluções de comunicação e segurança para os sensores e dispositivos IoT precisam respeitar a limitação dos recursos, como tempo de bateria, baixo poder computacional e memória. Desta forma, os processos que são executados nos dispositivos precisam ser simples e ter baixo consumo energético. A preocupação com o consumo energético impacta na relação entre a disponibilidade e outros princípios de segurança, como confidencialidade. Para garantir confidencialidade, geralmente são utilizados mecanismos criptográficos, mas o processo de criptografar os dados coletados pelos sensores antes do envio gera a necessidade de mais processamento nos dispositivos. Esse processamento adicional implica em consumo energético, que pode diminuir o tempo de vida

da bateria e afetar a disponibilidade.

Outro aspecto fundamental relacionado à disponibilidade é a necessidade de monitoramento e detecção de falhas. A variação nas características de coleta e envio de dados pelos sensores e dispositivos IoT pode gerar a necessidade de uso de diferentes mecanismos de monitoramento. O monitoramento e a detecção de falhas é mais simples em dispositivos que realizam coletas e envio de dados constantes ou em uma frequência predefinida, visto que a própria falta de recebimento dos dados pela aplicação pode indicar uma falha na rede ou no equipamento. No entanto, alguns dispositivos são programados para enviar informações apenas em situações específicas. Para esses casos, torna-se imprescindível a utilização de alternativas de monitoramento, como o envio de *heartbeats*. Nos casos em que é necessário gerar tráfego adicional para realizar o monitoramento, é importante considerar o *tradeoff* com o consumo energético e ajustar o algoritmo de acordo com as necessidades do sistema. Por exemplo, seria importante definir um intervalo ótimo entre o envio dos *heartbeats* e o consumo energético, de forma que fosse possível detectar as falhas em um período considerado adequado para a aplicação sem exaurir a bateria do dispositivo apenas com mensagens desse tipo.

Devido às características computacionais restritas, torna-se desafiador implementar mecanismos de segurança para proteger os dispositivos que compõem um sistema de vida assistida, deixando-os vulneráveis a uma série de problemas de segurança conhecidos. Como foi observado na Figura 2.5, as principais ameaças de segurança que impactam a disponibilidade em um sistema de vida assistida estão relacionadas a ataques de adulteração, ataques de travamento, drenagem de bateria, ataques de colisão e congestionamentos na rede. Essa gama de ataques é comumente conhecida como ataques de negação de serviço (do inglês, *Denial of Service* – DoS). Variações desses ataques ganharam notoriedade nos últimos anos na forma de ataques DoS distribuídos (DDoS). Tanto os ataques DoS quanto os DDoS oferecem riscos aos sistemas de vida assistida, porém em fases diferentes do processo de aquisição e armazenamento de dados. Os ataques direcionados aos dispositivos da IoHT têm o objetivo de desabilitar temporariamente ou em definitivo um dispositivo. A indisponibilidade do dispositivo pode ocasionar dados faltantes e prejudicar a acurácia da aplicação. Para lidar com ataques DoS e DDoS, é interessante que o problema seja tratado o mais próximo possível da fonte de ataque. Desta forma, é interessante aplicar estratégias de detecção e mitigação de ataques DoS no dispositivo coordenador e no roteador de borda da rede. Assim, se houver um ataque com origem interna, o coordenador pode auxiliar na detecção e mitigação do ataque. Em paralelo, se houver um ataque com origem externa, é possível fazer uso de regras no roteador de borda para tentar efetuar o bloqueio do tráfego malicioso.

Já a disponibilidade associada aos serviços de nuvem geralmente é de responsabilidade do próprio provedor. Apesar disso, em alguns casos é possível contratar serviços que aumentam as garantias de disponibilidade da conexão e da aplicação. Ataques DDoS em estruturas de nuvem exigem o uso de múltiplos dispositivos que são coordenados remotamente para gerar um grande fluxo de dados para um alvo específico com o objetivo de desabilitar temporariamente o serviço prestado. Esses ataques são menos comuns devido ao uso de soluções robustas de mitigação de DDoS pelos prestadores de nuvem e à complexidade dos ataques para que sejam efetivos. Em relação a esses ataques, é importante proteger os dispositivos IoHT utilizados no monitoramento do indivíduo para

que eles não sejam comprometidos por usuários maliciosos e usados para realizar ataques em outras estruturas. Esse tipo de ação evita o consumo energético dos dispositivos para atividades maliciosas e colabora com a segurança da Internet.

Além da garantia de disponibilidade dos sensores, dispositivos IoT e da nuvem, também é importante observar a disponibilidade dos mecanismos de comunicação. Como foi citado anteriormente, a garantia da disponibilidade da rede geralmente está associada ao uso de equipamentos, tecnologias e caminhos redundantes. Algumas discussões a respeito disso foram tratadas quando foram apresentados os requisitos de comunicação. Em relação ao monitoramento dos canais de comunicação, é preciso pensar em aplicações de monitoramento simples de serem instaladas e utilizadas em WBAN e/ou WSN. Neste sentido, há também a necessidade de definição de quem será responsável pelo monitoramento e suporte da rede, bem como quais são os níveis de acordo de serviço em relação à manutenção em caso de problemas. Esses aspectos são essenciais para o uso de aplicações de sistema de vida assistida por usuários comuns, em particular, por pessoas idosas.

2.4.4.2. Integridade

O princípio da integridade visa assegurar que os dados não sejam alterados ou destruídos de maneira não autorizada. No contexto dos sistemas de vida assistida o princípio da integridade busca preservar a exatidão das informações sobre um usuário, sejam dados de saúde, localização ou demais informações pessoais. Recentemente, a popularização dos dispositivos IoT aplicados aos sistemas de saúde impulsionou o debate acerca da necessidade de um maior cuidado com a integridade dos dados que trafegam nesse tipo de ambiente. As características dos dispositivos e canais de comunicação presentes nos sistemas de vida assistida, especialmente quando a mobilidade é um fator preponderante, cria uma diversidade de pontos de vulnerabilidade que podem ser explorados. As medidas corretivas e a utilização de programas, dispositivos e tecnologias de comunicação que estejam alinhadas com políticas de proteção da integridade dos dados são fatores cruciais para a aceitação e adoção em massa dos sistemas de vida assistida.

As principais ameaças para a integridade dos dados nos sistemas de vida assistida estão relacionadas à manipulação da informação de maneira indevida durante sua transmissão. Isso pode ocorrer através de um código malicioso que infecta um dispositivo e captura e modifica informações relevantes, ou através de um ataque direcionado aos canais de comunicação. Outra possibilidade é um ataque físico diretamente ao dispositivo, no qual o usuário malicioso consegue obter e manipular os dados coletados. Em geral, esse tipo de ataque requer alguma informação inicial para acesso ao dispositivo como um nome de usuário, senha ou ambos. Um dos ataques direcionados aos meios de transmissão é o ataque *Man-in-the-middle* (MITM), onde um atacante consegue personificar um dos agentes do sistema e a partir dessa ação passa a alterar as informações recebidas e enviadas. Por fim, também pode acontecer a inserção de um dispositivo malicioso no sistema para coletar informações relevantes e enviar dados manipulados. É importante ressaltar que no contexto da IoHT esses ataques são de difícil execução devido à mobilidade envolvida e a proximidade dos dispositivos em relação ao usuário. Algumas contramedidas que podem ser adotadas para evitar ataques de integridade estão relacionadas com o uso de canais de transporte criptografado, mecanismos de autenticação e autorização dos dispositivos e mudança de configurações padrões, como usuário e senha de acesso.

É comum que as preocupações relacionadas à integridade estejam associadas à ameaças externas, como usuários maliciosos. No entanto, a garantia da integridade também inclui aspectos como corretude e acurácia dos dados pois, na IoHT, a inacurácia dos dados coletados pode causar riscos superiores à ausência de dados. Esses problemas na acurácia dos dados podem ocorrer por vários motivos, como mau posicionamento do sensor, interferência de sinal, problemas no funcionamento do dispositivo, configurações de data e hora incorretas, falta de calibração, entre outros. Assim, é importante que sejam adicionados mecanismos de detecção de anomalias aos sistemas de detecção de falhas. Além disso, é importante que seja feito um período de adaptação de uso do sistema de vida, para determinação de um *baseline* que possa servir como parâmetro tanto para as aplicações de saúde, quanto para as aplicações de detecção de falhas e anomalias.

2.4.4.3. Confidencialidade

O princípio de confidencialidade estabelece que as informações de cunho confidencial não sejam compartilhadas com entidades não autorizadas. Nos sistemas de vida assistida a confidencialidade se refere, por exemplo, à proteção dos dados de um indivíduo que são compartilhados com um médico ou prestador de serviço de saúde e não devem ser repassados a terceiros e utilizados para fins distintos do que foi autorizado. Os dados do usuário devem ser protegidos ainda contra invasores que vasculham os canais de comunicação em busca de informações sensíveis, e quando armazenados devem ser protegidos contra intrusão. Ao adotar medidas de proteção que substanciam o princípio da confidencialidade torna-se mais complexa a tarefa de identificar um usuário a partir dos dados coletados de maneira indevida, auxiliando o princípio da privacidade dos dados.

As principais ameaças à confidencialidade em sistemas de vida assistida ocorrem quando um atacante monitora e subtrai informações dos canais de comunicação [Hasan et al. 2022]. Geralmente esse tipo de ataque ocorre em duas etapas: obtenção dos fluxos de dados através de monitoramento passivo e análise do tráfego capturado. Os ataques de espionagem (do inglês, *eavesdropping attacks*), ocorrem quando o atacante "escuta" os canais de comunicação disponíveis em busca de informações relevantes. Considerando as tecnologias de comunicação sem fio, o alcance da tecnologia será o principal fator para determinar o raio de ação do atacante. Sendo assim, tecnologias de alcance muito curto, como o Bluetooth, dificultam a ação dos atacantes, enquanto tecnologias como o Wi-Fi permitem que o atacante monitore a rede de uma distância maior.

Após capturar uma quantia significativa de dados, o atacante analisa os dados coletados em busca de informações importantes [Brezolin et al. 2022]. A filtragem inicial busca por palavras-chave, nomes de usuário, identificações únicas dos dispositivos, e demais informações que podem ser obtidas facilmente. Em uma verificação mais profunda, o atacante pode tentar correlacionar características específicas disponíveis no conjunto de dados para identificar um usuário ou um dispositivo. Caso o atacante tenha sucesso em identificar o usuário, os dispositivos que compõem o ambiente ou consiga definir o comportamento de ambos, além da confidencialidade o atacante estará atuando contra o princípio da privacidade dos dados. Outro tipo de ataque que pode afetar a confidencialidade são os ataques de personificação. Nesse caso, após a análise de tráfego e identificação de credenciais válidas um atacante personifica um usuário ou um dispositivo com o objetivo de obter informações continuamente. Caso tenha sucesso, o atacante pode repassar

informações distorcidas dentro do sistema, impactando o princípio da integridade.

A maioria das soluções relacionadas com confidencialidade envolve o uso de criptografia no canal de comunicação e nos dados. Do ponto de vista dos dados, é possível utilizar diferentes mecanismos criptográficos para garantir a confidencialidade e privacidade. Alguns desses mecanismos foram abordados na Seção 2.3.3 quando foi tratado do armazenamento e processamento em nuvem. O processo criptográfico deve ser realizado antes dos dados serem enviados para a nuvem. Como os sensores têm capacidade de recurso bastante limitada, a comunidade científica tem buscado soluções *lightweight* que garantam a confidencialidade sem exaurir os recursos dos sensores e dispositivos IoT. Na comunicação fim a fim, é comum a utilização de protocolos como o TLS (*Transport Layer Security*) ou alternativas de criação de túnel, como IPSec. Além disso, também há recomendação de segurança que devem ser adotadas de acordo com o tipo de protocolo sem fio utilizado [Souppaya and Scarfone 2012, Fan et al. 2017, Padgette et al. 2017].

Em [Kumar et al. 2018], há a sugestão de uso de *blockchain* para alcançar requisitos de segurança em aplicações de saúde. O uso da *blockchain* garante o princípio da integridade, visto que uma vez que a informação estiver no livro-razão, não poderá ser modificada. Além disso, também contribui com a disponibilidade ao distribuir os dados e processamento em diferentes nós. Por fim, auxilia na garantia da confidencialidade e privacidade ao fazer uso de contratos inteligentes para controlar o acesso aos dados. Além do uso de *blockchain*, tem-se sugerido o desenvolvimento de soluções de criptografia "pós computação quântica". Os algoritmos criptográficos atuais são baseados em problemas matemáticos complexos que não podem ser resolvidos por computadores tradicionais em tempo hábil para realização de ataques. No entanto, há a expectativa de que com a computação quântica esses algoritmos tornem-se quebráveis. Assim, espera-se o desenvolvimento de novos algoritmos que considerem esse cenário [Yang et al. 2020].

2.4.4.4. Privacidade

O princípio da privacidade é a propriedade de um sistema que busca assegurar que os dados particulares de um usuário sejam protegidos contra divulgação não autorizada ou tentativas de exploração ilegal dessa possível divulgação. As principais preocupações associadas à quebra da privacidade em sistemas de vida assistida estão relacionadas ao uso indevido dos dados e suas implicações, como a divulgação de informações em redes sociais, o uso das informações para ameaças, e a divulgação de dados não autorizada por parte de prestadores de serviço. Para um usuário do sistema de vida assistida qualquer vazamento de informação proveniente de um dispositivo participante pode representar a divulgação não autorizada de uma condição de saúde, da sua localização, de um comportamento ou rotina específica além de informações complementares que podem ser empregadas em golpes ou tentativas de extorsão. Para fabricantes de equipamentos ou prestadores de serviço, o vazamento de informações privadas pode representar a perda da credibilidade, trazendo danos imensuráveis.

A informação sensível armazenada por prestadores de serviço de saúde é um dos principais focos de vazamento de dados [Alhaj et al. 2022]. Sendo assim, todos os pontos de armazenamento devem ter mecanismos de controle de acesso e identificação de identidade para garantir a confidencialidade da informação sensível do usuário e evitar a violação da sua privacidade. Entretanto, como essa informação geralmente é confiada a

um sistema de armazenamento na nuvem, ela se torna vulnerável a um vazamento de dados. Como resultado, os dados armazenados podem ser obtidos através de um ataque ou serem intencionalmente expostos por qualquer entidade que tenha acesso ao sistema. Outra tendência atual que requer os mesmos cuidados com relação aos vazamentos de dados são as operações de análise em *Big Data*. Para inferir informações valiosas relacionadas à saúde dos usuários, um prestador de serviços de saúde pode confiar a informação a terceiros para análise e identificação de características. Nesses casos, todo o processo deve ser protegido de forma a garantir a privacidade do usuário [Onesimu et al. 2022]. As empresas que prestam serviços que estão relacionados com o armazenamento e processamento dos dados devem seguir as regulamentações para garantir que o direito à privacidade seja garantido. No Brasil, é preciso seguir as indicações da LGPD e deixar explícito quais dados serão coletados, a finalidade desses dados e como serão tratados.

2.4.4.5. Autenticação e Autorização

A complexidade do ambiente que apoia os sistemas de vida assistida torna a autenticação uma operação árdua, uma vez que ela deve acontecer considerando diversas redes e dispositivos heterogêneos. A autenticação deve acontecer tanto para as entidades que fazem parte do sistema quanto para a informação que será transmitida. A autenticação da entidade visa garantir que uma parte interessada e autenticada do sistema se conecte a outra parte interessada, se e somente se, a segunda parte também estiver autenticada. Já a autenticação da informação é o processo pelo qual uma entidade é verificada como a origem dos dados gerados. Atualmente, uma das principais tendências para os protocolos de autenticação é a chamada autenticação leve, uma vez que as limitações de memória e processamento dos dispositivos na IoHT torna impraticável a adoção de protocolos mais robustos. A autorização visa assegurar que somente entidades reconhecidas podem acessar um determinado serviço ou recurso, como um dispositivo ou os dados de um usuário.

Todos os dispositivos participantes do sistema devem ter mecanismos de autenticação nativos. Entretanto, devido a baixa capacidade computacional os dispositivos na IoHT não possuem mecanismos de autenticação ou possuem versões simplificadas de mecanismos tradicionais, sendo vulneráveis a uma série de ataques. Outro problema reside na incompatibilidade de mecanismos de autenticação entre os fabricantes. Uma vez que não existe consenso sobre uma forma exclusiva de autenticação, cada fabricante adota uma medida de segurança, tornando complexa a tarefa de adicionar e gerenciar os dispositivos conectados. Adicionalmente, tecnologias de comunicação para redes pessoais como o Bluetooth possuem várias versões vigentes, algumas sem nenhum mecanismo de autenticação para sincronizar um dispositivo. Os principais ataques visando interferir com mecanismos de autenticação e autorização consistem em ataques de falsificação, ataques de clonagem e replicação, ataques de engenharia social e ataques que envolvem a infecção através de códigos maliciosos [Papaioannou et al. 2022].

Nos ataques de falsificação, um invasor tenta utilizar um dispositivo, subsistema ou outra parte autenticada como suporte para construção de uma identidade válida. Depois de acessar o sistema o atacante passa a fraudar o sistema com informações falsas ou ganha acesso a funções privilegiadas. Em ataques de clonagem e replicação, um usuário compromete um dispositivo e cria um número significativo de clones para subverter o sistema. Nota-se que na IoHT esses ataques não são descartados mas são extremamente raros, uma

vez que a quantidade de dispositivos e a exclusividade dos mesmos dificultam sua execução. Um outro tipo de ataque direcionado à quebra da autorização é consideravelmente mais simples e eficiente: ataques de engenharia social. Muitos usuários dos dispositivos da IoHT possuem pouca ou nenhuma afinidade com tecnologia, sendo alvos relativamente fáceis para atacantes. Caso um atacante consiga acessar um dispositivo inteligente, como um relógio ou *smartphone*, ele terá acesso a uma infinidade de aplicações, incluindo as aplicações de saúde. Nesse sentido, existem vários problemas associados como: senhas simplificadas, senhas padrão nos dispositivos e falta de configuração em mecanismos que reforçam a segurança (e.g., autenticação em duas etapas e biometria). Finalmente, os ataques que envolvem a infecção através de códigos maliciosos podem colocar um atacante no controle parcial ou total de um dispositivo da IoHT, posteriormente o atacante pode desativar sensores, serviços e demais funções disponíveis no sistema de vida assistida.

2.5. Arquiteturas e Redes de próxima geração

Novas arquiteturas e redes de próxima geração podem contribuir com o desenvolvimento de aplicações de saúde ao endereçar alguns requisitos de comunicação e segurança. Nesta seção, serão expostos trabalhos relacionados ao uso de redes de próxima geração na área de saúde, explorando benefícios e desafios associados a esses requisitos. Além disso, também será discutido como algumas arquiteturas de Internet do Futuro podem auxiliar no desenvolvimento dessas aplicações.

2.5.1. Redes de próxima geração (e.g., 5G e 6G)

Pesquisas envolvendo 5G na área de saúde têm se tornado popular devido a algumas características da 5G, como baixa latência. A Tabela 2.9 resume as características entre as gerações de redes móveis e, ao comparar as características das redes 4G com 5G, observa-se que a rede 5G apresenta maior taxa de transmissão, latência baixa e suporte a maior quantidade de dispositivos por km^2. Portanto, o potencial de uso da 5G associado aos dispositivos IoT para prover soluções de saúde é alto. As características das redes 6G tornam esse cenário ainda mais promissor, principalmente por causa da sua integração com satélite, que pode facilitar o uso de serviços de saúde avançados em áreas remotas.

Tabela 2.9. Comparação entre as características das redes 4G, 5G e 6G.

Características	Desempenho		
	4G	5G	6G
Taxa de transmissão	1Gbps	20Gbps	1Tbps
Frequênca máxima	6GHz	90GHz	10THz
Latência fim a fim	10ms	1ms	100μs
Mobilidade	350km/h	500km/h	1000km/h
Dispositivos	100k/km^2	1000k/km^2	10^7/km^2
Arquitetura	MIMO	MIMO massivo	Superfície inteligente
Integração com satélite	Não	Não	Sim

Os benefícios das redes 5G podem ser observados em diferentes aplicações de saúde: monitoramento de indivíduos, prevenção de doenças infecciosas, realização de ci-

rurgias remotas, entre outros [Devi et al. 2023]. No caso dos serviços de monitoramento, as aplicações podem beneficiar-se do aumento na taxa e velocidade de transmissão de dados, baixa latência, rede com maior eficiência energética e uso de espectro de frequência mais eficiente. Alguns desses benefícios são explorados no desenvolvimento de uma solução de monitoramento remota em tempo real [Zhang et al. 2020]. A solução usa 5G, IA e computação de borda móvel para endereçar três questões: transmissão de dados contínua, baixa latência e utilização de mecanismos de análise de dados.

Os trabalhos envolvendo 5G e a área de saúde podem ser classificados de acordo com diferentes critérios, como tecnologias de comunicação, requisitos, objetivos, métricas de desempenho e abordagens [Ahad et al. 2019]. A classificação baseada em abordagem é categorizada em controle de congestionamento, agendamento e roteamento. Os trabalhos que buscam lidar com controle de congestionamento, geralmente tem benefícios como redução de perda de pacotes e do atraso fim a fim. Já as estratégias de agendamento são caracterizadas por propiciar otimização de recursos e garantias de QoS. Alguns desses trabalhos fazem uso de redes definidas por software (do inglês, *Software Defined Networking* – SDN) e funções de rede virtualizadas (do inglês, *Network Functions Virtualization* – NFV) para reservar e instanciar recursos. Por fim, os trabalhos associados ao roteamento têm o objetivo de melhorar a comunicação entre os dispositivos. Em muitas situações, as soluções acabam mesclando características de diferentes abordagens. A Tabela 2.10 apresenta um resumo de alguns desses trabalhos.

Tabela 2.10. Soluções de 5G associadas à comunicação e segurança.

Abordagem	Referência	Contribuições
Controle de congestionamento	[Tshiningayamwe et al. 2016]	Sistema de detecção e notificação de congestionamento baseado em limiares que definem três situações: sem congestionamento, congestionamento moderado e congestionamento elevado. No caso de congestionamento moderado, o algoritmo considera o tamanho do *buffer* e nível de energia do nó para tomada de decisões de encaminhamento. Caso haja congestionamento elevado, há a redução da taxa de dados trafegados. Os benefícios da solução são aumento da vazão, redução da perda de pacotes e do atraso fim a fim.
Priorização de tráfego	[Besher et al. 2022]	Estratégia de priorização de tráfego em redes 5G com congestionamento: os pacotes relacionados com a área de saúde são marcados com uma *flag*, que é utilizada pelo comutador SDN para classificá-lo de acordo com sua prioridade e encaminhá-lo através da rede. Os pacotes marcados como prioritários são processados primeiro e têm redução do atraso fim a fim. No entanto, nenhum mecanismo para minimizar o congestionamento é adotado.
Roteamento	[Ahad et al. 2021]	Protocolo de roteamento baseado em *cluster* para reduzir o atraso na transmissão dos dados, melhorar a eficiência energética e estender a vida útil da rede. O algoritmo seleciona o líder do agrupamento com base em critérios como distância entre os nós e a estação base, energia e velocidade dos nós. A partir da eleição do líder, os grupos são estabelecidos e os nós que fazem parte do agrupamento encaminham suas mensagens para o líder, que irá comunicar-se com uma das estações base. Nesse processo, os autores fazem uso de mecanismos de aprendizado por reforço para que os nós e o líder identifiquem a rota mais eficiente energeticamente.

Embora a implementação de 5G ainda esteja em seus estágios iniciais no Brasil e no mundo, a comunidade acadêmica e a indústria já tem realizado pesquisas a respeito da sexta geração de redes móveis. Em [Koren and Prasad 2022], os autores discutem questões de privacidade e segurança relacionados às aplicações de saúde em redes 6G. As principais ameaças sinalizadas pelos autores são uso de computação quântica para quebrar mecanismos criptográficos atuais, dispositivos IoT comprometidos ou não autorizados, roubo de dados de dispositivos IoT, espionagem nos canais de comunicação e bloqueio de sinal. Além desses pontos, os autores ainda ressaltam que as redes 6G podem herdar vulnerabilidades de segurança das redes 5G, como ataques direcionados ao controlador SDN e ataques relacionados à NFV. Na Tabela 2.11, é possível observar algumas das principais ameaças envolvendo as principais tecnologias utilizadas nas redes 5G e 6G.

Tabela 2.11. Ameaças associadas às tecnologias utilizadas nas redes 5G. Adaptado de [Mangla et al. 2022]

Ameaças	Alvo	Tecnologia afetada			
		SDN	NFV	Nuvem	MIMO
Ataque DoS	Elementos de controle centralizados	x	x	x	
Ataque de configuração	Switches e roteadores SDN	x	x		
Ataques hijacking	Controlador SDN e hypervisor	x	x		
Ataques de saturação	Controlador e switches SDN	x			
Eavesdropping	Canais de controle	x			x
Ataques TCP	Comunicação entre controlador e switch SDN	x			
MITM	Comunicação entre controlador e switch SDN	x			
Vazamento de dados	Sistemas de armazenamento em nuvem			x	
Intrusão na nuvem	Sistemas de nuvem			x	

Após identificar as principais ameaças relacionadas às redes 5G e 6G, os autores em [Mangla et al. 2022] destacam soluções baseadas em computação quântica para lidar com ataques DoS e DDoS [Price et al. 2020], MITM, *replay*, *eavesdropping* e roubo de sessão [Srivastava et al. 2020], segurança em SDN e NFV [Aguado et al. 2017] e segurança de redes heterogêneas [Kakkar 2020]. Técnicas de aprendizado de máquina e uso de criptografia homomórfica também são soluções que têm sido exploradas para promover segurança e privacidade em redes 5G e 6G [Koren and Prasad 2022].

2.5.2. Redes Definidas Por Software

O paradigma SDN [McKeown 2009] estabelece uma separação entre o plano de controle e o plano de dados dos dispositivos da rede. O plano de controle é centralizado em um nó, denominado controlador, que tem uma visão global da rede e define as regras de encaminhamento dos fluxos. Essas regras são enviadas aos comutadores de rede, que são responsáveis por encaminhar os pacotes. A centralização da rede possibilita a programabilidade do encaminhamento dos fluxos de forma mais flexível e dinâmica. Tais propriedades são favoráveis ao desenvolvimento de arquiteturas de redes para IoHT, visto que os nós encaminhadores não precisam realizar o processamento dos pacotes localmente. Além disso, a manutenção das regras de encaminhamento é feita no controlador de acordo com condições predefinidas, permitindo que o tráfego possa ser redirecionado automaticamente diante da detecção de gargalos [Cicioğlu and Çalhan 2019], para aplicar técnicas de

priorização de tráfego [Yaseen et al. 2022, Kamboj et al. 2021, Misra et al. 2020], fazer balanceamento de cargas e otimização da rede [Li et al. 2020], promover eficiência energética [Cicioğlu and Çalhan 2020], direcionar o tráfego com base na classe da aplicação [Kamboj et al. 2021], ou até mesmo realizar agregação de dados [Madureira et al. 2020].

Embora o uso de SDN tenha se tornado popular, uma das principais críticas está relacionada com a centralização da operação. No entanto, como em cenários IoHT é comum haver a presença de coordenadores locais, a centralização trazida pelo uso da SDN não adiciona complexidade ou desvantagens à aplicação. Neste sentido, a visão global da rede favorece o gerenciamento de dispositivos, principalmente em cenários de mobilidade ou na presença de falhas em função de esgotamento de recursos computacionais. O controlador SDN pode, portanto, monitorar os dispositivos e aplicar ações de gerenciamento em tempo real a partir das informações obtidas, definindo dinamicamente rotas por onde um fluxo deve passar [Cicioğlu and Çalhan 2019]. Essa definição das rotas pode, inclusive, considerar aspectos como temperatura e nível de bateria dos equipamentos.

Outro aspecto relevante no uso de soluções baseadas em SDN é a possibilidade de distribuir ou compartilhar o processamento e armazenamento de dados coletados por sensores em infraestruturas de névoa e nuvem. Recursos computacionais da névoa estão mais próximos da origem dos dados e mitigam os possíveis atrasos de transmissão, enquanto aumentam a disponibilidade e reduzem a sobrecarga. Nas aplicações IoHT em arquiteturas híbridas, as decisões de encaminhamento dos dados podem ser baseadas de acordo com níveis de prioridade, para implementar estratégias de priorização de tráfego, ou níveis de sensibilidade, para garantir requisitos de confidencialidade e privacidade. Nesses casos, é importante avaliar o *tradeoff* entre os recursos computacionais necessários para analisar os dados da aplicação e os seus requisitos de segurança [Misra et al. 2020]. Neste sentido, algumas propostas na literatura sugerem o gerenciamento de QoS de dados de saúde por meio do uso de SDN e técnicas de aprendizado de máquina [Kumari and Jain 2022, Misra et al. 2020]. Em tais propostas, as camadas de névoa e nuvens são adotadas para apoiar o armazenamento e processamento dos dados de saúde coletados por sensores na borda da rede.

Além do uso da SDN para melhorar a comunicação da rede e alcançar requisitos de QoS das aplicações, existem abordagens que exploram o modelo centralizado e a flexibilidade da programabilidade da SDN para prover serviços customizados ao usuário de sistemas de saúde. Por exemplo, em [Misra et al. 2023], os autores apresentam uma arquitetura de rede em que um controlador é utilizado para definição dinâmica de regras de encaminhamento dos *switches* de modo a apoiar a distribuição de módulos de *analytics* de um sistema de diagnóstico de pacientes à luz do QoS da rede. Assim como abordagens anteriores, a proposta faz um ranqueamento de prioridade dos fluxos. O conceito de programabilidade da rede também pode ser utilizado para desenvolver soluções de segurança. Em [Uddin et al. 2019], os autores desenvolveram um *framework* denominado *PrivacyGuard* que busca preservar a privacidade dos dados das aplicações através da construção de políticas de privacidade programáveis. As principais características desse *framework* é que ele utiliza informações de contexto relacionadas ao usuário, aplicação, dispositivo e rede para definir as políticas de privacidade. Além disso, há a premissa de que as políticas sejam transparentes para a aplicação, sem que haja necessidade de quaisquer mudanças no cliente ou no servidor.

2.5.3. Redes Centradas na Informação

As Redes Centradas na Informação (do inglês, *Information-Centric Networking* – ICN) [Jacobson et al. 2009, Sampaio et al. 2021] é um paradigma de Internet do Futuro que se baseia no fato de que os usuários estão interessados no conteúdo e não necessariamente onde ele está armazenado. Partindo dessa premissa, foi proposto um modelo de rede que desvincula o identificador e o localizador de um conteúdo e que executa funções de rede baseada em nome. As Redes de Dados Nomeados (do inglês, *Named-Data Networking* – NDN) [Zhang et al. 2014] é a arquitetura mais popular do paradigma ICN e suas principais características são: uso de um esquema de nomeação hierárquico e semântico para nomear dados e elementos da rede, *cache* nos dispositivos de rede, segurança a nível de dados e plano de encaminhamento com estado [Sampaio et al. 2021].

O esquema de nomeação semântico e o roteamento baseado em nome, característicos da NDN, possibilitam uma relação direta entre a aplicação e a rede. Essa associação oferece vantagens para adição de semântica ao tráfego, permitindo que nós da rede possam identificar classes dos dados e oferecer serviços diferenciados, tais como manutenção de dados em *caches*, priorização de tráfego ou até mesmo a adoção de estratégias de encaminhamento específicas e cientes de contexto. Esse tipo de funcionalidade pode ser utilizada em cenários de IoHT, para priorizar alertas médicos e o envio dos dados coletados pelos sensores. Além disso, o esquema de nomeação também auxilia outras funcionalidades úteis para aplicações de saúde, como a disponibilidade de serviços de *bootstrapping* e políticas de controle de acesso [Aboodi et al. 2019]. O processo de *bootstrapping* em dispositivos IoT está associado às configurações iniciais de conectividade e segurança dos dispositivos. Neste contexto, a NDN possibilita a divulgação desse serviço na rede e o envio de comandos para configuração dos dispositivos através do nome de pacotes de interesse. Esta mesma funcionalidade pode ser utilizada para enviar comando aos atuadores. Por exemplo, um pacote de interesse na camada de rede pode ter o nome "sala/luz/desligar" para indicar ao atuador localizado na sala que a luz deve ser desligada [Shang et al. 2016].

A NDN também tem sido explorada para endereçar requisitos de segurança na área da saúde [Saxena and Raychoudhury 2017, Boussada et al. 2019, Dulal et al. 2022]. Em [Dulal et al. 2022] os autores apresentam um sistema baseado em NDN, denominado *mGuard*, para desenvolver políticas de controle de acesso cientes de contexto e com alta granularidade. A solução faz uso de criptografia baseada em atributos e da semântica do esquema de nomeação para explorar os benefícios da NDN nesse cenário. Já em [Boussada et al. 2019], os autores demonstram os benefícios do uso do PP-NDNoT, um sistema desenvolvido para garantir requisitos de integridade, autenticação mútua e privacidade orientada ao conteúdo e baseada em contexto para aplicações de saúde em NDN.

As características da NDN, quando incorporadas à IoT, podem impulsionar o desenvolvimento de aplicações em diferentes áreas, como sistemas ciberfísicos, redes veiculares, *smart home*, *smart city* e saúde [Aboodi et al. 2019]. Na Tabela 2.12, há um resumo das características da NDN e uma breve discussão sobre como essas características podem beneficiar (ou não) as aplicações IoHT. Também foram listados desafios associados ao uso da NDN nesses cenários.

Tabela 2.12. Características e desafios da incorporação da NDN na área de saúde.

Características	Desafios
Esquema de nomeação semântico	
O esquema de nomeação é usado para nomear todos os componentes da rede, incluindo usuários, dispositivos e dados. Quando é adicionada semântica ao esquema de nomeação, é possível utilizá-lo para para suportar outros recursos e serviços, como encaminhamento de pacotes, *multicast*, suporte à mobilidade, roteamento, segurança e configuração de dispositivos.	-Quais características do esquema de nomeação podem ser utilizadas para auxiliar no desenvolvimento de serviços para saúde? -Qual o tipo de de esquema de nomeação mais adequado para esses cenários? -Quais são as implicações na privacidade ao fazer uso de nomeação semântica na camada de rede?
***Cache* nos dispositivos de rede**	
A NDN implementa *cache* oportunístico na rede para armazenar cópias dos dados. Em aplicações de saúde, o *cache* pode ser benéfico para garantir a entrega de conteúdo em caso de problemas na rede. Apesar disso, é preciso avaliar se o *cache* teria uma boa taxa de acerto, já que o envio de dados coletados pelos sensores é constante e irá gerar mudanças frequentes na *cache*.	-Quão benéfico é o uso de *cache* em sistemas de vida assistida? -Que tipo de políticas de *cache* seriam adequadas para aplicações de saúde? -Como o uso de *cache* poderia auxiliar no suporte à mobilidade nesses cenários?
Estratégias de encaminhamento e Diferenciação de tráfego	
É possível utilizar diferentes estratégias de encaminhamento com base no prefixo do nome do conteúdo ou mesmo, essas estratégias podem adaptar-se ao contexto, a depender das regras definidas. Essa característica, associada à semântica do esquema de nomeação, pode ser utilizada para proporcionar serviços de priorização de tráfego em aplicações de saúde.	-Quais requisitos devem ser levados em consideração na definição das estratégias de encaminhamento? –Como estratégias de encaminhamento podem melhorar o QoS em sistemas de vida assistida? -Como considerar as métricas da rede e a semântica da aplicação nas definições de encaminhamento e roteamento?
Configuração e Gerenciamento de dispositivos	
O esquema de nomeação semântico pode ser usado para registro de serviços, descoberta de vizinhos e configuração de dispositivos. Esses processos tem relação direta com requisitos de interoperabilidade, auto-configuração e escalabilidade em aplicações de saúde.	-Quais tipos de serviços devem ser divulgados na rede? -Que tipo de padrões devem ser incluídos no esquema de nomeação para dar suporte aos serviços de configuração e gerenciamento de dispositivos?
Segurança e Políticas de Controle de Acesso	
Os requisitos de integridade e autenticidade são alcançados na NDN através da assinatura de pacotes pelo produtor, mas a arquitetura não oferece suporte nativo à confidencialidade, sendo necessário o uso de soluções adicionais [Zhang et al. 2018]. A semântica do esquema de nomeação pode ser utilizada na definição de políticas de controle de acesso granulares e cientes de contexto. Neste sentido, serviços facilitados na NDN, como auto-configuração dos dispositivos, podem auxiliar no processo de *bootstrapping* das funções de segurança.	-Como a arquitetura NDN pode fornecer serviços de segurança e tolerância a falhas no nível de rede? -Quais são os principais ataques de rede que devem ser evitados em tais cenários? -Quais características dos dispositivos, comunicação, serviços e aplicativos devem ser consideradas ao endereçar os requisitos de segurança e as políticas de controle de acesso? -Como fornecer segurança a nível de dados considerando a limitação de recursos dos dispositivos?

2.6. Estudos de caso

Esta seção apresenta dois estudos de caso relacionados, respectivamente, à conectividade e segurança em aplicações IoHT. O primeiro estudo de caso trata do monitoramento de problemas cardíacos em idosos localizados em áreas remotas. O segundo estudo de caso descreve um mecanismo de autenticação que emprega biossinais para garantir segurança na transmissão de dados.

2.6.1. Monitoramento de idosos em áreas remotas

O estado do Amazonas, na região Norte do Brasil apresenta as piores qualidades de enlace[4]. Esse problema é ocasionado por fatores como dificuldade de acesso geográfico e falta de investimentos em infraestrutura. As consequências desses fatores não se limitam às redes de computadores e são observadas em áreas como educação e saúde. Com o intuito de oferecer melhores condições de saúde para a população, a Fundação Amazônia Sustentável[5] (FAS) desenvolveu um programa denominado Saúde na Floresta, cujo objetivo é promover o atendimento de atenção básica de saúde e auxiliar na formação de profissionais da área. Esse programa engloba ações de telessaúde, políticas públicas, educação e pesquisa em saúde. O uso da telessaúde traz grandes benefícios pois possibilita a oferta de serviços de saúde em áreas remotas. No entanto, outras aplicações também podem ser utilizadas para ampliar os serviços de saúde nessas regiões e melhorar as condições de vida dos indivíduos.

As doenças do aparelho circulatório é a maior causa de mortes e hospitalizações de idosos no Brasil [Heemann and Hermsdorf 2017] e, na região Norte, o percentual de óbitos em pessoas idosas com essa causa se aproximou de 25% em 2020. O diagnóstico e o acompanhamento dessas doenças ocorrem através do mapeamento de sintomas físicos (e.g., dor no peito, pernas inchadas e desmaios), do monitoramento de parâmetros de saúde do indivíduo (e.g., pressão arterial, saturação de oxigênio e frequência cardíaca) e através de eletrocardiogramas. Aplicativos, sensores e equipamentos vestíveis coletam esses dados. Como citado anteriormente, os sensores e equipamentos vestíveis têm poucos recursos computacionais e, geralmente, encaminham os dados coletados para o *gateway*. A comunicação física entre os sensores e o *gateway* segue protocolos de comunicação de curto alcance suportados pelos dispositivos. Além da comunicação local, essas aplicações necessitam da conectividade entre o *gateway* e a nuvem, onde serão realizados os processos de armazenamento permanente e análise dos dados. No entanto, a necessidade de comunicação com a nuvem pode impossibilitar o uso dessas aplicações em áreas remotas. Nessas regiões, o acesso à Internet comumente acontece através de comunicações a rádio ou via satélite. Enquanto a comunicação a rádio é mais suscetível a interferências de sinal e apresenta maior incidência de perda de pacotes, a comunicação via satélite tradicional resulta em alta latência. Essas duas características inviabilizam o uso de aplicações de monitoramento de doenças cardíacas, pois essas aplicações são sensíveis a perdas e precisam de baixa latência.

Neste estudo de caso, propusemos a adoção de redes de satélites de baixa órbita (do inglês, *Low Earth Orbit*) para possibilitar o uso de aplicações de monitoramento de

[4]Dados disponibilizados pelo NIC.br: https://qualidadedainternet.nic.br/
[5]https://fas-amazonia.org

condições cardíacas em áreas remotas, garantindo requisitos de alta capacidade e baixa latência para realização de monitoramento contínuo e em tempo real. Comparado com as redes de satélite tradicionais, a LEO apresenta características como baixo atraso de propagação, pequena perda de propagação e cobertura global. Além disso, há a expectativa de que o uso de constelações de satélite ofereçam latências inferiores às conexões via fibra óptica para distâncias superiores a 3000km [Handley 2018].

A Figura 2.6 ilustra o cenário proposto e dos seus elementos. Os sensores são responsáveis por coletar os dados e enviar para o *gateway* local, que irá se comunicar com o terminal de satélite do usuário. Após receber os dados do *gateway*, o terminal transmite o tráfego para o satélite mais próximo. A maioria desses terminais têm características favoráveis a sua adoção, como tamanho pequeno, tempo de vida útil longo e baixo consumo energético. Além disso, como os satélites estão em movimento, a comunicação entre o terminal e o satélite acontece mesmo quando há obstáculos próximos ao terminal [Qu et al. 2017]. Depois de receber o tráfego, os satélites comunicam-se através de *lasers* e encaminham os dados até chegar no terminal de destino. Assim, o usuário consegue acessar a Internet e a infraestrutura de nuvem pela aplicação. Na infraestrutura de nuvem é possível definir políticas de controle de acesso que permitam aos profissionais de saúde visualizarem os dados e avaliarem as condições de saúde do indivíduo. A partir dessa análise, os profissionais estão aptos a encaminhar tipo de *feedback* ao usuário, como solicitação de ajuste medicamentoso e encaminhamento do indivíduo para profissionais de saúde locais. Essas ações podem reduzir situações de emergência e mortes.

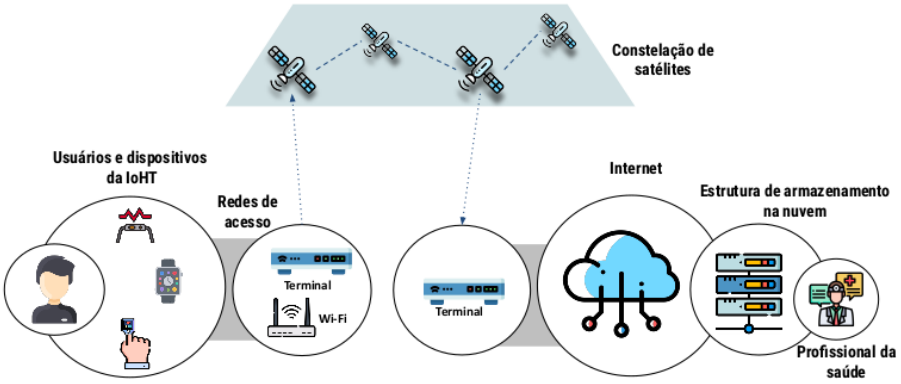

Figura 2.6. Utilização de LEO para possibilitar uso de aplicações IoHT em áreas remotas.

Alguns benefícios obtidos através da junção entre LEO e IoT são as características de QoS e velocidade da rede e a facilidade de uso dessa estrutura em ambientes remotos. Embora as redes LEO ainda estejam em processo de implementação, já existem trabalhos que relacionam o uso dessas redes para facilitar a comunicação em cenários IoT [Qu et al. 2017]. Alguns desafios de pesquisa que envolvem essa adoção incluem a adequação de protocolos IoT terrestres para comunicação via satélite.

2.6.2. Uso de biossinais na autenticação de usuários

A autenticação de usuários e dispositivos é um mecanismo fundamental para garantia de requisitos de segurança em sistemas de vida assistida. O processo de autenticação segue três abordagens: algo que se sabe (e.g., senha), algo que se tem (e.g.,

smartphone ou chave USB) e algo que se é (e.g., biometria). Como um dos principais públicos alvo dos sistemas de vida assistida são as pessoas idosas, a dependência de senhas e dispositivos físicos pode gerar aversão ao uso sistema devido ao esquecimento das senhas ou à dificuldade em interagir com o dispositivo de entrada. Paralelo a isso, alguns dispositivos na IoHT têm baixo poder de processamento e armazenamento, dificultando a implementação de mecanismos de autenticação de alta complexidade computacional. Assim, os sinais biométricos que capturam características de um indivíduo através de interfaces homem-máquina são candidatos promissores para promover a autenticação de um usuário. A impressão digital é o método mais comum de identificação biométrica, entretanto, pessoas idosas podem apresentar perda ou mudança da impressão digital e muitos dispositivos mais simples na IoHT, como monitores de atividade física e *smartwatches*, não possuem um leitor de impressão digital.

Neste estudo de caso, analisamos a adoção da variação da frequência cardíaca, medida através dos sensores pletismográficos encontrados nos dispositivos da IoHT, no processo de identificação e autenticação do usuário. A teoria empregada é que, assim como no caso das impressões digitais, cada ser humano tem uma variação de frequência cardíaca única e exclusiva, que pode ser utilizada em processos de autenticação. Embora essa alternativa possa resolver a autenticação de usuários, ainda há ameaças que podem afetar o processo de autenticação. Na Seção 2.4, foram expostas as principais ameaças considerando a autenticação e autorização dos dispositivos e um dos principais problemas está relacionado à transmissão das informações através de tecnologia de comunicação sem fio. Nesses casos, um atacante pode escutar o canal de transmissão e obter informações, incluindo senhas e dados biométricos empregados no processo de autenticação. Sendo assim, o processo de autenticação também deve considerar a segurança do canal de comunicação. A tecnologia de acoplamento galvânico permite que uma quantidade de dados seja transmitida utilizando o tecido corporal (pele) como meio de transmissão. Esse processo de autenticação é ilustrado na Figura 2.7 e segue as fases de (i) aquisição de dados e pré-processamento; (ii) comunicação galvânica e (3) processo de autenticação.

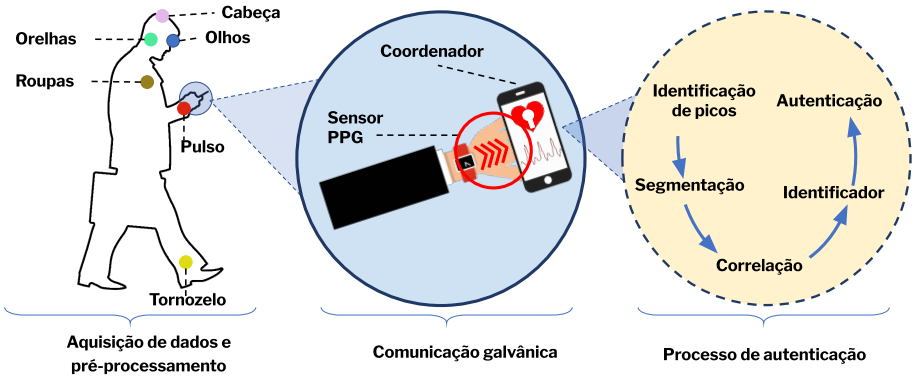

Figura 2.7. Mecanismo de autenticação através de sinal biométrico e transmissão segura de informações.

Na fase de aquisição de dados e pré-processamento os dispositivos com capacidade de coletar os biossinais do fotopletismograma (do inglês, *photoplethysmogram* – PPG), através dos sensores pletismográficos, coletam o sinal que será utilizado para inferência da variação da frequência cardíaca. Esses sensores podem estar posicionados

em diversos pontos do corpo humano de acordo com a necessidade e de modo a oferecer maior conforto para o usuário. Antes de ser enviado ao dispositivo coordenador para autenticação os dados coletados passam por uma etapa de pré-processamento básico para otimização do mecanismo. Uma vez que os dispositivos estão em contato com a pele do usuário, o envio dos dados até o coordenador ocorre através da tecnologia de acoplamento galvânico (fase de comunicação galvânica), utilizando a pele e tecidos como meio de transmissão. O acoplamento galvânico dificulta a interceptação da transmissão dos dados. De posse dos dados PPG, o coordenador executa o pós-processamento dos dados e identifica uma série de características do sinal (subfases no processo de autenticação). Finalmente, o coordenador autoriza o usuário a utilizar as funções do sistema, ou nega o acesso e encerra o processo de autenticação.

Essa alternativa de autenticação foi desenvolvida e avaliada no contexto do projeto *NSF/RNP US-Brazil Healthsense*[6], cujos dois objetivos principais são: *i)* analisar e explorar as características dos dispositivos vestíveis, aplicações e protocolos de rede, e *ii)* propor técnicas para resiliência através do uso do corpo para transmitir informações de forma segura. O detalhamento do mecanismo de autenticação utilizado como base neste estudo de caso e o repositório contendo informações relevantes e arquivos de coletas de sinais PPG estão publicamente disponíveis através dos trabalhos desenvolvidos no escopo do projeto [Nakayama et al. 2019a, Nakayama et al. 2019b].

2.7. Ambientes de Experimentação e *Datasets*

Nesta seção serão apresentadas ferramentas e *datasets* para experimentação relacionada à comunicação e segurança na área de saúde. É desafiador conseguir executar experimentos envolvendo os vários componentes de um sistema de vida assistida, que variam desde sensores em uma WBAN a recursos na nuvem. Geralmente o processo de experimentação é realizado através da segmentação da arquitetura para avaliar diferentes partes da rede, como a WBAN ou WSN, as redes de acesso e a comunicação com a nuvem. Simuladores de redes tradicionais, como o NS-2, NS-3 e o OMNeT++ podem ser utilizados para realizar experimentos envolvendo essas aplicações. No entanto, existem ferramentas que foram desenvolvidas especificamente para a experimentação de aplicações em IoT, incluindo cenários de saúde. Assim, é possível citar:

- **IoTFY**[7]: plataforma que facilita a prototipação, dimensionamento e gerenciamento de aplicações IoT na nuvem, em grande escala e de forma realista. A plataforma inclui particularidades da IoT, como o uso de comunicações M2M e a caracterização do tráfego dos dispositivos. Na comunicação M2M, o simulador suporta protocolos como MQTT, CoAP e HTTP. Além disso, implementa soluções de comunicação que consideram os princípios associados aos protocolos de comunicação sem fio de curta e longa distância e suas características em relação às métricas de rede, como latência. Esse simulador tem uma versão básica que é gratuita, mas para utilizar todas as funcionalidades desenvolvidas, é necessário utilizar a versão paga.

- **IotNetSim**: permite a execução de simulação em três níveis: camada IoT, borda e

[6] https://www.healthsenseproject.net/
[7] https://docs.iotify.io/

nuvem. A camada IoT é onde estão os sensores que geram dados para enviar para o *gateway*. O *gateway* processa os dados e os envia para a camada de borda, que trata os dados e encaminha para a nuvem [Salama et al. 2019].

- **MoSIoT**: permite a simulação de aplicações de monitoramento da saúde do indivíduo baseado no paradigma de engenharia orientada a modelos. Os autores validam a ferramenta utilizando uma aplicação associada à doença de Alzheimer e os códigos do simulador são disponibilizado em [Meliá et al. 2021].

Experimentos que envolvem o uso de arquiteturas e modelos de Internet do Futuro podem incluir simuladores e emuladores como o NDNSim[8], Mininet[9] e o MiniNDN[10]. No caso do uso de emuladores, é comum a integração com o módulo de rede sem fio do NS-3 para conseguir mapear as características do ambiente sem fio na emulação. Nos últimos anos, também foram apresentadas algumas soluções no salão de ferramentas do SBRC que podem ser utilizadas na criação de *testbeds*, experimentação e adoção de mecanismos de gerenciamento e segurança. Dentre as ferramentas, tem-se o OTALab [Cussuol et al. 2022], IoTFogSim [Pereira et al. 2021], IMAIoT [Heideker et al. 2019] e o SentryIoTOAuth [Andrade and Monteiro 2019].

O mapeamento das características do tráfego em aplicações de saúde varia muito a depender da aplicação. Algumas aplicações são caracterizadas por envio de tráfego contínuo, enquanto outras podem ter uma frequência e taxa de chegada de pacotes diferente. Além disso, é possível ter esses comportamentos variados em diferentes sensores que envolvem uma aplicação. Conseguir mapear essas características na simulação é um processo desafiador. Como alternativas, é possível criar *testbeds* utilizando dispositivos como Arduínos e Raspberry Pi para verificar as características do processo de geração e transmissão de dados ou utilizar *datasets* como insumo nas simulações. Alguns *datasets* relacionados à área de saúde são citados a seguir:

- **ECU-IoHT**[11]: *dataset* construído em um ambiente IoHT para possibilitar experimentos envolvendo segurança cibernética. Os dados incluem diferentes ataques e exploram várias vulnerabilidades.

- ***mHealth dataset***[12]: dados de movimento corporal e sinais vitais coletados por dez voluntários durante a realização de atividades físicas. Os dados incluem informações como monitoramento cardíaco.

- ***Healthsense dataset***[13]: dados de sinais PPG que podem ser utilizados para pesquisas envolvendo o processo de autenticação por meio de biossinais, conforme abordado no segundo estudo de caso [Nakayama et al. 2019b].

[8] https://ndnsim.net
[9] http://mininet.org/
[10] https://github.com/named-data/mini-ndn
[11] https://ro.ecu.edu.au/datasets/48/
[12] http://archive.ics.uci.edu/ml/datasets/mhealth+dataset
[13] https://github.com/Healthsense-Project

As métricas utilizadas na condução da avaliação dependem do tipo de solução proposta, mas é importante sempre avaliar o consumo energético atribuído ao custo da solução. Além disso, os parâmetros da rede, dos dispositivos e do tráfego podem variar dependendo do tipo de aplicação e é interessante emular ou simular essa heterogeneidade de dispositivos e recursos nas avaliações.

2.8. Considerações finais

O desenvolvimento de aplicações de vida assistida envolve a utilização de várias tecnologias que trazem consigo novas contribuições, possibilidades de aplicação, casos de uso e desafios de pesquisa. Nos últimos anos, tem-se observado a evolução da área de IoHT ao desenvolver soluções que não se limitem fisicamente a um lugar (e.g., a casa da pessoa) e que garantam o uso contínuo de serviços de vida assistida, independente de onde o indivíduo esteja. Essas aplicações são vistas como uma possível solução para lidar com o aumento dos custos na área de saúde e com o envelhecimento da população mundial. O desenvolvimento dessas soluções depende de diversas tecnologias, como IoT, computação em nuvem e Inteligência Artificial. Essas tecnologias são utilizadas como componentes em uma arquitetura que visa coletar dados, transmiti-los e processá-los de forma segura. Para que essas soluções sejam amplamente adotadas, é preciso atender aos requisitos das aplicações e dos usuários, que têm relação direta com os mecanismos de comunicação e segurança da rede e dos dispositivos.

Neste minicurso, foram apresentados os principais conceitos associados às aplicações de vida assistida, a partir da perspectiva da conectividade e da segurança. Assim, foi possível observar como os requisitos dos usuários e das aplicações têm impulsionado o desenvolvimento de dispositivos, protocolos de comunicação específicos para a IoT e alternativas mais seguras de armazenamento e processamento de dados. Os requisitos de conectividade e segurança em sistemas de vida assistida são transversais e precisam considerar a heterogeneidade dos dispositivos, tráfego e aplicações. De uma forma geral, observa-se que os requisitos associados à conectividade, QoS, integridade e disponibilidade têm uma relação direta com a confiabilidade do sistema pelos usuários. É preciso garantir que o sistema funcione de forma precisa e confiável em relação aos dados coletados pelos sensores e trafegados na rede. Também, a confiabilidade torna-se maior com a garantia de que o sistema trata questões de confidencialidade e privacidade dos usuários.

Cenários de aplicação associados à saúde de idosos impõem desafios relacionados a diferentes áreas. Embora seja desafiador experimentar soluções que considerem amplamente as características de cenários reais, é importante que os protocolos e soluções desenvolvidos sejam interoperáveis, pois o desenvolvimento de soluções compatíveis entre si impulsiona a criação de aplicações reais. Além disso, os protocolos precisam levar em conta as restrições de capacidade dos dispositivos para que seu uso seja realista. Existem soluções para esses cenários que se baseiam no uso de redes de próxima geração e arquiteturas de Internet do Futuro para alcançar requisitos de rede e segurança. Ao longo dos próximos anos, observaremos o crescimento de aplicações que necessitam da integração de dados e dispositivos para conseguir prover soluções inteligentes. Assim, o avanço de novos algoritmos e protocolos específicos para os sistemas voltados aos cuidados da saúde serão observados nas áreas de redes de computadores, sistemas distribuídos, segurança de redes e de informação, computação em nuvem e Inteligência Artificial.

Agradecimentos

Este trabalho foi realizado com o apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq – 432064/2018-4, 316208/2021-3, 402854/2022-5, 200404/2022-9, 313844/2020-8, 426701/2018-6), da Fundação de Aparo à Pesquisa do Estado da Bahia (FAPESB – TIC0004/2015), da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP – #2021/06733-6) e da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

Referências

[Aazam et al. 2020] Aazam, M., Zeadally, S., and Harras, K. A. (2020). Health fog for smart healthcare. *IEEE Consumer Electronics Magazine*, 9(2):96–102.

[Aboodi et al. 2019] Aboodi, A., Wan, T.-C., and Sodhy, G.-C. (2019). Survey on the incorporation of ndn/ccn in iot. *IEEE Access*, 7:71827–71858.

[Adil et al. 2022] Adil, M., Alshahrani, H., Rajab, A., Shaikh, A., Song, H., and Farouk, A. (2022). Qos review: smart sensing in wake of covid-19, current trends and specifications with future research directions. *IEEE Sensors Journal*.

[Aguado et al. 2017] Aguado, A., Lopez, V., Martinez-Mateo, J., Szyrkowiec, T., Autenrieth, A., Peev, M., Lopez, D., and Martin, V. (2017). Hybrid conventional and quantum security for software defined and virtualized networks. *Journal of Optical Communications and Networking*, 9(10):819–825.

[Ahad et al. 2021] Ahad, A., Tahir, M., Sheikh, M. A., Ahmed, K. I., and Mughees, A. (2021). An intelligent clustering-based routing protocol (crp-gr) for 5g-based smart healthcare using game theory and reinforcement learning. *Applied Sciences*, 11(21):9993.

[Ahad et al. 2019] Ahad, A., Tahir, M., and Yau, K.-L. A. (2019). 5g-based smart healthcare network: architecture, taxonomy, challenges and future research directions. *IEEE access*, 7:100747–100762.

[Alhaj et al. 2022] Alhaj, T. A., Abdulla, S. M., Iderss, M. A. E., Ali, A. A. A., Elhaj, F. A., Remli, M. A., and Gabralla, L. A. (2022). A survey: To govern, protect, and detect security principles on internet of medical things (iomt). *IEEE Access*, 10:124777–124791.

[Amiribesheli et al. 2015] Amiribesheli, M., Benmansour, A., and Bouchachia, A. (2015). A review of smart homes in healthcare. *Journal of Ambient Intelligence and Humanized Computing*, 6:495–517.

[Andrade and Monteiro 2019] Andrade, R. B. d. S. and Monteiro, J. A. S. (2019). Sentryiotoauth: um provedor de serviço de autenticação e autorização para casas inteligentes baseado no processo ace-oauth. In *Anais Estendidos do XXXVII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*, pages 73–80. SBC.

[Arcelus et al. 2007] Arcelus, A., Jones, M. H., Goubran, R., and Knoefel, F. (2007). Integration of smart home technologies in a health monitoring system for the elderly. In

21st International Conference on Advanced Information Networking and Applications Workshops (AINAW'07), volume 2, pages 820–825. IEEE.

[Attrapadung and Imai 2009] Attrapadung, N. and Imai, H. (2009). Attribute-based encryption supporting direct/indirect revocation modes. In *Cryptography and Coding: 12th IMA International Conference, Cryptography and Coding 2009, Cirencester, UK, December 15-17, 2009. Proceedings 12*, pages 278–300. Springer.

[Bansal and Kumar 2020] Bansal, S. and Kumar, D. (2020). Iot ecosystem: A survey on devices, gateways, operating systems, middleware and communication. *International Journal of Wireless Information Networks*, 27:340–364.

[Bardalai et al. 2022] Bardalai, P., Neog, H., Dutta, P. E., Medhi, N., and Deka, S. K. (2022). Throughput prediction in smart healthcare network using machine learning approaches. In *2022 IEEE 19th India Council International Conference (INDICON)*, pages 1–6. IEEE.

[Besher et al. 2022] Besher, K. M., OKidhain, I., Wick, L., and Ali, M. Z. (2022). Congestion control of healthcare packet routing in 5g edge computing networks. In *2022 International Conference on Engineering and Emerging Technologies (ICEET)*, pages 1–6. IEEE.

[Boneh and Franklin 2001] Boneh, D. and Franklin, M. (2001). Identity-based encryption from the weil pairing. In *Advances in Cryptology—CRYPTO 2001: 21st Annual International Cryptology Conference, Santa Barbara, California, USA, August 19–23, 2001 Proceedings*, pages 213–229. Springer.

[Boussada et al. 2019] Boussada, R., Hamdane, B., Elhdhili, M. E., and Saidane, L. A. (2019). Pp-ndnot: On preserving privacy in iot-based e-health systems over ndn. In *2019 IEEE Wireless Communications and Networking Conference (WCNC)*, pages 1–6. IEEE.

[Brezolin et al. 2022] Brezolin, U. Q., Prates Jr, N. G., Vergütz, A., and Nogueira, M. (2022). Um método para detecçao de vulnerabilidades através da análise do tráfego de rede iot. In *Anais do XL Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*, pages 447–460. SBC.

[Bui and Fonarow 2012] Bui, A. L. and Fonarow, G. C. (2012). Home monitoring for heart failure management. *Journal of the American College of Cardiology*, 59(2):97–104.

[Cicioğlu and Çalhan 2019] Cicioğlu, M. and Çalhan, A. (2019). Sdn-based wireless body area network routing algorithm for healthcare architecture. *ETRI Journal*, 41(4):452–464.

[Cicioğlu and Çalhan 2020] Cicioğlu, M. and Çalhan, A. (2020). Energy-efficient and sdn-enabled routing algorithm for wireless body area networks. *Computer Communications*, 160:228–239.

[Cornet et al. 2022] Cornet, B., Fang, H., Ngo, H., Boyer, E. W., and Wang, H. (2022). An overview of wireless body area networks for mobile health applications. *IEEE Network*, 36(1):76–82.

[Cussuol et al. 2022] Cussuol, E. B., Sachetti, L. L., Santos, B. P., and Mota, V. F. (2022). Otalab: um ambiente de experimentação remota de protocolos e aplicações em internet das coisas. In *Anais Estendidos do XL Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*, pages 73–80. SBC.

[Dang et al. 2019] Dang, L. M., Piran, M. J., Han, D., Min, K., and Moon, H. (2019). A survey on internet of things and cloud computing for healthcare. *Electronics*, 8(7):768.

[Demiris et al. 2004] Demiris, G., Rantz, M. J., Aud, M. A., Marek, K. D., Tyrer, H. W., Skubic, M., and Hussam, A. A. (2004). Older adults' attitudes towards and perceptions of 'smart home'technologies: a pilot study. *Medical informatics and the Internet in medicine*, 29(2):87–94.

[Devi et al. 2023] Devi, D. H., Duraisamy, K., Armghan, A., Alsharari, M., Aliqab, K., Sorathiya, V., Das, S., and Rashid, N. (2023). 5g technology in healthcare and wearable devices: A review. *Sensors*, 23(5):2519.

[Du et al. 2020] Du, Z., Wu, C., Yoshinaga, T., Yau, K.-L. A., Ji, Y., and Li, J. (2020). Federated learning for vehicular internet of things: Recent advances and open issues. *IEEE Open Journal of the Computer Society*, 1:45–61.

[Dulal et al. 2022] Dulal, S., Ali, N., Thieme, A. R., Yu, T., Liu, S., Regmi, S., Zhang, L., and Wang, L. (2022). Building a secure mhealth data sharing infrastructure over ndn. In *Proceedings of the 9th ACM Conference on Information-Centric Networking*, pages 114–124.

[Fan et al. 2017] Fan, X., Susan, F., Long, W., and Li, S. (2017). Security analysis of zigbee. *MWR InfoSecurity*, 2017:1–18.

[Fersi 2020] Fersi, G. (2020). Study of middleware for internet of healthcare things and their applications. In *The Impact of Digital Technologies on Public Health in Developed and Developing Countries: 18th International Conference, ICOST 2020, Hammamet, Tunisia, June 24–26, 2020, Proceedings 18*, pages 223–231. Springer.

[Ghayyur et al. 2020] Ghayyur, S., Pappachan, P., Wang, G., Mehrotra, S., and Venkatasubramanian, N. (2020). Designing privacy preserving data sharing middleware for internet of things. In *Proceedings of the Third Workshop on Data: Acquisition To Analysis*, pages 1–6.

[Handley 2018] Handley, M. (2018). Delay is not an option: Low latency routing in space. In *Proceedings of the 17th ACM Workshop on Hot Topics in Networks*, pages 85–91.

[Hasan et al. 2022] Hasan, M. K., Ghazal, T. M., Saeed, R. A., Pandey, B., Gohel, H., Eshmawi, A., Abdel-Khalek, S., and Alkhassawneh, H. M. (2022). A review on security threats, vulnerabilities, and counter measures of 5g enabled internet-of-medical-things. *IET Communications*, 16(5):421–432.

[Heart and Kalderon 2013] Heart, T. and Kalderon, E. (2013). Older adults: are they ready to adopt health-related ict? *International journal of medical informatics*, 82(11):e209–e231.

[Heemann and Hermsdorf 2017] Heemann, M. and Hermsdorf, M. (2017). Custo de internações de idosos é 30% maior para o sus. Disponível em https://infograficos.estadao.com.br/focas/planeje-sua-vida/custo-de-internacoes-de-idosos-e-30-maior-para-o-sus. Acessado: 2023-03-12.

[Heideker et al. 2019] Heideker, A., Ottolini, D., Zyrianoff, I., Kleinschmidt, J., and Kamienski, C. (2019). Imaiot-infrastructure monitoring agent for iot: Um agente monitor de infraestruturas para ambientes de iot. In *Anais Estendidos do XXXVII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*, pages 9–16. SBC.

[IBGE 2018] IBGE (2018). Projeção da população 2018: número de habitantes do país deve parar de crescer em 2047. Disponível em https://agenciadenoticias.ibge.gov.br/agencia-sala-de-imprensa/2013-agencia-de-noticias/releases/21837-projecao-da-populacao-2018-numero-de-habitantes-do-pais-deve-parar-de-crescer-em-2047. Acessado: 2023-03-12.

[Istepanian and Lacal 2003] Istepanian, R. S. and Lacal, J. C. (2003). Emerging mobile communication technologies for health: some imperative notes on m-health. In *Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (IEEE Cat. No. 03CH37439)*, volume 2, pages 1414–1416. IEEE.

[Jacobson et al. 2009] Jacobson, V., Smetters, D. K., Thornton, J. D., Plass, M. F., Briggs, N. H., and Braynard, R. L. (2009). Networking named content. In *Proceedings of the 5th international conference on Emerging networking experiments and technologies*, pages 1–12. ACM.

[Jones et al. 2010] Jones, V., Gay, V., and Leijdekkers, P. (2010). Body sensor networks for mobile health monitoring: Experience in europe and australia. In *2010 Fourth International Conference on Digital Society*, pages 204–209. IEEE.

[Kakkar 2020] Kakkar, A. (2020). A survey on secure communication techniques for 5g wireless heterogeneous networks. *Information Fusion*, 62:89–109.

[Kamboj et al. 2021] Kamboj, P., Pal, S., and Mehra, A. (2021). A qos-aware routing based on bandwidth management in software-defined iot network. In *2021 IEEE 18th International Conference on Mobile Ad Hoc and Smart Systems (MASS)*, pages 579–584.

[Kim et al. 2014] Kim, H.-S., Lee, K.-H., Kim, H., and Kim, J. H. (2014). Using mobile phones in healthcare management for the elderly. *Maturitas*, 79(4):381–388.

[Kitsiou et al. 2017] Kitsiou, S., Paré, G., Jaana, M., and Gerber, B. (2017). Effectiveness of mhealth interventions for patients with diabetes: an overview of systematic reviews. *PloS one*, 12(3):e0173160.

[Konečnỳ et al. 2016] Konečnỳ, J., McMahan, H. B., Yu, F. X., Richtárik, P., Suresh, A. T., and Bacon, D. (2016). Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*.

[Koren and Prasad 2022] Koren, A. and Prasad, R. (2022). Iot health data in electronic health records (ehr): Security and privacy issues in era of 6g. *Journal of ICT Standardization*, pages 63–84.

[Kumar et al. 2018] Kumar, T., Ramani, V., Ahmad, I., Braeken, A., Harjula, E., and Ylianttila, M. (2018). Blockchain utilization in healthcare: Key requirements and challenges. In *2018 IEEE 20th International conference on e-health networking, applications and services (Healthcom)*, pages 1–7. IEEE.

[Kumari and Jain 2022] Kumari, N. and Jain, V. K. (2022). Fog based healthcare monitoring system in sdn-iot networks. In *2022 Second International Conference on Advanced Technologies in Intelligent Control, Environment, Computing & Communication Engineering (ICATIECE)*, pages 1–6.

[Li et al. 2020] Li, J., Cai, J., Khan, F., Rehman, A. U., Balasubramaniam, V., Sun, J., and Venu, P. (2020). A secured framework for sdn-based edge computing in iot-enabled healthcare system. *IEEE Access*, 8:135479–135490.

[Li et al. 2013] Li, J., Li, J., Chen, X., Jia, C., and Lou, W. (2013). Identity-based encryption with outsourced revocation in cloud computing. *IEEE Transactions on computers*, 64(2):425–437.

[Lin et al. 2020] Lin, Y., Jiang, D., Yus, R., Bouloukakis, G., Chio, A., Mehrotra, S., and Venkatasubramanian, N. (2020). Locater: cleaning wifi connectivity datasets for semantic localization. *arXiv preprint arXiv:2004.09676*.

[Madureira et al. 2020] Madureira, A. L. R., Araújo, F. R. C., and Sampaio, L. N. (2020). On supporting iot data aggregation through programmable data planes. *Computer Networks*, 177:107330.

[Madureira et al. 2019] Madureira, P., Cardoso, N., Sousa, F., and Moreira, W. (2019). My-aha: middleware platform to sustain active and healthy ageing. In *2019 International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*, pages 21–26. IEEE.

[Mangla et al. 2022] Mangla, C., Rani, S., Qureshi, N. M. F., and Singh, A. (2022). Mitigating 5g security challenges for next-gen industry using quantum computing. *Journal of King Saud University-Computer and Information Sciences*.

[Maskeliūnas et al. 2019] Maskeliūnas, R., Damaševičius, R., and Segal, S. (2019). A review of internet of things technologies for ambient assisted living environments. *Future Internet*, 11(12):259.

[McKeown 2009] McKeown, N. (2009). Software-defined networking. *INFOCOM keynote talk*, 17(2):30–32.

[Mehrotra et al. 2020] Mehrotra, S., Sharma, S., Ullman, J. D., Ghosh, D., Gupta, P., and Mishra, A. (2020). Panda: Partitioned data security on outsourced sensitive and non-sensitive data. *ACM Transactions on Management Information Systems (TMIS)*, 11(4):1–41.

[Meliá et al. 2021] Meliá, S., Nasabeh, S., Luján-Mora, S., and Cachero, C. (2021). Mosiot: modeling and simulating iot healthcare-monitoring systems for people with disabilities. *International Journal of Environmental Research and Public Health*, 18(12):6357.

[Misra et al. 2023] Misra, S., Pal, S., Ahmed, N., and Mukherjee, A. (2023). Sdn-controlled resource-tailored analytics for healthcare iot system. *IEEE Systems Journal*, pages 1–8.

[Misra et al. 2020] Misra, S., Saha, R., and Ahmed, N. (2020). Health-flow: Criticality-aware flow control for sdn-based healthcare iot. In *GLOBECOM 2020 - 2020 IEEE Global Communications Conference*, pages 1–6.

[Mukherjee et al. 2018] Mukherjee, B., Wang, S., Lu, W., Neupane, R. L., Dunn, D., Ren, Y., Su, Q., and Calyam, P. (2018). Flexible iot security middleware for end-to-end cloud–fog communication. *Future Generation Computer Systems*, 87:688–703.

[Nakayama et al. 2019a] Nakayama, F., Lenz, P., Banou, S., Nogueira, M., Santos, A., and Chowdhury, K. R. (2019a). A continuous user authentication system based on galvanic coupling communication for s-health. *Wireless Communications and Mobile Computing*, 2019:1–11.

[Nakayama et al. 2019b] Nakayama, F., Lenz, P., Cremonezi, B., Banou, S., Rosário, D., Chowdhoury, K., Nogueira, M., Cerqueira, E., and Santos, A. (2019b). Autenticação contínua e segura baseada em sinais ppg e comunicação galvânica. In *Anais do XXXVII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*, pages 707–720. SBC.

[Nakayama et al. 2022] Nakayama, F., Lenz, P., and Nogueira, M. (2022). A resilience management architecture for communication on portable assisted living. *IEEE Transactions on Network and Service Management*, 19(3):2536–2548.

[Narra et al. 2019] Narra, K. G., Lin, Z., Wang, Y., Balasubramaniam, K., and Annavaram, M. (2019). Privacy-preserving inference in machine learning services using trusted execution environments. *arXiv preprint arXiv:1912.03485*.

[Nogueira et al. 2021] Nogueira, M., Borges, L. F., and Nakayama, F. (2021). Das redes vestíveis aos sistemas ciber-humanos: Uma perspectiva na comunicação e privacidade dos dados. *Sociedade Brasileira de Computação*.

[Onesimu et al. 2022] Onesimu, J. A., Karthikeyan, J., Eunice, J., Pomplun, M., and Dang, H. (2022). Privacy preserving attribute-focused anonymization scheme for healthcare data publishing. *IEEE Access*, 10:86979–86997.

[Padgette et al. 2017] Padgette, J., Scarfone, K., and Chen, L. (2017). Guide to bluetooth security. *NIST special publication*, 800(121).

[Papaioannou et al. 2022] Papaioannou, M., Karageorgou, M., Mantas, G., Sucasas, V., Essop, I., Rodriguez, J., and Lymberopoulos, D. (2022). A survey on security threats and countermeasures in internet of medical things (iomt). *Transactions on Emerging Telecommunications Technologies*, 33(6):e4049.

[Pereira et al. 2021] Pereira, R. S., Prazeres, C. V. S., Barbosa, M. T. M., Barros, E. B. C., and Peixoto, M. L. M. (2021). Iotfogsim: Um simulador orientado a eventos para avaliaçao de aplicaçoes baseadas em iot-fog-cloud. In *Anais Estendidos do XXXIX Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*, pages 25–32. SBC.

[Perez et al. 2022] Perez, A. J., Siddiqui, F., Zeadally, S., and Lane, D. (2022). A review of iot systems to enable independence for the elderly and disabled individuals. *Internet of Things*, page 100653.

[Philip et al. 2021] Philip, N. Y., Rodrigues, J. J., Wang, H., Fong, S. J., and Chen, J. (2021). Internet of things for in-home health monitoring systems: Current advances, challenges and future directions. *IEEE Journal on Selected Areas in Communications*, 39(2):300–310.

[Price et al. 2020] Price, A. B., Rarity, J. G., and Erven, C. (2020). A quantum key distribution protocol for rapid denial of service detection. *EPJ Quantum Technology*, 7(1):8.

[Qu et al. 2017] Qu, Z., Zhang, G., Cao, H., and Xie, J. (2017). Leo satellite constellation for internet of things. *IEEE access*, 5:18391–18401.

[Rashidi and Mihailidis 2012] Rashidi, P. and Mihailidis, A. (2012). A survey on ambient-assisted living tools for older adults. *IEEE journal of biomedical and health informatics*, 17(3):579–590.

[Rodrigues et al. 2018] Rodrigues, J. J., Segundo, D. B. D. R., Junqueira, H. A., Sabino, M. H., Prince, R. M., Al-Muhtadi, J., and De Albuquerque, V. H. C. (2018). Enabling technologies for the internet of health things. *Ieee Access*, 6:13129–13141.

[Salama et al. 2019] Salama, M., Elkhatib, Y., and Blair, G. (2019). Iotnetsim: A modelling and simulation platform for end-to-end iot services and networking. In *Proceedings of the 12th IEEE/ACM International Conference on Utility and Cloud Computing*, pages 251–261.

[Sampaio et al. 2021] Sampaio, L. N., Freitas, A. E. S., Araújo, F. R., Brito, I. V. S., and Ribeiro, A. V. (2021). Revisitando as icns: Mobilidade, segurança e aplicações

distribuídas através das redes de dados nomeados. In *Livro de Minicursos do XXXIX Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC) 2021*.

[Saxena and Raychoudhury 2017] Saxena, D. and Raychoudhury, V. (2017). Design and verification of an ndn-based safety-critical application: A case study with smart healthcare. *ieee transactions on systems, man, and cybernetics: systems*, 49(5):991–1005.

[Shang et al. 2016] Shang, W., Bannis, A., Liang, T., Wang, Z., Yu, Y., Afanasyev, A., Thompson, J., Burke, J., Zhang, B., and Zhang, L. (2016). Named data networking of things. In *2016 IEEE first international conference on internet-of-things design and implementation (IoTDI)*, pages 117–128. IEEE.

[Shi et al. 2015] Shi, Y., Zheng, Q., Liu, J., and Han, Z. (2015). Directly revocable key-policy attribute-based encryption with verifiable ciphertext delegation. *Information Sciences*, 295:221–231.

[Souppaya and Scarfone 2012] Souppaya, M. and Scarfone, K. (2012). Guidelines for securing wireless local area networks (wlans). *NIST Special Publication*, 800:153.

[Srivastava et al. 2020] Srivastava, G., Agrawal, R., Singh, K., Tripathi, R., and Naik, K. (2020). A hierarchical identity-based security for delay tolerant networks using lattice-based cryptography. *Peer-to-Peer Networking and Applications*, 13:348–367.

[Steele et al. 2009] Steele, R., Lo, A., Secombe, C., and Wong, Y. K. (2009). Elderly persons' perception and acceptance of using wireless sensor networks to assist healthcare. *International journal of medical informatics*, 78(12):788–801.

[Tshiningayamwe et al. 2016] Tshiningayamwe, L., Lusilao-Zodi, G.-A., and Dlodlo, M. E. (2016). A priority rate-based routing protocol for wireless multimedia sensor networks. In *Advances in Nature and Biologically Inspired Computing: Proceedings of the 7th World Congress on Nature and Biologically Inspired Computing (NaBIC2015) in Pietermaritzburg, South Africa, held December 01-03, 2015*, pages 347–358. Springer.

[Tun et al. 2021] Tun, S. Y. Y., Madanian, S., and Mirza, F. (2021). Internet of things (iot) applications for elderly care: a reflective review. *Aging clinical and experimental research*, 33:855–867.

[Uddin et al. 2019] Uddin, M., Nadeem, T., and Nukavarapu, S. (2019). Extreme sdn framework for iot and mobile applications flexible privacy at the edge. In *2019 IEEE International Conference on Pervasive Computing and Communications (PerCom*, pages 1–11.

[United Nations 2020] United Nations (2020). World population ageing 2020 highlights: Living arrangements of older persons (st/esa/ser.a/451). Disponível em https://www.un.org/development/desa/pd/sites/www.un.org.development.desa.pd/files/undesa_pd-2020_world_population_ageing_highlights.pdf. Acessado: 2023-03-12.

[Wang et al. 2022] Wang, Z., Xiong, H., Zhang, J., Yang, S., Boukhechba, M., Zhang, D., Barnes, L. E., and Dou, D. (2022). From personalized medicine to population health: a survey of mhealth sensing techniques. *IEEE Internet of Things Journal*.

[Xu et al. 2019] Xu, S., Yang, G., and Mu, Y. (2019). Revocable attribute-based encryption with decryption key exposure resistance and ciphertext delegation. *Information Sciences*, 479:116–134.

[Yang et al. 2020] Yang, P., Xiong, N., and Ren, J. (2020). Data security and privacy protection for cloud storage: A survey. *IEEE Access*, 8:131723–131740.

[Yaseen et al. 2022] Yaseen, F. A., Alkhalidi, N. A., and Al-Raweshidy, H. S. (2022). She networks: Security, health, and emergency networks traffic priority management based on ml and sdn. *IEEE Access*, 10:92249–92258.

[Zgheib et al. 2019] Zgheib, R., Conchon, E., and Bastide, R. (2019). Semantic middleware architectures for iot healthcare applications. In *Enhanced Living Environments: Algorithms, Architectures, Platforms, and Systems*, pages 263–294. Springer.

[Zhang et al. 2014] Zhang, L., Afanasyev, A., Burke, J., Jacobson, V., Claffy, K., Crowley, P., Papadopoulos, C., Wang, L., and Zhang, B. (2014). Named Data Networking. *SIGCOMM Comput. Commun. Rev.*, 44(3):66–73.

[Zhang et al. 2020] Zhang, Y., Chen, G., Du, H., Yuan, X., Kadoch, M., and Cheriet, M. (2020). Real-time remote health monitoring system driven by 5g mec-iot. *Electronics*, 9(11):1753.

[Zhang et al. 2018] Zhang, Z., Yu, Y., Ramani, S. K., Afanasyev, A., and Zhang, L. (2018). Nac: Automating access control via named data. In *MILCOM 2018-2018 IEEE Military Communications Conference (MILCOM)*, pages 626–633. IEEE.

Capítulo 3

Intrusion detection with Machine Learning in Internet of Things and Fog Computing: problems, solutions and research

Cristiano Antonio de Souza, Carlos Becker Westphall and Renato Bobsin Machado

Abstract

Intrusion detection is one of the key points in computer security, and it aims to identify attempted attacks by unauthorized users. Several researches are being developed to solve security problems in environments involving the Internet of Things, Fog Computing, and Cloud Computing. This mini-course has a theoretical and practical profile, aims to describe aspects of the context of intrusion detection in IoT and Fog Computing, presents Machine Learning techniques commonly used in intrusion detection, expose state-of-the-art approaches, and present some results obtained in developed research.

1.1. Introduction

With the development of technological resources and the popularization of the Internet, there has been significant growth in the number of computational applications. Faced with this new technological context, difficulties have arisen in maintaining the security of applications and data, given that the techniques for exploiting vulnerabilities in these computational infrastructures are constantly being improved to acquire access to systems and obtain and use improperly sensitive information.

Malicious users can exploit vulnerabilities in computer systems to carry out illicit activities. The attackers' main motivation is to obtain privileged digital content that can bring some benefit to the attacker and/or cause significant damage to the target of the attacks.

Currently, the Internet of Things (IoT) is spreading in all areas that apply computational resources. IoT devices allow everyday objects to be connected to the Internet, computers, and smartphones [Atzori et al. 2010]. The idea is to increasingly unite the physical and digital worlds by communicating objects with other devices, data centers, and clouds.

IoT devices have limited resources [Atzori et al. 2010]. There is a need to transfer, via the Internet, the data generated by these devices to process and store them in a computational center with greater capacity [Miorandi et al. 2012]. Many IoT applications use Cloud Computing to process and store data [Al-Fuqaha et al. 2015]. However, with the growth of IoT, applications began to deal with generating large amounts of data. Consequently, requiring heavy computational resources like bandwidth [Bonomi et al. 2012]. This large amount of data results in network congestion in the communication of IoT devices with data centers of cloud computing [Roman et al. 2018]. Fog computing provides services closer to the end devices [Bonomi et al. 2012, Roman et al. 2018]. Performing temporary data processing and storage close to IoT devices decreases the traffic sent to the cloud [Roman et al. 2018]. In addition, it allows applications that need real-time processing to obtain a faster response [Al-Fuqaha et al. 2015]. In this way, they do not send data from devices to the cloud.

IoT allows various physical objects to see, hear, feel, think, and communicate in the environment in which they are inserted. In this way, each object can share information with other objects and make decisions to perform certain tasks [Atzori et al. 2010]. As a consequence of this scenario, large amounts of information of the most varied natures are processed. Much of this data is confidential and private. Sensitive user data is collected, processed, transmitted, and stored via fog computing and IoT components [Roman et al. 2018]. In this context, attackers may be able to compromise a smart home system, for example, and discover confidential information about the family's habits, such as the time that residents sleep or that the house is empty, among others. This information can later be used to cause serious harm to victims [Ni et al. 2018].

The resource constraints of IoT devices make them susceptible to flaws and malicious data integrity attacks [Neshenko et al. 2019]. This can lead to unreliability and sometimes system collapse. One of the main objectives of an attack on an IoT network is to disrupt the availability of data sent from IoT devices to applications. This interruption can be achieved in several ways, such as overloading devices with information requests or compromising the network structure by dropping packets [Roman et al. 2018].

Smart environments are becoming real and possible through IoT. However, as mentioned above, they are also not free from security threats and vulnerabilities. In this context, in parallel with technological growth, there are also difficulties in maintaining the security of applications and computational infrastructures, considering that vulnerabilities also increase with the growth in the number of available services. A major incident involving IoT devices occurred in October 2016. An attack involving IoT devices via botnet Mirai against service provider Dyn took offline for several hours, hundreds from sites, including Twitter, Netflix, Reddit, and GitHub [Tanaka and Yamaguchi 2017].

This makes special security techniques indispensable in modern computer systems. According to [Roman et al. 2018], security is one of the biggest challenges to ensure an ideal IoT and fog computing environment, where devices can take advantage of the services provided by the paradigm. Intrusion detection is one of the key security points to identify attempted attacks.

There are several state-of-the-art approaches to detecting intrusions in an IoT environment. Some works focus on signature detection. These approaches fail to detect new

attacks or variations of known attacks [Arshad et al. 2019]. In addition, specification-based methods have also been proposed. However, these approaches require a human expert to specify the expected behavior of the network. Finally, other approaches have proposed anomaly-based methods to detect intrusions [Labiod et al. 2022, Rey et al. 2022, Souza et al. 2022a]. Anomaly detection considers that all abnormal behavior is an intrusion and thus can detect new attacks or variations of known attacks. Machine Learning (ML) methods are commonly applied in this context [Boukerche et al. 2007]. However, anomaly-based approaches often require features that IoT nodes lack. [Ahmad et al. 2021] highlight the resources consumed by complex models and the need for lightweight IDS for IoT. Thus, most of the studies work with fog computing.

Several state-of-the-art approaches focus on anomaly methods for binary detection (attack or non-attack) [Albdour et al. 2020, Kumar et al. 2021a, Rey et al. 2022]. They can detect an intrusion, but not the type or category of attack. In this context of intrusion detection, the approach must identify the attack category so that more specific countermeasures can be implemented for the given type of threat. Identifying the type or category of the attack is also important for the decision-making process of the person responsible for the network [Souza et al. 2022b]. Several multiclass approaches have been proposed in state of the art. However, multiclass detection approaches are generally more complex, have a higher computational cost, and have lower accuracy rates than binary methods [Nguyen et al. 2019]. This is justified by the difficulties in identifying specific types of attacks [Diro and Chilamkurti 2018, Kumar et al. 2020a].

The prohibitive cost is another important point, as the resource constraints present in IoT and Fog computing environments limit the design of robust approaches. Robust and slow multi-class analysis performed on IoT/Fog can overload the device and cause network flow delay [Nguyen et al. 2019]. Furthermore, the use of attribute selection and class balancing techniques, useful to improve detection performance, tend to increase the training cost of the approaches. Therefore, intrusion detection in these environments has challenges and research opportunities.

This mini-course presents the main concepts in this context, and several machine-learning techniques used to detect intrusions are addressed. From this, practical activity is proposed for the execution of experiments with the studied techniques. Subsequently, it is presented in the state of the art how these techniques have been applied to detect intrusions in the IoT environment. Finally, the main problems, challenges, and research opportunities in state of the art are discussed.

1.1.1. Organization of the minicourse

The remainder of the short course is organized as described below. Section 1.2 presents the fundamental concepts involved in the theme of this work. The Internet of Things (IoT), Cloud Computing, and Fog Computing concepts are presented. In addition, the main threats present in IoT environments are also discussed, and concepts related to Intrusion Detection Systems (IDS) are introduced. Next, Section 1.3 discusses applying machine learning in the context of intrusion detection. Some classification techniques that can be used for intrusion analysis and detection are discussed. In addition, some practical aspects of each technique are presented. This section also proposes to conduct simulation

experiments with the IoTID20 dataset to evaluate the techniques presented in an intrusion detection scenario with IoT traffic. In Section 1.4, state of the art is exposed, and the approaches proposed by the main related works are presented. Section 1.5 discusses some important aspects observed in state-of-the-art related to intrusion detection in an IoT/Fog/Cloud context. The objective is to instigate an initial reflection on this research topic's problems, challenges, and open questions. Finally, Section 1.6 concludes the mini-course and presents final considerations and direction for future work.

1.2. Concepts and Technologies

In this section, concepts related to the theme of this work are addressed. Initially, the concepts of the Internet of Things (IoT), Cloud Computing, and Fog Computing, present in the environment chosen for this scenario, are contextualized. Section 1.2.4 discusses the main threats in IoT environments. Then, important aspects of the Intrusion Detection System (IDS) are introduced.

1.2.1. Internet of Things

The Internet of Things (IoT) has as its basic characteristic the pervasive presence of a wide variety of intelligent objects in people's daily lives, such as sensors, tags of Radio-Frequency IDentification (RFID), mobile phones, among others [Atzori et al. 2010]. The IoT connects physical devices to the Internet, enabling them to communicate and act intelligently.

From a conceptual point of view, IoT is based on three basic principles related to the characteristics of smart objects: being identifiable, communicable, and capable of interacting with the environment in which they are inserted [Miorandi et al. 2012]. IoT allows various physical objects to see, hear, feel, think, and communicate to share information and make decisions to perform certain tasks.

IoT applications can improve people's lives and how they live, work, learn, and have fun. For example, smart homes can provide residents with certain practicalities, such as automatic garage openings, automatic coffee preparation, climate control systems, etc.

IoT devices are small physical objects with limited processing and storage capabilities. Due to the large amount of data generated by these devices, there is a need for greater computational capacity. Furthermore, the number of devices connected to the Internet continues to grow. Cisco predicts that the number of interconnected devices on the planet could reach the 500 billion mark by 2025 [Camhi 2015]. Cloud computing can be a solution to solve these needs for greater processing capacity.

1.2.2. Cloud Computing

The National Institute of Standards and Technology (NIST) defines Cloud Computing as a model for enabling ubiquitous and convenient network access to a shared pool of configurable computing resources that can be quickly provisioned and released with minimal management effort or interaction between service providers [Mell et al. 2011].

According to the authors [Takabi et al. 2010], Cloud Computing is an important paradigm with the potential to significantly reduce costs through optimization and increased operational and economic efficiencies. They point out that this paradigm can significantly improve collaboration, agility, and scalability, enabling a truly global computing model over the Internet's infrastructure.

Cloud Computing has five essential characteristics: on-demand self-service, ubiquitous network access, pooling of resources, location independence, rapid elasticity, and measured service, all aimed at transparently using clouds. In the cloud, the provider's computing resources are pooled to serve multiple consumers using a multi-tenancy model. Different physical and virtual components are dynamically assigned and reallocated ac-

cording to consumer demand. Providers must provide rapid elasticity, allowing the consumer to increase or decrease resources and on-demand service so that the customer can unilaterally allocate resources dynamically [Takabi et al. 2010]. Cloud providers can handle large amounts of data and high processing rates.

The basic characteristics of cloud computing make it an important processing mechanism for IoT applications that capture large amounts of information. However, its use also has disadvantages, as this centralization of processing and storage resources implies a great separation between the physical IoT devices and the data centers of the cloud. This fact, which according to [Satyanarayanan 2015], results in the growth of average latency and jitter.

Then came Fog Computing, capable of solving the abovementioned problems for IoT applications. It extends the cloud closer to the user so that data access, processing, and storage tasks are performed by local resources such as routers, gateways, and switches. Therefore, the processing and storage of temporary data and the execution of local analyzes are carried out without long transmissions over the Internet. This way, fog computing doesn't suffer from high latency and jitter problems [Ni et al. 2018].

1.2.3. Fog Computing

Authors [Bonomi et al. 2012] define Fog Computing as a highly virtualized platform that provides computing, storage, and network services between IoT devices and cloud data centers. In addition, it is generally close to IoT devices at the edge of the network, as seen in Figure 1.1.

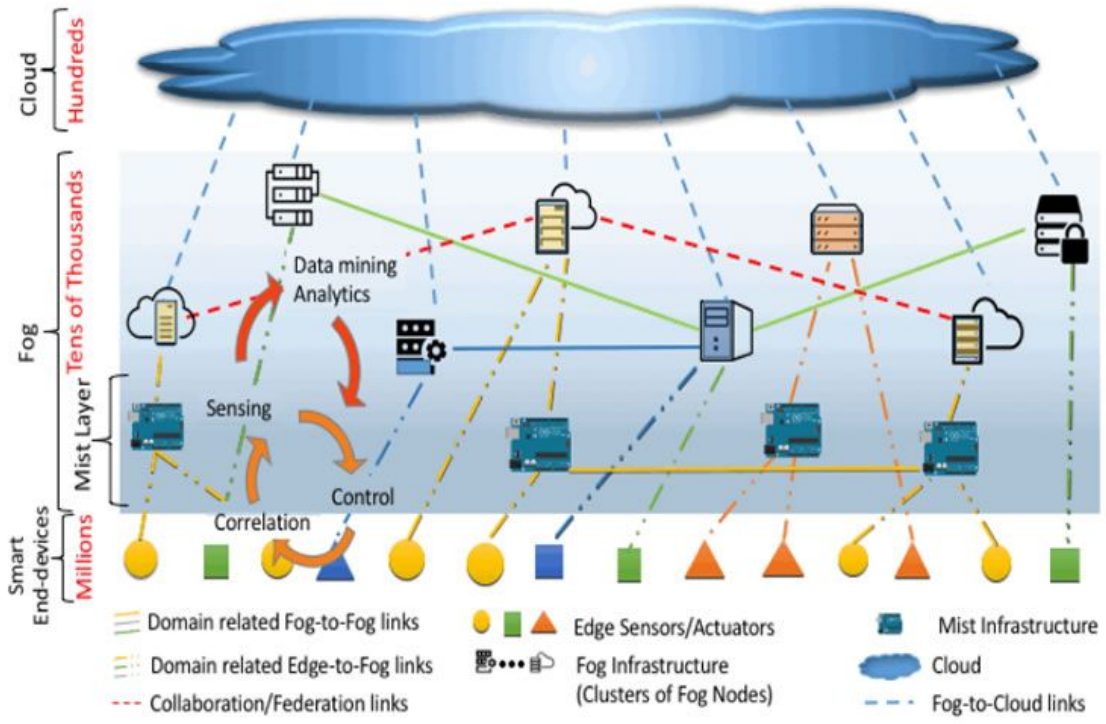

Figure 1.1. Illustration of existing layers in an environment based on the IoT-Fog-Cloud architecture [Iorga et al. 2018].

The main idea of this paradigm is to extend cloud computing closer to end devices to provide efficient data access, processing, and storage. Therefore, the hallmark of fog computing is the distribution of resources, communication services, processing, and storage close to users [Marín-Tordera et al. 2017].

Fog computing did not emerge to replace cloud computing for remote processing and storage but rather to complement it. Allowing the creation of a hierarchical infrastructure where local data is processed and stored by fog computing and permanent storage and global analysis are performed in the *datacenters* of the cloud [Ni et al. 2018].

As it is a recent paradigm, research into security and privacy issues is still at an early stage.

1.2.4. Threats in IoT and Fog Computing

This work considers the context of intelligent environments based on fog computing and IoT. Security in these environments is paramount, as IoT devices are often embedded in people's daily lives and handle sensitive information. In addition, some systems perform monitoring and critical actions, which need uninterrupted operation.

IoT and fog solutions comprise various technologies, services, and standards, each with security and privacy requirements [Zarpelão et al. 2017]. The IoT paradigm presents several security vulnerabilities that communication networks, cloud services, and the Internet have [Zarpelão et al. 2017]. However, traditional security tools have difficulties to be applied directly in this context due to three fundamental aspects: the limited computational power of IoT components, the high number of interconnected devices, and the sharing of data between objects and users [Sicari et al. 2015]. Furthermore, the rapid expansion of IoT solutions has left these networks vulnerable to security and privacy risks. Authors [Kolias et al. 2016] discovered several security vulnerabilities by creating IoT use cases using popular commercial products and services. Fog computing has emerged to provide greater computing resources for the IoT and low latency and compute-intensive use. These facts make it a great place to deploy IoT security applications like intrusion detectors [Nguyen et al. 2019]. However, security research on fog computing and IoT applications is still at an early stage [Ni et al. 2018]. This fact, combined with the great damage that attacks in this environment can cause, generates the need to concentrate efforts in this area. The authors [Garcia-Morchon et al. 2013] organized threats to the security of IoT environments into the following categories: device cloning, malicious replacement of devices, replacement of firmware, extraction of security parameters, interception, Man-In-The-Middle (MITM), routing attack, Denial of Service (DoS) and Distributed Denial of Service (DDoS). Threats related to cloning, replacement, and extraction usually occur during device manufacture, installation, maintenance, and updating [Zarpelão et al. 2017].

[Kolias et al. 2016] highlight the main attacks present in IoT: DoS, DDoS, MITM, routing attacks, and conventional attacks. Security threats related to conventional technologies that are part of the IoT environment may also apply, for example, insecure connections, malicious code injection, probing, interception, fabrication, and modification of messages [Muhammad et al. 2015].

As illustrated in Figure 1.2, these environments are subject to attacks from external sources, from the Internet, and internal attacks, by malicious devices in the IoT network. Malicious entities outside the network may attempt to gain privileged access to IoT devices to control them [Muhammad et al. 2015]. Through this access, it is possible to carry out botnet attacks, where compromised IoT devices can be used as bots or zombies to perform various malicious tasks [Aversano et al. 2021]. Furthermore, spoofing attacks involve impersonating legitimate devices, exploiting their identities to gain access to the IoT network, and then launching other types of malicious actions [Aversano et al. 2021], such as stealing confidential information handled in the IoT network. DoS attacks are quite common and aim to affect the victim's availability. This can be done by flooding with a large volume of requests or depleting resources such as memory and computing power. In the context of IoT, the device can be part of the network under threat or be used as a zombie to launch a DDoS on another network. Probing attacks, where the attacker scans a network to gather information and discover vulnerabilities, are common. This attack usually collects information from the target before launching another, more severe type of attack [Arshad et al. 2019].

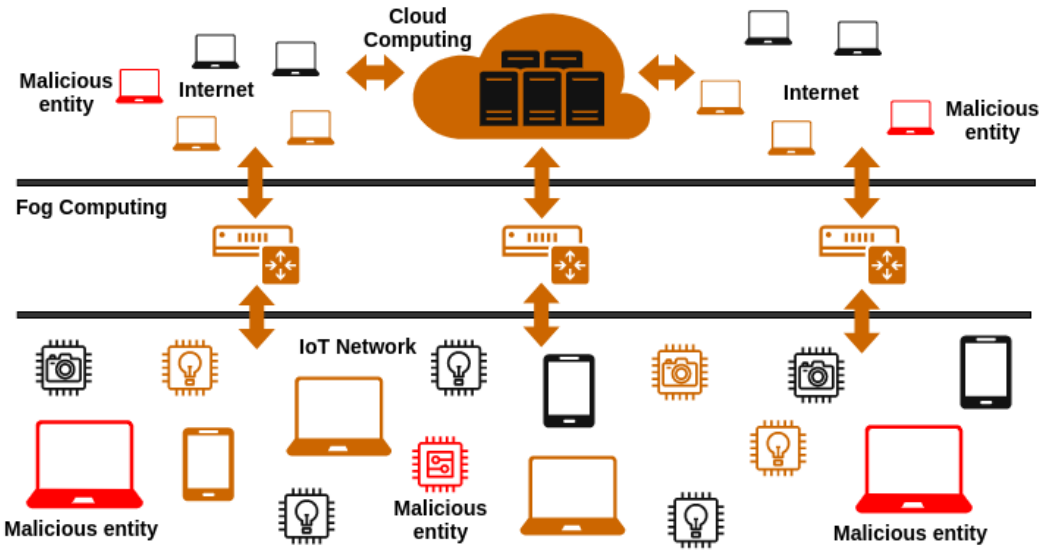

Figure 1.2. Illustration of potential threats in an IoT and fog computing environment.

In the case of an insider attack, the attacker is presumed to be a malicious entity that was successfully authenticated in the fog or a legitimate IoT device that has become malicious over time. It can carry out various attacks, including denial of service, by sending a high data packet rate in the fog. This allows overloading the devices and the upper layers, as illustrated in Figure 1.2, impairing or even causing the interruption of services provided by the systems, services that in many cases are extremely important. Therefore, these actions can damage the entire IoT system, and external users who access services and information generated by the IoT solution can be harmed.

Based on the classification presented by the [Zarpelão et al. 2017] authors, the main attacks found in the fog and IoT environment are presented below.

- Denial of Service (DoS): Denial of Service attacks aim to affect the availability of the victim. This can be done by flooding with a huge volume of requests or depleting resources like memory and computational power [Zarpelão et al. 2017]. Both an IoT node and the fog can fall victim to this attack.

- Distributed Denial of Service (DDoS): Distributed denial of service attacks have the same objective as a DoS. However, they are executed by a set of hosts, whereas in a common DoS, a single host is the attacker [Yan et al. 2016]. Furthermore, in the context of the Internet of Things, the IoT node can be part of the network under threat or be used as a zombie to launch a DDoS on another network.

- Man-In-The-Middle (MITM): a man-in-the-middle attack is performed when an attacker interferes with the communication between an entity A and an entity B, without A and B realizing it [Zarpelão et al. 2017]. Authors [Navas et al. 2018] demonstrate the risks of MITM attacks on IoT networks caused by malicious insider devices.

- Routing attacks: Routing attacks consist of spoofing, modifying network routing information to create loops, attracting or rejecting traffic, extending or shortening routes, etc. Other possible routing attacks include sinkhole attack, selective forwarding, wormhole attack, and sybil attack [Zarpelão et al. 2017].

- Conventional attacks: Security threats related to conventional technologies that are part of the IoT environment can also apply to IoT systems, for example, insecure connections, malicious code injection, interception, probing, fabrication, and modification of messages [Muhammad et al. 2015].

In addition to the existing threats related to computer networks, IoT needs to deal with the resource constraints that its devices have. Intrusion detection and prevention systems can be used to secure IoT networks. The basic concepts related to these systems are presented below.

1.2.5. Intrusion Detection

An intrusion can be defined as a set of actions to overcome an application's defense barriers to compromise the integrity, confidentiality, and availability of computational resources [Heady et al. 1990]. Intrusion Detection Systems (IDS) are intended to recognize intrusive actions and behavior to alert administrators or automatically execute countermeasures [Bace and Mell 2001].

Intrusion detection research efforts have been conducted since 1980. Around that time, [Anderson 1980] presented a threat model and security monitoring system based on detecting anomalies in user behavior. IDSs are inserted as the last line of defense within a computational architecture, making them of great importance, making it possible to infer the legitimacy of actions taken and having a proactive behavior in attack situations [Patel et al. 2010]. The structure of the IDSs can vary in relation to the way it is implemented, the frequency of operation, the data they analyze, or the analyzes carried out on them [Campello and Weber 2001].

IDSs can be classified according to the detection methods employed. Thus, they can be classified into analysis by signature or behavior, also known as analysis by anomaly.

In signature detection, monitored actions are compared with predefined intrusive events, normally stored in a database. These previously known patterns are called signatures. Signature detection allows quick detection and reduces the occurrence of false alarms. However, it has the limitation of detecting only known attacks, that is, only attacks with a signature known by the IDS [Northcutt et al. 2001]. Most commercial antivirus systems use this strategy [Bace and Mell 2001].

Anomaly detection assumes that any abnormal activity is necessarily an intrusion, and any activity that does not fit the defined normal behavior models is considered an attack. The great advantage of the anomaly detection technique is that it allows the detection of new attacks and/or variations of already known ones since it is not necessary to know about them previously. However, this technique is more likely to suffer from problems related to false positives [Boukerche et al. 2007]. This strategy is usually modeled using Machine Learning techniques. Section 1.3 presents more details about these techniques.

In addition, some works consider a branch of analysis by behavior called analysis by specification [Mitchell and Chen 2014]. This type of solution employs rules and thresholds that define the expected default behavior for monitored components. It is similar to anomaly detection, and both detect intrusions when network behavior deviates from specified. The main difference is that in specification-based analysis, a human expert sets the rules [Mitchell and Chen 2014, Zarpelão et al. 2017]. This type of analysis's major drawback is the specificity and domain knowledge required to specify benign behavior.

IDSs can analyze data from multiple sources and can be deployed in different locations. These data are generally related to how the approach is implemented. There are two main categories of implementation related to capturing information. The Host-Based Intrusion Detection System (HIDS) seeks to analyze the information captured from the very host where they are deployed, and Network-Based Intrusion Detection System (NIDS) analyzes traffic captured from the monitored network [Zarpelão et al. 2017]. Furthermore, in the context of IoT, approaches can be deployed at different levels: on IoT devices themselves, on devices in the fog, or in the cloud.

In a host-based IDS, all components, from event collection to classification, are located on the same host. The event monitoring and analysis mechanisms only use information from the host itself. Events can originate from system logs and data about users, services, and processes. This approach enables network independence and the detection of insider attacks. However, host-based solutions employed on IoT nodes may suffer from memory constraints. Those employed in fog devices allow the detection of attacks against the device itself but may have difficulty dealing with attacks on the network and IoT devices.

On the other hand, NIDS approaches are implemented in a device capable of capturing the network traffic intended to be monitored. Events and activities are obtained by capturing network traffic in promiscuous mode. These IDSs typically monitor a network made up of multiple devices. Sensors can also be used to capture information

in a distributed manner at various points in the network. One of this approach's difficulties is determining the best places to position the information capture sensors. The analysis method, in these approaches was generally included in the fog devices. This strategy allows the detection of external attacks and is more independent of the platform [Mukherjee et al. 1994]. Fog computing is one of the most promising alternatives for implementing IoT network monitoring approaches. Furthermore, it is interesting to divide the detection tasks along the complete architecture, considering IoT devices, fog, and cloud.

In addition to detecting intrusions, it is very important to have mechanisms to execute countermeasure actions, with the aim of blocking and preventing the intrusion from succeeding. Among the existing actions are issuing alerts to the network manager. Issuing just one alert does not configure a prevention action, as it only makes the manager aware that an intrusion has occurred, but does not prevent it. Issuing an alert is considered a passive post-detection. Another class of post-detection approaches is the active one, where the actions taken aim to stop an attack in progress and then block the attacker's access [Bace and Mell 2001]. IDSs that have active countermeasures are known as Intrusion Prevention Systems (IPS) [Birkinshaw et al. 2019].

In the following sections, other concepts involved in the context of this work are presented. Section 1.3 initially presents the basic concepts of ML and a brief description of its applicability for intrusion detection. Next, several classification techniques that are employed in behavioral detection approaches are presented.

1.3. Machine Learning techniques employed for intrusion detection

In [Russell and Norvig 2010], several definitions are presented for the term Artificial Intelligence (AI), which, in general, point to this as the ability to make machines reproduce intelligent activities and cognitive abilities found in humans. When we analyze more emphatically the currently proposed solutions in the area of computational security, there is increasingly stronger research and application of machine learning methods for improvements in the intrusion detection process.

According to [Goodfellow et al. 2016], ML is the ability of a given technique to acquire its knowledge, extracting information from raw data and representing it through some kind of mathematical model. ML has several sub-areas, including classification. A classification task consists of classifying data and objects into certain classes in an automated way.

It is observed that this ability of the machine learning classification methods fits perfectly with the context of intrusion detection since the detection approaches have the task of analyzing the information captured from the network or hosts, verifying the occurrence of abnormal behaviors, and performing the classification of information in benign or intrusive. In addition, there is also a need to classify attacks into types or categories. Therefore, classification techniques are great for composing anomaly-based detection approaches.

Machine learning methods usually need to train to acquire knowledge and generate a model with added knowledge. Next, the two most common types of learning employed in solutions found in the state of the art of intrusion detection are presented.

The main characteristic of approaches based on supervised learning is the existence of labels in the subset of training data. This type of learning reflects an algorithm's ability to generalize knowledge from available data with target or labeled cases so that the algorithm can be used to predict new unlabeled cases [Berry et al. 2019]. Thus, the method training process uses this prior knowledge to train and generate the classification models. After training, the methods can classify new data. This approach's difficulty lies in need for labeled data for training the models [Russell and Norvig 2009].

Unsupervised learning refers to grouping data into unlabeled data using automated methods or algorithms. In this situation, algorithms need to understand the underlying relationships or features of the available data and group cases with similar features or characteristics [Berry et al. 2019].

In this section, several Machine Learning (ML) techniques used in intrusion detection in fog computing and IoT environments are presented and discussed. They are often employed in behavior-based detection strategies. The main focus of this short course will be methods based on supervised learning, as they are the most used in this context of intrusion detection. Next, the K-Nearest Neighbors (KNN) method is presented first.

1.3.1. K-Nearest Neighbor (KNN)

The k-Nearest Neighbors (kNN) algorithm is one of the most basic instance-based learning methods. It assumes that all examples correspond to points in an n-dimensional plane R^n, where n is the number of attributes used to represent them. Despite its simple oper-

ation, KNN generally has a very low error rate. This method uses a distance function to determine one instance's proximity to another [Mitchell 1997].

When numerical attributes describe the data set, distance measures are used to calculate the similarity so that the smallest distance corresponds to the greatest similarity. The Euclidean Distance [Mitchell 1997] stands out among the commonly applied measures.

The Euclidean distance is calculated as the square root of the sum of the squared differences between points of instance p in relation to points of instance q, as can be seen in Equation 1. Since p_i and q_i, for $i = 1, 2, ..., n$, are the attributes n that describe the instances p and q, respectively.

$$Euclidean\ Distance\ (p,q) = \sqrt{(p_1 - q_1)^2 + ... + (p_n - q_n)^2} = \sqrt{\sum_{1}^{n}(p_i - q_i)^2} \quad (1)$$

The K-Nearest Neighbor (KNN) algorithm identifies the closest k neighbors to the new data point and classifies it according to the nearest neighbors [Mitchell 1997]. If $k > 1$, the classes of the closest k examples are considered to carry out the classification. In this case, the most common approach is to assign the new instance to the majority class present in the set of the closest k examples.

Figure 1.3 presents a basic implementation of the KNN algorithm for data classification. As can be seen, the method has some parameters that influence the classification process. The main one is *n_neighbors*, corresponding to the number of neighbors considered for the classification process.

```
from sklearn.neighbors import KNeighborsClassifier

method = KNeighborsClassifier(n_neighbors=1, weights='uniform', algorithm='kd_tree')

method.fit(training_data_samples, training_data_labels)

predictions = method.predict(testing_data_samples)
```

Figure 1.3. Example application of the KNN algorithm in Python using the scikit-learn library.

Furthermore, the parameter *weights* indicates the influence of *n_neighbors* on the ranking. In the case of "uniform" weights, all neighbors are weighted equally. On the other hand, with "distance" weights, among the k neighbors, the closest ones will have more influence than the farther neighbors. The *algorithm* parameter indicates the technique used to store the training data, usually transformed into a fast indexing structure, such as Ball Tree ("ball_tree") or KD Tree ("kd_tree"). A KD Tree is a binary tree where each node is a k-dimensional point. The standard metric used in this library to calculate the similarities between the data is the Minkowski distance. It is a metric in a normalized vector space that can be considered a generalization of the Euclidean distance. As can be

seen in Figure 1.3, the algorithm is trained using the *fit()* method, where training data and labels are passed as parameters. The *predict()* method is used to carry out the classification, and data without labels is provided. The method will then generate class predictions for this data as a result.

The KNN algorithm is commonly used in intrusion detection approaches due to its low error rate [Illy et al. 2019]. An existing disadvantage in KNN is the computational cost, which can become high because it is necessary to compare the new instances with all the instances stored in the example base.

1.3.2. Artificial Neural Networks (ANN)

The brain has densely interconnected neurons forming a highly complex structure. Inspired by it, the Artificial Neural Networks (ANN) was proposed [Haykin 2001]. The artificial neuron is a logical mathematical structure that aims to simulate the biological neuron's shape, behavior, and functions. Inputs replace the dendrites, and the connections of these inputs with the artificial cell body are known as weights, which simulate synapses. The summation function processes the stimuli received by the inputs. The firing threshold of the biological neuron is simulated by the activation function in the artificial neuron [Chua and Yang 1988]. According to [Dalton and Deshmane 1991], synaptic weights play an important role in artificial neurons. The purpose of the weights is to weigh the influence of input signals on postsynaptic neurons. Positive weights tend to increase a neuron's activation level, consisting of an excitatory connection. Negative weights, on the other hand, tend to decrease the level of activation, which are called inhibitory connections.

Figure 1.4 shows a simplified model of an artificial neuron. Where the neuron k receives an input x $(x_1, x_2, ..., x_n)$ that enters through the synapse j. Each signal x_j from the input x that enters the synapse j is multiplied by a weight w_{kj}. The result of this process passes through an adder that adds the input signals weighted by the weights of the respective synapses with an external *bias* (b_k). Concerning a given synaptic weight w_{kj}, the first index (k) refers to the neuron in question, and the second index (j) refers to the input of the synapse to which the weight is related to [Haykin 2001]. The previously mentioned bias b_k has the role of increasing or decreasing the value generated by the adder before passing it on to an activation function, which will then transform the output into a closed interval, usually between $[0, 1]$ or $[-1, 1]$ to be passed on to other neurons [Haykin 2001].

The activation function, which in Figure 1.4 is represented by $\varphi(.)$ is extremely important in the neural model, as it defines what the neuron's output will be according to the received input [Haykin 2001]. Therefore, the activation function $\varphi(v)$ defines the neuron's output according to the result of the sum (v) of the weighted inputs. Several activation functions are used in neural networks, including Binary Step, Logistic Sigmoid, Hyperbolic Tangent, ReLU, and Softmax.

The Binary Step activation function defines the output of the summation result in v. In this model, if the result of the adder of the neuron in question is greater than or equal to 0, that is, it is positive, the output of the neuron assumes a value of 1, and if the result of the adder is negative, the output value of the neuron will be 0.

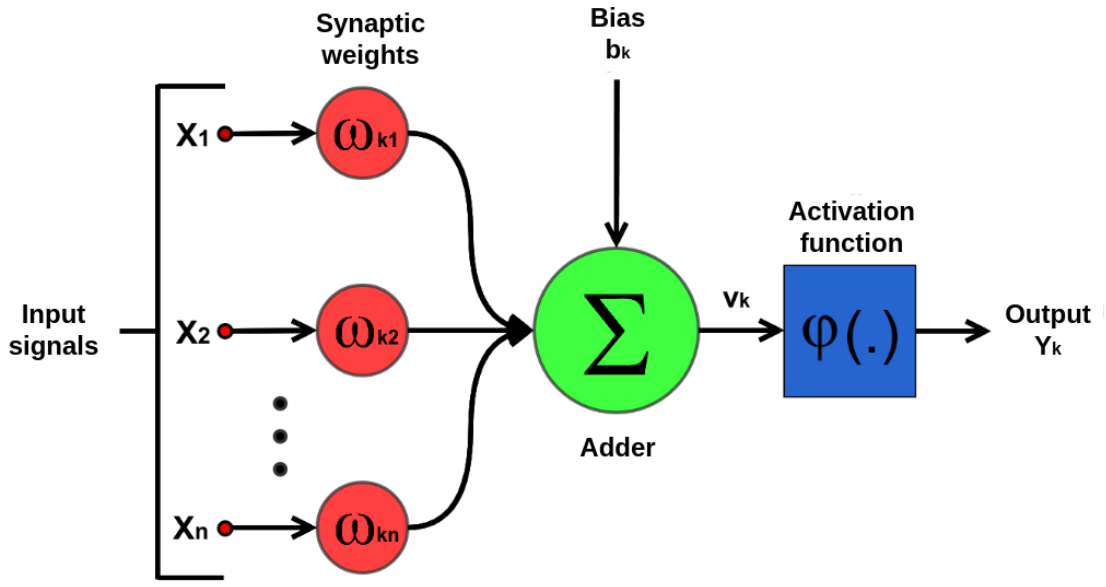

Figure 1.4. Simplified model of an Artificial Neuron. Adapted from [Haykin 2001].

The sigmoid activation function has an "S" shaped graph and assumes a continuous range of values between 0 and 1. It is one of ANN's most common activation functions and exhibits a good balance between linear and non-linear behavior. An example of a sigmoid function is the Logistic function, where a is the slope parameter of the sigmoid function. When this parameter tends to infinity, the function approaches the threshold function. While the threshold function assumes the value of 0 or 1, the logistic function assumes a continuous range of values between 0 and 1.

The hyperbolic tangent function is similar to the logistic sigmoid, but its continuous values range from -1 to 1. Allowing an activation function to assume negative values provides analytical benefits and advantages during the training phase [Haykin 2001].

Softmax generalizes the sigmoid function for non-binary cases. It is not usually applied to the hidden neural network layers but to the multiclass classification problems' output layer. The softmax function transforms the outputs for each class to values between 0 and 1 and divides them by the sum of the outputs. It is the probability that the input is in a given class.

Finally, ReLU is an abbreviation for Rectified Linear Unit (ReLU). It returns 0 for all negative values and the value itself for positive values. Thus, if the input is negative, the neuron will not be activated. This means that only a few neurons are activated simultaneously, making the network sparse and efficient. Therefore, it is a computationally light function widely used in hidden layers of neural networks [Goodfellow et al. 2016].

Artificial neural networks are composed of a large number of artificial neurons organized in the input layer, hidden layers, and output layer [Haykin 2001]. How neurons are arranged in an ANN is directly related to the learning algorithm. For this work, we will approach the class of networks Multilayer Perceptron (MLP) feedforward [Russell and Norvig 2010]. In this way, neurons in one layer are connected to neurons

in the next layer through weighted links. There are no recursive bindings present in this network topology. This model does not have feedback loops between neurons, so the flow of the synaptic process occurs from the input layer toward the output layer. The architecture displayed in Figure 1.5 illustrates a common basic structure of MLP neural networks. However, there may be several variations in the design of the hidden layers, both in terms of the number of layers and the number of neurons in each. These intermediate or hidden layers' design is crucial in defining a neural network, especially Deep Neural Networks (DNN). DNNs are deep networks, with more than one hidden layer [Goodfellow et al. 2016].

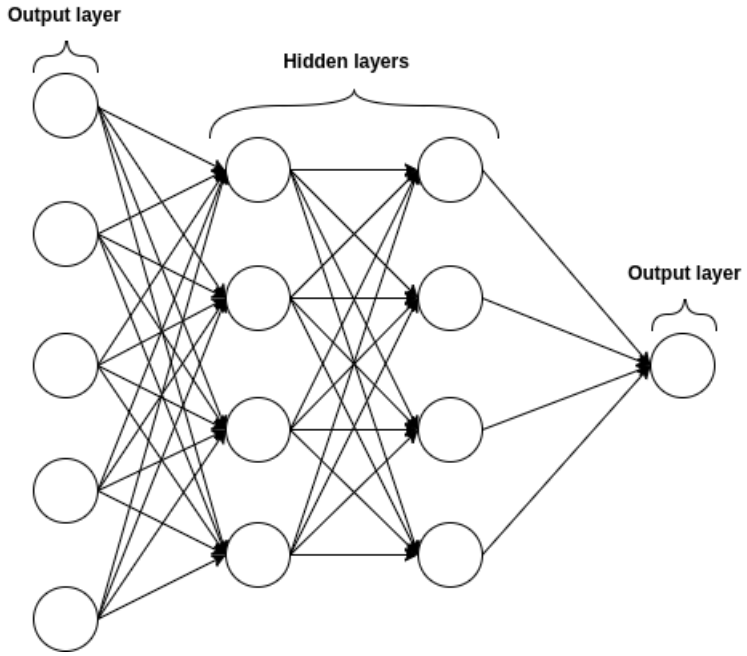

Figure 1.5. Simplified architecture of a deep feedforward neural network with two hidden layers.

An essential characteristic of an ANN is its ability to learn from its environment and improve its performance through training. An ANN learns more about its environment through an iterative process of adjusting its synaptic weights and bias levels, and this process is defined as training. The neural model is generated through the supervised training process, where the link weights are updated in several iterations based on the estimated error. ANN becomes more knowledgeable of its environment after each iteration of the training process [Haykin 2001]. ANN knowledge is represented through synaptic weights, forming a compact and distributed representation, thus providing generalization capabilities and adaptability to the neural network. ANN manage to achieve great ranking performances. However, they may suffer from instabilities caused by noise and variance in training. This instability means small changes to the training data used to build the model can result in very different models [Cunningham et al. 2000].

Training of MLP networks is usually performed using the backpropagation algorithm. The training takes place by propagating the data from the input layer to the output, passing through each of the hidden layers, and at that moment, the weights remain un-

changed. Afterward, based on the error calculated using the expected result (supervised learning) and the output value of the last layer, the weights are adjusted, and a new training iteration is performed. Training is considered completed when the error is small enough. From this, the network starts operating only in the forward direction for classifying new examples.

Figure 1.6 shows an implementation example of a basic DNN for classification, which can be used for intrusion detection. In this example, Keras and Tensorflow libraries are used. Keras *Sequential*() handles the ordering or sequencing of layers within a model. It makes the layers associated with neural networks work like a model that receives only one input as a feed and expects an output. The *add*() method adds layers to an already created layer stack. The *Dense*() layer is the regular deeply connected neural network layer and is the most common and frequently used. The *fit*() method is used to train the model, and the *predict*() method returns the values generated by the output neurons. The *np.argmax*() method obtains the neuron with the highest output value.

```
from keras.models import Sequential
from keras.layers.core import Dense, Dropout, Activation

method = Sequential()
method.add(Dense(10, input_dim=79, activation="relu"))
method.add(Dense(10, activation="relu"))
method.add(Dense(5, activation="softmax"))

print(method.summary())

method.compile(loss='sparse_categorical_crossentropy', optimizer='adam', metrics=['accuracy'])

method.fit(training_data_samples,training_data_labels, epochs=10, verbose=2)

predictions = np.argmax(method.predict(data_samples[test]),axis=1)
```

Figure 1.6. DNN implementation example using Keras and Tensorflow library.

1.3.3. Ensemble Learning

Ensemble Learning (EL) is the field of study of machine learning that works with Ensemble methods, which combine the decisions of various classification models to improve overall performance. The predictions from the various models are combined in some way to generate the final prediction [Breiman 1996].

Individual classifiers may experience instability. There is no guarantee that a classifier will always perform at its best in all situations. However, with ensemble learning, a better classification performance than any individual classifier can be achieved [Traganitis et al. 2018].

The main idea of ensemble learning is that combining several models decreases variation, especially in the case of unstable classifiers. In this way, it is possible to produce a more reliable classification than a single model [Breiman 1996].

1.3.3.1. Bagging

One of the main strategies for creating ensemble methods is Bagging (Bootstrap Aggregating). It generates multiple models of a classifier and uses them to obtain an aggregated classifier. Random samples of the training dataset are created for each model. Several subsets of the training dataset are created, and each model is trained with a subset. Finally, the results of these various models are combined using average or majority voting. Tests on real and simulated datasets using classification trees show that Bagging can provide substantial gains in accuracy [Breiman 1996].

Figure 1.7 presents an implementation example of a bagging method. The *n_estimator* parameter indicates the number of classifiers that will compose the method and the *estimator* indicates which is the base classifier.

```
from sklearn.ensemble import BaggingClassifier

method = BaggingClassifier(estimator=classifier,n_estimators=100)

method.fit(training_data_samples, training_data_labels)

predictions = method.predict(testing_data_samples)
```

Figure 1.7. Ensemble bagging application example in Python using the scikit-learn library.

Two specific ensemble methods for decision trees are presented below: the Random Forest and Extra Tree algorithms. These techniques create a diverse set of decision tree classifiers by introducing randomness in constructing the classifier.

1.3.3.2. Random Forest (RF)

Random Forest (RF) is an ensemble learning method based on Decision Tree (DT) created to reduce Overfitting. Decision Tree consists of a supervised machine learning algorithm based on the idea of recursively dividing a more complex problem into simpler problems. The input data are divided into homogeneous groups where each division performed represents a node of the tree where the data are separated according to a division criterion until reaching indivisible points. DTs have a tree-like structure composed of nodes. These nodes can be divided into a root node, a set of intermediate nodes, and a set of leaf nodes [Breiman et al. 1984]. The root node corresponds to the first division specifying how the data should be divided into separate parts. Successive intermediate nodes divide the data into smaller partitions until no further partitioning is needed. In this way, the leaf nodes of the structure represent the final partitions [Rokach 2016]. DTs classify using a hierarchical set of feature decisions. The decisions made in the internal nodes are the division criteria.

The basic decision tree induction algorithm builds decision trees recursively on a divide-and-conquer basis, starting from the top down. Each iteration seeks the attribute capable of best dividing the dataset. A good split in a decision tree corresponds to choos-

ing the attribute with the maximum separation power. In other words, the purpose of each node is to create child nodes dominated by a single class. The most suitable attribute is selected according to specific division criteria. Attributes are evaluated according to the division criterion, with the best attribute selected. The process is recursive, so each node further subdivides the training set into smaller subsets after selecting an appropriate split. For numeric attributes, there are many possible cut-off points. The induction algorithm looks for the best cut-off point by evaluating the split criterion at each possible cut-off point [Rokach 2016]. When the node satisfies the stopping rules, for example, because all instances of the current partition belong to the same class or no future split attributes can be determined, the DT terminates the splitting process, and the node is labeled [Rokach 2016]. One of the significant challenges of decision tree algorithms is finding the attribute that best divides the data into its corresponding classes. The main metrics used for this are the Gini Index and information gain based on entropy.

The Gini coefficient measures how well a given attribute separates the classes contained in a node. Possible values for the Gini index vary between 0 and 1, where 0 expresses the purity of the classification. All elements belong to a certain class, or only one class exists. Furthermore, 1 indicates a unequal distribution [Sundhari 2011]. The Gini index is determined by subtracting the sum of squared probabilities for each class from 1. It is expressed mathematically in Equation 2 [Breiman et al. 1984]. Where P_i denotes the probability that an element is classified into a class. When building the decision tree, resources with the lowest values of the Gini Index are chosen [Sundhari 2011].

$$GiniIndex = 1 - \sum_{i=1}^{C} (p_i)^2 \tag{2}$$

Another way to measure the quality of an attribute is to evaluate its degree of association with the class through the Information Gain measure. It evaluates the degree of association of attributes with the class to find the values with the highest degree of use and importance by calculating the entropy reduction. The greater the entropy, the greater the degree of impurity. The information gain indicates the entropy reduction. Thus, the attributes with the highest information gain will be the most useful for detection.

The measure of information gain is based on the concept of entropy. Entropy is a measure of the impurity and inhomogeneity of an attribute. The formula presented in Equation 3 corresponds to the entropy calculation for an attribute A, whose domain is $(a_1, a_2, ..., a_k)$, with $k >= 1$. The values p_i, with $1 <= i <= k$, correspond to the ratio between the number of instances of the base in which the value a_i occurs for the attribute A and the total number of instances.

$$Entropia(A) = \sum_{i=1}^{m} p_i \log_2(p_i) \tag{3}$$

Decision trees can have problems related to overfitting, which can degrade their predictive power when applied to new data [Breiman 2001]. In addition, they are considered models that can be unstable, where small variations in the training data can result in completely different trees. This can be avoided by training several different trees and

aggregating their predictions. Below are several methods based on the aggregation of ML models. This strategy is known as Ensemble Learning (EL) and seeks to generate methods with lower variance and more reliability.

RF builds a "forest" with a large number of uncorrelated decision trees and combines the results yielding the final classification results. Provides an additional layer of randomness over Bagging. In addition to building each tree using a different bootstrap sample of the data, RF changes how classification trees are built. Unlike standard DT, where each node is split using the best split among all features, in RF, each node is split using the best among a subset of features chosen randomly on that node [Breiman 2001]. It can get very good performance compared to many other classifiers like SVM and neural networks. Furthermore, it is robust against overfitting [Breiman 2001].

One of the main characteristics of RF is using a degree of randomness in selecting attributes to be considered for the division. Unlike DT, which applies impurity metrics across the entire set of attributes to find the best, RF applies these metrics only to a randomly selected subset of candidate attributes. Furthermore, it uses only a subset of the training data, with replacement, to construct each structure tree. The algorithm searches for the attributes that generate the best separability in each tree node. It randomly selects a set of candidate attributes and applies the measures to find the best cutoff point for each attribute and the best attribute among the candidates. This process ensures that each tree generates a different model. After RF training, the structure can perform the classification of new data. Each generated tree will classify the record, and its results will be combined through average or majority voting.

The RF method has several parameters, as shown in Figure 1.8, which presents a basic implementation of using RF for classification.

```
from sklearn.ensemble import RandomForestClassifier

method = RandomForestClassifier(n_estimators=100, max_features="sqrt",
                          criterion="gini", max_depth=100,
                          min_samples_split=2, min_samples_leaf=1)

method.fit(training_data_samples, training_data_labels)

predictions = method.predict(testing_data_samples)
```

Figure 1.8. Random Forest application example in Python using the scikit-learn library.

In RF, it is possible to define the parameters related to the internal structures of the decision trees that make up the RF structure. One of the parameters is the *criterion* that defines the metric used to choose the best attributes, where the metrics mentioned above can be used: Gini Index ("gini") and Entropy ("entropy"). The parameter *max_depth* indicates the maximum allowed depth of the tree. It is used to control this growth because normally, they can grow until all the leaves are pure or until all the leaves contain less than *min_samples_split* samples. However, this can lead to extremely long and costly trees. The parameter *min_samples_split* indicates the minimum size of the training set to split a node, and the (*min_samples_leaf*) indicates the minimum number of samples needed

to form a leaf node. A split point in any depth will only be considered if you leave at least *min_samples_leaf* samples in each of the left and right branches.

Furthermore, there are some additional parameters specific to the ensemble approach, one of which is *n_estimators*, which corresponds to the number of trees that will be created in the RF structure. Another important parameter is *max_features*, which indicates the number of randomly selected features in each node, where *max_features* is responsible for the intensity of the feature selection procedure and *n_estimators* the strength of variance reduction of the aggregation of the ensemble model [Geurts et al. 2006]. The RF is trained through the *fit()* method and performs the classification with the *predict()* method, as can be seen in Figure 1.8.

1.3.3.3. Extra Tree (ET)

Extra Tree (ET) [Geurts et al. 2006] classifiers are important tools in classification tasks. Like the RF, the Extra Tree consists of an ensemble method aggregating the results of several uncorrelated DTs accumulated in a "forest" to produce the classification results.

ET focuses on heavily randomizing the choice of attributes and the cutoff point while splitting a node in the tree. Therefore, a random sample of resources from the resource pool is selected at each intermediate node. Each decision tree must select the best feature to split the data based on some mathematical criteria, usually the Gini index. In the extreme case, it builds totally random trees whose structures are independent of the learning sample output values [Geurts et al. 2006]. The prediction trees are aggregated to produce the final prediction by majority vote in classification problems and arithmetic means in regression problems [Geurts et al. 2006]. It is very similar in operation to RF and varies mainly in the way of building the DTs inside the forest. In ET, randomness goes a step further in how divisions are calculated. Cutpoints are randomly drawn for each candidate attribute, and the best of these generated cutpoints is randomly chosen as the splitting rule.

```
from sklearn.ensemble import ExtraTreesClassifier

method = ExtraTreesClassifier(n_estimators=100, max_features="sqrt",
                          criterion="gini", max_depth=100,
                          min_samples_split=2, min_samples_leaf=1)

method.fit(training_data_samples, training_data_labels)

predictions = method.predict(testing_data_samples)
```

Figure 1.9. Example application of ExtraTree in Python using the scikit-learn library.

The rationale behind the method is that precise slicing and attribute randomization combined with ensemble mean should reduce variance more strongly than the weaker randomization schemes used by other methods. Using original and complete training data rather than bootstrap replicates is motivated to minimize bias [Verma and Ranga 2020]. In addition to precision, the main strength of the resulting algorithm is computational

efficiency, because, given the simplicity of the node split procedure, the constant factor can be much less than in other ensemble methods that locally optimize the cut points [Geurts et al. 2006]. The ET method also allows the definition of parameters of the internal decision trees that compose the structure. Furthermore, as it is a very similar technique to RF, it presents similar parameters, as seen in Figure 1.9.

1.3.3.4. Boosting

Boosting is another ensemble technique. It produces a highly accurate classifier by combining several "weak" models, each of which may not be good for the whole data set but is good for part of the data set so that the performance of these classifiers is improved [Schapire 1990]. Like bagging, boosting trains each model using a different training set. It is an iterative approach that adjusts the weight of an observation based on the last ranking. The performance of previously generated models influences each generated model. The boosting strategy is to focus on poorly classified examples. Each new model is created to classify well the examples poorly classified by previous models [Meir and Rätsch 2003]. Boosting generally decreases bias error and creates strong predictive models.

Decision trees are susceptible to overfitting, and to solve this problem, the Gradient Boosting Decision Tree (GBDT) can be used. It consists of a machine learning algorithm with effective implementations like XGBoost. Although many engineering optimizations were adopted in these implementations, the efficiency and scalability still needed improvement when the resource dimension was high, and the data size was large. One of the main reasons is that they have to scan all data instances for each feature to estimate the information gain of all possible split points, which is very time-consuming. To solve this problem, the authors [Ke et al. 2017] proposed LightGBM, which is based on Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) techniques. The two techniques form the characteristics of the LightGBM algorithm, and they integrate to make the template work efficiently and provide an advantage over other frameworks. These techniques work around the limitations of the histogram-based algorithm primarily used in all GBDT structures.

```
from lightgbm import LGBMClassifier

method = LGBMClassifier(n_estimators=100)

method.fit(training_data_samples, training_data_labels)

predicoes = method.predict(testing_data_samples)
```

Figure 1.10. Ensemble LightGBM application example in Python using the scikit-learn library.

The GOSS technique excludes a significant proportion of data instances with small gradients and uses only the remainder to estimate the information gain. According to the information gain definition, those instances with larger gradients will contribute more to the information gain. Therefore, when downsampling the data instances, to maintain the

accuracy of the information gain estimate, one should better keep those instances with large gradients, for example, greater than a predefined threshold or between the upper percentiles, and randomly eliminate instances with small gradients. GOSS can get a very accurate estimate of the information gain with a much smaller data size [Ke et al. 2017].

EFB groups mutually exclusive features, which rarely assume non-zero values simultaneously, to reduce the number of features. Finding the optimal grouping of unique features is NP-hard, but according to [Ke et al. 2017] authors, a greedy algorithm can achieve a good approximation ratio and effectively reduce the number of features without greatly impairing the accuracy of the split point determination. Figure 1.10 presents the implementation with the LightGBM technique in an example in Python.

1.3.3.5. Voting

Combining conceptually different machine learning classifiers and using a voting mechanism to generate the final classification is also possible. In this scheme, all classifiers are trained with the complete dataset, and their predictions are combined through voting.

The types of voting that can be used are hard and soft. In hard voting, each classification technique votes for a class, and the class that obtains the most votes is the final classification. The other strategy consists of a combiner based on the maximum sum of prediction probabilities. In a simplified way, each classification technique provides a probability value that the instance belongs to a given class. The predictions are then summed, and the class with the highest sum of probabilities is defined as the final classification.

Figure 1.11 presents an example of implementing an ensemble voting strategy. In the case presented, 'hard' voting is used, however, the voting parameter (*voting*) can also receive the value 'soft'.

```
from sklearn.ensemble import VotingClassifier

method = VotingClassifier(estimators= [('name1', classifier1), ('name2', classifier2)],
                          voting='soft', weights=None)

method.fit(training_data_samples, training_data_labels)

predicoes = method.predict(testing_data_samples)
```

Figure 1.11. Ensemble voting application example in Python using the scikit-learn library.

The base classifiers used to compose the voting-based method are passed as a parameter *estimators* in sequential form following the notation: name of the classifier and the object of the classifier. Although the example depicts only two base classifiers, the approach can work with more. Finally, the parameter *weights* indicates weights to weight the votes of the base classifiers. By default, it assumes the value 'None' where all base classifiers have a vote with the same weight.

1.3.3.6. Stacking

The Stacking technique consists of an ensemble strategy that combines several machine learning algorithms through a metamodel. The various base-level algorithms are trained on a complete training dataset, and the metamodel is trained on the final results of the base-level models. The predictions made by the basic models serve as a resource for the metamodel. In this way, the metamodel is responsible for learning to combine the individual results of each base classifier into an overall final result [Kumar et al. 2021a].

Figure 1.12 shows an example of implementing the ensemble stacking strategy. The base classifiers used to compose the stacking-based method are passed as a parameter *estimators* in sequential form following the notation: name of the classifier and the object of the classifier. The $stack_method$ parameter indicates the method called for each base classifier. In the case of the example, each of the base classifiers will perform class predictions through their *predict()* methods. The predictions generated by the base classifiers will be submitted to the final classifier, which is responsible for generating the final classification. This classifier is defined through the $final\_estimator$ parameter. The prediction generated by the $final\_estimator$ classifier is considered the final classification of the stacking method.

```
from sklearn.ensemble import StackingClassifier

method = StackingClassifier(estimators=[('name1', classifier1),('name2', classifier2)],
                            final_estimator=classifier, stack_method='predict')

method.fit(training_data_samples, training_data_labels)

predicoes = method.predict(testing_data_samples)
```

Figure 1.12. Ensemble stacking application example in Python using the scikit-learn library.

1.3.4. Practical simulation experiment with machine learning techniques for intrusion detection

Next, details are presented regarding the proposal for a practical simulation experiment with the IoTID20 dataset to evaluate the techniques presented in the section in an intrusion detection scenario with IoT traffic. First, a brief discussion of the existing datasets for intrusion detection simulation is presented.

1.3.4.1. Datasets

This section presents some datasets commonly used in research related to intrusion detection in IoT environments. The objective is to provide a survey of the validation strategies used in state of the art and bring an updated list that can serve as a basis for future researchers, providing indications that can help in deciding which datasets are most suitable for the context of their respective areas. In addition, the dataset chosen to be used in this mini-course is highlighted.

Table 1.1 presents the main databases used in intrusion detection problems. In addition, the main characteristics of the datasets are presented below. The **N.F.** column indicates the **N**umber of **F**eatures and the **L.** column indicates whether the dataset has labels for all records.

Table 1.1. Datasets for evaluating intrusion detection methods.

Dataset	Year	N.F.	L.	IoT	Comments
NSL-KDD [Tavallaee et al. 2009]	2009	42	✓	no	Reduces KDD Cup 99 redundancy problems, but is too old to represent current network standards environment
CTU-13 Botnet [García et al. 2014]	2013	33	✓	no	Focused only on Botnet attacks
RPL-NIDS17 [Verma and Ranga 2019]	2017	21	✓	✓	Differential focus on routing attacks, however, does not address other common IoT attacks
CICIDS-2017 [Sharafaldin et al. 2018]	2017	80	✓	no	It does not include some specific IoT features
CICIDS-2018 [Sharafaldin et al. 2018]	2018	80	✓	no	It does not include some specific IoT features
N-BaIoT [Meidan et al. 2018]	2018	115	no	✓	Only botnet attacks
DS2OS [Aubet 2018]	2018	13	✓	✓	Recent dataset focused on IoT
ToN-IoT [Alsaedi et al. 2020]	2019	7	✓	✓	Recent dataset focused on IoT
Bot-IoT [Koroniotis et al. 2019]	2018	46	✓	✓	Recent data set focused on IoT, but lacks some types of attacks
IoT-23 [Garcia et al. 2020]	2020	21	✓	✓	Real IoT environment traffic, but lacks some types of attacks
IoTID20 [Ullah and Mahmoud 2020a]	2020	12	✓	✓	Recent dataset focused on IoT
MQTT-IoT-IDS2020 [Hindy et al. 2021]	2020	44	✓	✓	Recent data set focusing on IoT, however, exclusively on the MQTT protocol
MQTTset [Vaccari et al. 2020]	2020	33	✓	✓	Recent data set focusing on IoT, however, exclusively on the MQTT protocol
NetFlow Datasets [Sarhan et al. 2021]	2021	43	✓	✓	Recent dataset, has some subsets in the context of IoT, feature standardization

The database used in this minicourse was IoTID20 [Ullah and Mahmoud 2020a]. This choice is mainly because the base has IoT traffic, is fully labeled, and is reasonably small compared to other datasets, which usually have millions of records. This will facilitate the execution of the experiments proposed in the minicourse.

It is one of the latest intrusion detection bases focused on IoT devices. Which was generated through a combination of IoT devices and interconnection structures, simulating a typical smart home environment having two IoT devices, namely, a smart speaker SKT NGU and an EZVIZ Wi-Fi camera [Ullah and Mahmoud 2020b].

These two IoT devices were connected to a home Wi-Fi router, which in turn interconnected with other devices connected to Smart Home, such as laptops, tablets, and smartphones. IoT SKT NGU and EZVIZ devices are victim devices and all other malicious devices. Figure 1.13 shows a simplified version of the architecture of the test environment.

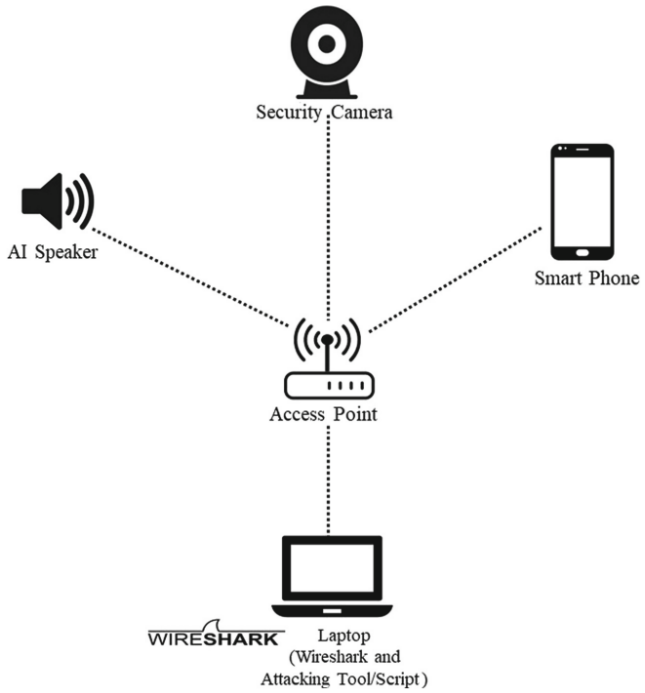

Figure 1.13. The Figure shows the architecture of the monitored environment for creating the IoTID20 dataset [Ullah and Mahmoud 2020b].

The IoTID20 dataset has 80 attributes of network characteristics and three attributes corresponding to labels. The first label is binary, the second corresponds to attack categories, and the last to attack subcategories. In Table 1.2, it is possible to observe the number of instances of traffic present in each type of label.

Table 1.2. Information about the number of instances of the IoTID20 dataset. Information is presented separated by normal traffic, anomalous traffic, their classes and subclasses.

Binary	Instances	Category	Instances	Subcategory	Instances
Benign	40073	Benign	40073	Benign	40073
Anomalous	585710	DoS	59391	DoS	59391
		Mirai	415677	*Ack Flooding*	55124
				Brute force	121181
				HTTP Flooding	55818
				UDP Flooding	183554
		MITM	35377	MITM	35377
		Scan	75265	*Host Port*	22192
				Port OS	53073

Figure 1.14 presents the implementation for reading the dataset for a DataFrame Pandas structure. DataFrame is a 2-dimensional labeled data structure with columns of potentially different types.

```
url_dataset = 'IoTID20/IoT Network Intrusion Dataset.csv'

dataset = pd.read_csv(url_dataset, header=0)
```

Figure 1.14. Example of reading the dataset into a Pandas DataFrame structure.

As part of the data pre-processing, the flow identifiers such as IDs, source IP, destination IP, and timestamps are dropped to avoid learning bias towards attacking and victim-end nodes. Also, the "Cat" label will be considered in this work. The traffic will be classified as benign or in some attack category: DoS, Mirai, MITM, or Scan. The other two labels ("Label" and "Sub_cat") will be taken from the dataset. Figure 1.15 presents an example of removing these columns from the DataFrame.

```
dataset.pop('Timestamp')
dataset.pop('Flow_ID')
dataset.pop('Src_IP')
dataset.pop('Dst_IP')
dataset.pop('Label')
dataset.pop('Sub_Cat')
```

Figure 1.15. Example of removing columns from the DataFrame.

Figure 1.16 presents the implementation for transforming categorical columns into numerical ones. In the case of the IoTID20 dataset, it will only be necessary to transform the labels column (Cat).

```
from sklearn import preprocessing

le = preprocessing.LabelEncoder()
le.fit(dataset['Cat'])
dataset['Cat']= le.transform(dataset['Cat'])
```

Figure 1.16. Example of reading the dataset into a Pandas DataFrame structure.

Next, sanitizing the dataset and removing records with infinite values (inf) and invalid or missing values (NaN) is necessary. For this, the $pd.option\_context$ command will be used to temporarily define options in the context within a block of code. And the option $'mode.use\_inf\_as\_na'$ is used to consider all infinite values (inf) as (NaN). Within this block, the $df.dropna(inplace = True)$ method is used to remove records with NaN from the dataset, as seen in Figure 1.17.

```
with pd.option_context('mode.use_inf_as_na', True):
  dataset = dataset.dropna()
```

Figure 1.17. Example in Python for removing records with infinite and invalid or missing values.

Another important task in dataset pre-processing is standardization. It helps to improve the performance of some classifiers based on machine learning that needs their resources to be normally distributed. They can misbehave if the individual features don't more or less resemble standard normally distributed data. The method used to standardize the data was standard scaling, given by the equation 4, where x is the sample, u is the mean, and s is the standard deviation. The mean and standard deviation are obtained based on the statistic for each attribute in the data set.

$$z = \frac{x - u}{s} \tag{4}$$

Figure 1.18 presents a basic implementation to standardize the data set using the Scikit-learn library. The *fit()* method computes the mean and standard deviation to be used for later scaling and the *transform()* performs standardization by centering and scaling.

```
from sklearn.preprocessing import StandardScaler

labels = dataset.pop('Cat')

scaler = StandardScaler()
scaler.fit(dataset.values)
dataset = scaler.transform(dataset.values)

dataset =  pd.DataFrame(dataset)
dataset['Cat'] = labels
```

Figure 1.18. Implementation example for dataset standardization.

Finally, it is necessary to separate the column of labels (*Cat*) from the dataset into a separate structure called *data_y*. The other columns correspond to traffic attributes and are assigned to *data_x*. This can be done according to Figure 1.19.

```
data_y = dataset.pop('Cat').values
data_x = dataset.values
```

Figure 1.19. Implementation example for label separation.

Once the preparation of the dataset for the experiment is complete, some information on the metrics commonly used to evaluate the ML method in intrusion detection experiments is presented below.

1.3.4.2. Metrics

The evaluation of detection methods is essential to examine the feasibility of applying them in a real environment. One of the evaluations that can be performed is the extraction of detection metrics through experiments, where the methods are trained and tested. The classifications of database events performed by the methods in these experiments can be categorized in the terms presented below.

- **False Negative (FN):** Events classified as normal by the detection method and which are intrusions;
- **False Positive (FP):** Non-intrusive events classified as intrusive by the intrusion detection technique;
- **True Negative (TN):** This category covers non-intrusive events, which were correctly classified by the detection method;
- **True Positive (VP):** This class includes events correctly reported as intruders by the intrusion detection technique.

From these terms, it is possible to construct confusion matrices. They consist of tables that present a summary of the classification performed by the method, indicating the number of events classified in relation to the predicted class and the true class of the element. The confusion matrix itself is not a metric for evaluating classification methods. However, several analyzes and metrics can be made from such information. In Table 1.3, it is possible to observe an example of a confusion matrix.

Original Label	**Prediction**	
	Normal (+)	Ataque (-)
Normal (+)	VN	FP
Ataque (-)	FN	VP

Table 1.3. Example of a confusion matrix.

Based on the abovementioned terms, it is possible to calculate different metrics that help evaluate detection methods built by machine learning algorithms. Below are the main metrics used [Liu and Lang 2019].

1. **Accuracy:** This metric corresponds to the proportion of correctly classified instances in relation to the total number of existing instances. It may not be a good aspect to consider in cases of large class imbalance. The accuracy is calculated from the Equation 5, presented below:

$$ACC = \frac{VP + VN}{VP + VN + FP + FN} \tag{5}$$

2. **Precision:** Another widely used metric for evaluating machine learning methods. This rate indicates the proportion of instances correctly detected as intrusive out of all those detected as intrusive, as can be seen in Equation 6.

$$PRE = \frac{VP}{VP + FP} \tag{6}$$

3. **Recall:** Also known as sensitivity, consists of the number of instances correctly classified as intrusive among all intrusive instances, calculated according to Equation 7.

$$Recall = \frac{VP}{VP + FN} \tag{7}$$

4. **F1-Score:** The F1 score is the harmonic mean of precision and recall, where an F1 score is best at 1 (perfect precision and recall) and worst at 0. The formula for the F1 score is given in Equation 8.

$$F1 - Score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{8}$$

5. **Balanced Accuracy (BACC):** Balanced Accuracy is an interesting metric to evaluate detection performance on unbalanced datasets. It is defined as the average of the recall obtained in each class, as seen in Equation 9. Where R_i corresponds to the recall (R) obtained considering the i-th (i) class present in the dataset, and n is the number of existing classes.

$$Balanced\ Accuracy = \frac{\sum_{i=1}^{n} R_i}{n} \tag{9}$$

Figure 1.20 shows the basic Python implementation for calculating detection metrics using the methods provided by the Scikit-learn library.

```
import sklearn.metrics

acc = sklearn.metrics.accuracy_score(testing_data_labels, predictions)
print('Accuracy: {}'.format(acc))

acc_balanced = sklearn.metrics.balanced_accuracy_score(testing_data_labels, predictions)
print('Accuracy balanced: {}'.format(acc_balanced))

precision = sklearn.metrics.precision_score(testing_data_labels, predictions, average=None)
print('Precision: {}'.format(precision))

reacll = sklearn.metrics.recall_score(testing_data_labels, predictions, average=None)
print('Recall: {}'.format(recall))

f1_score = sklearn.metrics.f1_score(testing_data_labels, predictions, average=None)
print('F1_score: {}'.format(f1_score))
```

Figure 1.20. Example of implementation of detection metrics using the Scikit-learn library.

1.3.4.3. Experiments

Intrusion detection databases are large datasets that can be used to validate and evaluate detection models. These data can be used to train and test the models. However, ideally, models should be evaluated with samples not used to build, train, or tune the model. In order to obtain an unbiased assessment of the model's effectiveness.

Thus, it is proposed in this work to carry out experiments with the Hold-Out 70-30 technique to divide the data set and carry out the evaluation. In this strategy, the data set (*data_x* and *data_y*) is divided into 70% of the data for training (*training_x* and *training_y*) the models and 30% of the data for testing ($test_x$ and $test_y$) the performance of the generated models. This division is performed to prevent data used to train the model from being used to test it. Furthermore, a stratified strategy divides these data so that the training and test sets maintain equal proportions of instances per class.

Figure 1.21 presents the basic implementation in Python for experimenting with the Hold-Out 70-30 technique and the pre-processed IoTID20 dataset. In this experiment, the data is split, and the generated training data is used to train the detection method.Before the training, it is necessary to define which classification method will be used in the experiment. For this, it is necessary to assign the classifier to the *method* variable. This can be done according to the definitions presented in each of the previous sections. After training, the test data, without the labels, is submitted to the method for classification. From this, the generated classifications can be used to calculate the detection metrics of the method. The example shows the calculation of the accuracy metric, however the other metrics presented can also be used in the experiment.

```
from sklearn.model_selection import train_test_split

training_X,test_x,training_y,test_y = train_test_split(data_X,data_y,test_size=0.3)

#Definition of the classification method to be used

method.fit(training_X, training_y)

predictions = method.predict(test_x)

acc_balanced = sklearn.metrics.balanced_accuracy_score(test_y, predictions)
print('Accuracy balanced: {}'.format(acc_balanced))

#Other metrics can be used below
```

Figure 1.21. Example of experiment implementation using the hold-out 70-30 technique.

The objective is to use the points mentioned in this practical section to implement, execute, evaluate, and compare each of the presented machine learning methods.

1.4. State of the Art in intrusion detection in the context of IoT/Fog/Cloud

In state-of-the-art, several works proposed intrusion detection approaches for IoT, Fog, and Cloud environments. Some approaches are based on analysis by signature and others by behavior. In addition, some works have inserted the detection module in the layer of IoT devices and others in the upper layers. Some of these approaches are presented and discussed below.

In signature-based detection, monitored actions are compared to predefined intrusive events. Signature-based solutions enable rapid detection and reduce the occurrence of false alarms. Thus, they are interesting options to be deployed on [Arshad et al. 2019] IoT devices. However, they have the limitation of not being able to detect attacks that do not have a signature known to the IDS.

Most state-of-the-art works have proposed behavior-based detection approaches. Some works consider a branch of analysis by a behavior called analysis by the specification. This type of solution employs rules and thresholds that define the expected default behavior for monitored components. It is similar to anomaly detection, and both detect intrusions when network behavior deviates from specified. The main difference is that in specification-based analysis, a human expert sets the rules [Mitchell and Chen 2014]. [Yaseen et al. 2017] proposed a threshold-based approach for detecting selective forwarding attacks on Wireless Sensor Networks (WSNs). In [Aliyu et al. 2018], a challenge or response based fog detection approach is proposed. The detection nodes periodically interrogate nearby nodes, sending interrogation packets and waiting for a response according to a previously specified calculation. The major disadvantages of this type of analysis are the necessary specificity and the need for a human expert to define the system's expected behavior.

Anomaly detection approaches, on the other hand, are usually modeled using Machine Learning (ML) techniques. There are threats against IoT devices and even against services deployed in the fog and cloud layer that need robust methods to be detected. These complex detection approaches often cannot be applied to IoT devices due to their computational constraints [Arshad et al. 2019]. In addition, approaches that only perform analysis on IoT nodes only consider a restricted view of events, thus limiting their ability to deal with complex, multi-stage, and distributed attacks. In this way, research on detection techniques implemented in devices in the fog layer has been growing, seeking to avoid the problem of latency, implement a distributed strategy, and take advantage of the privileged position of these devices. Many works have proposed solutions deployed in the fog to perform network monitoring of IoT devices through network packet analysis. This location allows the IDS to have a global view of the IoT network, monitoring all traffic [Souza et al. 2020, Lawal et al. 2021, Labiod et al. 2022].

The previously presented KNN method has been used in some works due to its good classification performance. In [Lawal et al. 2021], the proposed mitigation framework for fog computing uses a database that stores signatures of previously detected attacks and an anomaly-based detection scheme that uses the K-Nearest Neighbor (KNN) classification algorithm to detect DDoS attacks. The Euclidean, Manhattan, and Chebychev distances were evaluated, and the first two achieved the best results. An existing disadvantage in KNN is the computational cost, which can become high because it is

necessary to compare the new instances with those stored in the example base. The authors [Souza et al. 2020] proposed a hybrid approach called DNNKNN, composed of a DNN model and the K-Nearest Neighbor algorithm, to improve the method's robustness and reduce the prediction cost of the kNN algorithm. This approach made it possible to maintain a good detection performance and obtain a reduction of approximately 90% of the KNN prediction cost.

Another technique commonly used in intrusion detection research is the Deep Neural Network (DNN). In state-of-the-art, it is possible to find several works that propose detection approaches based on DNN. Most architectures consider two hidden layers. Inserting more layers did not bring detection benefits [Lalouani and Younis 2021]. In addition, architectures with a greater number of layers are more complex and demand more costs for the training process. [Sahar et al. 2021] used hidden ReLu layers with 768 and 512 neurons, respectively. However, a large number of neurons per layer may also be unnecessary. Several other works sought to build less complex neural models with fewer neurons. [Kumar et al. 2021b] used two hidden layers with 32 and 16 neurons, respectively, with an activation function ReLU. In [Lalouani and Younis 2021], the function used in the two hidden layers was the ReLU in the 64 neurons in each layer. In [Kumar and Tripathi 2021], a layer with 16 neurons and another with 12 neurons were used, with ReLU activation function. ReLU is more efficient for training large-scale data in terms of time and cost [Sahar et al. 2021]. On the other hand, the Soft-Max activation function is one of the most used in the output layer [Souza et al. 2020, Lalouani and Younis 2021, Kumar and Tripathi 2021, Kumar et al. 2021b]. Several other techniques based on neural networks can be used for intrusion detection and have state-of-the-art works. However, this work focuses only on feed-forward models.

One of the major challenges regarding neural model-based detection approaches is the high cost of training complex deep learning architectures during model updates. Motivated by this challenge, some works used the distributed characteristic of fog computing to propose distributed training approaches between fog devices through information sharing [Diro and Chilamkurti 2018, Labiod et al. 2022].

The authors [Diro and Chilamkurti 2018] proposed a new distributed approach based on DNN feed-forward to detect intrusions in the IoT environment. The approach is deployed distributed in fog devices and has two levels. It uses a fog device as the master responsible for training and placing the model on the other fog nodes. Second-level nodes send model updates to the master node. The master node updates the global model and spreads the updates to other nodes. The experiment demonstrated that the distributed approach could detect cyber attacks better than centralized algorithms due to the sharing of parameters that can avoid local minima in training [Diro and Chilamkurti 2018]. However, this primary node can be considered a single point of failure (SPOF), which is easier to compromise than a cloud-based parameter update approach.

In this context of distributed training of detection models, it is important to discuss Federated Learning (FL), which consists of a strategy to form a global ML model from several local data-driven models [Lalouani and Younis 2021]. This strategy decentralizes machine learning, eliminating the need to gather data on a single device. Instead, the model is trained through multiple iterations on different devices. FL is an ideal framework

for aggregating distributed models, preserving privacy and allowing convergence to a distributed learning engine with precision close to that of a centralized implementation. The FLIDS [Lalouani and Younis 2021] approach employs federated learning to enable privacy-preserving distributed aggregation in training deep neural models deployed in fog computing.

The framework architecture proposed by [Rey et al. 2022] is composed of clients that monitor IoT devices and a server that coordinates a Federated Learning process. The approach provides for intrusion detection through neural network models. Considering that IoT devices generally have limited resources and modest reliability, the clients responsible for training the models are not the devices to be protected, but other entities capable of collecting traffic from IoT devices present on the same network, such as fog nodes [Rey et al. 2022].

[Abbasi et al. 2021] presents extensive research on DL methods to detect anomalies in network traffic. They point out that federated learning approaches are promising to overcome the challenges of training a DL approach with many samples and training parameters in environments with resource constraints, such as IoT and fog environments. However, they are also susceptible to threats, model poisoning attacks can be carried out through updates of corrupted models sent to the server. Furthermore, due to the privacy issues employed in FL, it is difficult to verify whether the received models really correspond to the local training data or not [Lalouani and Younis 2021]. Furthermore, using a device as a server centralizes the process of aggregating models on a device. This centralization can bring some concerns, such as the need to trust a central device and the possibility of this central device becoming a single point of failure, where the failure or compromise of this device by an attacker, could harm the collaboration network completely.

In addition to deep neural methods, research on ensemble approaches is also promising for intrusion detection. Classifiers based on Ensemble Learning can be proposed to improve adaptability and generalizability in multiclass classification. Several recent works have investigated the use of the ensemble Random Forest method in intrusion detection approaches [Illy et al. 2019, Farukee et al. 2020, Kumar et al. 2022]. The authors [Farukee et al. 2020], however, used RF to select the main characteristics of the traffic to be submitted to another classifier. An important aspect of RF is the calculation of the importance of the resource. The Gini impurity criterion index is used. Thus, they used the RF property to classify resources according to their importance. Prioritizing accuracy, features with feature importance less than 0.005 were discarded. The RF used in the ensemble method proposed by Kumar et al. [Kumar et al. 2021a] had 100 estimators, maximum depth equal to 3, a minimum number of examples for split equal to 10, a minimum number of samples needed to be a leaf node equal to 6 and Entropy criteria. [Hosseini and Sardo 2022] proposed an approach with RF classification and feature selection by Spider-Monkey Optimization (SMO). The authors [Kumar et al. 2020b] evaluated RF in the context of DDOS attack detection and found that the fact that RF is insensitive to outliers, missing values, overfitting, and having the ability to handle a large number of incoming traffic makes it suitable in the process anomaly detection tool for the blockchain-IoT environment. However, no details are provided regarding the number of trees used in the structure. The performance of the proposed distributed structure is evaluated using a BoT-IoT dataset. The proposed distributed structure with RF and 10

base attributes surpasses some current state-of-the-art techniques, reaching recall close to 100% in all classes. However, it is observed that several techniques have already presented similar performance with this Bot-IoT dataset.

The Extra Tree, a method very similar to RF, was also used in detection approaches. In ET, randomness goes a step further in how divisions are calculated. Instead of looking for the most discriminating cut points, they are drawn randomly in ET, making the ET training process faster. In [Albdour et al. 2020] is proposed an ET-based intrusion detection approach for the fog layer. The approach has 10 DTs and uses the Gini criterion. The approach showed 98.3% accuracy with the UNSW-NB15 dataset. However, considering only the binary detection, not being able to identify the attack categories. In [Souza et al. 2022a] ET was used in a first level of binary detection, obtaining a high detection rate. The ET used is composed of 10 estimators DT, the minimum sample size for division is 2, the number of attributes considered for better division is the root of the number of existing attributes, and the Gini Index is used as a criterion.

Following the line of ensemble classifiers with Decision Tree, [Lawal et al. 2020] presented an approach with Extreme Gradient Boosting (XGBoost) for anomaly detection in an IoT framework. XGBoost is based on the ensemble boosting technique, where weak classifier predictions are combined to develop a strong classifier, employing additive techniques. In addition to the speed and performance benefits of XGBoost, additional advantages include the avoidance of overfitting and the full utilization of [Lawal et al. 2020] computational resources. In the experiments performed by [Lawal et al. 2020] with the Bot-IoT dataset, XGBoost was able to achieve 99.96% average accuracy, 98% average recovery, 97% average accuracy, and 97% average f1-score in the multiclass ranking. Superior performance compared to other classifier algorithms such as DT, kNN, and Naive Bayes.

Several proposed papers have also focused on ensemble voting approaches. The authors [Illy et al. 2019] proposed a voting-based ensemble approach composed of KNN, RF, DT bagging, and DT boosting. [Alhowaide et al. 2021] report two main strategies to combine the results of the base classifiers in ensemble methods by voting: hard and soft voting. They highlight that soft voting can perform better than hard voting because it takes into account more information and uses the uncertainty of each classifier in the final decision.

The ensemble technique proposed in [Al-Khafajiy et al. 2021] is composed of three base classifiers which together provide analysis of collected traffic for intrusion detection. The decisions generated by the three classifiers are combined using the majority voting rule (hard), where the class with the highest number of votes of the three classification systems is defined as the final classification. The authors point out that the majority voting rule is the simplest and most effective voting scheme in this case.

In [Souza et al. 2022a], a binary detection is applied, and only the events detected as intrusive are submitted for multiclass analysis. The proposed multiclass analysis model is a soft voting set comprising three classification models: ET, RF, and DNN. The combination strategy employed was soft voting, that is, the complete method implements a combinator that predicts the class label based on the argmax of the sums of the probabilities predicted by each of the three classification models.

In addition, the proposal of methods based on ensemble stacking strategy also stands out in state of the art. [Kumar et al. 2020a] proposed an approach where the KNN, XGBoost, and Naive Bayes classifiers are trained in parallel and act as basic classifiers. As a result, three prediction results $P1$, $P2$, and $P3$ are obtained, which an RF uses for the final classification. Thus, the RF is responsible for learning to combine the individual results of each weak classifier into a final overall result. The experiments with the data set showed that the approach performed well but presented difficulties in identifying some attacks.

The authors Kumar et al. [Kumar et al. 2021a] also present another similar approach; in this case, they proposed an ensemble method with Naive Bayes, DT, RF, and XGBoost to detect attacks on Internet of Medical Things (IoMT) networks. Naive Bayes, DT, and RF classifiers operate in parallel on the first level. As a result, three prediction outputs $P1$, $P2$, and $P3$ are obtained and are used by XGBoost to build the final predictive model.

Ensemble techniques can be useful in intrusion detection, allowing you to build a strong classifier to identify a specific attack's specific class. However, as these techniques use several classifiers, they may present difficulties related to the processing and training time of the models.

Some of the approaches mentioned above result in classifying events as normal or malicious, making it impossible to identify the type of attack. The methods that perform only the identification that an intrusion has occurred, that is binary detection, are not enough to provide efficient security. The mechanism must be able to mitigate the invasion so that it does not succeed. Therefore, it is important to classify the attack in its category so that specific countermeasures are executed for the given type of threat. In addition, the classification of the type or category of the attack is important for the decision-making of the person responsible for the network.

It is essential to identify more information about the attack so that specific countermeasures can be carried out for each type of threat. For example, a probing attack is usually performed before more powerful attacks such as DoS, DDoS, remote access attacks, etc [Nguyen et al. 2019]. Thus, running additional detection mechanisms to reinforce security when detecting a probing category threat may be interesting. Also, classifying attack types or category is important for the network manager. From identifying the category of a certain attack that occurs with a specific frequency, the person responsible for the network can decide to implement actions to correct the vulnerability used by the attack.

To improve the accuracy of multiclass detection without overloading the IoT-Fog environment, the approach proposed in [Souza et al. 2020, Souza et al. 2022a] presents a two-step hierarchical detection method. A binary detection analysis (Step 1 - Detection) is performed on fog computing devices to detect intrusive events. Only events detected as intrusive by the first step are sent to the multiclass analysis (Step 2 - Identification) in the cloud. The analysis module of Step 2 is responsible for identifying the attack category and providing further information to the countermeasures module. If the event is detected as non-intrusive by Step 1, it is automatically sent to the module output to free up the flow. The analysis performed at the cloud computing layer aims to classify the event

into a specific attack category or normal behavior. This step allows you to correct first-level false positives. The classifier consists of a more robust method that requires more processing than the first stage's analysis. This method will be activated only when the first level detects the event as intrusive. In this way, it is possible to apply a complex analysis that can more accurately classify the event into a specific class of attack for the execution of countermeasures.

For the binary detection module of the first stage, the authors proposed a hybrid DNNKNN approach previously mentioned [Souza et al. 2020]. A soft voting ensemble analysis was proposed for the second stage, consisting of three robust classification models: ExtraTree, Random Forest, and Deep Neural Network [Souza et al. 2022a]. The architecture also applies attribute selection and class balancing techniques. The obtained results provided superior detection performance than several state-of-the-art approaches.

This hierarchical detection architecture proposed in [Souza et al. 2020] has already served as a basis for other works. In [Labiod et al. 2022], some modifications are proposed in the approaches throughout the architecture, mainly concerning the first level of detection. The VAE-MLP method, a binary detection approach based on Variational AutoEncoder and DNN, was proposed.

As presented, several state-of-the-art works were found. However, there are still many challenges in this context. Next section discusses some important aspects observed in state-of-the-art related to intrusion detection in an IoT/Fog/Cloud context. The objective is to instigate an initial reflection on this research topic's problems, challenges, and open questions.

1.5. Discussions, Reflections and Questions

This section discusses some important aspects observed in state-of-the-art related to intrusion detection in an IoT/Fog/Cloud context. The aim is to instigate an initial reflection on this research topic's problems, challenges, and open questions.

1.5.1. Deployment strategy

The deployment location of the detection solution is an important aspect that must be considered when designing the network or host-based approach. Resource constraints, usually existing in devices inserted in the context of IoT applications, make it difficult to implement robust detection approaches in the devices themselves. However, some works have proposed approaches partially implemented in IoT devices through lighter signatures-based techniques. Other approaches proposed their solutions to operate entirely in the fog computing layer. In addition, some works have also delegated part of the analysis to the cloud computing layer.

The processing time and cost of robust machine learning models can be high, especially when considering strategies that use multiple detection methods. Implementing a robust approach in IoT devices and even in the fog computing layer can be a problem. Robust and slow multiclass analysis performed in the fog can overwhelm the device and slow network flow. It is necessary to investigate new intrusion detection approaches capable of managing and optimizing the analysis process in the various layers of the IoT environment. In other words, optimizing available resources on IoT devices, fog, and cloud.

1.5.2. Detection method category

The categories of detection methods can be anomaly, signature, and specification. Some detection approaches found in the state of the art focus on signature-based detection [Arshad et al. 2019, Lawal et al. 2021]. They are not able to detect new attacks or variations of known attacks.

On the other hand, some works have proposed approaches with analysis by specification, which detects intrusions when the network behavior deviates from the specified [Yaseen et al. 2017, Aliyu et al. 2018]. In this type of detection, a human expert specifies normal behavior. Thus, the approaches based on specifications found are closely linked to specific protocols or attacks. The weakness of these approaches is the difficulty of generalizing the approaches to a broader context with other attacks and protocols.

Finally, anomaly detection considers that all abnormal behavior is an intrusion and can detect new attacks. Several approaches have proposed anomaly-based methods to detect intrusions into the fog computing layer. Usually, they are machine learning-based approaches. IoT devices often have limited computational resources [Ni et al. 2018]. These restrictions make it difficult to conduct analyses based on complex anomaly techniques on IoT devices, thus preventing new attacks [Zarpelão et al. 2017]. Furthermore, anomaly-based approaches may suffer from problems related to false positives. Also, there is only a narrow view of events. Thus, research proposing hybrid detection approaches, combining detection categories, are interesting and necessary to maximize the advantages and minimize the disadvantages of these types of analysis and obtain a complete solution.

1.5.3. Machine Learning approachs

It is essential to categorize the attack to take specific countermeasures for the threat. For example, a probe category attack is usually performed before more powerful attacks such as DDoS, remote access attacks, etc. [Nguyen et al. 2019]. Thus, running additional detection mechanisms to strengthen security when detecting an attack from the probe attack category is interesting. Also, the type of attack or category classification is important to the network owner. From identifying the category of a certain attack that occurs with a specific frequency, the person responsible for the network can decide to implement actions to correct the attacker's vulnerability.

Many approaches found in the state-of-the-art focus on performing binary detection (attack or non-attack). However, binary methods cannot identify the type or category of attack [Albdour et al. 2020]. The approaches, which aim to classify the attack into specific categories, are multiclass. However, it is observed in the state-of-the-art that these approaches have lower accuracy rates than the binary detection methods [Nguyen et al. 2019, Kumar et al. 2020b]. In addition, these approaches may present difficulties related to false-positive problems and low detection of some types of attacks [Diro and Chilamkurti 2018, Kumar et al. 2020b, Kumar and Tripathi 2021].

Several works sought to propose ML-based solutions with more restricted models to respect the resource capacity of the devices involved in this context. The approaches obtained interesting results related to detection; however, single classifiers are subject to inconsistencies and need to be improved in detecting some types of attacks [Diro and Chilamkurti 2018, Kumar and Tripathi 2021].

Individual classifiers may experience instability. There is no guarantee that a classifier will always perform at its best in all situations. However, with Ensemble Learning (EL), a better classification performance than any individual classifier can be achieved [Traganitis et al. 2018]. Classifiers based on DL and ensemble approaches have been the object of research recently and have achieved promising results in intrusion detection. Ensemble methods can be proposed to improve adaptability and generalizability in multiclass classification [Traganitis et al. 2018]. Thus, combining different machine learning models for optimal performance and attack detection is another research trend. However, ensemble methods have greater computational complexity and require more training time and resources. Thus, it is important to consider the characteristics of the devices where the approach will be implemented in the design of detection methods.

In addition, using more complex techniques such as DL and Ensemble also makes efforts to improve training strategies, optimize resources and reduce cost and computational time. Collaborative IDS based on federated learning in fog computing emerges as an interesting alternative to optimize the training of complex DL and ensemble learning approaches.

Another interesting point is that the performance of the methods is related to the quality and quantity of the training data. This can be challenging, as obtaining training data can be extremely arduous. In this context, hybrid approaches that combine supervised machine learning techniques with other techniques that work with unlabeled data, such as unsupervised or reinforcement learning, are promising.

1.5.4. Collaborative IDSs approachs

Collaborative detection approaches are very promising solutions in the IoT-Fog-Cloud context, mainly due to the distributed nature of fog computing. However, these solutions have limitations, such as vulnerability to insider attacks [Li et al. 2021]. Insider attacks occur when attackers compromise a device that is part of the Collaborative Intrusion Detection System (CIDS) and, from there, perform false collaborative actions to undermine the functioning of the collaborative detection approach. As a solution, one can investigate trust management approaches to defend against insider attacks.

Another highlight is that although a server is responsible for aggregating the models, it would be possible to pass the steps from the server to the devices themselves, decentralizing the server into several entities [Rey et al. 2022]. This decentralization would avoid the need to have a trusted device as a central server and hence the single point of failure problem. However, a reliable decentralized approach would require a Blockchain-based architecture to be used as a decentralized database, where each device would share its local model and reliably retrieve models from other devices.

1.5.5. Detection models update

Another research point is training detection models based on ML e DL. This intrusion detection subtopic still has several points that need to be further studied and improved. Machine learning-based detection approaches must be retrained over time to prevent them from becoming obsolete. In the IoT context, the network changes over time. New devices can be inserted, and others removed. Therefore, detection models based on machine learning techniques need to be updated. Considering the new components, new data must be collected from the network to generate an updated model capable of identifying truly abnormal behaviors in the new IoT network. Furthermore, IoT applications can be inserted in a context of high device mobility, which undoubtedly makes the intrusion detection process even more challenging. Thus, another open question is understanding the maximum time a model can operate without becoming obsolete in this IoT context [Abbasi et al. 2021]. The ideal strategy would be to retrain the discovery models whenever there is any change in the network, such as the insertion and removal of devices. However, the training process can cost a lot of resources and overload the network. Machine learning models often suffer from high complexity in the training phase as they consume many resources and time. Devices in the IoT context have resource constraints, so the complexity of the detection models to be retrained can be considered a major challenge. Some works proposed a distributed training mechanism between fog devices with the exchange of parameters of the machine learning method [Diro and Chilamkurti 2018].

Approaches based on Federated Learning are also very interesting [Rey et al. 2022, Lalouani and Younis 2021]. However, they are also susceptible to threats. Model poisoning attacks can be carried out through corrupted model updates sent to the server. Trust management approaches can be investigated to defend against these insider attacks. Furthermore, due to the privacy concerns employed in FL, it isn't easy to verify whether the received models match the local training data or not [Lalouani and Younis 2021]. However, more studies are needed to propose detection approaches that are less susceptible to changes in the network and lighter to optimize detection models in terms of time com-

plexity and resource consumption so as not to overload fog computing.

Another important point is that machine learning approaches' good detection performance depends on the training data's quality and quantity. The vast majority of works found proposed approaches based on supervised learning. To train these approaches, having a large number of labeled data is usually necessary. This becomes an issue as you need to capture network traffic from the network and label it for use in the detection model training process. Therefore, this task can become extremely costly due to the large amount of data required. Some works have proposed promising approaches for the training data labeling problem, using data sampling and clustering methods [Ravi and Shalinie 2020]. As points for future studies, the need for research on new hybrid approaches based on supervised and unsupervised learning to overcome the problem of data labeling stands out. Therefore, proposing approaches that combine supervised machine learning techniques with other techniques capable of working with unlabeled data, such as unsupervised learning or reinforcement learning, can be a great solution.

Updating detection approaches is a major challenge, which generates the need for further studies to find the ideal strategy. They must consider how often it takes to update the model, training costs, data acquisition, and labeling.

1.6. Conclusion

IoT is spreading in all areas due to its ability to make objects smart. In this way, they can monitor and act in the environment in which they operate. IoT devices have limited resources and must send information to places with more computing resources. Fog computing then emerged as an excellent processing solution close to devices. IoT and fog are not free from security threats and vulnerabilities. Added to the significant damage generated by attacks in this environment, this fact generates the need to concentrate efforts in this area. Intrusion detection systems are an essential tool to ensure IoT security.

In this mini-course, the fundamental concepts involved in the theme of this work were presented, the main threats present in IoT environments were discussed, and concepts related to Intrusion Detection Systems were introduced. In addition, several machine learning techniques that can be used to analyze and detect intrusions were presented. This section also proposes to conduct simulation experiments with the IoTID20 dataset to evaluate the machine learning techniques presented in an intrusion detection scenario with IoT traffic. In Section 1.4, state-of-the-art surveying through a literature review is exposed, and the approaches proposed by the main related works are presented. Finally, Section 1.5 discusses some important aspects observed in state-of-the-art related to intrusion detection in an IoT/Fog/Cloud context. The objective is to instigate an initial reflection on this research topic's problems, challenges, and open questions.

For future research, the following points are highlighted: (1) investigate hybrid approaches combining detection categories; (2) investigate solutions capable of managing the analysis process in the various layers of the IoT-Fog-Cloud environment; (3) investigate new approaches for multiclass detection to achieve accuracy greater than or similar to binary detection; (4) investigate ensemble and hybrid methods to improve multiclass detection considering resource constraints; (5) investigate strategies to update/retrain detection models with federated learning.

References

[Abbasi et al. 2021] Abbasi, M., Shahraki, A., and Taherkordi, A. (2021). Deep learning for network traffic monitoring and analysis (ntma): A survey. *Computer Communications*, 170:19–41.

[Ahmad et al. 2021] Ahmad, Z., Shahid Khan, A., Wai Shiang, C., Abdullah, J., and Ahmad, F. (2021). Network intrusion detection system: A systematic study of machine learning and deep learning approaches. *Transactions on Emerging Telecommunications Technologies*, 32(1):e4150.

[Al-Fuqaha et al. 2015] Al-Fuqaha, A., Guizani, M., Mohammadi, M., Aledhari, M., and Ayyash, M. (2015). Internet of things: A survey on enabling technologies, protocols, and applications. *IEEE Communications Surveys & Tutorials*, 17(4):2347–2376.

[Al-Khafajiy et al. 2021] Al-Khafajiy, M., Otoum, S., Baker, T., Asim, M., Maamar, Z., Aloqaily, M., Taylor, M., and Randles, M. (2021). Intelligent control and security of fog resources in healthcare systems via a cognitive fog model. *ACM Trans. Internet Technol.*, 21(3).

[Albdour et al. 2020] Albdour, L., Manaseer, S., and Sharieh, A. (2020). Iot crawler with behavior analyzer at fog layer for detecting malicious nodes. *Int. J. Commun. Networks Inf. Secur.*, 12(1).

[Alhowaide et al. 2021] Alhowaide, A., Alsmadi, I., and Tang, J. (2021). Ensemble detection model for iot ids. *Internet of Things*, 16:100435.

[Aliyu et al. 2018] Aliyu, F., Sheltami, T., and Shakshuki, E. M. (2018). A detection and prevention technique for man in the middle attack in fog computing. *Procedia Computer Science*, 141:24 – 31. The 9th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN-2018) / The 8th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare (ICTH-2018) / Affiliated Workshops.

[Alsaedi et al. 2020] Alsaedi, A., Moustafa, N., Tari, Z., Mahmood, A., and Anwar, A. (2020). Ton_iot telemetry dataset: A new generation dataset of iot and iiot for data-driven intrusion detection systems. *IEEE Access*, 8:165130–165150.

[Anderson 1980] Anderson, J. P. (1980). Computer security threat monitoring and surveillance. *Technical Report, James P. Anderson Company*.

[Arshad et al. 2019] Arshad, J., Azad, M. A., Abdellatif, M. M., Rehman, M. H. U., and Salah, K. (2019). Colide: a collaborative intrusion detection framework for internet of things. *IET Networks*, 8(1):3–14.

[Atzori et al. 2010] Atzori, L., Iera, A., and Morabito, G. (2010). The internet of things: A survey. *Computer networks*, 54(15):2787–2805.

[Aubet 2018] Aubet, F.-X. (2018). *Machine Learning-Based Adaptive Anomaly Detection in Smart Spaces*. PhD thesis.

[Aversano et al. 2021] Aversano, L., Bernardi, M. L., Cimitile, M., and Pecori, R. (2021). A systematic review on deep learning approaches for iot security. *Computer Science Review*, 40:100389.

[Bace and Mell 2001] Bace, R. and Mell, P. (2001). Nist special publication on intrusion detection systems. Technical report, BOOZ-ALLEN AND HAMILTON INC MCLEAN VA.

[Berry et al. 2019] Berry, M. W., Mohamed, A., and Yap, B. W. (2019). *Supervised and unsupervised learning for data science*. Springer.

[Birkinshaw et al. 2019] Birkinshaw, C., Rouka, E., and Vassilakis, V. G. (2019). Implementing an intrusion detection and prevention system using software-defined networking: Defending against port-scanning and denial-of-service attacks. *Journal of Network and Computer Applications*, 136:71 – 85.

[Bonomi et al. 2012] Bonomi, F., Milito, R., Zhu, J., and Addepalli, S. (2012). Fog computing and its role in the internet of things. In *Proceedings of the first edition of the MCC workshop on Mobile cloud computing*, pages 13–16. ACM.

[Boukerche et al. 2007] Boukerche, A., Machado, R. B., Jucá, K. R., Sobral, J. B. M., and Notare, M. S. (2007). An agent based and biological inspired real-time intrusion detection and security model for computer network operations. *Computer Communications*, 30(13):2649–2660.

[Breiman 1996] Breiman, L. (1996). Bagging predictors. *Machine learning*, 24(2):123–140.

[Breiman 2001] Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

[Breiman et al. 1984] Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

[Camhi 2015] Camhi, J. (2015). Former cisco ceo john chambers predicts 500 billion connected devices by 2025. *Business Insider*.

[Campello and Weber 2001] Campello, R. S. and Weber, R. F. (2001). Sistemas de detecçao de intrusao. *Minicurso procedente do 19º Simpósio Brasileiro de Redes de Computadores*.

[Chua and Yang 1988] Chua, L. O. and Yang, L. (1988). Cellular neural networks: Applications. *IEEE Transactions on circuits and systems*, 35(10):1273–1290.

[Cunningham et al. 2000] Cunningham, P., Carney, J., and Jacob, S. (2000). Stability problems with artificial neural networks and the ensemble solution. *Artificial Intelligence in Medicine*, 20(3):217 – 225.

[Dalton and Deshmane 1991] Dalton, J. and Deshmane, A. (1991). Artificial neural networks. *IEEE Potentials*, 10(2):33–36.

[Diro and Chilamkurti 2018] Diro, A. A. and Chilamkurti, N. (2018). Distributed attack detection scheme using deep learning approach for internet of things. *Future Generation Computer Systems*, 82:761 – 768.

[Farukee et al. 2020] Farukee, M. B., Shabit, M. Z., Haque, M. R., and Sattar, A. S. (2020). Ddos attack detection in iot networks using deep learning models combined with random forest as feature selector. In *International Conference on Advances in Cyber Security*, pages 118–134. Springer.

[García et al. 2014] García, S., Grill, M., Stiborek, J., and Zunino, A. (2014). An empirical comparison of botnet detection methods. *Computers & Security*, 45:100 – 123.

[Garcia et al. 2020] Garcia, S., Parmisano, A., and Erquiaga, M. J. (2020). Iot 23: A labeled dataset with malicious and benign iot network traffic.

[Garcia-Morchon et al. 2013] Garcia-Morchon, O., Kumar, S., Keoh, S., Hummen, R., and Struik, R. (2013). Security considerations in the ip-based internet of things draft-garciacore-security-06. *Internet Engineering Task Force*.

[Geurts et al. 2006] Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Machine learning*, 63(1):3–42.

[Goodfellow et al. 2016] Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press.

[Haykin 2001] Haykin, S. S. (2001). *Redes neurais: Princípios e Práticas*. Bookman.

[Heady et al. 1990] Heady, R., Luger, G., Maccabe, A., and Servilla, M. (August 1990). The architecture of a network level intrusion detection system. Technical report, University of New Mexico, Department of Computer Science.

[Hindy et al. 2021] Hindy, H., Bayne, E., Bures, M., Atkinson, R., Tachtatzis, C., and Bellekens, X. (2021). Machine learning based iot intrusion detection system: An mqtt case study (mqtt-iot-ids2020 dataset). In Ghita, B. and Shiaeles, S., editors, *Selected Papers from the 12th International Networking Conference*, pages 73–84, Cham. Springer International Publishing.

[Hosseini and Sardo 2022] Hosseini, S. and Sardo, S. R. (2022). Network intrusion detection based on deep learning method in internet of thing. *Journal of Reliable Intelligent Environments*, pages 1–13.

[Illy et al. 2019] Illy, P., Kaddoum, G., Moreira, C. M., Kaur, K., and Garg, S. (2019). Securing fog-to-things environment using intrusion detection system based on ensemble learning. In *2019 IEEE Wireless Communications and Networking Conference (WCNC)*, pages 1–7.

[Iorga et al. 2018] Iorga, M., Feldman, L., Barton, R., Martin, M. J., Goren, N. S., and Mahmoudi, C. (2018). Fog computing conceptual model. Technical report.

[Ke et al. 2017] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 3149–3157, Red Hook, NY, USA. Curran Associates Inc.

[Kolias et al. 2016] Kolias, C., Stavrou, A., Voas, J., Bojanova, I., and Kuhn, R. (2016). Learning internet-of-things security" hands-on". *IEEE Security & Privacy*, 14(1):37–46.

[Koroniotis et al. 2019] Koroniotis, N., Moustafa, N., Sitnikova, E., and Turnbull, B. (2019). Towards the development of realistic botnet dataset in the internet of things for network forensic analytics: Bot-iot dataset. *Future Generation Computer Systems*, 100:779–796.

[Kumar et al. 2020a] Kumar, P., Gupta, G. P., and Tripathi, R. (2020a). A distributed ensemble design based intrusion detection system using fog computing to protect the internet of things networks. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–18.

[Kumar et al. 2021a] Kumar, P., Gupta, G. P., and Tripathi, R. (2021a). An ensemble learning and fog-cloud architecture-driven cyber-attack detection framework for iomt networks. *Computer Communications*, 166:110–124.

[Kumar et al. 2020b] Kumar, P., Kumar, R., Gupta, G. P., and Tripathi, R. (2020b). A distributed framework for detecting ddos attacks in smart contract-based blockchain-iot systems by leveraging fog computing. *Transactions on Emerging Telecommunications Technologies*, n/a(n/a):e4112.

[Kumar et al. 2021b] Kumar, P., Tripathi, R., and P. Gupta, G. (2021b). P2idf: A privacy-preserving based intrusion detection framework for software defined internet of things-fog (sdiot-fog). In *Adjunct Proceedings of the 2021 International Conference on Distributed Computing and Networking*, ICDCN '21, page 37–42, New York, NY, USA. Association for Computing Machinery.

[Kumar et al. 2022] Kumar, R., Kumar, P., Tripathi, R., Gupta, G. P., Garg, S., and Hassan, M. M. (2022). A distributed intrusion detection system to detect ddos attacks in blockchain-enabled iot network. *Journal of Parallel and Distributed Computing*, 164:55–68.

[Kumar and Tripathi 2021] Kumar, R. and Tripathi, R. (2021). Dbtp2sf: a deep blockchain-based trustworthy privacy-preserving secured framework in industrial internet of things systems. *Transactions on Emerging Telecommunications Technologies*, 32(4):e4222.

[Labiod et al. 2022] Labiod, Y., Amara Korba, A., and Ghoualmi, N. (2022). Fog computing-based intrusion detection architecture to protect iot networks. *Wireless Personal Communications*, 125(1):231–259.

[Lalouani and Younis 2021] Lalouani, W. and Younis, M. (2021). Robust distributed intrusion detection system for edge of things. In *2021 IEEE Global Communications Conference (GLOBECOM)*, pages 01–06.

[Lawal et al. 2020] Lawal, M. A., Shaikh, R. A., and Hassan, S. R. (2020). An anomaly mitigation framework for iot using fog computing. *Electronics*, 9(10).

[Lawal et al. 2021] Lawal, M. A., Shaikh, R. A., and Hassan, S. R. (2021). A ddos attack mitigation framework for iot networks using fog computing. *Procedia Computer Science*, 182:13–20. Learning and Technology Conference 2020; Beyond 5G: Paving the way for 6G.

[Li et al. 2021] Li, W., Au, M. H., and Wang, Y. (2021). A fog-based collaborative intrusion detection framework for smart grid. *International Journal of Network Management*, 31(2):e2107.

[Liu and Lang 2019] Liu, H. and Lang, B. (2019). Machine learning and deep learning methods for intrusion detection systems: A survey. *Applied Sciences*, 9(20).

[Marín-Tordera et al. 2017] Marín-Tordera, E., Masip-Bruin, X., García-Almiñana, J., Jukan, A., Ren, G.-J., and Zhu, J. (2017). Do we all really know what a fog node is? current trends towards an open definition. *Computer Communications*, 109:117–130.

[Meidan et al. 2018] Meidan, Y., Bohadana, M., Mathov, Y., Mirsky, Y., Shabtai, A., Breitenbacher, D., and Elovici, Y. (2018). N-baiot—network-based detection of iot botnet attacks using deep autoencoders. *IEEE Pervasive Computing*, 17(3):12–22.

[Meir and Rätsch 2003] Meir, R. and Rätsch, G. (2003). An introduction to boosting and leveraging. In *Advanced lectures on machine learning*, pages 118–183. Springer.

[Mell et al. 2011] Mell, P., Grance, T., et al. (2011). The nist definition of cloud computing.

[Miorandi et al. 2012] Miorandi, D., Sicari, S., De Pellegrini, F., and Chlamtac, I. (2012). Internet of things: Vision, applications and research challenges. *Ad hoc networks*, 10(7):1497–1516.

[Mitchell and Chen 2014] Mitchell, R. and Chen, I.-R. (2014). A survey of intrusion detection techniques for cyber-physical systems. *ACM Comput. Surv.*, 46(4).

[Mitchell 1997] Mitchell, T. M. (1997). Machine learning. *McGraw Hill Science/Engineering/Math*, page 432.

[Muhammad et al. 2015] Muhammad, F., Anjum, W., and Mazhar, K. S. (2015). A critical analysis on the security concerns of internet of things (iot). *International Journal of Computer Applications*, 111(7).

[Mukherjee et al. 1994] Mukherjee, B., Heberlein, L. T., and Levitt, K. N. (1994). Network intrusion detection. *IEEE network*, 8(3):26–41.

[Navas et al. 2018] Navas, R. E., Le Bouder, H., Cuppens, N., Cuppens, F., and Papadopoulos, G. Z. (2018). Do not trust your neighbors! a small iot platform illustrating a man-in-the-middle attack. In *International Conference on Ad-Hoc Networks and Wireless*, pages 120–125. Springer.

[Neshenko et al. 2019] Neshenko, N., Bou-Harb, E., Crichigno, J., Kaddoum, G., and Ghani, N. (2019). Demystifying iot security: an exhaustive survey on iot vulnerabilities and a first empirical look on internet-scale iot exploitations. *IEEE Communications Surveys & Tutorials*, 21(3):2702–2733.

[Nguyen et al. 2019] Nguyen, T. G., Phan, T. V., Nguyen, B. T., So-In, C., Baig, Z. A., and Sanguanpong, S. (2019). Search: A collaborative and intelligent nids architecture for sdn-based cloud iot networks. *IEEE Access*, 7:107678–107694.

[Ni et al. 2018] Ni, J., Zhang, K., Lin, X., and Shen, X. (2018). Securing fog computing for internet of things applications: Challenges and solutions. *IEEE Communications Surveys & Tutorials*.

[Northcutt et al. 2001] Northcutt, S., Cooper, M., Fearnow, M., and Frederick, K. (2001). *Intrusion Signatures and Analysis*. New Riders, New Jersey.

[Patel et al. 2010] Patel, A., Qassim, Q., and Wills, C. (2010). A survey of intrusion detection and prevention systems. *Information Management & Computer Security*.

[Ravi and Shalinie 2020] Ravi, N. and Shalinie, S. M. (2020). Semi-supervised learning based security to detect and mitigate intrusions in iot network. *IEEE Internet of Things Journal*, pages 1–1.

[Rey et al. 2022] Rey, V., Sánchez Sánchez, P. M., Huertas Celdrán, A., and Bovet, G. (2022). Federated learning for malware detection in iot devices. *Computer Networks*, 204:108693.

[Rokach 2016] Rokach, L. (2016). Decision forest: Twenty years of research. *Information Fusion*, 27:111 – 125.

[Roman et al. 2018] Roman, R., Lopez, J., and Mambo, M. (2018). Mobile edge computing, fog et al.: A survey and analysis of security threats and challenges. *Future Generation Computer Systems*, 78:680–698.

[Russell and Norvig 2009] Russell, S. and Norvig, P. (2009). *Artificial Intelligence: A Modern Approach*. Prentice Hall Press, USA, 3rd edition.

[Russell and Norvig 2010] Russell, S. J. and Norvig, P. (2010). *Artificial intelligence: A Modern Approach*. Pearson Education, Inc.

[Sahar et al. 2021] Sahar, N., Mishra, R., and Kalam, S. (2021). Deep learning approach-based network intrusion detection system for fog-assisted iot. In *Proceedings of international conference on big data, machine learning and their applications*, pages 39–50. Springer.

[Sarhan et al. 2021] Sarhan, M., Layeghy, S., Moustafa, N., and Portmann, M. (2021). Towards a standard feature set of nids datasets. *arXiv preprint arXiv:2101.11315*.

[Satyanarayanan 2015] Satyanarayanan, M. (2015). A brief history of cloud offload: A personal journey from odyssey through cyber foraging to cloudlets. *GetMobile: Mobile Computing and Communications*, 18(4):19–23.

[Schapire 1990] Schapire, R. E. (1990). The strength of weak learnability. *Machine learning*, 5(2):197–227.

[Sharafaldin et al. 2018] Sharafaldin, I., Lashkari, A. H., and Ghorbani, A. A. (2018). Toward generating a new intrusion detection dataset and intrusion traffic characterization. In *Proceedings of the 4th International Conference on Information Systems Security and Privacy - Volume 1: ICISSP*, pages 108–116.

[Sicari et al. 2015] Sicari, S., Rizzardi, A., Grieco, L., and Coen-Porisini, A. (2015). Security, privacy and trust in internet of things: The road ahead. *Computer Networks*, 76:146 – 164.

[Souza et al. 2022a] Souza, C. A., Westphall, C. B., and Machado, R. B. (2022a). Two-step ensemble approach for intrusion detection and identification in iot and fog computing environments. *Computers & Electrical Engineering*, 98:107694.

[Souza et al. 2022b] Souza, C. A., Westphall, C. B., Machado, R. B., Loffi, L., Westphall, C. M., and Geronimo, G. A. (2022b). Intrusion detection and prevention in fog based iot environments: A systematic literature review. *Computer Networks*, page 109154.

[Souza et al. 2020] Souza, C. A., Westphall, C. B., Machado, R. B., Sobral, J. B. M., and dos Santos Vieira, G. (2020). Hybrid approach to intrusion detection in fog-based iot environments. *Computer Networks*, 180:107417.

[Sundhari 2011] Sundhari, S. S. (2011). A knowledge discovery using decision tree by gini coefficient. In *2011 International Conference on Business, Engineering and Industrial Applications*, pages 232–235.

[Takabi et al. 2010] Takabi, H., Joshi, J. B. D., and Ahn, G. (2010). Security and privacy challenges in cloud computing environments. *IEEE Security Privacy*, 8(6):24–31.

[Tanaka and Yamaguchi 2017] Tanaka, H. and Yamaguchi, S. (2017). On modeling and simulation of the behavior of iot malwares mirai and hajime. In *2017 IEEE International Symposium on Consumer Electronics (ISCE)*, pages 56–60.

[Tavallaee et al. 2009] Tavallaee, M., Bagheri, E., Lu, W., and Ghorbani, A. A. (2009). A detailed analysis of the kdd cup 99 data set. In *Computational Intelligence for Security and Defense Applications, 2009. CISDA 2009. IEEE Symposium on*, pages 1–6. IEEE.

[Traganitis et al. 2018] Traganitis, P. A., Pagès-Zamora, A., and Giannakis, G. B. (2018). Blind multiclass ensemble classification. *IEEE Transactions on Signal Processing*, 66(18):4737–4752.

[Ullah and Mahmoud 2020a] Ullah, I. and Mahmoud, Q. H. (2020a). A scheme for generating a dataset for anomalous activity detection in iot networks. In *Canadian Conference on Artificial Intelligence*, pages 508–520. Springer.

[Ullah and Mahmoud 2020b] Ullah, I. and Mahmoud, Q. H. (2020b). A scheme for generating a dataset for anomalous activity detection in IoT networks. In *Advances in Artificial Intelligence*, pages 508–520. Springer International Publishing.

[Vaccari et al. 2020] Vaccari, I., Chiola, G., Aiello, M., Mongelli, M., and Cambiaso, E. (2020). Mqttset, a new dataset for machine learning techniques on mqtt. *Sensors*, 20(22).

[Verma and Ranga 2019] Verma, A. and Ranga, V. (2019). Evaluation of network intrusion detection systems for rpl based 6lowpan networks in iot. *Wireless Personal Communications*, 108(3):1571–1594.

[Verma and Ranga 2020] Verma, A. and Ranga, V. (2020). Machine learning based intrusion detection systems for iot applications. *Wireless Personal Communications*, 111(4):2287–2310.

[Yan et al. 2016] Yan, Q., Yu, F. R., Gong, Q., and Li, J. (2016). Software-defined networking (sdn) and distributed denial of service (ddos) attacks in cloud computing environments: A survey, some research issues, and challenges. *IEEE Communications Surveys Tutorials*, 18(1):602–622.

[Yaseen et al. 2017] Yaseen, Q., Albalas, F., Jararwah, Y., and Al-Ayyoub, M. (2017). Leveraging fog computing and software defined systems for selective forwarding attacks detection in mobile wireless sensor networks. *Transactions on Emerging Telecommunications Technologies*, 29(4):e3183.

[Zarpelão et al. 2017] Zarpelão, B. B., Miani, R. S., Kawakani, C. T., and [de Alvarenga], S. C. (2017). A survey of intrusion detection in internet of things. *Journal of Network and Computer Applications*, 84:25 – 37.

Capítulo 4

Aplicações Críticas Habilitadas pela Tecnologia 5G: Oportunidades, Tendências e Desafios

Francisco Carvalho Neto (UFF), Alessandro Aparecido Milan (UFF), Natalia Castro Fernandes (UFF) e Alberto G. Guimarães (UFF)

Abstract

This chapter presents the key critical applications that envision 5G as an enabler, discussing the essential concepts, key requirements, challenges, research trends, and technologies that make 5G meet the needs of these applications. Critical applications stand out because they are crucial for the safety and development of productive sectors and have strict performance requirements that, if not met, can cause social or environmental catastrophes. The chapter seeks to present the main critical applications made possible by 5G, discussing the enabling technologies described in releases 15, 16, 17, and 18 of 5G, in addition to discussing the challenges related to each technology and the research carried out internationally related to the applications and 5G. Furthermore, it discusses the main research trends and challenges associated with emerging technologies for 6G network development and deployment.

Resumo

Este capítulo apresenta as principais aplicações críticas que vislumbram a rede 5G como habilitadora, discutindo os conceitos essenciais, principais requisitos, desafios, tendencias de pesquisa e tecnologias que fazem o 5G atender às necessidades dessas aplicações. As aplicações críticas se destacam por serem cruciais para a segurança e desenvolvimento de setores produtivos e possuem requisitos de desempenho estritos que, caso não sejam atendidos, podem provocar catástrofes sociais ou ambientais. O capítulo busca apresentar as principais aplicações críticas viabilizadas pelo 5G, discutindo as tecnologias habilitadoras descritas nos releases 15, 16, 17 e 18 do Projeto de Parceria de Terceira Geração - Third Generation Partnership Project (3GPP), além de discutir os desafios relacionados a cada tecnologia e às pesquisas realizadas internacionalmente relacionadas às aplicações críticas e 5G. Ademais, discute as principais tendências e desafios de pesquisa associados a tecnologias em ascensão para o desenvolvimento e implantação da rede 6G.

Este capítulo foi realizado com recursos do CNPq, CAPES e FAPERJ.

4.1. Introdução

As redes móveis de comunicação vêm passando por uma transformação digital resultante de demandas impostas por novas aplicações cada vez mais complexas. Essa evolução melhora o desempenho e garante segurança e qualidade de serviço, possibilitando que novas aplicações sejam desenvolvidas buscando automatizar funções majoritariamente dependentes de ação humana, como dirigir ou realizar uma cirurgia. Historicamente, as principais evoluções nas redes móveis se deram na Rede de Acesso por Rádio - *Radio Access Network* (RAN), responsável pela conexão entre o dispositivo do usuário e a rede interna da operadora de telefonia, ou Núcleo da Rede - *Core Network* (CN). Os avanços tecnológicos que viabilizaram essa evolução permitiram uma melhoria na eficiência da utilização do espectro e o desenvolvimento de novas técnicas de acesso múltiplo ao meio. Além disso, as faixas de frequência e largura de banda foram ampliadas, possibilitando uma maior taxa de transferência de dados. Entretanto, com a implementação de novos serviços atendidos pelas aplicações móveis, novos requisitos de Qualidade de Serviço - *Quality of Service* (QoS) - passaram a ser exigidos. Estes novos requisitos estão relacionados à latência da rede, à densidade de dispositivos a serem atendidos por km² e à taxa de transferência de dados ofertada. Logo, para suportar o atendimento de tais aplicações é necessário uma evolução não só na RAN, mas uma mudança no tratamento das solicitações no CN, para que a rede possa prover o suporte aos serviços sem violar os diferentes requisitos de QoS. Além disso, uma infraestrutura que atenda a diversas aplicações deve ser flexível e de fácil gerenciamento, visto que os recursos físicos são limitados e devem ser utilizados de maneira eficiente.

As prominentes redes 5G, por exemplo, apresentam modificações expressivas na RAN e CN. Dentre as principais evoluções da RAN pode-se destacar uma melhoria na utilização do espectro, ou seja, o aumento da eficiência espectral com o aperfeiçoamento da forma de onda transmitida, o aperfeiçoamento de técnicas de acesso múltiplo ao meio, e uma densificação das Estações Rádio Base, visando aumentar a vazão de transmissão e o número de dispositivos por km² atendidos por uma mesma antena [Shaik e Malik, 2021]. Já no Núcleo da Rede 5G – 5G CN – , há uma mudança completa na arquitetura, que passa a ser orientada a serviços e a utilizar com mais eficiência os recursos computacionais e de comunicação disponíveis, promovendo maior flexibilidade no atendimento de requisições. O objetivo dessa mudança é prover melhor atendimento às necessidades específicas de cada aplicação e utilizar de forma eficiente e compartilhada os recursos disponíveis. Isso só é possível com o uso de equipamentos genéricos, que não executem funções específicas na rede. Deste modo, as redes 5G disponibilizam recursos computacionais e de comunicação da rede através das Redes Definidas por *Software* - *Software Defined Network* (SDN) aliadas à Virtualização de Funções de Rede - *Network Function Virtualization* (NFV).

O uso das SDNs e da NFV permite que os recursos disponíveis, que antes eram dedicados a serviços específicos, sejam abstraídos e se comportem como recursos genéricos utilizados por qualquer aplicação. Estas funções são digitalizadas e executadas sob *softwares* (por exemplo, máquinas virtuais, *docker*), utilizando com mais eficiência e dinamicidade os recursos da rede [Khorsandroo et al., 2021, Ahvar et al., 2021]. Além disso, o uso das SDNs permite a separação do Plano de Controle e do Plano de Dados, que facilita o gerenciamento e configuração destas redes. As redes 5G ainda implementam a

entrega de conteúdos e de recursos de processamento mais próximos do usuário, através da Computação de Borda de Acesso Múltiplo - *Multi-Access Edge Computing* (MEC), que aproxima do usuário o poder computacional e os arquivos comumente buscados que antes eram disponíveis somente no CN [Spinelli e Mancuso, 2020]. Estas tecnologias possibilitam a implementação de sub-redes específicas para cada aplicação, através do Fatiamento da Rede - *Network Slicing*, separando logicamente uma aplicação da outra, o que facilita o gerenciamento e atendimento de cada requisito heterogêneo de QoS [Barakabitze et al., 2020].

Estes avanços tecnológicos são determinantes para utilizar as redes 5G como principal meio de comunicação e gerenciamento sem fio de diversas aplicações. Dentre essas, as aplicações críticas despertam grande interesse da comunidade científica e da indústria, devido aos requisitos de desempenho estritos, e por abrangerem desde aspectos cotidianos até os mais diversos setores produtivos da sociedade, como agricultura, serviços, indústria e tecnologia. São consideradas críticas as aplicações que dependem fortemente da qualidade e continuidade de serviço oferecidas pela rede de suporte, apresentando requisitos estritos de parâmetros de desempenho, como latência, vazão e confiabilidade da conexão. As redes de suporte disponíveis até o surgimento do 5G não suprem as necessidades dessas aplicações, sendo necessário, portanto, evoluir a tecnologia de rede para a próxima geração.

A implantação das aplicações críticas é desafiadora devido às limitações de grande parte das redes de comunicação ainda hoje utilizadas. A rede LTE 4G, por exemplo, não supre as necessidades de aplicações que precisam de uma alta densidade de dispositivos conectados por km², se limitando a ordem de 10^5 conexões simultâneas. A rede 4G também não suporta, de maneira satisfatória, aplicações que demandam elevada vazão, uma vez que a vazão média, concebida no *International Mobile Telecommunications - Advanced (IMT-Advanced)*, por dispositivo da rede 4G está limitada a cerca de 10 Mbit/s [Cox, 2020]. Posteriormente no LTE *Advanced* PRO foi possibilitada uma vazão média de 1 Gbps por dispositivo que é difícil de ser alcançada em um cenário com vários dispositivos conectados. Assim, evidencia-se a necessidade de evolução da rede para a próxima geração.

A rede 5G é capaz de atender aos requisitos mínimos das aplicações críticas devido às mudanças na RAN, como a utilização de novas bandas de frequências, novas larguras de faixa para transmissão, e novas técnicas de acesso ao meio pelo uso de *beamforming*, Entrada Múltipla Saída Múltipla - *Multiple Input Multiple Output* (MIMO) - massivo e a Multiplexação por Divisão de Frequências Ortogonais - *Orthogonal Frequency Division Multiplexing* (OFDM) - com variadas numerologias; e mudanças no CN, como a utilização de SDN e NFV. A evolução tecnológica aumenta a complexidade da rede, de forma que é necessário compreender as novas técnicas utilizadas para propiciar o desenvolvimento ou aprimoramento de novas aplicações, e para o gerenciamento daquelas já existentes [Gupta e Jha, 2015].

Historicamente, as principais evoluções nas redes móveis se deram na RAN, responsável pela conexão entre o dispositivo do usuário e a rede interna da operadora de telefonia, ou CN. Entretanto, com a implementação de novos serviços atendidos pelas aplicações móveis, novos requisitos de QoS passaram a ser exigidos. Estes novos requi-

sitos estão relacionados ao tempo de resposta do servidor, à densidade de dispositivos a serem atendidos por km² e à taxa de transferência de dados ofertada. Logo, para suportar o atendimento de tais aplicações, é necessária uma evolução não só na RAN, mas uma mudança no tratamento das solicitações na CN, para que a rede possa prover o suporte aos serviço sem violar os diferentes requisitos de QoS. Além disso, uma infraestrutura que atenda a diversas aplicações deve ser flexível e de fácil gerenciamento, visto que os recursos físicos se tornam escassos e devem ser utilizados de maneira eficiente.

As aplicações críticas permitem, por exemplo, o atendimento de saúde ou realização de cirurgias remotamente (*eHealth*) [Moglia et al., 2022]; o controle ou automação de plantas industriais remotas (plataformas de petróleo) [Maroufkhani et al., 2022]; a implantação de tecnologias de manufatura colaborativa em indústrias (*smart manufacturing*) [Wu et al., 2021a]; o controle e automação de redes elétricas *Smart Grid* [Esenogho et al., 2022]; o fornecimento de conexão e serviços a áreas remotas como zonas rurais e centros de pesquisa isolados [Cavalcante et al., 2021]; o controle de tráfego em ruas e rodovias *Internet of Vehicles* (IoV) e de *Unmanned Aerial Vehicle* (UAV) [Sehla et al., 2022, Wei et al., 2022].

Este capítulo tem como objetivo apresentar as principais aplicações críticas que vislumbram a rede 5G como habilitadora, discutindo como esta rede pode suprir as necessidades dessas aplicações, com apresentação dos principais desafios, soluções e tendências de pesquisa existentes na literatura. Para isso, serão apresentados os conceitos essenciais e principais requisitos técnicos das aplicações críticas, apontando os principais desafios existentes e como a rede 5G pode solucioná-los. A atividade prática foca na realização de simulações de cenários com a implementação de aplicações em redes de 4ª Geração (4G) e 5G possibilitando a comparação de desempenho.

Ao final deste capítulo, espera-se que os participantes sejam capazes de: (i) compreender os desafios associados à implantação de aplicações críticas, (ii) identificar as principais aplicações críticas viabilizadas pelo 5G, (iii) compreender as principais características e tecnologias habilitadoras da rede 5G e porque essa rede habilita a implantação das aplicações críticas, e (iv) identificar e compreender as principais tendências e desafios de pesquisa associados a tecnologias em ascensão para o desenvolvimento e implantação da rede de 6ª Geração (6G).

O restante desse capítulo está organizado como descrito a seguir. Na Seção 4.2, são apresentados os principais conceitos sobre a evolução das redes móveis celulares, padronização, categorias de serviço e *Key Performance Indicator*s (KPIs) do 5G. Além disso, são apresentadas as tecnologias habilitadoras do 5G. Na Seção 4.3, são apresentadas as principais aplicações críticas que vislumbram as redes 5G como habilitadoras. Na Seção 4.4, são discutidas as tendências das tecnologias habilitadoras, desafios de pesquisa e é apresentada uma demonstração de simulação para pesquisas em redes 5G. Por fim, a Seção 4.5 traz as considerações finais do texto.

4.2. Conceitos Fundamentais das Redes 5G

Com o passar dos anos, as redes móveis celulares se tornaram o principal meio de comunicação usado no mundo. Estas redes, que antes ofereciam apenas serviços de voz através da comutação de circuitos, agora atendem a aplicações heterogêneas com requi-

sitos distintos de operação. Isso só foi possível com os esforços de setores industriais e acadêmicos, a fim de proporcionar estudos e padronizações que permitissem o desenvolvimento de novas tecnologias e o uso de tecnologias já existentes e consolidadas nos mais diversos setores. As redes 5G se caracterizam como a convergência de tecnologias atuais resultantes das evoluções tecnológicas nas áreas de comunicação e computação. Esta convergência proporciona um avanço expressivo nos indicadores de desempenho das redes 5G.

Dessa forma, esta seção aborda os conceitos fundamentais das redes móveis de 5ª Geração. Ressalta-se brevemente a evolução e os indicadores de desempenho das redes móveis celulares. Em seguida, destaca-se a padronização, as categorias de serviço e os requisitos técnicos das redes 5G. Por fim, são ressaltadas as principais tecnologias que permitem ao 5G atingir os indicadores de desempenho desejados, dentre elas a Computação de Borda de Acesso Múltiplo, as Redes Definidas por *Software* e o Fatiamento de Rede.

4.2.1. Evolução das redes móveis

O avanço nas redes móveis celulares vem impulsionando o surgimento de diversas aplicações desde a implantação da 2ª Geração (2G). Os avanços tecnológicos que viabilizaram essa evolução permitiram uma melhor eficiência espectral e o desenvolvimento de novas técnicas de acesso múltiplo ao meio [Gupta et al., 2019, Arshad et al., 2019]. Além disso, as faixas de frequência e largura de banda foram ampliadas, aumentando a taxa de dados, e foram aplicadas novas técnicas de compartilhamento espectral, melhorando a eficiência no uso de recursos computacionais. Os avanços tecnológicos nessas redes foram determinantes para utilizá-las como principal meio de comunicação e gerenciamento sem fio de aplicações críticas.

A Figura 4.1 apresenta as referências temporais do surgimento das gerações das redes celulares, indicando que, aproximadamente, a cada década, há uma evolução geracional dessas redes.

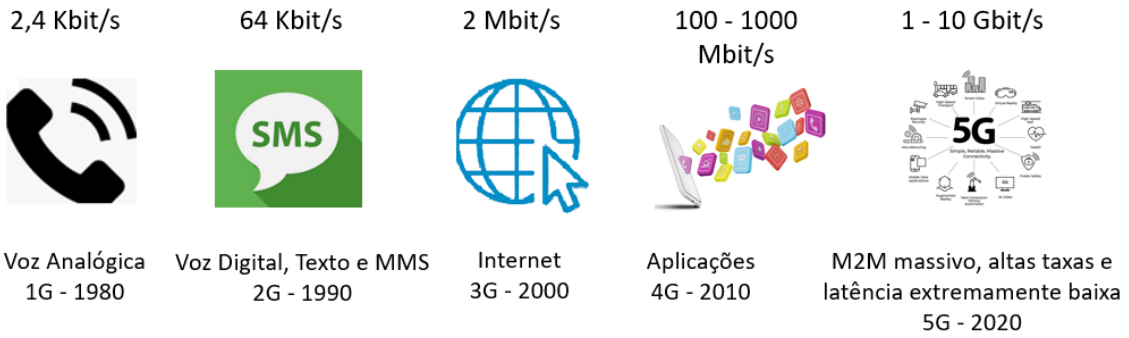

Figura 4.1. Evolução cronológica das gerações de redes celulares. As redes móveis, que antes só ofereciam serviços de voz, hoje suportam diversas aplicações com distintos requisitos de serviço.

As redes móveis de 1ª Geração surgiram nos anos 1980, com transmissão de informação analógica na interface rádio e baseadas na comutação por circuitos. O padrão mais utilizado no mundo foi o norte-americano, chamado de *Advanced Mobile Phone Sys-*

tem (AMPS). Essas redes disponibilizavam apenas serviços de voz, com frequências de operação entre 800 – 900 MHz (bandas "A"e "B"), utilizando modulação em frequência – *Frequency Modulation* (FM) – e com largura de canal de 30 KHz. As redes de 2ª Geração foram lançadas principalmente nos anos 1990 e representaram um grande avanço, sendo um sistema de comunicação digital e com preocupações relacionadas à segurança da interface rádio (enlaces criptografados). Nessa geração, já estavam disponíveis serviços de dados, além dos de voz, com o desenvolvimento do padrão *General Packet Radio Service* (GPRS), também conhecido como 2,5G, e o *Enhanced Data rate for GSM Evolution* (EDGE), usualmente chamado de 2,75G. O GPRS proporcionava taxas de dados até 144 Kbps e, o EDGE, de até 384 Kbps. As redes de 3ª Geração foram projetadas para aumentar a segurança das comunicações e efetuar melhorias nos serviços de voz e dados. Foram lançadas nos anos 2000, utilizando parcialmente a comutação por pacotes e alcançando taxas de até 2Mbps. Uma evolução nessas redes foi permitida através da introdução do padrão *High Speed Packet Access* (HSPA), que é resultado da combinação entre o *High Speed Downlink Packet Access* (HSDPA) e o *High Speed Uplink Packet Access* (HSUPA). Posteriormente, o HSPA recebeu aperfeiçoamentos, permitindo um aumento da taxa nominal máxima de transmissão, da capacidade do sistema, e redução da latência, sendo então denominado de HSPA+, o qual é, usualmente, chamado de padrão 3,75G da telefonia móvel. O HSPA proporciona taxas de até 42 Mbps de *downlink* e 11,5 Mbps de *uplink* em um canal de 5 MHz [De Alwis et al., 2021, Shaik e Malik, 2021].

O início da implantação do HSPA coincide com o aparecimento dos *smartphones* com mais recursos e aplicativos. Em paralelo, em 2008, a União Internacional de Telecomunicações - *International Telecommunications Union* (ITU) - publicou os requisitos para o IMT-Advanced, que estabelece os níveis de desempenho da rede a serem atingidos no padrão de 4a geração de redes móveis. A Evolução a Longo Prazo - *Long-Term Evolution* (LTE), padrão cujo desenvolvimento teve como objetivo atender ao IMT-Advanced e que foi em grande parte coordenado pelo 3GPP, tornou-se amplamente dominante. O LTE foi desenvolvido de forma a otimizar a comunicação de dados. Nele, pela primeira vez em redes móveis, a comutação é totalmente por pacotes e as chamadas de voz acontecem preferencialmente por comutação IP, através da funcionalidade conhecida como *Voice over LTE* (VoLTE). Ao longo dos anos, foram implementados aperfeiçoamentos ao LTE, como a agregação de portadora - *Carrier Aggreagation* (CA), resultando nos padrões LTE *Advanced* e LTE *Advanced Pro* (usualmente denominados de padrões 4,5G e 4,75G respectivamente), com capacidade de transmissão aumentada em relação à primeira versão do LTE. Além disso, no LTE *Advanced Pro* as possibilidades de comunicação entre máquinas é expandida, sendo um marco importante do *roadmap* desta forma de comunicação que viria a se consolidar no 5G [Cox, 2020].

Entretanto, a rede 4G não supre as necessidades de aplicações que precisam de uma alta densidade de dispositivos conectados por km². Além disso, essa rede não suporta aplicações que demandam elevada vazão, alcançando no máximo 100 Mbit/s. Cabe observar que, apesar de a vazão média de pico por dispositivo conectado à rede 4G poder chegar, conceitualmente no LTE *Advanced* PRO, a 1 Gbit/s, ela não é alcançada em um ambiente densamente conectado. Assim, evidencia-se a necessidade de evolução da rede para a próxima geração. A rede 5G é capaz de atender aos requisitos mínimos das aplicações críticas devido ao aperfeiçoamento da RAN e à implementação de uma nova

arquitetura no CN.

Em 2012, a *International Telecommunications Union - Radiocommunication Sector* (ITU-R) deu início ao programa IMT-2020, que desencadeou atividades de pesquisa, no mundo inteiro, para a concepção de redes móveis de 5ª geração. De acordo com a visão estabelecida, os sistemas IMT-2020 deveriam ampliar a capacidade e funcionalidades previstas nos objetivos do IMT-Advanced, conforme mostrado na Figura 4.2. Especificamente, os sistemas da próxima geração deveriam suportar aplicativos de baixa a alta mobilidade, e uma ampla gama de taxas de dados, de acordo com os requisitos do usuário e do serviço em vários cenários de utilização. Além disso, deveriam dispor de recursos para aplicativos multimídia de alta qualidade em uma ampla gama de serviços e plataformas, proporcionando uma melhoria significativa no desempenho e qualidade de serviço.

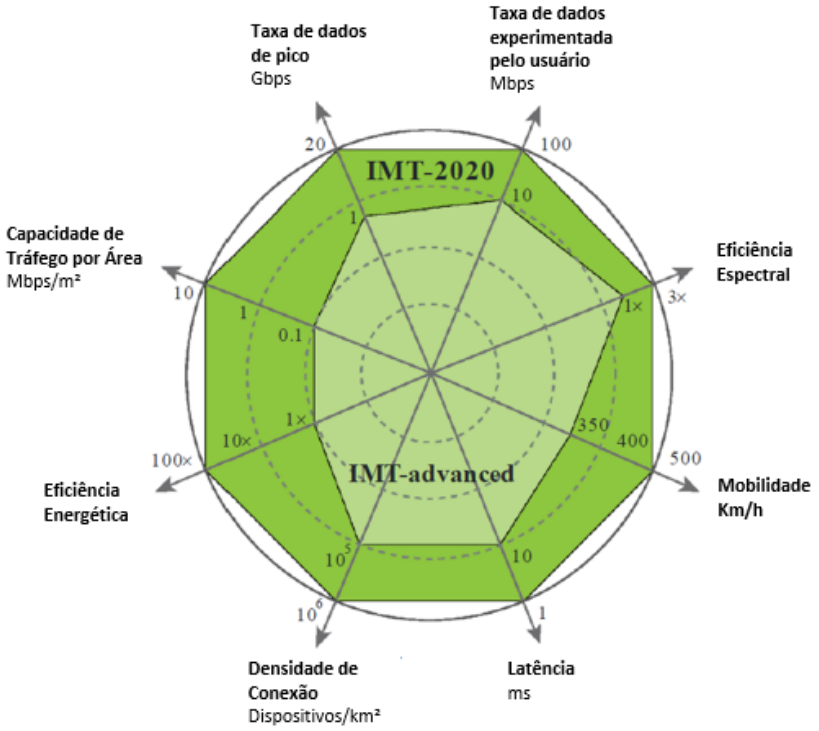

Figura 4.2. Comparação entre os indicadores de desempenho das redes móveis 4G (IMT-Advanced) x 5G (IMT-2020). Avanços expressivos na taxa de dados, densidade de dispositivos conectados e eficiência espectral e energética são resultado dos avanços tecnológicos implementados nas redes móveis. Adaptado de [ITU-R, 2015b]

Após a definição e refinamento do programa IMT-2020, o 3GPP, órgão encarregado de desenvolver as especificações técnicas para o sistema aderente ao IMT-2020, iniciou, em 2015, o trabalho gradual de definição das tecnologias que deveriam compor este novo sistema. O *Release* 15 do 3GPP, finalizado em junho de 2019, está associado à primeira fase das especificações 5G que definem características básicas desses sistemas, e o *Release* 16 especifica a segunda fase das especificações, definindo recursos adicionais que, combinados com os da primeira fase, satisfazem aos requisitos principais do

IMT-2020.

A implantação do 5G vem sendo realizada em dois modos possíveis: não autossuficiente - *Non-Standalone* (NSA), no qual o 5G aproveita parte da infraestrutura (principalmente o CN) do 4G para ser implementado; e o autossuficiente - *Standalone* (SA), no qual a rede 5G é implementada independentemente da rede 4G existente. O 5G SA apresenta desempenho melhor que o 5G NSA, mas o último é uma boa opção para as operadoras fazerem a migração para a nova geração com custos diluídos ao longo do tempo.

Na RAN da rede 5G, as principais inovações compreendem: (i) o uso de novas bandas de frequência (bandas sub-6 GHz e de ondas milimétricas) com larguras de faixa significativamente ampliadas; (ii) novas formas de onda utilizando a técnica OFDM com numerologia adaptativa e um novo e mais eficiente código (*Low-Density Parity-Check* (LDPC)) para controle de erro na camada física; (iii) técnicas multi antenas expandidas (MIMO massivo), que aumentam a eficiência espectral e permitem a implantação de técnicas de *beamforming* com multiplexação espacial [Shaik e Malik, 2021, Cox, 2020]. Já no CN, há mudança completa na arquitetura, sendo baseada em serviços, fatiamento de rede e NFV. Além disso, a Computação de Borda de Acesso Múltiplo - MEC- é implementada para fornecer processamento mais próximo ao usuário, reduzindo assim o tempo de resposta pela rede de suporte [Plachy et al., 2021a]. Essa mudança provê atendimento às necessidades específicas das categorias de aplicações e permite utilizar de forma eficiente e compartilhada os recursos computacionais disponíveis. Também são usadas técnicas para separação do Plano de Controle, que é responsável pelo gerenciamento e controle da sinalização dos dispositivos conectados, do Plano de Usuário, responsável pela entrega do tráfego de dados, (através da Separação do Plano de Controle e do Usuário - *Control and User Plane Separation* (CUPS)) possibilitando que os KPIs de cada categoria de aplicações sejam atingidos.

O 3GPP identificou vários mercados potenciais e um grande número de casos de uso para o 5G [3GPP, 2016], os quais foram referendados pelo ITU-R [ITU-R, 2015b], e, a partir deste estudo, ficaram consagradas as três famílias de casos de uso para o 5G: a *Enhanced Mobile Broadband* (eMBB), que atende a casos similares aos atendidos pelo LTE 4G, mas com capacidades (taxa de dados principalmente) ampliadas; a *Ultra-Reliable and Low Latency Communication* (uRLLC), para cenários que possuem requisitos de alta confiabilidade na conexão e baixa latência; e a *Massive Machine Type Communications* (mMTC), que possibilita a comunicação entre máquinas (dispositivos autônomos) com altíssima densidade de unidades por área de cobertura [ITU-R, 2015b].

4.2.2. Padronização e órgãos reguladores

Em se tratando de padronização do 5G, órgãos padronizadores e associações desempenham uma papel de grande relevância. A padronização visa implementar uma tecnologia e visão de rede em consonância com as necessidades do mercado e da sociedade. Os órgãos de maior destaque na padronização do 5G são: 3GPP, ITU, Institute of Electrical and Electronics Engineers (IEEE), *European Telecommunications Standards Institute* (ETSI), *Global System for Mobile Communications Association* (GSMA) e *Telecommunications Industry Association* (TIA) [Barakabitze et al., 2020].

Em 2015, o ITU-R finalizou a recomendação M.2083-0 (2015) que define o arcabouço e os objetivos globais para o desenvolvimento das redes móveis 5G, conforme apresentado na Figura 4.2. Na visão do ITU-R, a evolução das redes celulares proporciona a evolução social e tecnológica contribuindo, para o desenvolvimento econômico, e foi com base nisso que o arcabouço foi construído. Já o 3GPP se encarregou do desenvolvimento e padronização das tecnologias necessárias e das classes de serviço eMBB, uRLLC e mMTC, através dos *Releases* 15 e 16. Especificamente, o *Release* 15 se relaciona aos serviços de eMBB, em relação ao CN, e busca desenvolver e padronizar uma arquitetura baseada em serviços, o projeto de um sistema multiacesso, a característica nativa de nuvem e o fatiamento da rede, além de iniciar a especificação para serviços de uRLLC. O *Release* 16 se concentrou em finalizar as especificações para os serviços de uRLLC, além de permitir suporte total para a Internet das Coisas Industriais [Ghosh et al., 2019].

Conforme mencionando anteriormente, foi definido que a implantação do 5G poderia ser realizada em dois modos: modo não autossuficiente – NSA, no qual o 4G e o 5G coexistem, e o modo autossuficiente – SA, no qual há total soberania da rede 5G. Especificamente, no modo NSA, o equipamento do usuário é ancorado em antenas da rede 4G e as antenas da rede 5G – *New Radio* (NR) são utilizadas quando existe a exigência e cobertura de conexão. Além disso, no modo NSA o núcleo da rede utilizado continua sendo o núcleo da rede 4G – *Evolved Packet Core* (EPC). De maneira oposta, no modo SA, o equipamento do usuário se conecta diretamente à NR e utiliza o núcleo das redes 5G – *5th Generation Core* (5GC) [Liu et al., 2020a]. Estes dois modos são habilitados principalmente pela separação do plano de controle e do usuário. A Figura 4.3 demonstra esta estrutura.

No *Release* 17, foi buscada uma melhoria do desempenho da rede para os serviços e cenários de uso existentes, através da proposição de soluções de rede inteligentes. Esse melhoria foi chamada de 5G *Advanced* e possui como uma das principais novidades a utilização de inteligência artificial baseada em aprendizado de máquina no gerenciamento da rede, solucionando problemas de otimização. A inteligência artificial também é utilizada na melhoria da interface rádio, melhorando o desempenho de sistemas complexos de antenas. Também foram adicionadas melhorias em funcionalidades já existentes da NR como gerenciamento da formação de feixes e pontos de múltiplas transmissões e recepções. Foram feitas melhorias no compartilhamento dinâmico do espectro entre as redes 4G e 5G, economia de energia em equipamentos de usuário, posicionamento, entre outros. As novas funcionalidades ficaram por conta da criação de uma classe de equipamentos de usuário com capacidade reduzida que possui requisitos de serviços intermediários, na introdução das topologias de rede baseadas em satélite *Non-Terrestrial Networks* (NTN), e na utilização de frequências superiores a 71GHz na NR.

O *Release* 18 começou a ser discutido em junho de 2021, prevendo melhorias nos cenários de uso eMBB (consumo de energia, MIMO e mobilidade), aplicações (*Extended Reality* (XR), segurança pública e dispositivos com capacidade reduzida) e aplicação de inteligência artificial na camada física [Rahman et al., 2021]. A Figura 4.4 resume os principais avanços obtidos ao longo dos Releases 15 a 17, e a previsão dos futuros Releases para a evolução do 5G até a geração seguinte de redes móveis.

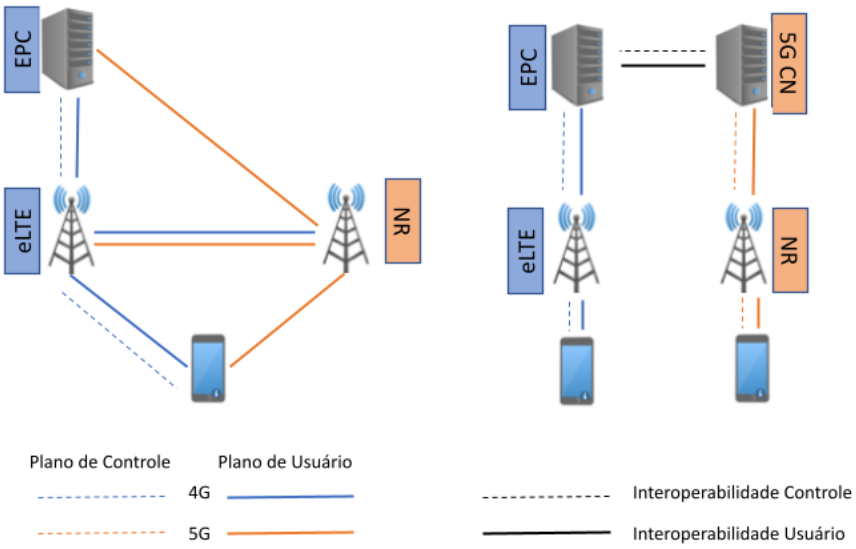

Figura 4.3. Modos de operação: NSA, à esquerda, e SA, à direita. No modo NSA as redes 4G e 5G coexistem, e o dispositivo móvel se conecta à rede através das antenas da rede 4G ou antenas da rede 5G, quando disponível, porém o núcleo da rede utilizado ainda é o EPC. No modo SA existe total separação entre as redes 4G e 5G, e o núcleo da rede 5G já é usado para o fluxo de dados.Adaptado de [Liu et al., 2020b]

4.2.3. Categorias de serviço nas redes 5G

O ITU-R, quando começou em 2015 a definição da estrutura e objetivos gerais do novo padrão de internacional de telecomunicações móveis para 2020 e além (resultando no conjunto de requisitos contidos no IMT-2020), considerou a evolução do papel que as comunicações móveis alcançaram no cotidiano da sociedade. O novo padrão deve considerar o surgimento de novas necessidades que demandem grande uso de dados a taxas elevadas, uma grande quantidade de dispositivos conectados, latências extremamente baixas e alta confiabilidade. Portanto, do novo padrão são esperados alguns comportamentos que não eram suportados pelos padrões de redes móveis anteriores [ITU-R, 2015b]. Entre eles podemos destacar:

1. Conectividade instantânea: algumas aplicações (saúde, realidade virtual, segurança) demandam uma resposta imediata após uma simples solicitação. Essa resposta só é possível com a autenticação e processamento dos dados mais próximo ao usuário.

2. Comunicação em tempo real: previsão de comunicação em tempo real entre máquinas (carros autônomos, comunicação em nuvem, controle de tráfego, *smart grid*, *e-health, indústrias* sem necessidade de um usuário humano.

3. Alta densidade de usuários: apesar da grande quantidade de usuários e dispositivos conectados, é necessário que todos experimentem um serviço com boa qualidade e com baixo consumo de energia.

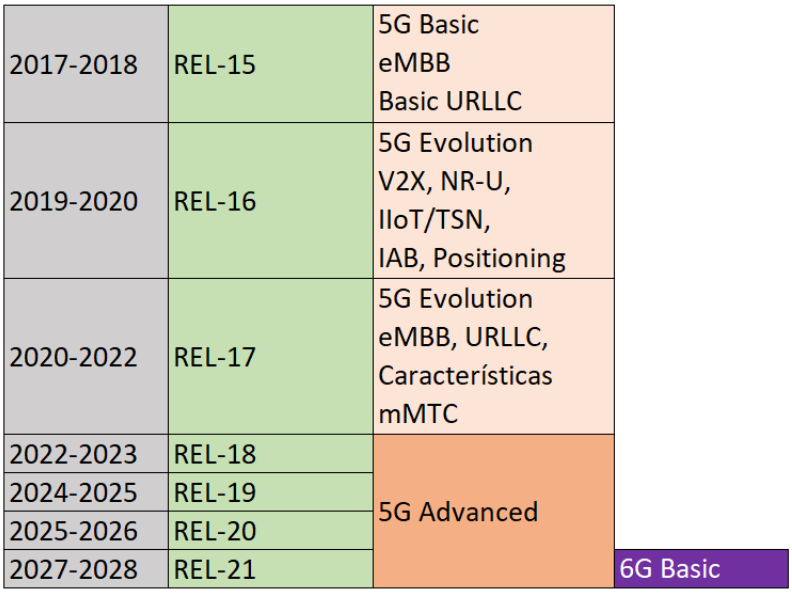

Figura 4.4. Cronograma Previsto do 3GPP para Evolução do 5G. As especificações para o desenvolvimento e padronização das tecnologias necessárias do 5G e dos cenários de eMBB, uRLLC e mMTC se dividem em cerca de 6 *Releases*, onde cada uma oferece o aprimoramento de tecnologias previamente inseridas no cenário de redes móveis em *Releases* anteriores, além da padronização de novo casos de uso. Adaptador de [Rahman et al., 2021]

4. Alta mobilidade: mesmo usuários em altas velocidades, em carros ou trens, precisam experimentar um serviço de qualidade e similar aos usuários estáticos.

5. Serviços Multimídia Melhorados e aumento do tráfego de dados: o consumo de dados irá aumentar e aplicações de mídia serão as principais responsáveis (definições ultra altas, projeções 3D, vídeos imersivos, entre outros).

6. Internet das Coisas: Com a perspectiva de que, em um futuro próximo, quase todos objetos (eletrodomésticos, sensores, atuadores, carros, casas) estejam conectados entre si, um novo universo para desenvolvimento de aplicações e serviços emerge. Essa necessidade de conexão para um número incontável de dispositivos precisa ser levada em consideração.

Com base nesses comportamentos, foram definidos 3 classes de serviços básicos que balizaram o desenvolvimento das redes de quinta geração. São eles:

1. ***Enhanced Mobile Broadband* (eMBB)**: no cenário de aprimoramento de banda larga móvel, são atendidas as necessidades de comunicação multimídia dos usuários humanos das redes 5G. Foi buscada uma melhoria no desempenho e experiência para os usuários desse tipo de aplicação.

2. ***Ultra-Reliable and Low Latency* (uRLLC)**: o cenário de comunicações ultra-confiáveis e latências extremamente baixas possui requisitos estritos de transmissão, atraso e disponibilidade. Com isso, ele é voltado para aplicações de controle industrial, segurança de transportes, saúde (p.ex. cirurgia remota), entre outros.

3. ***Massive Machine-Type Communication* (mMTC)**: cenário de uso caracterizado pela comunicação massiva entre máquinas. É previsto um grande número de dispositivos conectados transmitindo, com pequeno consumo de energia, um volume da dados relativamente baixo com dados não sensíveis ao atraso. Os dispositivos são de baixo custo e com alta vida útil da baterias.

A Figura 4.5 apresenta alguns dos exemplos vislumbrados para cada categoria de uso das redes 5G. Aplicações como carros autônomos e automação industrial demandam alta confiabilidade e disponibilidade da rede, sendo atendidas pelos cenários de uRLLC. Já as cidades inteligentes impõem à rede sem fio uma alta densidade de dispositivos conectados, sendo principalmente atendidos pelos cenários de mMTC.

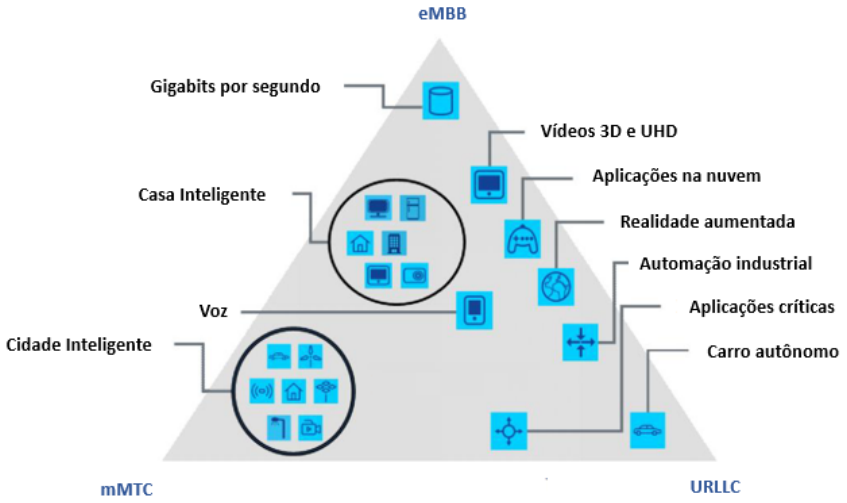

Figura 4.5. Categorias de uso do 5G, considerando as classes de serviço eMBB, mMTC e uRLLC, que mapeiam os novos tipos de serviço que são habilitados com o uso do 5G. Adaptado de [ITU-R, 2015b]

4.2.4. Capacidades–chave das redes 5G

Diante das categorias de uso definidas para o 5G, ficou claro que as redes deveriam passar por uma evolução significativa quanto aos serviços oferecidos e aos KPIs traçados em relação à geração anterior. Havia necessidade de melhoria no pico da taxa de dados (Gbps), na taxa de dados experimentada pelos usuários (Mbps/Gbps), na latência (ms), mobilidade (km/h), densidade de conexão (usuários/km²), eficiência energética (b/joule), eficiência espectral (b/s/Hz) e capacidade de tráfego por área (Mbps/s/m²). A Figura 4.2 apresenta a evolução dos requisitos entre as redes 4G e 5G (definidos respectivamente no IMT *Advanced* e IMT 2020), mostrando em ordens de grandeza a evolução do 5G com relação ao 4G de diversos parâmetros de rede.

Apesar de não ser considerada uma capacidade chave a melhoria na capacidade de localização de dispositivos na rede 5G é aproveitada por diversas aplicações críticas. Essa melhoria é possível por conta de novas características do 5G que tornam mais fácil e exato esse processo. Os 3 fatores básicos são: o menor tamanho médio das células (podem existir picocélulas), o uso de larguras de banda e frequências maiores torna mais preciso

o cálculo do tempo de chegada do sinal e a maior densidade de dispositivos conectados facilita a troca de informações de localização [Zhang et al., 2017].

As capacidades da rede 5G possuem pesos diferentes quando correlacionadas com os cenários de uso previstos. Apesar de algumas capacidades serem aplicáveis a mais de um cenário de uso, os valores necessários para alcance de um desempenho satisfatório podem ser extremamente diferentes. Por essa razão, foi realizada uma classificação em 3 níveis de importância (alto, médio e baixo) de acordo com a dependência que cada cenário de uso possui em relação a determinada capacidade da rede. A Tabela 4.1 apresenta um resumo dos KPIs versus classes de serviço.

Como pode ser observado, o eMBB possui grande dependência de uma boa taxa de dados experimentada pelo usuário e boa mobilidade. Para os cenários de uRLLC, a baixa latência e alta mobilidade será de extrema importância.Quando analisa-se as aplicações mMTC, a alta densidade de conexões é essencial, sendo o KPI mais importante para essa classe [ITU-R, 2015b].

Tabela 4.1. Importância dos KPIs nas classes de serviço. Adaptado de [ITU-R, 2015b]

KPI	eMBB	uRLLC	mMTC
Taxa de dados de pico	Alta	Baixa	Baixa
Taxa de dados experimentada pelo usuário	Alta	Baixa	Baixa
Eficiência espectral	Alta	Baixa	Baixa
Mobilidade	Alta	Alta	Baixa
Latência	Média	Alta	Baixa
Densidade de conexões	Média	Baixa	Alta
Eficiência energética da rede	Alta	Baixa	Média
Capacidade de tráfego por área	Alta	Baixa	Baixa

4.2.5. Tecnologias habilitadoras

Para atender aos requisitos mencionados anteriormente, as redes 5G implementam diversas evoluções na Rede de Acesso por Rádio - RAN - e no CN. A arquitetura passou de rígida e monolítica para flexível e baseada em serviços, resultando no uso eficiente dos recursos computacionais disponíveis e um gerenciamento fim-a-fim das aplicações para atender aos requisitos heterogêneos. O paradigma das **Redes Definidas por *Software*** - SDNs - aliado à **Virtualização das Funções de Rede** - NFV - são as principais tecnologias habilitadoras das redes 5G que permitem o atendimento rápido das demandas e o uso eficiente dos recursos. A **Computação de Borda de Acesso Múltiplo** - MEC - aproxima do usuário o poder computacional antes disponível apenas no núcleo da rede, diminuindo a latência das aplicações e o fluxo de dados nos enlaces de *backhaul* (por exemplo, enlaces entre as estações rádio–base (que no 5G são denominadas de *next Generation Node B* (gNB) e o núcleo da rede). Todas estas tecnologias habilitam o **Fatiamento da Rede** - *Network Slicing* - que isola de maneira lógica uma aplicação de outra, o que permite a flexibilidade das operações [Barakabitze et al., 2020]. A arquitetura e interoperabilidade destas tecnologias pode ser observada na Figura 4.6. Serviços distintos (por exemplo, aplicações de processamento de dados, entrega de conteúdo, etc.) coexistem nas redes

5G, sendo estruturados pelo Encadeamento de Funções de Serviço - *Service Function Chaining* (SFC). Cada instância dessas funções são distribuídas próximas ou distantes do usuário, de acordo com a sua função. Em alguns cenários de uso, a funções são instanciadas na MEC, uma vez que não é necessária a comunicação com o CN. A fatia de rede de cada serviço garante que os recursos alocados para o serviço sejam consumidos somente por ele, assegurando o atendimento aos requisitos de QoS.

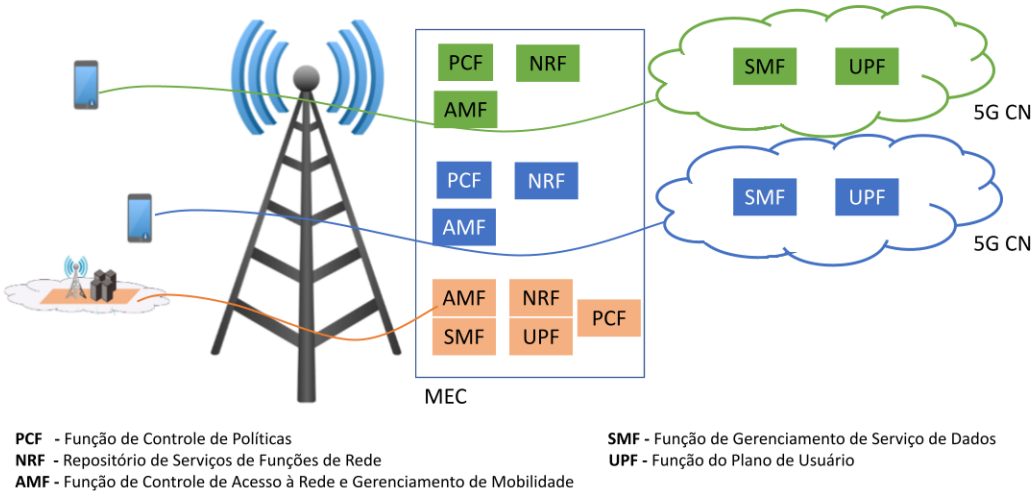

Figura 4.6. Interoperabilidade das tecnologias habilitadoras. O uso das SDNs e da NFV permite a alocação das funções de serviço, que são encadeadas e executadas próximas ou distantes do usuário, compondo a fatia da rede reservada para a aplicação. Em alguns casos, todas as funções de rede são alocadas na MEC, capaz de executá-las na borda da rede. Adaptado de [3GPP, 2017].

4.2.5.1. SDN

O principal objetivo das **SDNs** é a separação do Plano de Controle e do Plano de Dados [Fernandes et al., 2011]. Nas redes tradicionais, o gerenciamento e a manutenção da rede são complexos e custosos, uma vez que as políticas de encaminhamento de pacotes (Plano de Controle) e o fluxo de dados (Plano de Dados) estão centralizados no mesmo equipamento de rede (por exemplo, *switch*, roteador), e as políticas de encaminhamento só podem ser alteradas com uma nova configuração do equipamento. Além disso, os comandos de configuração são específicos de cada fornecedor, não existindo um padrão de linguagem. Nas SDNs, o controle das políticas de encaminhamento de pacotes é migrado para um controlador, que gerencia os recursos do Plano de Dados [Ahvar et al., 2021].

A definição de separação do Plano de Controle e de Dados nas SDNs é o conceito fundamental da separação dos Planos de Controle e de Usuário [Khorsandroo et al., 2021], responsáveis pelo gerenciamento e controle da sinalização dos dispositivos conectados e pela entrega do tráfego de dados aos usuários, respectivamente. Nas redes 5G, a CUPS permite que as Funções de Redes Virtualizadas - *Virtualized Network Functions*s (VNFs) - relacionadas ao controle e sinalização do dispositivo do usuário sejam alocadas no núcleo da rede, e as VNFs da aplicação sejam alocadas próximas ao usuário.

A Figura 4.7 demonstra a dinâmica do uso das redes 5G com o CUPS. Dentre as funções de controle destaca-se a Função de Acesso e Mobilidade – *Access and Mobility Management Function* (AMF), responsável pelo controle de acesso e de mobilidade do usuário por prover serviços de comunicação e localização para outras funções de rede; a Função de Controle de Políticas – *Policy Control Function* (PCF), que incorpora as politicas de mobilidade, fatiamento da rede e *roaming*; e a Função de Gerenciamento de Sessão – *Session Management Function* (SMF), que cria e mantém as sessões de conexão, selecionando e alocando o endereço de IP e configurando as regras de tráfego da Função do Plano de Usuário. A Função do Plano de Usuário – *User Plane Function* (UPF) é a principal função responsável pelo roteamento dos dados para o usuário [Shah et al., 2021]. Para facilitar o entendimento sobre as funções da rede, a Tabela 4.2 apresenta os elementos da rede 4G e a respectiva correspondência na rede 5G.

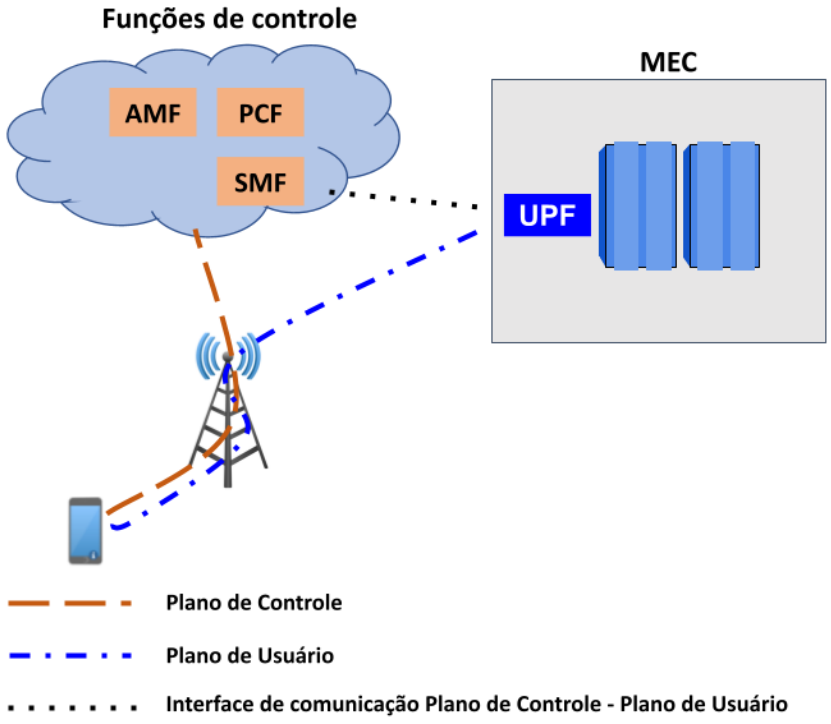

Figura 4.7. Separação do Plano de Controle e de Usuário. No dispositivo do usuário, o acesso à rede, a sessão, a mobilidade e as politicas de acesso são controladas pelas Funções de Controle, no Plano de Controle. O tráfego de dados é controlado é controlado pela Função do Plano de Usuário. Adaptado de [Shah et al., 2021]

4.2.5.2. NFV

A **NFV** é a virtualização das funções de rede que antes eram realizadas por equipamentos específicos (por exemplo, *firewalls*, NAT, TCP). Estas funções se tornam VNFs que podem ser executadas em máquinas virtuais sob recursos computacionais heterogêneos, reduzindo o Custo de Capital – *Capital Expenditures* (CAPEX) – e o Custo Opera-

Tabela 4.2. Relação entre elementos da rede 4G e funções das redes 5G

Equipamento 4G	Tarefa	Função 5G CN
Entidade de Gestão de Mobilidade	Gestão de Mobilidade	AMF
	Autenticação do usuário	AUSF
	Gestão da sessão	SMF
Gateway de rede de dados de pacote	Encaminhamento de dados do plano do usuário	UPF
Função de Política e Recurso de Cobrança	Politica de QoS	PCF
Servidor de Assinante Doméstico	Armazenamento do perfil do usuário	UDM

cional – *Operational Expenditures* (OPEX) – das redes. Neste sentido, as aplicações são construídas a partir do encadeamento destas VNFs, com enlaces lógicos interligando as funções de rede [Barakabitze et al., 2020] e todo o ciclo de vida da aplicação gerenciado por uma entidade da rede. A Figura 4.8 demonstra o encadeamento de funções de rede de um serviço de solicitação de conteúdo. O usuário ingressa na rede e a VNF1 realiza a autenticação com base em seus dados. A VNF2 inspeciona os pacotes de solicitação, evitando assim, invasões na rede, enquanto a VNF3 checa a disponibilidade dos dados na borda da rede. A VNF4 é executada caso os dados solicitados se encontrem somente no CN, caso contrário ela não é encadeada na aplicação. Por fim a VNF5 entrega o conteúdo solicitado. Vale ressaltar que o uso da NFV possibilita um controle fino das funções que serão executadas pela aplicação como demonstrado na VNF5, que pode dar origem a duas VNFs, uma de cobrança – VNF5-1, e outra de aceleração de vídeo –VNF5-2, sem a necessidade de instalação de novos equipamentos na rede. Todo o ciclo de vida da aplicação é gerenciado pelo Gerenciado de Aplicações, responsável por alocar os recursos necessários para a execução das VNFs.

A NFV vem ocupando lugar de destaque pela facilidade no uso de recursos heterogêneos. Sua inclusão no projeto de Gerenciamento e Orquestração NFV - NFV MANO (*Management and Orchestration*) - ressaltou os esforços do ETSI na padronização da MEC [Spinelli e Mancuso, 2020].

4.2.5.3. MEC

A **MEC** se caracteriza por entregar recursos de processamento ou conteúdos em *cache* (por exemplo, vídeos frequentemente buscados, mapas em alta definição) na borda da rede móvel que antes eram disponíveis somente no núcleo da rede. O processamento e entrega de conteúdos comumente buscados mais próximo do usuário diminui a latência de atendimento do serviço e o fluxo de dados nos enlaces entre as *evolved Node B*s (eNBs) e o núcleo da rede de acesso (por exemplo, enlaces de *backhaul*). Ela se divide em três níveis, como demonstrado na Figura 4.9. O Nível de Sistema tem uma visão geral da arquitetura MEC e coordena todos os outros níveis e entidades, sendo composto: pelo *Proxy* de Gerenciamento do Ciclo de Vida da Aplicação do Usuário - (PGCAU), responsável por

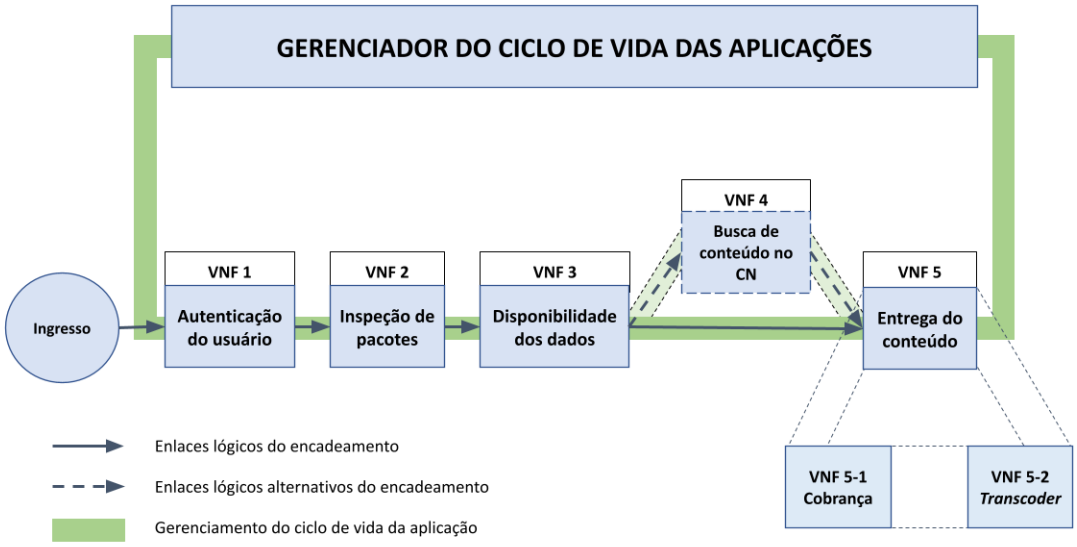

Figura 4.8. Encadeamento de VNFs em um serviço de entrega de conteúdo. As aplicações são construídas a partir do encadeamento das VNFs, que se conectam com enlaces lógicos e podem ser divididas para customização da aplicação. O Gerenciador do ciclo de vida das aplicações é responsável por alocar recursos e monitorar a execução das VNFs. Adaptado de [Wang et al., 2016]

.

checar a existência ou não da aplicação solicitada pelo usuário e por encaminhá-la ao Sistema de Suporte Operacional; pelo Sistema de Suporte Operacional - (SSO), responsável por receber os pedidos de uso da infraestrutura e encaminhá-los para o Orquestrador da MEC caso forem aceitos; e pelo Orquestrador MEC - (OMEC), que retém toda informação do sistema MEC, como os recursos disponíveis, os serviços, os *hosts* instanciados, as políticas das operadoras, e monitora a topologia da rede, além de escolher o melhor *host* para instanciar a aplicação considerando os requisitos de QoS. O Nível de *host* MEC também é composto por três entidades: o Gerente de Virtualização de Infraestrutura - (GVI), que gerencia da alocação e liberação dos recursos virtualizados e prepara a infraestrutura para o uso de uma imagem de *software*; o Gerente da Plataforma MEC - (GPMEC), que gerencia a plataforma MEC e o ciclo de vida das aplicações, repassando ao OMEC informações relevantes recebidas do GVI sobre falhas ou medidas de desempenho. A última entidade do Nível de *host* MEC são os *Hosts* MEC, dividido em três sub-entidades: a Infraestrutura Virtualizada - (IV), que fornece os recursos de rede e computação (por exemplo, servidores) e o Plano de Dados; a Plataforma MEC, que é o controlador da SDN, recebendo as regras de tráfego e configurações de DNS do GPMEC e instruindo o Plano de Dados a segui-las; e as Aplicações MEC, que são instanciadas em máquinas virtuais ou *containers* sob a infraestrutura virtualizada. O último nível da estrutura MEC é o de Redes de Comunicações, que é composto pelas entidades de comunicação que se conectam à estrutura MEC (por exemplo, as redes 5G).

É importante destacar que uma revisão do ETSI em 2018 incluiu a estrutura do Gerenciador e Orquestrador - (MANO) - no padrão MEC. Esta inclusão basicamente altera o GPMEC para Gerente de Plataforma MEC NFV, que agora delega as atividades de gerenciamento do ciclo de vida de funções de rede para um Gerente de Função de Rede

dedicado a cada *host* ou até a cada aplicação. Além disso, o OMEC se subdividiu entre Orquestrador de Aplicação MEC - OAMEC - e Orquestrador de NFV - ONFV [Spinelli e Mancuso, 2020]. Estas alterações na estrutura MEC proporcionaram uma sensibilidade maior em relação às fatias de rede, promovendo o isolamento fim-a-fim entre aplicações.

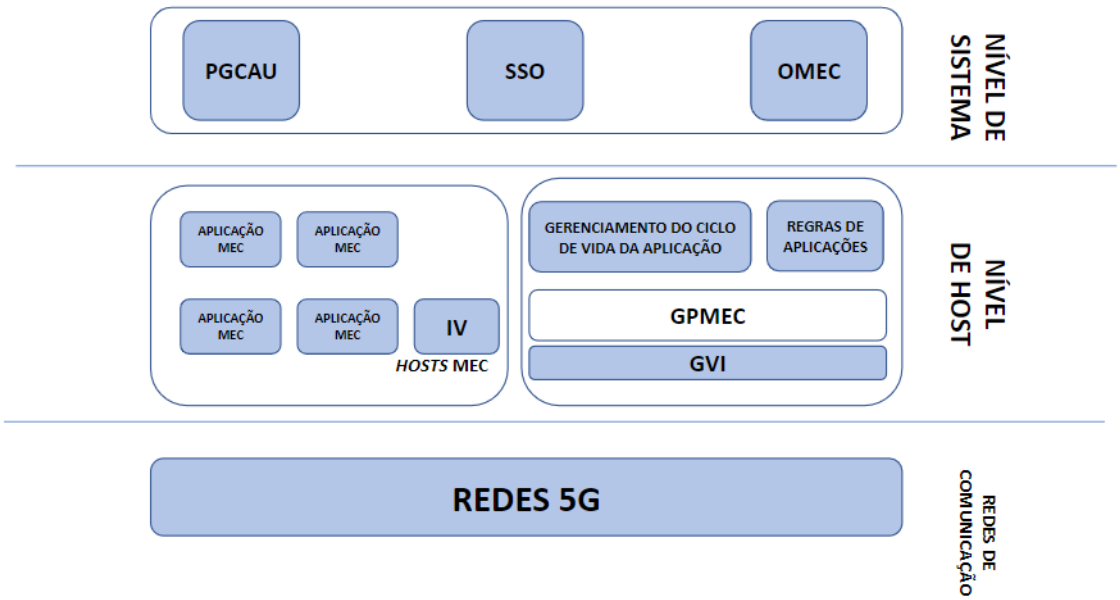

Figura 4.9. Estrutura MEC proposta pelo ETSI. A MEC é composta pelo nível de Sistema, que recebe as solicitações de uso da infraestrutura e, com uma visão geral da arquitetura MEC, coordena todos os outros níveis e entidades; pelo nível de *Host*, que gerencia diretamente o uso dos recursos virtualizados das SDNs, preparando a imagem das VNFs e alocando as funções de rede nos respectivos *hosts*; e pelo nível de Comunicação, responsável pela interface dos dispositivos móveis com a rede 5G. Adaptado de [Spinelli e Mancuso, 2020].

4.2.5.4. Fatiamento da Rede

O **Fatiamento de Rede** consiste em separar logicamente uma rede de modo que várias fatias coexistam isoladamente na mesma infraestrutura, com controle e gerenciamento independentes sem interferência uma na outra. O Fatiamento de Rede provê escalabilidade e alta confiabilidade, uma vez que os recursos são usados de forma eficiente e as fatias são monitoráveis e gerenciáveis, sendo reconfiguradas e alocadas em novos recursos em caso de falhas.Além disso, ele permite que operadores da rede 5G customizem as fatias para atendimento dos requisitos estritos de cada aplicação.

A Figura 4.10 denota a visão do *Next Generation Mobile Network* (NGMN) acerca das camadas do fatiamento de rede da rede 5G. O gerenciador da rede recebe uma solicitação para o uso de recursos. Nesta solicitação é descrito o tipo de aplicação que será usada e seus requisitos. Na Camada de Instância de Serviço, se encontram a imagem de *software* e o conjunto das VNFs, além do encadeamento necessário para o atendimento da aplicação. Com base nos requisitos das aplicações e politicas da operadora da rede,

o gerenciador decide se é necessário a customização da fatia. Após a criação dos enlaces lógicos entre as VNFs, a alocação das funções é iniciada. Na Camada de Recursos, as VNF são alocadas sob os recursos computacionais (por exemplo, roteadores, enlaces, partições em servidores) abstraídos. Por fim, a Camada de Fatia de Rede 5G oferece as características de rede necessárias para o uso e gerenciamento do serviço. Uma instância da fatia da rede pode ser composta por nenhuma, uma ou várias sub-redes. A Camada de Recursos consiste nos recursos físicos ou lógicos (por exemplo, roteadores, enlaces, partições em servidores) [Barakabitze et al., 2020].

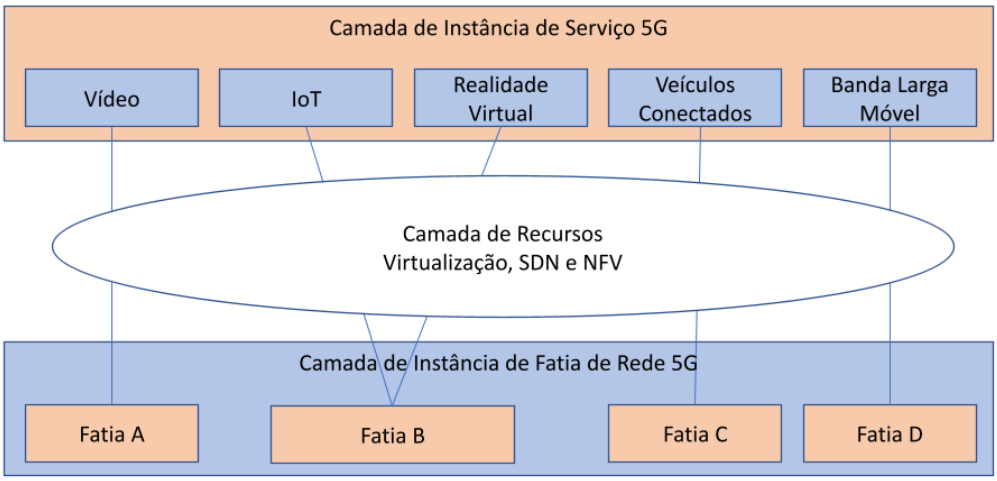

Figura 4.10. Visão do NGMN acerca do fatiamento de rede. O gerenciador da rede encadeia as VNFs e verifica a possibilidade de customização da fatia. Os recursos da rede são alocados e em caso de falha, uma nova fatia é criada para atender a solicitação. Adaptado de [Barakabitze et al., 2020]

As tecnologias citadas anteriormente aumentam o controle do operador de telecomunicações e possibilitam a alta disponibilidade esperada das redes 5G. O uso das SDNs associado às funções virtualizadas proporciona uma granularidade fina aos serviços, possibilitando a utilização da MEC e de seus recursos de maneira eficiente e rápida. Ademais, o Fatiamento de Rede garante a customização da aplicação, com atendimento aos requisitos heterogêneos de QoS, e um isolamento entre diferentes instâncias de serviços, proporcionando assim o atendimento a diferentes aplicações críticas.

4.3. Aplicações Críticas e seus Principais Requisitos e Desafios

Esta seção aborda as principais aplicações que são habilitadas pelas redes de 5ª Geração e os desafios que movimentam a comunidade acadêmica e industrial. Cada aplicação possui requisitos extremamente específicos, que só podem ser atendidos por uma arquitetura que seja flexível, robusta e confiável. Dentre os desafios abordados, destaca-se a eficiência energética, a baixa latência, a confiabilidade das conexões e o grande fluxo de dados.

4.3.1. Internet dos Veículos - IoV

A **Internet dos Veículos** é uma área de pesquisa ativa que integra as redes Ad Hoc Veiculares - *Vehicular Ad Hoc Network*s (VANETs) - e a Internet das Coisas - *Intenet of Things* (IoT), contribuindo diretamente com os Sistemas de Transporte Inteligentes no desenvolvimento das Cidades Inteligentes. Essa integração adiciona novas capacidades às VANETs e amplia o ecossistema da comunicação entre veículos, proporcionando mais segurança para pedestres e veículos, economia de combustível, controle do tráfego e prevenção de colisões [Agbaje et al., 2022]. No ecossistema de IoV, o veículo é um dispositivo inteligente capaz de se comunicar diretamente com os outros dispositivos integrantes da rede veicular, por meio do novo conceito de comunicação chamado Veículo-para-Tudo – *Vehicle-to-Everything* (V2X). Especificamente, o conceito de V2X engloba a capacidade do veículo se comunicar diretamente: com outros veículos, por intermédio da comunicação Veículo-para-Veículo - Vehicle-to-Vehicle (V2V); com dispositivos móveis de pedestres, mediante a comunicação Veiculo-para-Pedestre - *Vehicle-to-Pedestrian* (V2P); com a infraestrutura fixa na beira de estradas, com o uso da comunicação Veículo-para-Infraestrutura - *Vehicle-to-Infrastructure* (V2I); ou com a infraestrutura da rede celular, através da comunicação Veículo-para-Rede - *Vehicle-to-Network* (V2N) [Sehla et al., 2022]. A Figura 4.11 apresenta o ecossistema IoV e o conceito de comunicação V2X.

É possível destacar duas tecnologias que buscam atender aos requisitos de QoS necessários para a IoV: a tecnologia WiFi e a tecnologia celular - *Cellular Vehicle-to-Everything* (C-V2X).

A tecnologia WiFi se baseia na comunicação sem fio para o envio/recepção de mensagens e dados entre os dispositivos, necessitando a de uma infraestrutura auxiliar na beira das ruas e estradas (por exemplo, *Road-Side Unit* (RSU)), no modo infraestruturado, ou da comunicação par a par entre carros ou entre carro e infraestrutura, no modo ad hoc.

A tecnologia C-V2X se baseia no uso da infraestrutura celular para habilitar as comunicações V2X. No ecossistema IoV, a existência de dispositivos distintos gerando dados com estruturas diversas promove a coexistência de diferentes tecnologias de comunicação com o mesmo objetivo: garantir que o usuário possa utilizar os serviços disponíveis com segurança, conforto e eficiência, mantendo a privacidade dos dados trocados pelas entidades da rede [Agbaje et al., 2022].

Nas tecnologias de comunicação WiFi, o primeiro protocolo padronizado para uma rede veicular (por exemplo, VANET) foi o protocolo de Comunicações Dedicadas de Curto Alcance - *Dedicated Short-Range Communication* (DSRC), que é baseado no padrão IEEE 802.11p. Este protocolo habilita apenas as comunicações V2V e V2I, sendo que o alcance é de 300 metros para uma taxa de dados efetiva de 18,06 Mb/s. Além disso, o DSRC não suporta os cenários de alta velocidade buscados pela IoV. Deste modo, o IEEE criou o Grupo de Estudo IEEE 802.11 Próxima Geração V2X - *IEEE 802.11 Next Generation V2X Study Group* - para padronização do protocolo de 802.11bp, baseado na evolução dos protocolos 802.11a ao 802.11ax. Os aprimoramentos na camada de Controle de Acesso ao Meio - *Media Access Control* (MAC) - e na Camada Física permitem que 802.11bp alcance o dobro da taxa de dados e da distância máxima de cobertura em cenários de alta mobilidade e com velocidades desejadas pela IoV para alguns casos de uso [Noor-A-Rahim et al., 2020, Sehla et al., 2022].

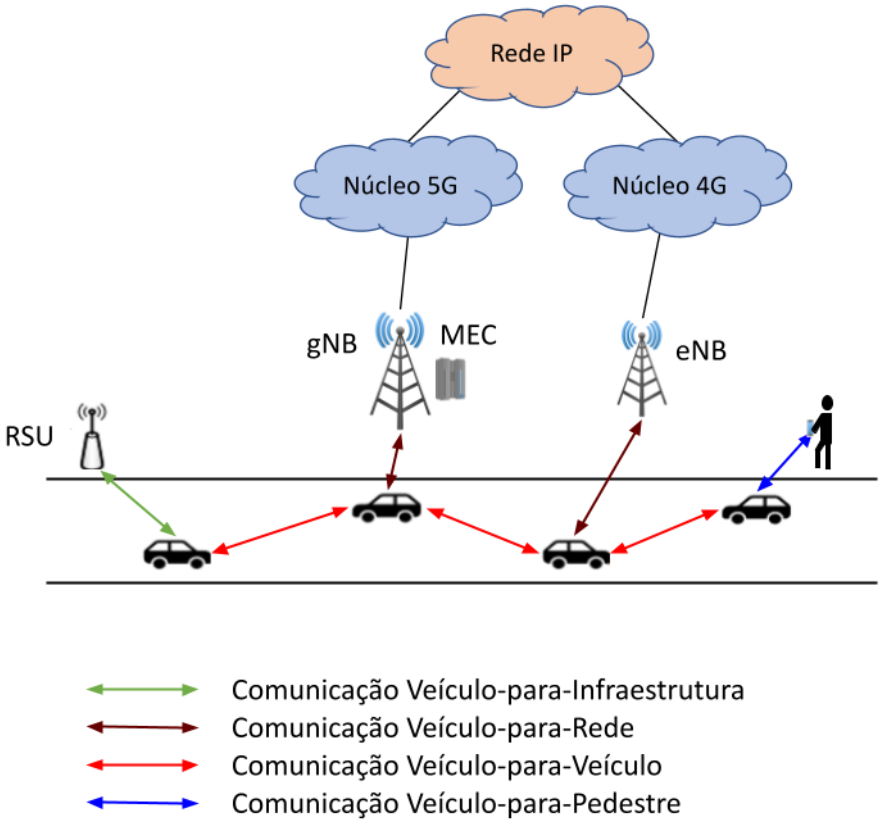

Figura 4.11. Ecossistema da IoV. Neste novo ecossistema o veículo se torna um dispositivo inteligente e se comunica com todos os agentes a sua volta através da V2X, que engloba a comunicação entre o veículo e: sensores e radares na beira de estradas – V2I, a rede 5G – V2N, outros veículos – V2V, os pedestres – V2P. Adaptado de [Soto et al., 2022].

Nas Tecnologias C-V2X, cabe destaque que o *Release 14* do 3GPP foi a primeira especificação contemplando esta tecnologia. No contexto da IoV, ela trata da tecnologia de comunicação *Long-Term Evolution-V2X* (LTE-V2X), que habilita as comunicações V2N, V2V, V2P e V2I, através da tecnologia *Device to Device* (D2D). O uso da D2D, por meio da interface *Sidelink*, permite a comunicação direta entre dois dispositivos próximos sem necessidade de interação com a infraestrutura celular, gerando dois modos de comunicação: o modo 3, usado em regiões cobertas pelas eNBs; e o modo 4, usado em regiões que não são cobertas pelas eNBs. No modo 3 as eNBs agendam e alocam os recursos de rádio para os veículos. Já no modo 4, os veículos reservam automaticamente seus recursos de rádio através de algoritmos semi-persistentes de alocação baseados em detecção. O projeto inicial da LTE-V2X se baseia apenas na troca de mensagens que garantem a segurança em cenários de tráfego moderado, ficando limitados pela tecnologia subjacente da rede Fourth Generation (4G). Entretanto, para cenários V2V, a LTE-V2X já oferta uma latência de $20 \sim 100$ ms com alta confiabilidade e velocidades relativas dos veículos de até 500 km/h. Uma vez que a tecnologia LTE-V2X não é capaz de suprir os requisitos de QoS dos casos de uso mais desafiadores da IoV, o 3GPP realizou melhorias na tecnologia *Sidelink* da LTE-V2X e iniciou a especificação da nova inter-

face aérea (por exemplo, NR) na *Release* 15, tecnologia *Sidelink* para *Fifth Generation - Vehicle-to-Everything* (5G-V2X). Completamente especificada na *Release* 16, a *New Radio - Vehicle-to-Everything* (NR-V2X) foi projetada para atender à aplicações avançadas de V2X [Bazzi et al., 2021, Sehla et al., 2022].

Dentre estas aplicações pode-se destacar: o uso da direção cooperativa entre Veículos em Pelotão - (*Vehicle Platooning*); a Direção Avançada - (*Advanced Driving*), e a Direção Remota - (*Remote Driving*).

Na direção cooperativa entre **Veículos em Pelotão**, ou Veículos em Grupo, os veículos trafegam cooperativamente, proporcionando mais eficiência no trânsito a medida que reduzem a distância entre os outros veículos de seu pelotão e aumentam a velocidade de deslocamento, sem comprometer a segurança do trânsito. Isso só é possível devido a disseminação de informações extremamente rápida por parte do líder do pelotão. Para a troca de informações entre os veículos de um pelotão, por exemplo, é necessário uma latência de comunicação máxima de 25 milissegundos com uma confiabilidade de 90%, no nível mais baixo de automação dos veículos (por exemplo, o motorista controla toda a movimentação do veículo, mesmo com o suporte de sistemas de segurança), e uma latência fim-a-fim máxima de 10 milissegundos com uma confiabilidade de 99,99%, no nível mais alto de automação do veículo (por exemplo, o sistema de automação controla toda a movimentação do veículo, sem a supervisão humana, em qualquer ambiente).

A **Direção Avançada** habilita a condução de veículos semi-automáticos, ou completamente automáticos, através do compartilhamento de dados coletados de dispositivos de beira de estrada ou de outros veículos, o que possibilita a coordenação da trajetória ou de manobras (por exemplo, conversão à direita/esquerda, estacionamento). Ademais, neste caso de uso, os veículos compartilham com os outros integrantes da IoV manobras ou trajetórias que pretendem executar, aprimorando a prevenção de colisões, a eficiência do trânsito e proporcionando viagens mais seguras. Para aplicações de prevenção cooperativa de colisão, a latência máxima aceitável é de 10 milissegundos, com uma confiabilidade de 99,99%.

Na **Direção Remota**, um motorista ou uma aplicação V2X controla um veículo remotamente, principalmente em ambientes que representem perigo para humanos, e em casos em que as trajetórias possuem poucas variações (por exemplo, transporte publico) é habilitado o uso da condução baseada em nuvem. A Direção Remota requisita 5 milissegundos de latência máxima com 99.999% de confiabilidade e uma taxa de dados de 25 Mbps no *Uplink* e 1 Mbps no *Downlink* [Sehla et al., 2022].

A Tabela 4.3 denota os principais requisitos dos casos de uso habilitados pela IoV.

Tabela 4.3. Requisitos de QoS das comunicações C2

Caso de uso	Latência	Confiabilidade	Taxa de dados
Veículos em Pelotão	10-500 ms	90-99,99%	50-65Mbps
Direção avançada	3-100 ms	90-99,99%	10-50Mbps
Direção Remota	5 ms	90-99,99%	*Uplink* (UL):25Mbps
			Downlink (DL):1Mbps

4.3.2. Veículos Aéreos não Tripulados (*Unmanned Aerial Vehicle* – UaV)

O uso de **Veículos Aéreos Não Tripulados** - UAVs - (por exemplo,*drones*) vem ganhando destaque em vários setores da sociedade, uma vez que estes veículos podem realizar uma série de tarefas devido à sua versatilidade, mobilidade e baixa altitude de operação. Um UAV é um dispositivo voador inteligente que possui uma unidade de processamento, sensores que auxiliam a navegação e movimentação, uma fonte energética (por exemplo, bateria, combustível), dispositivos de comunicação e mecanismos de vôo. Ele é capaz de se movimentar cooperativamente em grupo ou isolado, se comunicando com as estações de controle para troca de dados e recepção e/ou sinalização de comandos. Os UAVs se dividem em duas categorias principais quanto ao mecanismo de vôo: de asa fixa ou de asa rotativa. [Fotouhi et al., 2019]. Alguns modelos de UAV de asa fixa podem atingir velocidades próximas a 452 km/h, altitudes de 15 km e suportarem uma carga de até 1700 kg, tornando-se propícios para o carregamento de mercadorias e pessoas ou para aplicações militares ou de resgate (por exemplo, reconhecimento de território, magnitude de desastres), com modelos menores e mais leves. O voo vertical e a aerodinâmica deste modelo permite que a aeronave "deslize"no ar, sendo energeticamente eficiente. No entanto, estes modelos necessitam de pista de voo e pouso, não são capazes de sobrevoar um local fixo e seu custo é mais elevado. Já os modelos de asa rotativa possuem mais versatilidade quanto a movimentação, uma vez que a decolagem e pouso são verticais. Alguns modelos desta categoria podem atingir de 50km/h a 100km/h e cerca de 3km de altitude. A grande vantagem desta categoria é a facilidade de manobra e a possibilidade de sobrevoar em pontos fixos e em baixa altitude.

A alta mobilidade e versatilidade permite o emprego dos UAVs em cenários como: entrega de conteúdo; monitoramento e vigilância; coleta de dados de sensores espalhados em uma região e processamento de dados, além de estender a cobertura das redes celulares [Wei et al., 2022, Geraci et al., 2022]. O amplo uso dos UAVs também aumenta o volume de dados transmitidos entre as estações de controle e os UAVs. Neste cenário, as redes 5G desempenham um papel crucial para manter a comunicação rápida, garantindo o comando e controle em tempo real dos veículos aéreos. O 3GPP identificou três cenários de comunicação de Comando e Controle - C2 *Communication*, sendo: Comunicação C2 Direta - *Direct* C2, quando o controlador e o UAV estabelecem um enlace direto utilizando os recursos de rádio e se registrando na rede 5G; Comunicação C2 Assistida pela Rede - *Network-Assisted* C2, quando o UAV e o controlador estabelecem uma comunicação *unicast* com a rede 5G, permitindo que o controlador e o UAV se conectem em diferentes antenas; e a Comunicação C2 Assistida pelo Sistema de Navegação de Tráfego – (SNT), quando o UAV possui um plano de voo pré determinado (por exemplo, para um voo autônomo), mas uma aplicação de controle acompanha em tempo real o *status* do voo e, quando necessário, realiza modificações na trajetória [Geraci et al., 2022]. A Figura 4.12 denota os cenários de C2 que utilizam as redes 5G como a rede de transporte.

Classifica-se, basicamente, três modos de controle de voo, e cada modo necessita de um requisito de QoS específico. Considere uma aplicação de reconhecimento de território. O UAV decola e o controlador envia uma mensagem de controle contendo os **pontos de passagem** que o UAV deve seguir para um reconhecimento inicial. Durante o voo, o UAV identifica uma região de interesse específica e solicita que o controlador realize o **controle direto da direção**, e envia imagens ao controlador que realiza as ma-

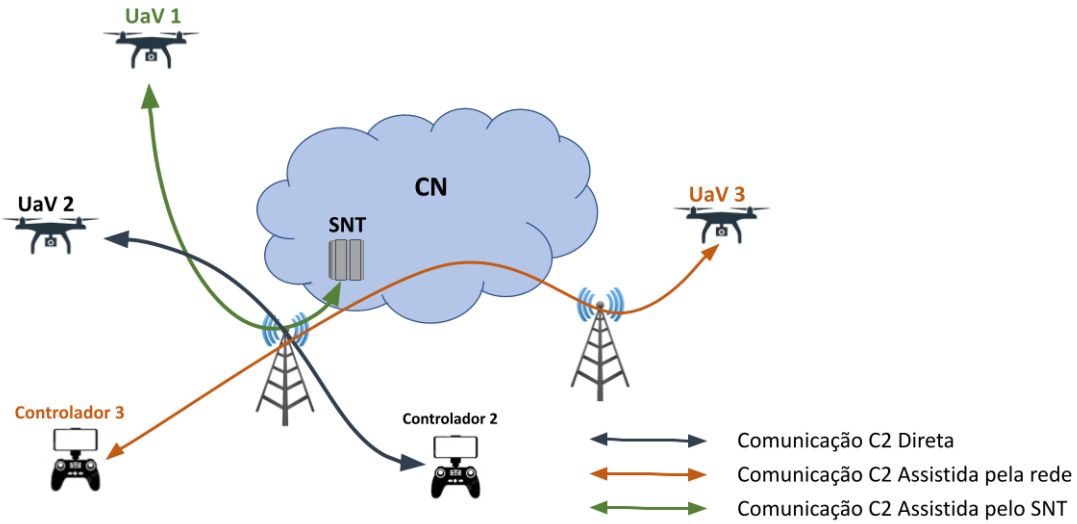

Figura 4.12. Comunicação de Controle e Comando. O UAV é diretamente controlado por um controlado autenticado na mesma RAN, por uma aplicação de voo autônomo ou por um controlador autenticado em uma RAN diferente. Adaptado de [ETSI, 2022].

nobras do UAV. Além disso, considere um **voo autônomo** do UAV supervisionado por uma aplicação de gerenciamento de tráfego de UAV. O modo de controle direto da direção é o mais desafiador, requerendo uma latência inferior a 40 ms, além de 99,9% de confiabilidade somente para controle do UAV. Para o controle direto da direção é mandatório o retorno de video, sendo requisitada uma taxa mínima de 2 Mbps no cenário com linha de visada direta - *Visual Line-of-Sight* (VLoS) e 4 Mbps além da visada direta, 1 s e 140 ms de latência e 99,9% e 99,9% de confiabilidade respectivamente [Geraci et al., 2022]. A Tabela 4.4 detalha os requisitos de QoS para as comunicações C2.

Tabela 4.4. Requisitos de QoS das comunicações C2.

C2	Latência	Velocidade do UAV	Confiabilidade
Pontos de Passagem	1s	300 km/h	99,9%
Controle Direto da Direção	40 ms	60km/h	99,9%
Voo Autônomo	1s	300km/h	99,9%

Além dos requisitos específicos para a comunicação de comando e controle, cada cenário de uso de um UAV possui seus próprios requisitos deQoS. Para transmissão ao vivo de video em 8K para uso em óculos de Realidade Virtual, a aplicação necessita de 100 Mbps de UL (para transmissão das imagens) e 600 kps de DL de taxa de bits e uma latência de UL e DL de 200 ms e 20 ms, respectivamente. Apesar de não apresentar requisitos tão relevantes de latência e taxa de bits, estas aplicação necessitam uma precisão de 0,5 m da posição do UAV [Geraci et al., 2022].

A Tabela 4.5 denota os requisitos dos principais casos de uso dos UAVs [Geraci et al., 2022]. Para além dos desafios relacionados ao grande volume de dados e localização precisa dos UAVs, a comunidade cientifica e setores da industria buscam solucionar os

desafios relacionados à eficiência energética e planejamento da trajetória dos veículos aéreos em cenários como os de coleta de dados de sensores [Luo et al., 2020], entrega de conteúdo [Zhao et al., 2021,Zhang et al., 2021] extensão da cobertura da rede celular [Lyu et al., 2022].

Tabela 4.5. Requisitos de QoS das aplicações UAV.

Aplicação	Taxa de Bit	Latência	Altitude	Precisão de posição
Transmissão de video 8K	UL: 100 Mbps	UL: 200 ms	<100 m	0,5 m
	DL: 600 Kbps	DL: 20 ms		
Mapeamento a Laser/ Patrulhamento em HD	UL: 120 Mbps	UL: 200 ms	30-300 m	0,5 m
	DL: 300 Kbps	DL: 20 ms		
Controle Remoto de UAV com video HD	UL: 25 Mbps	UL: 300 ms	<300 m	0,5 m
	DL: 300 Kbps	DL: 20 ms		

4.3.3. Aplicações de *eHealth*

As aplicações ***eHealth*** ganharam bastante destaque nos últimos anos, impulsionadas principalmente pela pandemia de COVID-19 (2019-2021), que remodelou as interações interpessoais e motivou a comunidade médica a buscar alternativas ao atendimento presencial [Santos et al., 2020]. Em cenários pandêmicos, o uso da tecnologia através de sensores inteligentes e das Redes de Sensores sem fio - *Wireless Sensor Network*s (WSNs) - para o acompanhamento remoto de sinais vitais e da recuperação do paciente é primordial para o controle da disseminação de doenças. Em especial, as Redes Corporais sem fio - *Wireless Body Area Network*s (WBANs) - são responsáveis pela coleta dos sinais vitais, emoções e hábitos de pacientes e pelo envio dos dados ao corpo clínico para acompanhamento e prescrição de tratamentos [Santos et al., 2020]. A Figura 4.13 denota a arquitetura básica destas redes. Minissensores com baixo consumo de energia, invasivos ou não, são colocados dentro, sobre ou em volta do corpo humano. Cada sensor é responsável por medir um sinal vital (por exemplo, frequência cardíaca, pressão arterial, etc) e enviar os sinais vitais ao nó coordenador (por exemplo, *smartphone, smartwatch*), através da comunicação Intra-BAN. Os nós coordenadores podem se conectar a outros dispositivos pessoais, a robôs domésticos ou a outras WBANs por meio da comunicação Inter-BAN. O grande volume de dados é enviado pela Internet para serem processados, na camada Além-da-BAN, para a equipe médica [Santos et al., 2020], que acompanha o estado de saúde do paciente e propõe tratamentos necessários. Com a proliferação das WBAN, o volume de dados a serem processados se torna imensurável e as redes 5G com o paradigma MEC podem ofertar recursos computacionais sob demanda e de larga escala [Alenoghena et al., 2022].

Além do monitoramento remoto de sinais vitais, vale destacar a Telemedicina como aplicação fundamental para os cuidados com a saúde. Consultas à distância foram adotadas durante a pandemia de COVID-19 para manter o acompanhamento médico de rotina, evitando visitas desnecessárias aos hospitais e mitigando a sobrecarga do sistema de saúde [Moglia et al., 2022]. Ademais, a telemedicina possibilita: as **cirurgias**

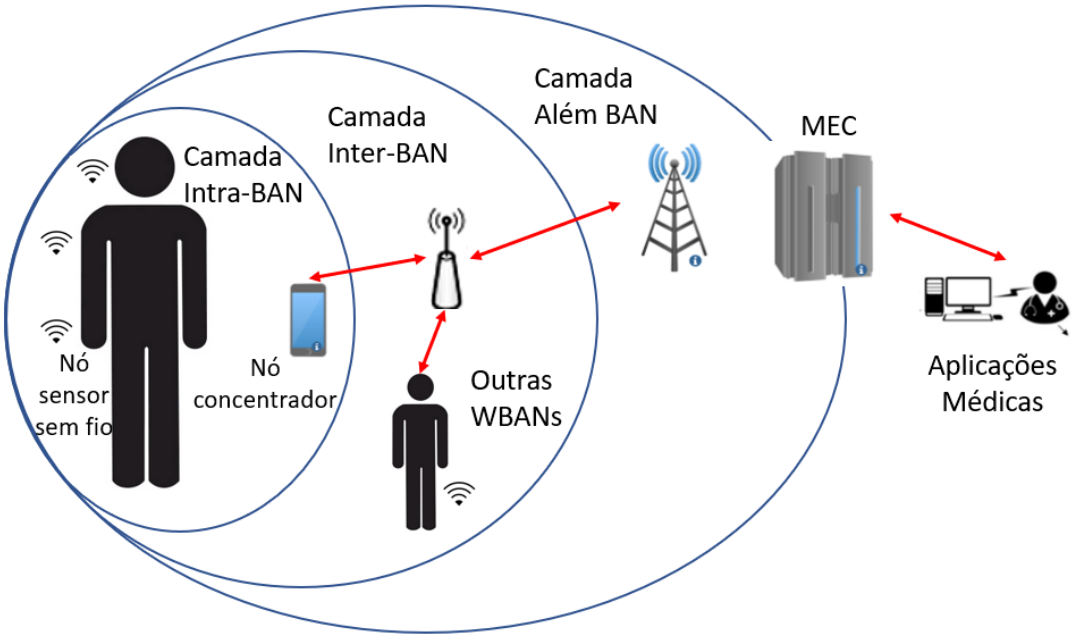

Figura 4.13. Arquitetura de Rede WBAN. Os dados coletados dos sensores são enviados para a MEC a fim de serem processados e encaminhados à equipe médica, que analisa os sinais vitais dos pacientes em tempo real. Adaptado de [Santos et al., 2020]

educativas remotas, quando residentes médicos acompanham em tempo real cirurgiões especialistas que realizam telecirugia, a realização de **exames remotamente**, o **suporte paramédico** durante o transporte de pacientes para o hospital, entre diversos outros. A Tabela 4.6 apresenta as métricas de serviço obtidas a partir do uso da Telemedicina durante a pandemia de COVID-19 [Moglia et al., 2022]. É possível observar que altas taxas de dados em *uplink* (UL) e *downlink* (DL) e baixas latências são necessárias para manter a segurança das atividades médicas. Em especial para realização de cirurgias, uma vez que é utilizado um sistema robótico mestre-escravo. Além disso, para a realização de exames, a taxa de DL é extremamente alta, sendo atendida apenas pelas redes 5G. A eficiência energética é outro desafio enfrentado pelas aplicações *eHealth*. No contexto de monitoramento de sinais vitais, a vida útil das baterias de sensores invasivos é um tema relevante estudado pela comunidade acadêmica [Saba et al., 2020].

4.3.4. Manufatura Inteligente (*Smart Manufacturing*)

A evolução dos setores industriais acompanha os ciclos de desenvolvimento das tecnologias. O surgimento das máquinas a vapor, no século XVIII, proporcionou a primeira geração industrial, que foi marcada pela mecanização de determinados processos. Na segunda geração, a linha de produção desenvolvida por Henry Ford, no início do século XX, e a massificação do uso da eletricidade aprimorou as técnicas de produção até então utilizadas. A terceira geração industrial foi marcada pelo surgimento da computação e automação, além da globalização dos mercados. A quarta geração, também chamada de Industria 4.0 ou Quarta Revolução Industrial, representa a evolução dos setores industriais, que passam a utilizar a conexão de diferentes máquinas, objetos e dispositivos para

Tabela 4.6. Métricas alcançadas com a Telemedicina

Aplicação	Latência	Velocidade
Cirurgia assistida por robô	264 ms	1 Gbs
Laparoscopia (cirurgia educativa)	146 ms- 202 ms	UL:98-101 Mb/s
		DL:98-101 Mb/s
Cirurgia cardíaca (telementoria)	30 ms	25 Mb/s
Ultrassom	23 - 30 ms	UL:130 Mb/s
		DL:930 Mb/s
Cirurgia a Laser Telerretiniana	20 ms	UL:88 Mb/s
		DL:854 Mb/s

facilitar a coleta de dados e automatizar a manufatura de um produto. Essa evolução permite um sistema de produção mais eficiente, aumentando a capacidade de customização dos produtos e reduzindo o ciclo de produção. A indústria 5.0 será a próxima revolução que permitirá uma customização em massa através da criatividade dos especialistas trabalhando em conjunto com máquinas inteligentes, eficientes e extremamente precisas [Mahmood et al., 2020] e [Maddikunta et al., 2022]. A Tabela 4.7 mostra as fases da evolução industrial desde o começo aos dias de hoje.

Tabela 4.7. Ciclos da Evolução Industrial. Adaptado de [Maddikunta et al., 2022]

Geração	Avanço	Ano
Indústria 1.0	Mecanização, vapor como energia	1784
Indústria 2.0	Produção em massa, linha de produção, eletrificação	1870
Indústria 3.0	Computação e automação	1969
Indústria 4.0	Sistemas físico-cibernéticos	2011
Indústria 5.0	Customização massiva e sistemas físico-cibernéticos cognitivos	–

A **Manufatura Inteligente** consiste em aplicações baseadas em *Cyber-Physical Manufacturing Systems* (CPMS) e no paradigma da IoT. O tema se tornou relevante por permitir a evolução na automação, no controle e na monitoração de equipamentos e processos industriais em tempo real [O'Connell et al., 2020]. O CPMS possibilita controle com precisão próxima a do tempo real a partir de qualquer local. Para tanto, utilizam-se as redes sensores sem fio, a computação na nuvem, a computação na borda e a computação em névoa. Nesse cenário industrial, há ainda a possibilidade de implantação da comunicação máquina para máquina – *Machine to Machine* (M2M) – e da manufatura colaborativa, que ocorre quando máquinas e humanos coexistem. Essas aplicações demandam alta confiabilidade, alta cobertura, baixa latência, dentre outras características que não estão disponíveis na rede 4G, mas se tornam acessíveis na rede 5G [Wu et al., 2021a].

O conceito de manufatura inteligente vai além da fábrica em si. É necessário considerar a logística inteligente (distribuição e transporte). Surge, portanto, o conceito de Internet das Coisas Industriais – *Industrial Internet of Things* (IIoT), que é uma rede

de dispositivos e sistemas industrias que compartilham informações entre si. Além dos requisitos específicos das aplicações, o ambiente industrial é desafiador para a rádio propagação do 5G. A presença de de grande número de maquinário com superfície metálica lisa gera múltiplas reflexões de sinal, assim como o considerável tamanho dificulta a propagação direta. O processo industrial e a presença de grande quantidade de motores gera muita interferência eletromagnética aleatória, alterando as características do meio sem fio e, consequentemente, o modelo de canal de propagação em comparação com ambientes de escritórios. Essas diferenças ambientais fazem necessário o estudo detalhado das características de propagação e do modelo de canal sem fio antes de ser desenhado um sistema confiável e eficiente. Em outubro de 2019, o 3GPP lançou uma atualização (V16.0.0) do relatório técnico TR 38.901 (*Study on channel model for frequencies from 0.5 to 100 GHz*, contendo o modelo e características do canal para IIoT [Jiang et al., 2021].

A revisão para inclusão das características do canal IIoT incluiu alguns itens como: descrição do cenário, modelo de perda de propagação do trajeto, probabilidade de comunicação com linha de visada, parâmetros para calibração, entre outros. As principais características de ocasionam essas mudanças em relação a um ambiente fechado de escritórios são o maior espaço físico de uma fábrica, alto teto, maior presença de objetos metálicos, objetos com tamanhos irregulares e os bloqueios de propação que são feitos por máquinas ao invés de paredes [Jiang et al., 2021].

A comunicação com latência extremamente baixa (menores que 10 ms) e a utilização de grandes larguras de banda proporcionadas pela utilização de tecnologias como uso de ondas milimétricas, fatiamento de rede, virtualização de funções de rede, redes definidas por *software* e computação na nuvem, névoa e borda são primordiais para o atendimento dos requisitos das novas aplicações industriais [Wu et al., 2022]. O grande volume de dados gerados pelos diversos dispositivos instalados e a necessidade de análise rápidas desses dados tornam as indústrias um ambiente que podem ser extremamente beneficiados pelas redes 5G [Mourtzis et al., 2021]. A Tabela 4.8 apresenta casos de tecnologias IIoT habilitadas pelo 5G.

4.3.5. Gêmeos Digitais (*Digital Twins*)

O conceito de gêmeos digitais não é recente. Ele foi criado há mais de meio século e está relacionado ao programa espacial americano, quando a Agência Espacial Norte Americana – *National Aeronautics and Space Administration* (NASA) – usou ideias iniciais de replicação digital durante os anos 1960 [Nguyen et al., 2021]. Nesse período, a NASA, proporcionalmente à tecnologia da época, utilizou simuladores de alta fidelidade para treinar os astronautas. O paradigma de gêmeos digitais atual está diretamente associado a necessidade, após uma explosão inesperada durante a missão Apollo 13 que causou um desvio da rota de aterrissagem, do controle da missão em terra simular, quase em tempo real, o comportamento da aeronave e a partir disso tomar decisões para proporcionar uma volta segura para os astronautas. Foram utilizados simuladores de voo disponíveis e que foram alimentados com os dados reais vindos da aeronave [Mihai et al., 2022].

A partir dessa ideia de simular o real com alta fidelidade, surgiu o conceito de gêmeo digital, que é um programa de computador que pega os dados e contextos do mundo

Tabela 4.8. Exemplos de aplicações industriais habilitadas pelo 5G, apontando os benefícios gerados e as características mais importante da rede para o pleno funcionamento da aplicação. Adaptado de [Mourtzis et al., 2021]

Aplicação	Caso de Uso	Benefícios	Característica 5G
Manutenção preditiva avançada	Uso de inúmeros sensores para fornecer uma representação em tempo real do estado de uma máquina, possibilitando a execução de manutenção preventiva e preditiva	Redução de tempo inoperante Redução de reposição de maquinário Redução no tempo de manutenção	Confiabilidade Custo dos dispositivos Densidade de dispositivos
Controle e monitoração precisos	Controle e monitoração em tempo real de robos e máquinas	Redução de defeitos Aumento da produtividade	Latência extremamente baixa Densidade de dispositivos
Realidade aumentada	Uso de visores de realidade aumentada para guiar o trabalhador local ou remoto.	Redução dos gastos com manutenção Redução do tempo de treinamento	Latência extremamente baixa Taxa de dados
Controle robótico remoto	Controle remoto de maquinário robótico	Melhoria de sergurança e saúde Aumento da produtividade	Latência extremamente baixa
Manufatura como serviço	Reduzir tempo de configuração permitindo que a fabricação seja flexível, permitindo o uso da planta por vários atores.	Aumento da taxa de inovação Inovação Redução dos custos Aumento da produtividade	Flexibilidade Custo dos dispositivos Onipresença
Veículo guiado automatizado	Controlar veículos autônomos de maneira mais flexível sem a necessidade de configuração de rotas pré-definidas	Aumento da eficiência e produtividade	Baixa latência Confiabilidade Conhecimento de localização
Inspeção por drone	Uso de drone para tarefas perigosas e difíceis para os humanos	Melhoria de sergurança e saúde Diminuição dos custos das inpeções	Baixa latência Confiabilidade Conhecimento de localização

real sobre um sistema e os processa de forma a reproduzir como o sistema ou processo real se comportaria diante de diferentes entradas. Essa tecnologia emergente tornou-se possível graças a possibilidade de conexão de um número massivo de sensores IoT. Apesar da criação de uma representação virtual do objeto, rede ou sistema real, essa representação virtual pode ou não estar conectada ao objeto real. Baseado nas informações que recebem em tempo real ou que são reproduzidas conforme captura real, é possível simular ações no ambiente virtual antes de fazê-las no objeto real e prevendo o comportamento correspondente. É possível ainda reproduzir situações críticas, para compreender melhor as falhas que geraram problemas e para buscar novas soluções funcionais para a arquitetura que já está em uso. Para que isso seja possível, considerando os casos nos quais o gêmeo digital conversa em tempo real com o gêmeo físico, é necessário uma rede extremamente confiável e robusta, com baixa latência e alta conectividade [Wu et al., 2021a].

Como exemplo de gêmeos digitais, pode-se citar o *The Spirent 5G DT*[2] que emula em software a réplica de uma rede 5G para teste de comportamento e performance de diversas aplicações. No contexto de emulação de redes 5G, há também a iniciativa da Huawei que, em 2020, lançou a primeira solução de engenharia para criar um *site* digital 5G que é uma réplica de um *site* físico [Huawei, 2020]. Na indústria, pode-se citar a iniciativa da Petrobras, que está utilizando gêmeos digitais para criar réplicas de suas instalações (refinarias de petróleo, plataformas, reservatórios, poços, sistemas navais e instalações submarinas) de forma a otimizar a operação, simular situações e reduzir custos. Só com

[2]Disponível em: https://www.spirent.com/assets/u/video-the-5g-network-digital-twin

a aplicações de réplicas de refinarias, a Petrobras conseguiu aumentar a rentabilidade em U$ 200 milhões no ano de 2020 [Petrobras, 2020].

As características fundamentais que permitem identificar um gêmeo digital verdadeiro conectado em tempo real, e não um simples modelo digital, são a autoadaptação, autorregulação, automonitoramento e autodiagnóstico. A autoadaptação permite uma reação automática do gêmeo digital a mudanças de ambiente e configuração do sistema real. A autorregulação não permite que o gêmeo digital assuma características que o gêmeo físico não pode atingir. O automonitoramento permite que o gêmeo digital sempre saiba o que está acontecendo com o gêmeo físico, através do monitoramento dos parâmetros relevantes. Por fim o autodiagnóstico permite que a versão digital saiba sobre a sua saúde baseado nos dados históricos e atuais quando não é possível manter o estado ótimo de operação [Wu et al., 2021a].

A perda de conectividade ou mesmo atraso da conexão pode representar um desvio do gêmeo digital em relação ao sistema real. A computação de borda no 5G contribui muito para a diminuição da latência possibilitando que as redes de 5ª geração surjam como candidata natural para habilitar a expansão da utilização de gêmeos digitais [Zhou et al., 2021].

A tecnologia de gêmeos digitais usufrui em maior ou menor escala de todas as características dos cenários de uso do 5G. O foco maior é no cenário de mMTC e uRLLC, por conta da necessidade de conexão de diversos dispositivos e inúmeros sensores dentro do ambiente industrial, além da necessidade de troca de informações quase que em tempo real (1 ms), sendo idêntico ao tempo de reação e interação entre homem e máquina, permitindo a utilização de soluções de automatização mais complexas com operação remota de equipamentos e máquina. A confiabilidade da conexão também é extremamente importante não podendo haver perda de conexão [Ramirez et al., 2022] e [Isto et al., 2020].

A Figura 4.14 mostra uma representação de uma rede de distribuição utilizando gêmeo digital em uma rede 5G. Pode-se observar a divisão em 3 camadas: dispositivos sensores, camada de borda onde há a agregação dos dados em servidores e a camada virtual do gêmeo digital que é alimentada com as informações da rede real.

4.3.6. Atendimento de áreas remotas ou de difícil acesso

Alguns locais, como plataformas de petróleo, centros de pesquisa avançados em locais remotos, ou mesmo áreas rurais ou de baixa densidade demográfica, usualmente carecem de uma infraestrutura de telecomunicações. A dificuldade de acesso e o alto custo de instalação e manutenção da infraestrutura impedem a disponibilização de redes de alta velocidade baseada em tecnologias como fibra ótica.

Devido a essas dificuldades, se torna necessário a busca de outras soluções para atender a esses cenários, que usualmente precisam da comunicação de forma confiável para diversas aplicações de saúde, sensoriamento e controle. Nesse contexto, as redes 5G surgem com suporte a adaptações capazes de atender a essas necessidades com um custo viavel.

As redes 5G para comunicação em áreas remotas preveem como principais cenários de uso os listados na Figura 4.15. Para o cenário brasileiro podemos destacar 4

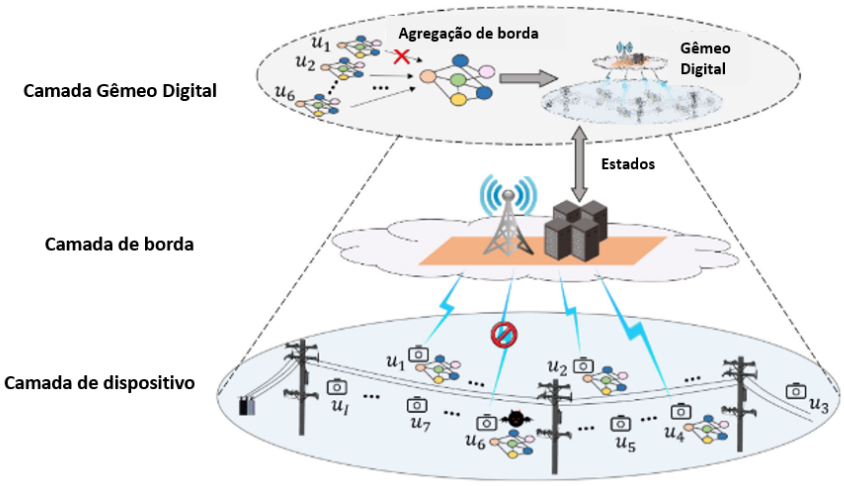

Figura 4.14. Aplicação de Gêmeo Digital com utilização de computação de borda, com a representação do gêmeo real, do gêmeo digital e da camada de borda entre eles. Adaptado de [Zhou et al., 2021].

casos de uso principais: voz e dados através da compra direta de dispositivos por usuários; *backhaul* provendo conexão para os sites 4G padrão; *e-Health* provendo acesso a esse tipo de serviço e *smart farming* [Cavalcante et al., 2021].

Figura 4.15. Principais Cenários de Uso das redes 5G em áreas remotas. Adaptado de [Mendes et al., 2020]

4.3.6.1. Atendimento a plataformas de petróleo e navios

As indústrias de energia, notadamente a de óleo e gás, são conhecidas por serem relativamente lentas na adoção em larga escala de novas soluções relacionadas a tecnologia da informação e telecomunicações. Por essa razão, a transformação digital ainda é um tema que possui ampla margem para crescimento nessas indústrias e que é extremamente relevante em cenários de transição para uma matriz energética mais sustentável.

Dados mostram que mesmo com a crise energética ocasionada pela guerra da Ucrânia, a demanda por combustíveis fósseis irá diminuir nos próximos anos. A participação desse tipo de fonte de energia na matriz energética mundial passará dos atuais 80% para 75% em 2030 e 60% em 2050. Essa queda na demanda e a necessidade na transição energética, impulsionam a busca pela redução dos custos de produção e adoção de novas tecnologias com o objetivo de tornar a indústria mais limpa. [International Energy Agency - IEA, 2022].Nesse contexto, a Figura 4.16 apresenta um sumário com os principais benefícios operacionais trazidos pela transformação digital na indústria de óleo e gás. Esse levantamento foi realizado através da classificação e levantamento dos principais artigos científicos relacionados ao tema [Maroufkhani et al., 2022].

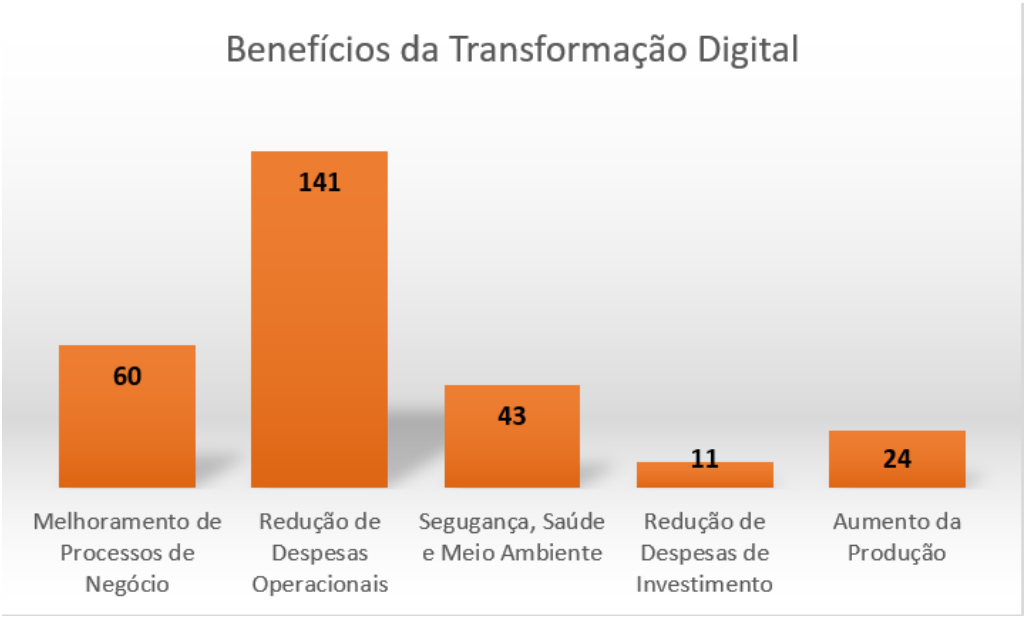

Figura 4.16. Principais benefícios operacionais proporcionados pela transformação digital na indústria de óleo e gás, compilados a partir da análise dos principais artigos científicos sobre o tema. O eixo Y apresenta a quantidade de artigos sobre cada tema apresentado no eixo X. Adaptado de [Maroufkhani et al., 2022]

Apesar de ter muitas características em comum com a manufatura inteligente, as plataformas de petróleo não são ambientes comuns. Elas estão localizadas em áreas remotas e inóspitas que na maioria dos casos não possuem infraestrutura de comunicação, havendo grande dependência de conexões por satélite. Por demandarem baixa latência, alta disponibilidade e grande quantidade de dispositivos conectados, o atendimento a plataformas de petróleo por meio de conexão por satélite representa um desafio para os cenários de uso eMBB, uRLLC e mMTC.

Diferentemente do ambiente terrestre, onde normalmente há grande disponibilidade e a infraestrutura de comunicação é relativamente fácil de ser implantada, o ambiente marítimo proporciona uma série de desafios e problemas para uma comunicação precisa e de qualidade. Pode-se citar desafios relacionados a eficiência de comunicação (taxa, latência, confiabilidade), integração multidisciplinar entre tecnologias e a deterioração aos equipamentos causadas pelo ambiente marítimo. Além disso, plataformas são metálicas o que prejudica a adoção de soluções por radio–propagação.

Em plataformas há demandas para aplicações de automação, IoT, operação remota, robótica, realidade aumentada, telemedicina, telepresença e gêmeos digitais. Atender a todos os requisitos dessas aplicações (latência, confiabilidade de conexão, conexões simultâneas, eficiência energética) em um ambiente hostil e remoto é extremamente difícil. Há sérias limitações de espaço físico em plataformas que dificultam a instalação de soluções diferentes e concorrentes por espaço e recursos. Essas plataformas podem operam a distâncias de mais de 300 km da costa o que gera um grande desafio para sua conexão com toda a infraestrutura de tecnologia da informação e telecomunicações em terra. A conexão com a infraestrutura terrestre pode ser feita basicamente das seguintes formas:

1. Através de enlaces de rádio a partir da costa (mais próximas à costa);
2. Através de malha óptica;
3. Através de enlaces de rádio com outras plataformas concentradoras;
4. Através de satélites de média órbita e geoestacionários;
5. Através de conexão com enlace *backhaul* 4G com outras plataformas concentradoras.

Nesse cenário, as redes 5G surgem como grande candidatas para suprirem as necessidades desse ambiente, facilitando a integração entre as diversas aplicações e provendo os requisitos necessários para o pleno funcionamento de todas. Uma visão básica da operação de uma plataforma e alguns dos desafios desse ambiente, que já começam na forma de embarque, podem ser vistos no vídeo[3], produzido através de uma parceria entre a Petrobras e o canal Manual do Mundo.

4.3.7. Atendimento a zonas rurais

Nos dias atuais há uma demanda reprimida por conectividade sem fio em áreas remotas e rurais. O atendimento a **Zonas Rurais ou Plataformas de Pesquisa Remotas** é desafiador porque essas áreas normalmente não possuem infraestrutura de comunicação. A rede 4G, desenvolvida para atender principalmente os requisitos e necessidades do ambiente urbano, e a cobertura padrão viabilizada pelas células 5G padrão, utilizadas em ambientes urbanos, são ineficazes para atendimento as áreas rurais e remotas. Algumas tecnologias habilitadoras adotadas nas redes 5G, por exemplo o uso de frequências mais altas e MIMO, podem ocasionar uma cobertura limitada das células 5G. Em áreas urbanas isso não é um problema, mas em áreas remotas não é alcançado um número suficiente de assinantes de forma a viabilizar um custo aceitável por assinante atendido. Soma-se a isso o investimento necessário em *CAPEX* para aquisição de equipamentos, licenças, aquisição de espectro de radiofrequência e infraestrutura (torres, energia, *backhaul*) [Mendes et al., 2020].

As comunicações melhoradas para áreas remotas – *enhanced Remote Area Communications* (eRAC) – proporcionam o uso de diversas aplicações como agricultura inteligente, monitoramento de desastres, entre outras. Também possibilitam a conexão de uma

[3]Disponível em: https://www.youtube.com/watch?v=vOuuZJ5d4Ks

parcela da população que hoje não está conectada e não desfrutando e não fazendo parte da era da informação. Utilizando-se projetos de células 5G com o objetivo de alcançar maiores raios de cobertura (p.ex empregando-se frequências mais baixas de transmissão e numerologia adequada para o OFDM, células 5G podem ter raios de até 50 km com entrega de uma taxa de dados de 100 Mbps na borda, permitindo a entrega efetiva de conectividade em áreas remotas. Áreas rurais são pouco povoadas e Um raio de 50 km é necessário para que haja assinantes suficientes de forma a tornar a rede sustentável economicamente. Há necessidade de atendimento a aplicações de banda larga, agricultura inteligente e *backhaul* sem fio, explicam a necessidade de uma taxa de 100 Mbps. A comunicação deve ser feita prioritariamente em faixa de frequência abaixo de 1 GHz é necessário para que seja possível o aumento do raio de cobertura da célula. Um resumo dos requisitos para as redes dessas áreas podem ser vistos na Tabela 4.9 [de Souza Lopes et al., 2020].

Tabela 4.9. Principais Requisitos para Redes em Áreas Rurais. Adaptado de [Mendes et al., 2020]

Requisito	Valor
Taxa de dados	100 Mbps
Cobertura	50 km
Faixa de frequência	Sub 1 GHz e espectro não licenciado
Alocação do espectro	Fragmentada e secundária

Além da utilização de redes não terrestres, utilizando soluções com UAVs e satélites, há outras iniciativas para solução do problema de oferta de conectividade em áreas rurais. No Reino Unido o projeto *5G Rural First* busca, através de mudanças na regulamentação, permitir que outras partes interessadas utilizem o espectro eletromagnético não utilizado pelas operadoras de determinada região [Mendes et al., 2020].

Uma solução que merece destaque é a – *Remote Area Access Network for the 5th Generation* (5G-RANGE) –, um projeto de cooperação bilateral entre o Brasil e a Europa com objetivo de disponibilizar uma rede móvel projetada para prover comunicação confiável e com bom custo benefício para as áreas rurais e remotas. O 5G-RANGE consiste no uso de canais de TV não utilizados – *Tv White Space* (TVWS)– como rede secundária. A utilização de frequências na faixa de VHF e UHF somado ao uso de rádios cognitivos[4] possibilita a formação de células 5G maiores, possibilitando até 100 Mbps a uma distância de 50 Km da estação base [Mendes et al., 2020].

4.3.7.1. Aplicações críticas em áreas remotas e de difícil acesso

Ainda dentro do contexto do atendimento de telecomunicações em áreas remotas, tem-se cenários de grande relevância, como as estações de pesquisa na Antártica, ou

[4]Rádios cognitivos são capazes de fazer o sensoreamento do espectro na localidade em que são empregados e usar "brechas", momentâneas ou não, da faixa espectral para sua transmissão, evitando interferência nas demais transmissões [Wang et al., 2022].

as estações militares em florestas ou desertos. Dentro desses contextos, as aplicações críticas demandadas vão, usualmente, de telessaúde para as equipes ao sensoriamento de áreas de interesse [Mallorquí e Zaballos, 2021]. O uso de comunicações por satélite torna-se a única opção disponível, muito embora esse tipo de canal não seja facilmente compartilhado. Nesse sentido, o uso do 5G sobre enlaces de satélite pode se mostrar de grande valia.

As redes NTN surgem como principais candidatas para atendimento a essas áreas. Especificamente, satélites geoestacionários são amplamente usados em para comunicação, sendo uma boa possibilidade em termos de meio físico para comunicações militares em florestas e desertos. Contudo, esse tipo de satélite apenas provê cobertura de uma área nas zonas tropicais e equatoriais do planeta, além de gerar grandes atrasos na comunicação, devido à distância entre a estação base e o satélite. Em latitudes maiores, próximo dos polos, apenas satélites de baixa ou média órbita conseguem prover serviços [Ceriotti et al., 2012]. Considerando o requisito de baixos atrasos e a cobertura dos satélites, vários trabalhos discutem a integração entre satélites de baixa órbita – *Low Earth Orbit* (LEO) *Satellite Network* (SatNet) e o 5G [Darwish et al., 2022, Azari et al., 2022]. Esse tipo de aplicação do 5G pode revolucionar a comunicação em áreas remotas, permitindo maior flexibilidade nas aplicações e maior segurança nas operações.

4.3.8. Sensoriamento e controle em Redes Elétricas Inteligentes (*Smart Grids*)

Os sistemas de potência vem se modernizando nos últimos anos. As redes elétricas inteligentes transformaram sistemas eletromecânicos em uma Rede Físico-Cibernética de Sistemas de Potência - *Cyber-Physical Power System* (CPPS) [Xu et al., 2020a]. As redes elétricas inteligentes implementam sensores e tecnologias de controle e comunicação nos sistemas de potência para coordenar e gerenciar as atividades do setor elétrico [Esenogho et al., 2022]. Diferente dos sistemas tradicionais, onde a comunicação cobria somente os centros de controle das subestações de energia, nas redes elétricas inteligentes a comunicação bidirecional abrange desde a geração de energia até o consumidor, aumentando demasiadamente o fluxo de dados gerados pelas redes elétricas.

Uma rede elétrica inteligente pode ser descrita por sete domínios lógicos que se comunicam entre si, sendo os domínios: de Operação, do Mercado de Energia, de Geração, de Transmissão, de Distribuição, dos Consumidores e dos Provedores de Serviço. A Figura 4.17 demonstra os domínios de uma rede elétrica inteligente e o fluxo de comunicação e energia entre eles. O domínio de Operação se comunica e controla todos os domínios do sistema elétrico. O domínio do Mercado de Energia também se comunica com todos os domínios do sistema, uma vez que é responsável por manter a harmonia entre a compra e venda de energia. Os domínios de Geração, Transmissão e Distribuição fazem parte do fluxo de energia até o consumidor. O Domínio de Geração se comunica com o de Transmissão, que, por sua vez, se comunica com o de Distribuição. Além disso, o domínio de Distribuição se comunica com o domínio dos Consumidores. No domínio dos Consumidores, se encontram o consumo energético e eventualmente a geração de energia em escala reduzida. Este domínio se comunica com os Provedores de Serviço, que oferecem os serviços de faturamento e de operações de resposta às demandas [Lopes et al., 2016].

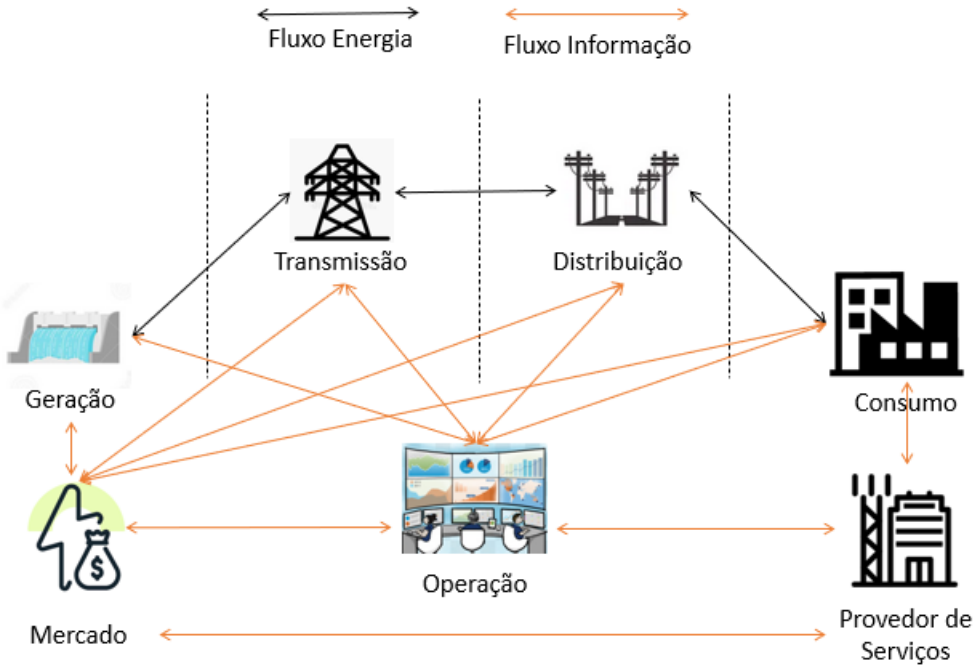

Figura 4.17. Domínios Lógicos das redes elétricas inteligentes. Sete domínios lógicos descrevem as *Smart Grid*. Diferente dos sistemas tradicionais, nas *Smart Grid*, a comunicação bidirecional abrange desde a geração de energia até o consumidor, aumentando demasiadamente o fluxo de dados. Adaptado de [Lopes et al., 2016].

Nesse contexto, o uso massivo de medidores inteligentes permite a leitura automatizada dos medidores, também conhecida como Leitura Automatizada de Medição - *Autommated Meter Reading* (AMR). Ela proporciona o encaminhamento dos dados obtidos dos medidores nos domínios de operação (por exemplo, geração, transmissão, distribuição, consumo, provisão de serviços) para os domínios de controle, como o centro de operação e o mercado de energia [Esenogho et al., 2022].

Em cada domínio das redes elétricas inteligentes, pode-se destacar diversos avanços em relação às redes elétricas tradicionais. No domínio da Geração de Energia, o principal avanço diz respeito à geração distribuída, que permite a produção de energia elétrica próxima aos locais de consumo energético, fomentando o uso de energia limpa (por exemplo, solar, eólica) e diminuindo as perdas em linhas de transmissão. Neste domínio, o paradigma de *microgrids* diz respeito a microgeradores de energia autossuficientes, localizados geralmente em uma região ou distrito, que se separam dos sistemas de energia em cenários de desastres naturais e são capazes de atender aos consumidores daquela microrregião, permitindo um sistema elétrico mais resiliente a falhas. Para suportar grandes consumidores ou picos no consumo de energia, as redes elétricas inteligentes podem agregar diferentes fontes da geração distribuída em uma Planta Virtual de Potência - *Virtual Power Plant* (VPP). Uma VPP trata as diversas fontes de geração como uma única entidade, sendo controlada por um Sistema de Gerenciamento de Energia que aproveita o controle descentralizado para verificar a demanda de energia nos microgeradores determinar se permanecerão em funcionamento ou se ficarão em *standby* [Lopes et al.,

2016, Esenogho et al., 2022].

Além disso, nos domínios de Geração, de Transmissão e de Distribuição, as redes elétricas inteligentes modificaram a natureza das subestações, que estão passando a ser digitalizadas. Essa digitalização implica na substituição de relés eletromecânicos por dispositivos eletrônicos inteligentes (*Intelligent Electronic Device*s (IEDs)) e também na inserção de redes Ethernet dentro do contexto de subestações para comunicação entre IEDs e entre IED e sistema *Supervisory Control and Data Acquisition* (SCADA). Tal comunicação implica em requisitos fortes de baixos atrasos, alta confiabilidade e alta segurança do tráfego. As subestações de geração e transmissão, muitas vezes, se encontram em locais de difícil acesso, sendo, em alguns casos, não assistidas, ou seja, operam sem supervisão *in loco* humana. A dificuldade de acesso a essas subestações e a necessidade de constante supervisão com requisitos estritos na comunicação enseja o estudo de novas tecnologias, tais como as redes 5G, para atender às demandas dessa nova geração das redes elétricas [Carrillo et al., 2022, Raussi et al., 2023, Iurii et al., 2022, Adrah et al., 2022].

Ainda no Domínio de Distribuição, a automação dos sistemas de distribuição integra o gerenciamento de dados, inteligência artificial, atuadores e sensores inteligentes para aumentar a confiabilidade e qualidade da energia entregue ao consumidor e diminuir a despesas operacionais. Para que toda essa comunicação funcione, incluindo a geração distribuída com suporte à VPP, é necessária uma infraestrutura de medição avançada (*Advanced Metering Infrastructure* (AMI)), capaz de interconectar milhões de dispositivos, entre eles medidores inteligentes e sensores das redes de distribuição [Lopes et al., 2016]. Com o crescimento da medição inteligente e a integração dos medidores inteligentes com centros de controle e casas inteligentes, torna-se necessária uma rede de comunicação sem fio capaz de lidar com um altíssimo número de dispositivos a baixo custo. As redes 5G surgem como uma forma técnica e economicamente viável de atender a essa demanda.

Com a evolução da rede elétrica, os Domínios dos Consumidores e dos Provedores de Serviços se tornam mais integrados, já que medidores inteligentes são instalados nos pontos de consumo de energia para fornecer acesso em tempo real do consumo de energia, além de permitir diferentes tarifações ao longo do dia (por exemplo, tarifação mais cara em horários de pico) [Lopes et al., 2016, Esenogho et al., 2022]. Ademais, a integração entre os dois domínios permite o desligamento de energia em casos de descumprimento das politicas de pagamento. O aumento na complexidade da alocação dos recursos de energia, através das VPPs, controle e gerenciamento das linhas de transmissão e do balanceamento de carga, leitura dos dados de medição ao longo dos domínios da rede e todas as outras vantagens do uso das redes elétricas inteligentes incentivam o uso de algoritmos de Inteligência Artificial e de Aprendizado de Máquina nos Centros de Controle. Especificamente, o uso das SDNs e NFV são os componentes chave na otimização do controle e gerenciamento dos domínios das redes elétricas inteligentes.

Fica evidente que os pilares das redes elétricas inteligentes são a automação inteligente e a comunicação entre os domínios da rede. Cada domínio implementa cenários de uso que possuem requisitos de QoS heterogêneos. A Tabela 4.10 denota os diferentes cenários de uso implementados nos domínios das redes elétricas inteligentes e seus requisitos de comunicação, além das classes de serviços das redes 5G que atendem aos

requisitos destes cenários de uso [Esenogho et al., 2022]. O uso de SDNs, base sob a qual as redes 5G foram projetadas, permite o isolamento, gerenciamento e atendimento dos requisitos heterogêneos destes cenários de uso. A Automação Distribuída com Sensores Inteligentes demanda uma alta confiabilidade da rede de comunicação e uma média densificação de dispositivos, sendo atendido pelas fatias de rede de uRLLC e mMTC, enquanto a aquisição de dados dos sistemas de distribuição, através de medidores inteligentes, demanda baixa latência de comunicação e alta quantidade de dispositivos espalhados na rede, proporcionadas por fatias de mMTC.

Tabela 4.10. Requisitos de QoS dos cenários de uso das redes elétricas inteligentes. Adaptado de [Esenogho et al., 2022]

Cenário de uso	Latência	Confiabilidade	Largura de Banda	Densidade de dispositivos	Tipo de fatia
Automação das subestações	Alta criticidade	Alta	Baixa	Média	uRLLC
Automação Distribuída	Alta criticidade	Alta	Média	Média	uRLLC mMTC
Controle de carga com alta precisão	Alta criticidade	Alta	Média Baixa	Média	uRLLC
Coleta de dados dos sistemas de distribuição	Baixa criticidade	Média	Média	Alta	mMTC
microgrids, VPPs	Média/ Alta criticidade	Alta	Baixa	Alta	uRLLC mMTC

4.4. Discussão, tendências, desafios e projetos pesquisa

Esta seção apresenta as tendências tecnológicas que complementam as tecnologias habilitadoras e o funcionamento das redes 5G de forma a torná-las a solução ideal atualmente para atendimento às aplicações críticas apresentadas na seção anterior. Também são apresentados alguns projetos de pesquisa que envolvem o desenvolvimento e melhoria das aplicações críticas. Por fim, é apresentado o cenário de simulação para a parte prática deste minicurso.

4.4.1. NTN

As **redes não terrestres** – NTN – desempenham papel fundamental na ubiquidade do 5G. As redes não terrestres são redes onde veículos espaciais (Satélites) ou aéreos (plataformas de altas altitudes – *High-altitude platform station* (HAPS) – ou veículos aéreos não tripulados UAV) atuam como nó de retransmissão ou uma estação base. A Tabela 4.11 apresenta os principais tipos de plataformas de comunicação das redes NTN.

As NTN complementam as redes terrestres, permitindo a oferta de serviço em áreas e localidades não atendidas pela infraestrutura de comunicação terrestre, aumentando a confiabilidade do 5G e garantindo comunicação contínua e viabilizando a escala-

Tabela 4.11. Principais Tipos de Plataformas de Comunicação NTN

Plataforma	Altitude	Órbita	Distância Típica de transmissão
Satélite LEO	300 – 1.500 km	Circular com velocidade de giro maior que a da Terra	100 – 1.000 km
Satélite GEO	35.786 km	Elevação/azimute fixos em relação a um ponto da terra	200 – 3.500 km
Sistema Aéreo não Tripulado	8 – 50 km	Elevação/azimute fixos em relação a um ponto da terra	5 – 200 km

bilidade de serviços, como por exemplo, *broadcasting*. Várias entidades, como *European Space Agency* (ESA), 3GPP, *Satellite and Terrestrial Network for 5G* (SaT5G), ITU-R, possuem iniciativas para desenvolver e integrar as redes NTN ao 5G. O 3GGP, através das recomendações técnicas 38.811 e 38.821, definiu quais devem ser atendidas, e cenários de uso que tornam possível a integração de redes NTN com a nova interface área do 5G - NR [3GPP, 2020a, 3GPP, 2021].

No *Release* 15 o 3GPP aprovou um item de estudo sobre NTN com o objetivo de identificar os principais desafios e propor as principais soluções para a integração das redes NTN ao 5G, ao contrátio da visão anterior onde sistemas de comunicação satélites eram soluções autônomas. Como resultado, foram publicados três relatórios técnicos no *Release* 16: TR 22.822, e os já mencionados anteriormente TR 38.811 e TR 38.821, cujos objetivos são [Hosseinian et al., 2021]:

1. TR 22.822 - "*Study on using Satellite Accessin 5G*": lista os casos de uso considerando a integração do satélite às redes 5G [3GPP, 2020b].
2. TR 38.811 - "*Study on New Radio (NR) to support Non-Terrestrial Networks*": visa adaptar o modelo de canal aéreo do 3GPP para a NTN [3GPP, 2020a].
3. TR 38.821 - "*Solutions for NR to supportNon-Terrestrial Networks (NTN)*": detalha uma série de adaptações que permitem os protocolos do NR funcionem na NTN [3GPP, 2021].

A grande maioria das pesquisas atuais, no que se refere as redes NTN, visa desenvolver a arquitetura das redes não terrestres baseadas em satélites. A razão para isso é que o foco, até o momento, de trabalho do 3GPP vem sendo a comunicação por satélite, com compatibilidade implícita para futuras integrações com veículos aéreos de comunicação [Lin et al., 2021]. O 3GPP definiu algumas opções de arquitetura baseada no papel que os satélites podem desempenhar em uma rede NTN. Elas são definidas a partir do tipo de tráfego do satélite e do tipo de *link* de acesso do usuário. O satélite pode operar no modo transparente, onde funciona apenas como retransmissor, e no modo regenerativo, onde possui a função de uma gNB embarcada. Quanto ao acesso do usuário, ele pode

ser direto ou através de um nó de retransmissão terrestre. A Figura 4.18 apresenta essas arquiteturas.

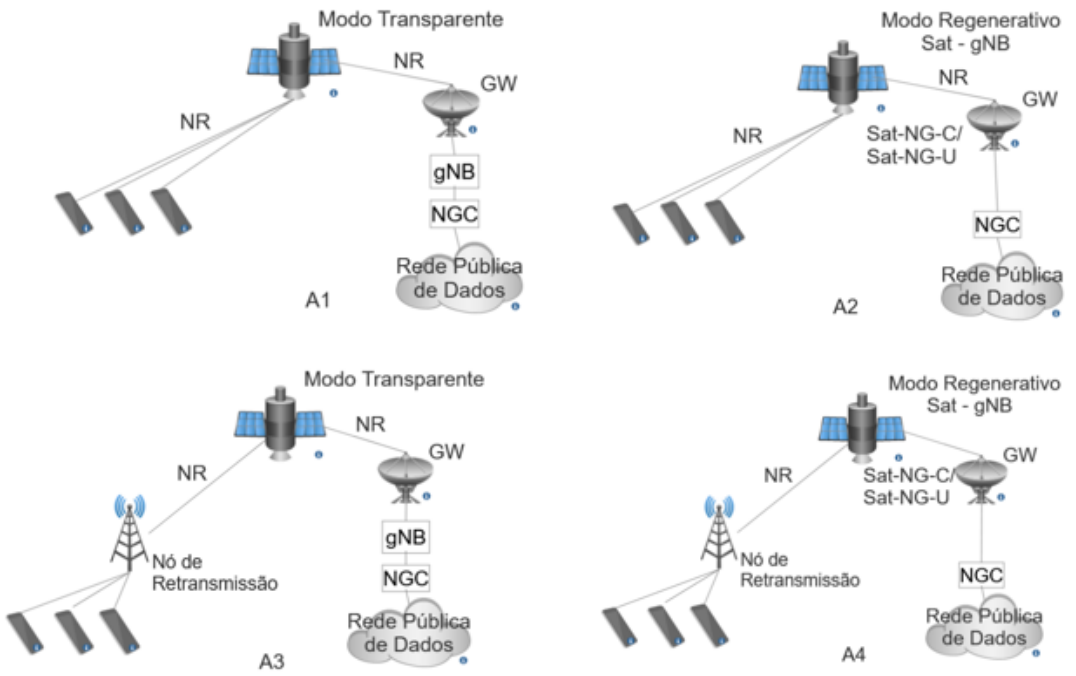

Figura 4.18. Arquitetura NTN baseada em satélites. Pode-se observar quatro cenários básicos que variam de acordo com o modo de atuação do satélite (transparente e regenerativo) e a presença ou não de um nó de retransmissão terrestre. Adaptado de [Guidotti et al., 2019].

No primeiro cenário, os dispositivos de usuário se comunicam diretamente com os satélites através da interface aérea da tecnologia 5G. Esse cenário é bastante desafiador pois submete uma interface que foi desenhada para sistemas terrestres às características peculiares do canal de comunicação por satélite. Nesse cenário, a gNB é conceitualmente localizada no *gateway* da rede satélite, provendo conexão ao *core* da rede 5G e a rede pública de dados. No cenário 2, o processamento regenerativo do sinal gera a forma de onda do 5G NR no satélite enquanto o *gateway* provê conectividade com o *core* da rede 5G e a rede pública de dados. Essa arquitetura é mais complexa e mais custosa, mas reduz os atrasos dos procedimentos das camadas físicas e MAC da NR. Nos cenários 3 e 4, a conexão entre os dispositivos de usuário é feita com nós de retransmissão terrestres que utilizam os satélites como link de *backhaul* [Guidotti et al., 2019].

Todos os 3 cenários de uso previstos para o 5G (eMBB, mMTC e uRLLC) devem ser suportados pelas redes não terrestres. Nesse contexto, os satélites geostacionários *Geostationary Earth Orbiting* (GEO) de alta capacidade *High Throughput Satellite* (HTS) e de baixa órbita *Low Earth Orbiting* (LEO) surgem como destaque. Os satélites GEO HTS com suas altas taxas de dados são ideais para serviços eMBB e as megaconstelações de satélites LEO proporcionam o baixo atraso de propagação, 14ms em um sentido e 50ms no percurso de ida e volta, necessário para atender aos requisitos das aplicações uRLLC não tão sensíveis [Rinaldi et al., 2020].

Apesar das boas perspectivas, a integração das NTN com o 5G traz uma série de

desafios. Questões relacionadas a não-linearidade do canal, atrasos de propagação gerando impactos direto nos mecanismos de acesso inicial ao meio (acesso aleatório, controle de *loop*, temporizadores, *Hybrid Automatic Repeat Request* (HARQ)) e a mobilidade do satélite (nos satélites LEO, a alta velocidade gera muitas mudanças e variação das taxas do efeito Doppler) são significativas e de difícil solução [Hosseinian et al., 2021].

As pesquisas em redes de satélites estão em franca ascensão. Um campo que surge com potencial de tornar essas redes mais acessíveis e universais é o de satélites menores e também de satélites miniaturizados ou *CubeSats*, em órbitas LEO. Exemplos de investimento para a formação de megaconstelações de satélites pequenos, proporcionando altas taxas de dados e baixa latência, não faltam. Estas redes, cujos principais exemplos são a Starlink, Oneweb e projeto Kuiper (Amazon), possibilitam conexão em áreas remotas e desassistidas por infraestrutura de conexão terrestre. Em relação a cubesats, que se constituem em satélites miniaturizados, em formato cúbico e com massa extremamente reduzida, podem também servir para formar redes provendo conexão, ainda que com menor capacidade, em áreas inóspitas. Um exemplo é o projeto KIPP da empresa *KEPLER Communication*[5]. Embora estas iniciativas ainda não estejam integradas a redes 5G, é possível dizer que, no futuro, há alta probabilidade de que suas infraestruturas sejam aproveitadas para esta finalidade [Saeed et al., 2020].

4.4.2. Desafios relacionados às tecnologias habilitadoras

Questões técnicas relacionadas a privacidade e segurança são importantes quando observamos o MEC. Aspectos de acesso, autenticação, criptografia e como prover acesso a apenas dispositivos confiáveis são pontos importantes e que deve ser devidamente endereçados. Uma falha de segurança pode permitir ataques a toda infraestrutura da NFV podendo ocasionar a sua manipulação ou desligamento.

Somam-se ao caso anterior aspectos relacionados ao uso de inteligência artificial que precisam ser melhor estudados para que seja possível extrair todo o potencial de forma a incrementar a eficiência de sistemas físico cibernéticos.

O consumo de energia também é preocupante para a vasta implementação de computação na borda e poucos trabalhos abordam a redução de consumo em *data centers* que são reconhecidamente grandes consumidores de energia [Spinelli e Mancuso, 2020]. Alguns trabalhos que buscam instanciar VNFs modelam os recursos da rede e as VNFs como grafos direcionados, a fim de implementar seus algoritmos de otimização [Xie et al., 2021, Quang et al., 2019, Zheng et al., 2018]. Especificamente, Xie *et al.* propõem a utilização da Aprendizagem por Reforço Profundo para a tomada de decisão ao instanciar as VNFs ao longo da rede, através do algoritmo proposto Kolin. Eles ainda utilizam uma Rede Neural em Grafos para explorar a estrutura das NFVs, que são representadas como Grafos Direcionados Acíclicos. Já Quang *et al.* modelam o problema de instanciação de VNFs como um Processo de Decisão de Markov, onde o espaço de estados S representa os requisitos computacionais de cada serviço e de cada enlace virtual entre funções de rede, o espaço de ações A que representa a prioridade em se instanciar uma função de rede em determinado nó e os pesos dos enlaces usados no encadeamento das funções. A taxa de aceitação é considerada como a recompensa r. Quang *et al.* utilizam o Aprendizado por

[5]Mais informações em: https://kepler.space/network/

Reforço Profundo para resolver o problema de instanciação.

Li *et al.*apresentam uma nova arquitetura de *Edge Computing* para IoT com módulos responsáveis por identificar o local que o serviço está sendo requisitado, realizar o processamento ou, quando necessário, a migração do processamento das requisições, gerenciar e otimizar o desempenho de uma borda especifica e gerenciar a mobilidade dos dispositivos. Li *et al.* ainda propõem um novo esquema de acesso e controle dos dispositivos [Li et al., 2018]. Já Song *et al.* apresentam um forma de alocação de recursos de rede em um ambiente de computação de borda baseada no contexto do usuário das aplicações, ou seja, baseado localização e requisitos das aplicações. Eles dividem os usuários em grupos de acordo com suas velocidades e alocam os recursos de rede para cada grupo, utilizando um algoritmo de particionamento de grafos [Song et al., 2019]. Xiong *et al.* utilizam técnicas de aprendizado por reforço para resolverem o problema de alocação de recursos em redes de computação de borda. Eles modificam uma rede Q Profunda (*Deep Q-Network*) para otimizar a aprendizagem e modelam o problema de alocação de recursos como um Processo de Decisão de *Markov*. Além disso, eles adicionam uma camada no final da Q-network para filtrar os valores de ação-estado inválidos [Xiong et al., 2020].

Liu *et al.* modelam o problema de mapeamento e roteamento dinâmico de VNF sensível a atraso, perda de pacotes e *jitter* como um problema de programação linear inteira e propõem uma heurística baseada em um método de relaxamento Lagrangeano melhorado para resolver o problema [Liu et al., 2021]. Já Wang *et al.* modelam o problema de posicionamento das encadeamentos de Funções de Rede, que visa a otimização na utilização dos recursos, como um problema de correspondência em Grafos Ponderados (*graph matching*). Eles modelam as requisições de SFC e a infraestrutura de rede como um grafo ponderado reformulando o problema de SFC como um problema de correspondência de grafos [Wang et al., 2019]. Xu et al. [Xu et al., 2020b] estudam o problema de encadeamento de funções e o mapeamento de VNFs sensível a QoS em redes IoT multicamadas, e formulam o problema de maximização na taxa de transferência como um problema de programação linear inteira. Wu *et al.* estudam o problema de mapeamento de VNFs considerando que as VNF de um serviço requisitado podem ser executadas em paralelo e representadas por grafos direcionados. Eles modelam o problema como um problema de Programação Linear Inteira com uma abordagem que leva em consideração a probabilidade de um mapeamento contribuir com a recompensa positiva [Wu et al., 2021b].

Diferente de trabalhos que modelam as VNFs como um grafo direcionado, os autores em [Agarwal et al., 2019] modelam as funções de rede como M/M/1, associando os requisitos de QoS a diferentes classes e associando a taxa de serviço a quantidade de CPU para executar a função. O trabalho conduzido em [Plachy et al., 2021b] busca alocar recursos computacionais e de comunicação considerando a movimentação dos usuários, instanciando maquinas virtuais e analisando a qualidade dos recursos de comunicação ao longo de um caminho que pode ser percorrido pelos usuários. Em um primeiro momento, é proposto alocar os recursos minimizando a latência das aplicações, definida como a soma dos atrasos de propagação no meio de rádio, nos enlaces de *backhaul*, na inicialização das máquinas virtuais e no tempo demandado de *handover*. Entretanto, obter o ótimo global neste problema de minimização é complexo, visto que deve considerar todas as possibilidades de instanciar e alocar os recursos. Deste modo, o problema de minimiza-

ção da latência é transformado em um problema de maximização da taxa de comunicação, considerando as restrições iniciais do problema.

Apesar dos vários benefícios (agilidade, flexibilidade, entre outros) proporcionados pelas tecnologias habilitadoras, há vários desafios a serem resolvidos. A mudança de uma arquitetura baseada em *hardware* para uma baseada em *software* simplificará o compartilhamento de recursos de uma mesma infraestrutura entre vários clientes e aplicações. Contudo, a configuração de várias VNFs em uma única plataforma NFV trás problemas de gerenciamento em fatias de rede muito grandes. Para haver uma utilização eficiente de recursos de rede, é necessário o uso de algoritmos de agendamento inteligentes. Esse problema de gerenciamento só aumenta quando se expande o problema para o gerenciamento entre e intra fatias distintas. A orquestração fim a fim da rede, bem como o gerenciamento de diferentes fatias, cada uma com requisitos de SLAs diferentes, ao mesmo tempo que é necessário um uso eficiente de recursos da rede é extremamente desafiador.

Uma preocupação básica devido ao compartilhamento de uma única infraestrutura é com a segurança e privacidade das informações. Existirão várias fatias da rede, cada uma com suas políticas de seguraça e privacidade de acordo com as aplicações e serviços que possuem.

Também há desafios de pesquisa relacionados à mobilidade dos dispositivos. Cada fatia de rede possuirá dispositivos com requisitos distintos de latência e mobilidade e isso fica claro quando se compara, por exemplo, a utilização da rede por um serviço para atendimento a carros autônomos e outra fatia para atender a trens de alta velocidade onde serão demandados muito mais *handovers* [Barakabitze et al., 2020].

4.4.3. Coexistência das Aplicações Críticas

No ecossistema das redes 5G, as aplicações coexistem e interagem umas com as outras. A Figura 4.19 demonstra um exemplo da interação entre diferentes aplicações habilitadas pelo 5G. Cabe destaque especial aos dispositivos UAV, que podem ser usados para a coleta de dados de pesquisa em áreas remotas ou fornecer conteúdos bastante buscados em determinada região. Além disso, as redes *Smart Grid* podem utilizar os dispositivos UAV para a coleta de dados dos sensores espalhados pelas linhas de transmissão e distribuição. As *microgrids*, abastecem as redes elétricas a partir da geração limpa de energia (por exemplo, eólica e solar). A concessionária de energia controla a geração distribuída de energia através de sensores incorporados às linhas de transmissão, e os centros de controle analisam a demanda de energia elétrica em tempo real.

As aplicações de IoV se beneficiam da conectividade disponível na borda da rede móvel e podem trafegar com segurança à velocidades elevadas com baixíssimo risco de colisão, uma vez que os veículos compartilham dados de obstáculos a sua volta. O grande volume de dados gerados a partir do monitoramento dos sinais vitais do pedestre, que caminha à beira da estrada, são processados na MEC e encaminhados à equipe médica que avalia remotamente o seu estado de saúde.

Assim, as aplicações críticas poderão funcionar, apesar da segregação em fatias, em uma mesma rede 5G e, além disso, interoperarem entre si. Além disso, algumas da aplicações podem funcionar como provedora de recursos ou conexão para outra. Cabe

destaque que todas as aplicações devem ser atendidas por fatias da rede que garantem o isolamento e confiabilidade do serviço prestado e que a criação dinâmica e o gerenciamento dessas fatias ainda são desafios de pesquisa.

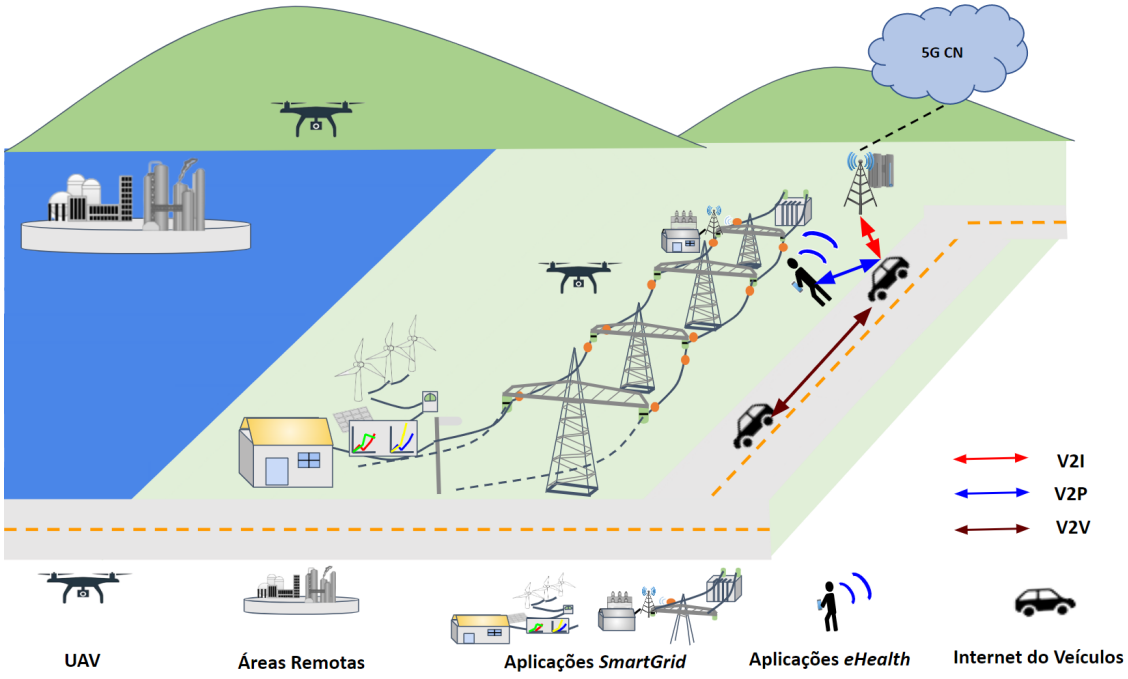

Figura 4.19. Coexistência das aplicações críticas nas redes 5G.

4.4.4. Simulação de redes 4G e 5G

Uma grande dificuldade na pesquisa em 5G é a disponibilidade de ferramentas de código aberto que permitam uma análise da comunicação em rede com a tecnologia 5G. Assim, essa parte do capítulo visa apresentar o uso do *Network Simulator 3* para o estudo das redes 5G.

A proposta prática deste minicurso é baseada na simulação de uma aplicação crítica utilizando como meio de conectividade as redes 4G e 5G, permitindo assim uma comparação dos resultados entre as tecnologias. No cenário proposto, a rede possui cinco estações rádio-base distribuídas ao longo de um trecho de 1 km de uma autoestrada. Nessa autoestrada, veículos se movem de uma ponta a outra. A rede possui conectividade com a Internet através da conexão de um *host* remoto conectado ao núcleo da rede. A quantidade de veículos é alterada e eles são divididos em 2 grupos que apresentam velocidades constantes diferentes (115 km/h e 120 km/h). A Figura 4.20 apresenta o cenário básico da simulação.

Existem atualmente vários simuladores de redes que estão em constante evolução e desenvolvimento para representar de maneira mais realista as novas redes 5G. Os mais relevantes são: *"5G K-Simulator"*[6] do *Korea Advanced Institute of Science and Technology*; o *"Vienna 5G Link Level Simulator"*[7] desenvolvido em Matlab feito pela *Technical*

[6]Disponível em: https://github.com/5GKSimSys/5G-K-SimSys
[7]Disponível em: https://www.tuwien.at/etit/tc/en/vienna-simulators/vienna-5g-simulators/

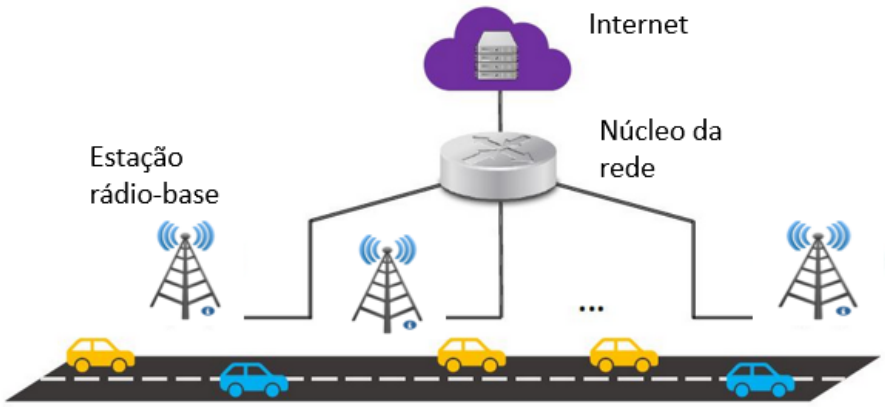

Figura 4.20. Cenário de simulação.

University of Vienna; o *"5G–air–simulator"*[8] e desenvolvido pela *Politechnico di Bari*; o *"Simu5G"*[9] que é uma extensão para o simulador *Omnet++* e foi desenvolvido pela *University of Pisa*; e o *"5G LENA"* que é um módulo para o *Network Simulator* – ns-3 desenvolvido pelo *Centre Tecnològic Telecomunicacions Catalunya (CTTC)*.

O simulador escolhido foi o ns–3[10] que é um simulador de eventos discretos e desenvolvido com fins educacionais e de pesquisa. Ele possui código aberto é amplamente reconhecido pela comunidade acadêmica, além de conter vasta documentação, desenvolvimento e manutenção constantes.

O módulo *5G LENA*[11] é baseado no módulo de ondas milimétricas, módulo totalmente aderente as recomendações do 3GPP, do ns–3 e uma evolução natural do simulador 4G, *"LTE/EPC Network Simulator"*, desenvolvido pelo mesmo centro técnológico. Os scripts utilizados e as instruções para execução da simulação se encontram em repositório[12] na internet.

4.4.5. Projetos de pesquisa

Diversos órgãos e entidades estão investindo recursos e tempo para na pesquisa de aplicações críticas e como as redes 5G podem atender seus requisitos.

O projeto 5G-MOBIX[13] é uma cooperação entre instituições de ensino, órgãos governamentais e empresas de países da União Europeia, Turquia, China e Coreia, que visa desenvolver e avaliar funcionalidades de veículos automatizados sob as principais inovações tecnológicas do 5G. O projeto conta com infraestrutura completa de conectividade 5G em 2 fronteiras entre países (Grécia-Turquia e Espanha-Portugal) e 6 sites de testes na França, Alemanha, Holanda, Finlândia, China e Coreia do Sul. Os sites supor-

[8]Disponível em: https://github.com/telematics-lab/5G-air-simulator
[9]Disponível em: http://simu5g.org/
[10]Disponível em: https://www.nsnam.org
[11]Disponível em: https://5g-lena.cttc.es
[12]Disponível em: https://github.com/fp-cn/Minicurso4_SBRC_2023
[13]Disponível em: https://www.5g-mobix.com/about

tam funcionalidades avançadas de mobilidade automatizada. São testadas uma variedade de tecnologias, funções e aplicações 5G.

O projeto PD_manager [14], desenvolvido na Europa, busca fornecer um acompanhamento para pacientes com *Parkinson* por uma equipe multidisciplinar de médicos. Os pacientes utilizam sensores simples e acessíveis, como relógios inteligentes e palmilhas para medir o equilíbrio, e os dados são enviados para a equipe de médicos, que pode fornecer o tratamento terapêutico especifico para cada paciente. Já o projeto SERAS [15], também desenvolvido na Europa, busca alertar pacientes que sofrem de epilepsia antes de suas crises epiléticas. O paciente utiliza um fone de ouvido que monitora a atividade cerebral e envia os dados para um aplicativo, que ao detectar uma possível crise epilética, com cerca de 1 minuto de antecedência, envia um alarme ao paciente, que pode se colocar em uma posição segura para evitar acidentes e lesões.

O HORSE [16] é um projeto financiado pelo programa de inovação e pesquisa da União Europeia *Horizon 2020* e busca uma integração, para pequenas e médias empresas, de sistemas robóticos integrados e inteligentes controlados por processos de fabricação dinâmicos baseados em IoT. Nele, é proposto um novo modelo de manufatura inteligente envolvendo a colaboração de humanos, robôs, veículos autônomos e maquinário. Apesar de toda a evolução em sistemas físico-cibernéticos, IoT e robótica, as linhas de produção ainda não são flexíveis. Questões de segurança não permitem a integração completa entre homens e máquinas e o processo de controle robótico ainda é isolado, não permitindo uma integração com o processo fim-a-fim de fabricação. Para resolver essas questões, o HORSE foca em múltiplas áreas de maneira a criar linhas de produção extremamente flexíveis e reconfiguráveis. Foi criado um pilar técnico buscando uma arquitetura de referência para um sistema físico cibernético que combina o processo de produção no espaço cibernético com agentes humanos em um ambiente de produção integrado. Alguns casos piloto que podem ser destacados são:

1. Experimento para inspeção visual robótica com co-manipulação humana e robótica em uma fábrica da BOSCH na Espanha.

2. Aplicações de trabalho compartilhando entre homens e robôs para controle híbrido de posicionamento em uma fábrica da Odlewnie Polskie SA na Polônia.

3. Montagem flexível com robôs móveis em uma fábrica da Thomas Regout International na Holanda.

Financiado pelo programa de inovação e pesquisa da União Europeia o 5G–Drones[17] busca testar vários casos de uso de veículos aéreos não tripulados envolvendo todos as categorias de uso definidas para o 5G (eMBB, uRLLC e mMTC), validando os KPIs para suportar essas aplicações. A ideia principal é demonstrar que o 5G garante as necessidades básicas para o uso dos UAV e provar que também suportam os casos de uso

[14]Disponível em: http://www.parkinson-manager.eu/
[15]Disponível em: https://d-lab.tech/mjn/
[16]HORSE project - Disponível em: http://www.horse-project.eu/
[17]Disponível em: https://5gdrones.eu/

desafiadores e que exigem o máximo dos recursos da rede, como latência, confiabilidade, grande quantidade de conexões e alta taxa de dados. O projeto atenderá aos diferentes cenários através do fatiamento de uma mesma infraestrutura de rede que compartilha os mesmos recursos. São definidos 4 casos de uso: gerenciamento de tráfego de UAV; segurança contra desastres; consciência situacional e conectividade durante grandes eventos.

A organização norte americana Aliança para Soluções Industriais de Telecomunicações *Alliance for Telecommunications Industry Solutions* (ATIS) [18] é composta por várias companhias com o objetivos de novas soluções tecnológicas para a indústria de tecnologia da informação e telecomunicações. Entre as tecnologias pesquisadas estão a evolução da rede 5G, comunicações e infraestruturas críticas, carros conectados, IoT, UAV, NFV, satélites e redes de 6ª geração.

O Comitê de Tecnologia de Telecomunicações (TTC) é uma organização de desenvolvimento de padrões no Japão [19]. O TTC está ligado ao ministério de comunicações japonês e representando o país em todas as discussões junto ao ITU. No campo das aplicações críticas, o TTC possui grupos de trabalho relacionados a IoV, *e-Health*, IoT, SDN e comunicações de emergência.

4.4.6. Caminho para o 6G

Com a implantação das redes 5G por todo o mundo, várias aplicações e casos de uso são criados utilizando todas as vantagens e recursos oferecidos por essas redes e desafiando os seus limites. Esse fato impulsiona os pesquisadores a trabalhar na próxima geração de redes móveis celulares, buscando uma grande evolução para suprir as demandas futuras da sociedade. Apesar dos avanços e qualidades das redes 5G, espera-se que as redes 6G sejam mais inteligentes, confiáveis, escaláveis, energeticamente eficientes, de forma a atender aplicações emergentes e atuais que estão em constante evolução e que não poderão ser atendidas pelas redes 5G. Os requisitos de taxa de dados ultra-altas, acesso em tempo real a recursos de computação, latências extremamente baixas, localização precisa, alta confiabilidade são muito mais estritos e superam em larga margem os das redes 5G. Essas aplicações (chamadas holográficas –*Holographic Telepresence* (HT), UAV, realidade estendida –XR, *smart grid 2.0*, Indústria 5.0 e Internet de Tudo –*Internet of Everything* (IoE)) demandarão o uso e implementação de novas tecnologias habilitadoras como: *Edge Intelligence* (EI), comunicação em frequências na casa dos THz, *Non-Orthogonal Multiple Access* (NOMA), *Large Intelligent Surfaces* (LIS), *Self-Sustaining Networks* (SSN) entre outras [De Alwis et al., 2021].

O lançamento do 6G está previsto para 2030 e ainda há muito a ser feito. A Figura 4.21 apresenta o cronograma previsto para implantação do 6G. Por conta desse ser um tema extremamente recente, ainda não há consenso na literatura sobre a nomenclatura das categorias de uso do 6G. De Alwis et al. [De Alwis et al., 2021] definem os 4 cenários de uso como sendo: *Further enhanced Mobile Broadband* (FeMBB), *Ultra-massive Machine-Type Communication* (umMTC), *Mobile BroadBand and Low-Latency* (MBBLL) e *massive Low-Latency Machine Type communication* (mLLMT). Já Chowdhury et al [Chowdhury et al., 2020] definem os quatro cenários de uso como: *Ubi-*

[18]Disponível em: https://www.atis.org/technologies/

[19]Disponível em: https://www.ttc.or.jp/e/org/workin-groups

quitous mobile ultra-broadband (uMUB), *Ultra-high-speed with low-latency communications* (uHSLLC), mMTC e *ultra-High Data Density* (uHDD). Porém uma questão é exatamente igual: os KPIs das redes 6G serão muito mais estritos.Fala-se de taxas de pico acima de 1 Tbps, atraso fim a fim menor que 0,1 ms, tempo de processamento de 10 ns, disponibilidade maior que 99,99999%, densidade de conexão de 10^7dispositivos/km², eficiência energética 5 vezes maior que no 5G e velocidades de deslocamento de até 1000 km/h.

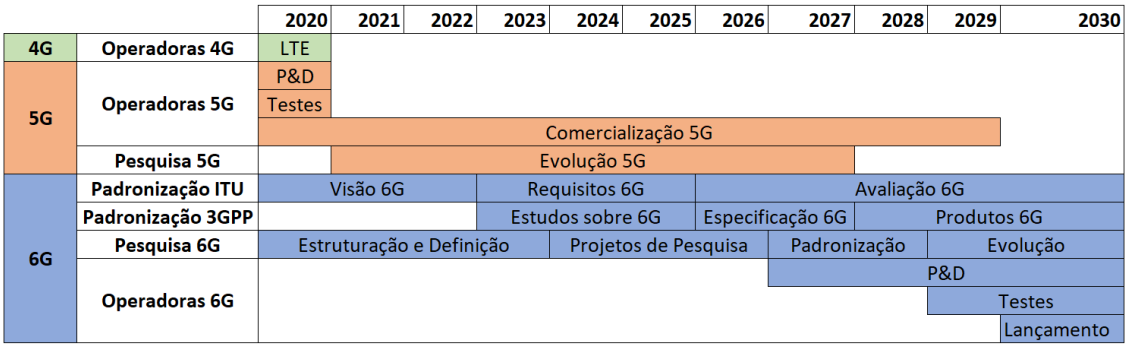

Figura 4.21. Cronograma Implantação 6G. Adaptado de [De Alwis et al., 2021]

A Figura 4.22 apresenta a evolução dos KPIs comparando-se as redes 4G, 5G e 6G.

4.5. Considerações Finais

As redes móveis têm se tornado um ponto de convergência de tecnologias de comunicação e computação. Desde a 1ª Geração, que tinha o núcleo da rede baseado em comutação por circuitos, à 5 Geração, que implementa um núcleo orientado a serviços, as redes móveis têm alcançado cada vez mais destaque como principal meio de comunicação sem fio. As redes 4G já implementavam avanços significativos em relação à geração anterior. Pela primeira vez em uma rede móvel, a comutação era totalmente por pacotes e as chamadas de voz eram por comutação IP, através da VoLTE, além dos avanços que ocorreram durante a maturidade da 4ª Geração, como a agregação de portadora e o aumento das possibilidades de comunicação entre máquinas, consolidando o protagonismo destas redes no cenário mundial. Esse protagonismo impulsiona o surgimento de novos serviços, que exigem cada vez mais recursos e disponibilidade das redes móveis.

Como evolução natural, as redes 5G implementam tecnologias novas e tradicionais para atender e habilitar aplicações que necessitam de alta confiabilidade, disponibilidade e conectividade, que não podem ser oferecidas por gerações anteriores. Dentre estas aplicações, as aplicações críticas geram grande interesse na indústria e comunidade científica devido à rigidez dos seus requisitos de desempenho. Taxa de dados (pico e média), eficiência espectral e energética, latência, densidade de conexão, capacidade de tráfego por área e mobilidade são todos indicadores chave que foram extremamente melhorados, tornando o 5G capaz de atender às aplicações críticas.

Para proporcionar essa evolução nos indicadores chave de desempenho, a arquitetura da RAN e do CN passa a ser orientada a serviços, implementando tecnologias que

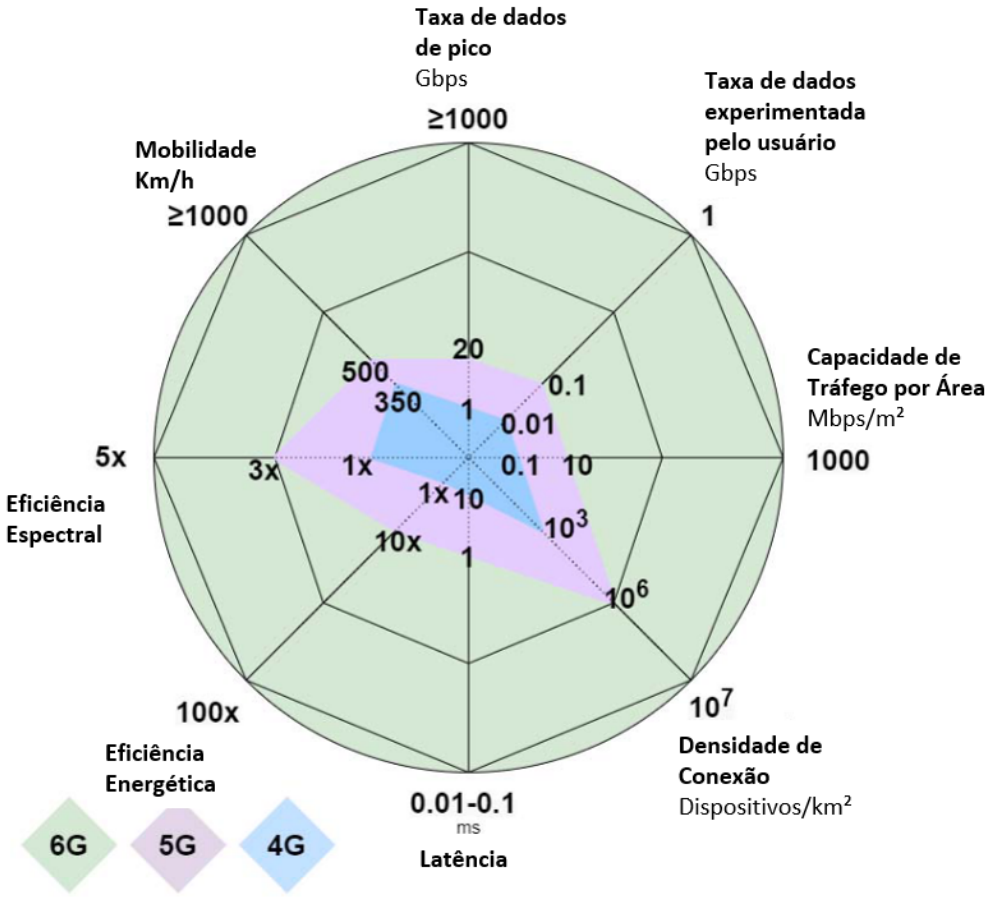

Figura 4.22. Evolução requisitos 4G x 5G x 6G. Adaptado de [De Alwis et al., 2021]

permitem fácil gerenciamento e que utilizam de maneira eficiente os recursos disponíveis. Em especial, o uso das SDNs aliadas à NFV proporciona um controle completo da rede, reduzindo o tempo de configuração de novas aplicações e o custo de CAPEX e OPEX. A implementação da MEC traz para perto do usuário o poder computacional e de armazenamento que antes era alocado somente no núcleo da rede, reduzindo o latência das aplicações. Todas essas tecnologias possibilitam o uso do Fatiamento da Rede, que permite um isolamento e customização das aplicações e o uso eficiente dos recursos, uma vez que somente a parcela necessária de recursos que atenda à aplicação é alocada. Além disso, em caso de falhas, uma nova fatia pode ser rapidamente configurada, garantindo a disponibilidade da aplicação.

Entretanto, problemas relacionados à escalabilidade das aplicações, privacidade dos dados e mobilidade das VNFs movimentam as pesquisas dos setores industriais e acadêmicos. Com aplicações extremamente complexas, o gerenciamento destas fatias e do ciclo de vida das VNFs passa de simples e eficiente para complexo e computacionalmente custoso. Neste mesmo sentido, o uso de redes comuns para aplicações heterogêneas com políticas de segurança distintas pode expor dados sensíveis dos usuários. A mobilidade dos usuários, característica evidente das redes móveis, impõe mobilidade também para as VNFs, aumentando demasiadamente o número de *handovers* na rede.

As aplicações críticas irão diferir em relação as características do 5G das quais usufruem e que atendem melhor aos seus requisitos. Dentre os 3 cenários de uso, o eMBB será o utilizado pelas aplicações que necessitam de uma maior taxa de dados transmitido. Um grupo de aplicações como, por exemplo, UAV, IoV e *smart grid* depende de conexões extremamente confiáveis, baixas latências e possuem como cenário predominante o uRLLC. Por último, tem-se as aplicações que possuem como característica mais marcante a necessidade de uma grande quantidade de dispositivos conectados, como as voltadas para industria (manufatura inteligente e gêmeos digitais).

Apesar dos desenvolvimentos previstos, as redes 5G não terão capacidade de atender à demanda futura de aplicações e serviços. O rápido desenvolvimento de aplicações emergentes como, inteligência artificial, realidade virtual e chamadas holográficas, além do aumento do volume de tráfego oriundo de redes móveis (era de 7,462 EB/mês em 2010 e é esperado que chegue a 5016 EB/mês em 2030 [ITU-R, 2015a]) apontam para a constante necessidade de evolução das redes móveis. Os estudos para a padronização e desenvolvimento das redes de 6ª geração ainda estão no começo (a previsão de lançamento é no ano de 2030), mas já começam a ser desenhados os requisitos para essas redes, assim como as novas tecnologias habilitadoras (EI, NOMA, entre outras). Aqui, fala-se, por exemplo, de taxas acima de 1 Tbps e atraso fim a fim de 0,1 ms [Chowdhury et al., 2020].

Com isso, conclui-se que ainda existe muitos campos de estudo nas redes 5G, em especial no que diz respeito ao planejamento das VNFs, gestão inteligente da rede, integração com tecnologias como a de satélite, entre diversos outros. Além disso, a aplicação dessas novas tecnologias dentro dos casos de uso específicos de cada aplicação crítica também requer estudos de viabilidade e custos. Por fim, a definição dos cenários não atendidos pelo 5G, mas que potencialmente serão demandados por novas versões de aplicações, gera insumos para novas proposições tecnológicas a serem avaliadas para o 6G.

Referências

[3GPP, 2017] 3GPP (2017). System architecture milestone of 5G phase 1 is achieved.

[3GPP, 2020a] 3GPP (2020a). Study on new radio (NR) to support non-terrestrial networks. Relatório técnico, 3GPP.

[3GPP, 2020b] 3GPP (2020b). Study on using satellite access in 5G. Relatório técnico, 3GPP.

[3GPP, 2021] 3GPP (2021). Solutions for NR to support Non-Terrestrial Networks (NTN). Relatório técnico, 3GPP.

[3GPP, 2016] 3GPP (2016). Feasibility study on new services and markets technology enablers; stage 1 (release 14). Relatório Técnico TR 22.891, Third Generation Partnership Project.

[Adrah et al., 2022] Adrah, C. M., Palma, D., Kure, O. e Heegaard, P. E. (2022). Deploying 5G architecture for protection systems in smart distribution grids. Em *2022 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT)*, p. 1–5.

[Agarwal et al., 2019] Agarwal, S., Malandrino, F., Chiasserini, C. F. e De, S. (2019). VNF placement and resource allocation for the support of vertical services in 5G networks. *IEEE/ACM Transactions on Networking*, 27(1):433–446.

[Agbaje et al., 2022] Agbaje, P., Anjum, A., Mitra, A., Oseghale, E., Bloom, G. e Olufowobi, H. (2022). Survey of interoperability challenges in the Internet of Vehicles. *IEEE Transactions on Intelligent Transportation Systems*, 23(12):22838–22861.

[Ahvar et al., 2021] Ahvar, E., Ahvar, S., Raza, S. M., Manuel Sanchez Vilchez, J. e Lee, G. M. (2021). Next generation of SDN in cloud-fog for 5G and beyond-enabled applications: Opportunities and challenges. *Network*, 1(1):28–49.

[Alenoghena et al., 2022] Alenoghena, C. O., Onumanyi, A. J., Ohize, H. O., Adejo, A. O., Oligbi, M., Ali, S. I. e Okoh, S. A. (2022). ehealth: A survey of architectures, developments in mhealth, security concerns and solutions. *International Journal of Environmental Research and Public Health*, 19(20):13071.

[Arshad et al., 2019] Arshad, Q. K. U. D., Kashif, A. U. e Quershi, I. M. (2019). A review on the evolution of cellular technologies. Em *2019 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST)*, p. 989–993. IEEE.

[Azari et al., 2022] Azari, M. M., Solanki, S., Chatzinotas, S., Kodheli, O., Sallouha, H., Colpaert, A., Mendoza Montoya, J. F., Pollin, S., Haqiqatnejad, A., Mostaani, A., Lagunas, E. e Ottersten, B. (2022). Evolution of non-terrestrial networks from 5G to 6G: A survey. *IEEE Communications Surveys & Tutorials*, 24(4):2633–2672.

[Barakabitze et al., 2020] Barakabitze, A. A., Ahmad, A., Mijumbi, R. e Hines, A. (2020). 5G network slicing using SDN and NFV: A survey of taxonomy, architectures and future challenges. *Computer Networks*, 167:106984.

[Bazzi et al., 2021] Bazzi, A., Berthet, A. O., Campolo, C., Masini, B. M., Molinaro, A. e Zanella, A. (2021). On the design of sidelink for cellular V2X: A literature review and outlook for future. *IEEE Access*, 9:97953–97980.

[Carrillo et al., 2022] Carrillo, D., Kalalas, C., Raussi, P., Michalopoulos, D. S., Rodriguez, D. Z., Kokkoniemi-Tarkkanen, H., Ahola, K., Nardelli, P. H. J., Fraidenraich, G. e Popovski, P. (2022). Boosting 5G on smart grid communication: A smart ran slicing approach. *IEEE Wireless Communications*, p. 1–8.

[Cavalcante et al., 2021] Cavalcante, A. M., Marquezini, M. V., Mendes, L. e Moreno, C. S. (2021). 5G for remote areas: Challenges, opportunities and business modeling for brazil. *IEEE Access*, 9:10829–10843.

[Ceriotti et al., 2012] Ceriotti, M., Diedrich, B. L. e McInnes, C. R. (2012). Novel mission concepts for polar coverage: An overview of recent developments and possible future applications. *Acta Astronautica*, 80:89–104.

[Chowdhury et al., 2020] Chowdhury, M. Z., Shahjalal, M., Ahmed, S. e Jang, Y. M. (2020). 6G wireless communication systems: Applications, requirements, technologies, challenges, and research directions. *IEEE Open Journal of the Communications Society*, 1:957–975.

[Cox, 2020] Cox, C. (2020). *An introduction to 5G: the new radio, 5G network and beyond*. John Wiley & Sons, 1 edição.

[Darwish et al., 2022] Darwish, T., Kurt, G. K., Yanikomeroglu, H., Bellemare, M. e Lamontagne, G. (2022). LEO satellites in 5G and beyond networks: A review from a standardization perspective. *IEEE Access*, 10:35040–35060.

[De Alwis et al., 2021] De Alwis, C., Kalla, A., Pham, Q.-V., Kumar, P., Dev, K., Hwang, W.-J. e Liyanage, M. (2021). Survey on 6G frontiers: Trends, applications, requirements, technologies and future research. *IEEE Open Journal of the Communications Society*, 2:836–886.

[de Souza Lopes et al., 2020] de Souza Lopes, C. H., Lima, E. S., Pereira, L. A. M., Borges, R. M., Ferreira, A. C., Abreu, M., Dias, W. D., Spadoti, D. H., Mendes, L. L. e Junior, A. C. S. (2020). Non-standalone 5G NR fiber-wireless system using FSO and fiber-optics fronthauls. *Journal of Lightwave Technology*, 39(2):406–417.

[Esenogho et al., 2022] Esenogho, E., Djouani, K. e Kurien, A. M. (2022). Integrating artificial intelligence Internet of Things and 5G for next-generation smartgrid: A survey of trends challenges and prospect. *IEEE Access*, 10:4794–4831.

[ETSI, 2022] ETSI (2022). Technical specification 5G: Unmanned aerial system (UAS) support in 3GPP (3GPP TS 22.125 version 17.6.0 release 17). Relatório Técnico ETSI TS 122 125 V17.6.0, ETSI, França.

[Fernandes et al., 2011] Fernandes, N. C., Moreira, M. D., Moraes, I. M., Ferraz, L. H. G., Couto, R. S., Carvalho, H. E., Campista, M. E. M., Costa, L. H. M. e Duarte, O. C. M. (2011). Virtual networks: Isolation, performance, and trends. *Annals of Telecommunications*, 66:339–355.

[Fotouhi et al., 2019] Fotouhi, A., Qiang, H., Ding, M., Hassan, M., Giordano, L. G., Garcia-Rodriguez, A. e Yuan, J. (2019). Survey on UAV cellular communications: Practical aspects, standardization advancements, regulation, and security challenges. *IEEE Communications surveys & tutorials*, 21(4):3417–3442.

[Geraci et al., 2022] Geraci, G., Garcia-Rodriguez, A., Azari, M. M., Lozano, A., Mezzavilla, M., Chatzinotas, S., Chen, Y., Rangan, S. e Di Renzo, M. (2022). What will the future of UAV cellular communications be? A flight from 5G to 6G. *IEEE communications surveys & tutorials*, 24(3):1304–1335.

[Ghosh et al., 2019] Ghosh, A., Maeder, A., Baker, M. e Chandramouli, D. (2019). 5G evolution: A view on 5G cellular technology beyond 3GPP release 15. *IEEE access*, 7:127639–127651.

[Guidotti et al., 2019] Guidotti, A., Vanelli-Coralli, A., Conti, M., Andrenacci, S., Chatzinotas, S., Maturo, N., Evans, B., Awoseyila, A., Ugolini, A., Foggi, T. et al. (2019). Architectures and key technical challenges for 5G systems incorporating satellites. *IEEE Transactions on Vehicular Technology*, 68(3):2624–2639.

[Gupta e Jha, 2015] Gupta, A. e Jha, R. K. (2015). A survey of 5G network: Architecture and emerging technologies. *IEEE access*, 3:1206–1232.

[Gupta et al., 2019] Gupta, R., Tanwar, S., Tyagi, S. e Kumar, N. (2019). Tactile Internet and its applications in 5G era: A comprehensive review. *International Journal of Communication Systems*, 32(14):e3981.

[Hosseinian et al., 2021] Hosseinian, M., Choi, J. P., Chang, S.-H. e Lee, J. (2021). Review of 5G NTN standards development and technical challenges for satellite integration with the 5G network. *IEEE Aerospace and Electronic Systems Magazine*, 36(8):22–31.

[Huawei, 2020] Huawei (2020). Huawei launches industry's first site digital twins based 5G digital engineering solution. https://www.huawei.com/en/news/2020/2/site-digital-twins-based-5g-digital-engineering-solution. Acessado em 23.04.2023.

[International Energy Agency - IEA, 2022] International Energy Agency - IEA (2022). World energy outlook 2022. Relatório técnico, International Energy Agency, Paris, França.

[Isto et al., 2020] Isto, P., Heikkilä, T., Mämmelä, A., Uitto, M., Seppälä, T. e Ahola, J. M. (2020). 5G based machine remote operation development utilizing digital twin. *Open Engineering*, 10(1):265–272.

[ITU-R, 2015a] ITU-R (2015a). IMT traffic estimates for the years 2020 to 2030. Relatório Técnico M.2370, ITU, Geneva, Switzerland.

[ITU-R, 2015b] ITU-R (2015b). IMT vision–framework and overall objectives of the future development of IMT for 2020 and beyond. Relatório Técnico ITU-R M.2083-0 - M Series, ITU, Geneva, Switzerland.

[Iurii et al., 2022] Iurii, D., Carrillo, M. D., Antti, P., Liana, A., Jari, J. e Kirsi, L. (2022). *IEC-61850 Performance Evaluation in a 5G Cellular Network: UDP and TCP Analysis*. Springer, Cham, Switzerland.

[Jiang et al., 2021] Jiang, T., Zhang, J., Tang, P., Tian, L., Zheng, Y., Dou, J., Asplund, H., Raschkowski, L., D'Errico, R. e Jämsä, T. (2021). 3GPP standardized 5G channel model for IIoT scenarios: A survey. *IEEE Internet of Things Journal*, 8(11):8799–8815.

[Khorsandroo et al., 2021] Khorsandroo, S., Sánchez, A. G., Tosun, A. S., Arco, J. M. e Doriguzzi-Corin, R. (2021). Hybrid SDN evolution: A comprehensive survey of the state-of-the-art. *Computer Networks*, 192:107981.

[Li et al., 2018] Li, S., Zhang, N., Lin, S., Kong, L., Katangur, A., Khan, M. K., Ni, M. e Zhu, G. (2018). Joint admission control and resource allocation in edge computing for Internet of Things. *IEEE Network*, 32(1):72–79.

[Lin et al., 2021] Lin, X., Rommer, S., Euler, S., Yavuz, E. A. e Karlsson, R. S. (2021). 5G from space: An overview of 3GPP non-terrestrial networks. *IEEE Communications Standards Magazine*, 5(4):147–153.

[Liu et al., 2020a] Liu, G., Huang, Y., Chen, Z., Liu, L., Wang, Q. e Li, N. (2020a). 5G deployment: Standalone vs. non-standalone from the operator perspective. *IEEE Communications Magazine*, 58(11):83–89.

[Liu et al., 2020b] Liu, G., Huang, Y., Chen, Z., Liu, L., Wang, Q. e Li, N. (2020b). 5G deployment: Standalone vs. non-standalone from the operator perspective. *IEEE Communications Magazine*, 58(11):83–89.

[Liu et al., 2021] Liu, L., Guo, S., Liu, G. e Yang, Y. (2021). Joint dynamical VNF placement and SFC routing in NFV-enabled SDNs. *IEEE Transactions on Network and Service Management*, 18(4):4263–4276.

[Lopes et al., 2016] Lopes, Y., Bornia, T., Farias, V., Fernandes, N. C. e Muchaluat-Saade, D. C. (2016). *Desafios de segurança e confiabilidade na comunicação para smart grids*, p. 142–186. SBC, Porto Alegre, Brasil.

[Luo et al., 2020] Luo, C., Satpute, M. N., Li, D., Wang, Y., Chen, W. e Wu, W. (2020). Fine-grained trajectory optimization of multiple UAVs for efficient data gathering from WSNs. *IEEE/ACM Transactions on Networking*, 29(1):162–175.

[Lyu et al., 2022] Lyu, Z., Zhu, G. e Xu, J. (2022). Joint maneuver and beamforming design for UAV-enabled integrated sensing and communication. *IEEE Transactions on Wireless Communications*, 22(4):2424 – 2440.

[Maddikunta et al., 2022] Maddikunta, P. K. R., Pham, Q.-V., Prabadevi, B., Deepa, N., Dev, K., Gadekallu, T. R., Ruby, R. e Liyanage, M. (2022). Industry 5.0: A survey on enabling technologies and potential applications. *Journal of Industrial Information Integration*, 26:100257.

[Mahmood et al., 2020] Mahmood, N. H., Marchenko, N., Gidlund, M. e Popovski, P. (2020). *Wireless Networks and Industrial IoT*. Springer, 1 edição.

[Mallorquí e Zaballos, 2021] Mallorquí, A. e Zaballos, A. (2021). A heterogeneous layer-based trustworthiness model for long backhaul NVIS challenging networks and an IoT telemetry service for antarctica. *Sensors*, 21(3446).

[Maroufkhani et al., 2022] Maroufkhani, P., Desouza, K. C., Perrons, R. K. e Iranmanesh, M. (2022). Digital transformation in the resource and energy sectors: A systematic review. *Resources Policy*, 76:102622.

[Mendes et al., 2020] Mendes, L. L., Moreno, C. S., Marquezini, M. V., Cavalcante, A. M., Neuhaus, P., Seki, J., Aniceto, N. F. T., Karvonen, H., Vidal, I., Valera, F. et al. (2020). Enhanced remote areas communications: The missing scenario for 5G and beyond 5G networks. *IEEE Access*, 8:219859–219880.

[Mihai et al., 2022] Mihai, S., Yaqoob, M., Hung, D. V., Davis, W., Towakel, P., Raza, M., Karamanoglu, M., Barn, B., Shetve, D., Prasad, R. V. et al. (2022). Digital twins: a survey on enabling technologies, challenges, trends and future prospects. *IEEE Communications Surveys & Tutorials*, 24(4).

[Moglia et al., 2022] Moglia, A., Georgiou, K., Marinov, B., Georgiou, E., Berchiolli, R. N., Satava, R. M. e Cuschieri, A. (2022). 5G in healthcare: from COVID-19 to future challenges. *IEEE Journal of Biomedical and Health Informatics*, 26(8):4187–4196.

[Mourtzis et al., 2021] Mourtzis, D., Angelopoulos, J. e Panopoulos, N. (2021). Smart manufacturing and tactile Internet based on 5G in Industry 4.0: challenges, applications and new trends. *Electronics*, 10(24):3175.

[Nguyen et al., 2021] Nguyen, H. X., Trestian, R., To, D. e Tatipamula, M. (2021). Digital twin for 5G and beyond. *IEEE Communications Magazine*, 59(2):10–15.

[Noor-A-Rahim et al., 2020] Noor-A-Rahim, M., Liu, Z., Lee, H., Ali, G. M. N., Pesch, D. e Xiao, P. (2020). A survey on resource allocation in vehicular networks. *IEEE transactions on intelligent transportation systems*, 23(2):701–721.

[O'Connell et al., 2020] O'Connell, E., Moore, D. e Newe, T. (2020). Challenges associated with implementing 5G in manufacturing. Em *Telecom*, p. 48–67. Multidisciplinary Digital Publishing Institute.

[Petrobras, 2020] Petrobras (2020). O que são digital twins e como podem aumentar a eficiência operacional. https://nossaenergia.petrobras.com.br/energia/22-digital-twins/. Acessado em 23.04.2023.

[Plachy et al., 2021a] Plachy, J., Becvar, Z., Strinati, E. C. e di Pietro, N. (2021a). Dynamic allocation of computing and communication resources in multi-access edge computing for mobile users. *IEEE Transactions on Network and Service Management*, 18(2):2089–2106.

[Plachy et al., 2021b] Plachy, J., Becvar, Z., Strinati, E. C. e di Pietro, N. (2021b). Dynamic allocation of computing and communication resources in multi-access edge computing for mobile users. *IEEE Transactions on Network and Service Management*, 18(2):2089–2106.

[Quang et al., 2019] Quang, P. T. A., Hadjadj-Aoul, Y. e Outtagarts, A. (2019). A deep reinforcement learning approach for VNF forwarding graph embedding. *IEEE Transactions on Network and Service Management*, 16(4):1318–1331.

[Rahman et al., 2021] Rahman, I., Razavi, S. M., Liberg, O., Hoymann, C., Wiemann, H., Tidestav, C., Schliwa-Bertling, P., Persson, P. e Gerstenberger, D. (2021). 5G evolution toward 5G advanced: An overview of 3GPP releases 17 and 18. *Ericsson Technology Review*, 2021(14):2–12.

[Ramirez et al., 2022] Ramirez, R., Huang, C.-Y. e Liang, S.-H. (2022). 5G digital twin: A study of enabling technologies. *Applied Sciences*, 12(15):7794.

[Raussi et al., 2023] Raussi, P., Kokkoniemi-Tarkkanen, H., Ahola, K., Heikkinen, A. e Uitto, M. (2023). Improving reliability of protection communication in a 5G slice. *Authorea*.

[Rinaldi et al., 2020] Rinaldi, F., Maattanen, H.-L., Torsner, J., Pizzi, S., Andreev, S., Iera, A., Koucheryavy, Y. e Araniti, G. (2020). Non-terrestrial networks in 5G & beyond: A survey. *IEEE access*, 8:165178–165200.

[Saba et al., 2020] Saba, T., Haseeb, K., Ahmed, I. e Rehman, A. (2020). Secure and energy-efficient framework using Internet of Medical Things for e-healthcare. *Journal of Infection and Public Health*, 13(10):1567–1575.

[Saeed et al., 2020] Saeed, N., Elzanaty, A., Almorad, H., Dahrouj, H., Al-Naffouri, T. Y. e Alouini, M.-S. (2020). Cubesat communications: Recent advances and future challenges. *IEEE Communications Surveys & Tutorials*, 22(3):1839–1862.

[Santos et al., 2020] Santos, A. C., Firmino, R. M., Soto, J. C., Medeiros, D. S., Mattos, D. M., Albuquerque, C. V., Seixas, F., Muchaluat-Saade, D. C. e Fernandes, N. C. (2020). Aplicações em redes de sensores na área da saúde e gerenciamento de dados médicos: tecnologias em ascensão. *Sociedade Brasileira de Computação*.

[Sehla et al., 2022] Sehla, K., Nguyen, T. M. T., Pujolle, G. e Velloso, P. B. (2022). Resource allocation modes in C-V2X: from LTE-V2X to 5G-V2X. *IEEE Internet of Things Journal*, 9(11):8291–8314.

[Shah et al., 2021] Shah, S. D. A., Gregory, M. A. e Li, S. (2021). Cloud-native network slicing using software defined networking based multi-access edge computing: A survey. *IEEE Access*, 9:10903–10924.

[Shaik e Malik, 2021] Shaik, N. e Malik, P. K. (2021). A comprehensive survey 5G wireless communication systems: open issues, research challenges, channel estimation, multi carrier modulation and 5G applications. *Multimedia Tools and Applications*, 80(19):28789–28827.

[Song et al., 2019] Song, S., Lee, C., Cho, H., Lim, G. e Chung, J.-M. (2019). Clustered virtualized network functions resource allocation based on context-aware grouping in 5G edge networks. *IEEE Transactions on Mobile Computing*, 19(5):1072–1083.

[Soto et al., 2022] Soto, I., Calderon, M., Amador, O. e Urueña, M. (2022). A survey on road safety and traffic efficiency vehicular applications based on C-V2X technologies. *Vehicular Communications*, 33:100428.

[Spinelli e Mancuso, 2020] Spinelli, F. e Mancuso, V. (2020). Toward enabled industrial verticals in 5G: A survey on mec-based approaches to provisioning and flexibility. *IEEE Communications Surveys & Tutorials*, 23(1):596–630.

[Wang et al., 2022] Wang, D., Zhou, F., Lin, W., Ding, Z. e Al-Dhahir, N. (2022). Cooperative hybrid nonorthogonal multiple access-based mobile-edge computing in cognitive radio networks. *IEEE Transactions on Cognitive Communications and Networking*, 8(2):1104–1117.

[Wang et al., 2016] Wang, L., Lu, Z., Wen, X., Knopp, R. e Gupta, R. (2016). Joint optimization of service function chaining and resource allocation in network function virtualization. *IEEE Access*, 4:8084–8094.

[Wang et al., 2019] Wang, M., Cheng, B., Feng, W. e Chen, J. (2019). An efficient service function chain placement algorithm in a MEC-NFV environment. Em *2019 IEEE Global Communications Conference (GLOBECOM)*, p. 1–6. IEEE.

[Wei et al., 2022] Wei, Z., Zhu, M., Zhang, N., Wang, L., Zou, Y., Meng, Z., Wu, H. e Feng, Z. (2022). UAV-assisted data collection for Internet of Things: A survey. *IEEE Internet of Things Journal*, 9(17):15460–15483.

[Wu et al., 2022] Wu, Y., Dai, H.-N., Wang, H., Xiong, Z. e Guo, S. (2022). A survey of intelligent network slicing management for industrial IoT: integrated approaches for smart transportation, smart energy, and smart factory. *IEEE Communications Surveys & Tutorials*, 24(2):1175–1211.

[Wu et al., 2021a] Wu, Y., Zhang, K. e Zhang, Y. (2021a). Digital twin networks: A survey. *IEEE Internet of Things Journal*, 8(18):13789–13804.

[Wu et al., 2021b] Wu, Y., Zheng, W., Zhang, Y. e Li, J. (2021b). Reliability-aware VNF placement using a probability-based approach. *IEEE Transactions on Network and Service Management*, 18(3):2478–2491.

[Xie et al., 2021] Xie, Y., Huang, L., Kong, Y., Wang, S., Xu, S., Wang, X. e Ren, J. (2021). Virtualized network function forwarding graph placing in SDN and NFV-enabled IoT networks: A graph neural network assisted deep reinforcement learning method. *IEEE Transactions on Network and Service Management*, 19(1):524–537.

[Xiong et al., 2020] Xiong, X., Zheng, K., Lei, L. e Hou, L. (2020). Resource allocation based on deep reinforcement learning in IoT edge computing. *IEEE Journal on Selected Areas in Communications*, 38(6):1133–1146.

[Xu et al., 2020a] Xu, S., Xia, Y. e Shen, H.-L. (2020a). Analysis of malware-induced cyber attacks in cyber-physical power systems. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 67(12):3482–3486.

[Xu et al., 2020b] Xu, Z., Zhang, Z., Liang, W., Xia, Q., Rana, O. e Wu, G. (2020b). QoS-aware VNF placement and service chaining for IoT applications in multi-tier mobile edge networks. *ACM Transactions on Sensor Networks (TOSN)*, 16(3):1–27.

[Zhang et al., 2017] Zhang, P., Lu, J., Wang, Y. e Wang, Q. (2017). Cooperative localization in 5G networks: A survey. *Ict Express*, 3(1):27–32.

[Zhang et al., 2021] Zhang, T., Wang, Z., Liu, Y., Xu, W. e Nallanathan, A. (2021). Joint resource, deployment, and caching optimization for AR applications in dynamic UAV NOMA networks. *IEEE Transactions on Wireless Communications*, 21(5):3409–3422.

[Zhao et al., 2021] Zhao, C., Liu, J., Sheng, M., Teng, W., Zheng, Y. e Li, J. (2021). Multi-UAV trajectory planning for energy-efficient content coverage: A decentralized learning-based approach. *IEEE Journal on Selected Areas in Communications*, 39(10):3193–3207.

[Zheng et al., 2018] Zheng, G., Tsiopoulos, A. e Friderikos, V. (2018). Optimal VNF chains management for proactive caching. *IEEE Transactions on Wireless Communications*, 17(10):6735–6748.

[Zhou et al., 2021] Zhou, Z., Jia, Z., Liao, H., Lu, W., Mumtaz, S., Guizani, M. e Tariq, M. (2021). Secure and latency-aware digital twin assisted resource scheduling for 5G edge computing-empowered distribution grids. *IEEE Transactions on Industrial Informatics*, 18(7):4933–4943.