Universidade Federal de Campina Grande

Centro de Engenharia Elétrica e Informática

Programa de Pós-Graduação em Engenharia Elétrica

Marcus Marinho Bezerra

Método para Disponibilização Dinâmica de Serviços Críticos de Saúde em Cenários de Mobilidade

Campina Grande - PB

2024

Universidade Federal de Campina Grande
Centro de Engenharia Elétrica e Informática
Programa de Pós-Graduação em Engenharia Elétrica

Método para Disponibilização Dinâmica de Serviços Críticos de Saúde em Cenários de Mobilidade

Marcus Marinho Bezerra

Tese de Doutorado submetida à Coordenação do Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal de Campina Grande como parte dos requisitos necessários para obtenção do grau de Doutor em Ciências no Domínio da Engenharia Elétrica.

Área de Concentração: Processamento da Informação
Linha de Pesquisa: Engenharia da Computação

Angelo Perkusich, D.Sc.
Danilo Freire de Souza Santos, D.Sc.
(Orientadores)

Campina Grande, Paraíba, Brasil
©Marcus Marinho Bezerra, Maio de 2024

B574m Bezerra, Marcus Marinho.
Método para disponibilização dinâmica de serviços críticos de saúde em cenários de mobilidade / Marcus Marinho Bezerra. – Campina Grande, 2024.
250 f. : il. color.

Tese (Doutorado em Engenharia Elétrica) – Universidade Federal de Campina Grande, Centro de Engenharia Elétrica e Informática, 2024.
"Orientação: Prof. Dr. Angelo Perkusich, Prof. Dr Danilo Freire de Souza Santos".
Referências.

1. Engenharia da Computação. 2. Processamento da Informação. 3. Inteligência Artificial (IA) – Personalização de Serviços Médicos. 4. Tecnologias Imersivas e Inteligentes – Realidade Aumentada (AR). 5. Internet das Coisas (IoT). 6. Saúde 4.0. I. Perkusich, Angelo. II. Santos, Danilo Freire de Souza. III. Título.

CDU 621.391:004(043)

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECÁRIA SEVERINA SUELI DA SILVA OLIVEIRA CRB-15/225

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
POS-GRADUACAO EM ENGENHARIA ELETRICA
Rua Aprigio Veloso, 882, - Bairro Universitario, Campina Grande/PB, CEP 58429-900

REGISTRO DE PRESENÇA E ASSINATURAS

1. ATA DA DEFESA PARA CONCESSÃO DO GRAU DE DOUTOR EM CIÊNCIAS, NO DOMÍNIO DA ENGENHARIA ELÉTRICA, REALIZADA EM 04 DE JUNHO DE 2024

(Nº 375)

CANDIDATO(A): **MARCUS MARINHO BEZERRA.** COMISSÃO EXAMINADORA: ANTONIO MARCUS NOGUEIRA LIMA, Dr., UFCG - Presidente da Comissão e Examinador interno, ANGELO PERKUSICH, D.Sc., UFCG, DANILO FREIRE DE SOUZA SANTOS, D.Sc., UFCG, Orientadores, ALEXANDRE JEAN RENÉ SERRES, D.Sc., UFCG, Examinador Interno, LEANDRO DIAS DA SILVA, Dsc., UFAL, JAIDILSON JÓ DA SILVA, D.Sc., UFCG e AUGUSTO JOSÉ VENÂNCIO NETO, Dr., UFRN, Examinadores Externos. TÍTULO DA TESE: Método para Disponibilização Dinâmica de Serviços Críticos de Saúde em Cenários de Mobilidade. ÁREA DE CONCENTRAÇÃO: Processamento da Informação. HORA DE INÍCIO: **08h00** – LOCAL: Auditório do Embedded e **Sala Virtual, conforme Art. 5º da PORTARIA SEI Nº 01/PRPG/UFCG/GPR, DE 09 DE MAIO DE 2022.** Em sessão pública, após exposição de cerca de 45 minutos, o(a) candidato(a) foi arguido(a) oralmente pelos membros da Comissão Examinadora, tendo demonstrado suficiência de conhecimento e capacidade de sistematização, no tema de sua tese, obtendo conceito APROVADO. Face à aprovação, declara o presidente da Comissão, achar-se o examinado, legalmente habilitado a receber o Grau de Doutor em Ciências, no domínio da Engenharia Elétrica, cabendo a Universidade Federal de Campina Grande, como de direito, providenciar a expedição do Diploma, a que o(a) mesmo(a) faz jus. Na forma regulamentar, foi lavrada a presente ata, que é assinada por mim, Leandro Ferreira de Lima, e os membros da Comissão Examinadora. Campina Grande, 04 de Junho de 2024.

LEANDRO FERREIRA DE LIMA
Secretário

ANTONIO MARCUS NOGUEIRA LIMA, Dr., UFCG
Presidente da Comissão e Examinador Interno

ANGELO PERKUSICH, D.Sc., UFCG
Orientador

DANILO FREIRE DE SOUZA SANTOS, D.Sc., UFCG,

Orientador

ALEXANDRE JEAN RENÉ SERRES, D.Sc., UFCG
Examinador Interno

LEANDRO DIAS DA SILVA, Dsc., UFAL

Examinador Externo

JAIDILSON JÓ DA SILVA, D.Sc., UFCG

Examinador Externo

AUGUSTO JOSÉ VENÂNCIO NETO, Dr., UFRN

Examinador Externo

MARCUS MARINHO BEZERRA

Candidato

2 - APROVAÇÃO

2.1. Segue a presente Ata de Defesa de Tese de Doutorado da candidato **MARCUS MARINHO BEZERRA**, assinada eletronicamente pela Comissão Examinadora acima identificada.

2.2. No caso de examinadores externos que não possuam credenciamento de usuário externo ativo no SEI, para igual assinatura eletrônica, os examinadores internos signatários **certificam** que os examinadores externos acima identificados participaram da defesa da tese e tomaram conhecimento do teor deste documento.

Documento assinado eletronicamente por **LEANDRO FERREIRA DE LIMA**, **SECRETÁRIO (A)**, em 04/06/2024, às 15:50, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de 2018.

Documento assinado eletronicamente por **DANILO FREIRE DE SOUZA SANTOS**, **PROFESSOR(A) DO MAGISTERIO SUPERIOR**, em 04/06/2024, às 15:58, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de 2018.

Documento assinado eletronicamente por **ALEXANDRE JEAN RENE SERRES**, **PROFESSOR(A) DO MAGISTERIO SUPERIOR**, em 04/06/2024, às 16:07, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de 2018.

Documento assinado eletronicamente por **ANTONIO MARCUS NOGUEIRA LIMA**, **PROFESSOR(A) DO MAGISTERIO SUPERIOR**, em 04/06/2024, às 16:08, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de 2018.

Documento assinado eletronicamente por **JAIDILSON JO DA SILVA**, **PROFESSOR(A) DO MAGISTERIO SUPERIOR**, em 04/06/2024, às 16:21, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de 2018.

Documento assinado eletronicamente por **ANGELO PERKUSICH**, **PROFESSOR(A) DO MAGISTERIO SUPERIOR**, em 04/06/2024, às 16:23, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de 2018.

Documento assinado eletronicamente por **Leandro Dias da Silva**, **Usuário Externo**, em 04/06/2024, às 21:14, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de 2018.

Documento assinado eletronicamente por **Marcus Marinho Bezerra**, **Usuário Externo**, em 05/06/2024, às 09:02, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de 2018.

A autenticidade deste documento pode ser conferida no site https://sei.ufcg.edu.br/autenticidade, informando o código verificador **4488342** e o código CRC **D503F134**.

Resumo

A implementação de tecnologias imersivas e inteligentes, como realidade aumentada (AR) e inteligência artificial (IA), é crucial na personalização de serviços médicos, permitindo diagnósticos precisos e tratamentos individualizados dentro do contexto da Saúde 4.0 e Internet das Coisas (IoT). Nos cenários de emergência, a aplicação de tecnologias avançadas em ambulâncias, incluindo teleatendimento, diagnósticos rápidos e transmissão de dados vitais em tempo real, é essencial. Essas tecnologias melhoram o apoio aos socorristas e são capazes de reduzir os tempos de resposta dos atendimentos emergenciais. Contudo, essa integração enfrenta desafios complexos, como limitações de recursos computacionais e autonomia dos dispositivos móveis, onde a computação na borda surge como alternativa para suprir os requisitos dessas aplicações. No entanto, a transferência contínua de serviços entre servidores, devido à mobilidade das ambulâncias, introduz desafios adicionais devido às variações de latência e largura de banda da rede, além da necessidade de otimização da utilização de recursos computacionais. Desse modo, neste trabalho de tese é apresentado um método para a disponibilização dinâmica de serviços críticos. Experimentos para validação de uma Prova de Conceito demonstram que o método consegue atender requisitos específicos de latência, taxa de transmissão, otimização de recursos computacionais e mobilidade, viabilizando uma alocação ótima.

Abstract

The implementation of immersive and intelligent technologies, such as augmented reality (AR) and artificial intelligence (AI), is crucial in customizing medical services, enabling precise diagnostics and individualized treatments within the context of Health 4.0 and the Internet of Things (IoT). In emergency scenarios, the application of advanced technologies in ambulances, including telehealth, rapid diagnostics, and real-time vital data transmission, is essential. These technologies enhance support for first responders and are capable of reducing response times in emergency care. However, this integration faces complex challenges, such as limitations in computational resources and the autonomy of mobile devices, where edge computing emerges as an alternative to meet the requirements of these applications. Nevertheless, the continuous transfer of services between servers, due to the mobility of ambulances, introduces additional challenges due to variations in network latency and bandwidth, as well as the need for optimization of computational resource utilization. Thus, in this thesis work, a method for the dynamic provisioning of critical services is presented. Experiments to validate a Proof of Concept demonstrate that the method can meet specific requirements for latency, transmission rate, computational resource optimization, and mobility, enabling optimal allocation.

Agradecimentos

Agradeço a Deus por ter sempre guiado meu caminho.

Aos meus pais, Claudenor e Eliane, por minha vida.

À minha esposa Layssa Melo, pelo incentivo e por entender os momentos de ausência para dedicação ao doutorado. Agradeço também pela companhia e pelos momentos de descontração durante toda a trajetória.

Aos meus irmãos, Patrícia e Claudenor Júnior, ao meu cunhado Thiago e minha cunhada Uilma Diniz, pelo incentivo e pela força.

Aos meus sobrinhos e afilhada Humberto, Anna Sophia, Clarice, Maria Elis e Anna Luiza, por todas as brincadeiras nos momentos de descontração.

Aos meus orientadores, Professores Angelo Perkusich e Danilo Santos, pelas discussões, direcionamentos e sugestões durante o desenvolvimento deste trabalho, sempre estiveram disponíveis quando precisei de suporte.

Aos professores da banca examinadora, Antonio Lima (UFCG), Jaidilson Silva (UFCG), Alexandre Serres (UFCG), Augusto José (UFRN), Leandro Dias (UFAL) e Dalton Valadares (IFPE) pelas críticas e sugestões.

Ao Professor Dalton Valadares e aos colegas de doutorado Romulo Omena e Daniel Macedo, pelo suporte e troca de conhecimentos nas nossas reuniões de acompanhamento semanal.

Aos amigos, Marcus Vinicius, Vitor Ribeiro, Andrei Patriota, pelos inúmeros momentos de aprendizado e descontração, pelo apoio recebido e suporte oferecido.

Ao VIRTUS UFCG - Núcleo de Pesquisa, Desenvolvimento e Inovação em Tecnologia da Informação, Comunicação e Automação pela rede de apoio, em especial Emyle Ferreira, Igor Silva e Mateus Maximo.

Aos servidores da COPELE-UFCG, pelo apoio administrativo.

Por fim, agradeço à CAPES pelo suporte financeiro durante a execução desse mestrado, no período enquanto bolsista.

Agradeço a todos.

Conteúdo

Lista de Símbolos, Abreviaturas e Acrônimos

5G – *Quinta Geração de Redes Móveis*

AR – *Realidade Aumentada*

AWS – *Amazon Web Services*

D2D – *Device-to-Device Communications*

GCP – *Google Cloud Platform*

IA – *Inteligência Artificial*

IoT – *Internet das Coisas*

kml – *Keyhole Markup Language*

MiB – *Mebibyte*

NFV – *Virtualização de Funções de Rede*

NS – *Network Slicing*

QoE – *Qualidade da Experiência*

QoS – *Qualidade do Serviço*

SFCs – *Funções de Cadeia de Serviço*

SDN – *Redes Definidas por Software*

TIC – *Tecnologias da Informação e Comunicação*

VMs – *Máquinas Virtuais*

YAML – *YAML Ain't Markup Language*

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Na contemporânea era da Internet das Coisas (IoT – *Internet of Things*), é notória a adoção de tecnologias emergentes orientadas à solução de problemas de negócios e melhoria da experiência dos usuários, promovendo o processo de transformação digital em diferentes setores, incluindo agricultura, educação, saúde, transporte, indústria, além de cidades e residências inteligentes [2–5]. Este processo envolve a modernização dos sistemas de tecnologias da informação e comunicação (TIC) e o desenvolvimento de novas estratégias, fomentando a tomada de decisões baseada em dados para impulsionar a inovação e a eficiência [6–9].

Na área da saúde, a digitalização de serviços emerge como uma estratégia crucial, visando não apenas a eficiência operacional, mas também a promoção de uma abordagem centrada no paciente [10]. A utilização de TICs para apoiar a saúde e suas áreas relacionadas é comumente conhecida por *e-Health*, do inglês *electronic health* [11]. Com o avanço da digitalização dos serviços médicos, novos cenários podem ser explorados, podendo apresentar requisitos de comunicação e computação mais restritos, principalmente em aplicações críticas, relacionadas à prestação de cuidados para a sobrevivência de pacientes [1], servindo como catalizador para o desenvolvimento de soluções inovadoras da Saúde 4.0 (do inglês, *Healthcare 4.0*).

De modo geral, a Saúde 4.0 integra tecnologias da Indústria 4.0, como inteligência artificial (IA), *Big Data Analytics*, Internet das Coisas (IoT), robótica e realidade aumentada (AR), no setor de saúde, possibilitando um atendimento personalizado em tempo-real [12]. Ao adotar, por exemplo, a telemetria de dados provenientes de redes de biossensores para monitoramento contínuo dos sinais vitais e detecção precoce de doenças [13], esse novo pa-

radigma não só melhora a qualidade do cuidado médico, mas também reduz custos e eleva a satisfação do paciente através da implementação de serviços inteligentes e sistemas avançados [14, 15].

Particularmente, em situações de urgência e emergência, a prestação de cuidados nos instantes iniciais do atendimento são determinantes para a sobrevivência dos pacientes [16]. A troca eficiente de informações entre paramédicos em serviço e equipes hospitalares é crucial para o atendimento de emergência, ocorrendo tipicamente apenas com a chegada da ambulância ao hospital [17]. Neste sentido, a utilização de soluções inteligentes baseadas em IoT, realidade aumentada (AR) e inteligência artificial (IA), apoiadas por redes de comunicações móveis de alta confiabilidade e sistemas de computação distribuídos, apresenta-se como uma oportunidade para melhoria de cenários críticos de saúde.

Conforme apresentado nos trabalhos de Schinle *et al.* [18] e de Martinez-Suarez e Alvarado-Serrano [19], novas tecnologias podem ser integradas a ambulâncias visando apoiar médicos emergencistas através do teleatendimento, rápido diagnóstico assistido por sistemas de visão computacional e a transmissão de dados de sinais vitais de pacientes em tempo-real para unidades de atendimento remotas. A integração de sistemas médicos avançados em veículos de emergência representa um progresso importante para diminuir o tempo de resposta no tratamento dos pacientes, um fator crítico onde cada segundo conta para a sobrevivência [20–22]. Essa abordagem viabiliza o desenvolvimento de 'ambulâncias conectadas', que prometem transformar o cenário do atendimento médico de urgência [23].

No entanto, para o desenvolvimento de soluções que promovam a Saúde 4.0, a personalização da prestação de serviços médicos depende fortemente da implementação de sistemas médicos inteligentes, de baixo consumo energético e adaptáveis a diferentes ambientes e necessidades [24, 25]. Um exemplo disso é a adoção de óculos inteligentes (*smart glasses*) em serviços médicos de emergência, como no caso de 'ambulâncias conectadas'. Esta integração representa uma grande inovação tecnológica, sendo capaz de melhorar a qualidade do serviço de emergência conforme apresentado nos trabalhos de Apiratwarakul *et al.* [26] e Thaijiam [27].

Com essas soluções tecnológicas, aproveita-se principalmente os recursos de realidade aumentada (AR) e inteligência artificial como ferramentas de suporte à tomada de decisões e orientações operacionais para casos críticos de saúde [28–30]. Esses sistemas exigem uma

capacidade de processamento e de transmissão de dados elevada para realizar a gestão, análise e o envio de informações. Além disso, o baixo tempo de resposta necessário para atender aos rigorosos requisitos de comunicação impostos por essas aplicações exige novas soluções frente às atuais, onde novas redes móveis de próxima geração e sistemas de Computação em Borda se apresentam como tecnologias promissoras [31].

A Computação em Borda é definida como um paradigma computacional onde o processamento de dados é realizado na borda da rede, próximo à fonte dos dados [32, 33]. A utilização dessa abordagem tem como foco a redução da latência e do uso de largura de banda ao processar os dados em ambientes mais próximos dos dispositivos, em vez de enviá-los para ambientes de computação centralizados, como os de Computação em Nuvem [34]. Isso permite o processamento local de dados, reduzindo o consumo de energia e melhorando a qualidade do serviço [35–38].

Ao tempo que a concepção de novas soluções visando a melhoria e a personalização da prestação de serviços médicos vão surgindo, alguns desafios são potencializados à medida que a criticidade das aplicações aumenta. Na grande maioria dos casos, essas aplicações demandam baixo tempo de resposta, que por sua vez podem ser determinantes para o desfecho do atendimento de paciente em situações de urgência e emergência [39–41]. Neste âmbito, o cenário de 'ambulâncias conectadas' tem o potencial de melhorar a eficiência operacional do atendimento de emergência, que pode ter impacto direto no estado de saúde de pacientes. Por exemplo, através de soluções imersivas sendo disponibilizadas por meio de óculos inteligentes, o paramédico poderá obter informações que incluem desde procedimentos operacionais simples, como a identificação de materiais e equipamentos médicos no interior da ambulância, até a visualização de sinais vitais do paciente e obtenção de diagnósticos assistidos por modelos de Visão Computacional e Inteligência Artificial.

Contudo, atender os requisitos desse tipo de aplicação se torna mais desafiador diante da necessidade de mobilidade das ambulâncias e considerando a limitação de recursos computacionais dos dispositivos imersivos. A utilização de técnicas de *offloading* para execução de tarefas em ambientes com maior capacidade computacional é uma abordagem bastante relevante e atual. Porém, ainda devem ser considerados aspectos como a dinamicidade das aplicações e a Qualidade da Experiência (QoE) dos usuários, nesse caso os paramédicos. Estudos recentes abordam a alocação dinâmica de serviços em ambientes de computação

distribuídos como uma a estratégia para redução da latência e aumento da autonomia dos dispositivos, mas ainda de maneira reativa, ou seja, dada a degradação da Qualidade do Serviço (QoS).

Nesse contexto, promover o desenvolvimento de soluções capazes de disponibilizar serviços dinamicamente em ambiente de Computação na Borda e redes móveis de próxima geração torna-se uma tarefa fundamental para a melhoria do atendimento de emergências em cenários de mobilidade assistidos por soluções de Saúde 4.0.

1.1 Justificativa

A mobilidade representa um aspecto crucial no contexto das ambulâncias conectadas, transcendendo a mera locomoção física e abrangendo a transmissão contínua e em tempo-real de dados médicos [42]. A garantia da continuidade dos serviços de saúde durante o deslocamento das unidades móveis de atendimento exige um robusto suporte à mobilidade de rede, como discutido por Siriwardhana *et al.* [43]. Desafios como a fraca qualidade de sinal, a atenuação causada pelas estruturas veiculares, frequentes transferências de conexão e interrupções do sinal são exacerbados pela mobilidade.

Nos cenários de emergência, a implementação de algoritmos inteligentes atrelados ao uso de óculos inteligentes imersivos, impulsionados pelos avanços em IA e AR, pode representar um avanço significativo, abrindo um leque de possibilidades para melhorar o cuidado ao paciente em cenários críticos, que incluem desde análise de dados em tempo-real até diagnósticos preditivos e suporte à decisão. Eles oferecem o potencial de aprimorar a maneira que os profissionais de emergência avaliam, diagnosticam e iniciam tratamentos para pacientes em trânsito.

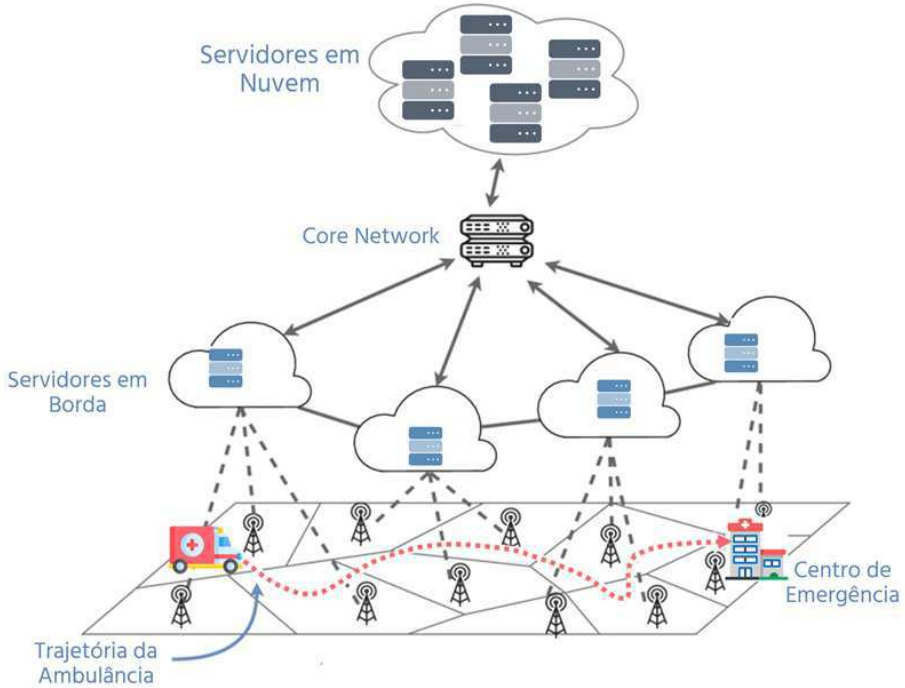
No contexto de mobilidade, a adoção em larga escala de soluções baseadas em de dispositivos de realidade aumentada, com os óculos inteligentes, ainda apresenta alguns desafios a serem superados. De acordo com Siriwardhana *et al.* [43], os dispositivos móveis de realidade aumentada apresentam limitações em sua capacidade computacional para executar tarefas de visão computacional para aplicações imersivas. Outro aspecto diretamente relacionado ao cenário de mobilidade e aplicações imersivas é o consumo de energia de dispositivos móveis. Segundo Chen *et al.* [44], o sistema imersivo como um todo precisa ser projetado

visando mitigar o consumo excessivo de energia, uma vez que isto afeta diretamente na autonomia do uso dos dispositivos.

Diante desses desafios, algumas possibilidades são destacadas na literatura como soluções. Dentre elas está a realização do *offloading* de serviços para servidores de computação em borda, conforme apontado por [43]. No entanto, a natureza distribuída da geolocalização, a heterogeneidade e as limitações de recursos inerentes à Computação em Borda, em comparação com a Computação em Nuvem, intensificam a complexidade da alocação de serviços. Serviços devem ser dinamicamente alocados e desalocados em servidores ao longo do percurso, adaptando-se ao padrão de mobilidade do veículo [45].

Dada essa mudança de paradigma da Saúde 4.0 em se ter a disponibilização de serviços médicos inteligentes e imersivos, encontra-se nesse cenário uma oportunidade de pesquisa no campo das 'ambulâncias conectadas', no que se refere à alocação inteligente de serviços em servidores de borda, sendo este o caso de uso motivador escolhido para o desenvolvimento desta tese. Na Figura 1.1 está ilustrado o diagrama para este cenário.

Figura 1.1: Diagrama do cenário de Ambulâncias Conectadas.



Fonte: Adaptado de [46].

Neste cenário, um dos potenciais problemas identificados está relacionado à disponibilização de serviços em servidores ao longo do trajeto, ainda atendendo os requisitos de

comunicação impostos para esse tipo de aplicação, que podem demandar serviços imersivos e inteligentes, incluindo a transmissão e processamento dos mais diversos tipos de dados, podendo-se citar, *streaming* de áudio e vídeo para assistências de especialista remoto, transmissão de dados de sinais vitais, transmissão de imagens de equipamentos de diagnóstico, entre outros.

Com isso, busca-se a utilização eficiente recursos computacionais, visando atender ainda diferentes requisitos de Qualidade de Serviço (do inglês, *Quality of Service* - QoS), visto que a alocação dos serviços pode ser realizada a partir de algoritmos inteligentes que adaptam os recursos disponíveis de acordo com a demanda e a criticidade dos casos atendidos, promovendo o desenvolvimento de soluções de Saúde 4.0. Na Figura 1.2 estão ilustradas diversas aplicações inovadoras no contexto da Saúde 4.0, que incluem:

- **Rastreamento de Doenças:** aplicações baseadas em IA e *Big Data Analytics* para classificação de grande volume de dados e análise de ocorrências de anomalias em tempo-real;
- **Rápida Triagem:** identificação rápida do paciente por meio de técnicas de reconhecimento facial e autorização de acesso a dados de histórico de pacientes;
- **Ambulâncias Conectadas:** soluções baseadas em realidade aumentada, inteligência artificial e IoT para suporte na identificação de quadros de saúde do paciente e suporte à tomada de decisão;
- **Cirurgia Robótica Teleoperada:** uso de robôs controlados remotamente que permitem realizar procedimentos cirúrgicos minuciosos à distância, com precisão aumentada e menor risco para os pacientes;
- **Diagnóstico Rápido:** implementação de algoritmos de aprendizado de máquina para análise instantânea de imagens médicas e testes laboratoriais, proporcionando diagnósticos precisos e acelerados;
- **Telemedicina:** streaming de vídeo com adição de ferramentas de diagnóstico em ambiente virtualizado;

- **Treinamento Imersivo:** uso de realidade virtual e simuladores avançados para treinar profissionais de saúde em procedimentos complexos, melhorando habilidades sem riscos para pacientes reais.

Figura 1.2: Casos de uso para aplicações de Saúde 4.0.

Fonte: Produzida pelo autor.

Em suma, o desenvolvimento de soluções para a Saúde 4.0 promove uma mudança de paradigma na prestação de cuidados médico. Contudo, avanços relacionados à infraestrutura de comunicação e computação são necessários para atender os requisitos impostos por essas aplicações.

1.2 Problemática

Considerando os cenários apresentados anteriormente, desafios específicos relacionados ao uso de 'ambulâncias conectadas', principalmente relacionados à disponibilização de serviços de computação e limitações de recursos dos dispositivos imersivos, podem ser identificados através da revisão bibliográfica. Os principais aspectos incluem requisitos específi-

cos de latência, taxa de transmissão, otimização de recursos computacionais e mobilidade, identificando-se a necessidade de se apresentar métodos processuais que viabilizem uma alocação ótima.

Métodos processuais na computação referem-se a abordagens e técnicas que utilizam procedimentos específicos ou algoritmos para gerar conteúdo de maneira automática e dinâmica, como em geração de conteúdo procedural em jogos ou em outros contextos similares [47]. No cenário de emergência, por exemplo, a adoção de métodos processuais pode ser utilizada para disponibilização de serviços virtualizados em infraestruturas de computação distribuídas mediante o contexto das aplicações. Atender aos requisitos necessários desses serviços implica na obtenção de ganhos significativos para o serviço de emergência, como a melhoria na eficiência da resposta médica, aumento da precisão no diagnóstico em situações críticas, e otimização da coordenação entre equipes de resgate e hospitais.

O tempo de disponibilização desses serviços é um fator a ser considerado, pois permite uma resposta mais ágil em situações de emergência, facilitando a comunicação e o suporte à tomada de decisões por parte dos profissionais de saúde. De modo geral, estes serviços precisam ser escaláveis e dinâmicos, adaptando-se rapidamente a diferentes cenários e aumentando sua capacidade conforme a demanda aumenta. Neste sentido, a utilização de arquiteturas monolíticas não são adequadas para atender a dinâmica e velocidade exigidas na disponibilização de serviços críticos, conforme estudo comparativo entre microsserviços e arquiteturas monolíticas apresentado por Benavente *et al.* [48].

Tecnologicamente, estes serviços são implementados utilizando conteineres Docker, uma abordagem de virtualização eficiente para empacotar e distribuir aplicações de forma isolada e consistente. A disponibilização desses conteineres é gerenciada através da ferramenta de orquestração Kubernetes, que permite automatizar a implantação, o dimensionamento e a gestão das aplicações. Essa configuração oferece flexibilidade e escalabilidade, facilitando a adaptação rápida aos requisitos dinâmicos de cenários de emergência médica. Além disso, Kubernetes promove uma maior disponibilidade e resiliência dos serviços, essenciais para manter as operações críticas funcionando sem interrupções.

Devido à sua alta demanda de processamento e necessidade de ultra-baixa latência, as aplicações de realidade aumentada são especialmente adequadas para serem executadas em servidores em borda, a fim de assegurar a qualidade dos serviços prestados [49]. Ren *et*

al. [50] propuseram uma arquitetura baseada em borda para diminuir atrasos e consumo energético em AR, em um cenário com múltiplos dispositivos móveis operando simultaneamente sob uma mesma estação rádio-base. Além disso, pesquisas recentes indicam que AR demanda baixíssimos tempos de resposta, não alcançados pela Computação em Nuvem, destacando a inadequação desta última para aplicações de AR [51–55].

Um outro fator crucial no contexto do atendimento de emergência é a mobilidade, que impõe desafios para a infraestrutura de comunicação e computação atuais, especialmente em termos de conectividade estável em áreas com cobertura de rede irregular [56]. A mobilidade é identificada como um desafio significativo na implementação de soluções de Computação em Borda, conforme sugerido por diversos estudos [57–61]. Além disso, é crucial aprimorar a mobilidade dos dados para aumentar a eficiência de aplicações críticas que operam com dados distribuídos geo-distribuídos.

Talpure e Mohan [62] destacam que soluções de alocação estáticas de serviços não são eficazes em cenários de mobilidade. Para isso, os autores propuseram um *framework* para alocação dinâmica de serviços em redes veiculares baseado em aprendizado por reforço (do inglês, *reinforcement learning*), com o objetivo de encontrar a alocação ideal de serviços nos servidores de borda considerando a mobilidade e a dinâmica de veículos para solicitações de diferentes tipos de serviços.

No trabalho de Malazi *et al.* [63] foi realizada uma revisão sistemática de literatura sobre alocação dinâmica de serviços em ambientes de Computação em Borda. Dentre os desafios em aberto associados à alocação dinâmica de serviços de aplicação estão o *cache* parcial de serviços e a dependência entre serviços em arquiteturas de microsserviço. Além disso, é destacado que, como diferentes aplicações precisarão atender a diferentes requisitos de Qualidade de Serviço (do inglês, *Quality of Service* - QoS), novas técnicas de alocação, implantação e execução dinâmica de serviços com capacidade aprimorada, controle de tráfego inteligente e soluções de direção terão que ser desenvolvidas e integradas para atender a requisitos rigorosos.

A migração contínua e fluida de uma aplicação entre diferentes servidores, mesmo com o usuário em movimento, é um mecanismo essencial conhecido como *seamless service delivery* [64]. No contexto de mobilidade, alcançar essa entrega ininterrupta é desafiador, pois a mobilidade afeta diretamente vários parâmetros da rede, como latência, largura de banda,

atraso e jitter, levando à deterioração do desempenho da aplicação [65].

A alocação de serviços virtualizados envolve a provisão e gestão de recursos computacionais para aplicações e dispositivos variados. Neste cenário, é essencial que a infraestrutura de TI aloque recursos para ambulâncias conectadas, suportando necessidades de processamento em tempo-real e realidade aumentada. A alocação eficaz em todos os servidores de borda que a ambulância se conectar ao trafegar deve considerar a disponibilidade de recursos, demanda variável, proximidade de dispositivos de borda e latência de rede, especialmente porque a demanda pode flutuar significativamente em casos de emergências.

Huang *et al.* [66] exploraram a eficiência de um sistema de computação em borda para serviços móveis de baixa latência e alta confiabilidade, mostrando que sistemas estáticos, como servidores únicos, não são suficientes para atender cenários críticos. Qiao *et al.* [67] discutem os desafios de minimizar a latência em AR, que incluem aprimoramento de algoritmos de vídeo e alocação dinâmica de recursos, ainda desafiados por limitações de tecnologia de redes móveis de nova geração.

Adicionalmente, no cenário de mobilidade em um ambiente de computação em múltiplas bordas, a replicação de serviços em diferentes servidores ao longo do trajeto da ambulância seria um caminho de solução possível. No entanto, apesar dessa abordagem atender aos requisitos de aplicações críticas e inteligentes, essa replicação gera desperdícios de recursos computacionais. Ao conhecer-se o padrão de mobilidade da ambulância, obtêm-se ganhos diretos associados à utilização eficiente de recursos computacionais, além dos benefícios citados anteriormente no âmbito da qualidade dos serviços médicos prestados. Portanto, a questão central é como disponibilizar dinamicamente serviços críticos de saúde virtualizados baseados em tecnologias de IA e AR, assegurando a qualidade de serviço (QoS) face à demanda e disponibilidade de recursos variáveis considerando a necessidade de mobilidade da ambulância.

Pesquisas recentes têm explorado a alocação dinâmica de serviços como método para reduzir a latência em ambientes de computação em múltiplas bordas. No trabalho apresentado por Panek et al. [68] foram abordadas estratégias de otimização baseadas na localização do usuário e requisitos da aplicação, visando melhorar a Qualidade do Serviço (QoS) e a eficiência no uso dos recursos de infraestrutura. Esta abordagem, trata-se de uma abordagem reativa, pois a estratégia de migração para o próximo nó de borda só é iniciada após a

detecção de mudanças na localização do usuário ou entre estações de rádio-base.

Além disso, Jiang et al. [69] e Vieira et al. [70] abordam a alocação de serviços com enfoque em aprendizado de reforço profundo e ajuste das Funções de Cadeia de Serviço (SFCs) em resposta à mobilidade do usuário. Contudo, ainda destaca-se que estas abordagens também operam de forma reativa, onde a alocação dos serviços ocorre somente após o reconhecimento de alterações na mobilidade ou no desempenho do usuário.

Com isso, os recursos e serviços só são disponibilizados em servidores de borda após a identificação de requisições provenientes da ambulância com o novo servidor. Conforme mencionado anteriormente, o *overhead* necessário para alocação de infraestrutura virtualizada e disponibilização dos serviços, acarreta que a ambulância, ao menos em um tempo inicial, fique desprovida dos serviços críticos necessários para o atendimento ao paciente no interior da ambulância em deslocamento para o hospital mais próximo. A depender de aplicação, esse intervalo de desassistência pode inviabilizar seu funcionamento adequado, considerando os rígidos requisitos de disponibilidade de recursos e serviços na borda. No mais, essa desassistência aconteceria sempre que conecta-se a uma nova borda. Ou seja, as aplicações críticas estariam sujeitas a falhas em diversas ocasiões ao longo do trajeto.

Já existem tecnologias e arcabouços que viabilizam o cenário proposto, entretanto, considerando os requisitos específicos de latência, taxa de transmissão, otimização de recursos computacionais e mobilidade, identifica-se a necessidade de se apresentar métodos processuais que viabilizem uma alocação ótima. Do ponto de vista do cenário e temática, alocação ótima significa a distribuição de recursos de modo a garantir a satisfação dos requisitos, visando cenários de *Always Best Connected and Best Served* [71].

Para direcionar este trabalho, o problema pode ser delineado de duas formas distintas:

- **Problema de Negócios** - definir um método para preparar uma infraestrutura virtualizada que viabilize a alocação proativa e dinâmica de serviços críticos de saúde em ambientes de computação de múltiplas bordas, visando a continuidade na disponibilização desses serviços.

- **Problema Técnico** - investigar, definir e implementar um método para alocação dinâmica de serviços e integrar uma infraestrutura virtualizada para validação da arquitetura em cenários de mobilidade, capaz de suportar a alocação dinâmica de serviços,

considerando a previsão e análise dos padrões de mobilidade, garantindo o cumprimento dos requisitos críticos de uma aplicação, mesmo havendo mudança de conexão de servidores de borda.

1.3 Objetivo geral

O objetivo geral neste trabalho é investigar, projetar e desenvolver um método para disponibilização proativa e dinâmica de serviços em ambientes de computação de múltiplas bordas, visando suprir requisitos de aplicações críticas de saúde, considerando o seu padrão de mobilidade.

1.3.1 Objetivos específicos

Para alcance do objetivo geral, os objetivos específicos são:

- Criar uma solução arquitetural para infraestrutura virtualizada, visando a alocação dinâmica e proativa de serviços críticos, considerando a gestão de mobilidade e a prioridade dos serviços;
- Implementar estruturas de *templates* visando o carregamento dinâmico de containers em infraestrutura virtualizada
- Investigar e implementar mecanismos de disponibilização dinâmica de serviços, utilizando tecnologias de virtualização e orquestração de containers, com foco na eficiência e na resposta rápida necessária para aplicações críticas de saúde em ambientes de computação em múltiplas bordas;
- Pesquisar, definir e configurar um ambiente virtual para simular a mobilidade de uma 'ambulância conectada', incluindo sua conectividade com redes móveis de nova geração e alocação de serviços em múltiplas bordas para uma região específica;
- Desenvolver algoritmos para identificação do padrão de mobilidade, que serão utilizados para a tomada de decisão da alocação proativa de recursos e serviços
- Elaborar cenários para testar e validar o método proposto.

- Criar cenários para validação do método;
- Validar o método utilizando um ambiente de simulação de mobilidade e ferramenta de orquestração de conteires.

1.4 Contribuição

A principal contribuição nesta tese é o desenvolvimento de um método para disponibilização proativa e dinâmica de serviços críticos de saúde em ambientes de computação de múltiplas bordas a partir do padrão de mobilidade do usuário. Este método define uma maneira de executar os procedimentos para a disponibilização de recursos considerando os requisitos de aplicação e mobilidade, onde mobilidade significa mudança de trajetória do usuário. O cenário motivador está na utilização de óculos e dispositivos médicos inteligentes, baseado nas tecnologias de AR, IoT e IA, onde os usuários são os paramédicos prestando assistência médica de urgência no interior de ambulâncias conectadas em movimento.

De modo geral, a utilização de métodos que contemplam a estratégia de alocação proativa e dinâmica em ambientes de computação na borda tem capacidade de atender aos estritos requisitos de comunicação, enquanto otimiza a utilização de recursos computacionais em ambientes de computação na borda distribuídos ao longo do trajeto. Esta solução não só viabiliza o desenvolvimento de soluções imersivas, orientadas à experiência dos usuários (UX), mas também representa uma mudança de paradigma que é capaz de melhorar a qualidade dos serviços de urgência e emergência médicas, através do rápido diagnóstico, recomendação de procedimentos médicos e orientações operacionais assistidos por sistemas inteligentes.

Para tanto, foi criada uma solução para infraestrutura virtualizada orientada à alocação dinâmica e proativa de serviços, que se baseia no padrão de movimento e rota da ambulância, considerando a prioridade dos seus serviços. Além disso, foram implementados *templates* e Provas de Conceito visando o carregamento dinâmico de containers em infraestrutura virtualizada com a utilização do método proposto, viabilizando a alocação dinamicamente dentro dos patamares de requisitos para serviços críticos.

De acordo com os resultados de experimentos realizados, a alocação dinâmica de serviços baseada no padrão de mobilidade da ambulância surge como uma solução eficaz para mitigar problemas de atrasos e perda de pacotes em redes de comunicação, que frequente-

mente ocorrem devido à degradação do sinal da rede. Desse modo, baseado na problemática apresentada previamente, o método de alocação dinâmica de serviços, denominado "*Make Way*", se baseia no padrão de mobilidade da ambulância para disponibilizar os serviços críticos de saúde dinâmica e proativamente. A ideia central consiste em uma analogia à expressão usada em inglês para pedir que os condutores abram espaço e permitam a passagem de ambulâncias em locais com tráfego intenso, garantindo a disponibilidade de recursos e serviços durante a conexão com todos os servidores de borda que se conectar.

O termo é uma metáfora para o processo de liberação de recursos de servidores de borda ocupados por aplicações de menor prioridade para garantir uma transmissão de dados rápida e eficiente para os serviços de emergência, especificamente ambulâncias. O algoritmo monitora a trajetória da ambulância em tempo real e antecipa a alocação de serviços nos próximos servidores de borda conforme a ambulância se move. Ele estima o padrão de mobilidade entre servidores de borda que a ambulância estará seguindo, sendo possível disparar ações de alocação e liberação de recursos e serviços baseado na demanda das aplicações imersivas e inteligentes demandadas no cenário de 'ambulâncias conectadas', "abrindo o caminho", caso os recursos disponíveis estejam sendo utilizados por serviços de menor prioridade. Para isso, foram utilizadas eurísticas que consideram a velocidade da ambulância, o padrão de deslocamento e o tempo necessário para que os serviços estejam disponíveis assim que a ambulância realize a transição para o próximo servidor de borda.

Assim, a estratégia do método "*Make Way*"implica em ajustar dinamicamente e de forma proativa a alocação de serviços baseando-se no padrão de mobilidade da ambulância, priorizando a comunicação e a alocação em rotas críticas e momentos de alta demanda. Isso viabiliza a alocação de serviços críticos conforme padrão de mobilidade, reduzindo o impacto de atrasos provenientes do tempo necessário para que os serviços estejam disponíveis, o que é essencial para a prestação de atendimento de urgência e emergência de modo mais eficiente. Nesse sentido, foram realizadas simulações em diferentes cenários, incluindo cenários que inclui um cenário prático de ambulâncias conectadas, onde o método pode ser implementado e avaliado.

1.4.1 Publicações

Durante o desenvolvimento do sistema apresentado nesta tese, os seguintes trabalhos foram publicados:

1. de Alencar, A.V., Bezerra, M.M., Valadares, D.C.G., Santos, D.F.S., Perkusich, A. (2023). *An Interoperable Microservices Architecture for Healthcare Data Exchange*. In: Barolli, L. (eds) Advanced Information Networking and Applications. AINA 2023. Lecture Notes in Networks and Systems, vol 655. Springer, Cham.

2. Macedo, D.E., Bezerra, M.M., Santos, D.F.S., Perkusich, A. (2023). *Orchestrating Fog Computing Resources Based on the Multi-dimensional Multiple Knapsacks Problem*. In: Barolli, L. (eds) Advanced Information Networking and Applications. AINA 2023. Lecture Notes in Networks and Systems, vol 654. Springer, Cham.

Além disso, o seguinte trabalho foi submetido e está aguardando o processo de revisão:

1. Bezerra, M. M.; Macedo, D.; Valadares, D.; Santos, D.; Perkusich, A. (2024). *Critical Health Applications in Connected 5G Edge Computing Scenarios: a Review*. In: IEEE/CAA Journal of Automatica Sinica.

1.5 Metodologia

Na Figura 1.3 está apresentado o fluxograma da metodologia utilizada. A partir da numeração do fluxograma, a metodologia foi dividida nas etapas descritas nos próximos parágrafos.

A metodologia foi conduzida inicialmente por uma etapa de estudo (Etapa 1). Nesta etapa, foram investigadas características de aplicações críticas de saúde, assim como os requisitos de comunicação associados. Além disso, foi necessário avaliar os requisitos associados à orquestração de serviços.

Na Etapa 2, buscou-se dar ênfase à investigação de soluções de alocação de serviços em ambientes de computação distribuídos, uma vez que sistemas de Computação na Borda em redes 5G, que podem ser utilizados para atender os requisitos de comunicação impostos pelas aplicações críticas de saúde, são sistemas distribuídos.

Visando explorar o problema de alocação de serviços para o cenário de Ambulâncias Conectadas na Etapa 3 foi definido e desenvolvido o ambiente de experimentação. Esta etapa incluiu a definição e desenvolvimento de ferramentas de comunicação e orquestração de serviços para avaliação e execução dos cenários de mobilidade.

Na Etapa 4, as ferramentas desenvolvidas e o ambiente experimental montado na etapa anterior foram utilizados em experimentos para analisar o impacto da mobilidade na alocação de serviços em cenários críticos de saúde.

A partir da avaliação dos resultados da etapa anterior, na Etapa 5, foi desenvolvido o método para disponibilização dinâmica de serviços em cenários críticos de saúde que apresentam a necessidade de mobilidade.

Na etapa 6, foi realizada a validação da solução desenvolvida. Onde foram definidos os cenários-alvo de teste, de modo que foi possível observar e avaliar o desempenho da solução.

Por fim, na Etapa 7, foi elaborada a documentação deste trabalho, assim como submissão dos resultados para a publicação.

1.6 Organização do documento

Neste capítulo inicial, foram apresentadas a justificativa, a problemática, os objetivos, as principais contribuições e a metodologia adotada pela tese. Definiu-se o escopo e os desafios que motivam a pesquisa, delineando as expectativas e a estrutura geral do documento.

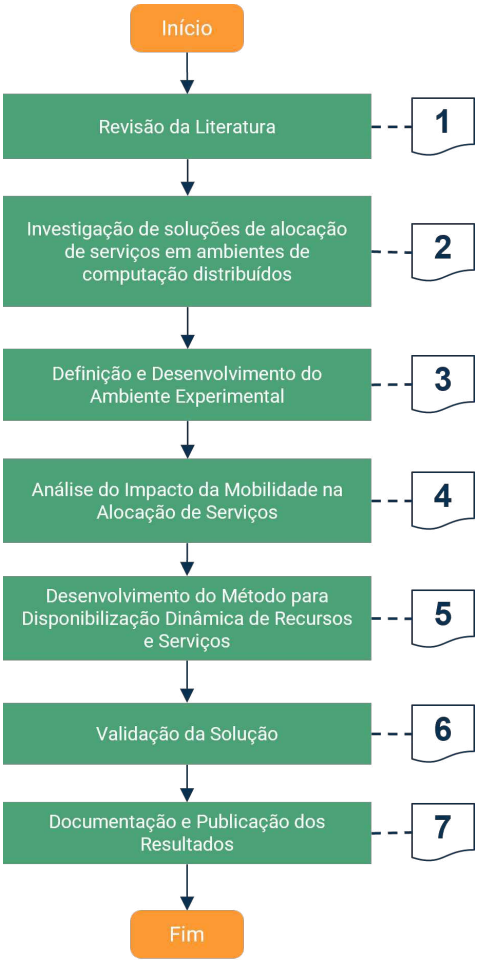
No Capítulo 2, é detalhada a fundamentação teórica, com a apresentação das ferramentas e conceitos que fundamentam a investigação. Este capítulo é essencial para compreender as bases técnicas e teóricas que suportam os experimentos e análises realizados.

No Capítulo 3, é apresentada a revisão bibliográfica, dividida em aplicações críticas de saúde e orquestração de serviços. Esta revisão destaca os estudos anteriores relevantes que moldam o entendimento atual e estabelecem a lacuna de conhecimento que a pesquisa visa preencher.

No Capítulo 4, são detalhados os requisitos de comunicação e computação no cenário de ambulâncias conectadas. Explica-se como as necessidades específicas de comunicação impactam o design e a implementação das soluções tecnológicas.

No Capítulo 5, são detalhados os passos metodológicos e as escolhas técnicas feitas

Figura 1.3: Fluxograma da metodologia.



Fonte: Produzida pelo autor.

durante o desenvolvimento do método *Make Way*.

No Capítulo 6, é descrita a Prova de Concito para Validação do Método, assim como o ambiente experimental utilizado para testar e validar a solução.

No Capítulo 7, são apresentados os resultados das simulações realizadas para validar o método. Analisam-se os dados coletados, discutindo-se a eficácia da solução e comparando-se os resultados com os objetivos inicialmente propostos.

Finalmente, no Capítulo 8, são apresentadas as conclusões da pesquisa e sugeridas futuras linhas de investigação.

Capítulo 2

Fundamentação Teórica

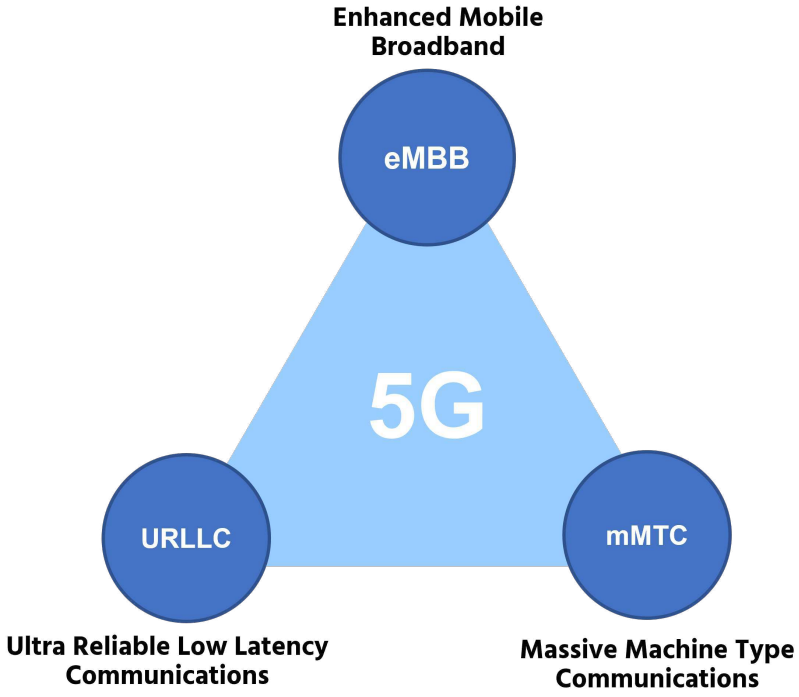
Neste capítulo está apresentada a fundamentação teórica, de forma a orientar o leitor com relação aos principais conceitos e tecnologias que são abordadas ao longo do trabalho. Inicialmente é apresentado o conceito da quinta geração de comunicação móvel (5G) e computação na borda, incluindo a suas utilizações no cenário da saúde. No restante do capítulo são apresentados os conceitos de Virtualização, Orquestração de Microsserviços, e Malha se Serviço, com as principais características, benefícios e ferramentas existentes.

2.1 Redes Móveis de Próxima Geração

Em um relatório publicado pela Nokia Communication [72] é afirmado que entre os anos de 2020 e 2030 haverá um aumento de 10 mil vezes mais no tráfego por meio de tecnologias de banda larga móvel. Um volume tão grande de tráfego de dados deverá implicar no design da arquitetura de redes móveis de próxima geração, como por exemplo o 5G, de maneira a permitir a implantação em uma nova e mais alta faixa de frequências. Alguns outros fatores como redução de consumo de energia e aumento da vida útil de baterias em termos de smartphones, tablets e laptops também deverão ser considerados no desenvolvimento da arquitetura 5G. Prevê-se que a 5G seja a principal tecnologia de desenvolvimento que levará a uma mudança radical na capacidade das redes móveis, principalmente no que diz respeito ao avanço de novas tecnologias de rádio, uso de frequências mais elevadas, redesenho de arquiteturas de redes, aprimoramentos de antenas principalmente associadas ao uso massivo de sistemas MIMO (Multiple-Input Multiple-Output), ultra baixa latência e altíssima confi-

abilidade [73].

Figura 2.1: Triângulo do 5G.



Fonte: [74].

As especificações da tecnologia 5G vem sendo tratadas pelo *3rd Generation Partnership Project* (3GPP), enquanto a sua implementação é feita por grandes empresas dos segmentos de hardware de rede e telecomunicações, tais como Nokia, Ericsson, Huawei, Qualcomm, entre outras. Comparado às tecnologias 4G atuais, amplamente difundidas em todo o mundo, o 5G deve ter maior largura de banda de até 10 Gbps, menor latência de 1 ms e maior densidade de dispositivos de até um milhão dispositivos por quilômetro quadrado [75]. Através do documento ITU-R M.2083-0 [76], os principais cenários de comunicação do uso do 5G (Figura 2.1) são definidas pela União Internacional de Telecomunicações (ITU – *International Telecommunication Union*):

- ***Ultra-Reliable and Low Latency Communications* (uRLLC):** aplicações que apresentam requisitos restritos de taxa de transferência, latência e disponibilidade. Aplicalçoes de Internet Tátil, aplicações de carros autônomos, cirurgias remotas, *Smart Grids*, proteção pública e atendimento emergencial em caso de catástrofes, controle

em rede sem fio de processos industriais, são exemplos de aplicações para este caso de uso [77].

- ***Massive Machine Type Communications* (mMTC):** caracterizado pela quantidade massiva de dispositivos conectados, onde o volume de transmissão de dados é relativamente baixo e não podem apresentar atraso. Além disso, característica como baixo custo e longa duração das baterias são fundamentais para estes casos de uso.

- ***Enhanced Mobile Broadband* (eMBB):** envolve os casos de uso de acesso de conteúdo multimídia, serviços e dados. O eMBB pode ser visto como uma evolução da existente Banda Larga Móvel (MBB - *Mobile Broadband*), de forma a atender aos requisitos das novas aplicações com mais desempenho e melhor experiência do usuário. A ampla cobertura e *hotspot* são cobertos nesses casos nos quais têm requisitos diferentes. No caso da ampla cobertura, espera-se uma experiência de uso sem falhas e alta mobilidade, com taxas de transferência de dados maiores em relação às que são oferecidas atualmente. Para o caso do *hotspot*, em uma área com alta densidade de usuários, por exemplo, alta capacidade de tráfego é necessária. O requisito de mobilidade é menos exigente e a taxa de transferência de dados é maior em relação ao da ampla cobertura.

Paralelamente, o desenvolvimento da tecnologia 5G não implica apenas em melhores velocidades de comunicação e todos os benefícios até aqui mencionados, mas também uma série de tecnologias com potencial para mudar o cenário da computação de maneira disruptiva. Entre essas tecnologias, é possível destacar as Redes Definidas por Software (SDN), a Virtualização de Funções de Rede (NFV), o *Network Slicing* (NS), as comunicações *Device-to-Device* (D2D) e Computação na Borda [78]. Redes Definidas por Software representam métodos para separar o plano de dados, que é responsável for manipular e encaminhar pacotes de rede e o plano de controle, responsável por estabelecer a rota dos pacotes. Virtualização de Funções de Rede representa, por sua vez, o uso de hardware comum executando serviços virtualizados para substituir hardware de rede personalizado. Por exemplo, um servidor básico pode executar serviços de firewall em vez de usar um firewall físico especializado. Se tratando do *Network Slicing*, esta é uma tecnologia que permite que várias redes lógicas compartilhem uma única infraestrutura de rede física. A comunicação D2D é um re-

curso do 5G que permite que os dispositivos se comuniquem diretamente, com ajuda mínima de uma autoridade central. Por exemplo, a estação base pode ajudar apenas no emparelhamento e autenticação, enquanto as etapas subsequentes, incluindo transferências de dados, são executadas sem seu envolvimento [79]. A introdução dessas tecnologias significa que as redes estão se tornando mais flexíveis e poderosas, uma vez que elas transferem grande parte da complexidade de uma rede do hardware para o software, da própria rede para seu gerenciamento e operação [80].

Diante de todo ecossistema tecnológico proporcionado pelo desenvolvimento da próxima geração da internet móvel, um dos grandes paradigmas que o 5G irá quebrar está associado a habilitar uma sociedade completamente conectada e móvel, dando espaço para amplas transformações socioeconômicas marcadas por melhorias na produtividade, sustentabilidade, eficiência e bem-estar geral da população mundial [81].

2.1.1 5G Aplicado na Saúde

Se tratando da particularidade da pandemia de COVID-19 vivenciada recentemente, por exemplo, tomar posse de soluções que venham beneficiar grande parte da população é de fundamental importância para controle, tratamento, diagnóstico, prevenção e combate a surtos, endemias, epidemias e pandemias. Aplicações de Inteligência Artificial, reconhecimento facial para detecção de equipamentos de proteção individuais, como a utilização de máscaras, e processamento de imagem para medição da temperatura corporal com câmeras de alta resolução apoiadas pela baixa latência e streaming de vídeo de alta qualidade proporcionados pela rede 5G foram utilizadas na China e Coreia do Sul como estratégias de controle da proliferação do COVID-19 [82]. Além disso, outras contribuições que a tecnologia 5G proporcionará à área de saúde são descritos em [83], onde são mencionadas aplicações em telemedicina, sensores ingeríveis, dispositivos vestíveis para auxílio em diagnósticos, terapia e cirurgia assistidas por robô, assistência domiciliar por meio de drones.

Foi publicado em um relatório da União Europeia (UE) [84] que as soluções de sistemas de saúde móvel podem economizar 99 bilhões de euros em gastos anuais totais com saúde e acrescentaria 93 bilhões de euros ao PIB da UE em 2017 com políticas de incentivo ao desenvolvimento de sistemas de saúde móvel (*mHealth*). Essas economias equivalem ao tratamento de 24,5 milhões adicionais pacientes com o mesmo número de médicos e instala-

ções. A cada ano, controle remoto móvel baseado em saúde o monitoramento de idosos em casa economiza 2,4 bilhões de euros na Suécia, 1,25 bilhão de euros na Dinamarca e 1,5 bilhões euros na Noruega [85]. Os sistemas de saúde integrados, utilizando 5G para aplicações de ciber-medicina (*e-Health*) e saúde móvel, têm sólidos benefícios econômicos promovidos pelo contínuo aumento da quantidade de smartphones e melhoria rápida da conectividade. Por exemplo, soluções de sistemas de saúde móvel permitem tratamento e monitoramento remotos de condições crônicas e equipa os médicos para tomar melhores decisões clínicas. Dessa forma, os sistemas de saúde móveis podem reduzir substancialmente os custos voltados à assistência médica [83].

Além disso, a telemedicina serve como uma clínica de baixo custo, estende a equipe clínica para outros locais e fornece um diagnóstico e tratamento conveniente, privado e integrado. O mercado global de telemedicina está crescendo rápido, o que representou uma receita de US$17,8 milhões em 2015 e deve crescer a uma taxa composta de crescimento anual (CAGR) de 18,7% entre os anos de 2016 e 2022 devido ao aumento da aceitação do espectro da tecnologia 5G. Apesar do grande aumento na cobertura de telefonia móvel e no acesso à Internet, 46,4% da população global ainda carece de acesso à Internet e há o dobro de assinaturas de banda larga móvel por 100 habitantes nos países desenvolvidos e nos países em desenvolvimento [86]. Em um cenário nacional, no mais recente Relatório de Gestão da Saúde [87] é possível observar que mais de 55% do valor empenhado para Gestão de Tecnologia da Informação foram destinados ao Suporte de Infraestrutura de Tecnologia da Informação e Comunicação (TIC), Comunicação de Dados e Redes em Geral, Manutenção e Conservação de Equipamentos de TIC e Suporte a Usuários de TIC, representando um montante de R$41,2 milhões.

No contexto de comunicações sem fio, é notório que o 5G apresenta grande contribuição na para os avanços na área de saúde, permitindo a personalização dessa por meio da coleta contínua de dados de monitoramento de pacientes para processamento e armazenamento centralizado. Além disso, o 5G permite mudar o local de atendimento de hospitais para residências e outras instalações de custo mais baixo, o que se traduz em economias adicionais. Outro exemplo que mostra que o 5G pode permitir a economia de custos exigida pela indústria médica pode ser encontrado dentro de hospitais onde a transmissão sem fio de fluxos de dados de baixa latência melhora o planejamento de salas de cirurgias, permitindo otimizar o

uso do equipamento e simplificando a implementação de centros cirúrgicos [88].

O 5G tem potencial para preencher significativamente essa lacuna, garantindo uma experiência consistente do usuário de áreas densas a aldeias ou mesmo áreas remotas. Dessa forma, as emergentes aplicações de assistência médica estarão disponíveis em uma disseminação muito mais ampla do que hoje.

2.1.2 5G Aplicado em Outros Segmentos da Economia

No ano de 2017, o governo brasileiro publicou um estudo em parceria com o BNDES e o CNPq chamado Plano Nacional de IoT [89], em que prevê grande impacto em quatro principais áreas: Cidades Inteligentes, Saúde, Agronegócio e Indústria, segmentos esses amplamente favorecidos pela tecnologia 5G, umas vez que é esperado que a nova geração de tecnologia de redes móveis transforme mercados verticais inteiros como governo, saúde, manufatura, automotivo, energia, alimentação e agricultura, administração de cidades, transporte e muitos outros.

O comércio eletrônico, como compras online, será beneficiado ainda mais pelas tecnologias 5G. O streaming de vídeo de alta qualidade e o feed de informações em tempo real fornecerão não apenas imersiva experiência de compra com rapidez e personalização recomendações, mas também permitem mistura e combinações dinâmicas de escolhas. Por exemplo, uma peça de mobiliário pode ser vista da perspectiva de um ambiente doméstico real. Enquanto o 4G popularizou as compras online, é provável que o 5G leve dá alguns passos adiante com realidade aumentada, verificação rápida de fatos, recomendações e experiência geral [90].

A área de logística será uma das grandes beneficiadas pelo desenvolvimento da tecnologia 5G: com mais velocidade, segurança e estabilidade, será mais fácil para o setor de logística planejar e acompanhar rotas e pacotes em tempo real. Com isso, será possível evitar congestionamentos que poderiam atrasar as entregas. Outro segmento que poderá ser impactado é o agronegócio [91]. A indústria automotiva será significativamente impactada pela 5G, pois abre o potencial de conexão de veículos a infraestruturas de borda das estradas, pedestres e outros veículos. Atualmente, veículos autônomos não são totalmente suportados pelo infraestrutura de TI devido à falta de antenas móveis e sensores, o que não permite comunicações com eficiência e estabilidade [92].

Nesse contexto, é possível perceber a gama de mudanças que a tecnologia 5G proporcionará em um cenário global e que investir no desenvolvimento dessa tecnologia apresenta benefício em diversos segmentos da geoeconomia global e aumento na qualidade de vida. A Quinta Geração de Comunicação Móvel (5G) está no caminho de ser adotada com grande expressividade em um cenário global. Atualmente, o 5G está sendo implantado em pequenas áreas em quase todos continentes, com um número maior de redes disponíveis na Europa e nos EUA [90].

2.2 Computação em Borda

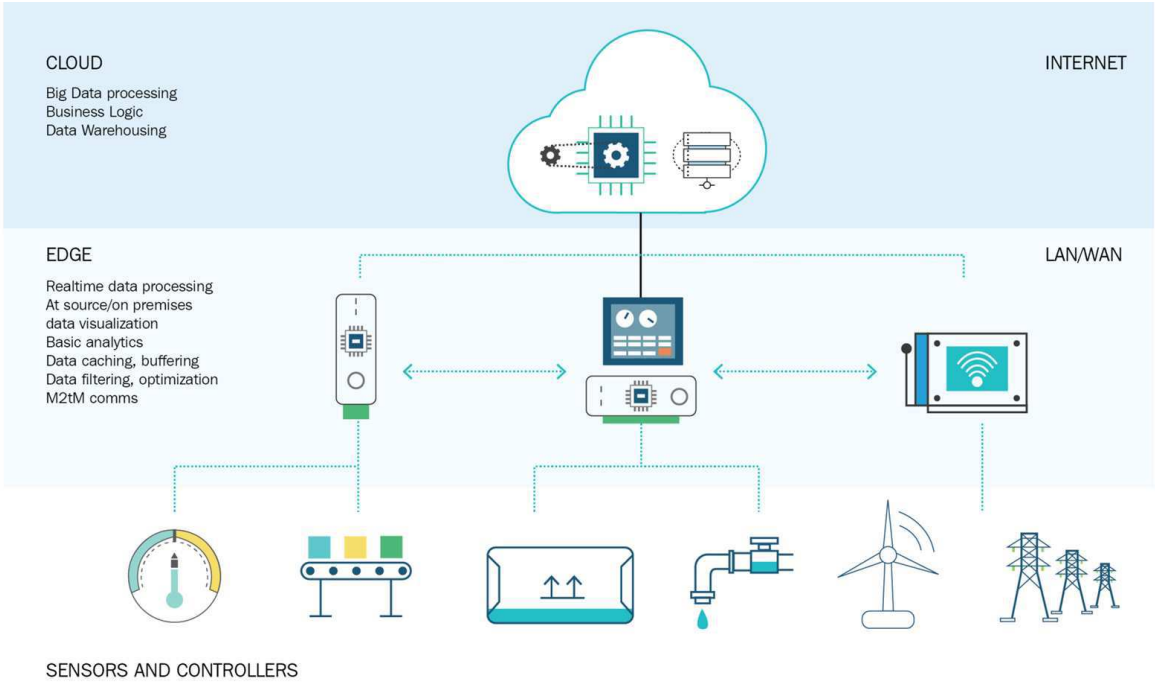
Vivemos em um mundo cada vez mais conectado. Hoje grande parte das aplicações são baseadas em tecnologias de computação em nuvem (Cloud Computing), onde busca-se fazer com que todos os dispositivos fiquem acessíveis de qualquer lugar, a qualquer hora. Estes dispositivos conectados, geram um grande volume de dados que podem ser processados para criação de aprendizado, geração valor e surgimento de novos modelos de negócios que podem envolver o gerenciamento, manutenção preventiva, controle remoto, monitoramento, análise, entre outros. A Internet das Coisas (IoT), indo além de conectar diversas "coisas"à internet, desempenha um papel fundamental para a existência de sistemas ciber-físicos, fazendo a ponte entre as pessoas e o ambiente com os milhares de sistemas e dispositivos conectados [93].

Atualmente, as arquiteturas de sistemas de IoT são fortemente embasadas em tecnologias de computação em nuvem, que fornecem os recursos necessários para o seu funcionamento. Porém, o aumento da quantidade de dispositivos conectados está conduzindo a limitações inevitáveis [94]. A grande quantidade de dados transmitidos a cada minuto e o grande processamento de dados realizado na nuvem geram um alto custo, insuficiência na confiabilidade, alta latência, utilização excessiva da rede e, também, tempo de inatividade imprevisível. A escalabilidade da nuvem, por apresentar uma arquitetura central, impede a evolução não planejada de soluções e instalações já existentes.

Neste sentido, destaca-se também um fator crucial para a criação de soluções de IoT: a segurança cibernética. Armazenar e assegurar o armazenamento de dados críticos e conexões em soluções de nuvem pública pode ser difícil. Na solução centralizada, sensores de dispo-

sitivos conectados carregam os dados diretamente para a nuvem ou transmitem por meio de um gateway. Hoje uma nova mudança de tecnologia já pode ser vista. As organizações estão movendo serviços de dados baseados na nuvem para as bordas da rede, permitindo que o armazenamento e o processamento se aproximem da fonte de geração dos dados para obtenção de resultados mais rápidos, sendo este novo paradigma de computação conhecido por a computação em borda (Edge Computing). A aproximação da capacidade de processamento dos dados em sua fonte pode proporcionar grandes benefícios aos negócios: insights mais rápidos, tempos de resposta melhores e disponibilidade de largura de banda aprimorada.

Figura 2.2: Arquitetura de computação em borda.



Fonte: [95].

2.2.1 Computação em Borda Aplicada à Saúde

Com o advento da Computação na Borda, o setor de saúde tem evoluído consideravelmente devido à sua capacidade de armazenar, processar e analisar os dados mais próximos de pacientes, hospitais e clínicas. De fato, a Computação na Borda está permeando o setor com tanta expressividade que profissionais da área de saúde estão começando a confiar cada vez mais nessa tecnologia para apoiar o tratamento de pacientes [96]. Atualmente, os smartpho-

nes estão desempenhando um papel cada vez mais importante na arquitetura de computação moderna graças a sua distribuição em massa e os crescentes recursos computacionais. Estes dispositivos são frequentemente usados como gateways IoT móveis [97] ou para oferecer suporte à integração heterogênea de redes de sensores sem fio [98]. A inclusão de dispositivos móveis pessoais na computação na borda podem acelerar o desenvolvimento de solução móvel centrada nas pessoas, não apenas na área da saúde, mas também na melhoria da qualidade da vida. Nesse sentido, o uso de dispositivos móveis pessoais é a maneira mais rápida e fácil de apoiar a integração de Redes de Sensores Corporais (*Body Sensor Networks*) e fornecer uma gama de serviços de saúde móvel.

Iniciativas como clínicas móveis [99] podem ser vistas como uma ótima estratégia quando a demanda por assistência médica aumenta, por exemplo, em casos de pandemia em que se é requerido a construção de pontos de atendimento emergenciais para atender a crescente demanda decorrente do alto índice de contaminação de doenças virais como é o caso da COVID-19. Como apresentado por [100], foi criado um dispositivo portátil de emergência médica capaz de ser usado no pescoço do paciente. O dispositivo oferece, entre outras funcionalidades, o serviço de chamada de emergências e a detecção automática contínua de queda que inicia uma chamada de emergência para centros atendimento de enfermagem especializados (em um canal predeterminado para esse tipo de serviço). Um outro sistema para monitoramento de pacientes é mostrado em [101], onde foi criado um dispositivo que deve ser usado preferencialmente sobre o tórax, logo abaixo dos músculos peitorais e monitora pelo menos o seguinte: dados de ECG, taxa de respiração, captação de oxigênio, frequência de pulso e temperatura corporal. Em [102] é apresentado um projeto e avaliação de um aplicativo móvel para auxiliar o automonitoramento da doença renal crônica nos países em desenvolvimento. Uma outra solução de integrada de serviços é apresentada em [103]. Nesse sistema são fornecidos: serviços de banco de dados móvel de gerenciamento de saúde, mecanismo de assistência educacional direcionada, compartilhamento seletivo de dados de assistência médica e interface gráfica do usuário em árvore genealógica.

A criação de uma estação de trabalho ou ponto de atendimento móvel com interface para smartphone foi uma solução apresentada por [104] para assistência ao profissional de saúde. Esta solução inclui um carrinho de ponto de atendimento móvel configurado para suportar pelo menos um dispositivo periférico, facilitando a entrada de dados médicos específicos de

um paciente e um módulo de interface montado no carrinho de ponto de atendimento móvel. O Hub de Dados de Saúde Pessoal (PHDH) foi uma solução desenvolvida em [105] que consiste em uma solução para recebimento, armazenamento e transmissão de dados de saúde pessoal pelo PHDH por meio de diversas tecnologias de comunicação como Bluetooth, Bluetooth Low Energy, ANT +, USB, entre outras. O PHDH pode ser usado por diferentes usuários, como várias sessões de usuário. Os usuários podem acessar e controlar o PHDH através de diferentes mecanismos de interface do usuário. Uma outra solução de gateway é apresentada em [106] que teve como foco o escalonamento dinâmico para sistemas embarcados para gerenciamento de saúde. Já em [107], é apresentada uma arquitetura de gateways inteligentes baseada em padrões e sensível ao contexto aplicada à cuidados pessoais.

Outras soluções relacionadas ao monitoramento de pacientes podem ser encontradas na literatura. Em [108], por exemplo, é mostrado um sistema ao qual um dispositivo vestível detecta uma variedade de dados do paciente. Os dados podem ser transmitidos por meio de uma rede para uma central onde os dados podem ser analisados e ações apropriadas podem ser tomadas. Já em [109], um sistema para monitorar remotamente o status do pessoal inclui uma pluralidade de sensores distribuídos em um paciente para gerar sinais que podem ser utilizados para determinar o status fisiológico do mesmo. Uma outra abordagem para aplicações médicas usando redes móveis é apresentada em [110], onde foi desenvolvido um sistema médico na área de desastre que pode ser usado principalmente para operações de resgate. Esse sistema é composto principalmente por vários dispositivos móveis de resgate médico, um servidor de informações de postos médicos e vários servidores do hospital. O dispositivo móvel de resgate médico pode formar um link de notícias com o servidor da plataforma de informações da estação médica por meio de conexão via satélite. Para que o pessoal da ambulância obtenha imediatamente informações a respeito de leitos hospitalares em hospitais próximos à área do desastre e informações do paciente para tratamento. Um sistema de saúde pessoal conectado à Internet das Coisas baseado no *Constrained Application Protocol* (CoAP) é apresentado em [111]. O CoAP é uma protocolo de transferência web especializado para uso com nós e redes restritos (por exemplo, baixa energia, perdas) [112].

De acordo com essa visão moderna, a computação em borda voltada à assistência médica está surgindo como uma maneira de as entidades adotarem resultados quase em tempo real,

processando os dados o mais próximo possível da fonte (pacientes), afirmando definitivamente que a computação em nuvem não é uma maneira eficiente de processar dados quando os dados são produzidos na borda da rede [113].

2.3 Virtualização

A virtualização é uma tecnologia que permite criar serviços de TI valiosos usando recursos que tradicionalmente estão vinculados a um determinado hardware. Esta tecnologia permite que o usuário use toda a capacidade de uma máquina física ao proporcionar a distribuição dos recursos entre diversos usuários ou ambientes [114].

Nos últimos anos, a aplicação da técnica de virtualização de servidores tem tomado uma representatividade considerável no mercado de engenharia de software, pois é um método que permite a execução de vários sistemas operacionais e aplicações em uma única máquina [115].

Atualmente, as tecnologias de virtualização existentes podem ser classificadas em duas categorias: *(i)* virtualização baseada em máquinas virtuais; ou *(ii)* virtualização baseada em containers. A virtualização por máquina virtual (VM) é o método utilizado pela maioria das empresas [116]. Porém, a virtualização por containers tem crescido consideravelmente nos últimos anos, pois é um método que exige menos recursos que as máquinas virtuais, sendo possível implantar e migrar serviços em um tempo menor [117]. A ferramenta de virtualização por containers denominada Docker é consideravelmente utilizada atualmente, sendo a líder mundial deste segmento no mercado [118].

2.3.1 Benefícios da virtualização

A virtualização pode aumentar a agilidade, a flexibilidade e o dimensionamento da TI e, ao mesmo tempo, proporcionar uma economia significativa. Alguns dos benefícios da virtualização, como a maior otimização do uso dos recursos computacionais, a redução de custos, a abstração dos recursos computacionais [119], mobilidade das cargas de trabalho, o aumento do desempenho e da disponibilidade dos recursos ou a automação das operações, simplificam o gerenciamento da infraestrutura de TI e permitem reduzir os custos de propriedade e operacionais. Além disso, podemos citar a flexibilidade no desenvolvimento de sistemas

devido à sua capacidade de alocação de recursos de hardware e software de forma dinâmica [117]. Mas também podem ser destacados outros benefícios:

- Redução dos custos operacionais e de capital;
- Redução ou eliminação do tempo de inatividade;
- Aumento de produtividade, eficiência, agilidade e capacidade de resposta da TI;
- Mais rapidez no aprovisionamento de aplicativos e recursos;
- Melhor continuidade de negócios e DR;
- Gerenciamento simplificado de data centers;
- Disponibilidade de um data center real definido por software. [120]

2.3.2 Tipos de virtualização

As tecnologias de virtualização tem se tornada fortes aliadas dos provedores de soluções e serviços de infraestrutura de comunicação e computação. Neste sentido, é possível beneficiar-se da aplicação do conceito e utilização de tecnologias de virtualização em diversos níveis de aplicação na cadeia de fornecimento de serviços comunicação e computação. Para tanto, conhecer as possibilidades e tipos de virtualização é de fundamental importância para correta solução e aplicação em negócios cada vez mais verticalizados. A seguir estão listados alguns tipos de virtualização existentes.

2.3.2.1 Virtualização de Dados

Dados distribuídos em vários locais são consolidados em uma única fonte. A virtualização de dados permite às empresas tratar os dados como um tipo de suprimento dinâmico, oferecendo recursos de processamento capazes de reunir dados de diversas fontes, acomodar facilmente novas fontes e transformar os dados de acordo com as necessidades dos usuários. As ferramentas de virtualização de dados funcionam como múltiplas fontes de dados e permitem que essas fontes sejam tratadas como uma só. Dessa forma, os dados necessários são fornecidos no formato e no momento certo para qualquer aplicação ou usuário. [114]

2.3.2.2 Virtualização de Desktop

Muitas vezes confundida com a virtualização do sistema operacional (a implantação de diversos sistemas operacionais em uma única máquina), a virtualização de desktop permite que um administrador central (ou ferramenta de administração automatizada) implante ambientes de desktop simulados em centenas de máquinas físicas de uma única vez. Diferente dos ambientes de desktop tradicionais, fisicamente instalados, configurados e atualizados em cada máquina, na virtualização de desktop, administradores podem realizar configurações, atualizações e verificações de segurança em massa em todos os desktops virtuais. [114]

2.3.2.3 Virtualização de Servidores:

Os servidores são computadores projetados para processar um grande volume de tarefas específicas, para que outros computadores, como laptops e desktops, realizem uma variedade de outras tarefas. A virtualização do servidor o libera para realizar mais funções específicas, pois se dá por meio do seu particionamento. Assim, os componentes podem ser utilizados para o processamento de várias funções. [114]

2.3.2.4 Virtualização do Sistema Operacional

A virtualização do sistema operacional é feita no kernel, o gerenciador de tarefas central dos sistemas operacionais. Essa é uma boa maneira de executar paralelamente ambientes em Linux e Windows. As empresas também podem implantar nos computadores sistemas operacionais virtualizados, que: (i) Reduzem os custos de hardware em massa, já que os computadores não requerem recursos prontos e sofisticados; (ii) Aumentam a segurança, pois todas as instâncias virtuais podem ser monitoradas e isoladas; e (iii) Limitam o tempo gasto com serviços de TI, como atualizações de software.[114]

2.3.2.5 Virtualização de Funções de Rede (NFV)

A virtualização de funções de rede separa as principais funções de uma rede (como serviços de diretório, compartilhamento de arquivos e configuração de IP) para distribuí-las entre os ambientes. Após separar as funções de software das máquinas físicas em que residiam, é possível reunir funções específicas em uma nova rede e atribuí-las a um ambiente. O resultado

da virtualização de redes é a redução do número de componentes físicos, como switches, roteadores, servidores, cabos e hubs, necessários para criar várias redes independentes. Esse tipo de virtualização é popular principalmente no setor de telecomunicações. [114]

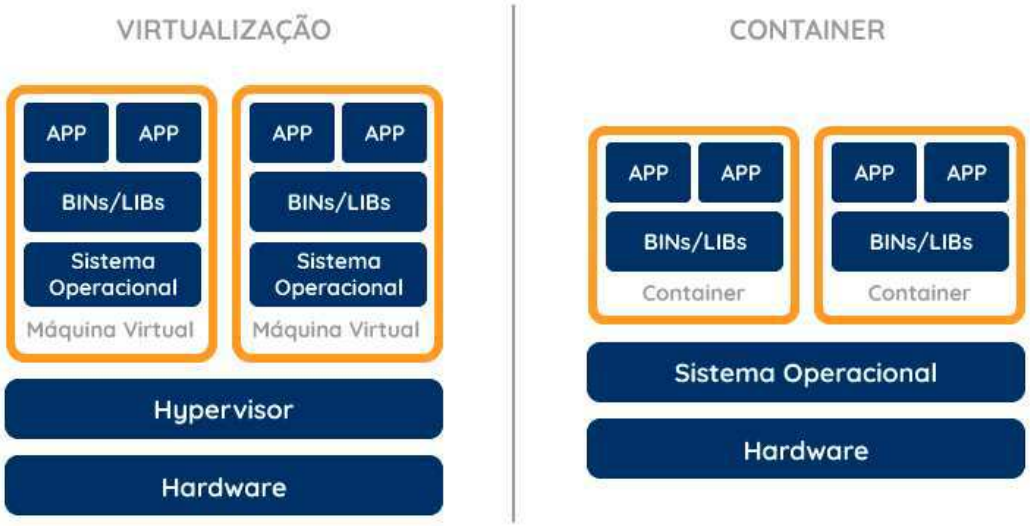
2.3.3 Conteiners

Um container é um conjunto de um ou mais processos organizados isoladamente do sistema. Todos os arquivos necessários para executá-los são fornecidos por uma imagem distinta. Na prática, os containers são portáteis e consistentes durante toda a migração entre os ambientes de desenvolvimento, teste e produção. Essas características os tornam uma opção muito mais rápida do que os pipelines de desenvolvimento, que dependem da replicação dos ambientes de teste tradicionais. Os containers também são uma parte importante da segurança da TI por conta da popularidade e da facilidade de uso deles. [95]

Apesar de serem duas tecnologias são distintas, a virtualização por meio de máquinas virtuais ou utilizando containers apresentam-se como tecnologias complementares. Mas qual é a diferença entre virtualização e os containers? Com a virtualização, é possível executar sistemas operacionais simultaneamente em um único sistema de hardware. Já os containers compartilham o mesmo núcleo do sistema operacional e isolam os processos da aplicação do restante do sistema. Por exemplo, os sistemas ARM Linux executam containers ARM Linux, os sistemas x86 Linux executam containers x86 Linux e os sistemas x86 Windows executam containers x86 Windows. Os containers Linux são extremamente portáteis, mas devem ser compatíveis com o sistema subjacente. [115]

Consequentemente, a virtualização usa um hipervisor para emular o hardware, o que permite executar vários sistemas operacionais simultaneamente. Essa não é uma solução tão leve quanto o uso de containers. Quando a capacidade e os recursos são limitados, é necessário usar aplicações leves que possam ser implantadas densamente. Os containers Linux são executados de maneira nativa no sistema operacional, compartilhando-o com todos os outros containers. Assim, as aplicações e os serviços permanecem leves e são executados em paralelo com agilidade. [95] Na figura 2.3 estão ilustradas as arquiteturas de virtualização por meio de VMs e containers.

Figura 2.3: Figura comparativa entre as arquiteturas de virtualização por meio de VMs e por meio de Containers.



Fonte: [95].

2.3.4 Docker

O Docker é a ferramenta de virtualização por containers. Os desenvolvedores de software a utilizam para eliminar problemas de compatibilidade em seus programas. Operadores utilizam o Docker para executar e gerenciar as suas aplicações lado a lado em containers isolados para obter uma melhor densidade computacional. Empresas usam o Docker para acelerar o desenvolvimento e a integração dos seus softwares (Docker, 2016). A ferramenta é instalada sobre o sistema operacional hospedeiro, conforme a Figura 3.

A tecnologia Docker usa o kernel do Linux e recursos do kernel como Cgroups e namespaces para segregar processos. Assim, eles podem ser executados de maneira independente. O objetivo dos containers é criar essa independência: a habilidade de executar diversos processos e aplicações separadamente para utilizar melhor a infraestrutura e, ao mesmo tempo, manter a segurança que você teria em sistemas separados.

As ferramentas de container, incluindo o Docker, fornecem um modelo de implantação com base em imagem. Isso facilita o compartilhamento de uma aplicação ou conjunto de serviços, incluindo todas as dependências deles em vários ambientes. O Docker também automatiza a implantação da aplicação (ou de conjuntos de processos que constituem uma

aplicação) dentro desse ambiente de container.

Essas ferramentas baseadas nos containers Linux (o que faz com que o Docker seja exclusivo e fácil de usar) oferecem aos usuários acesso sem precedentes a aplicações, além da habilidade de implantar com rapidez e de ter total controle sobre as versões e distribuição.

2.4 Arquitetura de Microsserviços

A busca por uma melhor separação de interesses em sistemas de informação tem impulsionado o desenvolvimento e evolução de arquiteturas de software, onde busca-se decompor e organizar sistemas em módulos logicamente coesos e com baixo acoplamento, ocultando sua implementação uns dos outros e disponibilizam serviços por meio de interfaces bem definidas [121, 122]. Atualmente, dois paradigmas de engenharia de software dominam o desenvolvimento de aplicações empresariais modernas: (*i*) arquitetura monolítica, na qual um aplicativo é construído com uma única base de código que inclui vários serviços, não executáveis de forma independente; e (*ii*) baseada em microsserviços, que decompõe um domínio de negócios em contextos pequenos e consistentemente delimitados implementados por serviços autônomos, independentes, acoplados de maneira flexível e implantáveis de forma independente. [123]

Arquiteturas de microsserviços têm sido amplamente adotadas por grandes empresas de tecnologia, podendo-se citar a Amazon, Spotify, Uber, LinkeIn, Twitter, eBay, Netflix, entre outras [124, 125]. O surgimento de tecnologias de criação e gestão de contêineres, como por exemplo, Kubernetes e Docker, favoreceu a adoção desse novo paradigma de arquitetura de engenharia de software, especialmente em ambientes baseados em computação em nuvem [126–129].

De acordo com a definição apresentada em [130], arquitetura de microsserviços é uma abordagem para desenvolver aplicações como um conjunto de pequenos serviços, cada um executando em seu próprio processo e se comunicando com mecanismos leves, geralmente por meio de uma API HTTP. Cada microsserviço é autocontido, ou seja, contem suas próprias lógicas de negócio, funções de manipulação do usuário e funções de backend, podendo também incluir seu próprio banco de dados. Apesar disso, é possível também realizar o compartilhamento de um único backend com múltiplos microsserviços. As principais carac-

terísticas dessa arquitetura são:

- **Responsabilidade única por serviço:** uma única unidade deve ter apenas uma responsabilidade e em nenhum momento duas unidades devem compartilhar uma responsabilidade ou uma unidade ter mais de uma responsabilidade.

- **Os microsserviços são autônomos** – Os microsserviços são autônomos – são serviços independentes e implementáveis de forma independente, totalmente responsáveis pela execução de um determinado negócio. Por causa de sua autonomia, eles contêm todas as dependências como: bibliotecas, ambientes de execução – servidores web e containers ou máquinas virtuais. Dessa forma, os microsserviços aumentam a possibilidade de monetização de partes do sistema, pois o acesso a APIs de microsserviços relevantes pode ser cobrado [38].

- **Os serviços são 'First-class citizen'** – eles expõem endpoints de serviço como APIs e abstraem todos os detalhes de implementação. A estrutura interna: lógica de implementação, arquitetura e tecnologias (incluindo linguagem de programação, banco de dados, etc.) estão completamente ocultas por trás da API.

Uma das características mais interessantes da arquitetura de microsserviços é a decomposição de aplicações complexas em componentes menores que são mais fáceis de serem desenvolvidos, gerenciados e mantidos do que uma única aplicação monolítica [131]. Desde que a API pública não seja alterada, as modificações internas de um serviço são mais diretas, fáceis e menos onerosas do que no caso de uma alteração semelhante em um modelo tradicional. Os microsserviços são autônomos e se comunicam por meio de protocolos abertos, portanto, podem ser desenvolvidos de forma bastante independente e até mesmo com diferentes tecnologias [132–134]. Outro benefício é que sua arquitetura fracamente acoplada os torna mais tolerantes a falhas [135] – a falha de um componente não resulta necessariamente na indisponibilidade de todo o sistema, pois os serviços em funcionamento ainda podem atender às solicitações do usuário. Também é possível identificar funcionalidades críticas de negócios e implantar microsserviços correspondentes em um ambiente mais redundante.

No âmbito da digitalização de serviços de saúde, os dados de saúde são fundamentalmente importantes para o fornecimento de serviços de saúde de qualidade. Para melhorar

o desenvolvimento, depuração, manutenção e funcionalidades de arquiteturas de sistemas médicos que demandam cada vez mais flexibilidade e confiabilidade, os microsserviços representam uma das mais recentes orientações na concepção de tais aplicações. [136]

2.4.1 Orquestração de Microsserviços

Apesar das inúmeras vantagens que a arquitetura de microsserviços apresenta, pode-se destacar algumas desvantagens principalmente relacionadas à sua natureza distribuída. A implantação, escalabilidade, gestão e monitoramento de um sistema multisserviço é uma tarefa mais complexa do que no caso de sistemas que utilizam arquiteturas monolíticas. Por esse motivo, vários procedimentos de automação no pipelines de integração contínua/entrega contínua (CI/CD), monitoramento e escalabilidade automática baseado em demanda são usados no desenvolvimento dessas aplicações. Para aproveitar totalmente o curto desenvolvimento até as operações, os testes de ciclo de vida também devem ser automatizados, o que é uma tarefa mais desafiadora em ambientes distribuídos [123]. Neste sentido, dispor de ferramentas e frameworks que facilitem a gestão e orquestração dos microsserviços, ainda assegurando os requisitos não funcionais das aplicações, é de fundamental importância à medida que a complexidade desses sistemas aumenta.

Em um ambiente como a Computação em Borda, composto potencialmente por dispositivos com restrição de recursos [137], o uso de containers, uma forma de virtualização leve e de tamanho reduzido [138], tem sido amplamente utilizado e indicado como sendo mais apropriado do que o uso de máquinas virtuais como ambientes de execução. Existem vários ambientes de execução de contêineres e o Docker é o mais conhecido. Além do próprio ambientes de execução, uma solução de orquestração de contêineres pode ser usada para gerenciar o ciclo de vida dos contêineres, dimensioná-los para cima ou para baixo, fazer self-healing, migrá-los e gerenciar uma infraestrutura heterogênea, abstraindo os dispositivos físicos onde eles executam. Assim, soluções de orquestração de contêineres podem ser adotadas para implementar e automatizar algumas das características essenciais da Computação em Borda: heterogeneidade, distribuição geográfica e escalabilidade [139]. Dentre as diversas ferramentas de orquestração existente, podemos citar: Kubernetes [140], Docker Swarm [141], Nomad [142] e Marathon on Mesos [143].

2.4.2 Kubernetes

O Kubernetes é uma plataforma open source que automatiza as operações dos containers Linux. Essa plataforma elimina grande parte dos processos manuais necessários para implantar e escalar as aplicações em containers. Em outras palavras, se você desejar agrupar em clusters os hosts executados nos containers Linux, o Kubernetes ajudará a gerenciar esses clusters com facilidade e eficiência. Esses clusters podem incluir hosts em nuvem pública, nuvem privada ou nuvem híbrida. Por isso, o Kubernetes é a plataforma ideal para hospedar aplicações nativas em nuvem que exigem escalabilidade rápida, como por exemplo a transmissão de dados em tempo real por meio do Apache Kafka [144].

Aplicações de produção abrangem múltiplos containers. Eles devem ser implantados em vários hosts do servidor. A segurança dos containers tem várias camadas e pode ser complexa. É aí que o Kubernetes entra em cena. Ele oferece os recursos de orquestração e gerenciamento necessários para implantar containers em escala para essas cargas de trabalho. Com a orquestração do Kubernetes, é possível criar serviços de aplicações que abrangem múltiplos containers, programar o uso deles no cluster, escalá-los e gerenciar a integridade deles com o passar do tempo. Com o Kubernetes, você toma medidas reais para aprimorar a segurança da TI [145].

No entanto, isso obviamente depende do uso que cada empresa faz dos containers em seus próprios ambientes. Uma aplicação rudimentar dos containers Linux os trata como máquinas virtuais rápidas e eficientes. Quando escalado para um ambiente de produção e diversas aplicações, fica claro que é necessário ter vários containers alocados funcionando em conjunto para disponibilizar serviços individuais. Isso multiplica substancialmente o número de containers no ambiente. À medida que eles se acumulam, a complexidade também aumenta [146].

O Kubernetes corrige vários problemas comuns que ocorrem com a proliferação de containers, organizando-os em "pods". Os pods adicionam uma camada de abstração aos containers agrupados. Assim, é mais fácil programar as cargas de trabalho e fornecer os serviços necessários a esses containers, como rede e armazenamento. Outros componentes do Kubernetes são úteis no balanceamento de carga entre os pods. Com isso, é possível garantir que o número de containers em execução seja suficiente para oferecer suporte às cargas de trabalho [144].

O Kubernetes possibilita:

- Orquestrar containers em vários hosts.
- Aproveitar melhor o hardware para maximizar os recursos necessários na execução das aplicações corporativas.
- Controlar e automatizar as implantações e atualizações de aplicações.
- Montar e adicionar armazenamento para executar aplicações com monitoração de estado.
- Escalar rapidamente as aplicações em containers e recursos relacionados.
- Gerenciar serviços de forma declarativa, garantindo que as aplicações sejam executadas sempre da mesma maneira como foram implantadas.
- Verificar a integridade e autorrecuperação das aplicações com posicionamento, reinício, replicação e escalonamento automáticos [144].

A principal vantagem que as empresas garantem ao usar o Kubernetes, especialmente se estiverem otimizando o desenvolvimento de aplicações para a cloud, é que elas terão uma plataforma para programar e executar containers em clusters de máquinas físicas ou virtuais. Em termos mais abrangentes, com o Kubernetes, é mais fácil implementar e confiar totalmente em uma infraestrutura baseada em containers para os ambientes de produção. Como o propósito do Kubernetes é automatizar completamente as tarefas operacionais, ele permite que os containers realizem muitas das tarefas possibilitadas por outros sistemas de gerenciamento ou plataformas de aplicações [145].

2.4.2.1 Arquitetura do Kubernetes

Na arquitetura do Kubernetes, podemos definir os seguintes componentes:

- **Master:** a máquina que controla os nós do Kubernetes. É nela que todas as atribuições de tarefas se originam.
- **Nó:** são máquinas que realizam as tarefas solicitadas e atribuídas. A máquina mestre do Kubernetes controla os nós.

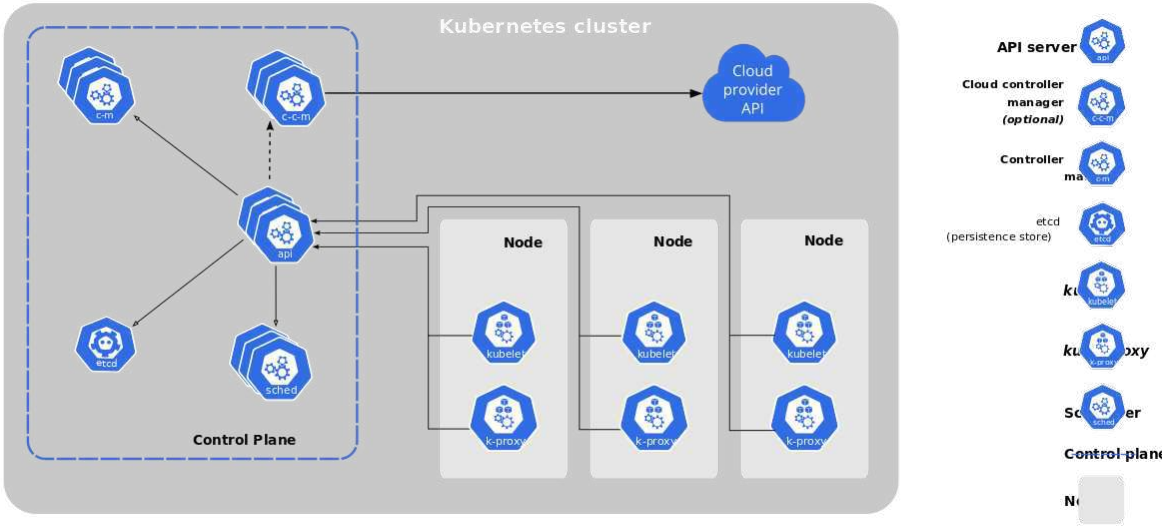
- **Pod:** um grupo de um ou mais containers implantados em um único nó. Todos os containers em um pod compartilham o mesmo endereço IP, IPC, nome do host e outros recursos. Os pods separam a rede e o armazenamento do container subjacente. Isso facilita a movimentação dos containers pelo cluster.

- **Controlador de replicações:** controla quantas cópias idênticas de um pod devem ser executadas em um determinado local do cluster.

- **Serviço:** desacopla as definições de trabalho dos pods. Os proxies de serviço do Kubernetes automaticamente levam as solicitações de serviço para o pod correto, independentemente do local do pod no cluster ou se foi substituído.

Na figura 2.4 está ilustrada a arquitura do Kubernetes com os componentes descritos anteriormente.

Figura 2.4: Arquitetura do Kubernetes.



Fonte: Extraído de [147].

O Kubernetes é executado em um sistema operacional e interage com pods de containers executados em nós. A máquina mestre do Kubernetes aceita os comandos de um administrador e retransmite essas instruções aos nós subservientes. Essa retransmissão é realizada em conjunto com vários serviços para automaticamente decidir qual nó é o mais adequado para a tarefa. Depois, são alocados os recursos e atribuídos os pods do nó para cumprir a

tarefa solicitada [145]. Portanto, o controle sobre os containers acontece em um nível superior, tornando-o mais refinado, sem a necessidade de microgerenciar cada container ou nó separadamente.

2.4.2.2 Componentes do Plano de Controle do Kubernetes

Os componentes do plano de controle são responsáveis por tomar decisões abrangentes acerca do cluster, tais como o agendamento de tarefas, além de detectar e responder a eventos pertinentes ao cluster, como por exemplo, a ativação de um novo pod quando o campo de réplicas de um Deployment se encontra insatisfatório.

É possível operar os componentes do plano de controle em qualquer máquina integrante do cluster. Contudo, visando simplificar o processo, os scripts de configuração habitualmente iniciam todos os componentes do plano de controle na mesma máquina, a qual não é utilizada para executar contêineres de usuários. Para ilustrações de configurações do plano de controle que se estendem por múltiplas máquinas, recomenda-se consultar o procedimento para criação de clusters altamente disponíveis utilizando o kubeadm [148].

Os componentes do plano de controle são os seguintes:

- **kube-apiserver**: O kube-apiserver é o front-end do plano de controle do Kubernetes, funcionando como a porta de entrada para todas as operações do cluster. Ele expõe a API do Kubernetes e processa as solicitações REST, validando e executando as operações correspondentes nos dados do cluster. É projetado para escalar horizontalmente, permitindo a adição de mais instâncias para aumentar disponibilidade e capacidade.

- **etcd**: O etcd é um armazenamento de chave-valor consistente e altamente disponível, utilizado como repositório para todos os dados do cluster do Kubernetes. Serve como a base para armazenar o estado de configuração e o estado de runtime, essencial para a recuperação e sincronização do estado do cluster.

- **kube-scheduler**: O kube-scheduler é responsável por monitorar pods recém-criados que ainda não foram atribuídos a um nó e selecionar o nó mais adequado para sua execução. A seleção é baseada em múltiplos fatores, incluindo requisitos de recursos, políticas de qualidade de serviço, afinidades, anti-afinidades, e localidade de dados.

- **kube-controller-manager**: Este componente executa os processos dos controladores. Cada controlador lida com diferentes aspectos do cluster, como manutenção do número correto de pods para cada objeto de replicação, gestão da lógica de orquestração de nó, criação de endpoints de serviço, entre outros. Todos os controladores são compilados em um único binário e executados como um único processo.

- **cloud-controller-manager**: Permite a integração do cluster Kubernetes com a infraestrutura de nuvem do provedor, gerenciando aspectos específicos como balanceadores de carga, volumes de armazenamento e rotas de rede. Este componente é essencial para ambientes que operam na nuvem, permitindo o desenvolvimento independente do código específico do provedor de nuvem e do código do Kubernetes.

Esses componentes são fundamentais para a operação e gestão eficaz de um cluster Kubernetes, permitindo que ele responda dinamicamente à carga de trabalho e às condições do ambiente.

2.5 Considerações Finais

Neste capítulo foram apresentados alguns conceitos e tecnologias que são abordadas ao longo do trabalho. Dentre as tecnologias apresentadas, quinta geração de comunicações móveis (5G) e a computação na borda promovem vários casos de uso, incluindo os de aplicações na área da saúde, promovendo o desenvolvimento da de sistemas de saúde móvel e aplicações de ciber-medicina, visto que os recursos computacionais da borda podem ser acessados através de uma rede de baixa latência.

Além disso, foram apresentados os conceitos de Virtualização, onde foram apresentados o conceito da tecnologia, benefícios e tipos de virtualização. Dentre as tecnologias de virtualização, foi destacado a virtualização baseada em conteires e a ferramenta utilizada para sua criação, o docker. Esta nova tecnologia de virtualização gera o desenvolvimento de serviços cada vez mais autocontidos, escaláveis e de funcionalidades bastante específicas, conhecidos por microsserviços.

À medida que os sistemas e arquiteturas vão evoluindo, gerenciar e operacionalizar os microsserviços se torna uma tarefa bastante desafiadora e bastante custosa para manuten-

ção, sendo este o cenário característico para disponibilização de serviços de saúde em um contexto de mobilidade. Neste sentido, também foram apresentados os conceitos de Orquestração de Microsserviços e Malha de Serviços, assim como duas ferramentas amplamente utilizadas para esta finalidade, o Kubernetes e o Istio, respectivamente. Estas ferramentas foram utilizadas no ambiente de simulação explorado neste trabalho.

Capítulo 3

Revisão Bibliográfica

A revisão bibliográfica apresentada neste capítulo está dividida em duas partes. Na primeira, são exploradas as Aplicações Críticas de Saúde, incluindo casos de uso, requisitos de comunicação e sobre mobilidade em aplicações críticas de saúde.

A segunda parte aborda o tópico de Orquestração de Serviços, onde são discutidos alguns tópicos relacionados a requisitos de orquestração, alocação de serviços e mobilidade na alocação de serviços. Uma discussão a respeito dos trabalhos incluídos na revisão, abordando as suas similaridades e comparações com a proposta deste trabalho, é apresentada no final de cada uma das partes desta revisão.

3.1 Aplicações Críticas de Saúde

Em um cenário mais específico no âmbito da medicina, existe uma necessidade particular que exige, além de outras características, um ambiente de comunicação de altíssima confiabilidade: aplicações médicas críticas. Aplicação Crítica de Saúde é um termo genérico usado para descrever as aplicações e seus dispositivos relacionados usados na prestação de cuidados de sobrevivência de pacientes [1]. Uma outra categorização acerca da criticidade de aplicações de saúde é feita em [149], onde são consideradas três categorias: *(i)* aplicações de segurança crítica, onde qualquer falha ou interrupção pode resultar em ferimentos graves ou perda de vida; *(ii)* aplicações de missão crítica, para os casos em que qualquer falha ou interrupção resultará em sério impacto sobre uma organização ou negócio; e *(iii)* aplicações não críticas, que englobam todas as criticidades que não se enquadram nas duas categorias anteriores.

Ao explorar alguns cenários de aplicações críticas de saúde, um dos principais casos de

uso suportados pela próxima geração de comunicações sem fio é a telessaúde, pois permitirá a disponibilização de serviços de saúde remotamente [150]. No entanto, onde está a criticidade em aplicações de telessaúde? A resposta para essa pergunta não é simples, mas ganhou muita atenção nos últimos anos, principalmente por causa da pandemia da coronavírus (COVID-19). Dadas as características de disseminação rápida do COVID-19, a telessaúde ajudou a desacelerar a propagação do vírus sob os benefícios de minimizar as interações pessoais [151]. Muitos outros benefícios da telessaúde podem ser levados em consideração. Aplicações como sistemas de monitoramento de saúde [152–154], reconhecimento de atividade humana [155] e sistemas de monitoramento humano baseados em visão computacional [156–160] apresentam sua relevância para melhoraria da qualidade dos serviços de saúde atendendo ao princípio da monitorabilidade.

Reforçando a definição apresentada de aplicações críticas de saúde, as aplicações mencionadas anteriormente, mas não se limitando a essas, referem-se à prestação de cuidados de sobrevivência do paciente. Dada a adição de novos dispositivos médicos para melhorar a prestação de serviços de saúde, os requisitos mínimos a serem atendidos para garantir a confiabilidade e disponibilidade dos serviços para aplicações médicas críticas precisam ser mapeados, que possuem altas demandas de computação e infraestrutura de rede.

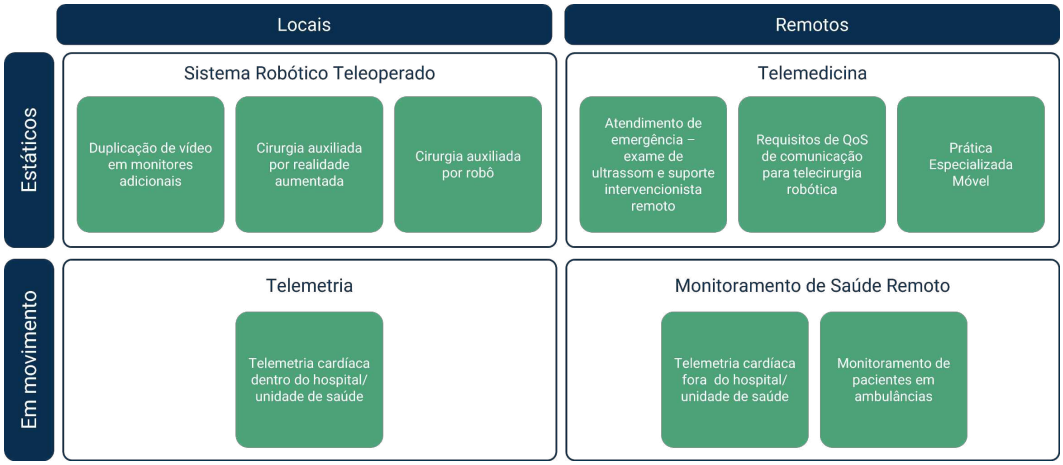
Neste sentido, um levantamento de requisitos de comunicação e casos de uso de aplicações críticas no cenário de comunicações móveis explorados na literatura foi realizado. Esses casos de uso e respectivos requisitos serão descritos nas próximas subseções.

3.1.1 Casos de Uso

Soluções e requisitos de computação devem ser atendidos para que aplicações críticas de saúde sejam confiáveis. Neste sentido, um ótimo ponto de partida para explorar casos de uso é o Relatório Técnico 3GPP 22.826 [1], onde são descritos um conjunto de casos de uso com a proposição de requisitos de comunicação para aplicações críticas de saúde. Para facilitar a exploração de novos cenários para pesquisas futuras, uma taxonomia de aplicativos críticos de saúde descritos em [1] é apresentada na Fig. 3.1. Basicamente, os casos de uso apresentados podem ser classificados sob as óticas da geolocalização, cenários onde os dispositivos e/ou soluções médicas estão localizadas localmente ou remotamente, e da movimentação, cenários onde os dispositivos e/ou soluções médicas estão estáticos ou em

movimento.

Figura 3.1: Mapeamento de casos de uso aplicações críticas de saúde



Fonte: Produzida pelo autor. Extraído de [1]

A partir desse agrupamento apresentado em [1], alguns casos de uso e características foram extraídos de trabalhos na literatura, sendo descritos nas subsessões subsequentes.

3.1.1.1 Cirurgia Robótica Tele-operada e Processamento Multimídia

A execução de uma gama de tarefas que são inviáveis ou que colocam em risco a vida de pessoas é uma oportunidade para a criação de soluções baseadas em Robôs Tele-operados (Tele-robôs). A aplicação de tele-robôs em sistemas, áreas ou processos específicos podem representar uma melhoria no desempenho geral da disponibilização de serviços, implicando em uma vasta aplicação para os tele-robôs, que podem demandar diferentes regras de negócios para seus sistemas de controle, comunicação e sensores [161].

Para o caso de aplicações críticas de saúde, um cenário de sala de cirurgia híbrida pode ser explorado. Minimizar incisões e realizar procedimentos cirúrgicos por meio de um ou vários pequenos cortes geralmente são menos traumáticos [162–164]. Desse modo, sistemas avançados de imagem podem dar suporte aos procedimentos médicos gerais nesse ambiente, por meio da extração e processamento de imagens de equipamentos de Tomografia Computadorizada, Ressonância Magnética e Raios-X, possibilitando a realização de cirurgias minimamente invasivas. Nesse cenário, dois operadores contribuem para o procedimento: um cirurgião, aquele que opera, e o assistente, responsável pelo controle do sistema de imagem.

Dependendo da posição dos operadores na sala de cirurgia, é necessária uma duplicação de vídeo em monitores adicionais, por exemplo, quando não estiverem posicionados do mesmo lado. Um exemplo de um Sistema de Cirurgia Robótica é o *da Vinci* Sistema Cirúrgico [165].

Outra tecnologia que vem sendo implementada gradativamente em sistemas de cirurgia robótica tele-operada é a Realidade Aumentada. O modelo 3D da anatomia do paciente obtido a partir de exames de ressonância magnética ou tomografia computadorizada fornecem condições para que os cirurgiões planejem o procedimento a ser realizado [1].

Nesse sentido, um requisito crítico de negócios a ser considerado é a importância de projetar bem os dispositivos médicos para fornecer serviços de alta qualidade aos pacientes. Em outras palavras, um dispositivo médico deve ser seguro e clinicamente eficaz, além de atender às necessidades dos usuários [166, 167]. Adotar novos modelos computacionais para processamento multimídia, como o paradigma de Computação em Borda, é fundamental para implementar com sucesso as aplicações de Cirurgia Robótica Tele-operada, uma vez que a borda da rede é potencializada com recursos computacionais.

3.1.1.2 Telemetria e Resposta em Tempo-Real

Ao falarmos em Telemetria, sistemas de monitoramento cardíaco é um tema relevante explorado no setor de saúde, como podemos ver em [19, 152, 154], por exemplo. Neste caso de uso, a condição do paciente é continuamente monitorada pela equipe de enfermagem através do dispositivo de telemetria de ECG sem fio que o paciente está usando. Uma solução comercial para monitoramento de ECG é o LevMed Mobile ECG [168]. Esta solução atende a esta necessidade de monitoramento remoto, no entanto, requer a intervenção de um operador ou do paciente para obtenção e envio dos dados para análise do especialista.

Para aplicações de telemetria, a capacidade de resposta em tempo-real é crucial para casos críticos. Um exemplo de um sistema de aquisição de telemetria médica inteligente foi explorado em [169]. Neste sistema, a medição da temperatura do gado foi implementada usando uma nova técnica de termometria não invasiva por infravermelho. Foram utilizados serviços de computação em nuvem e um servidor MQTT para armazenar e disponibilizar o EHR a qualquer momento e em qualquer lugar. Apesar desse sistema ter sido implementado em uma aplicação veterinária, a sua adaptação para utilização em humanos é totalmente factível.

Outro bom cenário é o monitoramento do Estresse Materno. O estresse excessivo durante a gravidez pode implicar em efeitos adversos na mãe e no feto, o que interrompe a adaptação materna ao longo da gravidez. Um sistema de monitoramento adaptativo para estimar os níveis de estresse foi apresentado em [170], que considerou a elevação da frequência cardíaca na gravidez determinante à adaptação materna. Eles propuseram um algoritmo de estimativa do nível de estresse em tempo-real com base na frequência cardíaca e nas variações da frequência cardíaca durante a gravidez.

O monitoramento em tempo-real dos pacientes é, também, um dos aspectos mais críticos para a equipe paramédica cuidar adequadamente dos pacientes, principalmente em situações de emergência. Nesse sentido, o projeto e a implementação de um sistema de monitoramento de saúde em tempo real, utilizando tecnologias 5G, representam boas oportunidades, atendendo aos requisitos rigorosos de confiabilidade e latência das aplicações.

3.1.1.3 Telemedicina e Alta Largura de Banda

Telemedicina refere-se ao uso de Tecnologias de Informação e Comunicação (TIC) com o objetivo de aumentar o acesso aos cuidados e informações médicas para melhorar as condições de saúde dos pacientes [171]. Uma grande contribuição da Telemedicina é a dissociação entre localização e qualidade do atendimento. Também ajuda a economizar inúmeras horas para médicos e cirurgiões, programando seus horários de forma eficiente em salas de cirurgia, locais de incidentes e centros médicos, em vez de ter que estar fisicamente presentes.

No trabalho desenvolvido em [172], foram apresentados casos de uso usando tecnologias de Realidade Aumentada (AR). Primeiro, um caso de uso de monitoramento robótico remoto baseado em AR (AR Bots) foi explorado como parte da solução de hospitais inteligentes para combater a pandemia de COVID-19. Um fato observado no início da disseminação do COVID-19 foi o contágio de profissionais de saúde, durante o atendimento aos pacientes [173]. Assim, os AR Bots podem ser controlados remotamente em estações de controle, reduzindo o contato de pacientes infectados com profissionais de saúde. O segundo caso de uso foi a aplicação da AR em cirurgia remota. As emergências se beneficiarão dessas aplicações dada a possibilidade de realização de cirurgias remotas quando um cirurgião especialista não estiver disponível localmente ou em casos de prevenção da equipe sob exposição a um determinado patógeno durante a realização da cirurgia.

Três cenários diferentes de Telemedicina também são vislumbrados em [1]: Atendimento de emergência, onde o exame de ultrassom apoiado por intervenção remota melhora positivamente a gestão do atendimento pré-hospitalar; na Tele-cirurgia Remota Estática, é possível ter acesso ao conhecimento nas mais diversas especialidades cirúrgicas a nível mundial sem precisar viajar; A partir de uma perspectiva de prestação de cuidados de saúde centrada no paciente, é introduzido, também, o conceito de Prática Móvel Especializada, onde caminhão equipado com uma sala para realização de exames e um conjunto de sistemas médicos avançados de imagem podem ser transportados para uma determinada cidade que pode carecer de serviços médicos ou que esteja passando por um pico de demanda de serviços médicos.

Os avanços nas tecnologias de comunicação são relevantes para promover o desenvolvimento desses casos de uso, principalmente falando de alta largura de banda, dadas as necessidades de *streaming* multimídia. Investigar os valores mínimos de operação desses requisitos é importante para viabilizar a implementação dessas soluções.

3.1.1.4 Monitoramento Remoto de Saúde e Análise de Dados

O monitoramento da saúde do paciente através da aquisição contínua de dados pode melhorar amplamente a prestação de serviços de saúde, uma vez que podemos aplicar técnicas de análise de dados para identificar anomalias nos registros de dados de saúde (EHR) dos pacientes. O desenvolvimento de pesquisas em diversas áreas aplicados ao setor da saúde contribuem para a realização de sistemas de Monitoramento Remoto de Saúde. Podemos citar as tecnologias de Redes Corporais (WBANs), a Internet das Coisas Médicas (IoMT), sensores biomédicos e sistemas embarcados de baixo consumo [174–179].

Diversas de aplicações de saúde podem se beneficiar com a adoção das tecnologias mencionadas. Por exemplo, em [180], foi proposto um equipamento de monitoramento de parâmetros fisiológicos em tempo-real, integrado a um detector de parâmetros fisiológicos e monitor corporal inteligente aplicado ao rastreamento de condicionamento físico, permitindo aos usuários desfrutar de soluções de monitoramento de condições saúde em tempo-real personalizada. Outra aplicação do sistema de monitoramento de saúde foi explorada em [181], onde um analisador de urina portátil contendo um sensor bioquímico foi utilizado como parte de um sistema de monitoramento remoto para avaliar a condição de pacientes.

Em cenários de cuidados intensivos, monitorar e fornecer cuidados contínuos a pacientes

feridos em uma ambulância em movimento é um caso de uso a ser explorado. Enquanto os pacientes são transportados para o hospital mais próximo, sua estabilização ainda ocorre no local do incidente. Além disso, será possível coletar e compartilhar dados críticos de pacientes em tempo-real por equipes de emergência com o hospital antes que eles cheguem. Assim, a equipe de emergência estará melhor preparada para receber o paciente, reduzindo o tempo de transferência [1].

A computação pervasiva está cada vez mais presente em nosso dia a dia. Nesse sentido, o grande volume de dados gerado pela introdução de tecnologias como IoMT, biossensores e WBANs representa um excelente cenário para a aplicação de técnicas de data analytics, haja vista a grande quantidade de dados de saúde coletados diariamente, abrindo inúmeras novas oportunidades e desafios na área da saúde.

3.1.2 Requisitos de Comunicação

A quinta geração de redes de comunicações móveis (5G) tem o potencial de dar suporte a novas aplicações de saúde e melhorar as aplicações já existentes. A integração do 5G com tecnologias digitais de saúde pode facilitar a prestação de serviços de saúde com recursos de comunicação expandidos, proporcionado pela alta velocidade de transmissão dados do 5G, ultrabaixa latência, conectividade massiva de dispositivos, confiabilidade, maior capacidade de rede e maior disponibilidade. [182–184]

No entanto, os casos de uso de saúde baseados em tecnologias 5G possuem uma variedade de requisitos técnicos de comunicação. Conhecer esses requisitos é de fundamental importância para todo ecossistema de soluções de tecnologia na área de saúde, que envolvem provedores de rede, autoridades regulatórias, desenvolvedores e fabricantes de produtos e aplicações hospitalares, com um interesse comum em facilitar a assistência médica de modo seguro e eficaz, onde é necessária uma compreensão dos requisitos de comunicação para correta seleção de tecnologias sem fio com recursos que suportem os requisitos das aplicações de saúde e o desempenho esperado. [185]

No trabalho apresentado em [186], são discorridos os requisitos e principais tecnologias habilitadoras de serviços avançados de saúde baseados em futuros sistemas de comunicação móveis, onde são identificados, também, os cenários de saúde mais representativos que podem se beneficiar das redes 5G e, por sua vez, são sintetizados os requisitos de comunicação

nesses cenários específicos. Embora as redes 5G ofereçam avanços para o suporte de comunicação de serviços de *mHealth* e *eHealth*, futuras aplicações e casos de uso podem exigir desempenhos de rede ainda mais desafiador. Por exemplo, visando proporcionar uma experiência ainda mais imersiva e natural que pode ser útil em aplicações de cirurgias tele-operadas e, também, na prestação de serviços de primeiros socorros remotos, no cenário de ambulâncias conectadas, a realidade estendida pode ser utilizada. A introdução desta tecnologia em aplicações de saúde aumenta significativamente os requisitos de latência, largura de banda e confiabilidade.

Identificar, descrever e comparar requisitos quantitativos e indicadores-chave de desempenho de comunicação em casos de uso de saúde foi o principal objeto de estudo do trabalho apresentado em [56]. Nesse estudo, foi possível observar que a comparação dos requisitos de saúde 5G com o status dos recursos 5G existentes revela que algumas aplicações de saúde podem ser suportadas pelos serviços 5G disponíveis atualmente, enquanto outros podem apresentar grandes desafios para serem alcançados, especialmente aquelas com estritos requisitos de latência. Algumas lacunas na literatura relacionadas a requisitos de comunicação para aplicações médicas também foram mapeadas, uma delas é que grande parte da literatura existente aborda qualitativamente os requisitos de comunicação para as aplicações de saúde adotando o uso de palavras como “grande”, “pequeno” e “extremamente baixo”, por exemplo.

Ambos os trabalhos apresentados em [186] e em [56] exploram e sumarizam características, requisitos e indicadores de desempenho para vários casos de uso de aplicações de saúde. Além disso, em seu Relatório Técnico 3GPP TR 22.826 v17.2.0 [1], a organização de padronização 3GPP publicou especificações associadas a aspectos de serviços de comunicação para aplicações críticas de saúde. A fins de direcionamento do caso de uso a ser explorado neste trabalho, serão explorados os requisitos de comunicação para o caso de uso de Ambulâncias Conectadas. Este cenário representa um estudo de caso relevante para explorar alguns dos principais desafios relacionados à saúde para a infraestrutura 5G em termos de latência, confiabilidade e mobilidade.

Para explorar o caso de uso de ambulâncias conectadas, em [1] é proposto um cenário de emergência suportado por exames de ultrassom e suporte intervencional remoto. A utilização de equipamentos de imagem por ultrassom podem melhorar substancialmente a assistência

médica em casos de emergência [187], uma vez que a principal missão do atendimento pré-hospitalar está associada ao fornecimento de suporte rápido e de alta qualidade e transportar os pacientes para o hospital, evitando complicações. Lidar com situações de emergência pode demandar a atuação de profissionais especialistas a depender da necessidade específica do paciente em atendimento, como cirurgiões, ortopedistas, cardiologistas, pediatras, entre outros.

A estabilização e administração dessas condições na maioria dos casos começa antes da chegada ao hospital. Um dos exemplos mais importantes a ser mencionado está associado à redução do "tempo porta-balão"(door-to-balloon time) para o tratamento do infarto do miocárdio com supra-desnivelamento do segmento ST (IAMCST). [188]. O "tempo porta-balão"representa o intervalo de tempo entre a chegada do paciente ao hospital e a primeira insuflação do cateter balão ou liberação do stent dentro da artéria coronária. Essa medida de tempo no atendimento cardíaco de emergência é tomada como referência para o reestabelecimento do fluxo sanguíneo do paciente [189, 190].

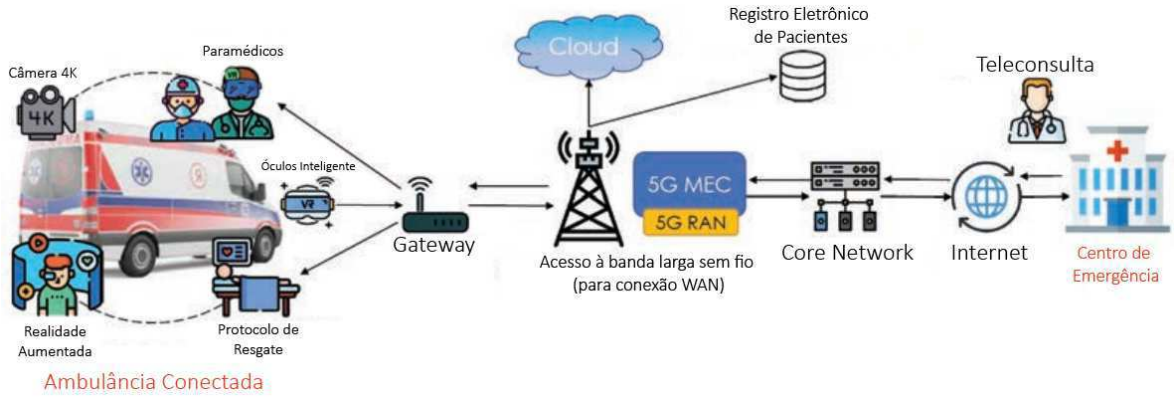
Por meio da Banda Larga Móvel Melhorada (eMBB) e da Comunicação Ultraconfiável de Baixa Latência proporcionadas pelas redes 5G, será possível a realização de uma teleorientação mais eficaz entre paramédicos na ambulância e especialistas remotos, visando a melhoria significativa do diagnóstico, o tratamento e a conformidade em casos de emergência médica. Nesta situação, os paramédicos na ambulância poderão realizar a captação de imagens baseadas em tecnologias de ultrassom e realizar exames tão precisos quanto os realizados pelos médicos.

Adicionalmente, um outro grande benefício está associado ao melhor direcionamento do paciente para o centro especializado mais adequado à sua condição, sendo possível informar a equipe do hospital sobre as lesões específicas, direcionando os preparativos adequados à situação do paciente, por exemplo, preparar a sala de cirurgia e chamar toda a equipe médica necessária, visando reduzir o tempo entre o momento do acidente e o do tratamento e garantir o transporte com maior segurança. O hospital pode estar longe do paciente e da ambulância, de modo que um paciente pode não sobreviver ou pode sofrer danos permanentes durante o transporte se uma interação em tempo-real entre paramédicos e especialistas remotos não for acionada imediatamente no local da intervenção. Nesses casos de uso, atenção especial é dada à segurança, confiabilidade da comunicação e confidencialidade dos dados do paciente.

Na prática, os dados que estão sendo comunicados precisam ser totalmente criptografados e protegidos por integridade, o que requer largura de banda de computação adicional em ambas as extremidades. [1]

A adição de novas tecnologias para realização do monitoramento de sinais vitais dos paciente, varredura e mapeamento do cenário do acidente por meio de câmeras de alta definição, que podem estar integradas a óculos de realidade aumentada, e o acesso em tempo-real ao histórico de registros médicos do paciente pode ajudar significativamente a reduzir o tempo de atendimento e assistência pré-hospitalar de pacientes em situações de emergência, por meio de um atendimento especializado e mais direcionado às necessidades particulares em questão. Para este cenário de ambulâncias conectadas, serão necessários dispositivos para monitoramento, exame e intervenções guiadas como, por exemplo, por meio de sonda de ultrassom, monitores de sinais fisiológicos e óculos inteligentes 4K portáteis. Também o acesso instantâneo aos registros médicos é importante para entender a condição do paciente antes do incidente. Na tabela 3.1 estão apresentados os requisitos de comunicação propostos em [1] para o caso de uso de Ambulâncias Conectadas.

Figura 3.2: Diagrama do cenário de Ambulâncias Conectadas



Fonte: Adaptado de [186]

A Figura 3.2 ilustra o cenário de Ambulâncias Conectadas apresentado em [186]. Apesar de passar um panorama geral do cenário, tecnologias e componentes arquiteturais de alto nível para a implementação do caso de uso de Ambulâncias Conectadas, no trabalho apresentado em [186] não são explorados os requisitos de comunicação baseado nos dispositivos integrantes deste cenário, podendo-se citar o ambiente de Realidade Aumentada, a utiliza-

ção de câmeras de alta definição e óculos inteligentes, que poderão demandar alta largura de banda e comunicação em tempo-real à rede de comunicação sem fio.

Tabela 3.1: Requisitos de comunicação para o cenário de Ambulâncias Conectadas propostos em [1].

Caso de Uso	Aplicações	Disponibilidade do serviço de comunicação: valor alvo em%	Confiabilidade do serviço de comunicação: Tempo médio entre falhas	Latência ponto-a-ponto: Máxima	Taxa de Transmissão	Direção
Monitoramento de Saúde Remoto: monitoramento de pacientes em ambulâncias.	Stream de vídeo em tempo-real em 4K com compressão (3840x2160 pixels) 60 fps 12bits por pixel (ex.: YUV4:1:1)	99.99	>1 mês	<100ms	25 Mbits/s	Upstream
	Stream de vídeo em tempo-real de sonda de ultrassom sem compressão 2048x2048 pixels 32 bits por pixel 20fps	99.999	>>1 mês (<1 ano)	<100ms	670 Mbits/s	Upstream
	Stream de dados de monitoramento de sinais vitais	99.999	>>1 mês (<1 ano)	<100 ms	1 Mbits/s	Upstream
	Stream de áudio de alta qualidade	99.99	>1 mês	<100 ms	128 kbits/s	Upstream/ Downstream

Requisitos de comunicação para assistência médica móvel em geral são discutidos em [191], onde foi implementada uma conectividade bidirecional entre ambulâncias e hospitais em todo o Reino Unido. Neste trabalho são definidas a máxima latência de ponta-a-ponta permitida para diferentes tipos de dados, assim como a taxa de transmissão e a relação entre o número de pacotes perdidos e o número total de pacotes enviados, conhecida por *Packet Loss Ratio* (PLR). A máxima latência observada para *streaming* de áudio e vídeo é de 150 ms, já para transmissão de dados de sinais vitais, 250 ms, e menos de 10 ms foi o valor definido para aplicações de força e vibração, como é o caso de aplicações que utilizem o *feedback* háptico. Se tratando dos requisitos relacionados à taxa de transmissão para diferentes, as aplicações de *streaming* de vídeo demandavam 10 Mbps, seguido de 400 Kbps para *streaming* de dados força e vibração e, em seguida, *streaming* de áudio com um requisito de 200 Kbps. Além disso, diferentes taxas de transmissão foram mapeadas para diversos tipos de sinais vitais, onde as aplicações de EEG apresentaram o maior requisito entre eles, sendo de até 86,4 Kbps. No tocante à PLR, foram mapeados os valores mínimos de perdas permitidos, sendo 10% para aplicações de áudio, 0,1% para aplicações de *streaming* de vídeo e de *streaming* de dados de sinais vitais e, por fim, um valor de aproximadamente 0,01% para aplicações de *streaming* de dados força e vibração. Os autores ainda ressaltam que com a adoção do URLLC, oferecendo uma confiabilidade de até 10^{-5}, os requisitos de PLR serão atendidos em as aplicações de ambulâncias utilizando redes 5G.

Requisitos baseados em medições para *streaming* de imagens de ultrassom de alta definição, *streaming* de vídeo 4K e dados de sinais vitais, também, foram estudados em [192]. As taxas de transmissão definidas para aplicações de *streaming* de imagens de ultrassom foram maiores que 20 Mbit/s para *uplink* e maiores que 5 Mbit/s para *downlink*, além da latência inferior a 80 ms. Para o caso de *streaming* de vídeo, os requisitos foram de 20 Mbit/s para *uplink* e *downlink* e a latência inferior a 50ms. Por fim, os requisitos de *streaming* de dados de sinais vitais foram de 4 Mbit/s para *uplink*, 2 Mbit/s para *downlink* e latência inferior a 50 ms. Baseado nos valores medidos, foram obtidos valores de 1.361,21 Mbit/s para download dentro da ambulância e 257,52 Mbit/s para *upload*. A solução apresentada utiliza tecnologias de comunicação 5G e uma arquitetura de computação em borda, no entanto é adotada uma abordagem de implantação de rede privada de emergência 5G para que seja garantida a transmissão de vídeo 4K HD dentro do veículo sem congestionamentos e atrasos.

3.1.2.1 Requisitos de Latência em ambientes de Computação em Nuvem versus Computação em Borda

Neste sentido, a investigação e avaliação de latências de servidores de Computação em Nuvem é fundamental. No trabalho apresentado em Charyyev et al. [193], foi realizada uma medição em larga escala para comparar as latências dos usuários finais a servidores em nuvem e em borda. Observa-se que os servidores de borda fornecem latência menor para 92% a 97% dos usuários finais em comparação com diferentes provedores de nuvem. Para a maioria significativa dos usuários finais, os servidores de borda estão mais próximos que os provedores de nuvem com uma diferença de latência de 10 a 100 ms. Baseado nos resultados apresentados, os servidores de borda podem fornecer latência consideravelmente menor do que os servidores de nuvem locais para a maioria significativa dos usuários finais. Além disso, a análise de latência baseada em regiões onde os *datacenters* estão implementados revela que os locais de nuvem podem corresponder ao desempenho de servidores de borda onde os data centers são abundantes, como na Europa Ocidental.

Em Yang et. al [194], foi realizado um comparativo entre quatro provedores de serviços de Computação em Nuvem de 200 nós do *PlanetLab* [1]. Para isso foram analisados alguns serviços em nuvem, como memória de instância virtual e E/S de disco, armazenamento de BLOB (do inglês, *Binary Large OBject*) e fila, e transferência de rede entre *datacenters*. No que se refere à medição de latências, para realizar o download de um BLOB de 1KB, por exemplo, foram necessários mais de 300 ms. Observou-se, também, que o número de *datacenters* e suas localizações desempenham um papel crucial na latência.

Outro estudo para medir e observar da latência entre regiões entre provedores para sistemas IoT baseados em nuvem foi apresentado em Vu et al. [195], onde foi observada uma latência em média de 100 ms nas análises realizadas.

No trabalho desenvolvido por Chen et al. [196], foi realizado um estudo empírico para analisar o desempenho de latência para aplicações de Computação na Borda para assistência cognitiva vestível. Além disso, foi mostrado que uma latência adicional de 100 a 200 ms foi introduzida ao realizar-se o *offloading* de serviços para os servidores em nuvem, em comparação com o *offloading* para um dispositivo de borda próximo.

[1] http://www.planet-lab.org/

3.1.3 Mobilidade em Aplicações Críticas de Saúde

A definição de mobilidade está relacionada à capacidade de locomoção com facilidade. Ela está intrinsecamente associada a quase todas aplicações do nosso dia a dia, principalmente devido aos avanços tecnológicos das redes de comunicação móveis. É possível identificar um crescimento acentuado no desenvolvimento e uso de tecnologias de comunicação móveis anos últimos anos, motivado pela busca de um novo padrão de vida associado à comodidade, conforto e mobilidade. [197]

No contexto da saúde, as tecnologias de comunicação móveis são capazes de proporcionar serviços de saúde em nível individual aos usuários. As estruturas de saúde móvel (mHealth) podem se apoiar na adoção e integração de diversas tecnologias, por exemplo, a utilização de *smartphones* no contexto de computação pervasiva, integradas com redes de sensores corporais sem fio (WBANs), para disponibilizar serviços de monitoramento de condições de saúde e acesso a atendimento médico quando necessário a pacientes. [198]

Se tratando de cenários de aplicações criticas de saúde, as aplicações de soluções de comunicações móveis se tornam fundamentais em cenários de prestação de serviços de emergência, podendo-se citar, o caso de uso de ambulâncias conectadas, tema bastante explorado na literatura e relevante para a sociedade. Em [199] foi realizado um estudo investigativo sobre as barreiras, facilitadores e demandas que afetam os profissionais de saúde em um ambiente integrado de telemedicina baseado em unidade de saúde móveis para atendimento de acidente vascular cerebral. Por meio desse estudo foi constatado que a implementação da telemedicina em ambulâncias pode reduzir o tempo de tratamento em pacientes com AVC, variável fundamental para minimizar os riscos de sequelas nos pacientes. Além disso, a comunicação e trabalho em equipe durante o atendimento de AVC por telemedicina em uma ambulância é fundamental para um atendimento bem-sucedido de pacientes com AVC. A adoção da telemedicina proporciona melhorias associadas à comunicação verbal e não verbal entre todos os membros da equipe usando sistemas de vídeo e áudio para fornecer cobertura efetiva do paciente para os médicos e vice-versa [200].

O papel da telemedicina no atendimento pré-hospitalar para o atendimento de pacientes com AVC em unidades móveis a caminho do hospital foi também explorado em [201]. Este tema fomenta o desenvolvimento de soluções utilizando redes de comunicações móveis como é possível observar no trabalho desenvolvido em [42], onde foi avaliada uma ambulân-

cia conectada em ambiente de redes 5G e *Network Slicing* usando uma plataforma de teste. A proposição de uma arquitetura para ambulância conectada é apresentada em [192], permitindo que os médicos do hospital prestem assistência e orientações remotamente à equipe da ambulância para fornecer cuidados médicos iniciais aos pacientes. A adoção de soluções de tecnologia móvel de quinta geração no atendimento de emergência pré-hospitalar pode ser, sem sombra de dúvidas, uma oportunidade para apoiar os paramédicos no transporte de pacientes com AVC em ambulâncias da área rural para o hospital, principal objeto de estudo realizado em [202].

A melhoria do desempenho do serviço de emergência de unidades de saúde móveis pode implicar em reduções significativas no índice de mortalidade para este tipo de atendimento. Neste sentido, uma das características singulares das aplicações de ambulâncias conectadas diz respeito à mobilidade, que, por sua vez, aumenta os problemas de conectividade ocasionados em veículos em movimento em alta velocidade. Dentre esses problemas, pode-se mencionar a baixa qualidade de sinal, perda de penetração de paredes do veículo, múltiplos *handovers* e maiores ocorrências de quedas de conexão [56].

Em [203] é proposto um sistema de saúde inteligente para melhorar as métricas de desempenho do serviço de emergência de ambulância, onde foram utilizadas as informações de tráfego em tempo-real e o tempo de espera de hospitais para reduzir o tempo de resposta da ambulância, o tempo de viagem da ambulância para os hospitais e o tempo de espera nos hospitais. Já em [204], uma solução baseada em Big Data aplicada em um estudo de caso de engenharia de sistemas médicos foi proposta com o objetivo de prever futuras localizações de ambulâncias para superar desafios baseados em mobilidade, onde o algoritmo proposto, denominado NextSTMove, apresentou um desempenho 300% maior que os algoritmos tradicionais, ainda alcançando uma precisão de 75% a 100%.

Visando superar esses desafios, no estudo apresentado em [205] foi avaliado o fluxo de dados no *uplink* entre uma ambulância e os nós do hospital com uma *small cell* dentro da ambulância viajando a uma velocidade de 120 km/h. No ambiente simulado, foi instalado um transceptor no teto da ambulância para transmissão e recepção de dados com a rede de macrocélulas de *backhaul*. A partir disso, foi possível estabelecer uma conexão sem fio entre a *small cell* instalada dentro da ambulância e o *small cell access point* (SAP). Para esta solução, o valor de PLR ao usar a *small cell* foi reduzido para 4,8% quando comparado

a 14% no caso de 10 usuários tentando se conectar à estação base de macrocélula externa, onde todos os 10 usuários estavam localizados na mesma ambulância. Foi possível observar, também, a melhoria na taxa de transferência com o uso da *small cell*. A conclusão dos autores para a adoção de *small cells* dentro da ambulância é que esta solução se pode ser bastante útil em cenários de congestionamento de alta largura de banda.

3.1.4 Discussão

A incorporação de tecnologias de comunicação móveis modernas no âmbito da digitalização de serviços na área da saúde é bastante promissor, dada as características de comodidade, conforto e mobilidade, proporcionando um novo padrão de vida. À medida que as tecnologias e aplicações médicas vão evoluindo, a busca por soluções cada vez mais centradas nos pacientes apresenta novas oportunidades e desafios associados, motivando o desenvolvimento de soluções com requisitos de comunicação e computação cada vez mais restritos.

Ao analisar aplicações críticas de saúde, é possível identificar restritos requisitos de latência, confiabilidade, disponibilidade e taxa de transmissão, ainda considerando a segurança e privacidade dos dados, uma vez que dados sensíveis de pacientes serão transmitidos. Tecnologias emergentes como 5G e Edge Computing se apresentam como soluções habilitadoras na promoção do desenvolvimento de soluções de tecnologias da informação e comunicação que sirvam de suporte às aplicações críticas de saúde.

Lidar com a mobilidade em cenários de emergência, que podem ser classificados como aplicações críticas de saúde, é inevitável, uma vez que a coleta e transmissão contínua de dados do paciente começará quando os paramédicos da ambulância de emergência chegarem ao local do incidente até a entrega do paciente ao departamento de emergência do hospital de destino [1]. Atrelado a isso, é possível observar a necessidade de transmissão de dados em tempo-real dos dados do paciente em atendimento para maior celeridade no atendimento. O evento do incidente pode ser visto como um evento esporádico, uma vez que não é possível prever onde e quando acontecerá.

Além disso, ao longo de um determinado trajeto, o deslocamento da ambulância seguirá, a priori, um padrão aleatório de rota. Considerando essas quatro variáveis (mobilidade, transmissão de dados em tempo-real, eventos esporádicos e rotas aleatórias), uma possível solução pode está associada à alocação dinâmica de serviços em servidores de borda, que

apresentam uma topologia de computação distribuída. Além disso, a limitação de recursos em servidores de borda exige soluções dinâmicas e eficientes de alocação de serviços e recursos, uma vez que o suporte à mobilidade deve ser parte da solução.

3.2 Orquestração de Serviços

Com o aumento no volume do tráfego da rede com os mais diversos requisitos de Qualidade do Serviço (do inglês, *Quality of Service* - QoS), vários desafios para o provisionamento de serviços de ponta-a-ponta são postos, dentre eles está o gerenciamento de recursos de rede e diferenciação de serviço são duas funcionalidades cruciais empregadas pelos provedores de serviços, essenciais para atender os requisitos do usuário final. [206]

À medida que as aplicações ganham escala, gerenciar gerenciar vários servidores, fluxos de trabalho e tarefas complexas se torna um desafio. Neste sentido, surge a necessidade da orquestração de serviços, que envolve tarefas de configuração, gerenciamento e coordenação automatizada de serviços, aplicações e sistemas de computador. Para desempenhar essa função são utilizados os orquestradores. [207]

Um orquestrador é um coordenador que fornece o controle de diferentes serviços, processos ou *threads* computacionais. Um orquestrador descreve o arranjo automatizado, coordenação, gerenciamento de recursos e é frequentemente discutido como tendo um controle inerentemente inteligente ou mesmo implicitamente automático. [208]

Velasquez et al. [209] propuseram alguns orquestradores para arquiteturas de Internet das Coisas baseadas em ambientes de Computação em Borda. Dentre os principais desafios que precisam ser abordados em ambientes de Computação na Borda, os autores destacam: agendamento de serviços (do inglês, *service scheduling*); computação de caminho (do inglês, *path computation*); descoberta e alocação (do inglês, *service discovery and placement*), interoperabilidade, latência, resiliência, previsão e otimização, segurança e privacidade.

De acordo com Fakude et. al, [210] o orquestrador não apenas melhora o desempenho do ambiente de computação, mas também adiciona uma camada extra que atua como controladora de toda a arquitetura, ou seja, o orquestrador mantém toda a visão geral da rede, que oferece possibilidades de implementação de várias funcionalidades associadas à segurança, escalonamento e gerenciamento de recursos, e monitoramento do sistema.

3.2.1 Requisitos de Orquestração

O orquestrador deve ser capaz de manter baixas latências, alta resiliência e confiabilidade de acordo com os requisitos das aplicações e expectativas dos usuários, mesmo em momentos de sobrecarga infraestrutura, dadas situações de cargas pesadas de tráfego. A questão principal aqui é que essas cargas são intrinsecamente específicas para aplicações específicas e não podem ser reutilizadas de outros cenários ou domínios. [209]

Conforme destacado na seção 2.4, de acordo com Ghofrani et al. [131], uma das principais características da arquitetura de microsserviços é a decomposição de aplicações complexas em componentes menores, onde são normalmente mais fácil de serem desenvolvidos, gerenciados e mantidos do que em um aplicação monolítica. No entanto, à medida que as aplicações são divididas em diversos componentes menores, sua implantação e gestão em produção se torna mais complexa.

Em aplicações baseadas microsserviços, as aplicações e serviços em contêineres devem ser capazes de aumentar e diminuir com base nos requisitos definidos de recursos disponíveis. Para isso, uma boa estrutura de gerenciamento e agendamento com eficiência dos contêineres é fundamental para atender aos requisitos das aplicações. Ao escolher uma ferramenta de orquestração de contêiner ou um serviço de orquestração de contêiner gerenciado, Wilson [211] sugere os principais pontos a serem considerados:

- Rede
- Alta disponibilidade
- Facilidade de implantação e manutenção
- Escalabilidade
- Descoberta de Serviço
- Segurança e Conformidade
- Suporte (Comunidade e Empresa)
- Sobrecarga Administrativa

Orquestração é um tema bastante relevante que tem mobilizado a indústria e a academia. Neste sentido, requisitos de orquestradores podem estar associados a várias camadas de gestão de infraestrutura, que envolve desde a gestão de infraestrutura de rede, por meio de Funções de Redes Virtualizadas (do inglê, *Network Function Virtualization* - NFV) e Redes Definidas por Software (do inglês, *Software Defined Networks*) à gestão a nível de aplicações. Segundo Carvalho e Araujo [212], atualmente, orquestração é uma palavra muito usada, sendo apresentada em diversos cenários para indicar o gerenciamento do ciclo de vida de um ou mais componentes distribuídos que juntos entregam um serviço ou funcionalidade.

Desse modo, visando elencar os requisitos de orquestradores associados às aplicações críticas de saúde, foram encontradas poucas referências com esta finalidade. Ainda, as que foram encontradas tinham finalidades distintas.

Em termos de gestão de recursos, por exemplo, podem ser explorados requisitos associados aos mecanismos que tratam de tarefas como Agendamento, Computação de Caminhos, Descoberta e Alocação e Interoperabilidade. Se falando de performance, podem ser abordados requisitos de latência e resiliência. Explorar requisitos e políticas para lidar com segurança, privacidade, acesso e autenticação é fundamental para o gerenciamento de segurança.

No documento apresentado em [213], cujo objetivo é a coleta de requisitos e *benchmarking* de desempenho de microsserviços para o projeto COLA[2], são mencionadas algumas métricas associadas à escalabilidade, tais como: tempos de criação e liberação da infraestrutura e tempos de *scale up* e *scale down*.

Já no trabalho apresentado em [214], são descritas lições aprendidas, descobertas e recomendações na construção de um ambiente de emulação adequado para experimentação para executar o Kubernetes em redes táticas. Neste estudo, foram comparados os desempenhos de 3 versões do Kubernetes: a versão padrão (K8S), a versão leve Kubernetes Lightweight (K3s), e a versão Kubernetes Native Edge Computing Framework (KubeEdge). Dentre outras métricas exploradas, as relacionadas à implantação de aplicações foram uma boa referência para este trabalho. Para este ensaio, o K8S foi o mais rápido (mas apenas um pouco mais rápido que o KubeEdge) em completar a fase de *bootstrap* sem efeitos de comunicação. À medida que as condições de rede pioram, o KubeEdge teve o melhor desempenho, com K8s mais rápidos que K3s. Por exemplo, K8s, K3s e KubeEdge levaram 46 s, 71 s e 46 s,

[2]Cloud Orchestration at the Level of Application - https://project-cola.eu/cola-project/

respectivamente, para implantar 10 réplicas por *worker* no cenário de LAN (5 trabalhadores no total).

Ao analisar métricas de performance, pode-se olhar para velocidade, confiabilidade e disponibilidade. Neste sentido, podem ser avaliadas métricas de velocidade de agendamento de contêineres e de tolerância a falhas, por exemplo. Desse modo, em [215] é feito um comparativo de desempenho do tempo de inicialização do *pod* em *clusters* Kubernetes de 100 nós que estão 10%, 25%, 50% e 100% cheios, conforme apresentado na Tabela 3.2.

Tabela 3.2: Desempenho do tempo de inicialização do *pod* em *clusters* Kubernetes de 100 nós.

Percentil	**10%-Cheio**	**25%-Cheio**	**50%-Cheio**	**100%-Cheio**
50º percentil	0,90 s	1,08 s	1,33 s	1,94s
90º percentil	1,29s	1,49s	1,72s	2,50 s
99º percentil	1,59s	1,86s	2,56 s	4,32 s

Com isso, destaca-se a necessidade de investigar outros parâmetros associados ao tempo de subir um serviço, ainda considerando a interconexão entre os microsserviços, que podem demandar um tempo maior para que a aplicação esteja disponível da íntegra.

3.2.2 Alocação de Serviços

O desenvolvimento de novas aplicações móveis têm ganhado cada vez mais notoriedade em diversos cenários e negócios fortemente baseados no crescimento explosivo de dispositivos móveis. Esses dispositivos são capazes de suportar uma infinidade de aplicativos, o que resulta em redução de energia e desempenho limitado [216]. A adoção de soluções que exijam pouco esforço computacional tem sua relevância no contexto de economia de energia em dispositivos móveis, que podem adotar rotinas de *offloading* de serviços a serem executados em servidores de nuvem e/ou de borda. No entanto, devido às limitações de recursos computacionais de servidores de borda, adotar estratégias eficientes na alocação de recursos se apresenta como uma alternativa bastante atrativa [217].

As estratégias de alocação de serviços para ambientes baseados em tecnologias de computação em nuvem são amplamente encontrados e explorados na literatura. Contudo, as

abordagens utilizadas são, em sua maioria, centralizadas e, portanto, não são adequadas para sistemas de borda descentralizados [218]. O problema de alocação de serviços tem sido bastante discutido na literatura e diversas propostas têm surgido. Baseado em diferentes cenários de aplicações, suposições de características de rede e resultados esperados, essas soluções geralmente são difíceis de comparar umas com as outras [219]. Pesquisas recentes abordam os problemas de gerenciamento de recursos e alocação de serviços. No trabalho apresentado em [220], além de ser discutido conceitos relacionados à Computação em Borda e paradigmas de computação relacionados, são realizadas discussões acerca do gerenciamento de recursos e implantação de serviços (orquestração e migração) no ambiente de Computação em Borda. Já no trabalho apresentado em [221], é discutido brevemente o problema de gerenciamento de serviços no contexto de sistemas computação distribuídos.

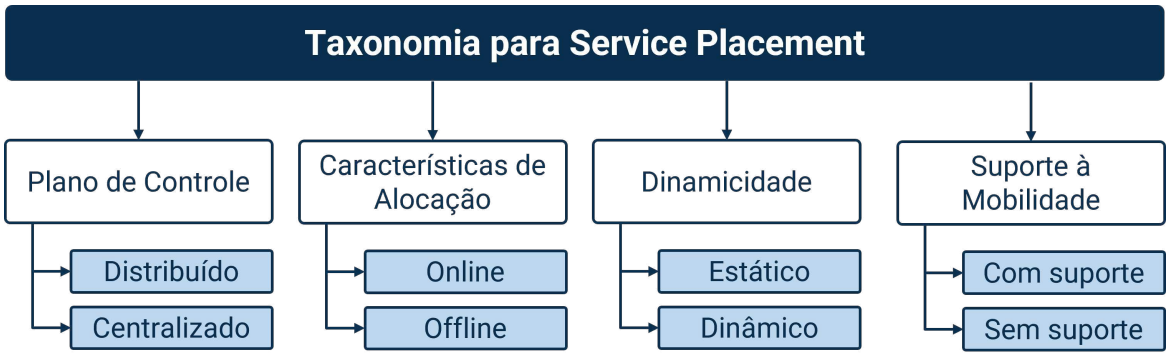
A Computação na Borda é uma tecnologia emergente utilizada para suportar aplicações sensíveis a atrasos na borda de redes móveis. Um dos desafios críticos para a Computação na Borda é determinar a seleção eficiente do servidor de borda e da entidade de serviço a ser executada [222]. A distribuição geográfica em grande escala e a heterogeneidade dos nós computacionais de borda tornam a alocação de serviços em tal infraestrutura uma questão desafiadora. Uma das principais barreiras para a adoção da Computação na Borda está associada a como alocar serviços de forma eficiente nos nós de Borda disponíveis [219]. Alocação de serviços, um dos temas centrais no MEC, ainda é um problema em aberto [46]. No cenário de MEC, uma das principais questões a serem consideradas é como garantir a continuidade do acesso dos usuários aos serviços durante o movimento [58].

Em aplicações de mobilidade, a utilização cooperativa de servidores de borda próximos é de fundamental importância em cenários com determinadas restrições de latência. Sem a cooperação de servidores de borda próximos, a transmissão dos resultados dos serviços pode ser interrompida. Assim, os serviços precisam ser migrados dinamicamente entre vários servidores de borda para garantir as métricas de desempenho do serviço. Além disso, é difícil atender a aplicações com determinadas restrições de latência com um único servidor MEC. Neste sentido, a alocação dinâmica de serviços em vários servidores na borda da rede pode aumentar a robustez das aplicações por meio de estratégias eficientes de *offloading* dos serviços [223]. A alocação dos serviços em servidores de borda deve ser dinâmica, pois a alocação otimizada dos serviços depende da localização e pode mudar ao longo do tempo. A

alocação dinâmica de serviços refere-se ao processo de configuração adaptativa de serviços em servidores de borda. O processo inclui alocação de recursos, serviço de armazenamento em cache e decisões de *offloading* [224].

Uma taxonomia associada ao problema de alocação de serviços é apresentada em [219], onde são elencadas especificidades e critérios a serem considerados nas estratégias de implantação. Esta taxonomia está ilustrada na Figura 3.3.

Figura 3.3: Taxonomia com especificidades e critérios para Service Placement



Fonte: Adaptado de [219]

Um dos pontos de partida a serem considerados no desenvolvimento de estratégias e serviços de gerenciamento está associado à estratégia de coordenação a ser adotada, cuja abordagem está diretamente associada à definição do **Plano de Controle**, podendo apresentar modelos de coordenação centralizados ou distribuídos. Na abordagem de coordenação *centralizada* é necessário ter o conhecimento sobre demandas de aplicativos e recursos de infraestrutura de modo global para se tomar e disseminar decisões de implantação, tendo como uma de suas principais vantagens uma abordagem de solução potencialmente ótima de um ponto de vista global do sistema. No entanto, as soluções centralizadas são vulneráveis em relação à escalabilidade e à questão da complexidade computacional. Por outro lado, uma abordagem *distribuída* leva em consideração múltiplos nós de autoridade e orquestração para controlar o mapeamento de serviços, tendo como principais vantagens a solução de problemas de escalabilidade e de fornecimento de serviços que melhor se ajustem ao contexto local. Em contrapartida, nem sempre será possível obter uma alocação de serviços ótima a um nível global.

A **Característica de Alocação** está associada ao método de implantação dos algoritmos de alocação de serviços. A estratégia de alocação *offline* a tomada de decisão para alocação de serviços acontece em tempo de compilação. Esta abordagem é adotada quando se tem posse todas informações necessárias para implementação estratégia. Já na alocação *online*, as decisões de implantação são realizadas em tempo de execução do sistema. As decisões são baseadas nas características do processo e no estado atual do sistema.

Do ponto de vista da **Dinamicidade**, uma abordagem é dita dinâmica se a estratégia de alocação de serviços disponibilizada for capaz de implantar novos serviços e substituir ou liberar serviços já implantados para atender às restrições de QoS e otimizar um determinado objetivo. Para lidar com a natureza dinâmica de aplicações e/ou infraestrutura de Computação em Borda, é necessário definir estratégias reativas capazes de determinar quando a adaptação será necessária, fornecendo um mecanismo transparente de acesso, ainda garantindo uma qualidade de serviço (QoS) satisfatória.

Gerenciar a **Mobilidade** é um grande desafio em ambientes de Computação em Borda. Fornecer uma solução que suporte a mobilidade de usuários finais e/ou dispositivos de borda, ainda garantindo que os usuários sempre recebam os serviços associados e o desempenho desejado sem interrupções, é uma questão complexa a ser resolvida para este novo paradigma de computação distribuído. Neste sentido, na próxima subseção será tratada a relevância da mobilidade para estratégias de alocação de serviços.

3.2.3 Relevância da Mobilidade na Alocação de Serviços

A próxima geração de redes sem fio englobará várias técnicas de comunicação heterogêneas para dar sustentação a diferentes requisitos de qualidade e restrições de provedores de serviços, aplicativos e usuários. Atualmente, as redes de comunicação móveis de quinta geração (5G) integradas a soluções de Computação em Borda são capazes de oferecer diferentes tipos de serviços, tais como, o gerenciamento de tráfego, aplicativos de infotenimento, serviços de segurança, entre outros. No entanto, possibilitar a mobilidade contínua em várias redes de acesso é uma questão fundamental para disponibilizar serviços sem interrupções [225].

Neste sentido, pode-se destacar um dos principais desafios da alocação de serviços na Computação em Borda: a gestão da mobilidade do usuário [46]. Os usuários mudam suas localizações dinamicamente e a atual alocação de serviços em servidores de borda pode não

ser a melhor em termos de custos envolvidos. Desse modo, alguns dos componentes de uma determinada aplicação poderão precisar ser alocados em diferentes nós de borda, visando minimizar os custos de processamento e comunicação relacionados a sua execução. [226]

Para lidar com a mobilidade dos usuários, uma estratégia de alocação de serviços é proposta em [227] baseada nas seguintes decisões: "Quando-Migrar", com base no parâmetro de latência; e "Para onde migrar", com base na distância do usuário de nós de borda, considerando o nó processamento atual e os nós de borda circundantes. Nas referências [228–232], por exemplo, os autores tentam resolver o problema da mobilidade do usuário final fornecendo abordagens de alocação dinâmica de serviços.

A crescente complexidade dos padrões de mobilidade e da dinâmica nas solicitações de diferentes tipos de serviços tornou a alocação de serviços uma tarefa desafiadora. Uma solução típica de alocação estática não é eficaz, pois não considera a mobilidade do tráfego e a dinâmica do serviço [62]. A maioria dos esquemas existentes determinam as posições dos controladores com base em uma estimativa aproximada das cargas de tráfego dos *switches*. Essa abordagem não é adequada para sistemas de natureza altamente dinâmica, ocasionada principalmente pela necessidade de mobilidade. [233]

3.2.4 Discussão

Apesar dos recentes avanços de pesquisa relacionados ao problema de alocação serviços, muitos desafios ainda encontram-se em aberto. Podendo-se citar: qual problema pode ser resolvido com a abordagem apresentada, ou seja, encaixa em qual cenário é aplicável; quais informações importantes precisam ser consideradas, por exemplo, informações de infraestrutura, descrição da aplicação, requisitos de mapeamento, entre outros; e sob quais aspectos abordá-lo do ponto de vista das especifidades e critérios a serem considerados nas estratégias de implantação, tais como, dinamicidade, plano de controle, característica de alocação, suporte à mobilidade.

De acordo com o trabalho apresentado por Salaht et al. [219], distribuir a tomada de decisão para vários nós de borda em vez de depender do mapeamento em um único nó central ajuda a distribuir a carga e possivelmente aumentar a escalabilidade. No entanto, é possível observar que a abordagem de alocação de serviços de forma distribuída ainda é pouco explorada. Um dos motivos explanados está associado à complexidade de construção,

gestão e orquestração de algoritmos e serviços distribuídos, onde suas implantações muitas vezes não são triviais de serem realizadas em um ambiente real devido à complexidade da comunicação e sincronização entre processos. Além disso, um outro desafio mapeado é que a maioria das técnicas de alocação de serviços que suportam a mobilidade são reativas. Adotar estratégias de um ponto de vista proativo, que leve em consideração o padrão de mobilidade de usuários e dispositivos, pode ser mais adequado no contexto de Computação em Borda.

3.3 Considerações Finais

A partir da revisão bibliográfica apresentada neste capítulo, foi verificada a necessidade de mecanismos de orquestração inteligentes para a gestão e alocação eficiente de recursos e serviços em um ambiente de aplicação crítica utilizando redes móveis, em particular, no cenário de Ambulâncias Conectadas, considerando ainda a adoção de um paradigma distribuído em Computação em Borda em redes 5G como uma abordagem para atender aos estritos requisitos de comunicação impostos pela necessidade de transmissão de dados em cenários de emergência. Observou-se, ainda, que grande parte das soluções existentes para este cenário ainda carecem de uma análise criteriosa do impacto da mobilidade e de soluções que levem em consideração esse impacto. Adotar estratégias inteligentes de alocação de serviços considerando o padrão de mobilidade, pode ter sua relevância na promoção de soluções de saúde móveis (mHealth), ainda garantindo uma solução otimizada e de altíssima confiabilidade em sistemas dinâmicos, distribuídos e ainda com capacidade computacional limitada.

Capítulo 4

Ambulâncias Conectadas

Em situações de emergência, a eficácia da troca de informações entre os paramédicos que transportam pacientes e as equipes hospitalares é um aspecto crucial do atendimento ao paciente, ocorrendo geralmente assim que a ambulância chega ao hospital [17]. Especificamente, as operações de ambulância são uma parte essencial dos serviços de emergência e, ao longo das últimas décadas, tornaram-se cada vez mais avançadas, contando com um número significativo de equipamentos e dispositivos médicos de alta tecnologia [234].

Visando a melhoria da eficiência nos serviços pré-hospitalares, com base na adoção de tecnologias de informação e comunicação, a solução de Ambulância Conectada é bastante promissora. Com o desenvolvimento de soluções para esse caso de uso, será possível, por exemplo, coletar dados críticos dos pacientes e compartilhá-los com o hospital em tempo real, mesmo antes da chegada. Como resultado, médicos e enfermeiros de emergência estarão melhor preparados para receber o paciente, o que significa um processo de transferência mais suave e eficiente.

Nesse contexto, neste capítulo estão detalhados o cenário de ambulâncias conectadas e os equipamentos e dispositivos utilizados para possibilitar soluções inteligentes em emergências médicas.

4.1 Visão Geral e Estado da Arte

As ambulâncias conectadas surgem como uma inovação promissora na área da saúde, integrando-se às cidades inteligentes para otimizar o atendimento pré-hospitalar. Neste sen-

tido, nesta seção são abordadas pesquisas do estado da arte deste tema, explorando suas arquiteturas, aplicações, requisitos e tecnologias habilitadoras.

Park *et al.* [235] analisaram os requisitos para a implementação de serviços médicos de emergência inteligentes, destacando a importância da interoperabilidade, padronização de dados e a adoção de tecnologias como computação em nuvem e a internet das coisas médicas (IoT). No entanto, seu estudo não considera a crescente importância do streaming de vídeo e da realidade aumentada em ambientes médicos, que exigem alta largura de banda e baixa latência. Essa omissão sugere que sua dependência apenas da computação em nuvem pode não atender a essas necessidades tecnológicas avançadas, potencialmente limitando a eficácia de seu quadro proposto em certos cenários médicos.

Cisotto *et al.* [186] exploraram os requisitos e tecnologias necessários para fornecer serviços de saúde avançados através de sistemas celulares futuros. Eles enfatizaram a necessidade crítica de alta largura de banda, baixa latência e confiabilidade, especialmente para aplicações essenciais de telemedicina e cirurgia remota. Entre os cenários-chave identificados, a ambulância conectada apresenta um desafio significativo às redes de comunicação. Ela exige transmissões de alta taxa de dados e latência mínima, tempos de resposta excepcionalmente breves e confiabilidade robusta de conexão. Esses fatores são essenciais para manter o fluxo de dados ininterrupto e consistente, crucial quando a ambulância está em movimento.

Uma arquitetura de ambulância inteligente habilitada pela rede 5G foi apresentada por Zhai *et al.* [42], permitindo o envio dos dados vitais dos pacientes em tempo real para os dispositivos médicos do centro de operações, possibilitando diagnósticos precoces e intervenções cirúrgicas remotas. Embora o cenário de emergência delineado para testar sua ambulância inteligente habilitada para 5G — focando em um paciente em trânsito a uma velocidade constante de ambulância e transmitindo dados de imagens de CT para um hospital — forneça insights valiosos sobre a conclusão e precisão das tarefas do sistema, levanta certas preocupações. O cenário parece excessivamente otimista em termos de estabilidade de rede e capacidades de transmissão de dados, especialmente em condições geográficas e infraestruturais variadas. Cenários do mundo real frequentemente envolvem velocidades de rede flutuantes e cobertura, o que pode impactar significativamente a confiabilidade e eficácia da transmissão de dados em tempo real, crucial para serviços médicos de emergência.

Ahmed *et al.* [236] propuseram um protocolo de recomendação de recursos de rota

(R3) baseado em inteligência artificial embarcada para ambulâncias autônomas conectadas (ACA), considerando o status do tráfego em tempo-real e os recursos médicos disponíveis em cidades inteligentes. O protocolo ACA-R3 é utilizados para fornecer uma abordagem confiável e ecologicamente correta para computação, visando reduzir o uso interno de energia e aumentar a eficiência energética das ambulâncias autônomas conectadas. O principal objetivo desta pesquisa é reduzir o tempo das transferências de pacientes e agilizar o processo de compartilhamento de informações dos pacientes, especialmente em situações de emergência.

Chowdhury *et al.* [237] propuseram uma nova arquitetura de e-health baseada em comunicação por câmera óptica para 5G, possibilitando a transmissão de dados para médicos usando sinais de luz visível, o que pode ser útil em situações de desastre ou em áreas remotas com baixa cobertura celular. Por outro lado, Abdeen *et al.* [203] investigaram como os sistemas de saúde inteligentes podem melhorar o desempenho dos serviços de emergência de ambulâncias ao reduzir o tempo de resposta e otimizar o uso de recursos médicos. Além disso, Singh *et al.* [238] avaliaram a cobertura celular para um caso de uso de ambulância inteligente, identificando desafios e oportunidades para garantir a conectividade confiável necessária para suportar aplicações críticas de saúde.

4.2 Definição do Cenário

Em uma perspectiva futurista, onde os serviços essenciais de saúde são substancialmente melhorados, o atendimento de emergência é transformado por ambulâncias conectadas em movimento, suportadas por sistemas avançados de recomendações médicas em tempo-real, impulsionado pelas tecnologias de realidade aumentada (AR). Estas ambulâncias são equipadas com dispositivos inteligentes, sensores e soluções de conectividade de alta velocidade, facilitando a comunicação instantânea e de baixa latência com equipes médicas e funcionários dos hospitais. Essa comunicação instantânea é crucial para acelerar respostas a emergências e viabilizar o compartilhamento de informações dos pacientes, como sinais vitais, histórico médico e imagens diagnósticas. O conceito de "ambulância conectada"[23] serve como um exemplo principal das necessidades complexas de comunicação que serão integradas à evolução dos futuros serviços de e-health.

Utilizando técnicas de análise de dados de última geração e inteligência artificial, um sistema de recomendação médica em tempo real pode ser usado para processar o influxo de dados dessas ambulâncias conectadas. Com base nessa análise, o sistema é capaz de fornecer recomendações precisas e personalizadas para profissionais de saúde, ajudando-os a tomar decisões rápidas e adequadas em situações de emergência.

Nesse contexto, a Tecnologia de Realidade Aumentada é fundamental, pois permite a sobreposição de informações virtuais no ambiente real, por meio de dispositivos como óculos inteligentes. Assim, os profissionais de saúde podem acessar informações importantes em tempo real, como guias de procedimentos, instruções de primeiros socorros e imagens médicas, sem precisar desviar a atenção do paciente.

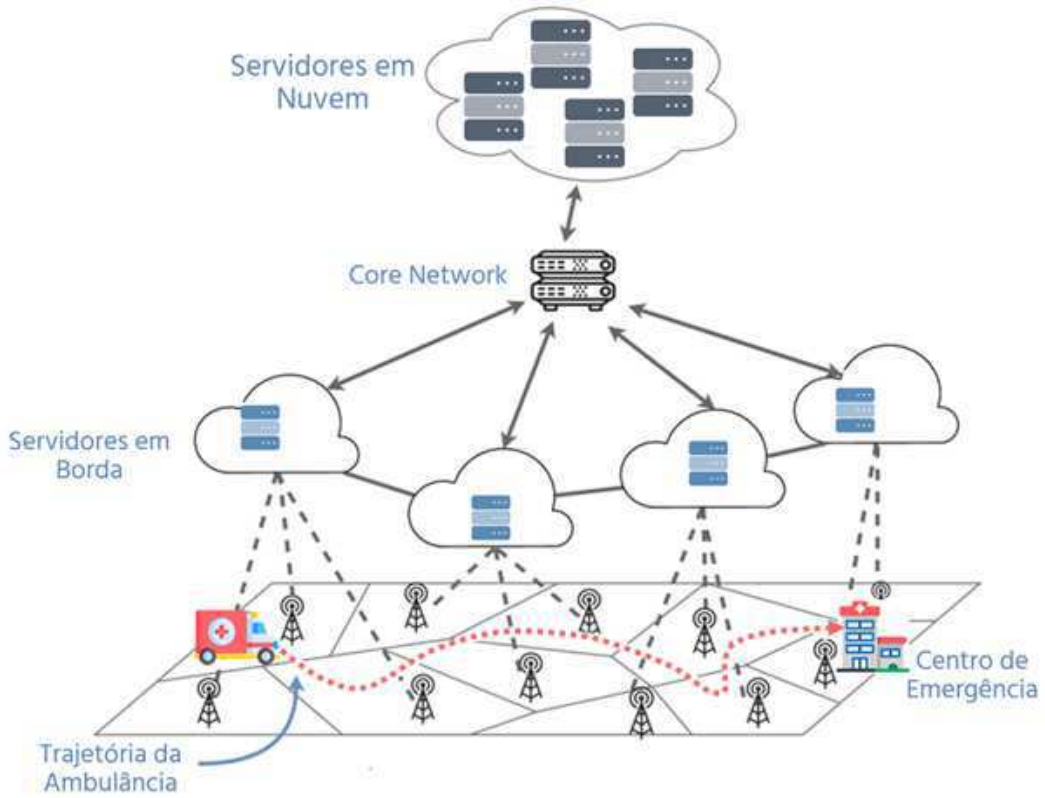
Na Figura 4.1 está apresentado o cenário de ambulância conectada a ser explorado neste trabalho. É possível observar a interconexão entre várias estações rádio-base e servidores de borda que, por sua vez, estão interligados entre si e, ainda, possibilitam a realização de *offloading* de serviços para servidores em nuvem.

Este cenário apresenta características distintas que exigem soluções personalizadas para elevar o padrão dos serviços de ambulância de emergência. Notavelmente, o ambiente da ambulância abrange tanto o cuidado médico quanto cenários de saúde urgentes. Nesse sentido, é imperativo empregar soluções tecnológicas adeptas ao manejo de aplicações críticas. Estas incluem sistemas avançados de monitoramento, equipamentos médicos especializados e conexão sem fio de muito baixa latência, onde a segurança e privacidade dos dados do paciente devem ser garantidas durante todo o processo.

Dada a natureza geo-distribuída das emergências, as ambulâncias devem ser móveis, prontas para lidar rapidamente com incidentes em locais diversos. As soluções de mobilidade abrangem sistemas de navegação em tempo real, roteamento otimizado e canais de comunicação eficazes ligando o pessoal médico, hospitais e instalações de tratamento.

A inovação é chave para avançar nos serviços médicos de emergência. As soluções de ponta podem incluir telemedicina para avaliação e triagem de pacientes à distância, inteligência artificial para apoiar escolhas diagnósticas e terapêuticas, bem como tecnologia médica avançada e dispositivos vestíveis para monitoramento contínuo dos sinais vitais dos pacientes. Dentro deste quadro, a implantação de sistemas de recomendação de procedimentos médicos em tempo real, potencializados por tecnologias imersivas como óculos inteligentes

Figura 4.1: Diagrama do cenário de Ambulâncias Conectadas.



Fonte: Adaptado de [46].

de AR, visa refinar a qualidade do serviço de emergência. Tais sistemas requerem capacidades robustas de processamento e transmissão de dados de alta velocidade para lidar com a análise instantânea e a transferência de dados audiovisuais.

4.3 Equipamentos e Dispositivos para Soluções Inteligentes em Emergências Médicas

O conceito inovador de ambulâncias conectadas apresenta uma oportunidade de explorar um conjunto abrangente de equipamentos e dispositivos que são cruciais para o avanço dessas soluções inteligentes de saúde. A integração dessas tecnologias é possível graças ao uso de algoritmos avançados de inteligência artificial e à fusão de dados provenientes de diversos sensores, frequentemente referida como fusão de sensores (do inglês, *sensor fusion*. Essa integração é fundamental na criação de sistemas inteligentes capazes de reagir autonomamente

a ambientes dinâmicos e às necessidades dos pacientes [239].

No cenário em questão, é possível listar uma variedade de tipos de dispositivos, finalidades e tipos de dados a serem transmitidos por meio de uma rede de comunicação sem fio, cuja definições implicarão na proposta de solução arquitetural para atender aos diferentes tipos as demandas. Dentre estes dispositivos e equipamentos podemos citar:

- Sensores de sinais vitais capazes de medir o nível de saturação de oxigênio, temperatura corporal, frequência respiratória, pressão arterial, entre outros;
- Eletrocardiograma (ECG) e/ou Eletroencefalograma (EEG);
- Equipamentos para diagnósticos por imagem, como os baseados em tecnologias de Ultrassom e de Tomografia Computadorizada;
- Equipamentos multimídia, incluindo microfones, câmeras de alta definição e óculos inteligentes, para comunicação com unidades de saúde remotas.

Na Tabela 4.1 está apresentado um mapeamento desses equipamentos e dispositivos, destacando não apenas suas funções e usos no contexto de Ambulâncias Conectadas, mas também detalhando sua implantação específica e aplicação prática em cada cenário respectivo. Na Tabela 4.2 estão listados o uso desses equipamentos e dispositivos no cenário de ambulâncias conectadas, ilustrando como eles contribuem para estratégias de resposta a emergências mais eficientes e eficazes.

A incorporação desses dispositivos não apenas transforma a abordagem médica em ambientes de emergência, mas também potencializa a interação entre a equipe de resgate e os especialistas em centros médicos avançados. Com o uso de tecnologia de transmissão de dados em tempo-real, é possível que médicos localizados em hospitais acompanhem o estado do paciente durante o transporte, oferecendo orientações precisas e decisões de tratamento baseadas em informações atualizadas e detalhadas. Essa colaboração remota não só melhora a qualidade do atendimento prestado, como também maximiza as chances de recuperação do paciente ao permitir intervenções médicas imediatas e fundamentadas, ainda que à distância. Isso representa um avanço significativo na gestão de situações críticas, onde cada segundo é crucial para a evolução do estado de saúde do paciente.

Equipamentos/ Dispositivos	Caso de uso	Entradas	Análise/Processamento	Saídas
Óculos de realidade mista	Óculos de realidade mista para auxiliar paramédicos em cenários de emergência em ambulâncias	· Stream de vídeo 4K (3840x2160 pixels) 60 fps 12 bits por pixel (por exemplo, YUV4:1:1)	· Análise automática de cenários com base em visão computacional · Reconhecimento facial do paciente baseado em visão computacional · Processamento de streaming de vídeo · Unificação de frames para exibição em óculos de realidade mista	· Visualização de hologramas 3D com procedimentos de orientação para paramédicos · Adição de indicadores médicos · Exibição de dados do paciente
Headset integrado com óculos de realidade mista	Fone de ouvido usado para comunicação de voz entre paramédicos e equipe de assistência remota	· Fluxo de áudio de alta qualidade do microfone · Entradas de comando de voz para óculos de realidade mista	· Extração de conteúdo de áudio para análise de contexto	· Sinal de áudio de alta qualidade transmitido para fones de ouvido (da equipe remota e do sistema de recomendação) · Sinais de controle para óculos de realidade mista
Sensores sem fio para medir sinais vitais (ECG, oxigenação, pressão arterial, temperatura corporal)	Aquisição de sinais vitais do paciente	· Fluxo de dados de monitoramento de sinais vitais	· Cálculos de scores médicos · Estimativa de condições clínicas do paciente	· Análise do estado de saúde do paciente
Sonda de ultrassom	Visualização detalhada das estruturas internas do paciente, bem como de seus órgãos e tecidos. Aplicado para identificar lesões internas	· Transmissão de vídeo de cronometragem sem ultrassom sem compressão 2048x2048 pixels 32 bits por pixel 20 fps	· Diagnóstico por imagem baseado em sistemas de classificação	· Identificação de lesões · Identificação de fraturas internas · Identificação de hemorragias

Tabela 4.1: Mapeamento de equipamentos e dispositivos para melhorias nos serviços de saúde.

Equipamentos/Dispositivos	Contribuições para melhorias nos serviços de saúde
Óculos de realidade mista	O uso de óculos de realidade mista em ambulâncias melhora significativamente o atendimento emergencial, oferecendo diagnósticos mais rápidos, visualização avançada de dados e suporte decisivo para navegação e planejamento de procedimentos médicos.
Headset integrado com óculos de realidade mista	O headset integrado aos óculos de realidade mista melhora os serviços de saúde ao permitir a comunicação direta e eficiente entre paramédicos e equipes de suporte remoto, utilizando extração de conteúdo de áudio e comandos de voz para melhorar a usabilidade do paramédico e análise contextual rápida e recomendações precisas em situações de emergência.
Sensores sem fio para medir sinais vitais (ECG, oxigenação, pressão arterial, temperatura corporal)	A utilização de sensores sem fio para medição de sinais vitais permite o monitoramento contínuo e preciso da saúde do paciente, facilitando cálculos de escores médicos e estimativas de condições clínicas para melhorar a tomada de decisões na área da saúde.
Sonda de ultrassom	O uso de uma sonda de ultrassom para visualizar detalhadamente as estruturas internas do paciente melhora o diagnóstico de lesões, fraturas internas e sangramentos, melhorando significativamente a precisão e a eficácia dos tratamentos médicos.

Tabela 4.2: Contribuições para melhorias nos serviços de saúde para o caso de uso de ambulâncias conectadas.

4.4 Serviços de Emergência auxiliados por Tecnologias Emergentes

A adoção de tecnologias emergentes, como Realidade Aumentada (AR) e Inteligência Artificial (IA), tem o potencial de transformar fundamentalmente os serviços de ambulâncias conectadas. A Realidade Aumentada possibilita que os socorristas visualizem informações críticas e dados de saúde em tempo-real, superpondo-os no campo de visão, o que agiliza as decisões no local da emergência.

Paralelamente, a Inteligência Artificial pode ser utilizada em diversas frentes neste cenário, que englobam desde questões de tráfego urbano, podendo realizar otimização das rotas, além de aplicações diretamente relacionadas ao estado de saúde dos pacientes, como a análise rápida de dados vitais e na assistência à decisão clínica, garantindo uma resposta mais eficaz e precisa.

Essas inovações não só melhoram a qualidade do atendimento em situações críticas, como também potencializam a coordenação e a eficiência operacional das equipes de emergência. A seguir, estão detalhadas a implementação e o impacto dessas tecnologias avançadas nas práticas modernas de serviços médicos de emergência.

4.4.1 Óculos Inteligentes: Serviços Médicos de Emergência sob uma nova Perspectiva

A integração de óculos inteligentes nos serviços médicos de emergência representa um desenvolvimento revolucionário na tecnologia da saúde. Essa integração utiliza principalmente as capacidades da Realidade Mista (MR), Realidade Aumentada (AR) e Realidade Virtual (VR), oferecendo abordagens transformadoras para aplicações críticas de saúde, especialmente em cenários com ambulâncias. A importância dessa integração é explorada através de diversos trabalhos de pesquisa, cada um contribuindo para uma compreensão abrangente deste avanço tecnológico.

No âmbito da MR, a taxonomia para assistência e treinamento remoto, como explorada em [240], destaca as diversas aplicações e o potencial da MR em serviços de emergência.

Esta taxonomia é crucial para entender como a MR pode ser usada tanto para aplicações educacionais quanto práticas em campo durante emergências médicas.

O papel dos displays montados na cabeça (do inglês, *head-mounted display* - HMDs), como os óculos inteligentes, é crucial nas aplicações de MR e AR. Esses dispositivos passaram por avanços significativos, tornando-se mais amigáveis ao usuário e ricos em recursos. Guo *et al.* [241] realizaram uma extensa pesquisa sobre o desenvolvimento de HMDs, apresentando vários modelos e tipos de HMDs, cada um adaptado para cenários específicos, incluindo serviços de saúde de emergência.

As aplicações de AR e VR, particularmente na saúde, exigem requisitos de comunicação rigorosos, como detalhado em [242]. Essas aplicações requerem alta largura de banda, latência ultra baixa e conexões confiáveis para garantir assistência em tempo real e sem interrupções durante emergências médicas.

No trabalho apresentado por Ishikawa *et al.* [243], uma investigação preliminar sobre a viabilidade do uso de óculos inteligentes no ambiente pré-hospitalar é conduzida, onde informações do paciente são compartilhadas por vídeo e voz sem afetar o tempo de resposta, potencialmente aprimorando o cuidado médico de emergência. No entanto, o estudo sublinha a necessidade de enfrentar desafios relacionados à instabilidade de comunicação para garantir a eficácia e a confiabilidade dessa tecnologia inovadora.

A interseção da AR com a saúde e o Metaverso é outro desenvolvimento intrigante. Bansal *et al.* [244] exploram essa área, investigando como o Metaverso pode facilitar ambientes de saúde imersivos para diagnósticos e tratamentos remotos.

No entanto, a integração da AR com as tecnologias web apresenta vários desafios. Qiao *et al.* [245] discutem esses desafios, incluindo o equilíbrio entre as capacidades computacionais limitadas dos dispositivos e as altas demandas computacionais das aplicações de AR, a latência da rede versus a necessidade de interação em tempo real, e o conflito entre as limitações de bateria e o alto consumo de energia das aplicações de AR.

Uma tecnologia chave que possibilita a AR na Web é o WebRTC, essencial para comunicação em tempo-real em emergências médicas. Soluções baseadas computação de borda são promissoras nesse sentido, pois podem processar dados mais próximos ao usuário, aumentando a velocidade e eficiência das aplicações de AR em cenários de emergência. Essas abordagens são destacadas em estudos como [246] e [245].

Conforme destacado em [247], a computação de borda desempenha o importante papel de processar dados em tempo-real, que viabiliza maior engajamento e iteratividade do usuário com soluções imersivas. Os autores ainda destacam que a computação em borda permitirá que os requisitos avançados de rede para aplicações do Metaverso sejam atendidos, que incluem a alta largura de banda, qualidade de serviço (QoS) e ultra baixa latência.

Considerando a necessidade de mobilidade das unidades móveis de atendimento, destaca-se a importância de disponibilizar serviços e aplicações nos servidores mais próximos das ambulâncias, implicando em estratégias que permitam migração entre servidores de borda. As tecnologias atuais nessa área ainda estão em desenvolvimento, conforme detalhado em [248], que demandam mais pesquisas para atender à QoS necessária em vários casos de uso envolvendo migração de serviços em redes de borda.

Em geral, a integração de óculos inteligentes e tecnologias AR/VR/MR nos serviços médicos de emergência cria um novo marco na área da saúde. Embora existam desafios na integração tecnológica e nas capacidades de rede atuais, a pesquisa e o desenvolvimento contínuos de novas soluções estão gradativamente superando esses obstáculos. Essas integrações resultam em uma abordagem mais eficaz no atendimento médico de emergência, sendo assistida por tecnologias avançadas para aprimorar a qualidade dos serviços médicos e os resultados dos pacientes.

4.4.2 Potencializando Serviços de Emergência em Ambulâncias com Inteligência Artificial

A Inteligência Artificial (IA) tem alcançado avanços notáveis nos últimos anos, revolucionando diversos setores, especialmente a saúde [249]. A capacidade da IA de processar e analisar grandes quantidades de dados rapidamente levou a avanços significativos no diagnóstico, planejamento de tratamento e gestão de cuidados com pacientes. Na saúde, as aplicações de IA têm sido fundamentais para aumentar a precisão diagnóstica, prever resultados dos pacientes, personalizar planos de tratamento e melhorar a eficiência geral na entrega de cuidados de saúde [250].

No trabalho apresentado por Uddin *et al.* [251], por exemplo, técnicas de aprendizado de máquina são utilizadas para otimizar a alocação de dados em repositórios de saúde. Este es-

tudo destaca o papel do aprendizado de máquina em acelerar o acesso a informações críticas de saúde, melhorando a gestão de recursos de dados e prevendo necessidades de armazenamento. A implementação dessas técnicas leva a respostas mais eficientes e oportunas às demandas de saúde, permitindo cuidados médicos mais eficazes e informados. Esse avanço é particularmente relevante para ambulâncias conectadas, onde o acesso rápido aos dados do paciente é crucial. Uma gestão eficiente de dados permite decisões médicas informadas em tempo-real em situações de emergência, impactando significativamente a qualidade do cuidado ao paciente durante o transporte.

Outra aplicação de IA na área da medicina é apresentada em Yue *et al.* [252], onde algoritmos de aprendizado de máquina analisam registros eletrônicos de saúde (EHR) para prever o estado de saúde dos pacientes. O estudo destaca a precisão melhorada na previsão de resultados de pacientes, detecção de doenças e planejamento de tratamento personalizado, utilizando técnicas como árvores de decisão, *random forests* e redes neurais. No cenário de ambulâncias conectadas, esses insights fornecidos por IA podem ser determinantes para a vida dos pacientes. Eles permitem que os paramédicos tomem decisões informadas e em tempo-real sobre o cuidado ao paciente durante o traslado, potencialmente melhorando a eficácia da resposta a emergências e preparando a equipe do hospital para a chegada do paciente com estratégias de tratamento personalizadas.

O uso de Modelos de Linguagem de Grande Escala (LLMs), como o ChatGPT [1], na medicina é examinado no trabalho de Sallam [253], que fornece uma revisão sistemática da aplicação dessa tecnologia no setor de saúde. O estudo enfatiza os benefícios da utilização do ChatGPT no ensino, pesquisa e prática clínica, incluindo aprendizado interativo, análise de dados eficiente e apoio em diagnósticos e comunicação com pacientes. A capacidade do modelo de analisar e interpretar dados médicos complexos em movimento é capaz de contribuir significativamente na tomada de decisões, melhorando a qualidade do cuidado prestado durante o transporte de pacientes em estado crítico. Ao mesmo tempo, aborda preocupações sobre precisão e ética, sublinhando a importância da supervisão regulatória na aplicação do ChatGPT em ambientes de saúde.

Tsoi *et al.* [254] discutem o uso da IA no cuidado da hipertensão, destacando seu papel em melhorar a precisão diagnóstica, personalizar o tratamento e melhorar o monitoramento

[1] https://chat.openai.com/

do paciente. Ele foca em algoritmos de aprendizado de máquina e análises preditivas para analisar dados de pacientes, auxiliando na detecção precoce e no gerenciamento eficaz da hipertensão, e enfatiza a capacidade da IA de lidar com grandes conjuntos de dados para decisões de saúde informadas. A integração da IA em ambulâncias poderia levar a respostas a emergências mais eficientes e personalizadas, especialmente para pacientes com emergências relacionadas à hipertensão.

O uso da IA na análise de ECG foi explorado por Attia *et al.* [255], destacando a melhoria na detecção de anormalidades cardíacas e capacidades preditivas para eventos cardíacos futuros. Eles enfatizam o uso de aprendizado de máquina, particularmente aprendizado profundo, para uma análise de dados de ECG mais eficaz, levando a diagnósticos mais antecipados e precisos de condições cardíacas. No contexto de ambulâncias conectadas, essa aplicação poderia revolucionar o cuidado cardíaco de emergência. Ao adotar-se modelos aprendizado de máquina para análise de ECG, os paramédicos podem diagnosticar condições cardíacas com mais precisão e rapidez durante o atendimento em movimento.

4.5 Requisitos

No contexto das Ambulâncias Conectadas, garantir que os paramédicos possam prestar cuidados de emergência eficazes com acesso a informações atualizadas e comunicação confiável durante o trânsito envolve enfrentar vários desafios e requisitos. Estabelecer uma infraestrutura de comunicação robusta é crucial para serviços de emergência ininterruptos, especialmente em saúde móvel. Isso envolve garantir cobertura de rede consistente e abordar questões como interferência, latência e perda de sinal [256].

A implementação de Comunicação Ultra Confiável e de Baixa Latência é chave para a transmissão suave de dados de áudio, vídeo e sinais vitais. Em ambientes com largura de banda limitada, otimizar a compressão de vídeo e adaptar-se às condições de rede são essenciais para manter a qualidade em aplicações de saúde. Além disso, proteger a integridade e a privacidade dos dados sensíveis do paciente por meio de protocolos de transmissão seguros é um aspecto fundamental dessa infraestrutura, garantindo comunicação eficiente, ininterrupta e segura nos serviços médicos de emergência [257].

Abordar a capacidade computacional em dispositivos de saúde móveis é crucial, especialmente para aplicações avançadas que exigem processamento intensivo de dados e análises em tempo real. Dispositivos móveis, frequentemente usados em ambulâncias conectadas, enfrentam limitações de poder de processamento em comparação com sistemas de computação tradicionais. Isso exige a otimização de algoritmos e arquiteturas de software para garantir desempenho eficiente. A alta capacidade de processamento é particularmente importante para aplicações multimídia em tempo real e sistemas de RA, que exigem recursos computacionais significativos para tarefas como compressão de dados e unificação de quadros. A otimização eficiente garante que esses dispositivos possam suportar aplicações de saúde complexas efetivamente em situações críticas.

O gerenciamento eficaz do consumo de energia é essencial em tecnologias de saúde móveis, especialmente para tarefas intensivas em energia, como transmissão de vídeo e processamento de dados em tempo real. Otimizar estratégias de gerenciamento de energia é chave para prolongar a vida útil da bateria e garantir a funcionalidade ininterrupta dos dispositivos móveis. Isso envolve a simplificação de sistemas para reduzir o uso de energia, particularmente em dispositivos como óculos inteligentes de AR, e garantir a utilização eficiente de recursos para manter a longevidade e a confiabilidade dos equipamentos de saúde móveis em configurações críticas.

A mobilidade das ambulâncias conectadas requer transições de rede contínuas e conectividade robusta para manter a qualidade do serviço em vários locais. Isso é essencial para garantir a integridade contínua dos dados e uma experiência de usuário contínua para os paramédicos, que dependem de soluções móveis para acesso a informações em tempo real e comunicação em situações de emergência. Portanto, é crucial ter sistemas portáteis e fáceis de usar e veículos bem equipados com tecnologia avançada para transmissão de dados eficiente, garantindo capacidades de resposta a emergências eficazes e flexíveis [258].

Para fornecer uma visão geral, a Tabela 4.3 detalha esses desafios e requisitos específicos para Ambulâncias Conectadas. Esta tabela é um recurso crucial, delineando áreas-chave que precisam de atenção, como infraestrutura tecnológica, segurança de dados e capacidades de comunicação em tempo real, para melhorar a eficiência e eficácia dos serviços de ambulância em situações críticas.

Desafios	Descrição	Requisitos	Justificativa
Infraestrutura de comunicação: Construindo uma rede robusta para contato ininterrupto durante emergências	Garantir uma infraestrutura de comunicações robusta e confiável é fundamental para fornecer conectividade adequada em ambientes móveis. Isto inclui garantir uma cobertura de rede adequada, especialmente em áreas remotas, bem como lidar com potenciais problemas de interferência, latência e perda de sinal.	Comunicação ultra confiável e de baixa latência	Entrega contínua e rápida de pacotes de dados de áudio, vídeo, sonda de ultrassom e sensores de monitoramento de sinais vitais. Reduzir a latência é fundamental para garantir uma experiência de visualização suave e sem interrupções.
		Alta largura de banda	Nas redes móveis, onde a largura de banda pode ser limitada, é necessário otimizar a compressão de vídeo e utilizar técnicas de adaptação para ajustar a qualidade do vídeo com base nas condições da rede. Para aplicações críticas de saúde, a garantia de qualidade torna-se essencial.
		Segurança da informação	Como estamos lidando com dados sensíveis de pacientes, garantir a integridade e a privacidade dos dados transmitidos torna-se uma tarefa fundamental.
Capacidade computacional: Garantir poder de processamento suficiente para aplicações médicas avançadas.	Os dispositivos móveis têm recursos de computação limitados em comparação com os sistemas de computação tradicionais. Isto pode representar um desafio para a execução de aplicações e serviços mais complexos, especialmente aqueles que requerem processamento intensivo de dados ou análises em tempo-real. É necessário otimizar algoritmos e arquiteturas de software para garantir um desempenho eficiente em dispositivos móveis.	Alta capacidade de processamento	Aplicações multimídia em tempo-real exigem altas taxas de processamento, devido à necessidade de executar algoritmos de compressão e descompressão de dados de áudio e vídeo. Além disso, aplicações que envolvem sistemas de realidade mista potencializam ainda mais essa necessidade de processamento, pois podem exigir tarefas de unificação de frames, por exemplo.
Consumo de energia: Gerenciando a energia de forma eficiente para suportar todas as tecnologias integradas.	O consumo de energia é um desafio significativo, especialmente quando se trata de aplicações que consomem muitos recursos, como streaming de vídeo ou processamento de dados em tempo-real. O desenvolvimento de estratégias de gestão de energia e otimização do consumo é essencial para prolongar a vida útil das baterias e garantir a funcionalidade contínua dos dispositivos móveis.	Autonomia para Dispositivos Móveis	A transmissão de vídeo em tempo-real e a utilização de algoritmos baseados em IA pode consumir muita energia, o que pode afetar negativamente a vida útil da bateria em dispositivos móveis, incluindo óculos de realidade mista. É importante otimizar o sistema para reduzir o consumo de energia, minimizando processamentos desnecessários e fazendo uso eficiente dos recursos do dispositivo.
Mobilidade: Manter uma qualidade de serviço consistente em diversas localizações geográficas.	A natureza móvel do cenário das Ambulâncias Conectadas apresenta desafios únicos, como a necessidade de lidar com a transição entre diferentes redes e pontos de acesso. Isto requer a capacidade de manter a conectividade e a integridade dos dados ao mudar de local ou alternar entre redes, garantindo uma ótima experiência de usuário.	Flexibilidade para paramédicos	Os paramédicos precisam de soluções móveis flexíveis que lhes permitam acessar informações críticas e se comunicar em tempo-real enquanto estão em movimento. Isto requer sistemas e dispositivos portáteis que sejam fáceis de transportar e operar durante situações de emergência.
		Necessidade de mobilidade	Os serviços de emergência envolvem a necessidade de deslocamento rápido para diferentes locais de atendimento. Isto implica na necessidade de veículos e equipamentos móveis adequados, dotados de tecnologias e sistemas de comunicação que permitam a transmissão de dados e informações em tempo-real.

Tabela 4.3: Desafios e Requisitos mapeados para o cenário de Ambulâncias Conectadas.

O cenário de Ambulâncias Conectadas apresenta diferentes requisitos de comunicação para diferentes equipamentos e dispositivos. Para a transmissão de sinais vitais, por exemplo, a latência deve ser inferior a 50 ms e taxa de transmissão menor que 10 kbps, com ressalvas para o caso de *streaming* de sinais de ECG e EEG, onde as taxas são de 72 kbps e 86,4 kbps, respectivamente. Para transmissão de *streaming* de vídeo de Ultrassom e Tomografia Computadorizada, a latência é inferior a 20 ms e as taxas de transmissão são, respectivamente, 160 Mbps e 670 Mbps. Por fim, para *streaming* multimídia de áudio, câmeras de alta definição e óculos inteligentes, as latências são inferiores a 100 ms, 50 ms e 20 ms, respectivamente. Já as taxas de transmissão são de 200 kbps para áudio e maior que 30 Mbps para vídeo e óculos inteligentes. Estes requisitos de comunicação apresentados para os dispositivos e equipamentos mencionados para este cenário foram extraídos dos trabalhos [1], [192] e [191], e estão sumarizados na Tabela 4.4.

Para muitos sistemas de Internet das Coisas baseados em ambientes de Computação em Nuvem, como controle de tráfego inteligente, casas inteligente, assistência cognitiva vestível, entre outros, a alta resiliência e baixa latência são dois requisitos principais [259]. No entanto, ao utilizar serviços baseados em ambientes em nuvem para análise de dados de IoT, controle remoto, monitoramento de sistemas, transmissão de informações em tempo-real, determinada latência de comunicação e processamento pode ser inaceitavelmente alta, como é o caso do cenário de Ambulâncias Conectadas.

Além desses desafios técnicos, a implementação eficaz de ambulâncias conectadas também requer consideração cuidadosa das questões de segurança e privacidade. A transmissão de dados sensíveis de saúde através de redes, especialmente em ambientes externos e muitas vezes inseguros, exige protocolos de segurança robustos para proteger contra acesso não autorizado e ataques cibernéticos. A conformidade com regulamentações globais de privacidade de dados, como o GDPR na Europa e a HIPAA nos Estados Unidos, é essencial para garantir que as informações dos pacientes sejam manipuladas com o mais alto nível de segurança e confidencialidade. A confiança do paciente e a integridade do sistema de saúde dependem da segurança dessas comunicações críticas, tornando as tecnologias de criptografia e autenticação componentes indispensáveis no desenvolvimento de soluções de ambulância conectadas [260].

Tabela 4.4: Requisitos de comunicação para o cenário de Ambulâncias Conectadas.

Fonte de Dados	**Características de Transmissão e Tipos de Dados**	**Disponibilidade do serviço de comunicação: valor alvo em%**	**Confiabilidade do serviço de comunicação: Tempo médio entre falhas**	**Máxima Latência ponto-a-ponto: Máxima**	**Taxa de Transmissão**	**Direção**	**Referências**
Temperatura Corporal	Streaming de dados com compressão sem perdas	99,999%	>>1 mês (<1 ano)	<50 ms	<10 kbps	Upstream	[1, 191, 192]
Frequência Cardíaca	Streaming de dados com compressão sem perdas	99,999%	>>1 mês (<1 ano)	<50 ms	<10 kbps	Upstream	[1, 191, 192]
Pressão Sanguínea	Streaming de dados com compressão sem perdas	99,999%	>>1 mês (<1 ano)	<50 ms	<10 kbps	Upstream	[1, 191, 192]
Nível de Saturação de Oxigênio	Streaming de dados com compressão sem perdas	99,999%	>>1 mês (<1 ano)	<50 ms	<10 kbps	Upstream	[1, 191, 192]
Frequência Respiratória	Streaming de dados com compressão sem perdas	99,999%	>>1 mês (<1 ano)	<50 ms	<10 kbps	Upstream	[1, 191, 192]
Eletrocardiograma	Streaming de dados com compressão sem perdas	99,999%	>>1 mês (<1 ano)	<50 ms	72 kbps	Upstream	[1, 191, 192]
Eletroencefalograma	Streaming de dados com compressão sem perdas	99,999%	>>1 mês (<1 ano)	<50 ms	86,4 kps	Upstream	[1, 191, 192]
Ultrassonografia	Streaming em tempo-real de vídeo descompactado da sonda de ultrassom 512x512 pixels 32 bits 20 fps	99,999%	>>1 mês (<1 ano)	<20ms	160 Mbps	Upstream	[1, 192]
Tomografia Computadorizada	Streaming em tempo-real de vídeo descompactado de CT 2048x2048 pixels 16 bits 10 fps	99,999%	>>1 mês (<1 ano)	<20ms	670 Mbps	Upstream	[1, 192]
Audio	Streaming de Audio de alta qualidade	99,99%	>1 mês	<100 ms	200 kbps	Upstream/ Downstream	[1]
Câmeras 4K	Streaming em tempo-real de vídeo 4K compactado (3840x2160 pixels) 60 fps 12 bits por pixel codificados por cores (ex. YUV 4:1:1)	99,99%	>1 mês	<50 ms	>30 Mbps	Upstream/ Downstream	[1]
Óculos Inteligentes	Streaming em tempo-real de vídeo 4K compactado (3840x2160 pixels) 60 fps 12 bits por pixel codificados por cores (ex. YUV 4:1:1)	99,99%	>1 mês	<20ms	>30 Mbps	Upstream/ Downstream	[1]

4.6 Melhorias e benefícios para os Serviços Pré-hospitalares

A integração de soluções de ambulâncias conectadas traz uma riqueza de benefícios que elevam profundamente a entrega dos serviços médicos de emergência. Isso inclui o aprimoramento da qualidade do atendimento de emergência, a redução dos tempos de resposta e o fornecimento de orientação e suporte em tempo real. Também garante acesso instantâneo aos registros médicos, otimiza o uso de recursos médicos e facilita o treinamento contínuo e o desenvolvimento profissional. Essas melhorias coletivas convergem para criar um ambiente de cuidados de emergência mais responsivo, eficiente e centrado no paciente.

Com o advento das ambulâncias conectadas, os profissionais de saúde ganham acesso instantâneo a dados essenciais do paciente, como sinais vitais, histórico médico e imagens diagnósticas, enquanto estão em movimento. Esse fluxo imediato de informações simplifica a tomada de decisões clínicas, permitindo a entrega de cuidados personalizados desde o início, melhorando significativamente o padrão dos serviços médicos de emergência. Barrett [261] destaca como a comunicação eletrônica das informações do paciente é percebida como útil pelo pessoal da ambulância, melhorando a colaboração com outros profissionais de saúde.

Permitindo que as equipes médicas iniciem a avaliação do paciente e comecem o tratamento durante o trânsito, as ambulâncias conectadas desempenham um papel crucial em reduzir o tempo crítico para intervenção, o que pode melhorar significativamente as taxas de sobrevivência dos pacientes e reforçar as perspectivas de recuperação para aqueles enfrentando condições ameaçadoras à vida. Uma estratégia de redução dos tempos de resposta de ambulâncias foi proposta por Ong et al. [262], onde uma análise geoespacial-temporal do posicionamento de ambulâncias foi proposta, enfatizando a relevância do tema no contexto dos serviços de saúde de emergência.

No estudo de Sonkin *et al.* [263], um exame baseado em simulação da comunicação de vídeo em tempo real entre paramédicos de ambulância e a cena de emergência revelou benefícios significativos. O estudo descobriu que esse modo de comunicação melhora a capacidade dos paramédicos de diagnosticar condições com precisão, fornecer instruções de tratamento oportunas e preparar medicamentos ou equipamentos necessários com antece-

dência. Além disso, as ambulâncias conectadas servem como um conduto para assistência visual e dados cruciais para profissionais de saúde que realizam procedimentos médicos, aumentando a precisão e eficiência dessas intervenções. Esse suporte avançado é chave para reduzir erros médicos e diminuir a probabilidade de complicações.

A tecnologia por trás das ambulâncias conectadas oferece acesso rápido e atualizado aos registros dos pacientes, incluindo alergias, medicamentos em uso e mais, permitindo que os tratamentos sejam finamente ajustados às necessidades individuais e evitando interações médicas potencialmente prejudiciais ou contraindicações. Nesse aspecto, Ben-Assuli *et al.* [264] enfatizam a importância de ter acesso ao histórico médico do paciente, incluindo informações sobre medicamentos, diagnósticos, procedimentos recentes e exames laboratoriais, para tomar decisões de admissão mais precisas em salas de emergência. Além disso, vale destacar a importância de sistemas e arquiteturas interoperáveis para realizar essa troca de dados, como apresentado em de Alencar *et al.* [265].

Ambulâncias de alta tecnologia modernas funcionam como plataformas de treinamento móveis, oferecendo ao pessoal médico a oportunidade de desenvolver habilidades complexas por meio de simulações em um ambiente realista e interativo. Este método de treinamento inovador aprimora a proficiência dos trabalhadores de saúde e expande seu conhecimento em medicina de emergência. Abelsson *et al.*[266] enfatizam que o aprendizado baseado em simulação, apresentando cenários realistas e de alta pressão, acelera a aquisição de habilidades sem a necessidade de extensa experiência de campo. Além disso, sistemas conectados em tempo real em ambulâncias enriquecem esse treinamento, permitindo exposição imersiva a cenários práticos, onde estudantes e profissionais podem observar e participar de uma gama completa de respostas a emergências.

No estudo apresentado por Gunnarsson *et al.* [267], a importância da experiência do profissional de emergência nos processos de tomada de decisão é destacada. O estudo sugere que uma compreensão mais profunda de como esses profissionais tomam decisões complexas em situações de emergência é necessária. Ele sublinha a importância da experiência para navegar eficazmente nesses ambientes de alta pressão e recomenda mais pesquisas para desvendar as complexidades da tomada de decisão no contexto do cuidado de emergência. Ambulâncias conectadas também facilitam a tomada de decisões sobre quando envolver especialistas ou determinar o melhor local para o cuidado do paciente, garantindo que os recursos de saúde

sejam alocados de maneira sábia e eficaz, o que agiliza o transporte do paciente e prepara os hospitais para emergências iminentes.

O advento das soluções de ambulâncias conectadas representa uma mudança transformadora nos serviços médicos de emergência. Esse salto tecnológico traz melhorias substanciais em múltiplos aspectos do cuidado ao paciente e operações clínicas. Ao integrar sistemas avançados de comunicação e gestão de dados, essas ambulâncias aprimoram a velocidade e precisão da resposta médica, levando a melhores resultados para os pacientes. Elas também melhoram a eficiência operacional, permitindo um gerenciamento de recursos mais eficaz e coordenação durante emergências. Em geral, a implementação de ambulâncias conectadas significa um marco importante no avanço da qualidade, eficiência e sucesso dos serviços clínicos no cuidado médico de emergência.

4.7 Limitações Tecnológicas

É fundamental reconhecer que, apesar dos avanços significativos em tecnologias de comunicação e computação, alguns desafios que afetam diretamente a implementação eficaz de soluções baseadas em inteligência artificial para serviços de emergência médica. Essas limitações tecnológicas não apenas influenciam a capacidade dos dispositivos em campo, como óculos inteligentes e outros dispositivos móveis, mas também impactam a infraestrutura de rede necessária para suportar a transmissão de dados em tempo-real e a execução de tarefas computacionalmente intensas em plataformas de computação em nuvem. A seguir estão detalhadas algumas limitações identificadas para o cenário de Ambulâncias Conectadas, do ponto de vista da disponibilização de serviços de computação em nuvem, limitação de recursos computacionais de dispositivos móveis, a necessidade de mobilidade e a alocação dinâmica de serviços virtualizados como uma alternativa para enfrentar esses desafios.

4.7.1 Provisionamento de Serviços de Computação

Considerando a demanda por alta capacidade computacional no contexto das Ambulâncias Conectadas para a execução de algoritmos de visão computacional e sistemas de classificação e recomendação, que utilizam técnicas de Inteligência Artificial e Aprendizado de Máquina [268–271], uma alternativa viável é o *offloading* de tarefas. Essa estratégia envolve

transferir tarefas que anteriormente seriam processadas em dispositivos móveis, como óculos inteligentes, para sistemas de computação em nuvem com maior capacidade. Tal abordagem minimiza a necessidade de processamento de algoritmos complexos diretamente nos dispositivos, promovendo maior autonomia e oferecendo uma solução mais eficiente em termos de custo [272–275].

Por outro lado, deve-se considerar que a transferência da execução de tarefas para ambientes com maior capacidade computacional ainda deve atender aos requisitos de comunicação impostos para o cenário crítico de Ambulâncias Conectadas. Isso se torna um desafio, pois a latência imposta para essas aplicações pode não ser atendida em casos onde a rede está congestionada e/ou em cenários onde os servidores de computação em nuvem estão localizados em locais muito distantes. Além disso, é essencial considerar as limitações das tecnologias de comunicação atuais, conforme destacado por Zhai *et al.* [42]. Eles demonstraram que as redes de comunicação existentes, como as redes sem fio de Longo Termo de Evolução (LTE), não atendem às exigências de ambulâncias inteligentes. Consequentemente, propuseram a implementação de uma rede sem fio baseada em 5G para suportar as operações de ambulâncias inteligentes.

Adicionalmente, a utilização de modelos de IA em aplicações de saúde em plataformas de nuvem apresenta desafios formidáveis. O modelo convencional de computação em nuvem luta para desbloquear todo o potencial da IA em várias organizações e casos de uso, principalmente devido a preocupações com latência e largura de banda, como detalhado em [276]. Além disso, o processo de treinamento ou ajuste fino de Modelos Grandes de IA (LAMs), para realizar diagnósticos de eletrocardiograma, por exemplo, exige uma quantidade extensa de tempo e recursos, ultrapassando as capacidades financeiras de muitas entidades de pesquisa e organizações, conforme descrito em [277]. Além disso, o custo associado à inferência também pode ser proibitivamente alto devido ao tamanho substancial do modelo. Isso torna impraticável para a maioria dos hospitais implantar esses LAMs localmente, utilizando sua infraestrutura computacional existente, como enfatizado em [278]. Consequentemente, há uma crescente necessidade por serviços compartilhados e acessíveis para atender a essa demanda.

Um uso das aplicações de Realidade Aumentada no setor de saúde é facilitar um espaço de trabalho compartilhado e livre de mãos, permitindo que os paramédicos prestem cuidados

enquanto simultaneamente recebem instruções de um especialista remoto [279]. No entanto, o desempenho dessas aplicações quando integradas a ambientes de computação em nuvem fica aquém de atender aos requisitos para aplicações de emergência críticas em serviços de ambulância, uma vez que o offloading remoto também introduz atrasos de transporte, que são considerados inaceitáveis para aplicações de RA de ultra-baixa latência [43]. Naqvi *et al.* [280], por exemplo, examinaram a utilização de recursos e os compromissos de desempenho ao estender uma aplicação de Realidade Aumentada (RA) com conscientização de contexto e computação em nuvem ao transferir a tarefa de reconhecimento de objetos para a nuvem. Apesar dessa abordagem, a latência de reconhecimento permaneceu alta, com um atraso máximo de 2 segundos. Zhang *et al.* [281] avaliaram a Realidade Aumentada em dispositivos móveis, focando especialmente no offloading para a nuvem para tarefas visuais computacionalmente intensivas em sistemas AR comerciais populares. Sua principal descoberta é que o reconhecimento baseado em nuvem não está bem otimizado para baixa latência, uso eficiente de dados e consumo de energia. Embora o offloading do reconhecimento de objetos para a nuvem seja recomendado, a intensidade computacional nos servidores comerciais ainda leva a tempos de execução de tarefas superiores a 150 ms.

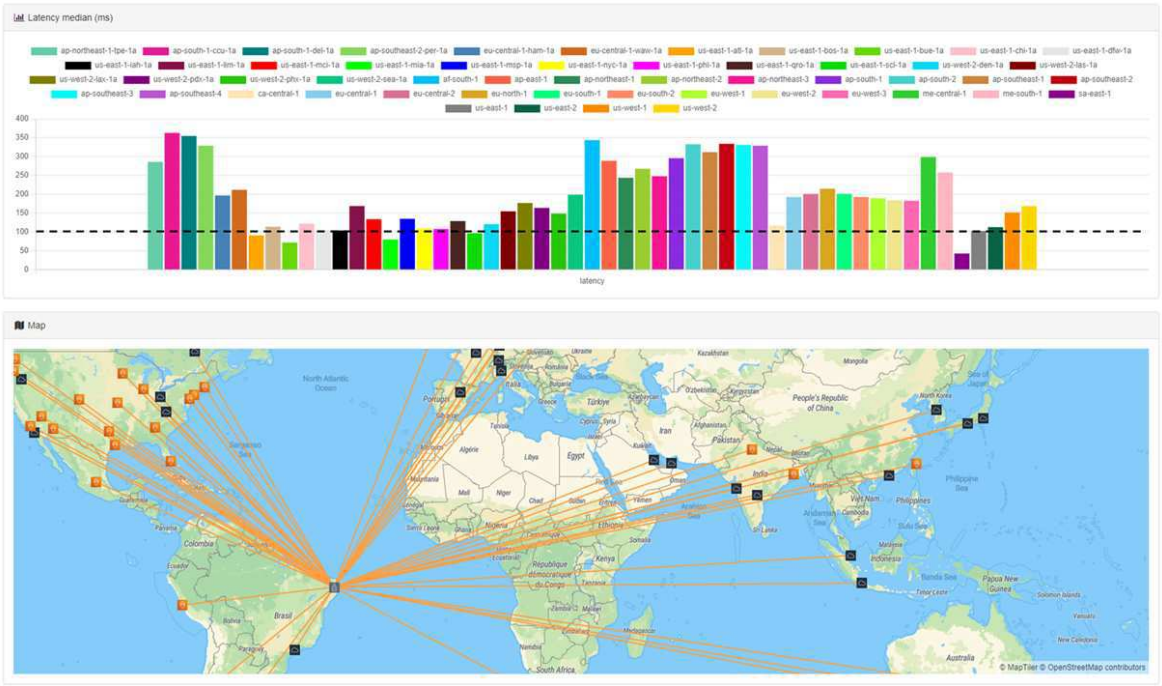
A latência da rede do Microsoft Azure [2] pode ser encontrada em [282]. Observamos que, enquanto algumas regiões alcançam baixas latências, muitas ainda experimentam atrasos acima de 100ms. Isso é significativo para aplicações críticas de saúde, onde alta latência pode impedir o desempenho. Na saúde, o processamento de dados em tempo real e tempos de resposta rápidos são cruciais, especialmente em telemedicina, monitoramento remoto e outros serviços sensíveis ao tempo. Portanto, a variabilidade no desempenho da rede do Azure sugere que pode não ser a solução mais eficiente para tais aplicações críticas de saúde.

Uma ferramenta para observar as latências dos servidores da Amazon Web Services (AWS)[3] é apresentada em [283], onde é medida a latência média para regiões e zonas locais a partir da localização do usuário, sendo essencial para avaliar o desempenho da rede. Embora não seja um projeto oficial, essa ferramenta traz uma avaliação significativa dos valores de latência em tempo-real para os servidores em nuvem. Os resultados de um teste realizado em Campina Grande-PB são ilustrados na Figura 4.2. As barras coloridas indicam

[2] https://azure.microsoft.com/

[3] https://aws.amazon.com/

Figura 4.2: Latências e Mapa obtidos do Teste de Latência da AWS.



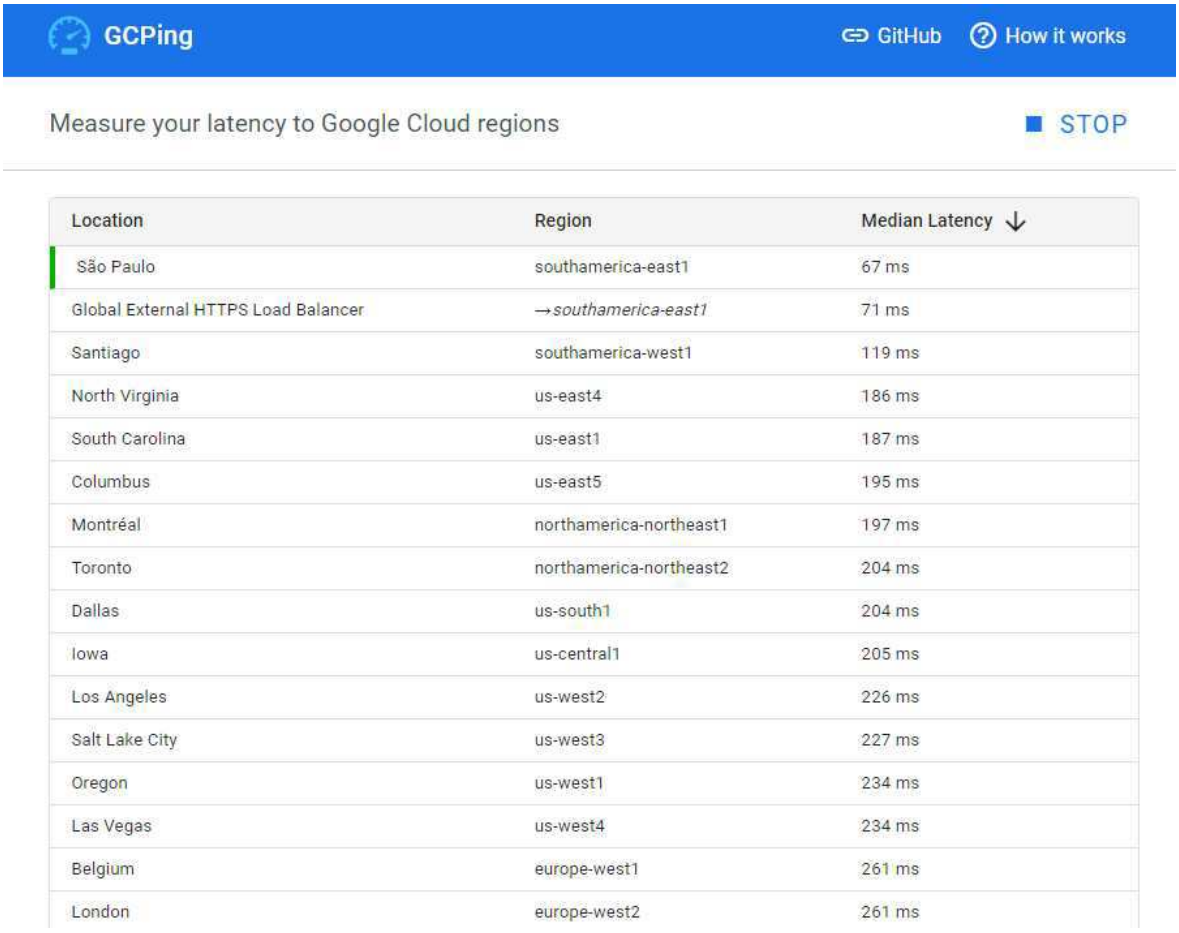
Fonte: Produzida pelo autor.

a latência média em milissegundos para diferentes regiões, enquanto o mapa mostra conexões de rede e tempos de resposta associados. A menor latência foi de 45ms para a região de São Paulo, mas a maioria das outras regiões teve latências acima de 100 ms. Isso indica que a proximidade geográfica é crucial para a eficiência da rede, e para aplicações onde baixa latência é fundamental, a escolha da região é determinante.

Em [284] é apresentada uma ferramenta que permite medir a latência para regiões do Google Cloud Platform (GCP)[4]. A partir do teste realizado na cidade de Campina Grande-PB, a menor latência registrada também foi para a região de São Paulo, com um valor médio de 67ms, conforme ilustrado na Figura 4.3. No entanto, assim como os resultados da AWS, as latências para outras regiões do Google Cloud também foram superiores a 100ms. Esta observação consistente entre diferentes provedores de serviços em nuvem destaca a necessidade crítica de estratégias alternativas para alcançar a redução da latência, especialmente para aplicações sensíveis à latência, onde mesmo unidades de milissegundos podem fazer uma diferença significativa.

[4]https://cloud.google.com/

Figura 4.3: Latências e Mapa obtidos do Teste de Latência da GCP.



Location	Region	Median Latency ↓
São Paulo	southamerica-east1	67 ms
Global External HTTPS Load Balancer	→*southamerica-east1*	71 ms
Santiago	southamerica-west1	119 ms
North Virginia	us-east4	186 ms
South Carolina	us-east1	187 ms
Columbus	us-east5	195 ms
Montréal	northamerica-northeast1	197 ms
Toronto	northamerica-northeast2	204 ms
Dallas	us-south1	204 ms
Iowa	us-central1	205 ms
Los Angeles	us-west2	226 ms
Salt Lake City	us-west3	227 ms
Oregon	us-west1	234 ms
Las Vegas	us-west4	234 ms
Belgium	europe-west1	261 ms
London	europe-west2	261 ms

Fonte: Produzida pelo autor.

A tecnologia de computação em nuvem, que se baseia em ferramentas para análise de dados em grande escala, é empregada para entregar desempenho ótimo, escalabilidade e suporte para soluções de IoT que não demandam por baixíssimas latências. No entanto, em situações críticas onde a disponibilidade de recursos e eficiência são primordiais, desconectar da rede principal ou encontrar diferenças significativas de latência pode ter consequências graves e potencialmente fatais durante emergências. A evolução contínua de *frameworks* explorando a sinergia de computação em nuvem e em borda ainda se apresenta como uma questão complexa e desafiadora [285].

4.7.2 Limitação dos Recursos Computacionais

Uma solução promissora para esses casos é a utilização da Computação na Borda. Ao discutir os benefícios da computação na borda, torna-se evidente que uma quantidade substancial de dados temporários pode ser processada na borda sem a necessidade de serem carregados para servidores em nuvem. Essa redução na transmissão de dados não só reduz a pressão sobre a largura de banda e diminui o consumo de energia dos *data centers*, mas também reduz substancialmente a latência e melhora a qualidade do serviço [36–38].

As vantagens do uso da Computação na Borda para reduzir a latência no processamento em tempo-real, especialmente em cenários de IoT sensíveis ao tempo, são apresentadas por Zen *et al.* [35]. O estudo destaca que a distância geográfica entre as fontes de dados e a infraestrutura de nuvem contribui significativamente para a latência da rede em casos de uso críticos de IoT. Para abordar essa questão, eles recomendam localizar servidores em nuvem mais próximos aos dispositivos IoT. Além disso, enfatizam o papel da computação na borda e na névoa em reduzir a distância física e recomendam-nas como soluções eficazes para minimizar a latência em aplicações sensíveis ao tempo.

No cenário em questão, ambulâncias conectadas se comunicam com nós de processamento na borda, onde os dados de saúde dos pacientes são coletados e processados em tempo real. No entanto, essa abordagem enfrenta desafios significativos. Um problema é a capacidade limitada de processamento e armazenamento nos dispositivos de borda, que podem não ter recursos suficientes para lidar com cargas de trabalho complexas ou que exigem muitos recursos. Isso pode levar a atrasos no processamento de dados ou capacidade insuficiente para lidar com grandes volumes de informações. Além disso, a manutenção e atualização desses dispositivos também podem ser problemáticas, pois estão distribuídos em diferentes localizações e podem exigir intervenções técnicas locais.

Aplicações de AR, em particular, emergem como fortes candidatas para offloading na borda, dada a sua demanda por computação intensiva e latência ultra-baixa para garantir a entrega de serviços de alta qualidade aos usuários finais [286]. Neste contexto, a Computação na Borda desempenha um papel crucial complementar, permitindo o processamento de dados em tempo-real e melhorando as capacidades de interação do usuário local [49]. Ren *et al.* [50], por exemplo, propuseram uma nova arquitetura baseada na Computação na Borda para abordar os desafios de mitigar longos atrasos de processamento e alto consumo de energia

em aplicações móveis de AR. Os autores consideraram um cenário onde vários dispositivos móveis no alcance do mesmo BS executam aplicações de RA simultaneamente.

Zhang *et al.* [287], propuseram uma abordagem de offloading de tarefas, baseada em Aprendizado por Reforço Profundo, para tarefas relacionadas à transmissão de vídeo (renderização e codificação) em um cenário de RA Móvel Multiusuário. Os principais objetivos desta abordagem são maximizar a Qualidade de Experiência (QoE) e reduzir os custos de recursos, incluindo a redução da latência de interação, despesas de largura de banda e custos de aluguel de servidores.

A utilização da computação na borda para melhorar a Qualidade de Experiência (QoE) em cenários de metaverso é explorada em [248], com um foco específico em serviços de Realidade Aumentada dependentes da localização dentro de sistemas de Metaverso habilitados para MEC. Neste sistema, cada usuário inicialmente realiza localização simultânea e comunicação para estimar sua posição e solicitar conteúdo de RA do servidor MEC. Dadas as restrições de recursos de computação e comunicação, o servidor MEC ajusta dinamicamente a resolução do conteúdo de RA para otimizar a experiência do usuário.

No estudo realizado por Park *et al.* [288], a Computação na Borda é utilizada para realizar tarefas de modelagem de objetos 3D intensivas em recursos, incluindo fatiamento de objetos virtuais e simplificação de malhas, aproveitando recursos de computação e armazenamento amplos. Além disso, a comunicação Direta-para-Dispositivo (D2D) é utilizada para mitigar atrasos de transmissão por meio do compartilhamento de objetos virtuais de interesse entre usuários próximos.

As descobertas de uma pesquisa indicaram que os desenvolvimentos recentes em aplicações de realidade aumentada (RA) necessitam de tempos de resposta de serviço genuinamente em tempo real [51–54]. Métodos computacionais tradicionais como a Computação em Nuvem são incapazes de atender às crescentes exigências para processamento de RA [55].

Kssinger *et al.* [289] melhoraram a velocidade das simulações de RA em tempo real por meio da utilização de um sistema de Computação na Borda chamado SimEdge. Este sistema melhorou efetivamente o desempenho do offloading em tempo real ao distribuir os cálculos para dispositivos situados próximos ao usuário final. Além disso, o experimento deles destacou a influência significativa da elevada latência da rede no desempenho de aplicações de

simulação pervasivas em tempo real. A latência de entrada observada de 45,84 ms, embora abaixo da média da indústria de 52,5 ms para dispositivos convencionais de realidade aumentada, ainda superou o limiar de atraso ideal que a maioria dos artigos científicos concorda [290, 291].

Segundo o relatório técnico apresentado em [292], a latência de ponta-a-ponta (E2E) da captura de uma nova imagem até a exibição da imagem aumentada deve ser inferior a 50 ms para prevenir a ciberdoença. Para alcançar isso, é necessário uma latência E2E unidirecional do sistema de comunicação dentro de 10 ms, abrangendo as latências combinadas das redes de acesso e *backhaul*. No entanto, segundo Akyildiz e Guo [293], como a Realidade Aumentada integra conteúdo virtual com o ambiente real inerentemente dinâmico — como cenários envolvendo paramédicos atendendo pacientes em locais de acidentes de trânsito — torna-se imperativo que as soluções de AR sejam altamente responsivas a essas situações dinâmicas. Consequentemente, aplicações caracterizadas por alta dinâmica necessitam de renderização de vídeo com uma latência ultra-baixa, visando tipicamente menos de 8,3 ms. Em contraste, em cenários com dinâmicas mais fracas ou ambientes estáticos, uma latência maior pode ser considerada aceitável.

No contexto da utilização de AR assistida por IA, um algoritmo de otimização para alocação de tarefas de IA é introduzido em [294]. Este algoritmo leva em conta a complexidade da tarefa, a capacidade do servidor e as condições atuais da rede. Os resultados experimentais ilustram que a abordagem unificada supera significativamente os métodos baseados em nuvem convencionais em termos de tempo de resposta e eficiência computacional, destacando assim o potencial dos sistemas de AR assistidos por computação em borda para aplicações práticas.

Wang et al. [295] revisaram os desenvolvimentos recentes na Inteligência Artificial Gerativa (GenAI) e suas implicações para a computação na borda e na nuvem. A GenAI é especializada em gerar conteúdo semelhante ao humano, resultando em um aumento de dados na Internet. Os serviços de GenAI atualmente dependem da computação em nuvem tradicional, levando a problemas de latência devido à transmissão de dados e às solicitações de usuários.

Considerando o estado-da-arte, embora as soluções existentes contemplem o uso de servidores na borda para reduzir a latência em aplicações de AR e IA, nenhum dos trabalhos

apresentados leva em conta a mobilidade do usuário entre nós de borda, o que necessita de estratégias eficientes de *offloading*, migração de cache e/ou alocação de serviços para atender aos requisitos dessas aplicações.

4.7.3 Necessidade de Mobilidade

No contexto do atendimento de emergência, a mobilidade é um fator crítico a ser considerado. As ambulâncias precisam ser capazes de se deslocar rapidamente até o local onde a assistência médica é necessária, e isso pode apresentar desafios para as infraestruturas de comunicação e computação envolvidas. Por exemplo, garantir uma conectividade estável e confiável em movimento pode ser complicado, especialmente em áreas com cobertura de rede irregular. Segundo Qureshi *et al.* [56], esse caso de uso particular pode enfrentar efeitos adversos em cenários caracterizados por mobilidade excepcionalmente alta, populações densas de usuários, situações de desastre com um aumento de ambulâncias convergindo para um único local, interrupções na rede celular ou a coexistência de múltiplos fluxos críticos de tráfego dentro da rede.

Khan et al. [296] destacam que o tratamento de emergência para um paciente gravemente doente em uma ambulância em movimento a caminho do hospital exige a migração de recursos entre servidores de borda. No entanto, o estado-da-arte atual em migração de recursos é insuficiente, indicando a necessidade de mais pesquisas para garantir a Qualidade de Serviço (QoS) necessária em vários casos de uso envolvendo migração de recursos entre servidores de computação em borda.

Em ambientes de Computação na Borda, gerenciar efetivamente a mobilidade do usuário representa um desafio significativo. Mehrabi *et al.* [297] enfatizam que a mobilidade dos usuários, caracterizada por *handovers* frequentes entre estações base, pode potencialmente contrariar as vantagens do *offloading*. Portanto, é imperativo incorporar considerações de mobilidade do usuário em esquemas de *offloading* para atender aos rigorosos requisitos de baixa latência. Adaptar-se aos padrões de mobilidade dinâmica dos usuários finais apresenta um desafio de pesquisa, necessitando do desenvolvimento de métodos adaptativos de *caching* e *offloading*. Além disso, a mobilidade desempenha um papel crucial em aplicações críticas de latência como Realidade Aumentada [298].

A mobilidade é identificada como um desafio significativo na implementação de solu-

ções de Computação na Borda, como sugerido por diversos estudos de pesquisa [57–61]. Para enfrentar esse desafio e garantir o acesso rápido aos serviços de saúde para usuários situados em locais diversos, é crucial considerar o recurso de mobilidade dos sensores de saúde inteligentes, dispositivos e equipamentos em múltiplos Servidores de Borda conectados. Surpreendentemente, esse aspecto foi negligenciado na maioria dos trabalhos de pesquisa examinados. Além disso, para melhorar a eficiência de aplicações críticas de tempo que lidam com dados distribuídos em locais geo-distribuídos, é imperativo focar e aprimorar a mobilidade de dados.

A migração contínua e suave de uma aplicação entre vários servidores de Computação na Borda, mesmo enquanto o usuário final está em movimento, é um mecanismo crucial conhecido como entrega de serviço contínua (do inglês, *seamless service delivery*) [64]. Alcançar a entrega de serviço contínua no contexto de mobilidade apresenta uma tarefa desafiadora porque a mobilidade impacta significativamente vários parâmetros de rede, incluindo latência, largura de banda, atraso e jitter, que, por fim, levam à degradação do desempenho da aplicação [65].

4.7.4 Alocação Dinâmica de Serviços Virtualizados

A alocação de serviços virtualizados refere-se ao fornecimento e gerenciamento de recursos computacionais virtualizados para diferentes aplicações e dispositivos. No cenário em questão, a infraestrutura de TI precisa alocar recursos virtualizados para as ambulâncias conectadas a fim de suportar o processamento em tempo real e as necessidades de RA.

No entanto, a alocação eficiente de recursos pode ser desafiadora, pois é necessário levar em conta fatores como disponibilidade de recursos, demanda variável em diferentes tempos e locais, proximidade com dispositivos de borda e latência de rede. Além disso, a escalabilidade da infraestrutura também é um problema, pois a demanda por serviços pode variar amplamente em emergências. Garantir uma alocação adequada de recursos virtualizados é crítico para assegurar o desempenho consistente e confiável do sistema de recomendação de procedimentos médicos em tempo real.

Huang *et al.* [66] investigaram a eficiência e o desempenho de um sistema de Computação Multi-Acesso em Borda (MEC) ao fornecer serviços de baixa latência e alta confiabilidade para usuários móveis, considerando cenários de mobilidade de usuários estáticos e

dinâmicos. Experimentos reais foram conduzidos, e os resultados revelaram que o sistema exibe baixa latência e alta confiabilidade ao lidar com tarefas intensivas em computação. No entanto, também ficou evidente que uma abordagem de alocação de recursos estática, como um único servidor, não aborda completamente os desafios, particularmente em cenários críticos.

Qiao *et al.* [67] apresentam desafios associados ao domínio da Realidade Aumentada (RA). O esforço para minimizar a latência em aplicações de RA é uma empreitada complexa que envolve aprimorar algoritmos de processamento de vídeo e alocar recursos dinamicamente entre nós de borda frontais e móveis. Embora a tecnologia 5G possa aliviar os gargalos de latência nas redes de acesso, permanece um desafio formidável.

Nesse sentido, o desafio principal é realizar a alocação dinâmica de serviços de streaming multimídia virtualizados em tempo-real para aplicações de AR e IA. O objetivo é garantir os atributos de qualidade de serviço, levando em conta a demanda variável e a disponibilidade de recursos ao longo de um determinado trajeto.

4.8 Considerações finais

Conforme apresentado, as tecnologias emergentes desempenham um papel crucial na transformação dos serviços médicos de emergência, destacando-se a implementação de óculos de realidade aumentada suportados por IA em ambulâncias. Essas tecnologias não apenas potencializam as capacidades diagnósticas e operacionais dos profissionais de saúde em emergências, mas também asseguram uma comunicação eficaz e protegida entre as equipes em campo e os especialistas localizados remotamente.

A introdução de dispositivos inovadores, como óculos de realidade mista e sensores sem fio para o monitoramento de sinais vitais, é capaz de revolucionar o atendimento médico pré-hospitalar ao possibilitar diagnósticos mais rápidos e precisos, além de aprimorar significativamente a gestão do tratamento em contextos críticos. A habilidade de projetar hologramas tridimensionais de procedimentos médicos e a aplicação de análises automáticas para avaliação de cenários emergenciais reforça a capacidade de tomar decisões rápidas e bem fundamentadas, que são essenciais em situações de urgência médica.

No entanto, os desafios tecnológicos e de infraestrutura mencionados anteriormente des-

tacam a necessidade de desenvolver soluções avançadas em sistemas de computação e comunicação. Dependências como a exigência de uma conectividade constante e confiável, uma gestão eficiente de energia e uma capacidade computacional adequada mostram-se fundamentais para a eficácia e o sucesso dessas inovações em campo. A superação desses obstáculos requer avanços técnicos contínuos e uma colaboração estratégica entre os desenvolvedores de tecnologia, profissionais médicos, indústrias de equipamentos médicos e de telecomunicações, e autoridades de saúde pública.

Ao considerar os avanços necessários em infraestrutura de computação e comunicação, o desenvolvimento de estratégias para a alocação dinâmica de serviços críticos em sistemas de computação em borda torna-se mandatório para otimização dos recursos disponíveis. Esta abordagem promissora contribuirá com o aumento da eficiência e a resposta em tempo-real dos serviços médicos de emergência, maximizando a utilização de recursos e a resiliência do sistema em cenários críticos.

Capítulo 5

Método para Disponibilização Dinâmica de Recursos e Serviços

A alocação dinâmica de serviços críticos de saúde tem se tornado uma questão de urgente atenção no contexto de uma sociedade cada vez mais rápida e conectada. Com o aumento constante de casos de emergências médicas em locais remotos ou de difícil acesso, a necessidade de sistemas eficientes que possam responder prontamente e com precisão a essas situações torna-se crucial. A utilização de tecnologias emergentes, como realidade mista, análise de dados em tempo real e computação em borda, destaca-se como uma abordagem promissora para enfrentar esses desafios.

Atualmente, existe uma demanda crescente por inovações que possam melhorar a eficácia e eficiência dos serviços de saúde de emergência. A necessidade de minimizar o tempo de resposta e aumentar a precisão do atendimento em situações críticas impulsiona a busca por novas soluções que integrem tecnologia e capacidade de resposta rápida. Estas inovações são essenciais para adaptar os serviços de saúde aos desafios impostos por cenários de emergência complexos e variáveis.

Entre os subdesafios específicos relacionados a este cenário, destacam-se:*(i)* Como garantir uma comunicação eficaz e de baixa latência entre os dispositivos de borda em ambientes móveis?; *(ii)* Como otimizar a alocação de recursos computacionais em tempo-real para garantir o processamento eficiente de dados críticos?; e *(iii)* Como integrar de forma eficaz as tecnologias de Realidade Aumentada para auxiliar os paramédicos durante o atendimento em campo?

Desse modo, neste capítulo está detalhado o método proposto para solução do problema de alocação dinâmica de serviços críticos de saúde, denominado *Make Way*. Este método foi desenhado para otimizar a coordenação e disponibilização de serviços em cenários de emergência, utilizando tecnologias emergentes para uma resposta ágil e precisa. Com isso, propõe-se a integração vários módulos funcionais que trabalham em conjunto para contribuir com a melhoria dos serviços de saúde emergenciais.

Para a construção da proposta de solução, foram considerados os requisitos de comunicação para os casos de uso de aplicações críticas de saúde, priorizando o cenário de ambulâncias conectadas. Após estas definições, foi possível realizar a definição dos requisitos de sistema e executar experimentos para validação do problema: disponibilização de serviços críticos de saúde em cenários de mobilidade. Com base nos resultados e requisitos, foi elaborada a proposta de arquitetura para o problema abordado.

5.1 Etapas de Construção do Método

Visando sistematizar e viabilizar a compreensão dos passos realizados para definição da proposta de solução, na Figura 5.1 está apresentado o fluxograma das etapas de construção da solução apresentada neste trabalho.

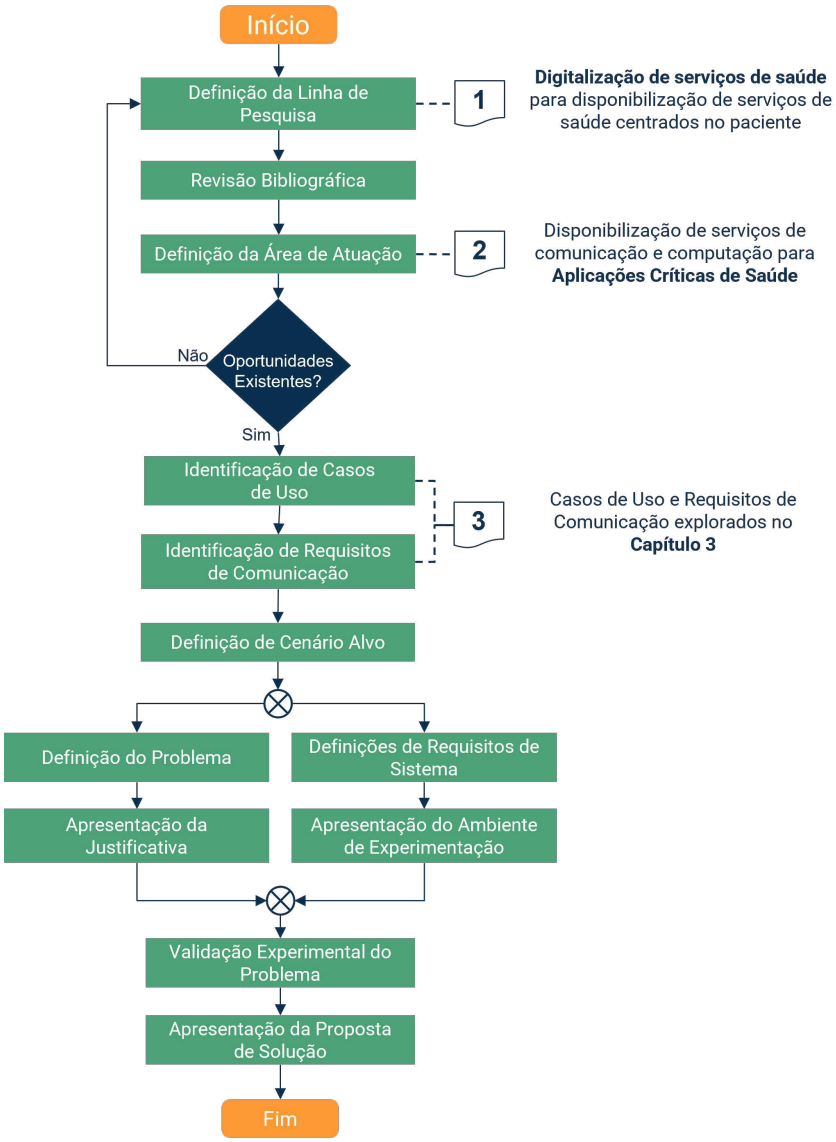
Inicialmente, na etapa **Definição da Linha de Pesquisa** foi identificada uma oportunidade relacionada à digitalização de serviços de saúde visando viabilizar a prestação de serviços médicos centrados no paciente, sendo esta a linha de pesquisa definida para este projeto, utilizada para direcionar as etapas subsequentes.

A partir da revisão bibliográfica apresentada no Capítulo 3, realizada na etapa **Revisão Bibliográfica**, foram identificadas características e classificações para aplicações de saúde que as tornam particulares e desafiadoras. Dentre essas características, podemos citar os requisitos associados à confiabilidade, segurança, latência, interoperabilidade, largura de banda e consumo de energia dos dispositivos médicos. Além disso, foi possível identificar dois grupos para aplicações de saúde: aplicações não críticas, geralmente envolvendo condições de bem-estar e monitoramento de saúde em situações de rotina, e aplicações críticas, aquelas envolvidas na prestação de cuidados para a sobrevivência do paciente.

De posse dessas informações, na etapa **Definição da Área de Atuação**, casos de uso

e requisitos de aplicações críticas de saúde foram investigados na literatura, que por sua vez apresentam estritos requisitos de comunicação. Disponibilizar serviços de comunicação e computação para Aplicações Críticas de Saúde ainda apresenta desafios e oportunidades associados, sendo esta a área de atuação definida para este trabalho.

Figura 5.1: Fluxograma das Etapas de Construção da Proposta de Solução.



Fonte: Produzida pelo autor.

Na etapa **Identificação de Casos de Uso**, casos de uso de aplicações críticas de saúde foram explorados na literatura, conforme apresentado na Seção 3.1.1 do Capítulo 3.

A partir disso, na etapa **Identificação de Requisitos de Comunicação**, foi possível identificar os principais requisitos de comunicação de aplicações críticas e cenários aborda-

dos, necessidades de tecnologias de comunicação sem fio e infraestruturas associadas, assim como características de comunicação de dados, por exemplo, *streaming* de dados de pacientes; *streaming* de áudio e vídeo, tanto para aplicações de consulta médica remota utilizando câmeras de alta definição e/ou óculos de realidade aumentada, como para transmissão de dados de dispositivos de diagnósticos por imagens; e *streaming* de dados de controle e atuação de dispositivos hápticos, que envolvem sinais de força e vibração. Cabe destacar que a transmissão de dados de saúde de pacientes pode apresentar requisitos de comunicação variados, a depender da criticidade do estado de saúde do paciente, da relevância dos parâmetros a serem monitorados para definição de condições clínicas, assim como do mecanismo ou dispositivo de medição utilizado.

Visando explorar o impacto da mobilidade em cenários de aplicações críticas de saúde, na etapa de **Definição de Cenário Alvo**, escolheu-se o caso de uso de Ambulâncias Conectadas. Atender os requisitos de comunicação e qualidade do serviço (do inglês, *Quality of Service* - QoS) disponibilizado é crucial em cenários críticos, uma vez que estão relacionadas diretamente com a prestação de cuidados para a sobrevivência do paciente.

Diante das possibilidades e benefícios identificados na adoção de sistemas de computação em borda e 5G, na etapa de **Definição do Problema**, foi identificado que a gestão e coordenação dos serviços se torna uma atividade complexa, à medida que a disponibilização desses serviços aumenta de escala. No contexto de disponibilização de serviços para Ambulâncias Conectadas, um outro fator agravante para a gestão e coordenação dos serviços de aplicações críticas de saúde é a mobilidade do usuário e/ou unidade de atendimento móvel, uma vez que a alocação de serviços acontecerá de maneira dinâmica em servidores localizados na borda da rede ao longo de um trajeto. Desse modo, na etapa de **Apresentação da Justificativa**, a inserção de serviços de saúde de modo inteligente e orquestrado por meio de serviços e infraestrutura na borda da rede pode auxiliar no acesso, distribuição e disponibilização de serviços de saúde à população de modo mais rápido e localizado. A problemática e a justificativa estão apresentadas nas Seções 1.2 e 1.1, respectivamente.

Paralelamente às etapas de **Definição do Problema** e **Apresentação da Justificativa**, na etapa de **Definições de Requisitos de Sistema** foram levantados os requisitos para implementação de um ambiente experimental, visando avaliar e simular a alocação de serviços em servidores na borda da rede em cenários de mobilidade para o caso de uso de Ambulâncias

Conectadas. Estes requisitos estão descrito na Seção 4.5. Posteriormente, na etapa de **Apresentação do Ambiente de Experimentação**, descrito na Seção 6.2, apresenta-se o ambiente experimental desenvolvido.

Uma vez concluídas estas etapas, foi realizada a **Validação Experimental do Problema**, na etapa subsequente, apresentadas na Seção 7.1.

A **Apresentação da Solução Proposta** está descrita nas próximas seções deste capítulo, onde também é a presentada a arquitetura do *Make Way*.

Considerando os requisitos de aplicações de saúde critica e os desafios em aberto destacados por Salaht et al. [219] e por Malazi et al. [63] referentes à alocação dinâmica de serviços em ambientes de Computação na Borda, identificou-se a possibilidade do uso de estratégias de alocação de serviços com base no padrão de mobilidade pois:

- A alocação baseada em padrões de mobilidade permite que os **serviços de computação se ajustem dinamicamente** conforme a localização da ambulância, melhorando a eficácia da entrega de serviços em tempo-real;
- A alocação de serviços com base no movimento dos usuários **reduz a latência e aumenta a eficiência do uso de banda larga e de processamento**, crucial para aplicações de saúde que demandam resposta rápida e dados precisos;
- Esta abordagem possibilita a resposta rápida a mudanças no ambiente ou no comportamento do usuário, essencial para manter a **continuidade e a qualidade do serviço** em cenários de saúde crítica;
- Ao sincronizar os recursos de computação com as necessidades em tempo-real das aplicações médicas, melhora-se significativamente a **experiência do usuário**, garantindo que as aplicações de saúde operem de forma eficiente e sem interrupções.

5.2 Premissas

No cenário de ambulâncias conectadas, um desafio crítico é a implantação de modelos de IA para recomendar ou inspecionar procedimentos médicos integrados com *streaming* multimídia de óculos inteligentes de AR em tempo-real enquanto a ambulância está em movimento.

Dadas as limitações computacionais dos óculos inteligentes e a necessidade de atualizações de modelos de IA, a execução desses modelos diretamente na ambulância torna-se um desafio. Portanto, considera-se que:

1. Devido às restrições computacionais dos óculos inteligentes e à necessidade de manter os modelos de IA atualizados, é essencial executá-los externamente à ambulância. Essa abordagem atende viabiliza maior autonomia para os dispositivos móveis imersivos, dado o *offloading* de tarefas computacionais, mas introduz um desafio adicional: a necessidade de um tempo de resposta baixo dos modelos de IA e do *streaming* multimídia para garantir a eficácia dos procedimentos recomendados ou inspecionados.

2. Para atender ao requisito de um tempo de resposta baixo, na ordem de 10 ms, torna-se necessário implantar esses modelos nos servidores de borda. Executá-los mais próximos ao ponto de uso pode reduzir a latência na comunicação e no processamento, o que é crucial em situações médicas de emergência.

3. Considerando a natureza móvel das ambulâncias, é crítico garantir que os serviços de IA e *streaming* multimídia estejam sempre em execução no servidor de borda mais próximo. Isso significa que, à medida que a ambulância se move, esses serviços devem ser alocados dinamicamente para manter a proximidade com o veículo. Essa abordagem dinâmica é vital para manter a eficiência necessária nos serviços de IA e *streaming* multimídia, garantindo que a assistência médica seja rápida e eficaz, independentemente da localização da ambulância.

Esse cenário destaca a importância de uma infraestrutura de TI flexível e responsiva, capaz de se adaptar continuamente às mudanças de localização das ambulâncias, garantindo que as tecnologias de IA e AR possam ser efetivamente utilizadas no atendimento de emergência.

5.3 Detalhamento do Método

O método proposto tem por finalidade a disponibilização de mecanismos de gerenciamento e alocação dinâmica de serviços, projetada para viabilizar a alocação dinâmica de serviços em

cenários de emergência. O caso de uso escolhido como objeto de estudo engloba veículos de emergência, como ambulâncias, e recursos tecnológicos avançados, como Realidade Aumentada (AR) e Inteligência Aumentada (IA). O funcionamento de alto nível desse sistema é ilustrado através de uma série de diagramas de sequência (Figuras 5.2, 5.3, 5.4 e 5.5) que delineiam o processo integrado desde a solicitação inicial até a configuração e alocação de recursos, com certo nível de abstração.

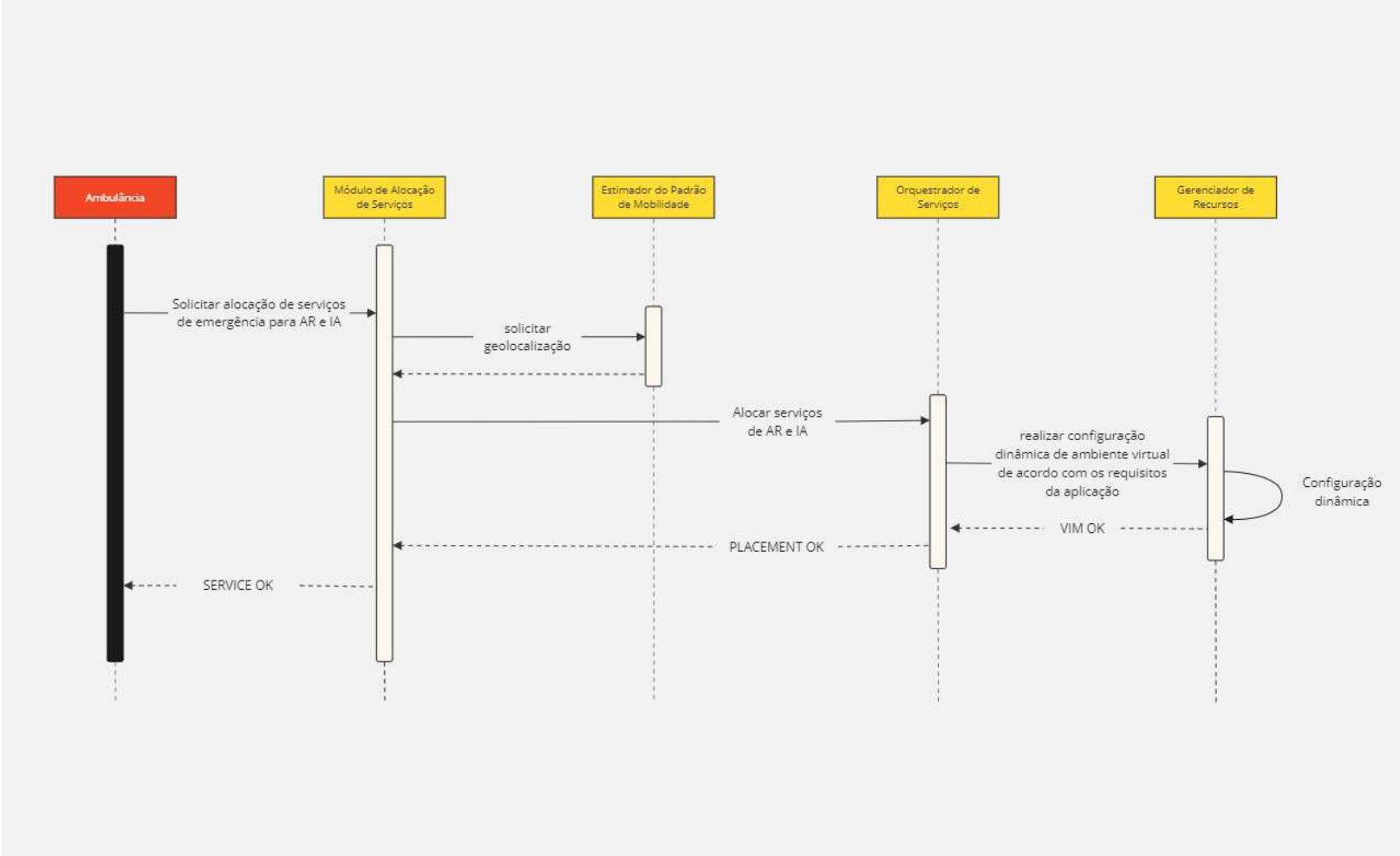
Inicialmente, o processo é ativado quando uma ambulância solicita serviços de emergência específicos ao Módulo de Alocação de Serviços. Essa solicitação inicial é fundamental para desencadear uma cadeia de operações automatizadas dentro do sistema. O primeiro passo subsequente envolve a obtenção da geolocalização da ambulância, uma funcionalidade fornecida pelo Estimador do Padrão de Mobilidade. Esta etapa é crucial, pois o sistema depende da localização precisa para estimar os padrões de mobilidade do veículo, uma informação que influencia diretamente a alocação dos serviços requeridos.

Após a determinação da localização, o Orquestrador de Serviços procede com a alocação efetiva dos serviços de AR e IA. Este processo é confirmado por um indicador "PLACEMENT OK", que assegura que os serviços foram alocados corretamente e estão prontos para serem utilizados pela ambulância.

Adicionalmente, o sistema inclui um Gerenciador de Recursos, que é responsável por realizar configurações dinâmicas do ambiente virtual de acordo com os requisitos da aplicação em uso. Esta configuração é essencial para assegurar que os recursos computacionais estão otimizados e adaptados às necessidades específicas da operação em curso, sendo esta etapa validada pelo retorno "VIM OK", indicando que a infraestrutura virtual, abrangendo máquinas virtuais e funcionalidades de redes virtualizadas, está preparada para utilização. Este fluxo inicial está detalhado na Figura 5.2.

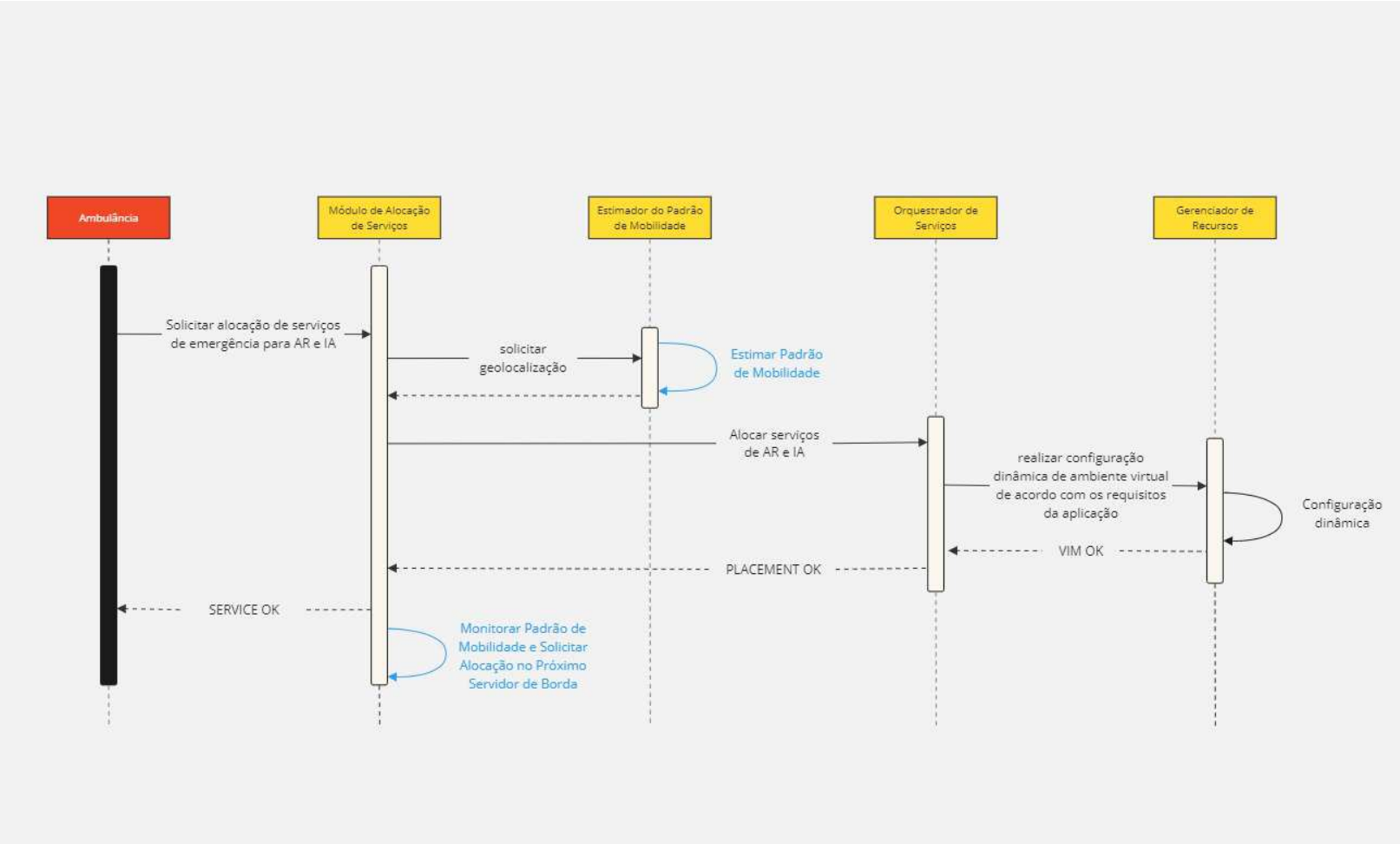
Durante o processo de inicialização, dois novos processos são disparados, conforme destacado em azul na Figura 5.3. O primeiro está relacionado à estimativa do Padrão de Mobilidade da Ambulância. Inicialmente, a ação de solicitar geolocalização ao Estimador do Padrão de Mobilidade é fundamental para obter a localização precisa da ambulância. Esta informação é essencial para que o sistema possa determinar com precisão os padrões de mobilidade, permitindo uma alocação mais eficiente dos recursos de Realidade Aumentada (AR) e Inteligência Aumentada (IA).

Figura 5.2: Diagrama de Sequência de Alocação de Serviços do método *Make Way*.



Fonte: Produzida pelo autor.

Figura 5.3: Diagrama de Sequência de Alocação de Serviços do método *Make Way* - Monitoramento do Padrão de Mobilidade.



Fonte: Produzida pelo autor.

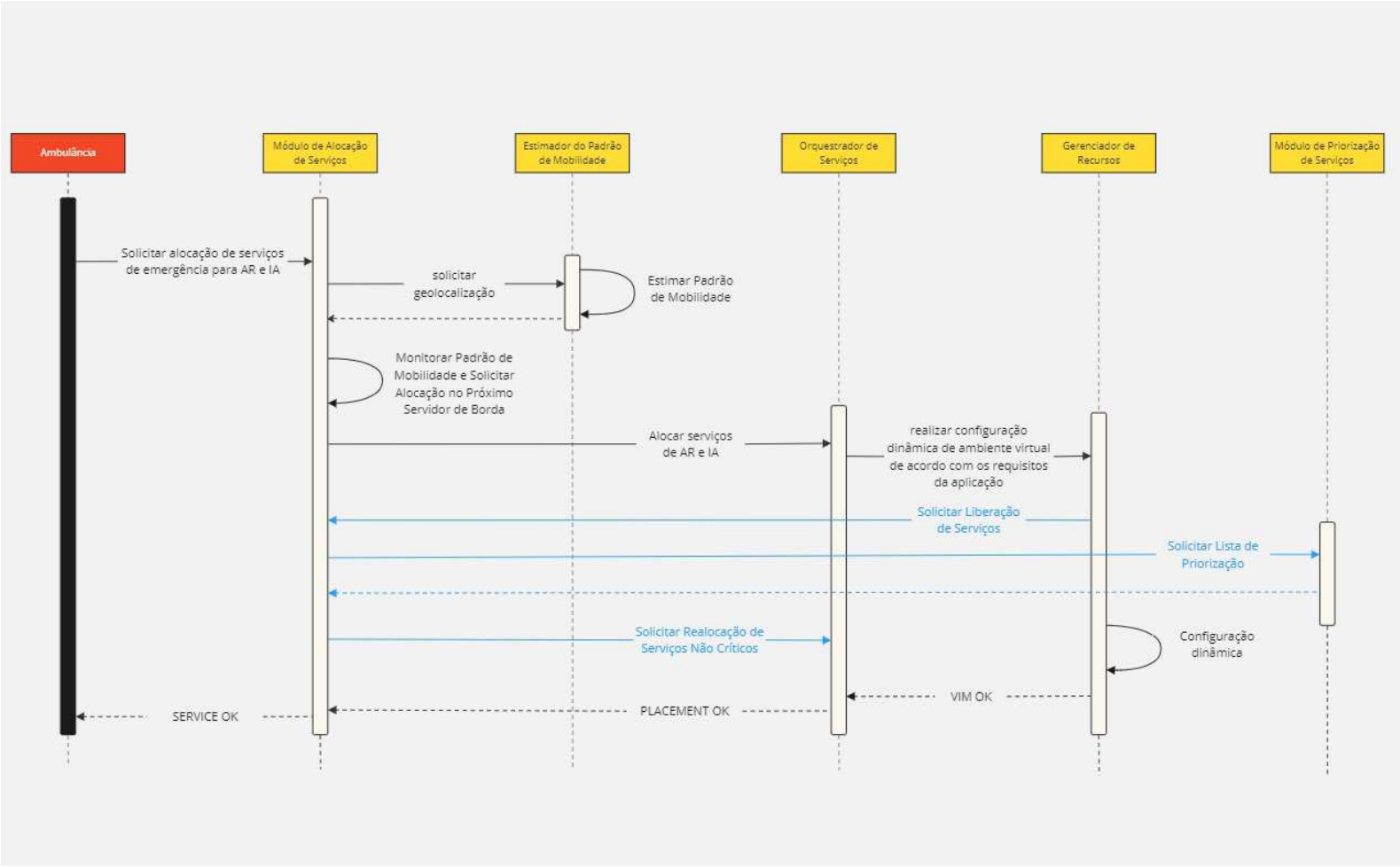
Posteriormente, há o processo de monitorar o padrão de mobilidade e solicitar alocação no próximo servidor de borda. Esta operação é de grande importância, pois que a alocação dos serviços seja disponibilizada, a priori, de acordo com o padrão de mobilidade do veículo, possibilitando uma resposta rápida e adequada às necessidades de emergência. Essa etapa assegura que os recursos necessários sejam provisionados de forma contínua e eficaz, minimizando latências e maximizando a eficiência do serviço prestado. Através deste mecanismo, o método *Make Way* mantém sua capacidade de responder prontamente a situações dinâmicas, adaptando-se às mudanças no ambiente e nas necessidades do usuário. Neste cenário descrito, assume-se como premissa que os recursos de computação em borda estão prontamente disponíveis para serem utilizados.

Contudo, considerando principalmente a necessidade de gerenciar a limitação de recursos em cenários onde múltiplas aplicações e serviços estão em operação, a funcionalidade de liberação e priorização de serviços se torna essencial. O Módulo de Alocação de Serviços e o Módulo de Priorização de Serviços, destacados em azul na Figura 5.4, desempenham papéis cruciais neste processo. A partir do momento em que surge a necessidade de alocação do serviços críticos em outros servidores de borda, conforme o padrão de mobilidade, e estes possuem disponibilidade de recursos, torna-se imperativo implementar mecanismos eficientes capazes de suprir as necessidades impostas pela dinâmica do funcionamento das aplicações.

O Módulo de Alocação de Serviços solicita a realocação dinâmica de serviços não críticos entre diferentes ambientes de execução, permitindo que o sistema adapte sua infraestrutura em resposta às flutuações de demanda e disponibilidade de recursos. Este módulo assegura que os serviços possam ser transferidos para outros servidores sem interrupção, mantendo a integridade e a disponibilidade das funções críticas.

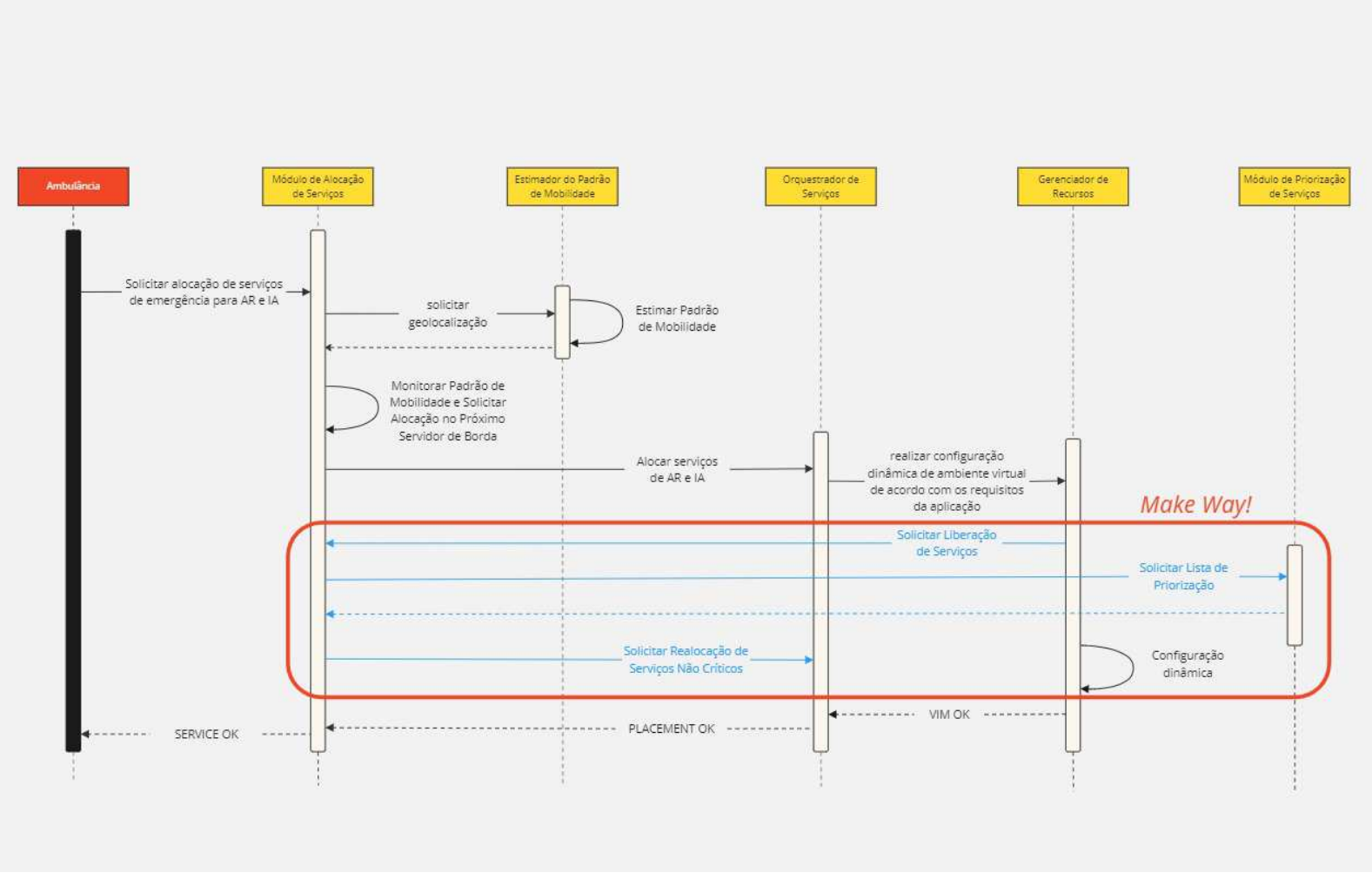
Paralelamente, o Módulo de Priorização de Serviços atua ajustando a ordem de priorização dos serviços baseado em critérios predefinidos, que podem incluir a urgência, a criticidade da aplicação e a disponibilidade de recursos. Esse módulo é particularmente vital em situações onde recursos limitados devem ser compartilhados entre várias aplicações operando simultaneamente, garantindo que os serviços mais críticos recebam a atenção e os recursos necessários sem atrasos.

Figura 5.4: Diagrama de Sequência de Alocação de Serviços do método *Make Way* - Solicitação de Liberação de Recursos.



Fonte: Produzida pelo autor.

Figura 5.5: Diagrama de Sequência de Alocação de Serviços do método *Make Way* - "Abrindo Caminho".



Fonte: Produzida pelo autor.

Destaca-se, na Figura 5.5, a ideia fundamental proposta com o método *Make Way*, que faz alusão ao conceito de "dar passagem", expressão usada em inglês para solicitar que motoristas abram caminho para ambulâncias em situações de tráfego intenso. Essa metáfora reflete o objetivo do método de facilitar um fluxo mais ágil e desobstruído de serviços críticos para os casos em que a infraestrutura de computação esteja com capacidade de alocação de novos serviços limitada.

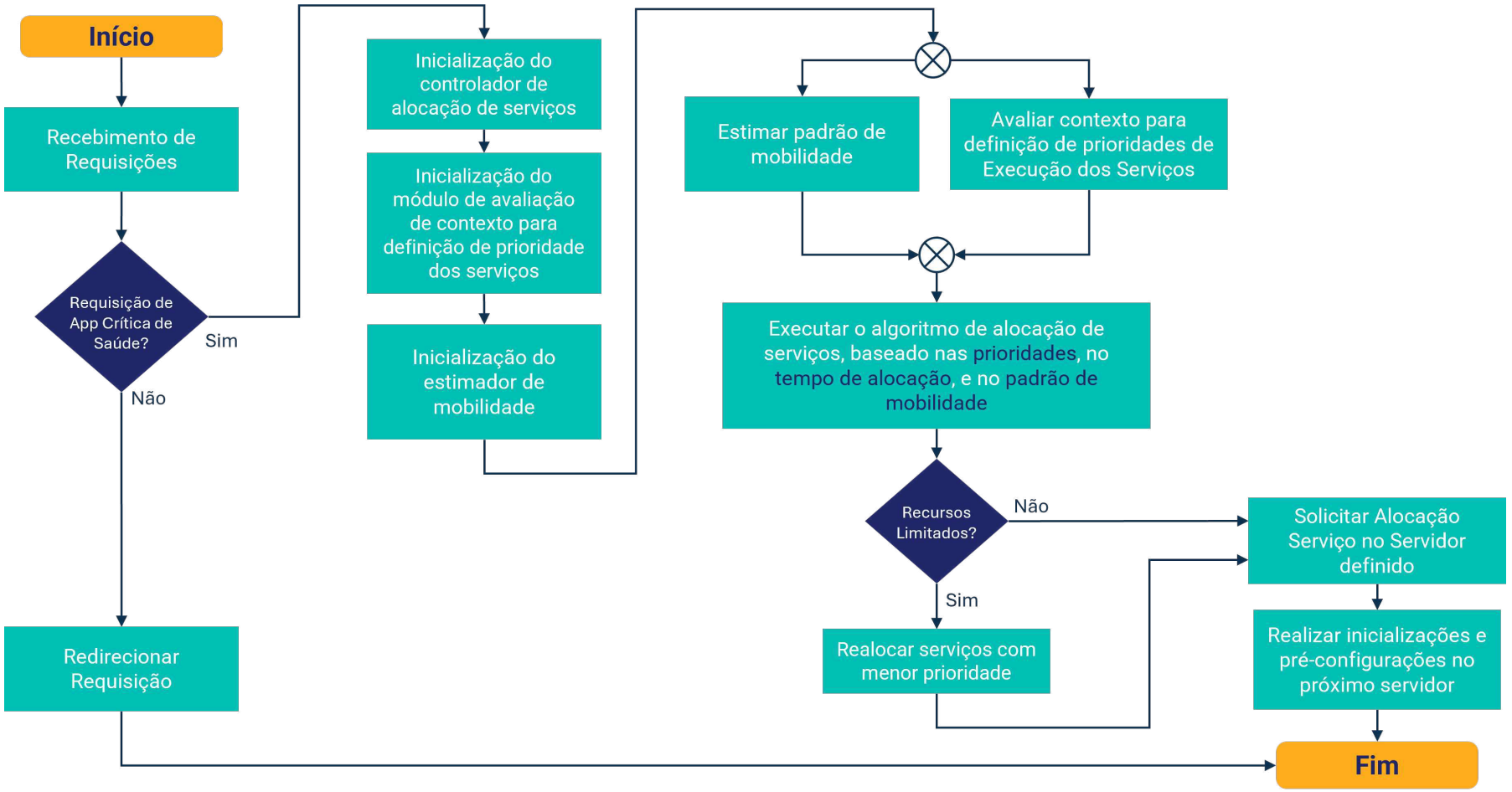
5.4 Protocolo de Alocação Dinâmica de Serviços baseado no Método *Make Way*

Na Figura 5.6 está apresentado um fluxo definido para o Protocolo de alocação dinâmica de serviços baseado no método *Make Way*. O protocolo inicia com o recebimento de requisições, onde é verificado se a solicitação vem de um aplicativo crítico de saúde, sendo redirecionada caso não seja. Em seguida, o sistema procede com a inicialização do controlador de alocação de serviços, do módulo de avaliação de contexto para determinação da prioridade dos serviços, e do estimador de mobilidade, preparando o sistema para avaliar e responder às demandas de mobilidade.

Paralelamente, são executados algoritmos de avaliação de contexto e a estimativa de padrão de mobilidade. O módulo de avaliação de contexto deve ser capaz de analisar as condições atuais para definir as prioridades na execução dos serviços, garantindo que as aplicações mais críticas sejam executadas prioritariamente. Além disso, o sistema utiliza o modelos eurísticos para estimar os padrões de mobilidade dos usuários, uma etapa fundamental para prever a demanda futura por serviços em diferentes locais. Esse modelo deve ser projetado para priorizar aplicações críticas, adaptando-se às mudanças dinâmicas na localização e nas necessidades dos usuários, o que viabilizando uma alocação de serviços proativa e dinâmica.

A etapa seguinte envolve a execução de algoritmos de alocação de serviços, visando determinar o servidor mais apto a executar o serviço baseado em prioridades, tempo de alocação e padrões de mobilidade. Após a seleção do servidor, o serviço é alocado e são realizadas as configurações iniciais necessárias para sua operação. Se necessário, serviços de menor prioridade devem ser realocados para outros servidores para liberar recursos para os serviços críticos de saúde.

Figura 5.6: Protocolo de Alocação Dinâmica de Serviços baseado no método *Make Way*.



Fonte: Produzida pelo autor.

5.5 Componentes-Chave do Método

O método foi idealizado com a finalidade de facilitar a disponibilização dinâmica de serviços críticos de saúde, proporcionando soluções customizadas e análises detalhadas que são fundamentais para apoiar a tomada de decisões em situações críticas. De modo geral, sua concepção foi inspirada na expressão em inglês usada para solicitar que motoristas **abram caminho** para a passagem de ambulâncias em áreas de tráfego denso, este sistema visa garantir que recursos de saúde essenciais cheguem rapidamente onde são mais necessários.

Desse modo, os componentes-chave do método são:

- **Orquestrador de Serviços**: Coordena a alocação de serviços médicos e de emergência com base na urgência e localização;
- **Gerenciador de Recursos**: Monitora e administra a capacidade de processamento, armazenamento e comunicação nos servidores de borda e nuvem;
- **Módulo de Alocação de Serviços**: Garante a alocação dinâmica dos serviços críticos entre diferentes nós de borda, assegurando a continuidade na disponibilização. Além disso, o módulo também é responsável por garantir a transferência de serviços não críticos para outros servidores, assegurando a liberação de recursos para execução dos serviços críticos;
- **Estimador do Padrão de Mobilidade**: Determina o padrão de mobilidade de ambulâncias através da análise de dados de tráfego e geolocalização em tempo real. Este componente visa fornecer informações que serão utilizadas para a alocação prévia de serviços em servidores de borda ao longo do trajeto, assegurando uma resposta ágil e eficaz conforme o padrão de mobilidade detectado;
- **Módulo de Priorização de Serviços**: Utiliza algoritmos de IA para priorizar casos com base na gravidade e nos recursos disponíveis.

5.6 Considerações Finais

Nesse capítulo foi apresentado o processo de desenvolvimento do método "*Make Way*", projetado para otimizar a alocação dinâmica de serviços críticos de saúde em ambientes de

emergência, utilizando tecnologias emergentes para resposta ágil e precisa. Foram descritos os desafios específicos do cenário, como garantir comunicação eficaz de baixa latência em dispositivos de borda e otimizar a alocação de recursos computacionais em tempo real.

Foram detalhadas as etapas de construção da solução, começando pela definição da linha de pesquisa, passando pela revisão bibliográfica e chegando à definição da solução proposta. Entre os componentes principais, destacam-se: um orquestrador de serviços, um gerenciador de recursos e módulos especializados como o de alocação de serviços e o estimador do padrão de mobilidade. Esses componentes são fundamentais para garantir a eficiência operacional do sistema em cenários de mobilidade, especialmente em serviços de saúde providos via ambulâncias conectadas.

Adicionalmente, foram apresentados o detalhamento do método e o protocolo de alocação dinâmica de serviços baseado no método *Make Way*, com a finalidade de atender de maneira proativa e eficiente às demandas em situações críticas. Este protocolo emprega algoritmos avançados para a gestão e priorização de recursos, assegurando que os serviços críticos sejam alocados e gerenciados com prioridade, minimizando a latência e maximizando a eficácia do atendimento.

A implementação do *Make Way* traz benefícios tanto técnico-científicos quanto sociais. Tecnicamente, a integração de novas tecnologias avança o tratamento de dados e a comunicação em tempo real, cruciais para decisões rápidas em emergências. Socialmente, essas inovações aceleram e melhoram o atendimento médico, elevando as taxas de recuperação dos pacientes e diminuindo os tempos de resposta em emergências.

O potencial de inovação do método *Make Way* reside na sua capacidade de transformar a maneira como os serviços de emergência são prestados. Através de uma infraestrutura robusta e da aplicação de tecnologias de ponta, é possível não apenas atender às necessidades atuais, mas também antecipar futuras demandas e adaptar-se a elas. Esta abordagem inovadora garante uma melhoria contínua na qualidade do serviço, sendo capaz de superar às expectativas atuais de pacientes e profissionais da saúde.

Capítulo 6

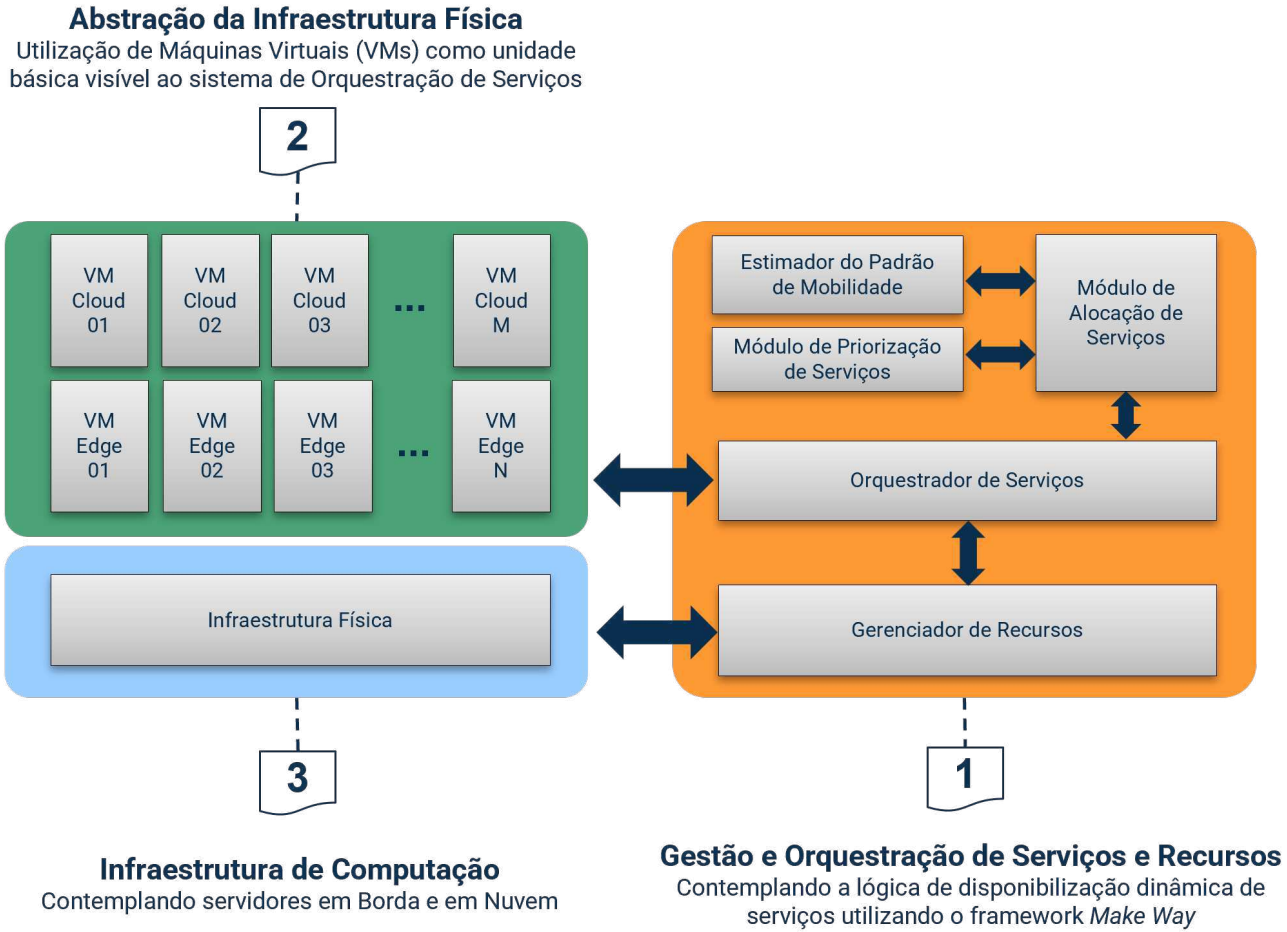
Prova de Conceito para Validação do Método

Neste capítulo está detalhado o ambiente experimental utilizado para validar o método *Make Way*, implementado através de Provas de Conceito. O objetivo principal é testar e demonstrar a eficácia deste método na simulação da mobilidade de ambulâncias em rotas específicas. O ambiente de simulação permite avaliar como a mobilidade de veículos de emergência pode ser otimizada utilizando o *Make Way*. Além disso, o capítulo explora a integração deste ambiente de simulação de mobilidade com um sistema de orquestração de serviços. Este sistema é baseado na ferramenta de orquestração Kubernetes, que gerencia a execução e interação dos diversos serviços que compõem a solução proposta.

6.1 Modelo Arquitetural da Prova de Conceito para Validação

O modelo arquitetural da Prova de Conceito para Validação da solução utilizando o método *Make Way* está apresentada na Figura 6.1, fornecendo uma visão de como os componentes interagem dentro do sistema proposto. A utilização deste modelo permite não apenas direcionar a implementação do método proposto, mas também oferece uma visão clara dos componentes necessários para implantação em casos reais. Nas próximas subseções serão detalhados os componentes e suas características.

Figura 6.1: Arquitetura do método *Make Way*.



Fonte: Produzida pelo autor.

6.1.1 Orquestrador de Serviços

O orquestrador de serviços desempenha um papel fundamental na gestão eficiente da implantação de aplicações e serviços em ambientes de computação dinâmicos, como sistemas de computação em nuvem e em borda. Isso envolve tarefas como mapear atividades para recursos, assegurar desempenho ótimo e gerenciar restrições de recursos [299].

De modo geral, o projeto do orquestrador é focado em flexibilidade, escalabilidade e resiliência, especialmente em cenários com dispositivos móveis de usuários e disponibilidade de recursos variável [300]. Ao tirar proveito da tomada de decisões distribuída e modelos baseados em agentes, os orquestradores podem aprimorar a escalabilidade, tolerância a falhas e latência de resposta em ambientes de computação em borda [301].

Além disso, orquestradores possibilitam a implantação rápido de serviços em ambientes de computação em nuvem, onde essa funcionalidade pode ser expandida para ambientes de computação em borda visando atender à necessidade de baixa latência e orquestração de serviços dinâmica. O desenvolvimento de orquestradores utilizando ferramentas de código aberto pode oferecer soluções custo-efetivas com flexibilidade aprimorada para gerir ambientes multiusuário e multimáquinas de maneira eficiente [302].

Nesse contexto, a adoção de uma ferramenta de orquestração de serviços, que é amplamente reconhecida em ambientes de produção e suportada por uma comunidade extremamente ativa, confere à solução uma significativa robustez e adaptabilidade. Esta escolha facilita integrações eficazes e promove uma evolução contínua, alinhando-se às exigências dinâmicas do mercado e às contínuas inovações tecnológicas. Portanto, considerando as características previamente delineadas, a ferramenta de orquestração de serviços selecionada para integrar a solução do método *Make Way* foi o Kubernetes.

6.1.2 Gerenciador de Recursos

O Gerenciador de Recursos é um componente fundamental dentro de uma infraestrutura de computação em borda e na nuvem. Este sistema é responsável por monitorar e administrar de forma eficiente a capacidade de processamento, armazenamento e comunicação dos servidores, garantindo que os recursos estejam disponíveis onde e quando necessário. Com a crescente demanda por aplicações que necessitam de baixa latência e alto desempenho, o ge-

renciamento eficaz de recursos torna-se um fator decisivo para o sucesso de operações tanto em ambientes de borda quanto na nuvem.

O gerenciador realiza o monitoramento dos recursos de modo contínuo afim de ajustar dinamicamente a alocação de recursos em resposta às mudanças na carga de trabalho e demanda dos serviços. Isso inclui a escalabilidade horizontal e vertical dos recursos, permitindo que as aplicações escalonem conforme necessário sem a intervenção manual, o que é essencial para lidar com picos de tráfego ou falhas em componentes de rede.

A capacidade de processamento é otimizada através de técnicas de agendamento e balanceamento de carga, assegurando que os processadores sejam utilizados de maneira eficiente e sem desperdício. No que tange ao armazenamento, o gerenciador coordena o provisionamento e a replicação de dados, promovendo a resiliência e o acesso rápido às informações. Além disso, a gestão da comunicação entre servidores é feita de forma a maximizar a largura de banda e minimizar a latência, permitindo uma troca de dados ágil e confiável entre a borda e a nuvem.

No que tange o método *Make Way*, o Gerenciador de Recursos pode ser dividido em duas camadas principais: a camada de Gestão de Infraestrutura Virtual (do inglês, *Virtual Infrastructure Manager*) e a camada de Gestão de Recursos de Serviços.

6.1.2.1 Camada de Gestão de Infraestrutura Virtual

A Camada de Gestão de Infraestrutura Virtual desempenha um papel crucial dentro de uma infraestrutura de computação em nuvem e em borda. Esta camada é essencial para a gestão eficiente da capacidade de processamento, armazenamento e comunicação dos servidores, o que é fundamental para garantir que os recursos estejam disponíveis onde e quando necessário [303].

Esta camada é responsável pelo provisionamento, gerenciamento e monitoramento dos recursos computacionais em um ambiente virtualizado. Isso inclui a administração de máquinas virtuais, armazenamento de dados e redes virtuais, essenciais para suportar a execução de aplicações distribuídas e escaláveis. Esta camada assegura que os recursos são alocados de forma eficiente, respondendo dinamicamente à demanda variável das aplicações, o que é particularmente importante em cenários de saúde crítica onde a disponibilidade e a resposta rápida são cruciais. [304]

Em contextos críticos de saúde, a capacidade de gerir dinamicamente a infraestrutura virtual é vital. Por exemplo, a alocação rápida de máquinas virtuais pode ser necessária para processar dados de emergência ou para escalar serviços de acordo com as demandas imediatas do cenário de saúde. Isso permite que os recursos computacionais acompanhem as necessidades em constante mudança das ambulâncias conectadas e de outros dispositivos médicos em uso no campo.

Entre as tecnologias empregadas na camada de gestão de infraestrutura virtual, o OpenStack se destaca. OpenStack é uma plataforma de código aberto que fornece serviços de Infraestrutura como Serviço (IaaS). Ele permite a gestão de larga escala de computação, armazenamento e recursos de rede em um data center, tudo gerenciado através de APIs ou uma interface de usuário baseada em tecnologias web. Sua arquitetura modular permite aos administradores fornecer e gerenciar esses recursos através de uma plataforma centralizada, o que facilita a implementação de uma infraestrutura virtual resiliente e adaptável.

O OpenStack tem sido amplamente utilizado como Gerenciador de Infraestrutura Virtual (VIM) nas implementações de NFV (Network Functions Virtualization) para ambientes de computação em borda e redes 5G. Nogales *et al.* [305] detalham o desenvolvimento de uma plataforma aberta de gerenciamento e orquestração para experimentação NFV em múltiplos sites com o OpenStack, destacando sua capacidade de suportar isolamento em NFVI, o que permite experimentações complexas em larga escala. Simultaneamente, Foresta *et al.* [306] exploram a integração de SDN com o OpenStack, observando que tal combinação pode potencializar o NFV, resultando em uma rede mais ágil e eficiente.

No estudo de Sechkova *et al.* [307], o OpenStack é avaliado em comparação com o OpenVIM, analisando suas capacidades como VIM em ambientes de edge computing. Apesar de não ser formalmente adotado pela ETSI, o OpenStack é amplamente utilizado devido à sua flexibilidade e robustez. Por sua vez, Asquini *et al.* [308] detalham uma implementação de NFV que segue os padrões da ETSI, empregando o OpenStack dentro do arcabouço de orquestração MANO, evidenciando sua eficácia em ambientes de rede móvel virtualizados, especialmente em automação e configuração. Além disso, Quintana-Ramirez *et al.* [309] exploram o uso do OpenStack no desenvolvimento de sistemas para 5G, integrando componentes de 3GPP e NFV da ETSI, ressaltando sua importância para a criação de redes 5G completas e com funcionalidades de ponta-a-ponta.

Basnet *et al.* [310] propuseram um novo modelo de Computação de Alto Desempenho (HPC) distribuído em uma nuvem pública OpenStack sobre infraestrutura SDN (Software-Defined Networking). A integração do OpenStack com o controlador SDN OpenDaylight demonstrou um aumento significativo no desempenho de velocidade, validando o OpenStack como uma solução eficiente para aplicações de alto desempenho.

Por outro lado, Jiang *et al.* [311] desenvolveram uma arquitetura eficiente para implantar o OpenStack no ambiente de supercomputação TianHe, usando uma abordagem de registro distribuído inteligente (IDRD). Este método reduziu consideravelmente o tempo de distribuição de imagens de componentes e melhorou a eficiência de implantação de clusters em larga escala. Com essa implementação, o projeto destacou as capacidades do OpenStack em gerenciar recursos computacionais de forma flexível e escalável, especialmente em cenários que exigem resposta rápida a demandas computacionais intensas.

Esses estudos reforçam o papel fundamental do OpenStack em simplificar e otimizar a gestão de infraestrutura virtualizada, evidenciando sua importância para a evolução das redes modernas e para atender necessidades de eficiência, escalabilidade e flexibilidade. Além disso, a sua capacidade de integração e compatibilidade com diferentes tecnologias e padrões de rede o torna uma escolha primordial para implementações de uma infraestrutura de computação e comunicação virtualizada, refletindo no motivo da escolha do OpenStack para composição do método *Make Way*, proporcionando uma base sólida para a disponibilização de serviços críticos, como aqueles necessários no contexto de saúde.

6.1.2.2 Camada de Gestão de Recursos de Serviços

A camada de Gestão de Recursos de Serviços é encarregada de alocar e manter os recursos computacionais essenciais para o funcionamento eficiente de aplicações que são encapsuladas em contêineres. Isso inclui a distribuição dinâmica de CPU, memória, armazenamento e capacidade de rede, garantindo que as aplicações tenham acesso aos recursos de que precisam para funcionar de maneira otimizada.

O planejamento e utilização eficiente dos recursos computacionais disponíveis recursos é vital, não apenas para garantir o desempenho adequado das aplicações, mas também para garantir a disponibilidade e reduzir custos operacionais. A capacidade de escalar automaticamente os recursos conforme a demanda permite que as aplicações sejam resilientes durante

picos de carga, sem necessidade de intervenção manual. Esta característica é fundamental em ambientes de produção altamente voláteis e dinâmicos, típicos de aplicações nativas da nuvem.

Dentro desta camada, o Kubernetes se destaca ao oferecer um modelo declarativo para a especificação dos recursos necessários, o que simplifica a gestão de carga e a escalabilidade das aplicações. Utilizando-se de controladores inteligentes, ele monitora constantemente o estado dos pods e ajusta os recursos automaticamente para atender às demandas em tempo real, sem interrupção ou degradação do serviço. Além disso, a camada de Gestão de Recursos de Serviços no Kubernetes ajuda a minimizar o desperdício através da alta densidade de utilização de conteineres por máquina, permitindo que múltiplos serviços compartilhem a mesma infraestrutura física ou virtual sem interferência mútua. Este modelo não apenas reduz os custos de hardware, mas também diminui a complexidade operacional, facilitando a gestão de ambientes de TI em grande escala.

A integração de políticas de qualidade de serviço (QoS) e a possibilidade de configurações flexíveis de priorização de recursos complementam esta camada, permitindo que administradores de sistema definam políticas que refletem as necessidades críticas dos negócios, assegurando que aplicações prioritárias tenham garantia de acesso aos recursos em cenários de escassez.

Desse modo, a camada de Gestão de Recursos de Serviços é um pilar crucial para a eficiência e estabilidade de ambientes baseados em microserviços, proporcionando uma fundação sólida para o desenvolvimento e operação de soluções escaláveis e resilientes em ambientes de nuvem, que podem ser expandidos para o desenvolvimento de soluções em computação em borda.

Alguns trabalhos recentes têm explorado estratégias para gestão de recursos integradas ao Kubenetes. Turin *et al.* [312] realizaram a modelagem do consumo de recursos em sistemas de contêineres Kubernetes, focando na otimização de recursos para aplicações nativas da nuvem. Eles desenvolvem um modelo preditivo baseado em Real-Time ABS, aproveitando dados de implantações menores para prever o comportamento em cenários de maior escala. Os resultados do estudo indicam que o modelo proporciona previsões precisas do consumo de recursos em diferentes configurações de cluster, ajudando na administração eficaz dos recursos. Apesar de algumas divergências nas previsões, especialmente em condições de carga

máxima, o modelo prova ser uma ferramenta robusta para a gestão de serviços e otimização de recursos em ambientes Kubernetes.

Uma abordagem arquitetural inovadora para realizar operações de agendamento em ambiente Kubernetes, que coleta métricas de desempenho na camada de aplicação em vez da camada de rede é apresentada por Centofanti *et al.* [313]. Esta abordagem se baseia em medições de latência realizadas internamente no serviço de interesse, em vez de utilizar serviços externos de medição, e toma decisões de agendamento com base na latência efetivamente percebida pelo usuário final, ao invés de considerar a latência entre os nós do cluster.

Lai *et al.* [314] propuseram um algoritmo de agendamento de conteineres com reconhecimento de atraso (DACS) para abordar a questão da heterogeneidade de nós em computação de borda, que considera não apenas os recursos residuais dos nós trabalhadores, mas também os atrasos potenciais causados pela atribuição de pods. O trabalho discute como o Kubernetes pode ser utilizado como uma plataforma para apoiar microsserviços em ambientes de computação de borda, considerando a heterogeneidade de nós para esse tipo de ambiente.

Essas pesquisas destacam o desenvolvimento contínuo e a importância das estratégias de gestão de recursos em ambientes de microserviços, particularmente em Kubernetes. A integração de tecnologias atuais e a adaptação às necessidades específicas dos ambientes de nuvem e de borda ilustram a evolução constante na otimização de recursos computacionais em arquiteturas modernas. Além disso, o ajuste fino de algoritmos de agendamento para atender às demandas dinâmicas e heterogêneas dos sistemas distribuídos garante a eficácia e eficiência operacional, reforçando a importância de abordagens preditivas e adaptativas em gestão de recursos.

Desse modo, para realizar a Gestão de Recursos de Serviços no método *Make Way*, optou-se por utilizar os próprios recursos de gestão existentes na ferramenta de orquestração do Kubernetes, dada a sua abrangência e maturidade em termos de funcionalidades e utilização em ambientes de produção.

6.1.2.3 Integração OpenStack e Kubernetes

A integração do OpenStack e do Kubernetes é um tópico relevante no âmbito da computação em nuvem, que pode ser expandido para os sistemas de computação em borda. O OpenStack, componente essencial na computação em nuvem [315, 316], fornece serviços de infraestru-

tura, enquanto o Kubernetes é uma plataforma de orquestração de contêineres muito popular [317]. O OpenStack pode ser usado para gerenciar máquinas virtuais e armazenamento, ao passo que o Kubernetes se destaca na gestão de aplicações em contêineres.

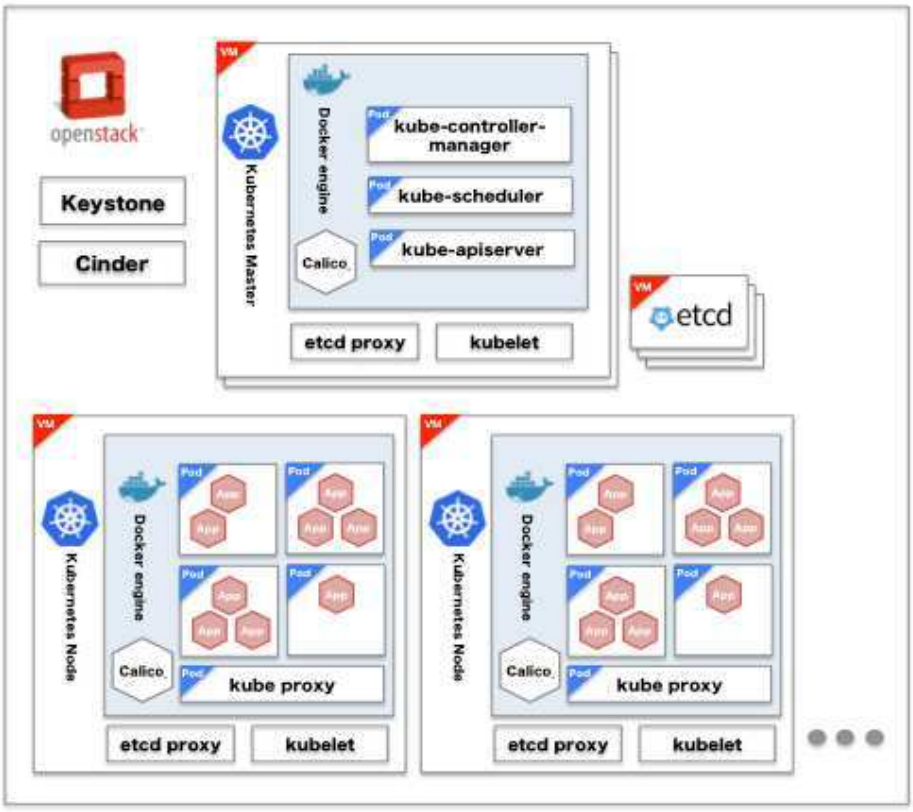
A combinação dessas duas tecnologias permite criar uma infraestrutura de nuvem abrangente que gerencia de forma eficiente tanto máquinas virtuais quanto conteineres. Enquanto o OpenStack foca em serviços de infraestrutura, o Kubernetes cuida da orquestração de conteineres, proporcionando uma plataforma robusta para a implantação e escalabilidade de aplicações em conteineres [318]. Essa integração pode aumentar a eficiência geral, escalabilidade e flexibilidade dos ambientes de nuvem ao aproveitar as facilidades de ambas as ferramentas.

A utilização conjunta dessas ferramentas permite que os recursos de infraestrutura do OpenStack sejam usados para provisionar e gerenciar clusters Kubernetes. Isso inclui a alocação de máquinas virtuais (VMs), redes, armazenamento e outros recursos. Um exemplo de arquitetura do Kubernetes com o OpenStack está ilustrado na Figura 6.2, onde os nós necessários para composição de um cluster Kubernetes são instanciados em máquinas virtuais do OpenStack.

Ao adotar essa abordagem, o problema de gestão de recursos é segmentado em diferentes camadas de abstração. Dessa forma, o escopo de gestão de recursos de cada ferramenta é limitado às funcionalidades específicas para as quais foi designada. Neste contexto, o OpenStack atua como um Gerenciador de Infraestrutura Virtualizada (VIM), gerenciando recursos computacionais, de armazenamento e de rede em uma camada de infraestrutura. Por outro lado, o Kubernetes (K8S) é utilizado para a gestão de serviços em um nível superior, orquestrando a distribuição e a operação de aplicações contêinerizadas, garantindo que os serviços se mantenham resilientes, escaláveis e acessíveis.

Um exemplo de sistema de monitoramento utilizando as ferramentas Openstack e Kubernetes é apresentado em [318]. Neste trabalho, os autores descrevem um sistema chamado Openstack-Reporter, que é composto por três microserviços principais: Openstack-Exporter, Prometheus e Grafana. Esses componentes são gerenciados dentro de um cluster Kubernetes, que simplifica o gerenciamento do ciclo de vida das aplicações através de mecanismos como descoberta de serviços baseada no DNS do Kubernetes e controle de acesso baseado em funções (do inglês, *Role-Based Access Control* - RBAC). Este sistema permite uma con-

Figura 6.2: Exemplo de arquitetura do Kubernetes integrada com o OpenStack.



Fonte: Extraída de [319].

figuração flexível e adaptação às demandas dos usuários, demonstrando sua aplicabilidade através da integração efetiva com uma plataforma OpenStack existente, onde é capaz de coletar e monitorar uma ampla gama de métricas do OpenStack.

Em contrapartida, Kouchaksaraei e Karl [320] exploram a orquestração de funções de serviço entre os domínios OpenStack e Kubernetes. Este estudo não apenas amplia as fronteiras da interoperabilidade entre diferentes plataformas de orquestração de contêineres e NFV, mas também propõe extensões aos padrões de computação em nuvem e NFV para um gerenciamento mais eficiente.

Em resumo, a combinação do OpenStack e do Kubernetes oferece uma solução abrangente para gerenciar recursos em ambientes de computação em nuvem, abordando tanto a infraestrutura quanto os serviços e aplicações. Essa abordagem híbrida permite otimizar o uso de recursos e garantir a escalabilidade e a eficiência operacional.

6.1.3 Módulo de Alocação de Serviços

Este módulo gerencia a alocação dinâmica de serviços a serem disponibilizados para aplicações críticas, baseando-se em variáveis como o padrão de mobilidade das unidades de atendimento móveis e a localização dos servidores de borda. A análise contínua dessas variáveis possibilita que o módulo intervenha de forma antecipada na disponibilização de tais serviços, assegurando que eles estejam disponíveis no local e momento precisos.

Este módulo opera através da coleta e análise de dados relacionados ao padrão de mobilidade, essencial para entender quais serão os próximos servidores responsáveis pela disponibilização dos serviços. Com esses dados, o módulo pode prever o momento e o local onde uma ambulância entrará em uma nova área de cobertura. Esta capacidade preditiva permite preparar os servidores de borda adequados antes da chegada da ambulância, reduzindo latências e otimizando a eficiência operacional.

Algumas abordagens existentes para este módulo é o trabalho desenvolvido por Jojoa et al. [321], onde são priorizadas aplicações críticas com base em uma regra preventiva que beneficia a latência e o tempo de espera para aplicativos críticos. O modelo desenvolvido, MAACO (*Mobility-Aware, Priority-Driven, ACO-based Service Placement Model*), pode ser parametrizado para priorizar requisições de serviços com base em requisitos específicos de QoS.

6.1.4 Estimador do Padrão de Mobilidade

A estimativa do padrão de mobilidade é um dos principais pontos do método *Make Way*, uma vez que a alocação de serviços ao longo de um trajeto apresenta um tempo de *warmup*. Desse modo, é necessário que a alocação dos serviços aconteça antes das requisições do serviço em execução ao futuro nó de borda mais próximo.

A função principal deste módulo é fornecer informações de geolocalização e o padrão estimado da trajetória da ambulância para que a alocação de serviços ocorra de maneira proativa, antecedendo as requisições efetivas do serviço em execução ao nó de borda mais próximo. Isso é vital para a manutenção da continuidade e eficiência dos serviços durante a movimentação dos veículos de emergência.

De modo geral, algoritmos inteligentes capazes de analisar e prever a mobilidade dos

veículos tornam a estimativa do padrão de mobilidade bastante promissora. Esses algoritmos processam conjuntos de dados que incluem informações como velocidades médias, rotas e condições de tráfego atualizadas, permitindo a modelagem precisa das trajetórias veiculares e a previsão dos tempos de chegada nos próximos nós de borda.

A precisão nas previsões realizadas por este módulo é crucial para a programação eficaz da alocação de serviços. Inexatidões podem resultar em atrasos ou na alocação ineficiente de recursos, afetando o desempenho do sistema em situações críticas. Consequentemente, os modelos preditivos são constantemente refinados com base no feedback operacional, visando otimizar sua acurácia e minimizar os impactos provenientes de estimativas imprecisas.

No trabalho apresentado por Wang et al. [322], por exemplo, foi proposta a predição de mobilidade em redes 5G utilizando método baseado em IA para melhorar a precisão e a confiança da previsão de mobilidade de usuários, onde foram obtidas maiores precisões de previsão de trajetória (cerca de 90%) com menor tempo de treinamento.

6.1.5 Módulo de Priorização de Serviços

O Módulo de Priorização de Serviços, integrado ao método *Make Way*, emprega algoritmos de decisão para avaliar e priorizar serviços de acordo com a gravidade das situações e a disponibilidade de recursos. Este módulo é relevante principalmente em cenários onde os recursos de computação são limitados ou estão previamente ocupados por aplicações menos críticas e a demanda por serviços emergenciais é alta.

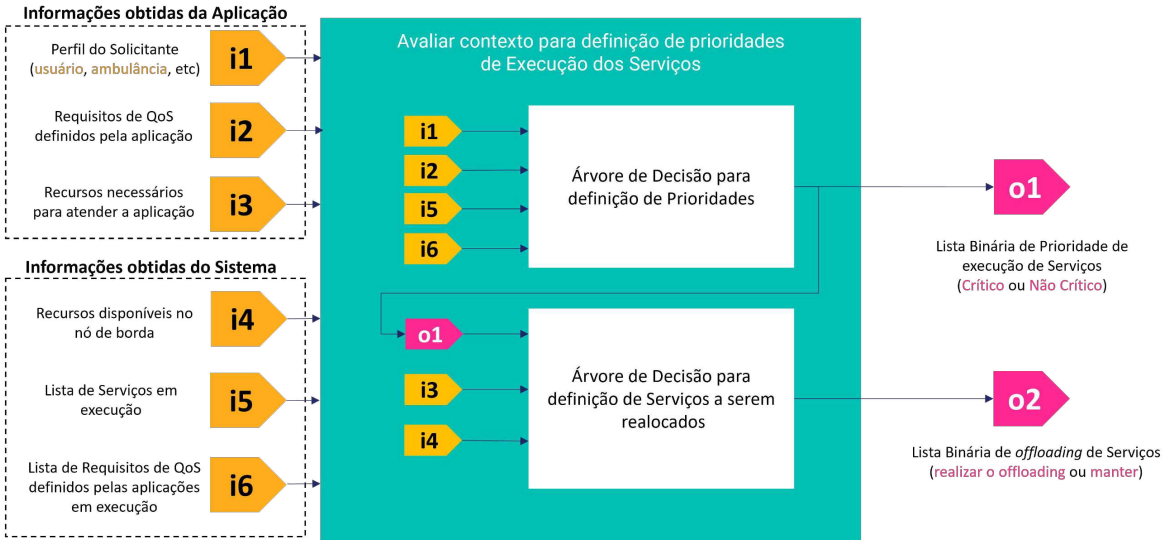
A função principal deste módulo é a análise do contexto em que os serviços serão executados, utilizando um conjunto de indicadores, tais como tipo de perfil do usuário, requisitos de QoS das aplicações, recursos necessários para atender a aplicação, recursos disponíveis no servidor de borda alvo, entre outros. Esses indicadores podem ser processados por algoritmos inteligentes capazes de realizar uma avaliação do contexto e, a partir disso, definir os serviços a serem priorizados. Por exemplo, utilizando uma árvore de decisão, o módulo categoriza os serviços em críticos e não críticos, o que permite uma alocação de recursos mais estratégica e fundamentada.

Além da classificação inicial, o Módulo de Priorização de Serviços também determina quais serviços devem ser mantidos no mesmo nó e quais podem ser transferidos para outros servidores. Esta segunda avaliação é crucial para otimizar o desempenho do ambiente de

computação e garantir a continuidade dos serviços críticos sem interrupções.

Os algoritmos utilizados no módulo são projetados para adaptar-se dinamicamente às mudanças nas condições de rede e nas prioridades dos casos de emergência. Isso inclui a reavaliação periódica das prioridades à medida que novos dados são recebidos, permitindo reajustes proativos nas estratégias de alocação de recursos. Na Figura 6.3 está ilustrado um diagrama do módulo de avaliação de contexto para definição de prioridades de Execução dos Serviços, baseada em Árvores de Decisões.

Figura 6.3: Módulo de avaliação de contexto para definição de prioridades de Execução dos Serviços.



Fonte: Produzida pelo autor.

6.2 Ambiente de Experimentação

A evolução contínua dos cenários de mobilidade exige estudo e desenvolvimento constante de novas soluções tecnológicas para garantir eficiência e segurança. A complexidade desses cenários, que inclui variáveis como movimentação contínua de dispositivos e variações de sinais de rede, torna essencial a simulação detalhada para entender e antecipar comportamentos e problemas potenciais.

Neste sentido, para a Gestão e Implantação de Aplicações e Serviços foi adotada uma abordagem moderna baseada em arquitetura de microsserviços, justificada pela capacidade

de recursos reduzidas em ambientes de Computação em Borda. Neste sentido, os serviços necessários para atender às aplicações de *streaming* de dados, vídeo e áudio foram implementados em forma de contêineres. Lidar com a gestão de contêineres se torna uma tarefa complexa à medida que a aplicação aumenta de escala. Neste sentido, é necessário utilizar ferramentas de orquestração de contêineres. A ferramenta escolhida para esta finalidade foi o Kubernetes, dada a sua abrangência e maturidade em termos de funcionalidades e utilização em ambientes de produção.

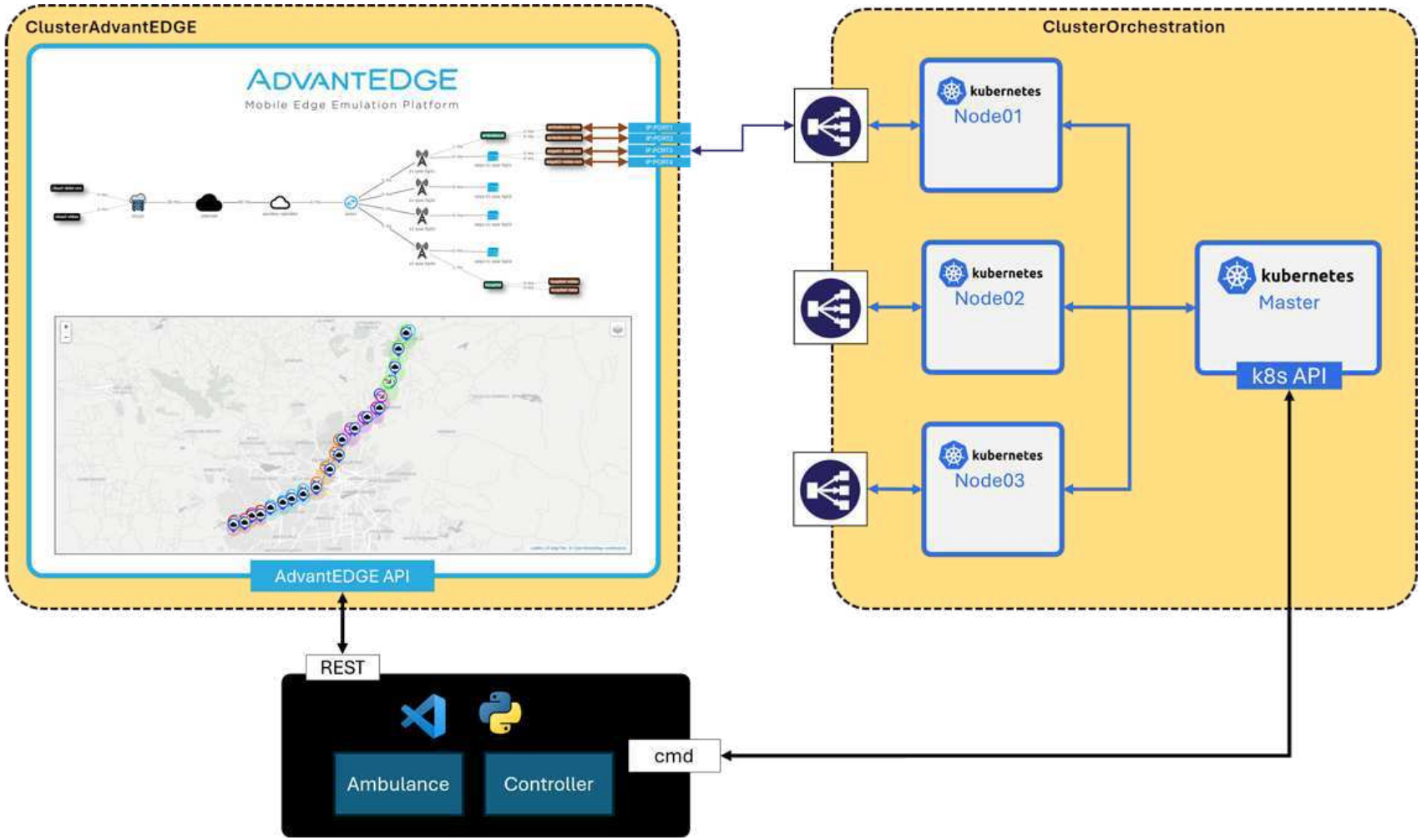
Além disso, para emular o efeito da mobilidade da Ambulância, foi utilizada a plataforma AdvantEDGE, com suas capacidades de configuração de rede, simulação de movimento e visualização de elementos em um ambiente controlado, facilitando a experimentação e o teste de soluções em condições que refletem características do mundo real de maneira precisa e escalável. Desse modo, conforme ilustrado nas Figuras 6.4, 6.5 e 6.6, o ambiente experimental foi montado baseado em ferramentas atuais e avançadas para simulação realista, e pode ser dividido em duas partes principais:

- O ambiente AdvantEDGE foi empregado para simular a mobilidade da ambulância, abrangendo a emulação das características dinâmicas de rede. Isso é essencial para testar a capacidade de comunicação em cenários de movimento rápido sob condições variáveis.

- Por meio do Kubernetes, foram disponibilizados microsserviços especializados, tais como streaming de multimídia e transmissão de dados, visando explorar a eficácia da orquestração de serviços em tempo-real, crucial para operações críticas como as de serviços de emergência.

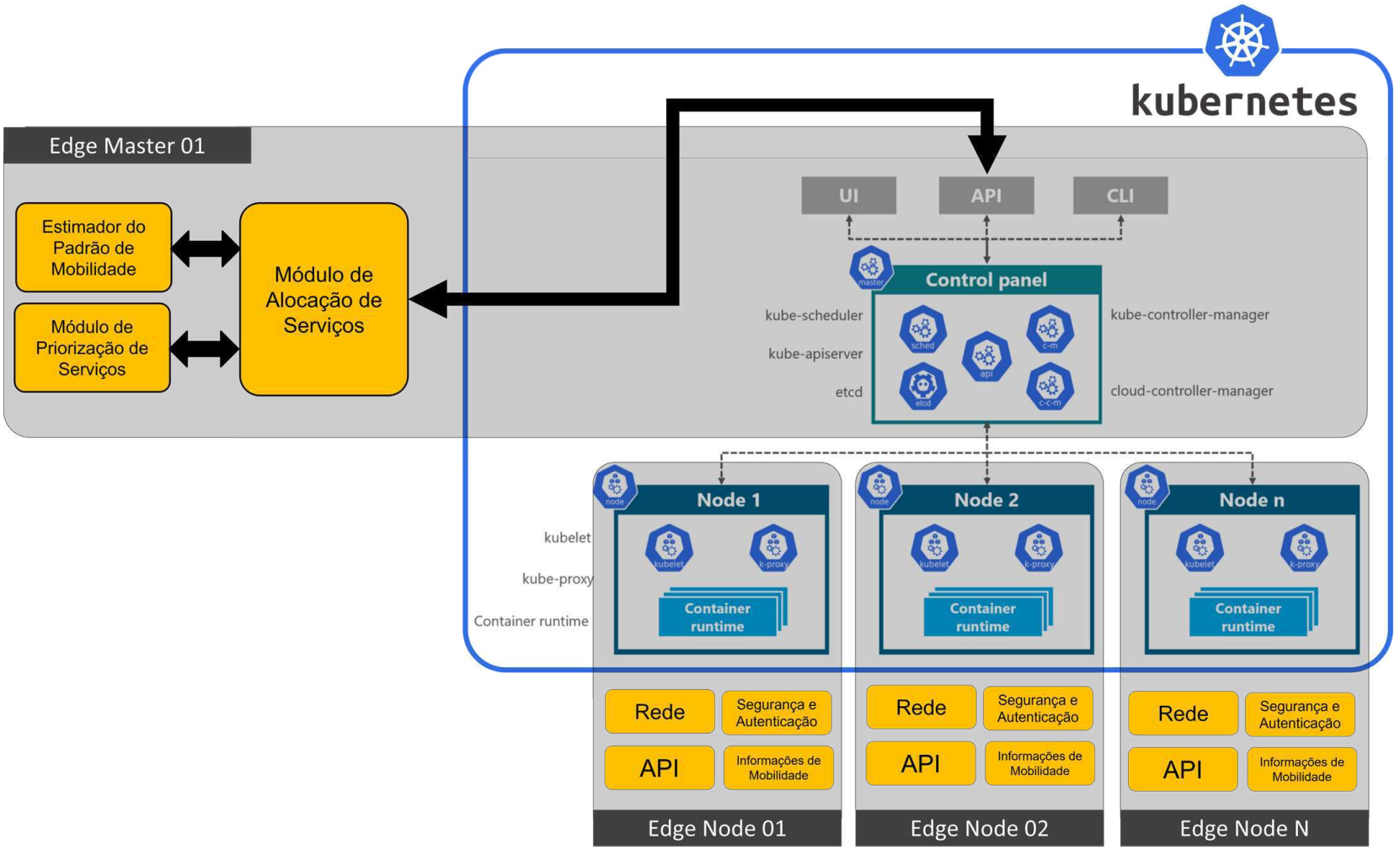
Conforme observa-se na Figura 6.4, o controlador (*Controller*) e a ambulância (*Ambulance*) se comunicam com os ambientes de simulação de mobilidade e disponibilização de serviços através das APIs do AdvantEDGE e do Kubernetes, respectivamente. Isto viabiliza a simulação integrada e automatizada de cenários complexos, onde a coordenação entre o veículo de emergência e o controle central para disponibilização de serviços é crítica, replicando de forma precisa as condições reais de uso e permitindo testes e ajustes em tempo-real nas estratégias de resposta a emergências.

Figura 6.4: Diagrama de alto nível com a representação dos ambientes de Mobilidade e de Gestão e Implantação de Aplicações e Serviços.



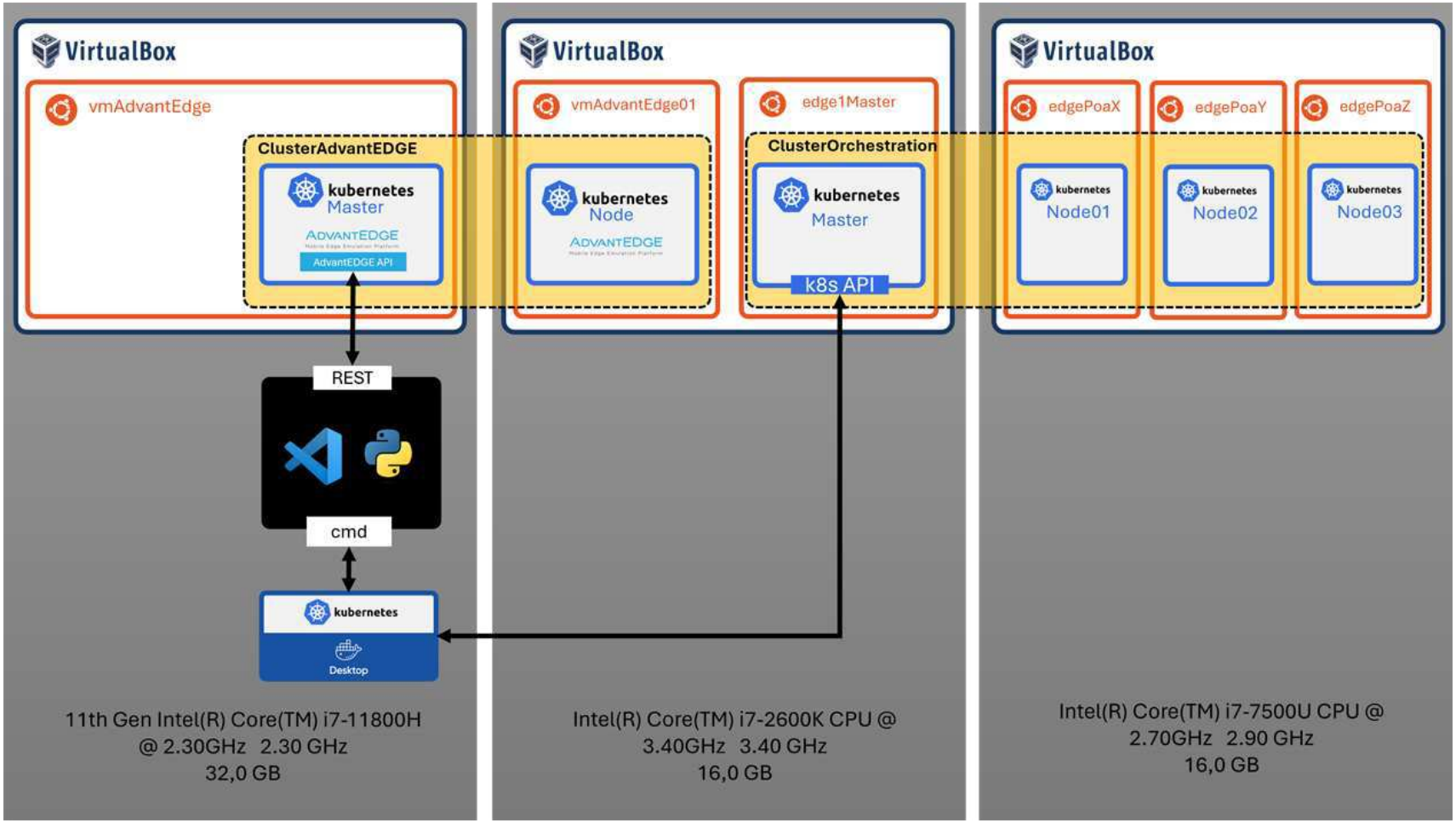
Fonte: Produzida pelo autor.

Figura 6.5: Diagrama do *ClusterOrchestration* com os componentes do framework *Make Way*.



Fonte: Produzida pelo autor.

Figura 6.6: Diagrama do ambiente físico e máquinas virtuais para execução experimental.



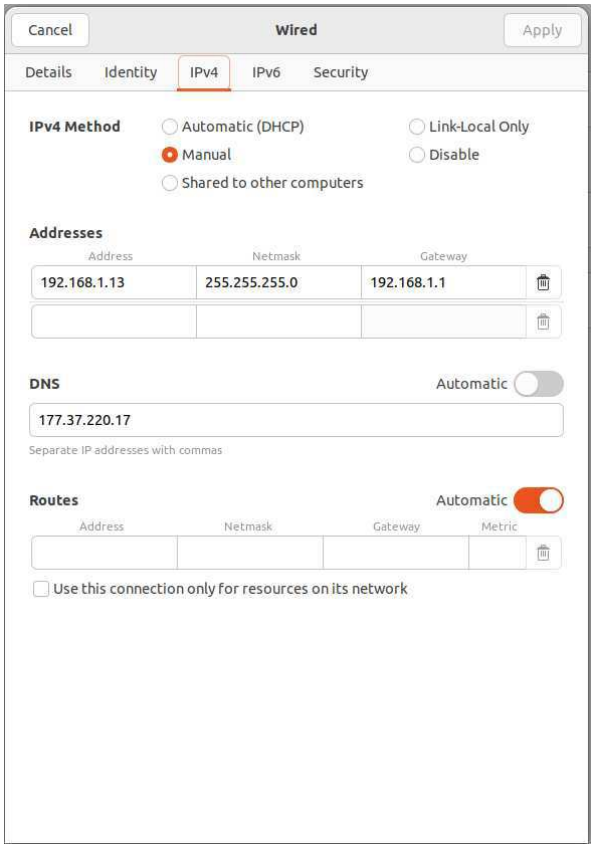
Fonte: Produzida pelo autor.

Na Figura 6.5 é apresentado um diagrama de arquitetura do *ClusterOrchestration* e sua interação com alguns componentes do framework *Make Way*. No topo, no servidor de borda "Edge Master 01"estão contidos o **Estimador do Padrão de Mobilidade**, **Módulo de Alocação de Serviços** e o **Módulo de Priorização de Serviços**, essenciais para a gestão dinâmica e adaptativa de recursos e serviços em ambientes móveis. Na parte central, o painel de controle do Kubernetes é destacado, mostrando componentes como o UI, API, CLI, kube-scheduler, kube-apiserver, etcd, entre outros, que gerenciam o estado e a operação do cluster. Na parte inferior, diferentes camadas de comunicação são destacadas nos servidores de borda "Edge Nodes", cada um equipado com seu próprio *runtime* de containeres, reforçando a escalabilidade e a distribuição de carga característica de ambientes Kubernetes. Além disso, estão inclusos serviços de Rede, API, e Segurança e Autenticação, sublinhando a importância de uma comunicação segura e eficiente entre os nós e o ambiente centralizado de controle.

A montagem do ambiente de simulação foi estruturada utilizando três computadores, conforme detalhado na Figura 6.6. Nestes computadores, foram configuradas máquinas virtuais (VMs) para executar funções específicas. Essas VMs foram divididas em dois principais grupos de funcionalidades: um grupo rodando o ambiente AdvantEDGE para simulação de redes de comunicação móveis e mobilidade, e outro dedicado à orquestração de serviços usando Kubernetes. Esta configuração permitiu a distribuição da carga computacional necessária para execução da simulação.

Cada VM foi configurada para realizar a conexão do adaptador de rede virtual a uma rede física, onde um adaptador de rede físico da máquina hospedeira do VirtualBox está conectado. Essa configuração é denominada de *Bridge Mode*. Ao configurar as VMs neste modo de operação, o servidor DHCP (*Dynamic Host Configuration Protocol*) do roteador, cujo adaptador de rede físico da máquina hospedeira está conectado, define automaticamente o IP de cada VM. Nesse caso, seria necessário atualizar os IPs das VMs nos clusters Kubernetes sempre que as máquinas fossem inicializadas. Para resolver essa questão, foram realizadas as configurações de IP em cada VM, atribuindo-se um valor fixo dentro da faixa definida para o servidor DHCP da rede local, conforme ilustrado na Figura 6.7.

Figura 6.7: Captura de tela da configuração de rede para uma das máquinas virtuais utilizada no ambiente experimental.



Fonte: Produzida pelo autor.

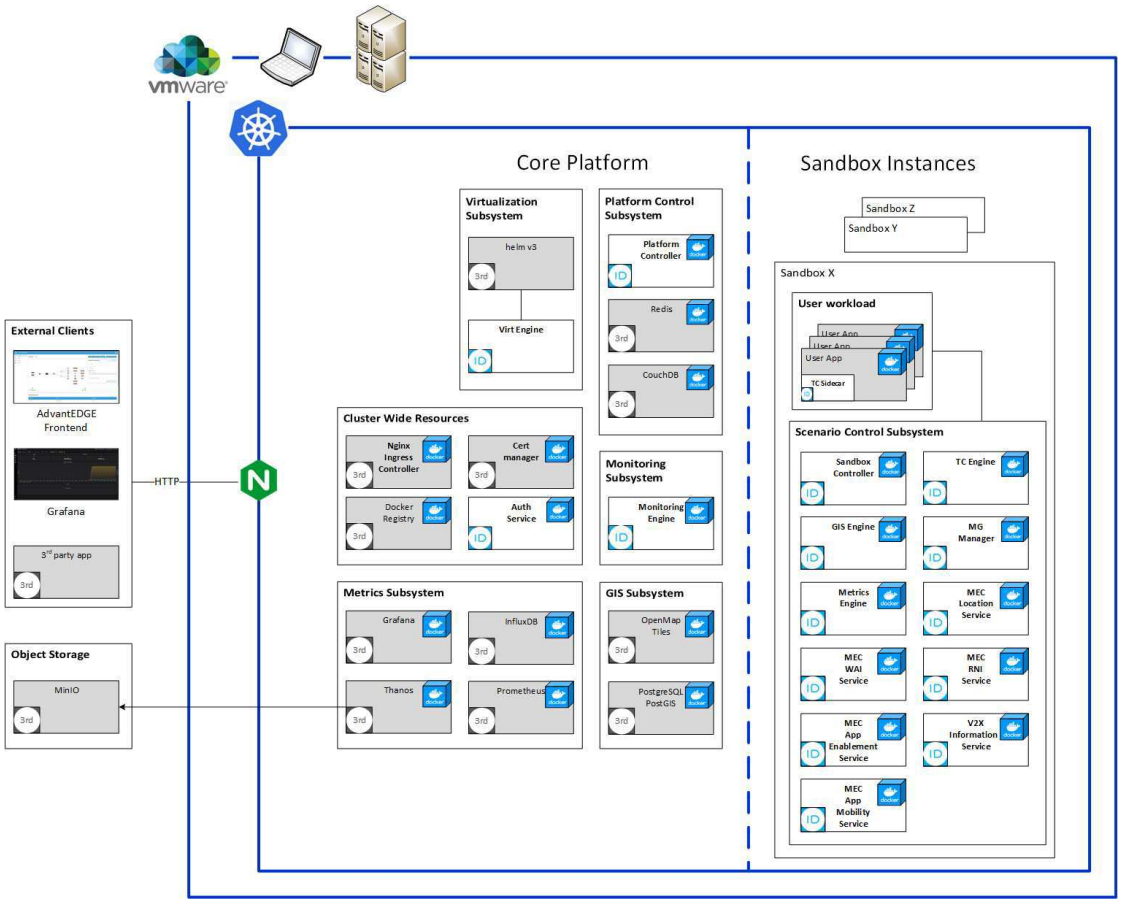
6.2.1 A plataforma AdvantEDGE

O AdvantEDGE é uma Plataforma de Emulação de Edge Móvel (*Mobile Edge Emulation Platform* - MEEP) que funciona com base nas plataformas do Docker e Kubernetes. Esta plataforma proporciona um ambiente de emulação robusto, que permite a experimentação com tecnologias, aplicações e serviços de computação em borda. O AdvantEDGE facilita a exploração de modelos de implantação em borda, permitindo a avaliação do impacto dessas configurações nas aplicações e serviços por meio de ciclos rápidos e ágeis de desenvolvimento [323].

6.2.1.1 Arquitetura de Microsserviços

A arquitetura do AdvantEDGE é baseada em microsserviços, facilitando interações eficientes e permitindo uma integração flexível dos componentes essenciais para simular cenários complexos de rede na borda. De modo geral, o AdvantEDGE é um software controlador especialmente desenvolvido para simplificar a implantação e o gerenciamento de aplicações de borda em ambientes de rede simulados. A Visão geral de alto nível da arquitetura de microsserviços do AdvantEDGE está apresentada na Figura 6.8.

Figura 6.8: Visão geral de alto nível da arquitetura de microsserviços do AdvantEDGE.



Fonte: Extraído de [324].

Cada um dos microsserviços do AdvantEDGE é encapsulado em contêineres Docker e são otimizados para operação dentro de um ambiente Kubernetes. Isso garante escalabilidade, gerenciamento simplificado e uma implantação eficiente, ideal para testar diferentes configurações e comportamentos de rede em tempo-real. Os contêineres são agrupados em quatro categorias principais, cada uma com funções específicas na plataforma:

- ***Core-Platform*:** Contém microsserviços que fornecem funcionalidades essenciais da plataforma do controlador AdvantEDGE, gerenciando a orquestração e a coordenação centralizada das operações.

- ***Core-Sandbox*:** Inclui microsserviços que apoiam a funcionalidade de sandbox do controlador, permitindo simulações isoladas e seguras de cenários específicos sem afetar o ambiente operacional principal.

- ***Dependency*:** Agrupa microsserviços de terceiros necessários para que os serviços centrais funcionem adequadamente, garantindo compatibilidade e extensibilidade.

- ***Scenario*:** Compõe contêineres que implementam casos de uso específicos para aplicações em borda, permitindo que desenvolvedores e pesquisadores testem aplicações e serviços em condições de rede simuladas que replicam ambientes de borda realistas.

Essa estrutura modular e escalável de microsserviços no AdvantEDGE permite que usuários e desenvolvedores explorem eficientemente o potencial das tecnologias de borda, melhorando o desenvolvimento e a implementação de soluções inovadoras na borda da rede [324].

6.2.1.2 Criação de Diagramas de Rede

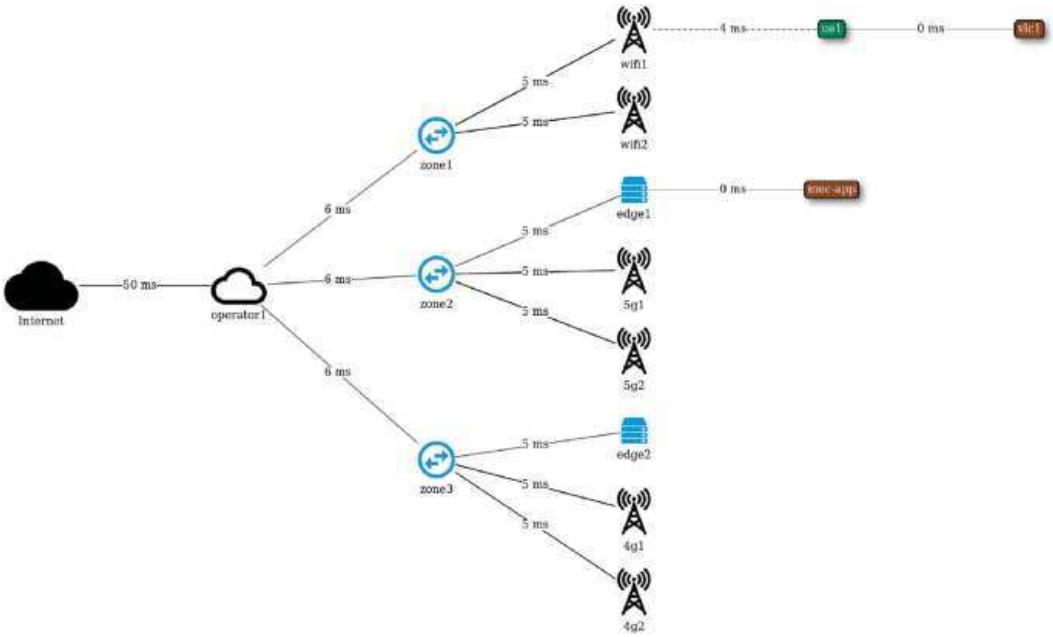
Uma das vantagens da utilização do ambiente AdvantEDGE está na possbilidade de criação de diagramas de rede a partir de uma ferramenta visual, que é baseada na ferramenta gráfica *vis.js*[1]. Esta ferramenta permite interações como arrastar e ampliar a visão do diagrama de rede utilizando controles providos ou o mouse. Na visualização do cenário, informações básicas dos elementos de rede como tipo, nome e características de rede são exibidas diretamente na interface gráfica. Detalhes adicionais podem ser visualizados ao passar o cursor sobre um elemento ou clicando nele para acessar o painel de configuração do elemento de rede. É possível ajustar a posição dos elementos de rede através de cliques e arrasto, facilitando a organização visual conforme necessário.

Esta funcionalidade não só permite uma visualização intuitiva e interativa dos elementos de rede, mas também simplifica a configuração e o ajuste das características de rede dire-

[1] https://visjs.org/

tamente através da interface gráfica. Isso proporciona uma maior agilidade e precisão no desenvolvimento e teste de cenários complexos, tornando-se uma ferramenta essencial para engenheiros e pesquisadores na simulação de ambientes de rede dinâmicos. Na Figura 6.9 está apresentada uma captura de tela do AdvantEDGE para a visão do Diagrama de Rede.

Figura 6.9: Captura de tela do AdvantEDGE para a visão do Diagrama de Rede.



Fonte: Extraída de [325].

6.2.1.3 Modelo de Rede

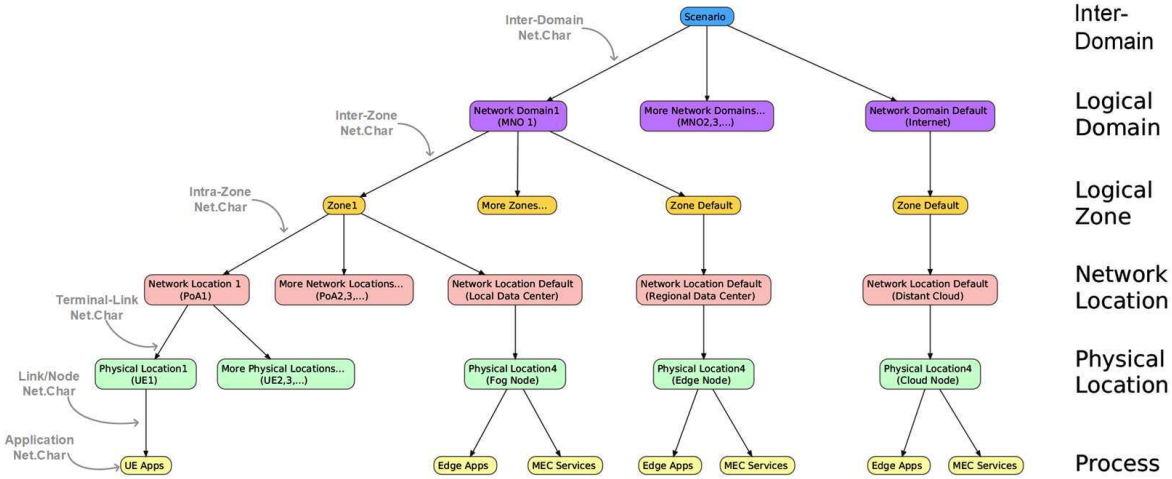
A plataforma AdvantEDGE utiliza um modelo detalhado para definir cenários de rede, permitindo a simulação e avaliação de diversas configurações de rede em uma plataforma estruturada. Este modelo oferece uma representação precisa dos componentes de rede, possibilitando a modelagem de interações complexas entre dispositivos, aplicações e infraestruturas. O uso deste modelo facilita a análise e resolução de problemas, a otimização de desempenho e o planejamento de implantações de rede, proporcionando um ambiente eficaz para o desenvolvimento de soluções de conectividade e computação em borda.

Na Figura 6.10 está ilustrado o diagrama do modelo de rede utilizado pelo AdvantEDGE, detalhando a estrutura hierárquica, desde o cenário até os componentes específicos de localizações de rede e processos, incluindo zonas lógicas, domínios e tipos de localização física.

O **cenário** é o componente de nível mais alto do modelo de rede. Os usuários do

AdvantEDGE criam cenários, e múltiplos cenários podem ser armazenados simultaneamente. Um cenário é implantado por vez, sendo o cenário ativo aquele que está em operação. Este define as características de rede Inter-Domínio para o tráfego que cruza entre domínios.

Figura 6.10: Diagrama do modelo de rede utilizado no AdvantEDGE.



Fonte: Extraída de [324].

O **Domínio Lógico** determina o número e os tipos de domínios dentro de um cenário, sendo a internet ou nuvem distante o domínio padrão. Cada Operadora de Rede Móvel (MNO) representa um domínio, e o Domínio Lógico define as características de rede Inter-Zona para o tráfego que cruza entre zonas.

A **Zona Lógica** permite que um domínio seja decomposto em diferentes zonas, agrupando múltiplas Localizações de Rede. Ela define as características de rede Intra-Zona para o tráfego que atravessa estas localizações de rede.

A **Localização de Rede** especifica os locais dentro de uma zona onde os nós se conectam à rede, também referida como Ponto de Acesso (PoA). O PoA representa o ponto de conexão de rede de uma localização física, como nós de borda, nevoeiro, nuvem ou dispositivos de usuário final (UE), e define as características de rede do link terminal.

A **Localização Física** define o local físico de um dispositivo, onde cada dispositivo no sistema ocupa sua própria localização física única, determinando o tipo de nó que ocupa a localização física. Tipos de nó incluem borda, nevoeiro, nuvem e UE. Todos os tipos de nó podem ser internos ou externos à plataforma (por exemplo, um telefone móvel físico ou um

nó de nevoeiro físico podem ser interconectados com o cenário da plataforma). Dispositivos de UE, nós de borda e nevoeiro podem mudar dinamicamente de PoA durante a execução do cenário por meio do envio de um Evento de Mobilidade. A localização física define as características de rede do nó, que representam o impacto do nó no tráfego, como limitações de latência/vazão de um nó de borda sobrecarregado.

O **Processo** representa a "folha" da árvore do modelo, onde uma aplicação está sendo executada em uma Localização Física específica. Cada processo é realizado pela implantação de um Pod (que possui conteineres específicos) no Kubernetes. Os processos podem ser impactados pelas características de rede definidas. Além disso, dispositivos de UE externos têm seus processos executando fora da plataforma AdvantEDGE. O processo define as características de rede da aplicação, que representam o impacto da aplicação no tráfego, como simular latência extra de um acesso lento ao disco ou banco de dados sobrecarregado.

6.2.1.4 Características de Rede

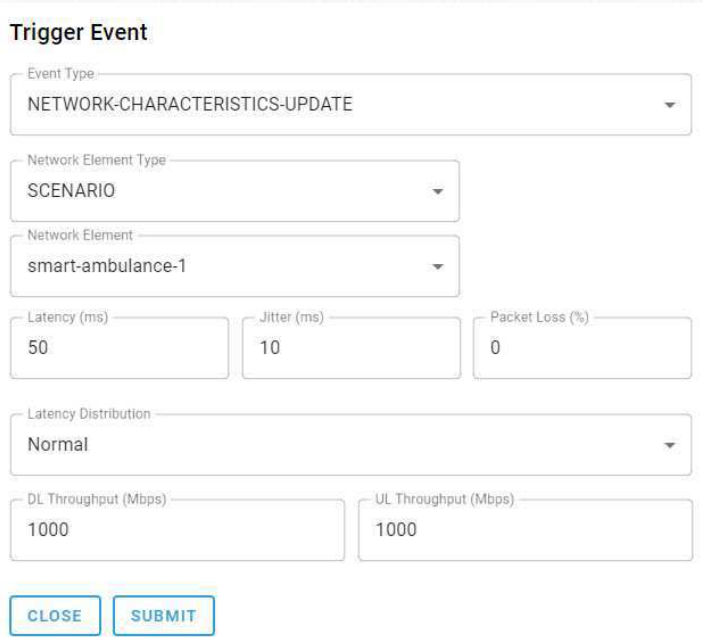
Na etapa de implantação do cenário de simulação, o AdvantEDGE insere um conteiner complementar (um *sidecar* denominado *Traffic Control Engine*, que se baseia na ferramenta Netem[2]) em cada Pod implantado. O *sidecar* tem a função de aplicar as características de rede pre-configuradas, assim como manter e atualizar constantemente as características de rede usando múltiplas entradas, que incluem os valores do cenário, as atualizações do cenário (eventos de mobilidade e características de rede), largura de banda atual utilizada por cada Pod, entre outras, fornecendo a cada *sidecar* do cenário valores a serem aplicados.

As características de rede podem ser aplicadas em todos os níveis detalhados no Modelo de Rede. O AdvantEDGE permite a configuração das seguintes características de rede: latência, jitter, vazão e perda de pacotes. Estas configurações podem ser realizadas diretamente na interface gráfica do ambiente de execução, conforme ilustrado na Figura 6.11, ou via API.

Sobre a emulação das características de rede, o modelo se baseia na comunicação de ponta-a-ponta entre processos, utilizando a hierarquia de rede definida pelo cenário. As características de rede resultantes aplicadas entre dois processos são então o resultado de cálculos baseados na topologia do cenário. Internamente, cada caminho possível entre processos é subdividido em segmentos nos quais as características de rede são aplicadas.

[2] https://wiki.linuxfoundation.org/networking/netem

Figura 6.11: Captura de tela do AdvantEDGE para a Visão de Envio de Eventos de Características de Rede.



Fonte: Produzida pelo autor.

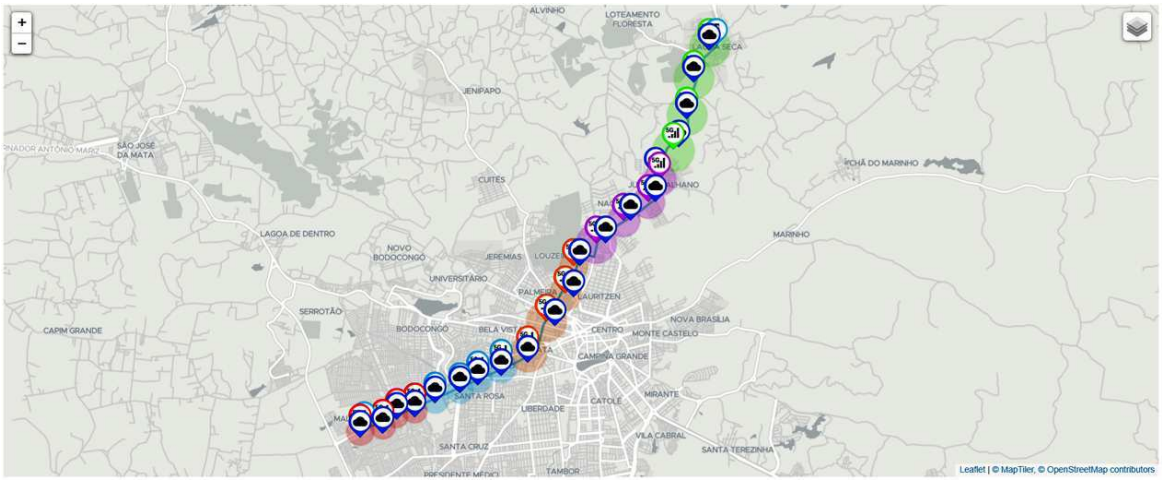
6.2.1.5 Definição e Visualização de Mapas

No ambiente AdvantEDGE, também é possível integrar funcionalidades avançadas de mapeamento com a gestão de cenários de rede, através do uso do *Leaflet*[3] para renderização e do *Leaflet-Geoman*[4] para edição interativa. Esta integração não apenas melhora a precisão na simulação de redes e dispositivos em localizações geográficas específicas, mas também facilita a visualização e manipulação dos elementos de rede diretamente no mapa, proporcionando uma ferramenta poderosa para testes e desenvolvimento em ambientes controlados e realistas. Na Figura 6.12 está apresentada uma captura de tela do AdvantEDGE para a visão do Mapa.

A simulação do cenário de ambulâncias conectadas no AdvantEDGE, com ênfase na definição e visualização de mapas e coordenadas geográficas, oferece benefícios significativos. A capacidade de mapear precisamente as rotas e localizações das ambulâncias em tempo real permite uma resposta mais rápida e eficiente em situações de emergência. Além disso, ao

[3] https://leafletjs.com/
[4] https://geoman.io/

Figura 6.12: Captura de tela do AdvantEDGE para a Visão do Mapa.



Fonte: Produzida pelo autor.

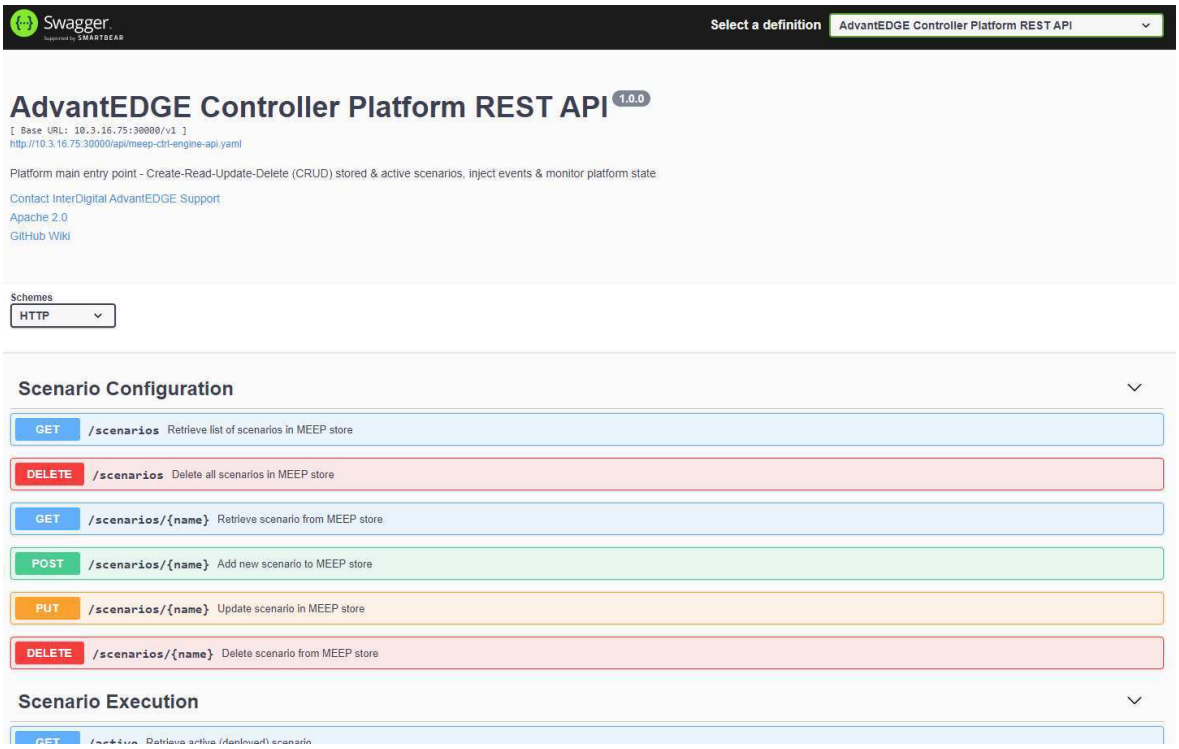
simular condições de rede em diferentes áreas geográficas, é possível antecipar e resolver problemas de conectividade antes que eles afetem a entrega de serviços críticos de saúde, otimizando assim a coordenação e a comunicação entre equipes de emergência.

6.2.1.6 Interface de Programação de Aplicações

O *backend* da plataforma AdvantEDGE oferece uma série de APIs REST, que permitem aos usuários interagir com a plataforma a partir de ambientes de navegador e de softwares externos. Estas APIs são usadas para atividades como acionar eventos conforme cenários específicos, criar sessões PDU em cenários celulares, e ler métricas para análise e experimentação. As APIs seguem a OpenAPI Specification (OAS), com algumas baseadas na versão 2.0 e outras na versão 3.0, facilitando a integração graças a um amplo ecossistema de ferramentas compatíveis [326]. Na Figura 6.13 está ilustrada a interface Swagger-UI da plataforma AdvantEDGE.

Particularmente, as APIs REST do backend da plataforma AdvantEDGE são extremamente vantajosas para facilitar a integração de outras ferramentas durante as simulações para o cenário de ambulâncias conectadas. Através dessas APIs, é possível modelar e simular com precisão o comportamento e a interação das ambulâncias em ambientes urbanos complexos. Por exemplo, as APIs permitem a interação de forma automatizada com o ambiente, o que é essencial para realizar a alocação dinâmica dos serviços críticos de saúde, mesmo

Figura 6.13: Captura de tela da Interface Swagger-UI da plataforma AdvantEDGE.



Fonte: Produzida pelo autor.

com a ambulância em movimento. Além disso, a capacidade de acionar eventos conforme cenários específicos possibilita a simulação de situações de emergência variadas, preparando os sistemas para responder de maneira eficaz em casos reais.

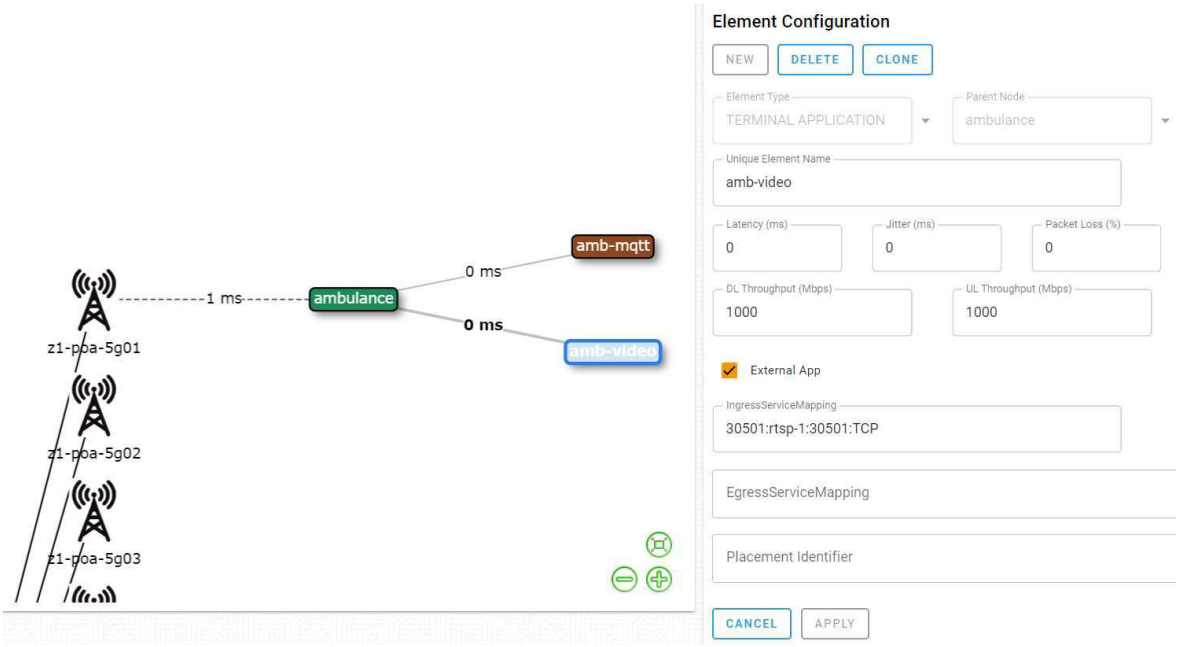
6.2.1.7 Suporte a nós externos

A plataforma AdvantEDGE permite a experimentação com aplicações e serviços executados em nós externos à plataforma. Essa funcionalidade abrange a integração de equipamentos de usuário (UE) externos e nós de computação externa (como *Fog*, *Edge* e *Cloud*), aplicando características de rede específicas aos fluxos de entrada e saída desses dispositivos. Além disso, eventos de cenário podem alterar essas características de rede, impactando a comunicação com os dispositivos externos. Esta capacidade amplia significativamente as possibilidades experimentais da plataforma, permitindo uma interação mais complexa e realista com ambientes de rede diversificados.

A funcionalidade de suporte a nós externos do AdvantEDGE é essencial para disponi-

bilização de serviços externos a partir do ambiente de orquestração de microsserviços. Utilizando a orquestração de microsserviços, é possível gerenciar dinamicamente os serviços críticos de saúde e responder a eventos do cenário em tempo-real, aumentando a eficácia e a resposta do sistema em situações críticas.

Figura 6.14: Captura de tela do AdvantEDGE para a Visão de Configuração de Elementos.



Fonte: Produzida pelo autor.

Na Figura 6.14 está ilustrada a Visão de Configuração de Elementos do AdvantEDGE. Neste exemplo, configura-se o elemento denominado "amb-video", do tipo TERMINAL APPLICATION e que faz o papel de fonte do streaming de vídeo na ambulância, para que seja possível realizar a transferência de vídeos a partir de aplicações externas ao AdvantEDGE. Destaca-se que a utilização dessa interface é de extrema importância para a disponibilização de serviços através do Kubernetes, conforme será descrito nas próximas subseções. Em resumo, através dessa interface é possível disponibilizar serviços no ambiente de computação externo ao ambiente de simulação de mobilidade.

6.3 Considerações Finais

A relevância de utilizar ferramentas como Docker e Kubernetes em ambientes experimentais não pode ser subestimada, já que essas tecnologias são fundamentais em muitas aplicações reais. Ao incorporar tais ferramentas no ambiente experimental, a pesquisa se aproxima das condições de aplicação real, permitindo não apenas uma maior fidelidade nos testes, mas também uma transição mais suave das soluções desenvolvidas para o uso prático, garantindo que os resultados sejam diretamente aplicáveis em cenários operacionais reais.

Além disso, os ambientes de simulação de redes de comunicação móveis e mobilidade são fundamentais para o desenvolvimento e aprimoramento de serviços críticos de saúde, incluindo os cenários que envolvem ambulâncias conectadas. Essas simulações permitem aos pesquisadores e engenheiros testar e validar tecnologias de comunicação em cenários diversos e dinâmicos sem os riscos associados aos testes em ambientes reais. Por exemplo, podem-se explorar as implicações de diferentes configurações de rede e movimentações de veículos em situações críticas, como áreas de cobertura variável e condições de tráfego intenso. Isso é de grande valia para a disponibilização de serviços de telemetria de dados médicos e sistemas de comunicação de emergência de modo confiável e eficaz, proporcionando decisões rápidas e seguras durante o transporte de pacientes. Em última análise, isso não apenas melhora os resultados de saúde dos pacientes, mas também otimiza a utilização de recursos e a coordenação de cuidados em situações de emergência.

Por fim, a escolha adequada de ferramentas e técnicas para a coleta de dados em simulações de parâmetros de rede e eficiência computacional é essencial para garantir a precisão e a confiabilidade dos resultados experimentais. Isso permite que pesquisadores e engenheiros analisem e projetem sistemas considerando as variações de redes móveis com fidelidade e avaliem corretamente o desempenho computacional. Esses aspectos são fundamentais principalmente em áreas que demandam alta precisão, como a disponibilização de serviços críticos de saúde.

Capítulo 7

Simulações e Resultados

No Capítulo 5 foi apresentado o método proposto para a disponibilização dinâmica de recursos e serviços críticos de saúde em cenários de mobilidade, onde são apresentados o detalhamento do método e a definição do protocolo de alocação dinâmica.

Além disso, no Capítulo 6 foi apresentada uma Prova de Conceito para Validação do método, incluindo o ambiente experimental utilizado. De modo geral, o ambiente pode ser dividido em duas partes principais: *(i)* simulação de mobilidade de dispositivos móveis, contemplando a inclusão de mapas e obtenção de dados georreferenciados e a conexão com redes de comunicação móveis de quinta geração (5G), e *(ii)* disponibilização de serviços através de um orquestrador de conteineres. Para isso, foram utilizados os ambientes AdvantEDGE e Kubernetes, respectivamente.

Neste capítulo, estão descritos os cenários utilizados para validação da Prova de Conceito definida no Capítulo 6. Inicialmente, estão apresentados dois cenários com a verificação experimental do problema, que inclui a disponibilização de serviços críticos por meio de ambientes de computação em nuvem e a disponibilização dos mesmos serviços em ambientes de computação em borda. Adicionalmente, estão apresentados três cenários com a orquestração dinâmica de serviços em ambientes de computação em borda para diferentes estratégias, que incluem o conhecimento da trajetória e a estimativa do padrão de mobilidade da ambulância.

Os resultados das simulações com os cenários incluem os dados de recursos computacionais e de rede. De modo geral, esses dados viabilizam a análise da contribuição do método proposto, onde é possível estabelecer um comparativo em termos de redução no uso de recursos computacionais e atendimento aos requisitos de QoS para casos críticos de saúde.

7.1 Validação Experimental do Problema

Visando abordar o problema de forma sequencial e validar o entendimento de que, ao conhecer-se o padrão de mobilidade da Ambulância Conectada, é possível observar ganhos significativos associados à utilização de infraestrutura de computação em borda. Esta abordagem não apenas atende aos restritos requisitos de comunicação para o cenário crítico de saúde explorado, mas também potencializa a eficiência no tratamento e resposta em situações emergenciais. A computação em borda permite que dados críticos sejam processados mais rapidamente, reduzindo a latência e melhorando a tomada de decisões em tempo real, o que é essencial em um contexto onde cada segundo pode salvar vidas.

Contudo, a validação experimental do problema é uma etapa fundamental que serve de referência comparativa para indicação dos ganhos obtidos com a implementação da solução proposta. Este processo é crucial para demonstrar não apenas a viabilidade técnica, mas também os benefícios práticos da infraestrutura de computação em borda em cenários críticos de saúde. Neste sentido, foram explorados dois cenários:

- **Cenário 1: Alocação Estática de Serviços de Streaming de Vídeo e Dados em servidores em Nuvem.** O principal desafio deste cenário está relacionado à latência da rede, que pode provocar atrasos significativos na entrega de dados, impactando negativamente a experiência do usuário final. Esta questão é particularmente crítica em aplicações que dependem de interações em tempo-real, como, por exemplo, aquelas que envolvem streaming de vídeo de alta definição, onde a continuidade e a agilidade na disponibilização e tempo de resposta dos serviços são cruciais.

- **Cenário 2: Alocação Estática de Serviços de Streaming de Vídeo e Dados nos servidores na Borda contemplados ao longo do trajeto.** Este cenário emerge como uma solução alternativa para redução da latência, ao disponibilizar serviços através dos servidores mais próximos fisicamente dos pontos de uso, permitindo um processamento mais rápido e uma resposta com menor tempo. Contudo, dois desafios significativos derivam neste cenário: *(i)* a incerteza quanto ao trajeto exato da ambulância, que pode variar devido a condições imprevistas como tráfego ou emergências adicionais, tornando a alocação antecipada dos serviços potencialmente problemática, e *(ii)* a limitação nos recursos dos servidores de borda, que não possuem a mesma capacidade

de processamento ou armazenamento que os servidores centrais em nuvem.

Para estes cenários, considera-se como **estática** a alocação realizada no início da simulação, sendo os serviços disponibilizados em servidores específicos previamente definidos. Esta abordagem implica que, uma vez estabelecida, a disponibilização dos serviços não se adapta dinamicamente às mudanças de demanda ou condições de rede que possam surgir ao longo do tempo, o que pode afetar a eficiência e a qualidade da entrega de serviços durante a simulação.

7.1.1 Cenário 1: Alocação estática de Serviços de Streaming de Vídeo em servidores em Nuvem

O principal objetivo desta simulação foi examinar e validar um problema específico associado aos prazos de entrega das requisições em aplicações que lidam com Streaming de Vídeo e Dados. Essas aplicações são particularmente relevantes no contexto de serviços que utilizam Realidade Aumentada (AR) e Inteligência Artificial (IA), especialmente no cenário inovador de ambulâncias conectadas. A investigação centrou-se na hipótese de que os prazos estabelecidos para as requisições não são cumpridos de maneira eficaz quando os serviços são hospedados em Servidores em Nuvem. Este problema é atribuído principalmente à latência da rede, que afeta a a eficiência com que os dados são transmitidos e recebidos, comprometendo assim a qualidade dos serviços, principalmente em cenários críticos.

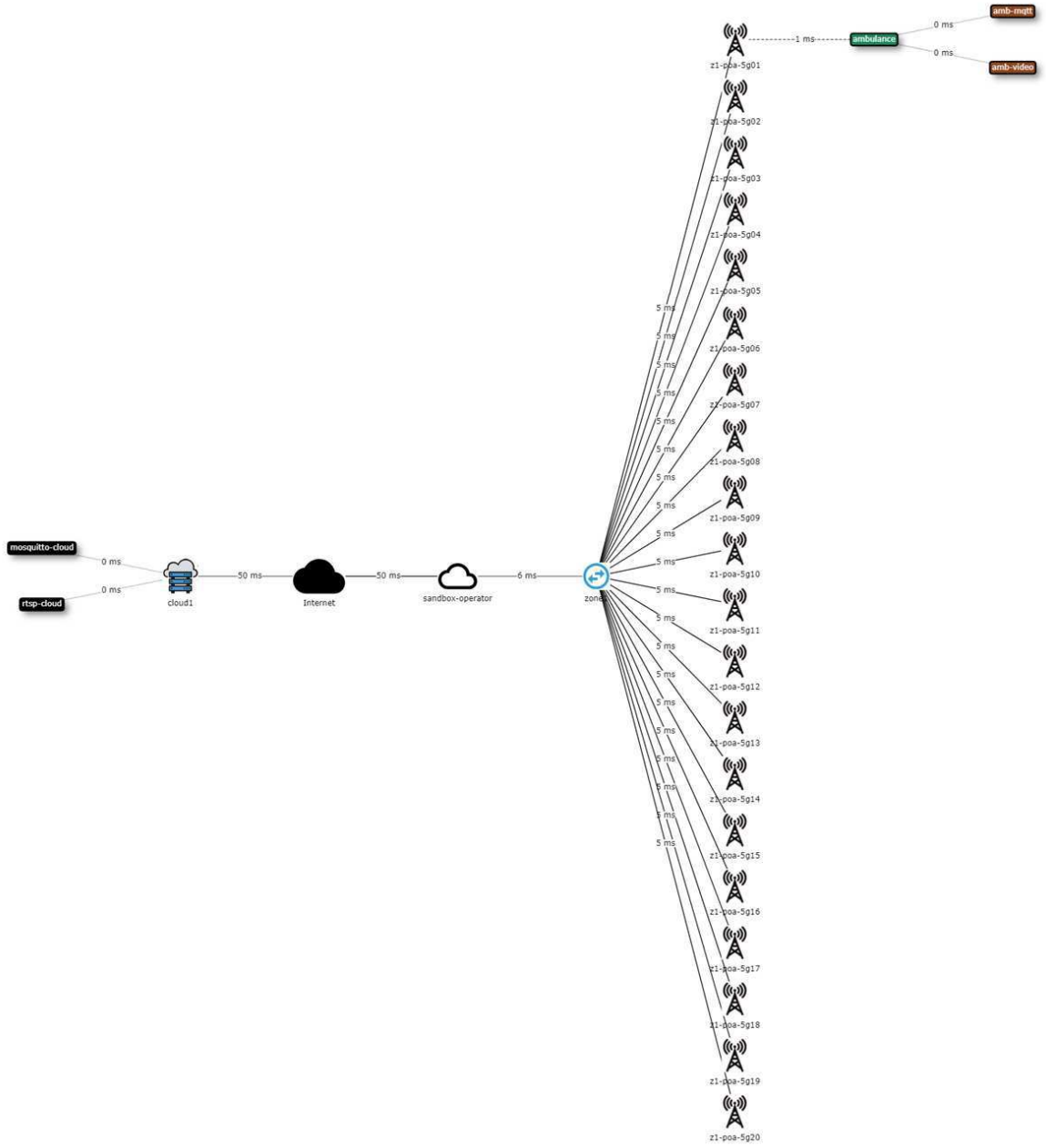
A distância física entre os usuários finais e os servidores centralizados em nuvem e a quantidade de dispositivos realizando um alto número de requisições, principalmente no contexto de IoT, pode agravar este problema. Tal cenário exige uma avaliação cuidadosa sobre as configurações de rede e as estratégias de alocação de serviços, enfatizando a necessidade de otimizar a infraestrutura de rede para minimizar o impacto da latência. Assim, embora a alocação estática em nuvem ofereça benefícios como escalabilidade e capacidade de gerenciamento centralizado, ela também apresenta limitações que podem ser cruciais dependendo das necessidades específicas da aplicação em questão.

Neste sentido, foi realizada a montagem do cenário no ambiente do AdvantEDGE. O detalhamento das configurações e simulações realizadas estão definidos nas próximas seções.

7.1.1.1 Configurações de Rede

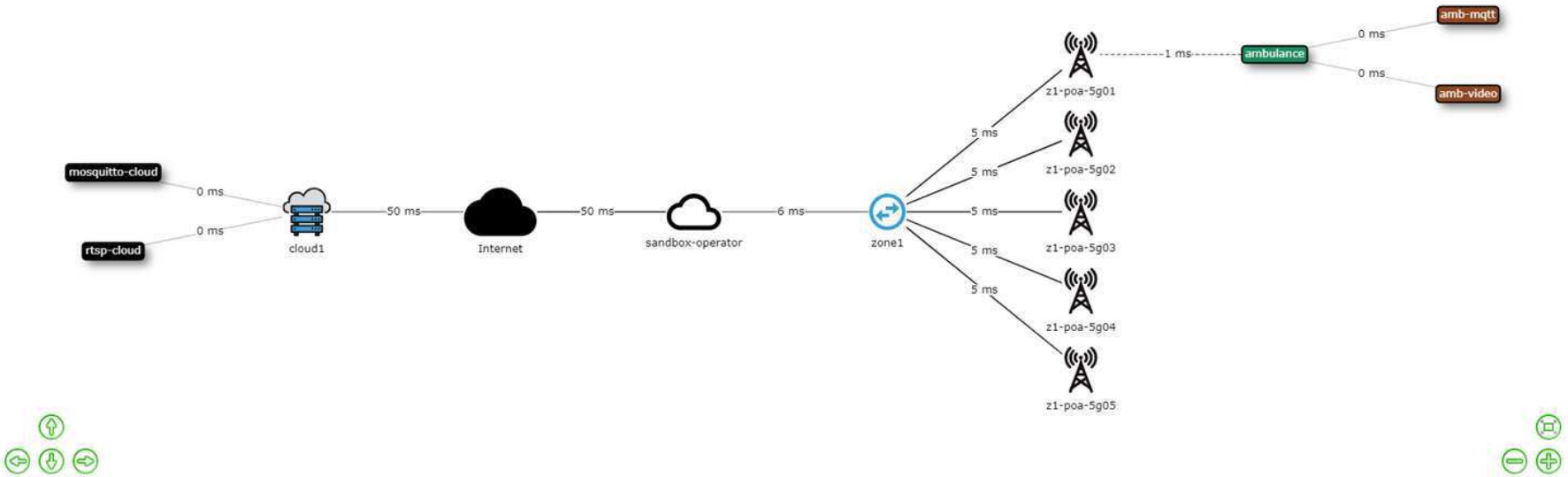
Inicialmente, foi definido o diagrama de rede para simulação do cenário proposto. Na Figura 7.1, está ilustrado o diagrama de rede para o cenário proposto. Para facilitar o entendimento, uma visão mais simplificada do diagrama de rede apresentado para este cenário está ilustrado na Figura 7.2, onde as configurações e descrição dos elementos serão apresentados a seguir.

Figura 7.1: Diagrama de rede para o Cenário 1.



Fonte: Produzida pelo autor.

Figura 7.2: Diagrama de rede simplificado para o Cenário 1

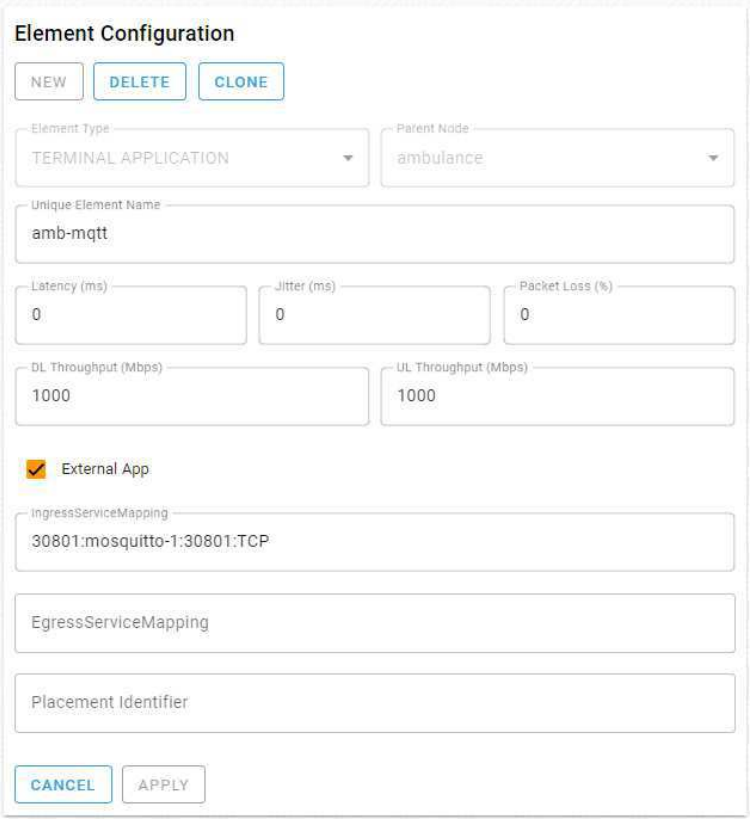


Fonte: Produzida pelo autor.

Conforme detalhado no Capítulo 6, o envio e recebimento de dados externos ao ambiente de AdvantEDGE é realizado pelas interfaces do tipo *TERMINAL APPLICATION*. Neste cenário estão representados dois elementos que utilizam esse tipo de interface **amb-mqtt** e **amb-video**, conforme configurações apresentadas nas Figuras 7.3 e 7.4, respectivamente. De modo geral, através dessas interfaces é possível emular, por exemplo, o envio e recebimento de dados de equipamentos, dispositivos médicos e óculos inteligentes no interior da ambulância, representada pelo elemento **ambulance** do tipo *TERMINAL*.

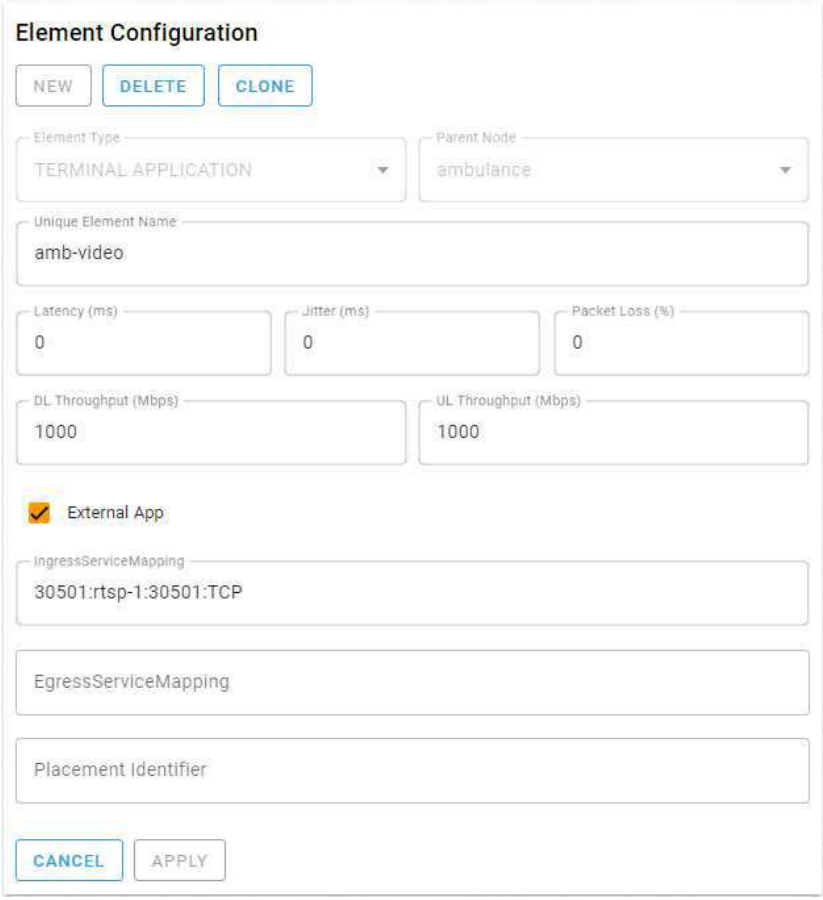
Para execução da plataforma de simulação de mobilidade, foi utilizada uma máquina virtual rodando na rede local com IP **192.168.1.13**. As requisições são direcionadas para o endereço de IP configurado para o ambiente do AdvantEDGE e direcionado para cada serviço, exclusivamente, a partir da porta definida na configuração de cada elemento. Conforme apresentado nas Figuras 7.3 e 7.4, as portas configuradas os elementos **amb-mqtt** e **amb-video** foram **30801** e **30501**, respectivamente.

Figura 7.3: Captura de tela do AdvantEDGE para a Visão de Configuração de Elementos do **amb-mqtt**.



Fonte: Produzida pelo autor.

Figura 7.4: Captura de tela do AdvantEDGE para a Visão de Configuração de Elementos do **amb-rtsp**.



Fonte: Produzida pelo autor.

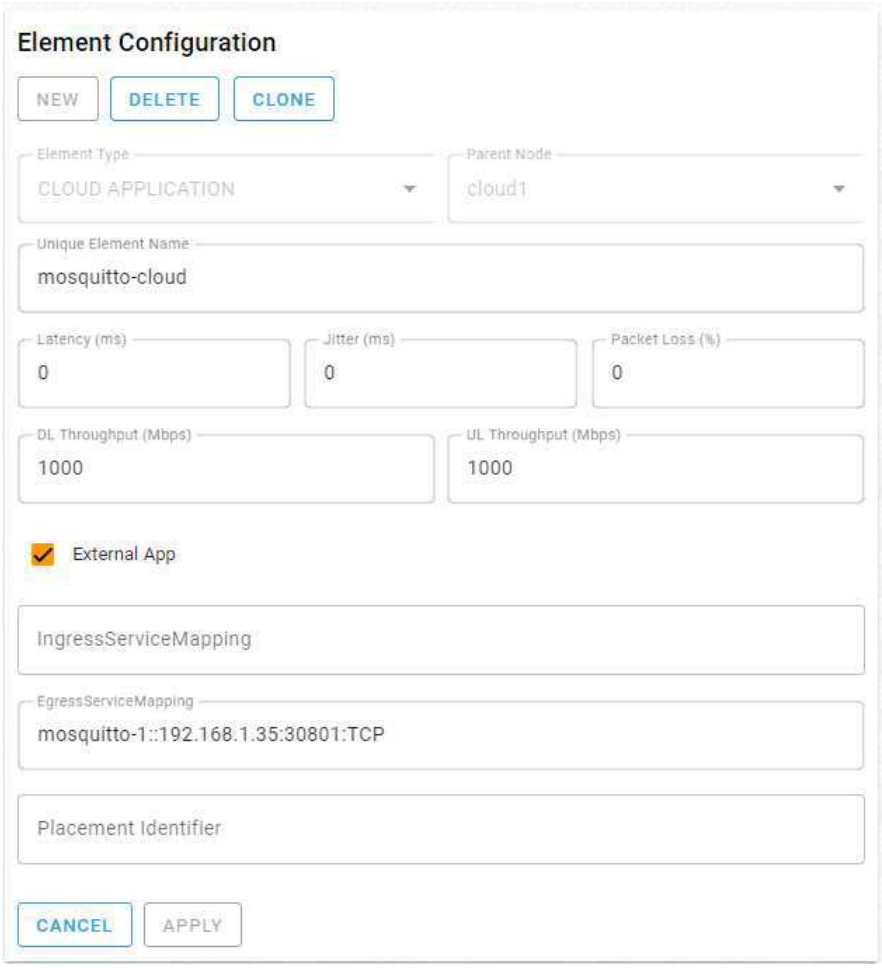
O acesso aos serviços pela ambulância, representada pela tag **ambulance**, é realizado através da conexão à infraestrutura de comunicação móvel do ambiente de simulação representadas pelas antenas. Na visão simplificada do diagrama de rede apresentada na Figura 7.2, os pontos de acesso possíveis são os elementos **z1-poa-5g01**, **z1-poa-5g02**, **z1-poa-5g03**, **z1-poa-5g04** e **z1-poa-5g05**.

Com relação às latências observadas no diagrama de rede, os valores utilizados são previamente definidos pela ferramenta, com base no tipo de elemento utilizado. Porém, é possível realizar a configuração conforme necessidade e ambiente de avaliação. Para o cenário em questão, foram utilizadas as configurações padrão, uma vez que, principalmente para o ambiente de computação em nuvem, o valor ficou condizente tanto com o que foi apresentado na

literatura, como os valores obtidos experimentalmente, conforme apresentado no Capítulo 4, sendo em torno de 100 ms.

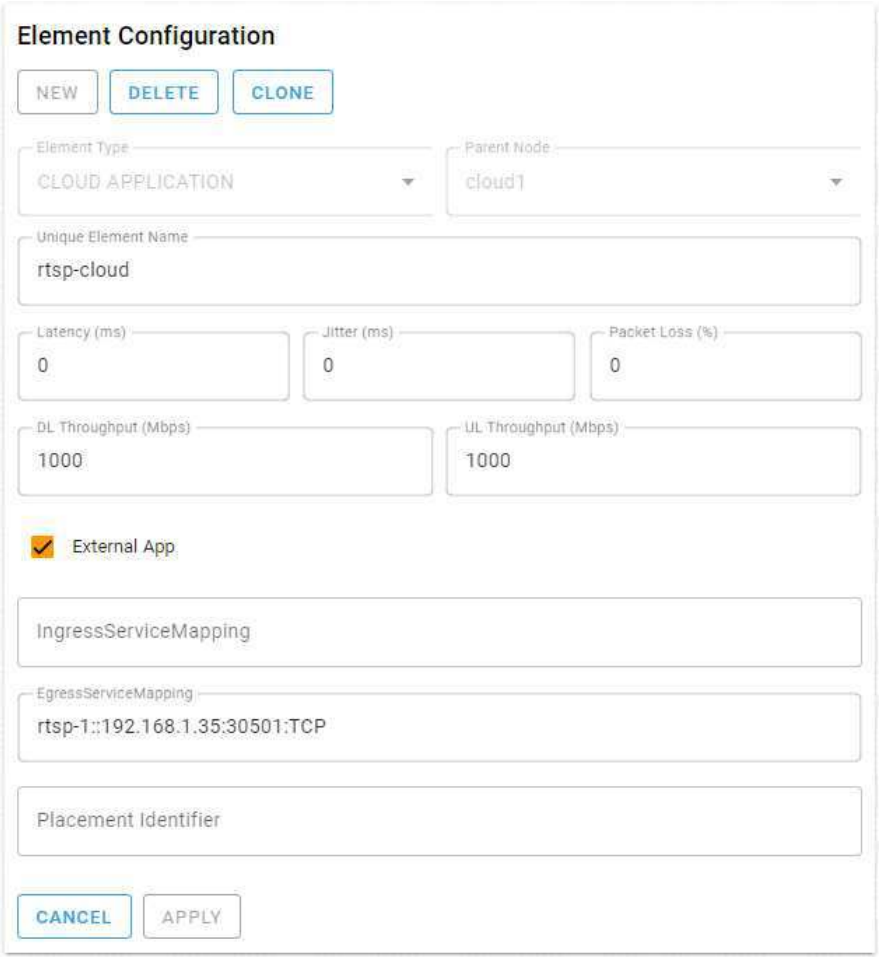
Similar à utilização das interfaces do tipo *TERMINAL APPLICATION*, a disponibilização de serviços externos ao ambiente de simulação de mobilidade é realizada a partir das interfaces do tipo *CLOUD APPLICATION*, representadas neste cenário pelas tags **mosquitto-cloud** e **rtsp-cloud**. As requisições foram direcionadas para os serviços em execução em uma máquina virtual na rede local, representando o servidor de computação em nuvem. Conforme apresentado nas Figuras 7.5 e 7.6, os mapeamentos de rede foram realizados para os endereços de IP e portas **192.168.1.35:30801** e **192.168.1.35:30501**, respectivamente.

Figura 7.5: Captura de tela do AdvantEDGE para a Visão de Configuração de Elementos do **mosquitto-cloud**.



Fonte: Produzida pelo autor.

Figura 7.6: Captura de tela do AdvantEDGE para a Visão de Configuração de Elementos do **rtsp-cloud**.



Fonte: Produzida pelo autor.

O uso da funcionalidade de Aplicação Externa no AdvantEDGE, com a configuração dos serviços *Ingress* e *Egress* torna possível o seu uso como uma plataforma de passagem simplesmente usando as interfaces - a entrada para um UE externo e a saída para um aplicativo de borda ou de nuvem externo.

Um Mapeamento de Entrada (serviços *Ingress*) é usado para redirecionar o tráfego externo para um nome de serviço interno ou aplicação externa. O *endpoint* final é determinado pela configuração do serviço. O formato para especificar um mapeamento de entrada é <*Ext-Port*>:<*Svc-Name*>:<*Svc-Port*>:<*Svc-Protocol*>. Além disso, vários mapeamentos de entrada podem ser especificados usando o separador de vírgula. Para o elemento **amb-mqtt**

(Figura 7.3), por exemplo, temos a seguinte configuração:

- <*Ext-Port*> = 30801
- <*Svc-Name*> = mosquitto-1
- <*Svc-Port*> = 30801
- <*Svc-Protocol*> = TCP

O tráfego TCP (configurado através do <*Svc-Protocol*>) recebido na porta <*Ext-Port*> = **30801** da plataforma AdvantEDGE é redirecionado para um serviço chamado <*Svc-Name*> = **mosquitto-1** na porta <*Svc-Port*> = **30801**. O serviço denominado <*Svc-Name*> = **mosquitto-1** pode ser implantado de forma independente como um serviço interno ou externo.

De modo análogo, um Mapeamento de Saída (serviços *Egress*) é usado para redirecionar o tráfego interno para um endereço IP e porta externos. O formato para especificar um mapeamento de saída é <*Svc-Name*>:<*ME-Svc-Name*>:<*IP*>:<*Port*>:<*Protocol*>. Também, vários mapeamentos de saída podem ser especificados usando o separador de vírgula Abaixo está um exemplo de um serviço de saída configuração para um aplicativo de servidor externo. Para o elemento **mosquitto-cloud** (Figura 7.5), por exemplo, temos a seguinte configuração:

- <*Svc-Name*> = mosquitto-1
- <*ME-Svc-Name*> = <vazio>
- <*IP*> = 192.168.1.35
- <*Port*> = 30801
- <*Protocol*> = TCP

Desse modo, para o cenário configurado, um equipamento médico transmitindo dados através da configuração de *Ingress* igual a **30801:*mosquitto-1*:30801:TCP** se conectará ao endereço IP do AdvantEDGE na porta 30801 (ou seja, no *endpoint* **192.168.1.13:30801**) - seu tráfego será redirecionado para o serviço interno **mosquitto-1** (definido tanto na configuração do *Ingress* como no *Egress*), que por sua vez é mapeado para um serviço externo através da configuração de *Egress* igual a ***mosquitto-1*::192.168.1.35:30801:TCP** - resultando na passagem do tráfego pelo AdvantEDGE e na aplicação das características de rede e padrão de mobilidade. Em outras palavras, a transferência de dados através da interface

amb-mqtt será redirecionada ao serviço em nuvem **mosquitto-cloud** através do serviço interno **mosquitto-1**. O mesmo ocorre entre os elementos **amb-video** e **rtsp-cloud**, que são interconectados através do serviço interno **rtsp-1**.

7.1.1.2 Configurações de Mapa

Visando a simulação da mobilidade da ambulância, a plataforma AdvantEDGE fornece um Sistema de Informação Geoespacial (GIS) integrado que se integra a cenários. Este sistema permite aos usuários criar simulações realistas, onde a dinâmica do deslocamento de veículos de emergência pode ser analisada e otimizada. Integrando-se a diferentes cenários, o GIS da AdvantEDGE facilita a avaliação do impacto de variáveis da mudança entre estações rádio-base, contribuindo assim para o desenvolvimento de serviços mais eficientes.

Desse modo, utilizou-se a ferramenta Google My Maps [1], que permite a criação de mapas customizados. Na Figura 7.7 está ilustrada a captura de tela do Google My Maps com a criação de um mapa customizado utilizado nas simulações. Neste trajeto, considerou-se que a ambulância estaria iniciando o atendimento de emergência na cidade de Lagoa Seca - PB e se deslocando em direção ao Hospital de Emergência e Trauma Dom Luiz Gonzaga Fernandes, na cidade de Campina Grande - PB.

Utilizando essa ferramenta, foi possível extrair com precisão os dados georreferenciados da rota especificada para o ensaio experimental. Esses dados foram exportados no formato de arquivo KML, que é utilizado para a visualização de informações geográficas em diversas plataformas de mapeamento. Este formato é compatível com aplicativos como Google Earth [2] e Google Maps [3], o que é particularmente vantajoso para o desenvolvimento de aplicações móveis e serviços digitais que requerem integração de dados geográficos para funcionalidades de localização e navegação em tempo-real.

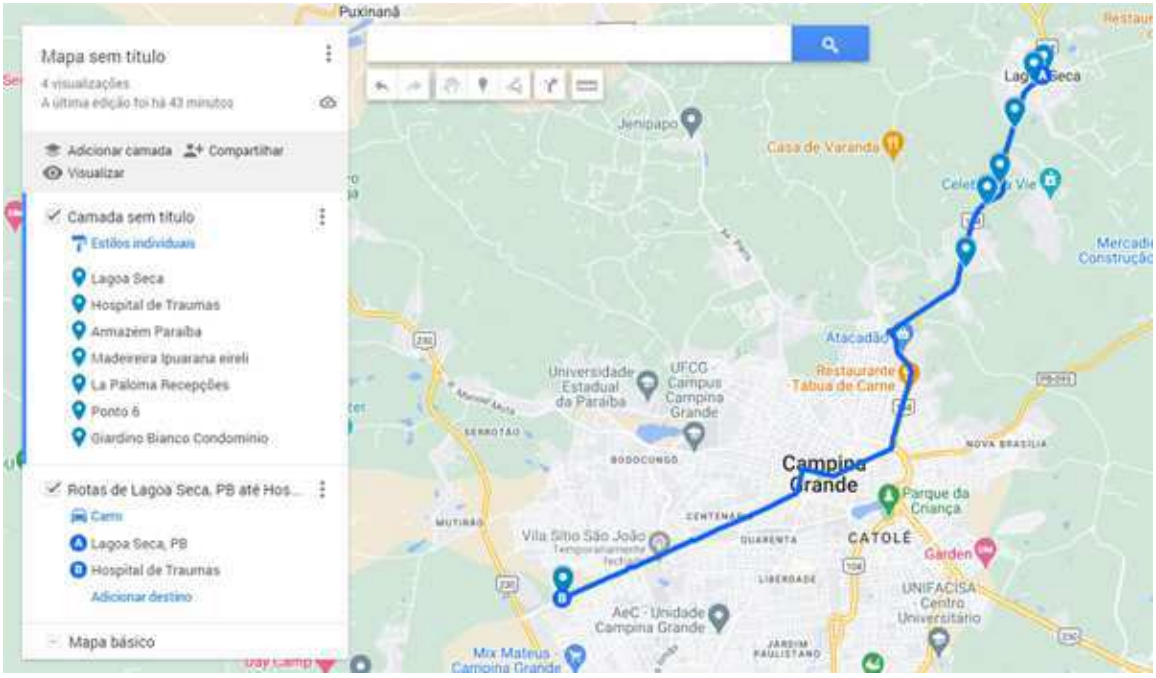
Na Figura 7.8 está ilustrada uma captura de tela de trecho do arquivo *.kml obtido através do Google My Maps com a criação de um mapa customizado, onde é possível observar as coordenadas de longitude e latitude selecionadas na imagem.

[1] urlhttps://www.google.com/mymaps/

[2] https://www.google.com/intl/pt-BR/earth/about/

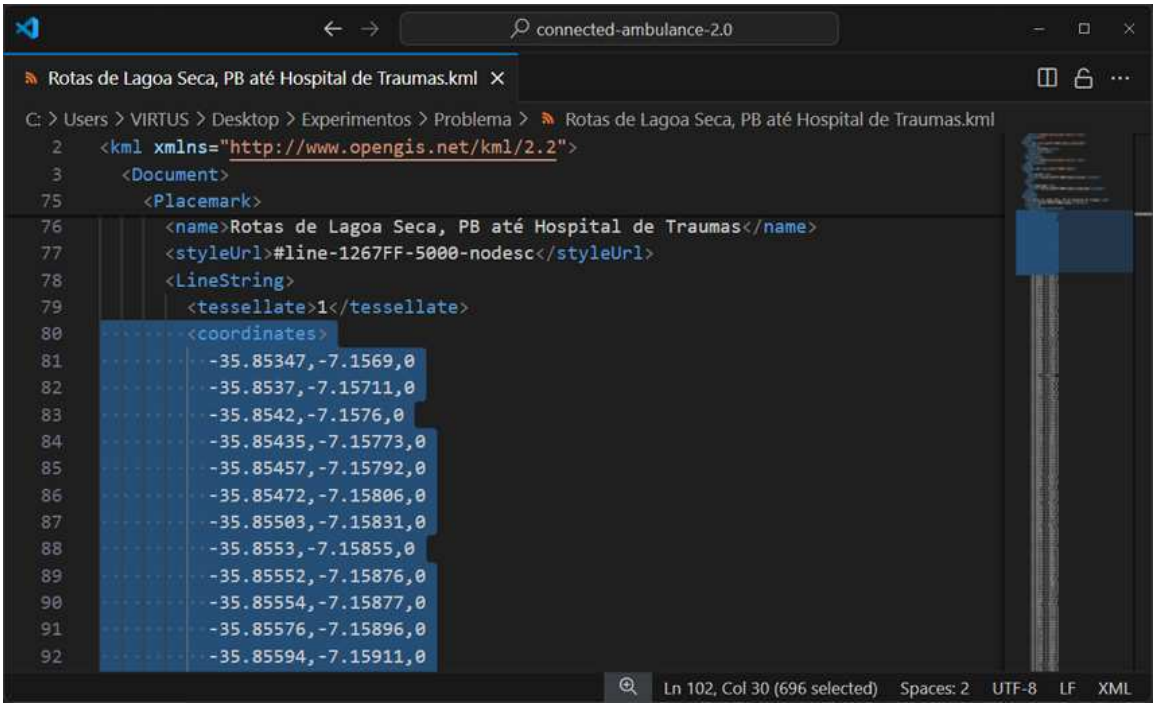
[3] https://maps.google.com/

Figura 7.7: Captura de tela do Google My Maps com a criação de um mapa customizado.



Fonte: Produzida pelo autor.

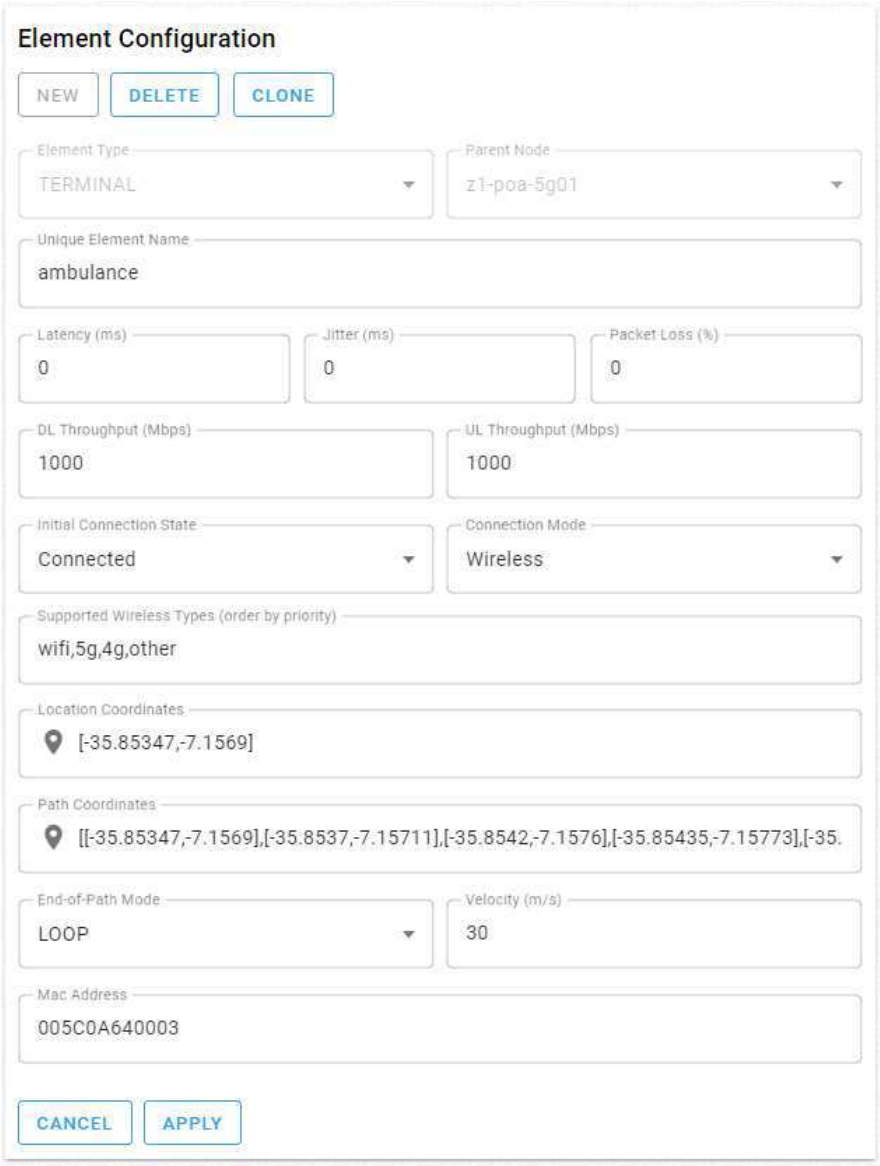
Figura 7.8: Captura de tela de trecho do arquivo *.kml obtido através do Google My Maps com a criação de um mapa customizado.



Fonte: Produzida pelo autor.

Utilizando os dados georreferenciados disponíveis no arquivo *.kml extraídos do mapa elaborado, foi possível definir uma rota específica para a ambulância na plataforma AdvantEDGE. Essa configuração está ilustrada na Figura 7.9, onde a rota estabelecida para a ambulância é claramente indicada no campo *Path Coordinates*, facilitando o entendimento e a implementação prática deste itinerário no contexto operacional em questão.

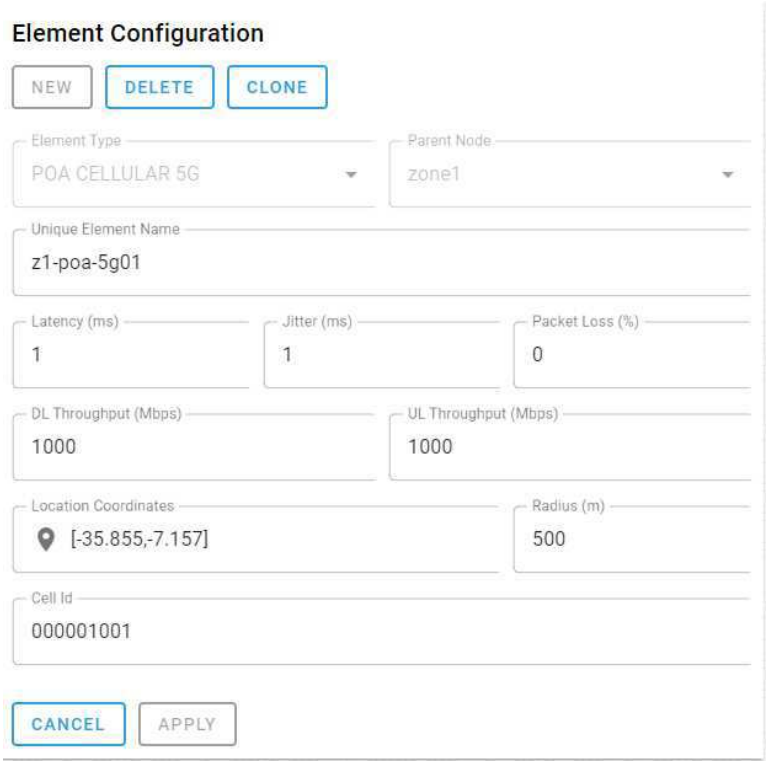
Figura 7.9: Captura de tela do AdvantEDGE para a Visão de Configuração de Elementos do **ambulance**.



Fonte: Produzida pelo autor.

A partir da definição da rota, foi realizada a configuração da localização geográfica dos demais elementos. Foi considerada a distribuição das antenas ao longo do trajeto, atribuindo-se as coordenadas específicas e o raio de alcance das antenas, nos campos *Location Coordinates* e *Radius*, respectivamente. A fins de exemplificação, na Figura 7.10 podem ser observadas as respectivas configurações.

Figura 7.10: Captura de tela do AdvantEDGE para a Visão de Configuração de Elementos do **z1-poa-5g01**.

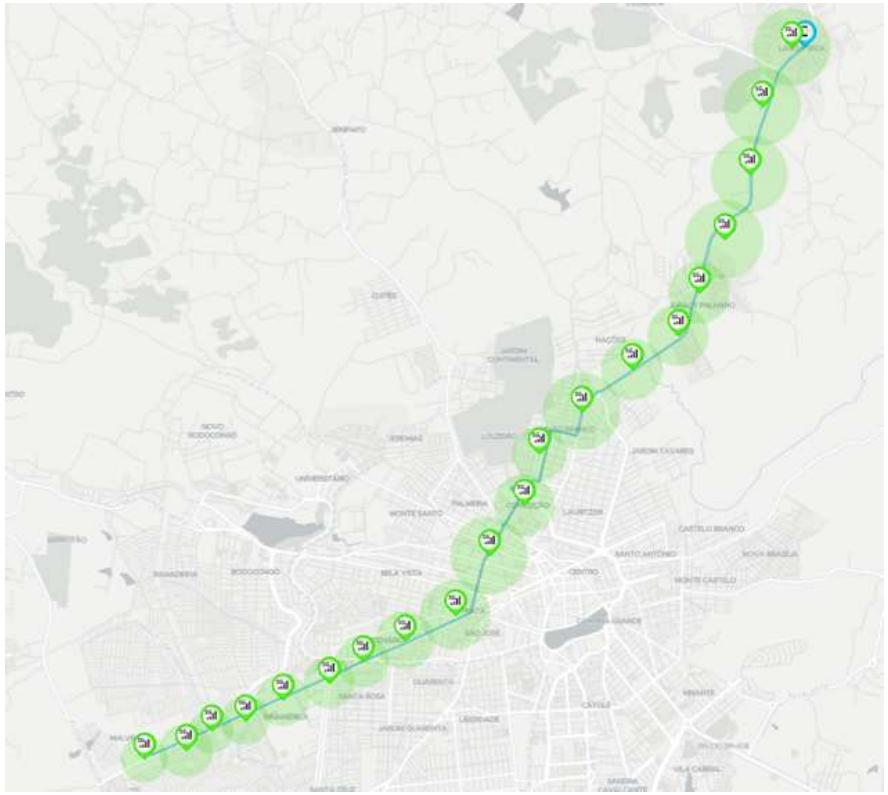


Fonte: Produzida pelo autor.

Com isso, a utilização da ferramenta possibilitou a geração de um mapa detalhado para simulação, que inclui a distribuição da cobertura das antenas ao longo do percurso especificado, conforme ilustrado na Figura 7.11b. A fim de conferir maior realismo à simulação, optou-se por posicionar o servidor de dados em nuvem na cidade de São Paulo, conforme ilustrado na Figura 7.11a. Essa escolha geográfica reflete as condições reais de conectividade e infraestrutura tecnológica. Tais configurações são cruciais para a análise e simulação das variações de desempenho da rede em diferentes segmentos do trajeto e facilita a identificação de potenciais pontos de melhoria no sistema de comunicação.

Figura 7.11: Captura de tela do AdvantEDGE para a Visão de Configuração de Mapas.

(a) Mapa do cenário com destaque para a localização do servidor em nuvem na cidade de São Paulo.



(b) Mapa do cenário com destaque para a distribuição das antenas 5G ao longo do trajeto.

Fonte: Produzida pelo autor.

7.1.1.3 Execução Experimental

Para a execução dos experimentos, foram disponibilizados serviços de streaming de vídeo e de dados. A implementação dos serviços de streaming de vídeo envolveu o a investigação, definição e disponibilização de protocolos de streaming com foco na baixa latência. Para isso, o protocolo utilizado foi o RTSP. A mesma abordagem foi utilizada para o streaming de dados, onde, no contexto de Internet das Coisas, o MQTT tem sido amplamente utilizado.

Para o streaming de vídeo RTSP, foi utilizada a imagem Docker `bluenviron/mediamtx`, disponível no Docker Hub [4]. A configuração e orquestração do serviço foram realizadas através do Kubernetes, conforme integração descrita no Capítulo 6 e detalhamento apresentado nas subseções anteriores. O serviço de *streaming* de vídeo foi configurado para transmitir fluxos de vídeo, que foram acessados e visualizados utilizando o *VLC Media Player*[5]. Para isso, foi empregado o modo de exibição de transmissão de rede do VLC, permitindo a avaliação da qualidade e estabilidade do vídeo sob diferentes condições de rede.

O streaming de dados utilizando o protocolo MQTT. A imagem Docker `eclipse-mosquitto:latest`, disponível no Docker Hub[6], foi utilizada para configurar o *broker* MQTT. Similarmente ao caso do streaming de vídeo, a gestão do serviço foi realizada via Kubernetes com o auxílio da plataforma AdvantEDGE, conforme integração descrita no Capítulo 6 e detalhamento apresentado nas subseções anteriores. Um código desenvolvido em Python foi utilizado para publicar e subscrever tópicos e dados no *broker* MQTT. Este código permitiu a interação em tempo real com o *broker*, facilitando o envio e recebimento de mensagens e obtenção de dados experimentais em diferentes condições experimentais.

Cada um dos experimentos foi submetido a uma série de testes para avaliar o desempenho e a resiliência dos serviços sob várias configurações de rede. Os resultados desses testes são discutidos em detalhes na próxima seção deste documento, incluindo métricas específicas de desempenho e análises comparativas.

[4] https://hub.docker.com/r/bluenviron/mediamtx
[5] videolan.org/vlc/index.pt_BR.html
[6] https://hub.docker.com/_/eclipse-mosquitto

7.1.1.4 Resultados

A avaliação inicial do experimento concentrou-se na validação da latência da rede dentro do ambiente experimental configurado. Para isso, utilizou-se a ferramenta *Wireshark*[7], conhecida por sua eficácia na análise de tráfego de rede. A mensuração da latência foi essencial para assegurar que a infraestrutura de rede atendesse às demandas do experimento, garantindo a integridade dos resultados obtidos.

Neste sentido, foram realizadas simulações de streaming de dados utilizando o protocolo MQTT para três diferentes configurações de latências para o servidor de nuvem no ambiente experimental, sendo elas 50ms, 200ms e 500ms. Os resultados experimentais das três configurações de latência revelam algumas diferenças no comportamento da rede, conforme observado através dos pacotes capturados pelo *Wireshark*, apresentados nas Tabelas 7.1, 7.2 e 7.3. A seguir, descrevemos os resultados específicos para cada configuração de latência.

- **Latência de 50ms:** Conforme observa-se na Tabela 7.1, a latência média de ponta-a-ponta para esta simulação foi de 62,45ms. Destaca-se que, além da latência definida para o servidor de nuvem, devem ser observadas as definições de todo o percurso de dados na infraestrutura de comunicação. Por exemplo, na Figura 7.2 é possível observar uma latência de ponta-a-ponta igual a 112ms. Ou seja, considerando o caso de a latência da nuvem ser configurada para 25ms, isso implicaria no valor de 62ms, que está bem próximo do valor médio obtido experimentalmente. Além disso, a comunicação entre o serviço e aplicação foi majoritariamente estável, com sequências de mensagens MQTT consistentes e sem retransmissões excessivas. A conexão foi estabelecida e mantida com mínimas interrupções, evidenciando um ambiente de rede otimizado para esta latência.

- **Latência de 200ms:** Fazendo-se a mesma validação do cenário com a latência de 50ms, o valor médio obtido para esta simulação foi de 214,62ms, conforme é possível observar na Tabela 7.2. Calculando-se o valor configurado no AdvantEDGE, a latência de ponta-a-ponta é de 212ms, que também está bem próximo do valor médio obtido experimentalmente. Observou-se, também, um aumento no número de retransmissões, especialmente nas mensagens MQTT, indicando uma deterioração na eficiência da

[7] https://www.wireshark.org/

transmissão de dados. A latência aumentada provocou atrasos notáveis na entrega de pacotes, o que poderia afetar aplicações críticas que dependem da rápida entrega na transmissão de dados.

- **Latência de 500ms:** A latência média obtida experimentalmente para este cenário foi de 516,78ms, frente a 512ms configurados no ambiente de simulação. A configuração com esta latência resultou em um desempenho significativamente degradado. Retransmissões frequentes e múltiplas transmissões espúrias foram registradas, sugerindo uma perda considerável de pacotes e uma baixa confiabilidade na comunicação. A latência elevada comprometeu severamente a integridade e a eficácia do fluxo de dados.

Os testes realizados permitiram uma análise detalhada do comportamento da rede. Esta fase inicial de resultados validou a configuração do ambiente experimental, estabelecendo uma base sólida para a continuação da pesquisa e a execução de testes mais direcionados ao foco do estudo. Além disso, estes resultados demonstram claramente o impacto direto da latência na performance da rede, ressaltando a necessidade de otimizar as configurações de rede para cada aplicação específica, a fim de minimizar problemas de transmissão e garantir a eficiência operacional.

Esta observação motivou a avaliação mais detalhada da relação entre a latência e a taxa de transmissão. Para isso, foram realizadas simulações com diferentes latências de ponta-a-ponta para streamings de vídeo em FULL HD (1920x1080 25fps) e 4K (3840x2160 25fps). A simulação consistiu na disponibilização do serviço de servidor de streaming RTSP, onde os streamings dos vídeos foram enviados através da ferramenta *FFmpeg* [8] e exibidos com o *VLC Media Player*. Com o estabelecimento da conexão entre a aplicação, sendo encaminhada através do ambiente AdvantEDGE, foi realizado o incremento gradual da latência de ponta-a-ponta no ambiente de simulação, de 10 em 10ms para a faixa de 0 a 200ms. Após isso, o passo foi de 50ms para a faixa de 200 a 500ms. Os resultados obtidos para estes experimentos estão ilustrados nas Figuras 7.12, 7.13 e 7.14, para o streaming de vídeo em FULL HD, e nas Figuras 7.15, 7.16 e 7.17, para o streaming de vídeo em 4K. Nas Figuras 7.18 e 7.19 estão apresentados os consumos de CPU e memória para este cenário, respectivamente.

[8] https://ffmpeg.org/

Tabela 7.1: Dados obtidos a partir do *Wireshark* - Latência = 50ms.

No.	Time	Source	Destination	Protocol	Length	Info
1033	19,250720057	192.168.1.5	192.168.1.13	TCP	66	61156 > 31001 [SYN] Seq=0 Win=64240 Len=0 MSS=1460 WS=256 SACK_PERM=1
1036	19,393501545	192.168.1.13	192.168.1.14	TCP	66	24856 > 1883 [SYN] Seq=0 Win=64240 Len=0 MSS=1460 WS=256 SACK_PERM=1
1039	19,411615024	192.168.1.5	192.168.1.13	TCP	60	61156 > 31001 [ACK] Seq=1 Ack=1 Win=131328 Len=0
1040	19,411615158	192.168.1.5	192.168.1.13	TCP	68	61156 > 31001 [PSH, ACK] Seq=1 Ack=1 Win=131328 Len=14
1046	19,512893413	192.168.1.13	192.168.1.14	TCP	54	24856 > 1883 [ACK] Seq=1 Ack=1 Win=131328 Len=0
1047	19,534550574	192.168.1.13	192.168.1.14	MQTT	68	Connect Command
1053	19,555650723	192.168.1.5	192.168.1.13	TCP	128	61156 > 31001 [PSH, ACK] Seq=15 Ack=1 Win=131328 Len=74
1054	19,603981719	192.168.1.5	192.168.1.13	TCP	60	61156 > 31001 [ACK] Seq=89 Ack=5 Win=131328 Len=0
1055	19,608565302	192.168.1.13	192.168.1.14	MQTT	128	Publish Message [ambulance/ecg], Subscribe Request (id=2) [ambulance/ecg], Publish Message [ambulance/ecg]
1066	19,711497051	192.168.1.13	192.168.1.14	TCP	54	24856 > 1883 [ACK] Seq=89 Ack=5 Win=131328 Len=0
1070	19,789877530	192.168.1.5	192.168.1.13	TCP	60	61156 > 31001 [ACK] Seq=89 Ack=47 Win=131328 Len=0
1101	19,932971900	192.168.1.13	192.168.1.14	TCP	54	24856 > 1883 [ACK] Seq=89 Ack=47 Win=131328 Len=0
1109	20,420185863	192.168.1.5	192.168.1.13	TCP	91	61156 > 31001 [PSH, ACK] Seq=89 Ack=47 Win=131328 Len=37
1110	20,531710816	192.168.1.13	192.168.1.14	MQTT	91	Publish Message [ambulance/ecg]
1115	20,604954495	192.168.1.5	192.168.1.13	TCP	60	61156 > 31001 [ACK] Seq=126 Ack=84 Win=131072 Len=0
1116	20,747283153	192.168.1.13	192.168.1.14	TCP	54	24856 > 1883 [ACK] Seq=126 Ack=84 Win=131072 Len=0
1153	21,416202144	192.168.1.5	192.168.1.13	TCP	92	61156 > 31001 [PSH, ACK] Seq=126 Ack=84 Win=131072 Len=38
1156	21,547688010	192.168.1.13	192.168.1.14	MQTT	92	Publish Message [ambulance/ecg]
1159	21,610346974	192.168.1.5	192.168.1.13	TCP	60	61156 > 31001 [ACK] Seq=164 Ack=122 Win=131072 Len=0
1168	21,738125585	192.168.1.13	192.168.1.14	TCP	54	24856 > 1883 [ACK] Seq=164 Ack=122 Win=131072 Len=0
1209	22,416710347	192.168.1.5	192.168.1.13	TCP	91	61156 > 31001 [PSH, ACK] Seq=164 Ack=122 Win=131072 Len=37
1219	22,558324071	192.168.1.13	192.168.1.14	MQTT	91	Publish Message [ambulance/ecg]
1222	22,616792002	192.168.1.5	192.168.1.13	TCP	60	61156 > 31001 [ACK] Seq=201 Ack=159 Win=131072 Len=0
1223	22,710539618	192.168.1.13	192.168.1.14	TCP	54	24856 > 1883 [ACK] Seq=201 Ack=159 Win=131072 Len=0
1256	23,423531254	192.168.1.5	192.168.1.13	TCP	91	61156 > 31001 [PSH, ACK] Seq=201 Ack=159 Win=131072 Len=37
1265	23,538463171	192.168.1.13	192.168.1.14	MQTT	91	Publish Message [ambulance/ecg]
1270	23,611957720	192.168.1.5	192.168.1.13	TCP	60	61156 > 31001 [ACK] Seq=238 Ack=196 Win=131072 Len=0
1279	23,770222836	192.168.1.13	192.168.1.14	TCP	54	24856 > 1883 [ACK] Seq=238 Ack=196 Win=131072 Len=0
1303	24,027749203	192.168.1.5	192.168.1.13	TCP	60	61156 > 31001 [PSH, ACK] Seq=238 Ack=196 Win=131072 Len=2
1304	24,027749439	192.168.1.5	192.168.1.13	TCP	60	61156 > 31001 [FIN, ACK] Seq=240 Ack=196 Win=131072 Len=0
1310	24,151285643	192.168.1.13	192.168.1.14	MQTT	56	Disconnect Req
1311	24,155672863	192.168.1.13	192.168.1.14	TCP	54	24856 > 1883 [FIN, ACK] Seq=240 Ack=196 Win=131072 Len=0
1315	24,156974502	192.168.1.5	192.168.1.13	TCP	60	61156 > 31001 [ACK] Seq=241 Ack=197 Win=131072 Len=0
1317	24,246477076	192.168.1.13	192.168.1.14	TCP	54	24856 > 1883 [ACK] Seq=241 Ack=197 Win=131072 Len=0

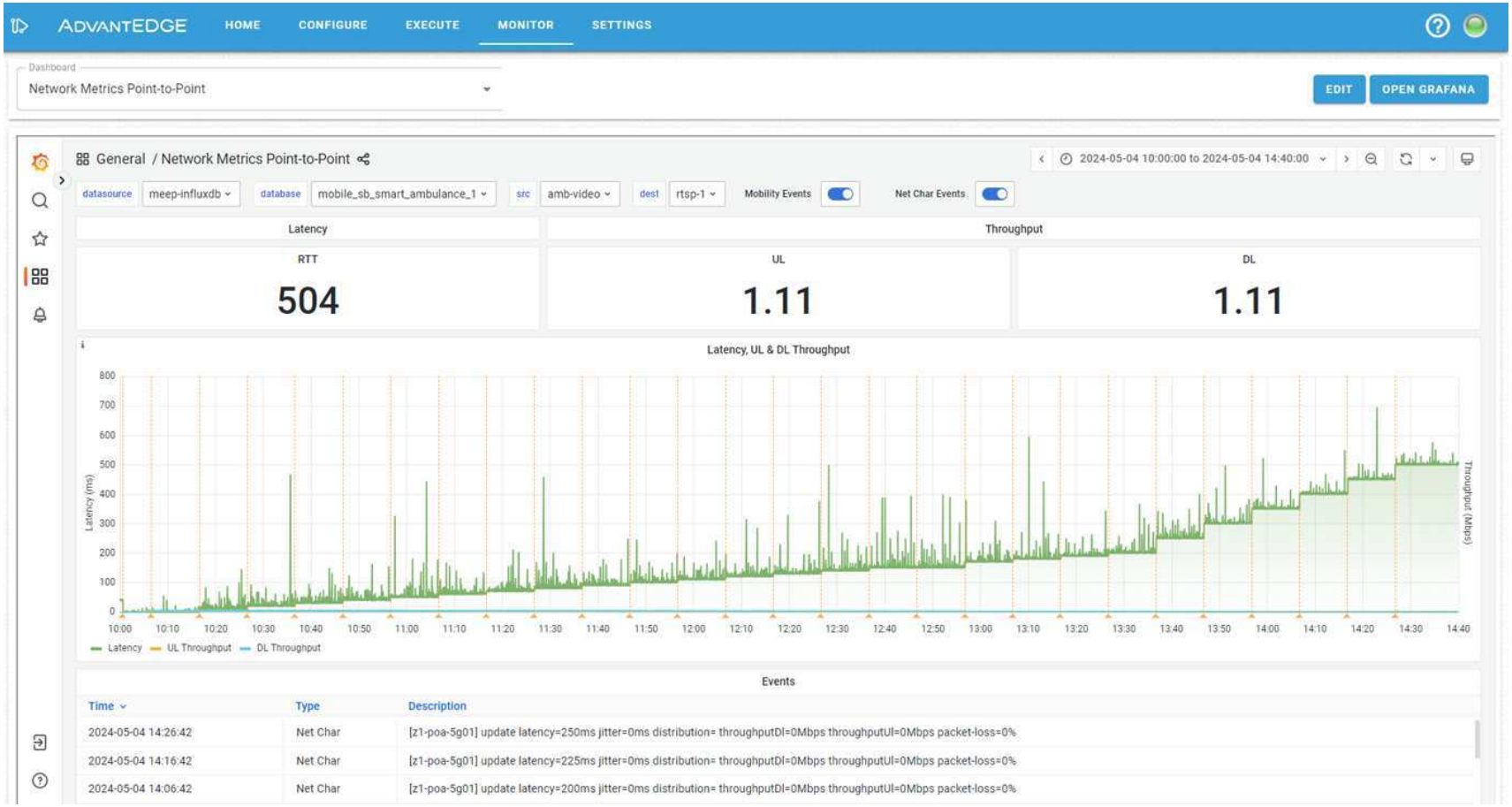
Tabela 7.2: Dados obtidos a partir do *Wireshark* - Latência = 200ms.

No.	Time	Source	Destination	Protocol	Length	Info
1247	28,449701097	192.168.1.5	192.168.1.13	TCP	66	63502 > 31001 [SYN] Seq=0 Win=64240 Len=0 MSS=1460 WS=256 SACK_PERM=1
1264	28,611607056	192.168.1.13	192.168.1.14	TCP	66	25135 > 1883 [SYN] Seq=0 Win=64240 Len=0 MSS=1460 WS=256 SACK_PERM=1
1267	28,616159582	192.168.1.5	192.168.1.13	TCP	60	63502 > 31001 [ACK] Seq=1 Ack=1 Win=131328 Len=0
1268	28,616159675	192.168.1.5	192.168.1.13	TCP	68	63502 > 31001 [PSH, ACK] Seq=1 Ack=1 Win=131328 Len=14
1290	28,716678480	192.168.1.13	192.168.1.14	TCP	54	25135 > 1883 [ACK] Seq=1 Ack=1 Win=131328 Len=0
1291	28,723858964	192.168.1.13	192.168.1.14	MQTT	68	Connect Command
1296	28,744445754	192.168.1.5	192.168.1.13	TCP	128	63502 > 31001 [PSH, ACK] Seq=15 Ack=1 Win=131328 Len=74
1297	28,791490614	192.168.1.5	192.168.1.13	TCP	60	63502 > 31001 [ACK] Seq=89 Ack=5 Win=131328 Len=0
1298	28,855106041	192.168.1.13	192.168.1.14	MQTT	128	Publish Message [ambulance/ecg], Subscribe Request (id=2) [ambulance/ecg], Publish Message [ambulance/ecg]
1299	28,859980908	192.168.1.13	192.168.1.14	TCP	54	25135 > 1883 [ACK] Seq=89 Ack=5 Win=131328 Len=0
1302	28,906819677	192.168.1.5	192.168.1.13	TCP	60	63502 > 31001 [ACK] Seq=89 Ack=47 Win=131328 Len=0
1305	29,032463734	192.168.1.13	192.168.1.14	TCP	54	25135 > 1883 [ACK] Seq=89 Ack=47 Win=131328 Len=0
1329	29,622042067	192.168.1.5	192.168.1.13	TCP	91	63502 > 31001 [PSH, ACK] Seq=89 Ack=47 Win=131328 Len=37
1347	30,062782440	192.168.1.5	192.168.1.13	TCP	91	[TCP Retransmission] 63502 > 31001 [PSH, ACK] Seq=89 Ack=47 Win=131328 Len=37
1348	30,085321505	192.168.1.13	192.168.1.14	MQTT	91	Publish Message [ambulance/ecg]
1356	30,180512506	192.168.1.5	192.168.1.13	TCP	60	63502 > 31001 [ACK] Seq=126 Ack=84 Win=131072 Len=0
1359	30,442214073	192.168.1.5	192.168.1.13	TCP	66	[TCP Dup ACK 1356#1] 63502 > 31001 [ACK] Seq=126 Ack=84 Win=131072 Len=0 SLE=47 SRE=84
1360	30,480664880	192.168.1.13	192.168.1.14	TCP	91	[TCP Spurious Retransmission] 25135 > 1883 [PSH, ACK] Seq=89 Ack=47 Win=131328 Len=37
1365	30,572675683	192.168.1.13	192.168.1.14	TCP	54	25135 > 1883 [ACK] Seq=126 Ack=84 Win=131072 Len=0
1372	30,629610986	192.168.1.5	192.168.1.13	TCP	92	63502 > 31001 [PSH, ACK] Seq=126 Ack=84 Win=131072 Len=38
1401	30,887779801	192.168.1.13	192.168.1.14	TCP	66	[TCP Dup ACK 1365#1] 25135 > 1883 [ACK] Seq=126 Ack=84 Win=131072 Len=0 SLE=47 SRE=84
1402	31,052055389	192.168.1.13	192.168.1.14	MQTT	92	Publish Message [ambulance/ecg]
1405	31,064319010	192.168.1.5	192.168.1.13	TCP	92	[TCP Retransmission] 63502 > 31001 [PSH, ACK] Seq=126 Ack=84 Win=131072 Len=38
1408	31,132198366	192.168.1.5	192.168.1.13	TCP	60	63502 > 31001 [ACK] Seq=164 Ack=122 Win=131072 Len=0
1411	31,420921459	192.168.1.5	192.168.1.13	TCP	66	[TCP Dup ACK 1408#1] 63502 > 31001 [ACK] Seq=164 Ack=122 Win=131072 Len=0 SLE=84 SRE=122
1412	31,535880514	192.168.1.13	192.168.1.14	TCP	92	[TCP Spurious Retransmission] 25135 > 1883 [PSH, ACK] Seq=126 Ack=84 Win=131072 Len=38
1415	31,555598686	192.168.1.13	192.168.1.14	TCP	54	25135 > 1883 [ACK] Seq=164 Ack=122 Win=131072 Len=0
1430	31,625432693	192.168.1.5	192.168.1.13	TCP	91	63502 > 31001 [PSH, ACK] Seq=164 Ack=122 Win=131072 Len=37
1447	31,880635514	192.168.1.13	192.168.1.14	TCP	66	[TCP Dup ACK 1415#1] 25135 > 1883 [ACK] Seq=164 Ack=122 Win=131072 Len=0 SLE=84 SRE=122
1448	32,030744669	192.168.1.13	192.168.1.14	MQTT	91	Publish Message [ambulance/ecg]
1453	32,104923257	192.168.1.5	192.168.1.13	TCP	60	63502 > 31001 [ACK] Seq=201 Ack=159 Win=131072 Len=0
1460	32,390622655	192.168.1.5	192.168.1.13	TCP	66	[TCP Dup ACK 1453#1] 63502 > 31001 [ACK] Seq=201 Ack=159 Win=131072 Len=0 SLE=122 SRE=159
1468	32,487403637	192.168.1.13	192.168.1.14	TCP	54	25135 > 1883 [ACK] Seq=201 Ack=159 Win=131072 Len=0
1481	32,628327540	192.168.1.5	192.168.1.13	TCP	91	63502 > 31001 [PSH, ACK] Seq=201 Ack=159 Win=131072 Len=37
1510	32,832344670	192.168.1.13	192.168.1.14	TCP	66	[TCP Dup ACK 1468#1] 25135 > 1883 [ACK] Seq=201 Ack=159 Win=131072 Len=0 SLE=122 SRE=159
1512	33,081724558	192.168.1.13	192.168.1.14	MQTT	91	Publish Message [ambulance/ecg]
1515	33,142883499	192.168.1.5	192.168.1.13	TCP	60	63502 > 31001 [ACK] Seq=238 Ack=196 Win=131072 Len=0
1518	33,434631490	192.168.1.5	192.168.1.13	TCP	66	[TCP Dup ACK 1515#1] 63502 > 31001 [ACK] Seq=238 Ack=196 Win=131072 Len=0 SLE=159 SRE=196
1519	33,568322470	192.168.1.13	192.168.1.14	TCP	54	25135 > 1883 [ACK] Seq=238 Ack=196 Win=131072 Len=0
1533	33,628689719	192.168.1.5	192.168.1.13	TCP	91	63502 > 31001 [PSH, ACK] Seq=238 Ack=196 Win=131072 Len=37
1548	33,891519868	192.168.1.13	192.168.1.14	TCP	66	[TCP Dup ACK 1519#1] 25135 > 1883 [ACK] Seq=238 Ack=196 Win=131072 Len=0 SLE=159 SRE=196
1549	33,905972832	192.168.1.5	192.168.1.13	TCP	60	63502 > 31001 [FIN, PSH, ACK] Seq=275 Ack=196 Win=131072 Len=2
1550	34,030438921	192.168.1.13	192.168.1.14	MQTT	91	Publish Message [ambulance/ecg]
1553	34,036781852	192.168.1.5	192.168.1.13	TCP	60	63502 > 31001 [RST, ACK] Seq=278 Ack=233 Win=0 Len=0
1554	34,328882093	192.168.1.13	192.168.1.14	MQTT	56	Disconnect Req
1562	34,463393711	192.168.1.13	192.168.1.14	TCP	54	25135 > 1883 [RST, ACK] Seq=278 Ack=233 Win=0 Len=0

Tabela 7.3: Dados obtidos a partir do *Wireshark* - Latência = 500ms.

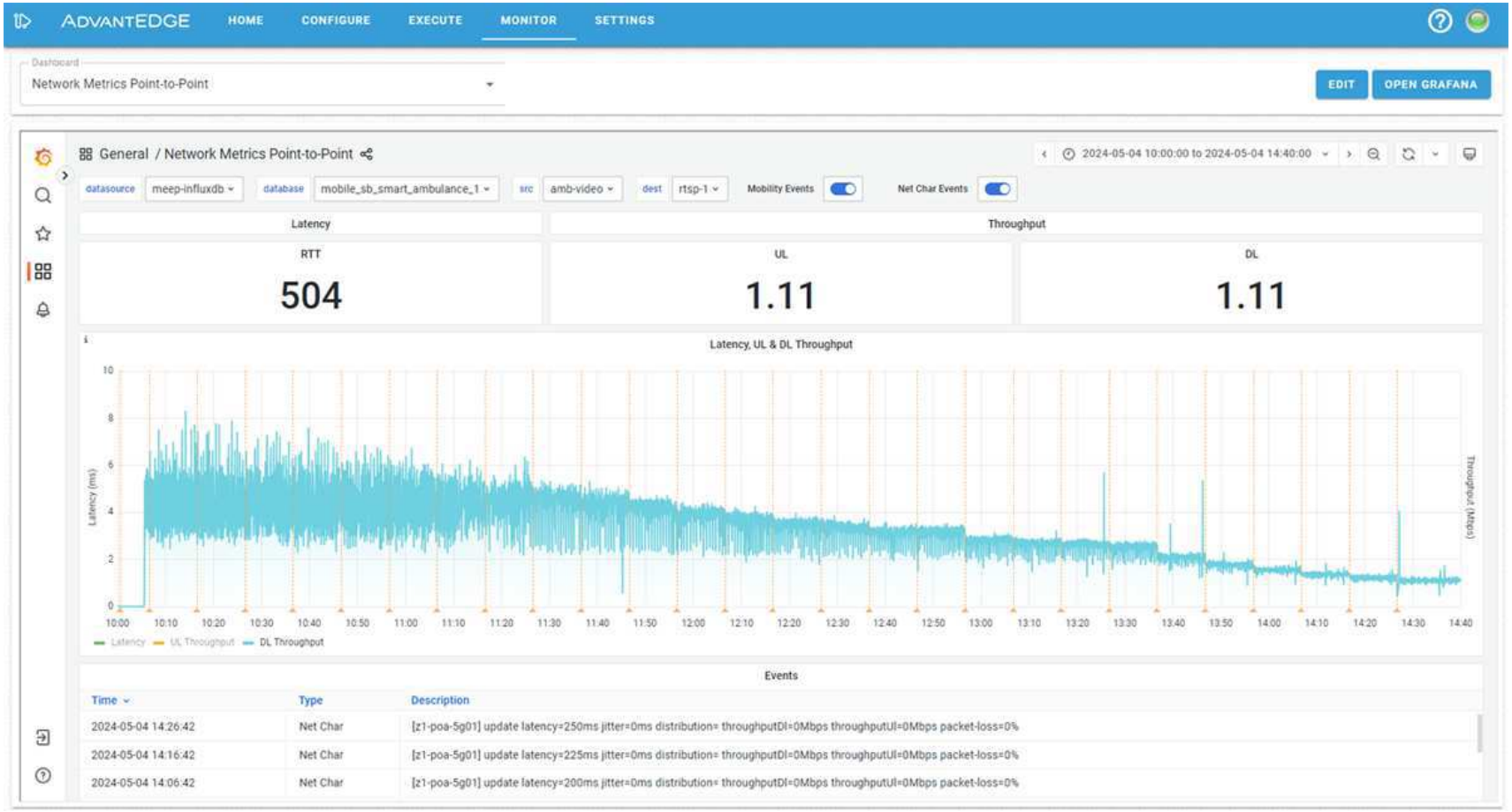
No.	Time	Source	Destination	Protocol	Length	Info
872	18,165794213	192.168.1.5	192.168.1.13	TCP	66	65191 > 31001 [SYN] Seq=0 Win=64240 Len=0 MSS=1460 WS=256 SACK_PERM=1
914	19,170939673	192.168.1.5	192.168.1.13	TCP	66	[TCP Retransmission] [TCP Port numbers reused] 65191 > 31001 [SYN] Seq=0 Win=64240 Len=0 MSS=1460 WS=256 SACK_PERM=1
925	19,238132882	192.168.1.13	192.168.1.14	TCP	66	13055 > 1883 [SYN] Seq=0 Win=64240 Len=0 MSS=1460 WS=256 SACK_PERM=1
928	19,239243413	192.168.1.5	192.168.1.13	TCP	60	65191 > 31001 [ACK] Seq=1 Ack=1 Win=131328 Len=0
930	19,257937866	192.168.1.5	192.168.1.13	TCP	68	65191 > 31001 [PSH, ACK] Seq=1 Ack=1 Win=131328 Len=14
952	19,491707397	192.168.1.5	192.168.1.13	TCP	142	[TCP Retransmission] 65191 > 31001 [PSH, ACK] Seq=1 Ack=1 Win=131328 Len=88
970	19,788054011	192.168.1.5	192.168.1.13	TCP	142	[TCP Retransmission] 65191 > 31001 [PSH, ACK] Seq=1 Ack=1 Win=131328 Len=88
974	20,173957068	192.168.1.13	192.168.1.14	TCP	66	[TCP Retransmission] [TCP Port numbers reused] 13055 > 1883 [SYN] Seq=0 Win=64240 Len=0 MSS=1460 WS=256 SACK_PERM=1
977	20,201203949	192.168.1.5	192.168.1.13	TCP	66	[TCP Dup ACK 928#1] 65191 > 31001 [ACK] Seq=89 Ack=1 Win=131328 Len=0 SLE=0 SRE=1
1002	20,308331020	192.168.1.13	192.168.1.14	TCP	54	13055 > 1883 [ACK] Seq=1 Ack=1 Win=131328 Len=0
1003	20,308802140	192.168.1.13	192.168.1.14	MQTT	68	Connect Command
1009	20,337052275	192.168.1.5	192.168.1.13	TCP	91	65191 > 31001 [PSH, ACK] Seq=89 Ack=1 Win=131328 Len=37
1011	20,384858203	192.168.1.5	192.168.1.13	TCP	60	65191 > 31001 [ACK] Seq=126 Ack=5 Win=131328 Len=0
1014	20,505126865	192.168.1.13	192.168.1.14	TCP	142	[TCP Retransmission] 13055 > 1883 [PSH, ACK] Seq=1 Ack=1 Win=131328 Len=88
1018	20,795293197	192.168.1.5	192.168.1.13	TCP	91	[TCP Retransmission] 65191 > 31001 [PSH, ACK] Seq=89 Ack=5 Win=131328 Len=37
1019	20,799694038	192.168.1.13	192.168.1.14	TCP	142	[TCP Spurious Retransmission] 13055 > 1883 [PSH, ACK] Seq=1 Ack=1 Win=131328 Len=88
1041	21,103764874	192.168.1.5	192.168.1.13	TCP	91	[TCP Retransmission] 65191 > 31001 [PSH, ACK] Seq=89 Ack=5 Win=131328 Len=37
1042	21,214190158	192.168.1.13	192.168.1.14	TCP	66	[TCP Dup ACK 1002#1] 13055 > 1883 [ACK] Seq=89 Ack=1 Win=131328 Len=0 SLE=0 SRE=1
1072	21,381687523	192.168.1.13	192.168.1.14	MQTT	91	Publish Message [ambulance/ecg]
1074	21,391041231	192.168.1.5	192.168.1.13	TCP	60	65191 > 31001 [ACK] Seq=126 Ack=47 Win=131328 Len=0
1076	21,409104495	192.168.1.5	192.168.1.13	TCP	92	65191 > 31001 [PSH, ACK] Seq=126 Ack=47 Win=131328 Len=38
1077	21,445858297	192.168.1.13	192.168.1.14	TCP	54	13055 > 1883 [ACK] Seq=126 Ack=5 Win=131328 Len=0
1078	21,636019576	192.168.1.5	192.168.1.13	TCP	92	[TCP Retransmission] 65191 > 31001 [PSH, ACK] Seq=126 Ack=47 Win=131328 Len=38
1081	21,837276373	192.168.1.13	192.168.1.14	TCP	91	[TCP Spurious Retransmission] 13055 > 1883 [PSH, ACK] Seq=89 Ack=5 Win=131328 Len=37
1090	21,932969085	192.168.1.5	192.168.1.13	TCP	92	[TCP Retransmission] 65191 > 31001 [PSH, ACK] Seq=126 Ack=47 Win=131328 Len=38
1091	22,101793739	192.168.1.13	192.168.1.14	TCP	91	[TCP Spurious Retransmission] 13055 > 1883 [PSH, ACK] Seq=89 Ack=5 Win=131328 Len=37
1133	22,395759870	192.168.1.13	192.168.1.14	TCP	54	13055 > 1883 [ACK] Seq=126 Ack=47 Win=131328 Len=0
1137	22,443379821	192.168.1.13	192.168.1.14	MQTT	92	Publish Message [ambulance/ecg]
1139	22,448092149	192.168.1.5	192.168.1.13	TCP	60	65191 > 31001 [ACK] Seq=164 Ack=84 Win=131072 Len=0
1141	22,448787250	192.168.1.5	192.168.1.13	TCP	91	65191 > 31001 [PSH, ACK] Seq=164 Ack=84 Win=131072 Len=37
1147	22,626736053	192.168.1.13	192.168.1.14	TCP	92	[TCP Spurious Retransmission] 13055 > 1883 [PSH, ACK] Seq=126 Ack=47 Win=131328 Len=38
1153	22,687676399	192.168.1.5	192.168.1.13	TCP	91	[TCP Retransmission] 65191 > 31001 [PSH, ACK] Seq=164 Ack=84 Win=131072 Len=37
1154	22,949932734	192.168.1.13	192.168.1.14	TCP	92	[TCP Spurious Retransmission] 13055 > 1883 [PSH, ACK] Seq=126 Ack=47 Win=131328 Len=38
1157	23,006829126	192.168.1.5	192.168.1.13	TCP	91	[TCP Retransmission] 65191 > 31001 [PSH, ACK] Seq=164 Ack=84 Win=131072 Len=37
1187	23,487148460	192.168.1.13	192.168.1.14	TCP	54	13055 > 1883 [ACK] Seq=164 Ack=84 Win=131072 Len=0
1188	23,487536184	192.168.1.13	192.168.1.14	MQTT	91	Publish Message [ambulance/ecg]
1193	23,488822989	192.168.1.5	192.168.1.13	TCP	91	65191 > 31001 [PSH, ACK] Seq=201 Ack=122 Win=131072 Len=37
1211	23,659149102	192.168.1.13	192.168.1.14	TCP	91	[TCP Spurious Retransmission] 13055 > 1883 [PSH, ACK] Seq=164 Ack=84 Win=131072 Len=37
1215	23,724712494	192.168.1.5	192.168.1.13	TCP	91	[TCP Retransmission] 65191 > 31001 [PSH, ACK] Seq=201 Ack=122 Win=131072 Len=37
1244	24,020676992	192.168.1.5	192.168.1.13	TCP	91	[TCP Retransmission] 65191 > 31001 [PSH, ACK] Seq=201 Ack=122 Win=131072 Len=37
1245	24,048818035	192.168.1.13	192.168.1.14	TCP	91	[TCP Spurious Retransmission] 13055 > 1883 [PSH, ACK] Seq=164 Ack=84 Win=131072 Len=37
1285	24,248611322	192.168.1.5	192.168.1.13	TCP	60	65191 > 31001 [FIN, PSH, ACK] Seq=238 Ack=122 Win=131072 Len=2
1327	24,505398403	192.168.1.13	192.168.1.14	MQTT	91	Publish Message [ambulance/ecg]
1330	24,507528264	192.168.1.5	192.168.1.13	TCP	60	65191 > 31001 [RST, ACK] Seq=241 Ack=159 Win=0 Len=0
1342	24,751195881	192.168.1.13	192.168.1.14	TCP	91	[TCP Spurious Retransmission] 13055 > 1883 [PSH, ACK] Seq=201 Ack=122 Win=131072 Len=37
1345	25,043764653	192.168.1.13	192.168.1.14	TCP	91	[TCP Spurious Retransmission] 13055 > 1883 [PSH, ACK] Seq=201 Ack=122 Win=131072 Len=37
1359	25,218239611	192.168.1.13	192.168.1.14	MQTT	56	Disconnect Req
1402	25.501.918.540	192.168.1.13	192.168.1.14	TCP	54	13055 > 1883 [RST, ACK] Seq=241 Ack=159 Win=0 Len=0

Figura 7.12: Captura de tela do AdvantEDGE para a Visão de Monitoramento de Métricas de Rede - Latência FULL HD.



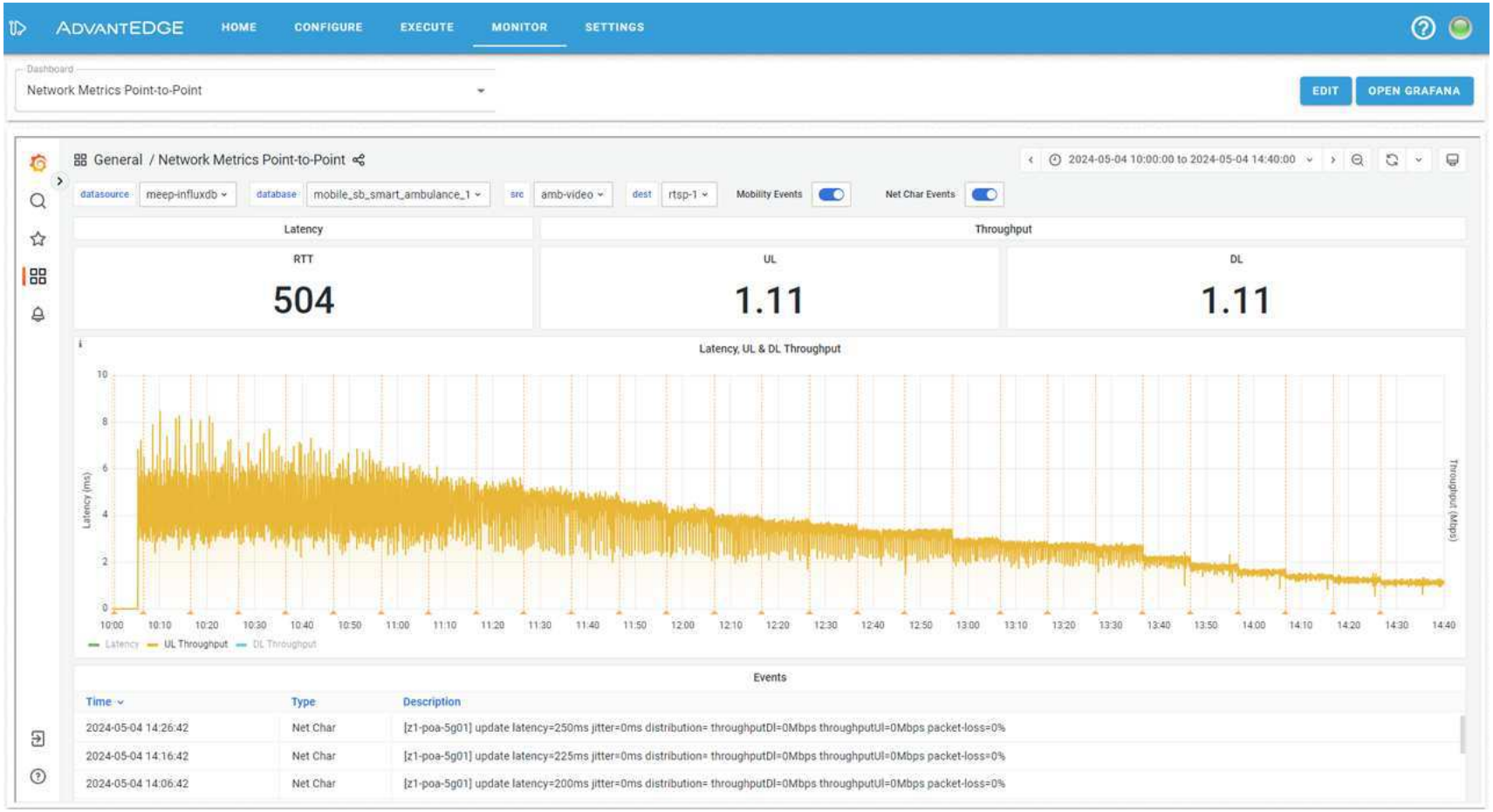
Fonte: Produzida pelo autor.

Figura 7.13: Captura de tela do AdvantEDGE para a Visão de Monitoramento de Métricas de Rede - Downlink FULL HD.



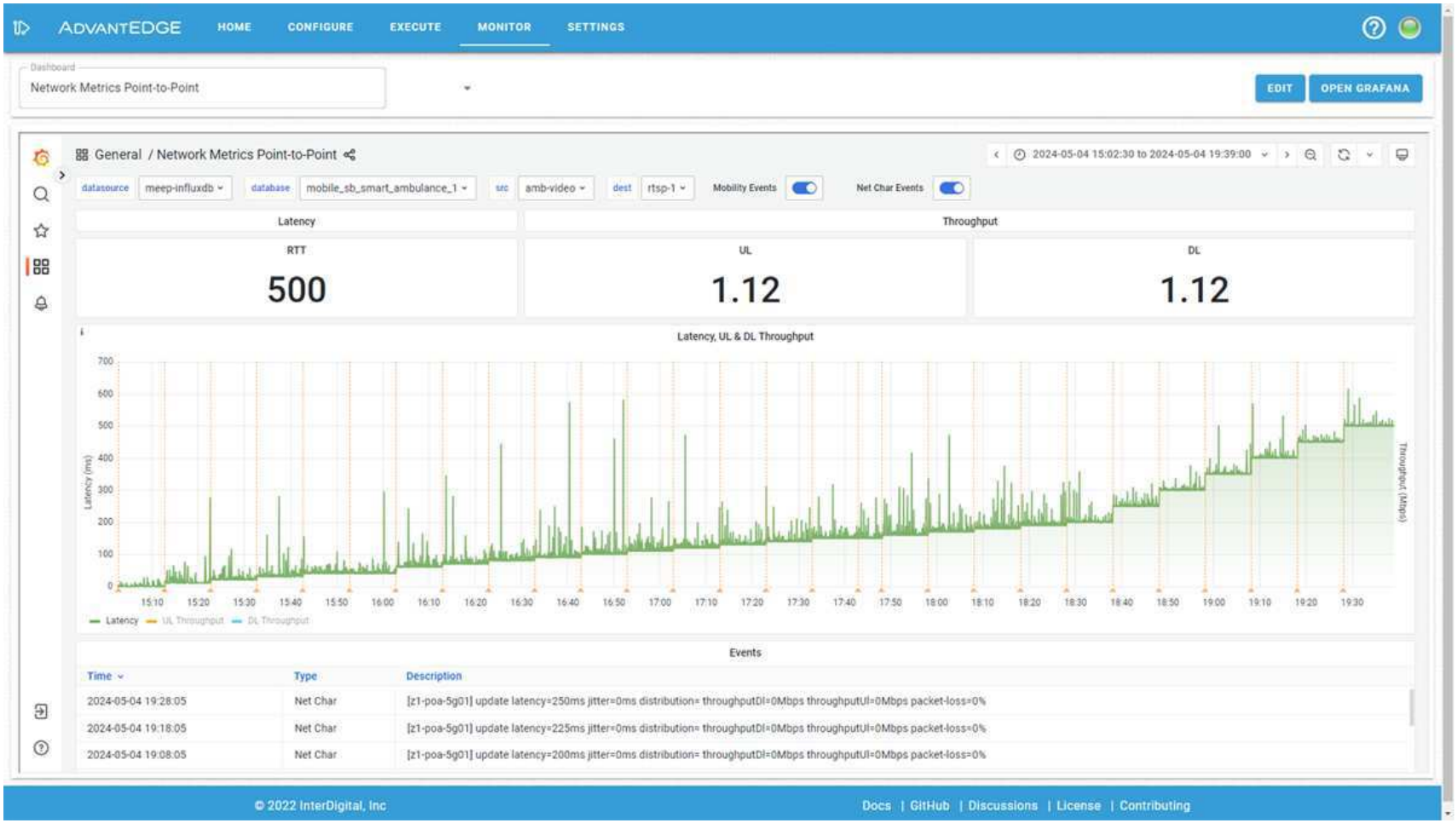
Fonte: Produzida pelo autor.

Figura 7.14: Captura de tela do AdvantEDGE para a Visão de Monitoramento de Métricas de Rede - Uplink FULL HD.



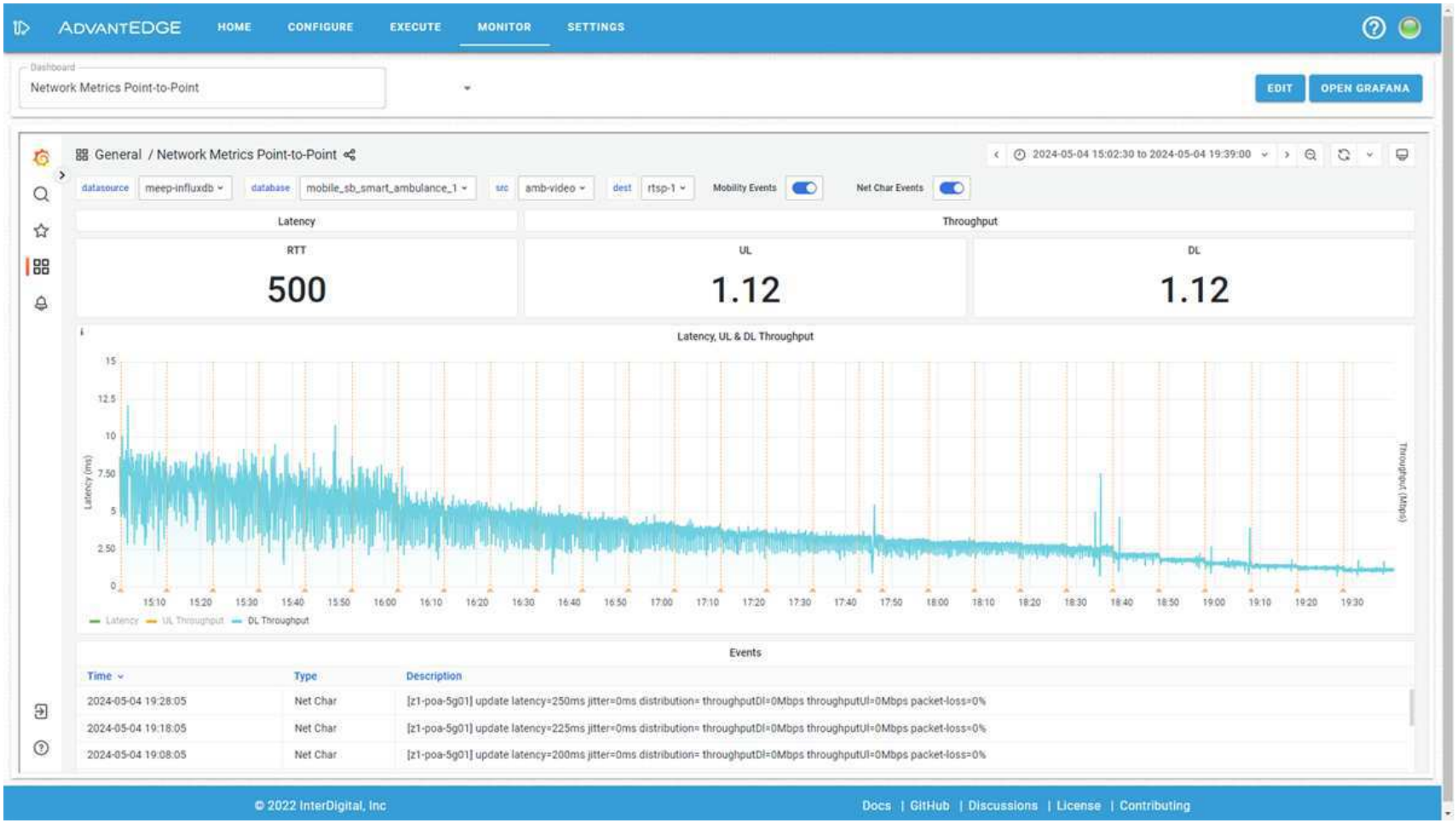
Fonte: Produzida pelo autor.

Figura 7.15: Captura de tela do AdvantEDGE para a Visão de Monitoramento de Métricas de Rede - Latência 4K.



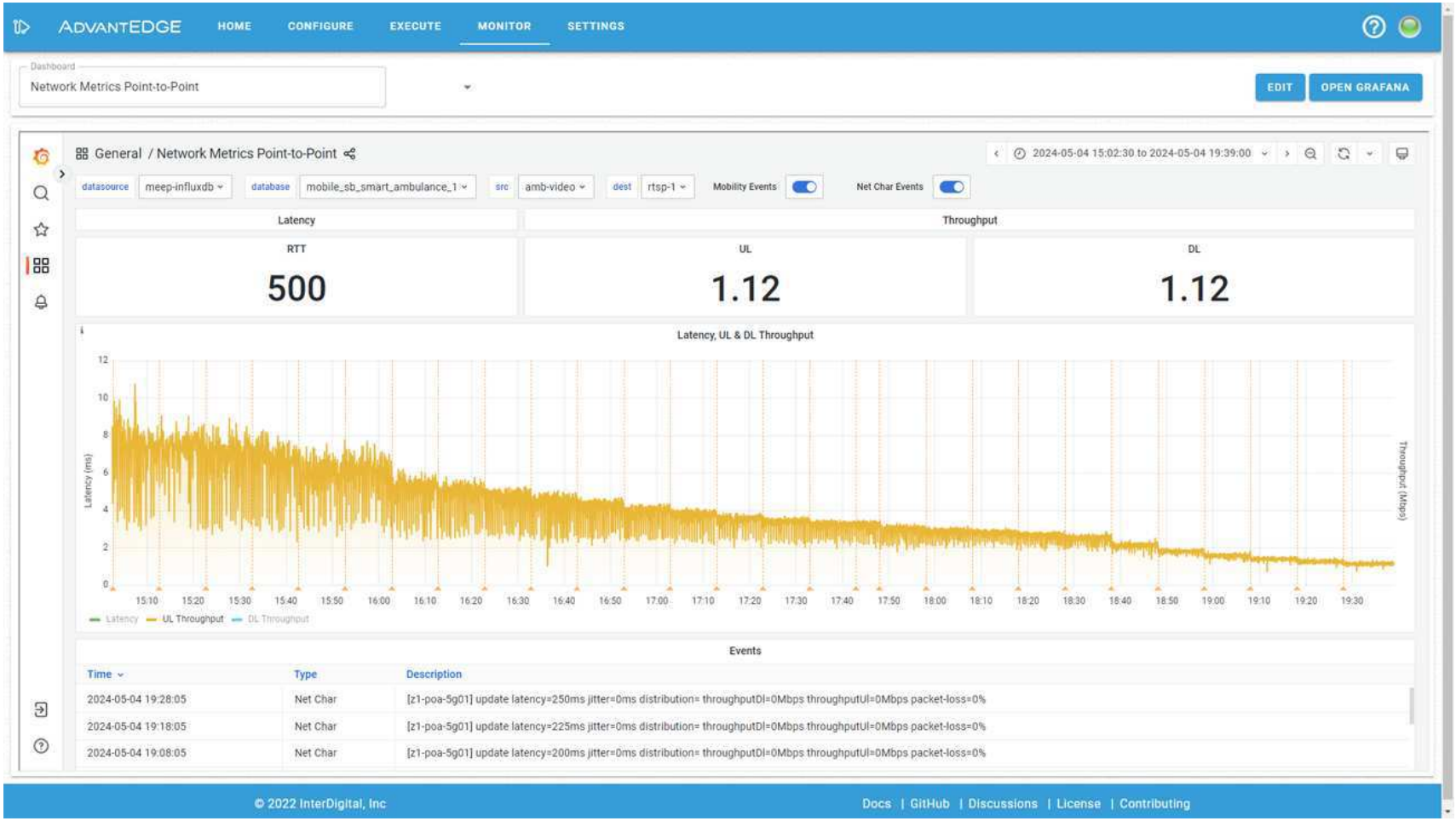
Fonte: Produzida pelo autor.

Figura 7.16: Captura de tela do AdvantEDGE para a Visão de Monitoramento de Métricas de Rede - Downlink 4K.



Fonte: Produzida pelo autor.

Figura 7.17: Captura de tela do AdvantEDGE para a Visão de Monitoramento de Métricas de Rede - Uplink 4K.



Fonte: Produzida pelo autor.

Figura 7.18: Captura de tela do Grafana para utilização de CPU no Cenário 1.

Fonte: Produzida pelo autor.

Figura 7.19: Captura de tela do Grafana para utilização de Memória no Cenário 1.

Fonte: Produzida pelo autor.

7.1.1.5 Discussão

A partir dos experimentos realizados para este cenário, foi possível reproduzir o problema da latência para ambientes de computação em nuvem. Os resultados indicaram que, à medida que a latência aumenta, observa-se frequentemente uma diminuição correspondente na taxa de transmissão. Esse aumento na latência pode ser causado por diversos fatores, como congestionamento de rede, problemas de roteamento, ou distâncias físicas maiores entre a origem e o destino dos dados. Quando a latência é alta, os pacotes de dados levam mais tempo para serem entregues, resultando em uma taxa de transmissão efetiva mais baixa.

Este impacto na taxa de transmissão é particularmente crítico para aplicações sensíveis ao tempo, como streaming de vídeos em aplicações de AR para o cenário de ambulâncias conectadas, onde um atraso significativo pode degradar a experiência do usuário de maneira notável. Em tais cenários, não apenas a fluidez da interação é afetada, mas também a percepção de imediatismo e responsividade, que são cruciais para a satisfação do usuário.

Dado que a experiência do usuário em muitas aplicações modernas é extremamente sensível à latência, é crucial implementar soluções que minimizem este atraso. Uma abordagem é a utilização de ambientes de Computação em Borda, onde os dados são processados o mais próximo possível do usuário final, reduzindo a distância que os dados precisam trafegar na rede e, consequentemente, a latência.

Essa estratégia de disponibilização de serviços em ambientes de computação em nuvem, embora simplifique a configuração inicial e reduza a complexidade operacional, pode não ser ideal para cenários onde a carga demanda é altamente variável ou onde a qualidade da experiência do usuário é crítica. Portanto, é essencial avaliar cuidadosamente as implicações da disponibilização de serviços críticos a partir de servidores de computação em nuvem, especialmente em ambientes que exigem alta disponibilidade e desempenho.

7.1.2 Cenário 2: Alocação Estática de Serviços de Streaming de Vídeo nos servidores na Borda contemplados ao longo do trajeto

Uma solução alternativa ao problema explicitado no Cenário 1 é a implantação dos serviços de Streaming de Vídeo e Dados em Servidores na Borda da Rede. Neste cenário é considerada a alocação dos serviços em todos os servidores de borda contemplados ao longo

do trajeto. Neste contexto, a implantação de serviços em servidores localizados fisicamente mais próximos aos pontos de utilização emerge como uma estratégia promissora para a diminuição da latência. Este arranjo facilita um processamento de dados acelerado e uma redução significativa no tempo de resposta em aplicações críticas.

No entanto, a implementação deste modelo enfrenta desafios substanciais. Primeiramente, a incerteza relacionada ao trajeto exato de veículos em missões críticas, como ambulâncias, representa uma complexidade notável. O percurso de uma ambulância pode ser alterado sem aviso prévio devido a variáveis não previstas, como condições de tráfego intensificado ou o surgimento de emergências adicionais. Esta variabilidade no trajeto pode complicar a alocação eficiente de recursos de processamento, visto que a previsão exata do servidor de borda mais apropriado para atender a demanda torna-se um processo incerto.

Além disso, existe uma limitação técnica considerável nos recursos disponíveis nos servidores de borda. Estes servidores, apesar de sua utilidade em reduzir a latência, não possuem a mesma capacidade de processamento ou de armazenamento que os servidores centrais localizados em infraestruturas de nuvem. Esta restrição impõe limites à quantidade e à complexidade dos dados que podem ser processados localmente, o que pode comprometer a eficácia da solução em cenários que exigem um processamento intensivo de dados.

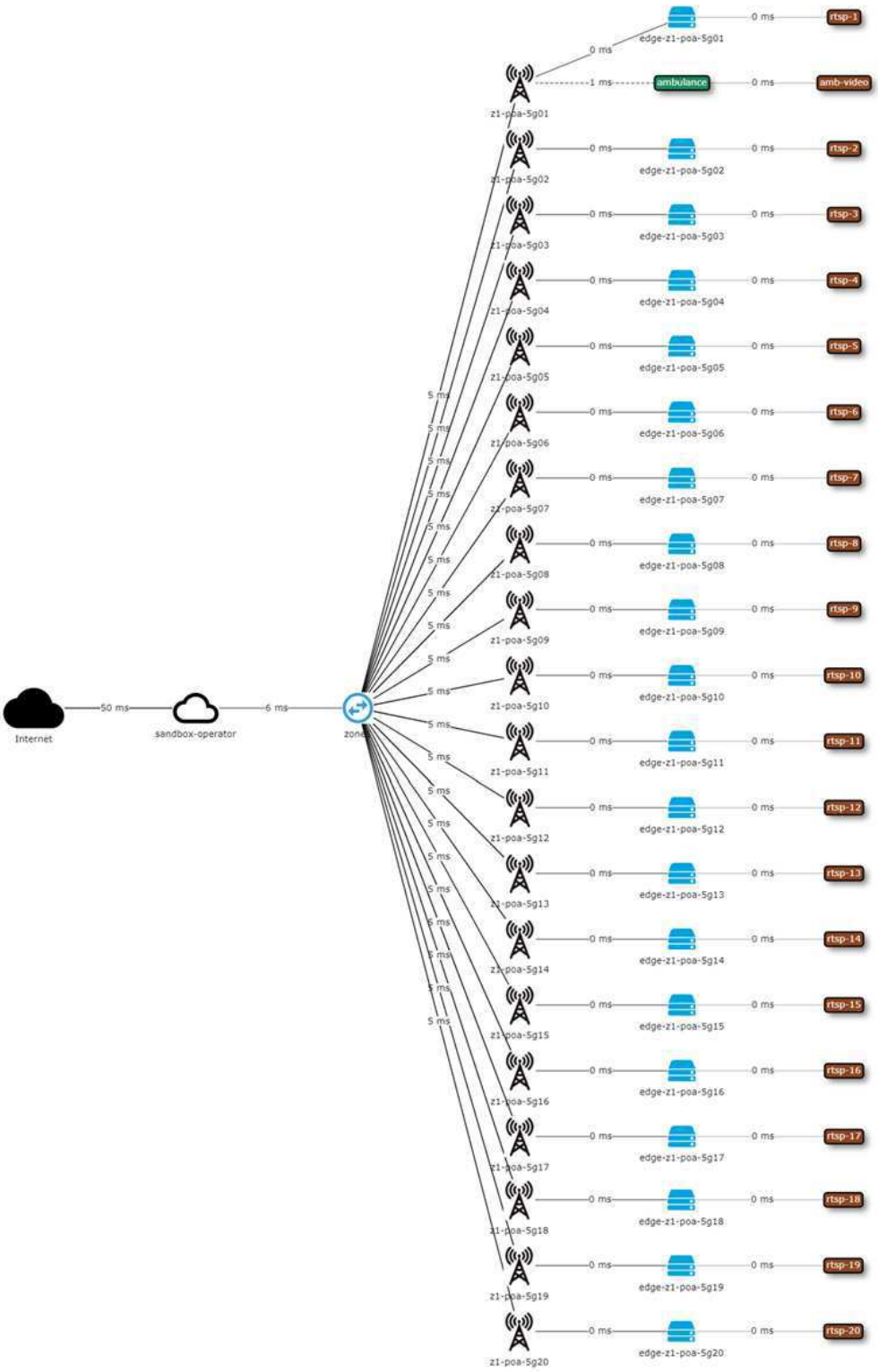
Desse modo, tomando como base o Cenário 1, foram adicionadas elementos do tipo FOG para representar os servidores na borda vinculados a cada antena. Na Figura 7.20 está ilustrado o diagrama de rede atualizado para o Cenário 2.

Adicionalmente, no contexto da distribuição georreferenciada, observa-se que cada servidor de borda foi estrategicamente localizado próximo às antenas correspondentes, o que é evidenciado na Figura 7.21. Esta configuração visa otimizar a eficiência da transmissão de dados e reduzir a latência. A disposição geográfica dos servidores de borda próximo às antenas facilita uma comunicação mais rápida e eficiente, crucial para a operacionalidade do sistema. A Figura 7.21 oferece uma representação visual detalhada dessa estratégia de posicionamento, destacando a importância da localização na infraestrutura de rede.

7.1.2.1 Resultados

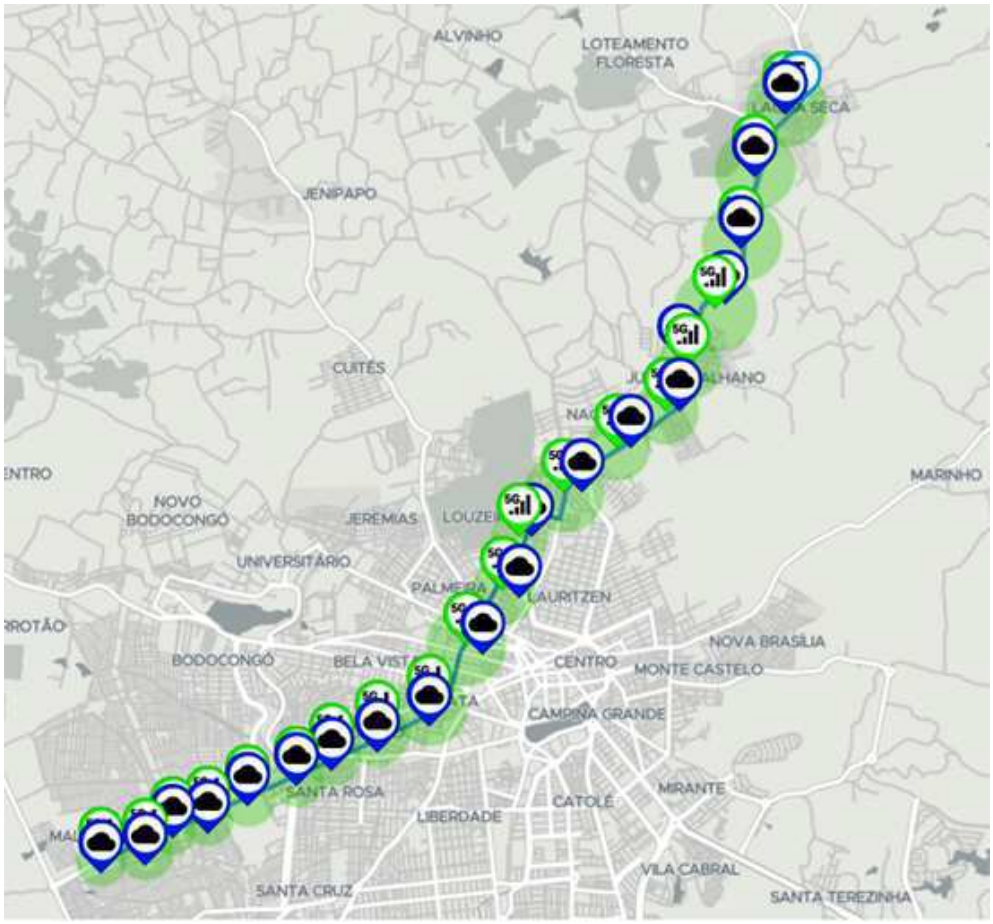
A implementação de serviços de Streaming de Vídeo e Dados em Servidores de Borda, conforme delineado no Cenário 2, representa uma abordagem significativa na busca por eficiên-

Figura 7.20: Diagrama de rede para o Cenário 2.



Fonte: Produzida pelo autor.

Figura 7.21: Mapa do cenário com destaque para a distribuição das antenas dos servidores em borda ao longo do trajeto.



Fonte: Produzida pelo autor.

cia em transmissão e processamento de dados em tempo-real. A alocação de tais serviços em todos os servidores de borda ao longo do trajeto pretendido visa diminuir a latência de comunicação, especialmente em aplicações críticas onde o tempo de resposta é crucial.

Por outro lado, esta estratégia impacta diretamente o consumo de memória nos servidores de borda, conforme observa-se na Figura 7.23. Dados coletados durante a implementação indicam um aumento substancial no uso de memória, uma consequência direta da necessidade de armazenar e gerenciar grandes volumes de dados de vídeo e outros tipos de conteúdo digital em tempo real. Caso seja utilizada uma estratégia de alocação dinâmica, espera-se uma redução de aproximadamente *(n-2)* vezes a quantidade de servidores, onde os 2 servidores ativos seriam o atual (atendendo as demandas da aplicação no servidor mais perto) e o próximo (visando antecipar a alocação para que o serviço já esteja disponível no momento

em que a ambulância chegue na respectiva localização).

Esta elevação no consumo de memória pode ser atribuída à necessidade de cada servidor de borda manter uma instância completa dos serviços de Streaming de Vídeo e Dados para garantir uma transmissão contínua e sem interrupções. Além disso, a redundância de dados, necessária para manter a integridade e a disponibilidade dos serviços em face de possíveis falhas de hardware ou de rede, também contribui para este aumento no uso de recursos de memória, sendo necessário uma avaliação minuciosa da necessidade de redundância para cada aplicação.

Em relação ao uso de CPU, a distribuição dos serviços nos servidores de borda, conforme observado na Figura 7.22, demonstra a utilização eficiente em termos de consumo de recursos computacionais. Observou-se que o percentual de uso da CPU nos servidores de borda aumenta proporcionalmente com o número de requisições recebidas. Esta correlação indica que a alocação dos serviços de Streaming de Vídeo e Dados, apesar de estarem alocados em todos os servidores em borda, é capaz de responder dinamicamente à demanda, sem causar um uso excessivo ou contínuo de CPU em momentos de baixa demanda. Na Figura 7.24 está ilustrado um comparativo dos consumos de recursos computacionais (CPU e Memória) entre os Cenários 1 e 2.

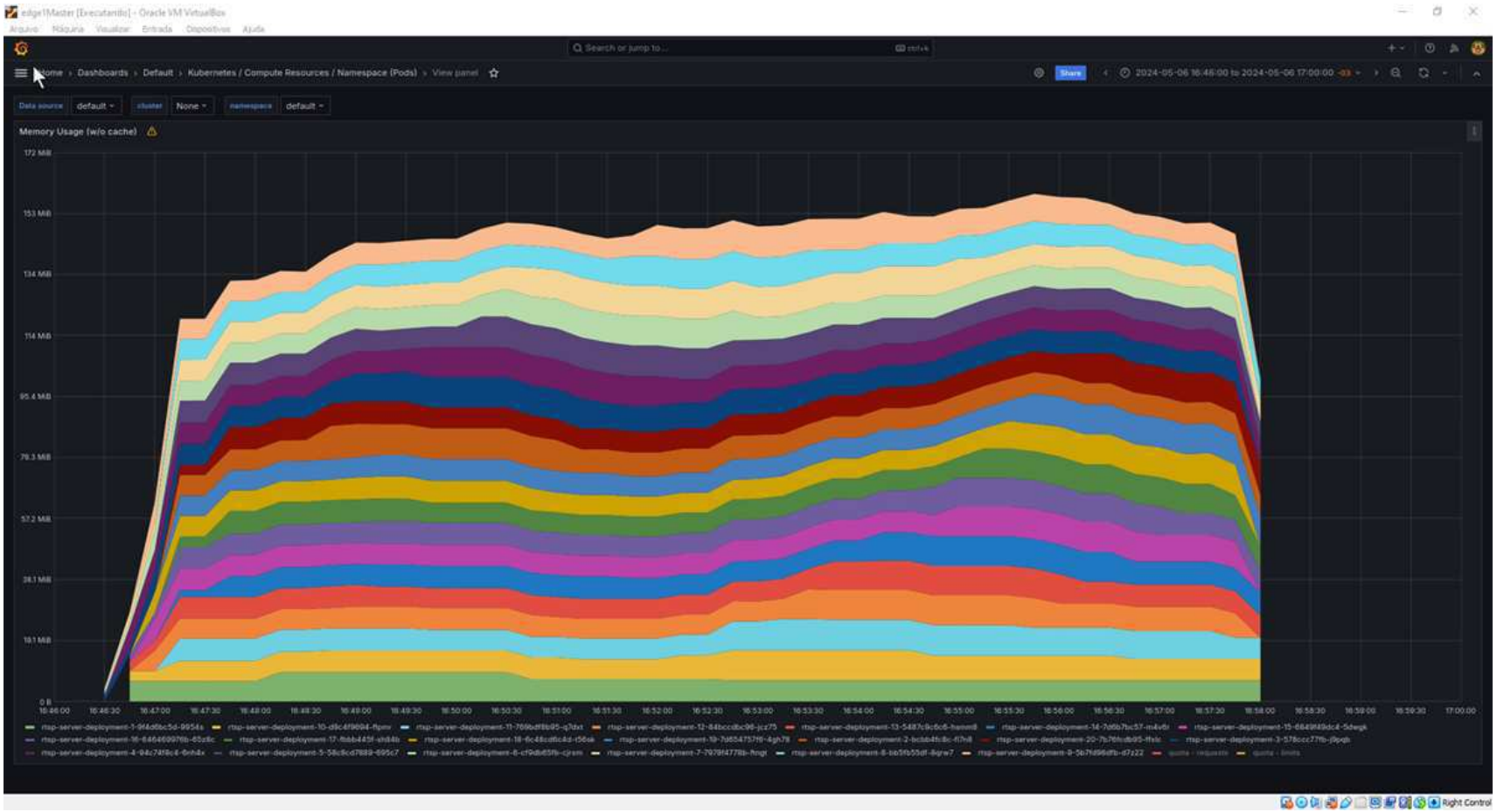
Para viabilizar a melhor visualização da redução da latência ao disponibilizar-se os serviços nos servidores mais próximos da ambulância, o mesmo cenário foi executado com a redução da velocidade da ambulância, de modo que a transição entre estações rádio-base fosse mais lenta, resultando em um maior espaçamento entre as transições ao longo do tempo. Na Figura 7.25 estão ilustrados os gráficos das latências entre a ambulância e os serviços RTSP disponibilizados nos respectivos servidores de borda. Adicionalmente, no gráfico de barras está representada a latência do servidor **rtsp-04**, onde a redução na latência, entre os instantes 23:10 e 23:14, aproximadamente, está relacionada ao período cujo a ambulância estava conectada à respectiva estação rádio-base.

Além disso, na Figura 7.26 estão ilustradas as latências médias registradas no momento inicial da simulação. Neste cenário específico, a ambulância estabelece uma conexão com a primeira estação rádio-base, acessando o serviço denominado **rtsp-1**. Esta representação gráfica fornece uma visualização quantitativa das latências médias envolvidas, que são fundamentais para avaliar a eficiência da comunicação em tempo real requerida pela aplicação.

Figura 7.22: Captura de tela do Grafana para utilização de CPU no Cenário 2.

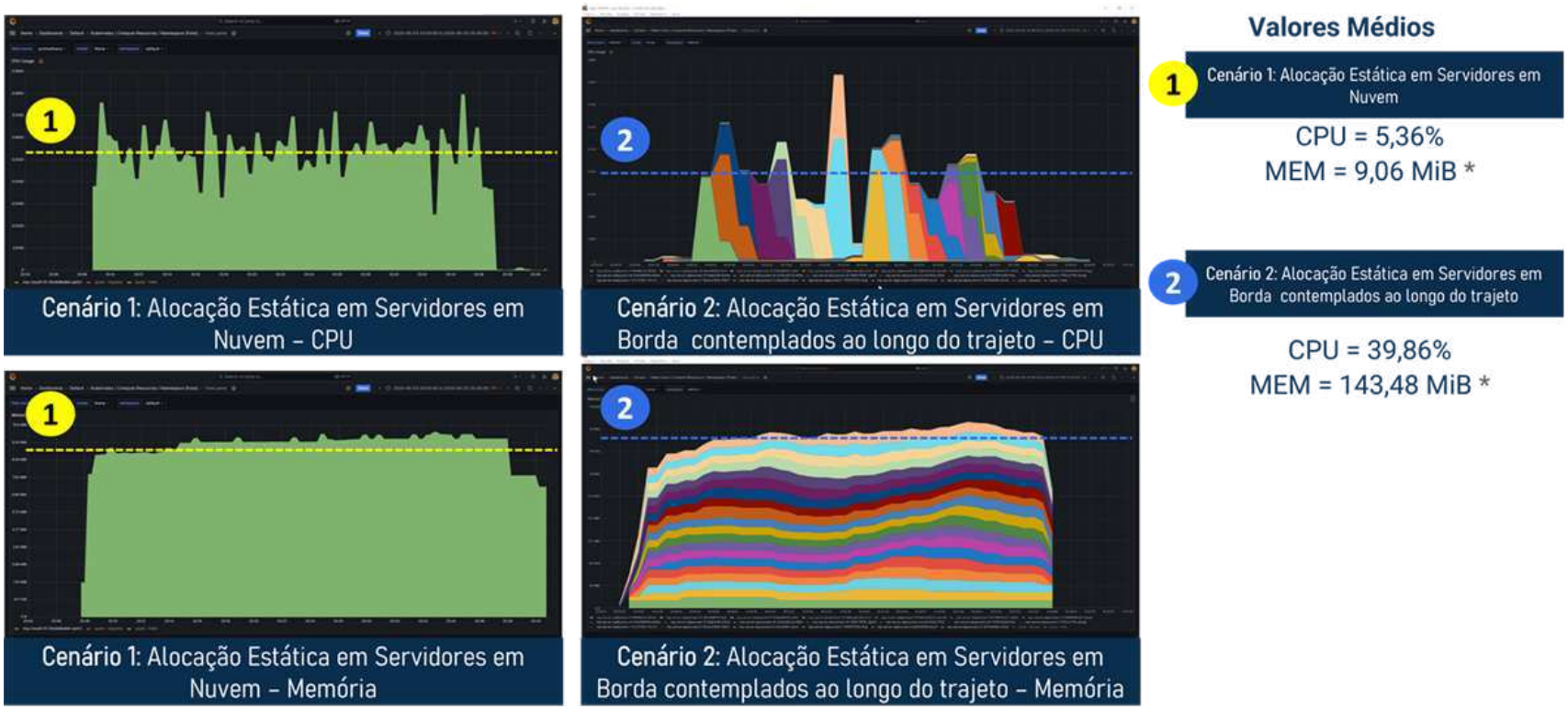
Fonte: Produzida pelo autor.

Figura 7.23: Captura de tela do Grafana para utilização de Memória no Cenário 2.



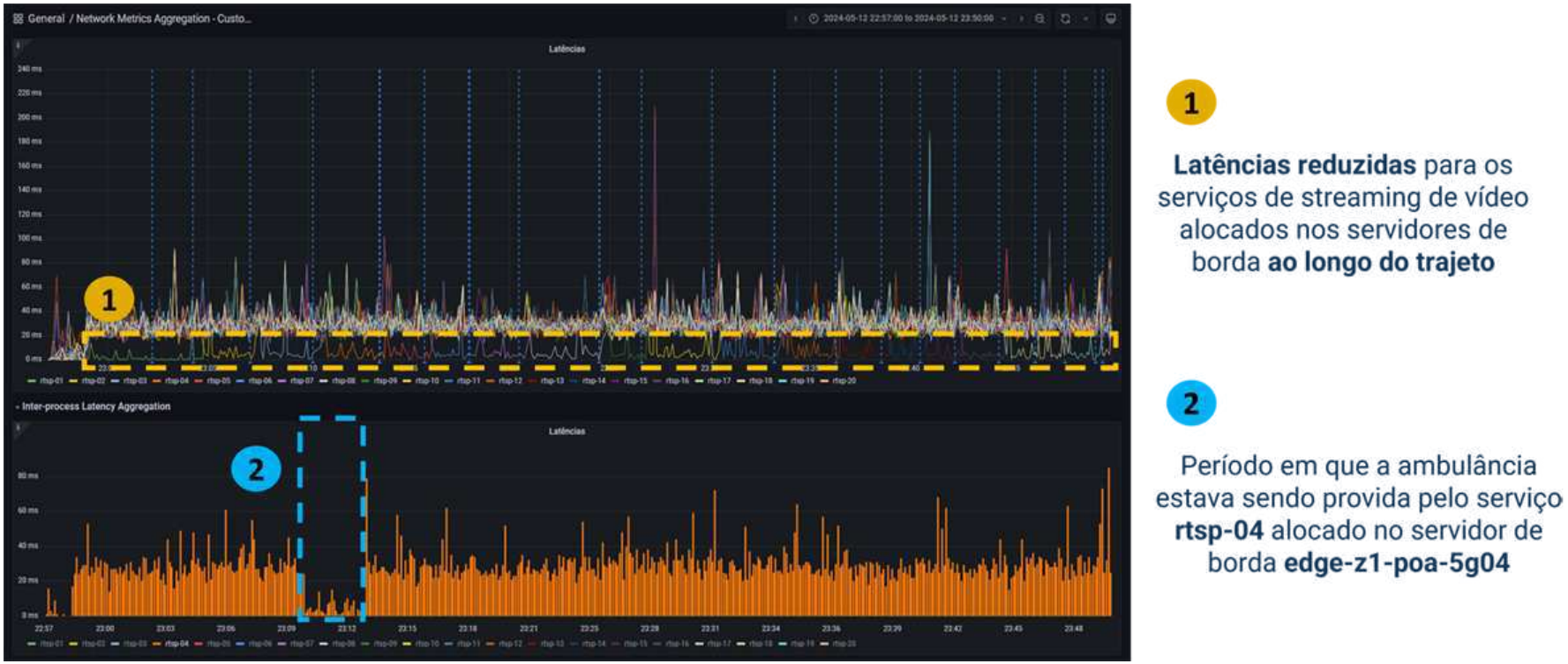
Fonte: Produzida pelo autor.

Figura 7.24: Comparativo de consumo de recursos computacionais para os cenários 1 e 2.



Fonte: Produzida pelo autor.

Figura 7.25: Captura de tela do Grafana com as latências da simulação do Cenário 2.



Fonte: Produzida pelo autor.

Figura 7.26: Captura de tela do Grafana com as latências médias no início da simulação do Cenário 2.

Fonte: Produzida pelo autor.

7.1.2.2 Discussão

Apesar dos benefícios em termos de redução de latência, o cenário expõe limitações significativas relacionadas à capacidade dos servidores de borda. Estes dispositivos, embora eficazes na aceleração do processamento de dados próximo aos pontos de uso, não possuem a mesma capacidade de armazenamento ou processamento dos servidores centrais em infraestruturas de nuvem. Portanto, a implementação deste modelo requer uma avaliação criteriosa do balanço entre a capacidade de recursos computacionais disponíveis e as demandas de processamento dos serviços de dados distribuídos.

Em resumo, enquanto a estratégia de alocação distribuída de serviços nos servidores de borda facilita melhorias na latência e na eficiência de processamento de dados, ela impõe desafios significativos ao consumo de recursos computacionais dos servidores. A gestão eficiente deste recurso é fundamental para sustentar a viabilidade operacional do sistema em longo prazo.

7.2 Validação do Método para a Prova de Conceito Definida

Uma vez validado o problema experimentalmente, foram definidos mais três cenários para validação experimental do método proposto. Considerando que serão executados diversos serviços em Servidores na Borda, a limitação de recursos implicar no compartilhamento de recursos computacionais, influenciando desde na latência de processamento até a disponibilidade. Desse modo, alocar serviços dinamicamente ao longo de um trajeto pode trazer ganhos associados à utilização de recursos em Servidores na Borda, devendo-se ponderar o custo associado à disponibilidade de recursos computacionais e a sua utilização, de fato.

Desse modo, para a validação da solução proposta, foram considerados três cenários:

- Cenário 3: É o cenário "ótimo", em que se conhece a rota exata da ambulância, viabilizando maior precisão na definição do servidor subsequente;
- Cenário 4: Utiliza a estratégia de alocar em todas as bordas vizinhas, garantindo a disponibilidade do serviço, qualquer que seja a direção que a ambulância siga;

- Cenário 5: Neste cenário, identifica-se o padrão de mobilidade a partir de um algoritmo eurístico implementado com base na topologia da rede e no histórico de deslocamento da ambulância.

Além disso, para estes cenários foram definidos *templates* para carregamento dinâmico de conteiners, conforme será descrito a seguir.

7.2.1 Template de Carregamento Dinâmico

A utilização dessa abordagem está em reduzir o tempo de carregamento dos serviços ao longo do trajeto da ambulância na infraestrutura virtualizada. Para isso, foram criados arquivos *template* contemplanto toda a configuração necessária para implantação de serviços relacionados à aplicação de 'ambulâncias conectadas', tais como, servidores RTSP, Brokers MQTT, entre outros. Na Lista 1 está apresentado o código fonte do arquivo de configuração para servidores RTSP.

Na definição do *template*, são utilizadas *tags* que viabilizam o recebimento de informações dinâmicas do ambiente (por exemplo, servidor responsável pela disponibilização do serviço, porta de acesso a ser utilizada etc) e a partir disso são gerados, em tempo de execução, os arquivos com as configurações necessárias. Na Lista 2 está apresentado o código fonte da função de implantação para os servidores RTSP gerados a partir do *template* apresentado na Lista 1. Na Figura 7.27 está ilustrada a captura de tela do VSCode com os arquivos de implantação gerados para o Cenário 2, ou seja, a implantados no início da simulação.

7.2.1.1 Discussão

A implementação do carregamento dinâmico de serviços utilizando arquivos *template* demonstrou ser uma estratégia factível para a implantação de serviços em infraestruturas virtualizadas no cenário de ambulâncias conectadas. Os *templates* facilitaram a configuração automática e adaptável de servidores RTSP e Brokers MQTT. A inclusão de *tags* dinâmicas nos *templates* permitiu que as configurações dos serviços se ajustassem automaticamente às variações do ambiente, como mudanças nas configurações de acesso à rede ou na localização dos servidores próximos à ambulância, agilizando o processo de carregamento dos serviços.

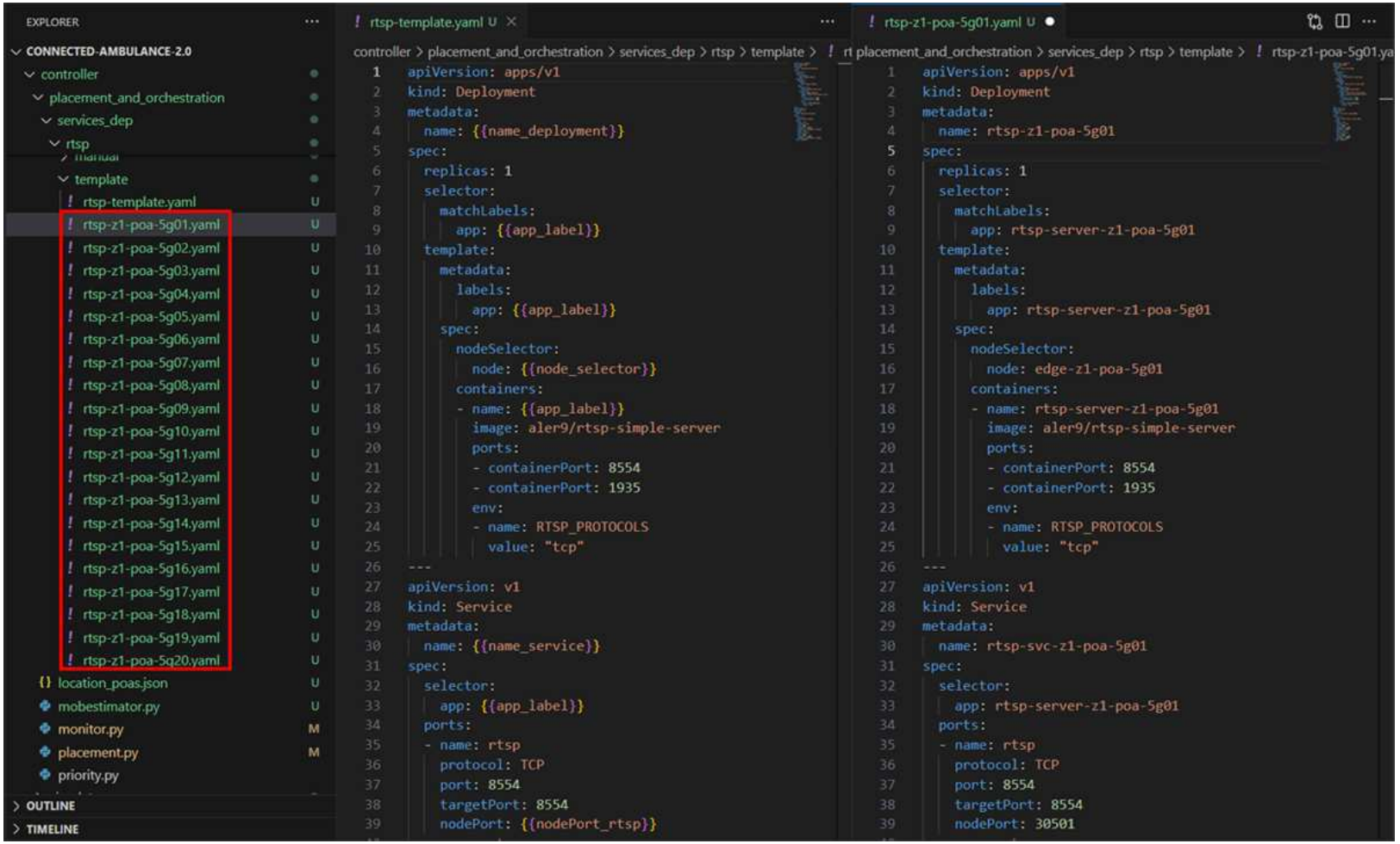
```
apiVersion: apps/v1
kind: Deployment
metadata:
  name: {{name_deployment}}
spec:
  replicas: 1
  selector:
    matchLabels:
      app: {{app_label}}
  template:
    metadata:
      labels:
        app: {{app_label}}
    spec:
      nodeSelector:
        node: {{node_selector}}
      containers:
      - name: {{app_label}}
        image: aler9/rtsp-simple-server
        ports:
        - containerPort: 8554
        - containerPort: 1935
        env:
        - name: RTSP_PROTOCOLS
          value: "tcp"
---
apiVersion: v1
kind: Service
metadata:
  name: {{name_service}}
spec:
  selector:
    app: {{app_label}}
  ports:
  - name: rtsp
    protocol: TCP
    port: 8554
    targetPort: 8554
    nodePort: {{nodePort_rtsp}}
  - name: rtmp
    protocol: TCP
    port: 1935
    targetPort: 1935
    nodePort: {{nodePort_rtmp}}
  type: NodePort

```

Listing 1: Código fonte to *template* para implantação dinâmica de serviços.

```python
def deploy_rtsp_server_template(server_name, port_rtsp, port_rtmp, command):

    # Implantação de serviços ###############
    current_directory =
        os.path.dirname(os.path.abspath(__file__)) +
        '\\services_dep\\rtsp\\template'

    template_path = os.path.join(
            current_directory,
            "rtsp-template.yaml"
    )

    file_path = os.path.join(
            current_directory,
            f"rtsp-{server_name}.yaml"
    )

    # Carregar o template
    with open(template_path, 'r') as file:
        template = file.read()

    mapping = {
        '{{name_deployment}}': f'rtsp-{server_name}',
        '{{app_label}}': f'rtsp-server-{server_name}',
        '{{node_selector}}': f'{server_name}',
        '{{name_service}}': f'rtsp-svc-{server_name}',
        '{{nodePort_rtsp}}': f'{port_rtsp}',
        '{{nodePort_rtmp}}': f'{port_rtmp}'
    }

    for key, value in mapping.items():
        template = template.replace(key, value)

    # Salvar o arquivo YAML modificado
    with open(file_path, 'w') as file:
        file.write(template)


    kubectl_apply_command =
        f"kubectl {command} -f {file_path}
        --grace-period=0 --force > nul 2>&1"

    os.system(kubectl_apply_command)

```

Listing 2: Código fonte da função de implantação dinâmica de serviços.

Figura 7.27: Captura de tela do VSCode com os arquivos de implantação gerados para o Cenário 2.



Fonte: Produzida pelo autor.

7.2.2 Cenário 3: Alocação Dinâmica de Serviços de Streaming de Vídeo nos servidores na Borda, conhecendo-se a trajetória

Para este cenário, a configuração de rede e a geolocalização implementadas no Cenário 2 foram replicadas. A diferença principal entre os dois cenários reside na estratégia adotada para a alocação dos serviços. Uma vez atrelada ao padrão de mobilidade, a alocação dinâmica é capaz de melhorar eficiência na utilização dos recursos, através da redução no consumo de memória, ainda atendendo a capacidade de resposta do sistema às necessidades dos usuários em cenários críticos.

7.2.2.1 Resultados

A validação da solução proposta focou em examinar a eficácia da alocação dinâmica de serviços nos servidores de borda, particularmente em relação à utilização de recursos de memória. Os resultados demonstraram que a alocação dinâmica, ajustada de acordo com o padrão de mobilidade da ambulância, resultou em uma redução significativa no consumo de memória nos servidores, conforme apresentado na figura 7.29. Essa otimização é atribuída à capacidade do sistema de alocar os recursos com base na demanda da ambulância, ao invés de manter as aplicações alocadas e distribuída em todos os servidores, que geralmente resulta em subutilização ou sobrecarga em pontos específicos da rede. Na Figura 7.30 está ilustrado um comparativo dos consumos de memória entre os Cenários 2 e 3.

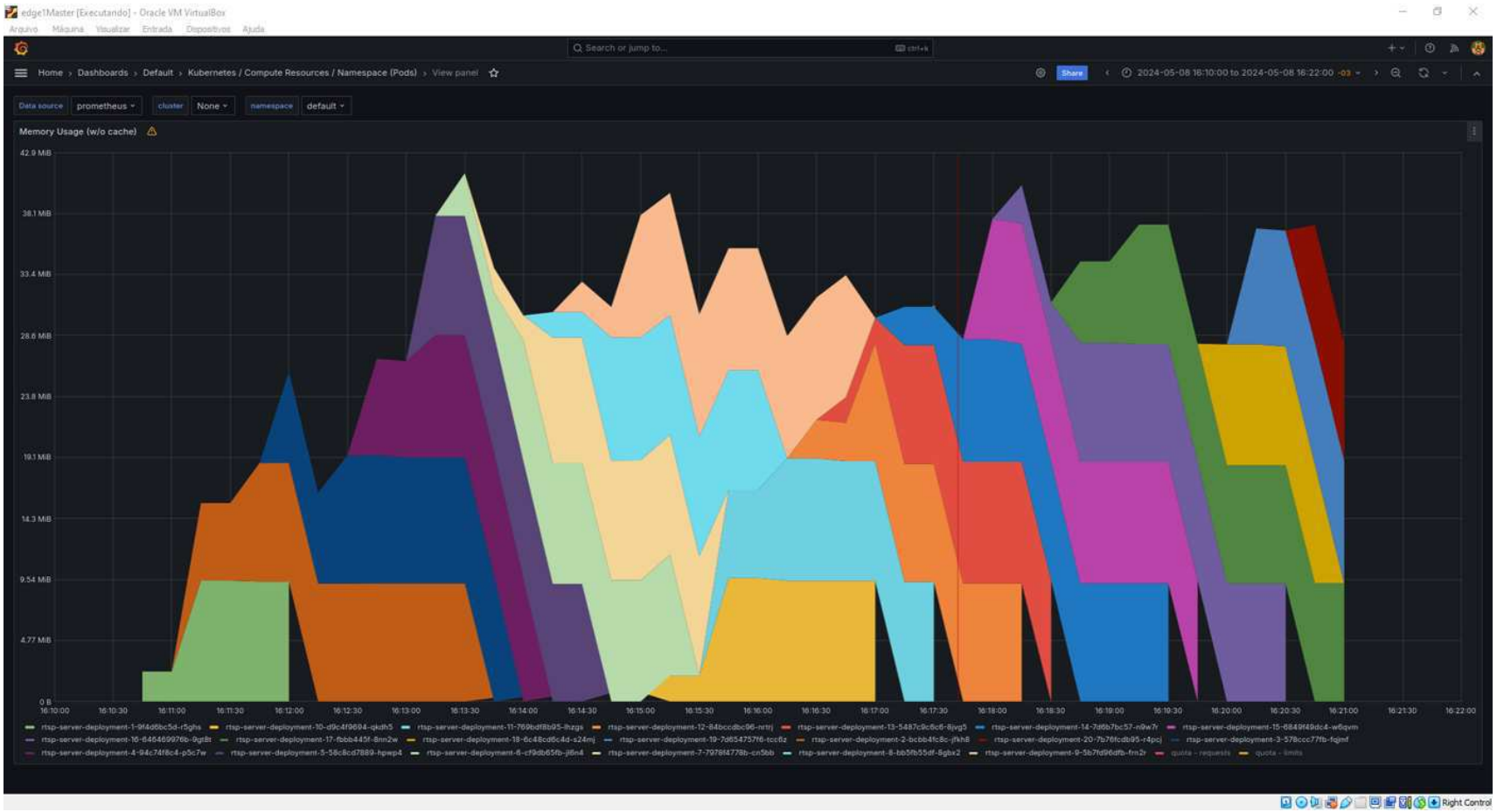
A partir da análise do uso de CPU nos servidores de borda, conforme apresentada na Figura 7.28, não foram identificadas alterações significativas ao comparar as simulações dos Cenários 2 e 3. Esta constância no consumo de CPU sugere que a estratégia de alocação foi eficaz na distribuição uniforme da carga de processamento, evitando flutuações excessivas que poderiam afetar negativamente o desempenho do sistema.

A implementação dessa estratégia implica em ganhos relacionados à limitação de recursos nos servidores de borda, contribuindo para a manutenção da latência de processamento dentro dos parâmetros de aplicações críticas de saúde, uma vez que são executados em ambientes de computação na borda. Os dados coletados durante a simulação indicam que, comparativamente à alocação estática de serviços, a alocação dinâmica reduz o consumo de memória, promovendo uma solução mais eficiente de gestão de recursos e reduzindo os custos operacionais relacionados ao consumo de energia e manutenção.

Figura 7.28: Captura de tela do Grafana para utilização de CPU no Cenário 3.

Fonte: Produzida pelo autor.

Figura 7.29: Captura de tela do Grafana para utilização de Memória no Cenário 3.



Fonte: Produzida pelo autor.

Figura 7.30: Comparativo de consumo de memória para os cenários 2 e 3.

Fonte: Produzida pelo autor.

7.2.2.2 Discussão

A análise dos resultados obtidos na validação da solução proposta reforça a viabilidade da alocação dinâmica de serviços em servidores de borda, especialmente em cenários que exigem respostas rápidas e eficientes, como é o caso das aplicações críticas de saúde. A redução significativa no consumo de memória, evidenciada na Figura 7.29, destaca a eficácia dessa estratégia em otimizar o uso dos recursos. Este resultado é particularmente relevante, pois demonstra que a alocação dinâmica, ajustada em tempo real de acordo com o padrão de mobilidade e as demandas da ambulância, evita a subutilização e a sobrecarga dos servidores, problemas comuns em sistemas com alocação estática.

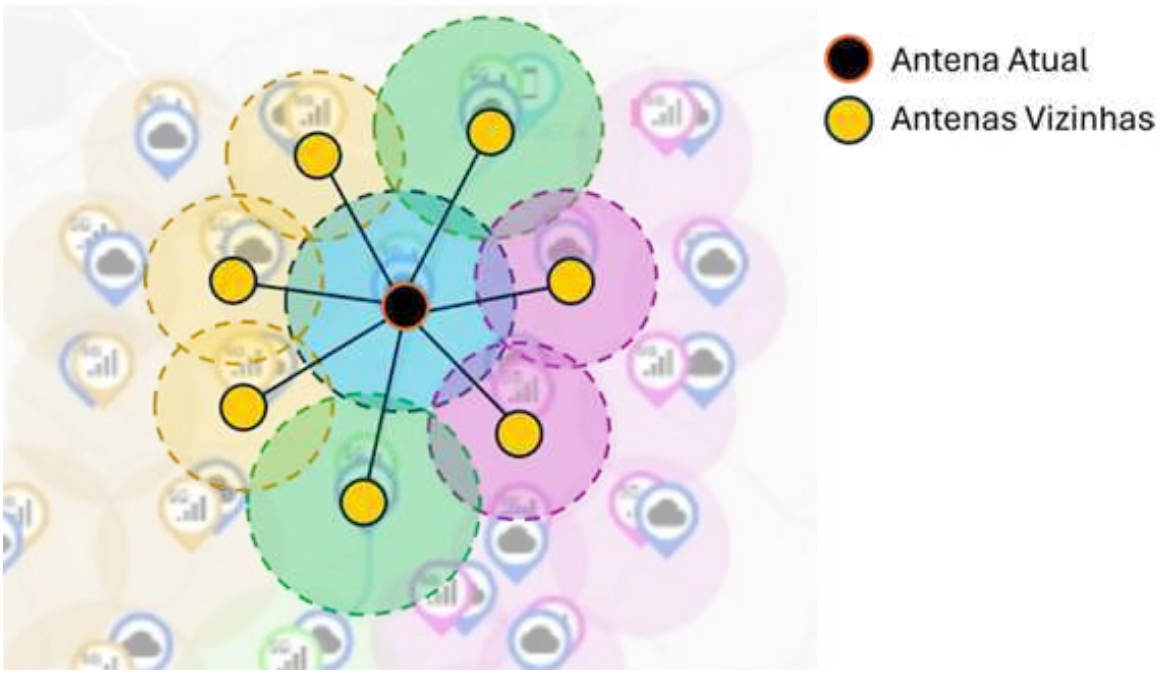
Por fim, os benefícios adicionais da redução o consumo de memória se traduzem em melhorias diretas na eficiência energética e na redução dos custos operacionais, além de prolongar a vida útil do hardware. Estes aspectos são fundamentais para a sustentabilidade de longo prazo de infraestruturas de TI em contextos de saúde, onde a confiabilidade e a eficiência são imperativas. Os resultados, portanto, não apenas confirmam o entendimento inicial de que a alocação dinâmica é necessária para disponibilização de aplicações críticas, mas também fornecem uma base sólida para futuras investigações e desenvolvimentos que visem aprimorar ainda mais o desempenho e a eficácia dos sistemas de computação em borda.

No entanto, considerando o cenário de não conhecer-se o padrão de mobilidade da ambulância, uma solução possível seria disponibilizar os serviços em todos os servidores ao redor da ambulância, servindo como uma estratégia visando garantir a disponibilidade dos serviços críticos. Este cenário será detalhado a seguir.

7.2.3 Cenário 4: Alocação Dinâmica de Serviços de Streaming de Vídeo nos servidores na Borda vizinhos

Conforme detalhado anteriormente, para os casos onde não se conhece a rota da ambulância, um desafio associado está relacionado a como garantir a disponibilidade dos serviços críticos para as ambulâncias. Uma estratégia possível é a alocação em todos os servidores vizinhos, conforme ilustrado na Figura 7.31, ao custo de utilizar-se mais recursos do que o necessário.

Figura 7.31: Estratégia utilizada para o Cenário 4.



Fonte: Produzida pelo autor.

Desse modo, tomando como base o Cenário 3, foram adicionados novos servidores e antenas ao cenário,visando avaliar a utilização dos recursos computacionais nesta possível solução. Na Figura 7.32 está ilustrado o diagrama de rede atualizado para o Cenário 4.

Após isso, foi realizado o mapeamento de rede de todas as antenas com base nas coordenadas geográficas e respectivos raios de alcance. Em termos práticos, essa informação do mapa de rede poderia ser fornecida pelo Operador de Rede ou utilizar-se alguma estratégia de *service discovery* para obtenção automatizada do mapa. De posse desses mapas de rede e utilizando como base a antena cuja ambulância está conectada, os serviços foram alocados nos nós vizinhos e liberados, conforme deslocamento da ambulância ao longo do trajeto.

7.2.3.1 Resultados

As principais métricas de referência utilizada para comparação desses cenários consistiram na avaliação do consumo de memória e taxa de sucesso na disponibilização dos serviços ao longo do trajeto. Com base nessa simulação, observou um aumento no consumo de memória proporcional à alocação em servidores vizinhos. Conforme detalhado anteriormente,

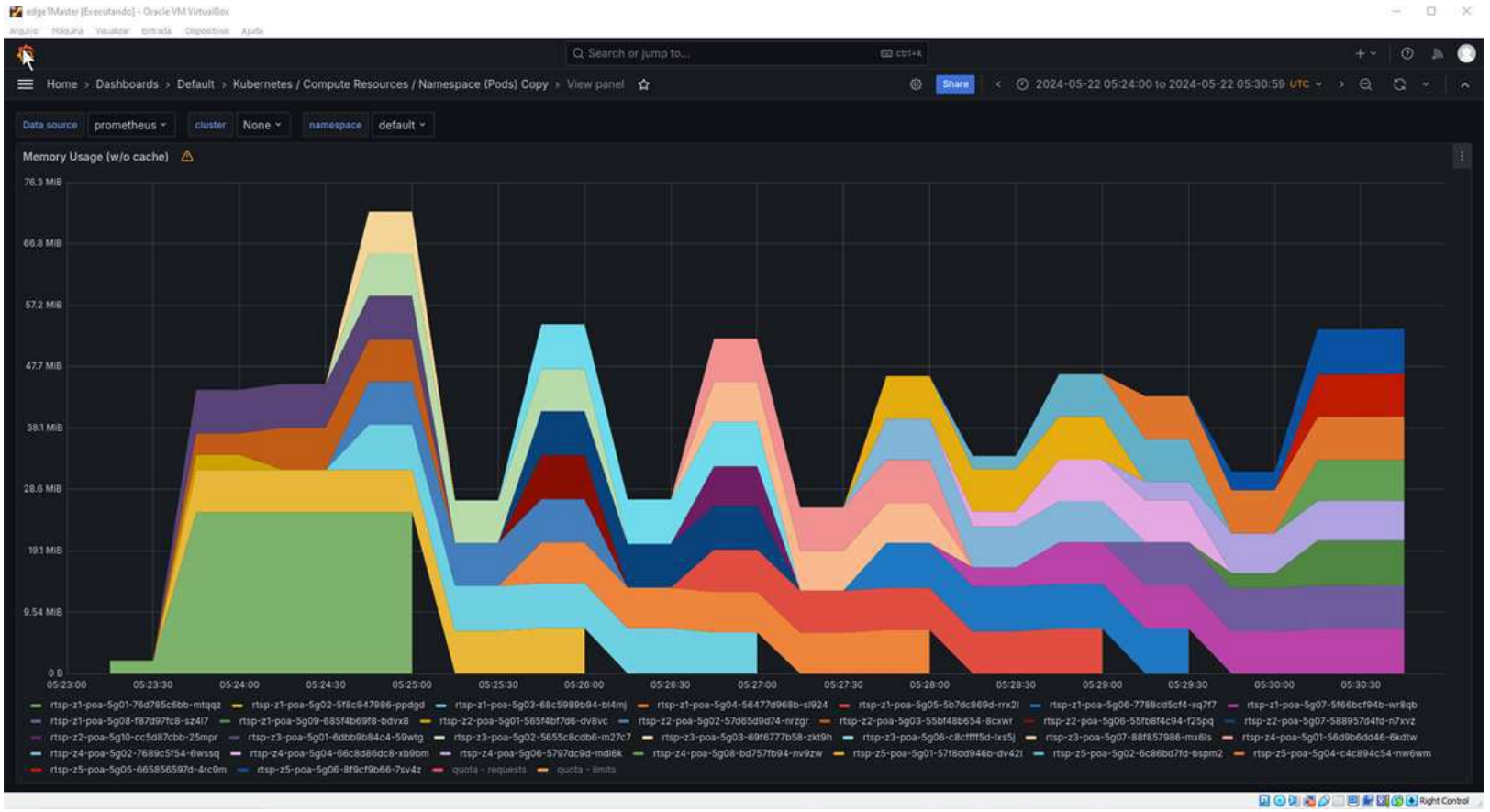
Figura 7.32: Mapa do cenário em ambiente de múltiplas bordas.

Fonte: Produzida pelo autor.

esta resposta estava dentro do esperado, sendo este o custo para garantir a continuidade na disponibilização dos serviços. Na Figura 7.33 está ilustrada a captura de tela do Grafana para utilização de Memória no Cenário 4. Na Figura 7.34 está ilustrado um comparativo dos consumos de memória entre os Cenários 3 e 4.

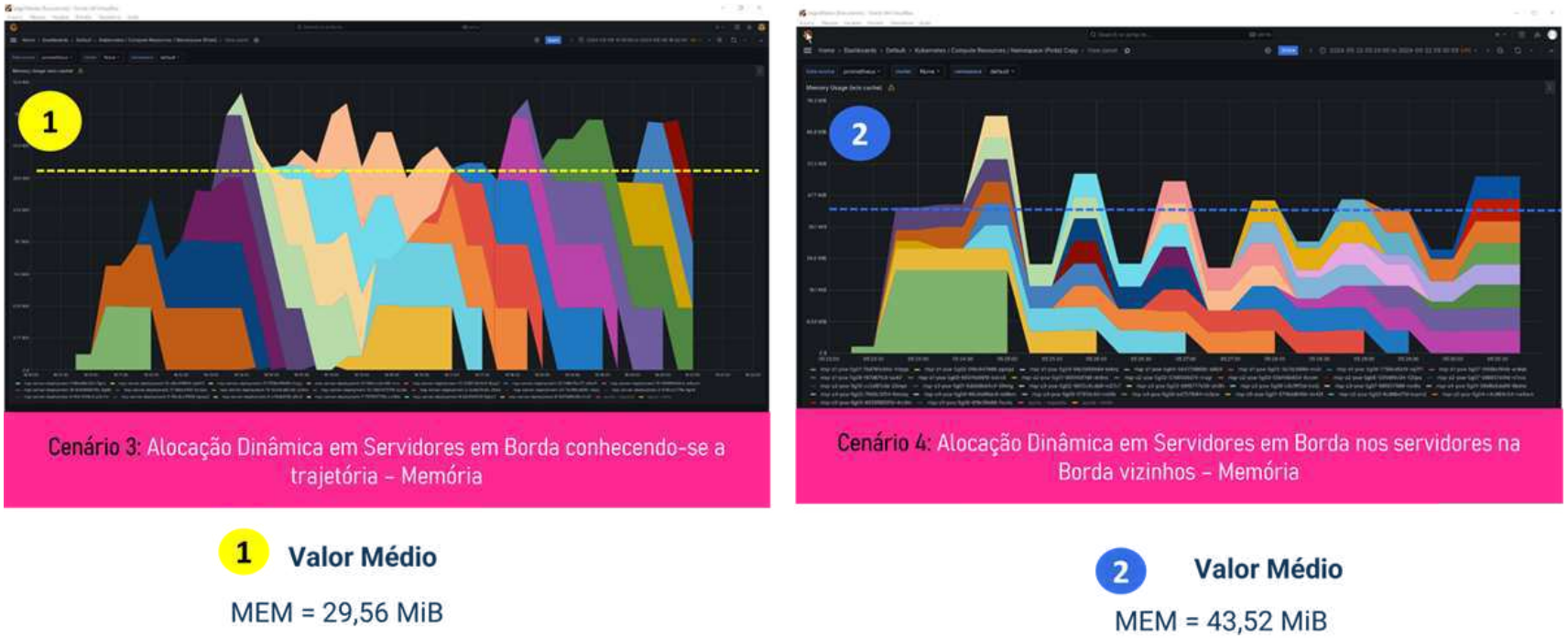
Com relação à taxa de sucesso na disponibilização dinâmica dos serviços, na Figura 7.35 está ilustrado o momento de transição entre os servidores de borda **z1-poa-5g16** e **z1-poa-5g17**, também ilustrado na visão geográfica na Figura 7.37, onde o serviço respectivo ao servidor destino já estava em execução para utilização no momento da chegada da ambulância. Além disso, na Figura 7.36 está ilustrada a liberação, manutenção e alocação de serviços.

Figura 7.33: Captura de tela do Grafana para utilização de Memória no Cenário 4.



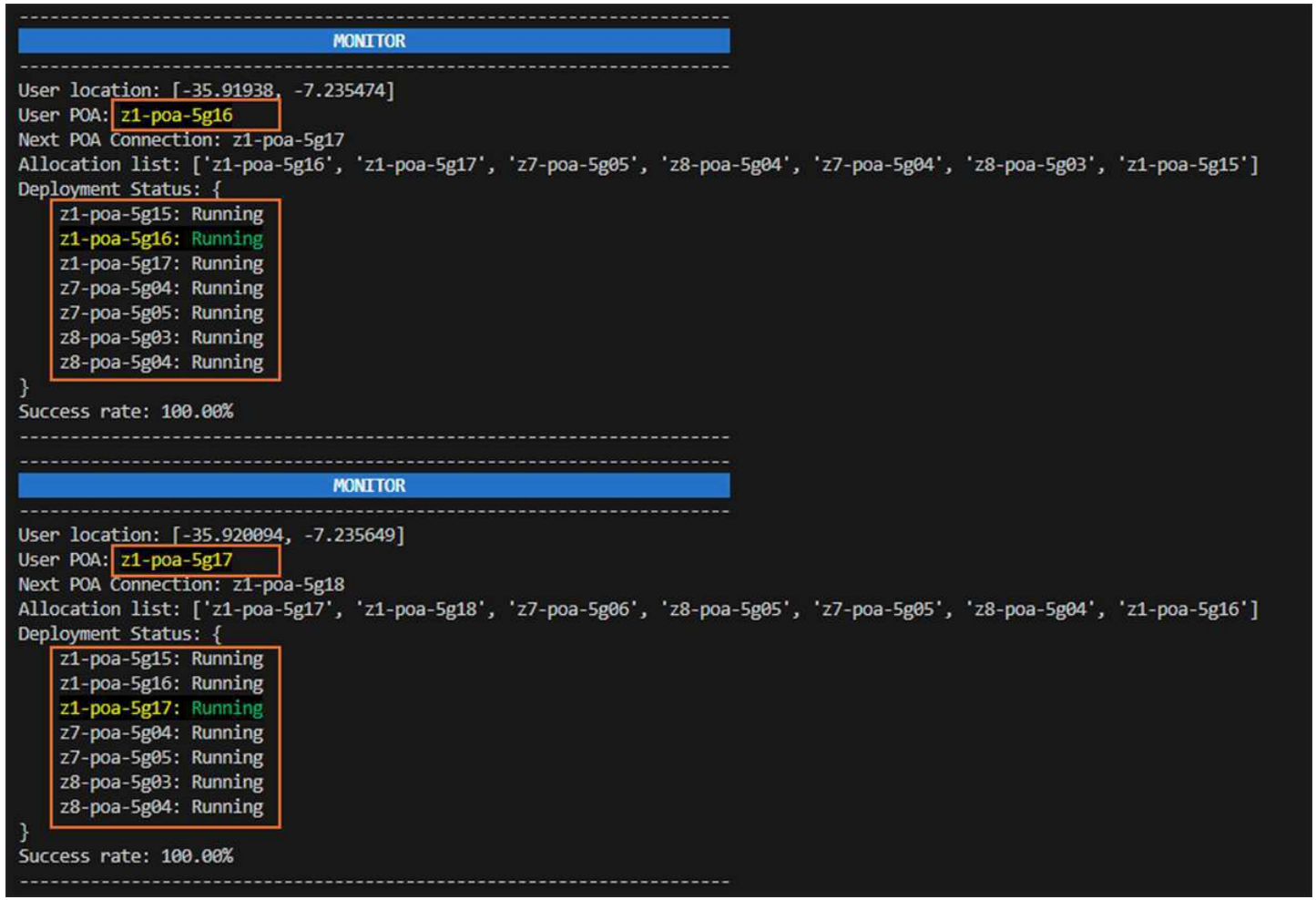
Fonte: Produzida pelo autor.

Figura 7.34: Comparativo de consumo de memória para os cenários 3 e 4.



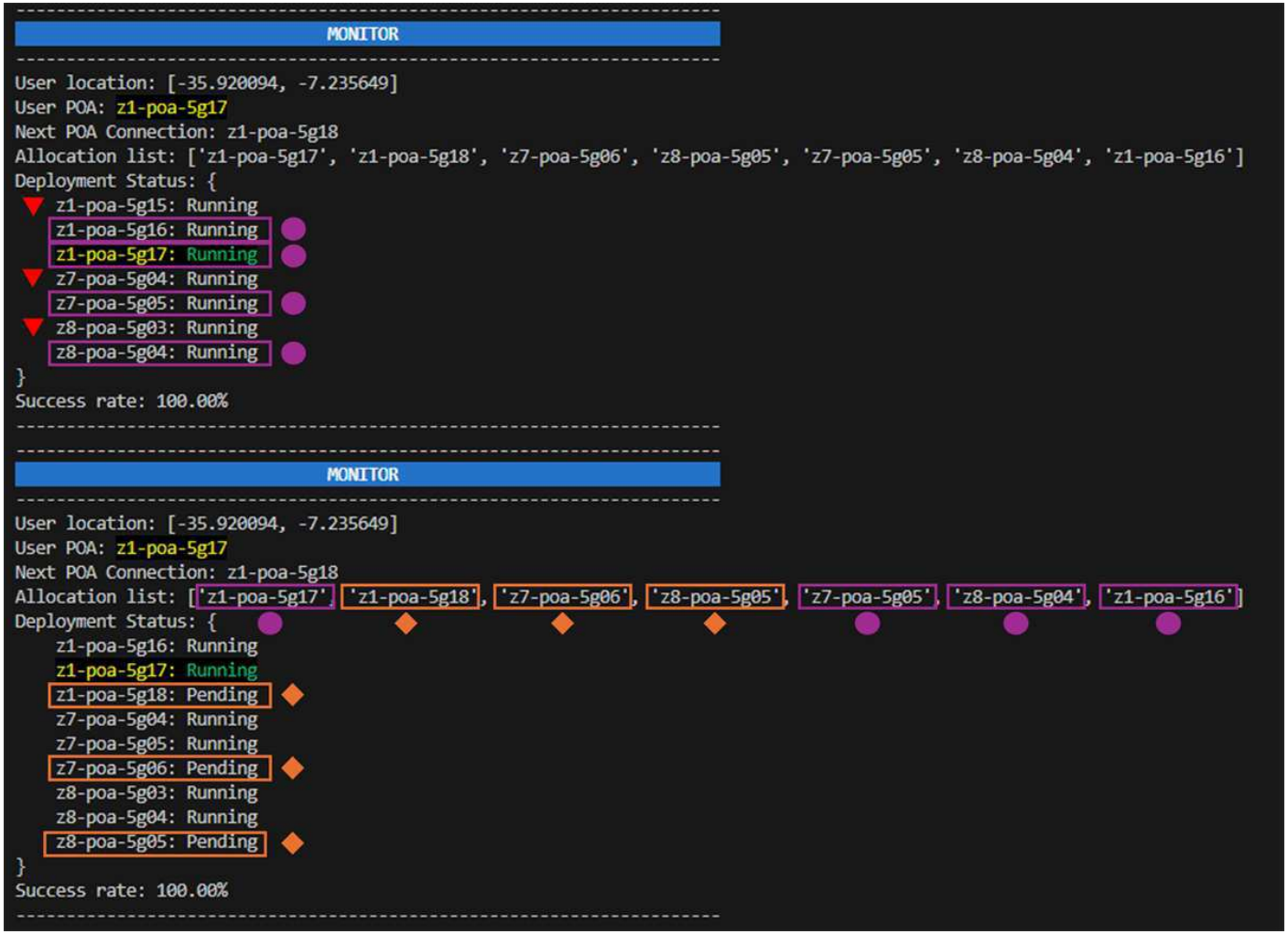
Fonte: Produzida pelo autor.

Figura 7.35: Captura de tela do terminal para o Cenário 4 - Transição entre Bordas.



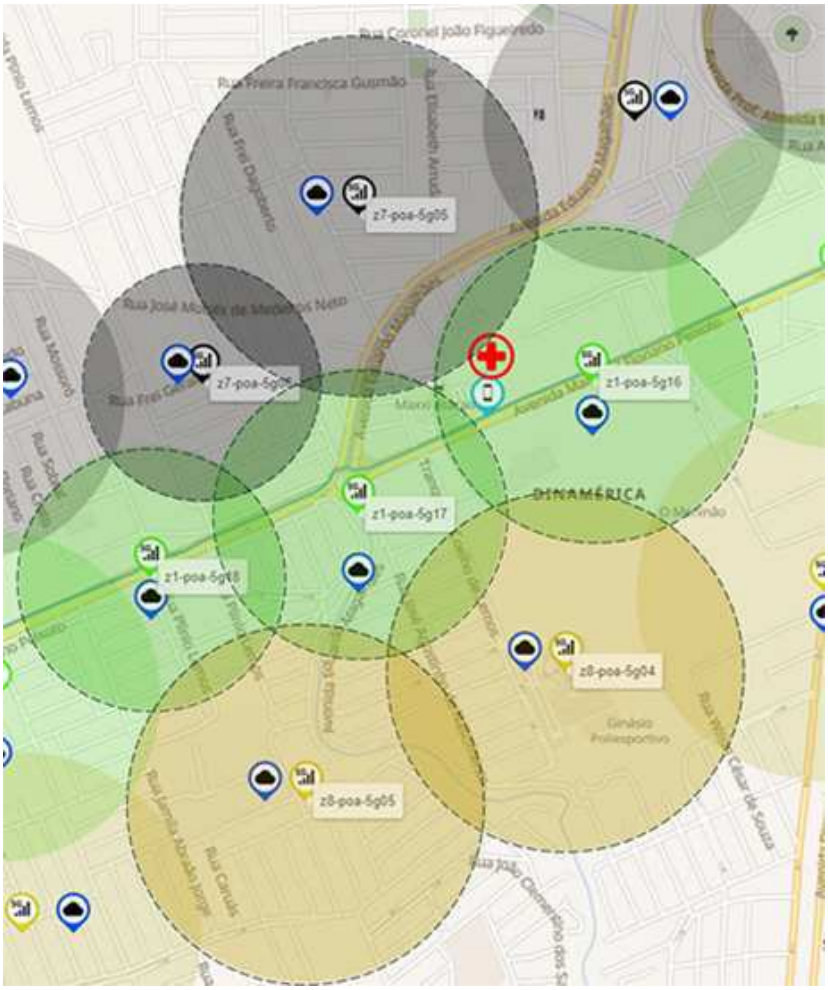
Fonte: Produzida pelo autor.

Figura 7.36: Captura de tela do terminal para o Cenário 4 - Alocação de Serviços.



Fonte: Produzida pelo autor.

Figura 7.37: Captura de tela Mapa do cenário em ambiente de múltiplas bordas para o instante de transição.



Fonte: Produzida pelo autor.

7.2.3.2 Discussão

Neste cenário foi explorada a estratégia de alocação dinâmica de serviços de streaming de vídeo em servidores de borda visando a continuidade dos serviços críticos ao longo do trajeto, especialmente para aplicações críticas sem conhecimento prévio das rotas. A implementação dessa estratégia envolveu alocar os serviços em todos os servidores vizinhos, conforme ilustrado nas Figuras 7.31 e 7.32, aumentando assim o uso de recursos computacionais, quando comparado ao caso "ótimo"do Cenário 3.

A utilização dessa abordagem, embora resulte em maior consumo de memória, conforme

evidenciado pela simulação e capturado na Figura 7.33, demonstra ser uma solução factível para garantir a disponibilidade dos serviços ao longo do trajeto da ambulância. A taxa de sucesso na disponibilização dos serviços, ilustrada na Figura 7.35, indica que os serviços estavam ativos e prontos para uso no momento da transição entre os servidores de borda. Isso sugere que, apesar do aumento no consumo de recursos, a estratégia é capaz de atender às exigências de continuidade de serviço necessárias em cenários críticos.

Além disso, o mapeamento de rede realizado e/ou o uso de estratégias de *service discovery*, mencionados anteriormente, são fundamentais para a implementação eficaz dessa abordagem de alocação dinâmica. Estes elementos contribuem para a adaptação automática da rede e a alocação de recursos conforme a necessidade, fundamentais para o sucesso da estratégia em ambientes de múltiplas bordas.

No entanto, ao conhecer-se o padrão de mobilidade ou pelo menos uma tendência é possível minimizar a alocação de serviços em servidores que não estão na mesma direção que a ambulância está se deslocando. Para isso, foi desenvolvido um algoritmo para definição dos servidores que estariam na tendência do deslocamento.

7.2.4 Cenário 5: Alocação Dinâmica de Serviços de Streaming de Vídeo com estimador do padrão de mobilidade

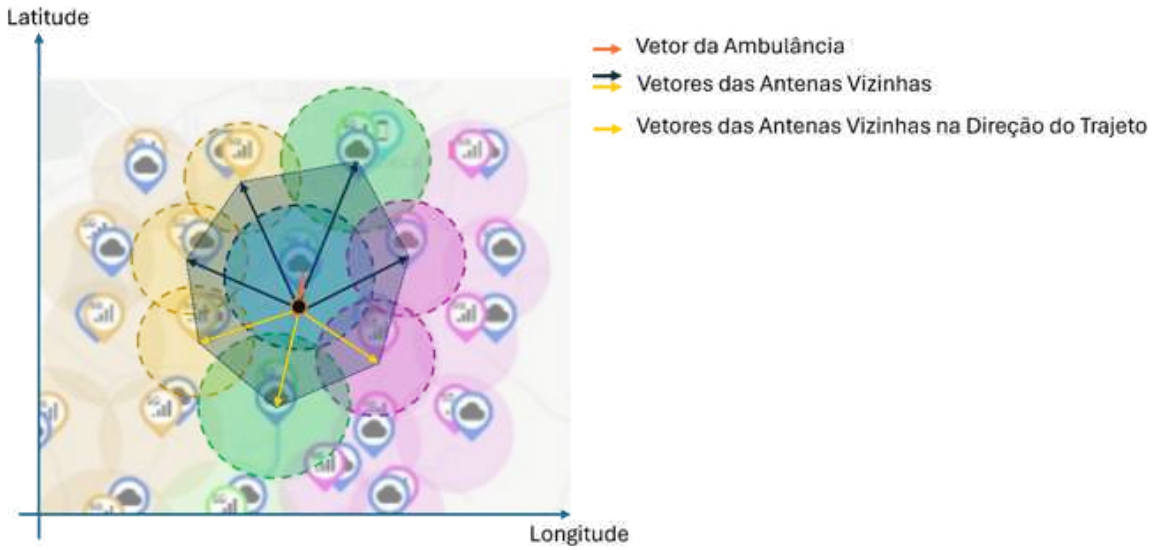
Para este cenário, a configuração de rede e a geolocalização implementadas no Cenário 4 foram replicadas. A diferença principal entre os dois cenários reside na estratégia adotada para a alocação dos serviços. Esta estratégia é uma das contribuições deste trabalho, uma vez que envolve o uso de uma abordagem alternativa e prática para identificar o padrão de mobilidade de ambulâncias, crucial para a otimização de serviços críticos em movimento, sem aumentar o *overhead* ocasionado por algoritmos de estimativa de padrão de mobilidade mais avançados. Esta solução alternativa emprega um algoritmo eurístico projetado para prever o deslocamento futuro da ambulância com base no mapa de rede e no histórico de deslocamentos passados da ambulância.

O algoritmo utilizado opera calculando a distância entre vetores. Dois conjuntos de vetores são considerados neste cálculo: o primeiro conjunto consiste nos vetores que representam as distâncias da ambulância para as antenas vizinhas, e o segundo conjunto é formado pelo

vetor que representam o deslocamento da ambulância de um ponto anterior para o ponto atual, ou seja, com base no seu histórico de deslocamento.

Este algoritmo eurístico, ao integrar tanto a informação geográfica do ambiente quanto o comportamento histórico da ambulância, visa fornecer uma previsão praticável do padrão de mobilidade. A abordagem visa não apenas a eficácia na identificação da trajetória da ambulância, mas também a eficiência na gestão dos recursos de computação em múltiplas bordas, ao focar a alocação de serviços críticos nas antenas que provavelmente estarão na rota da ambulância. Na Figura 7.38 está ilustrado o diagrama da estratégia utilizada para este cenário.

Figura 7.38: Estratégia utilizada para o Cenário 5.



Fonte: Produzida pelo autor.

Além disso, nas Listas 3 e 4 estão apresentados os códigos fonte das principais funções utilizadas no estimador do padrão de mobilidade. A estimativa é baseada no cálculo do ângulo entre retas e no mapa de rede.

```
def calcular_angulo_entre_retas(p1_r1, p2_r1, p1_r2, p2_r2):
    """
    Calcula o ângulo entre duas direções definidas por pontos de
    latitude e longitude.

    :param p1_r1: Tupla contendo as coordenadas (lat, lon) do
    primeiro ponto da primeira reta.
    :param p2_r1: Tupla contendo as coordenadas (lat, lon) do
    segundo ponto da primeira reta.
    :param p1_r2: Tupla contendo as coordenadas (lat, lon) do
    primeiro ponto da segunda reta.
    :param p2_r2: Tupla contendo as coordenadas (lat, lon) do
    segundo ponto da segunda reta.
    :return: Ângulo entre as duas direções em graus.
    """
    vetor1_inicio = latlon_to_vector(*p1_r1)
    vetor1_fim = latlon_to_vector(*p2_r1)
    vetor2_inicio = latlon_to_vector(*p1_r2)
    vetor2_fim = latlon_to_vector(*p2_r2)

    vetor1 = (vetor1_fim[0] - vetor1_inicio[0], vetor1_fim[1] -
            vetor1_inicio[1], vetor1_fim[2] - vetor1_inicio[2])

    vetor2 = (vetor2_fim[0] - vetor2_inicio[0], vetor2_fim[1] -
              vetor2_inicio[1], vetor2_fim[2] - vetor2_inicio[2])

    prod_escalar = produto_escalar(vetor1, vetor2)
    mag_vetor1 = magnitude(vetor1)
    mag_vetor2 = magnitude(vetor2)

    cos_theta = prod_escalar / (mag_vetor1 * mag_vetor2)

    # Limitar o valor do cosseno para evitar erros de arredondamento
    fora do intervalo [-1, 1]
    cos_theta = max(min(cos_theta, 1), -1)

    angulo_rad = math.acos(cos_theta)
    angulo_graus = math.degrees(angulo_rad)

    return angulo_graus

```

Listing 3: Código fonte do ângulo entre os vetores.

```
def define_online_antennas_commands( antenna_posicao_raio_dict, grafo_dict,
        posicao_anterior, posicao_atual, servidor_atual_name, range_erro = 1):

        servidores_vizinho = grafo_dict[servidor_atual_name]

        list_angulos_vizinhos = []
        for servidor_vizinho_name in servidores_vizinho:
                if servidor_vizinho_name != servidor_atual_name:
                        x_vizinho_aux =
                            antenna_posicao_raio_dict[servidor_vizinho_name][0]

                        y_vizinho_aux =
                            antenna_posicao_raio_dict[servidor_vizinho_name][1]

                        centro_servidor_vizinho =
                            (x_vizinho_aux, y_vizinho_aux)

                        angulo = calcular_angulo_entre_retas(
                            posicao_anterior,
                            posicao_atual,
                            posicao_anterior,
                            centro_servidor_vizinho)

                        list_angulos_vizinhos.append(
                            (servidor_vizinho_name, angulo))

        sorted_list = sorted(list_angulos_vizinhos, key=lambda x: x[1])
        list_servidores_escolhidos = sorted_list[:range_erro]
        first_elements = [t[0] for t in list_servidores_escolhidos]

        #resposta --> [vetores em ordem de prioridade]
        return first_elements

```

Listing 4: Código fonte da função do estimador do padrão de mobilidade.

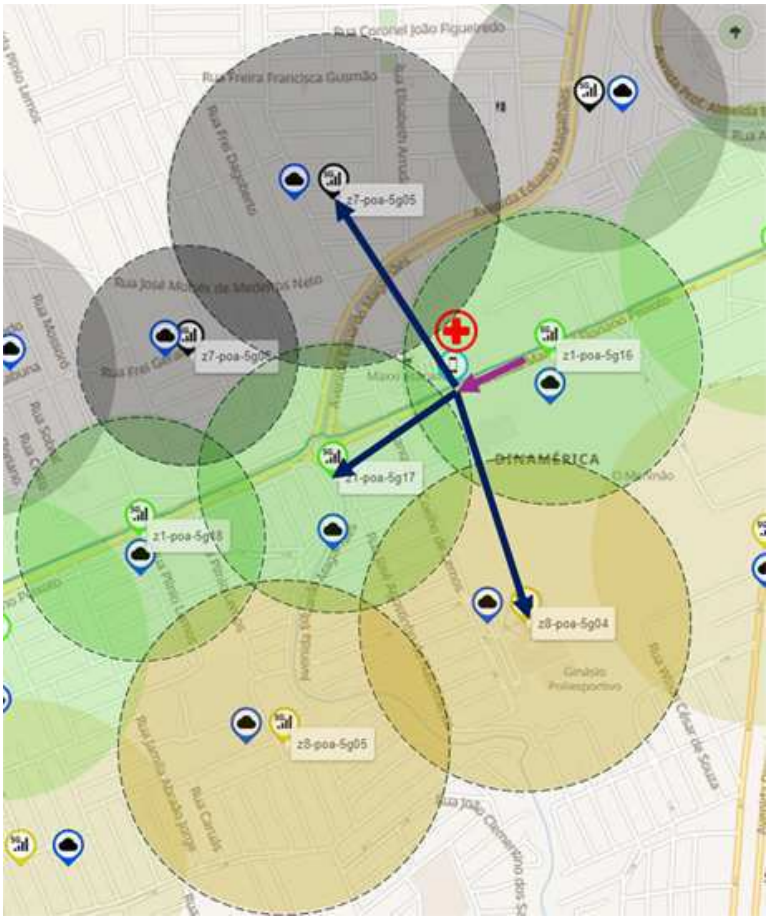
7.2.4.1 Resultados

Para este cenário também foram utilizadas as métricas a avaliação do consumo de memória e taxa de sucesso na disponibilização dos serviços ao longo do trajeto. Com base nessa simulação, observou uma redução no consumo de memória proporcional aos serviços não alocados nos servidores vizinhos fora do padrão de mobilidade da ambulância. Representando um ganho, principalmente nos cenários onde não se conhece o trajeto da ambulância,

garantindo, ainda, a continuidade na disponibilização dos serviços. Na Figura 7.40 está ilustrada a captura de tela do Grafana para utilização de Memória no Cenário 5. Na Figura 7.41 está ilustrado um comparativo dos consumos de memória entre os Cenários 4 e 5. Além disso, em uma simulação adicional, foi feita uma análise comparativa para os cinco cenários, onde os resultados estão ilustrados na Figura 7.43.

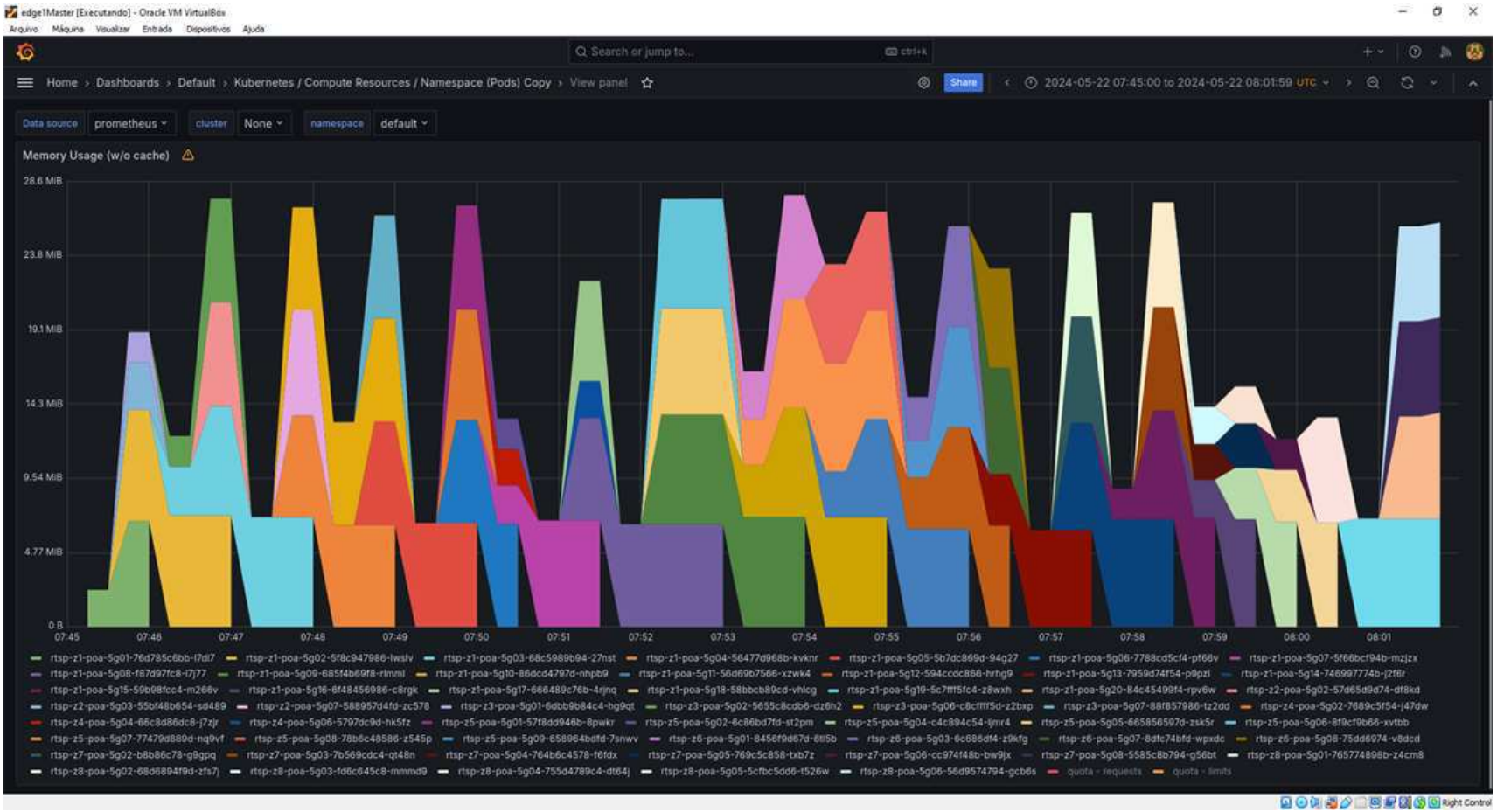
Com relação à taxa de sucesso na disponibilização dinâmica dos serviços, na Figura 7.42 está ilustrado o momento de transição entre os servidores de borda **z1-poa-5g17** e **z1-poa-5g18**, também ilustrado na visão geográfica na Figura 7.39, onde o serviço respectivo ao servidor destino já estava em execução para utilização no momento da chegada da ambulância, destacando-se a menor quantidade de serviços alocados e também a liberação, manutenção e alocação de serviços de modo dinâmico.

Figura 7.39: Captura de tela Mapa do cenário em ambiente de múltiplas bordas para o instante de transição.



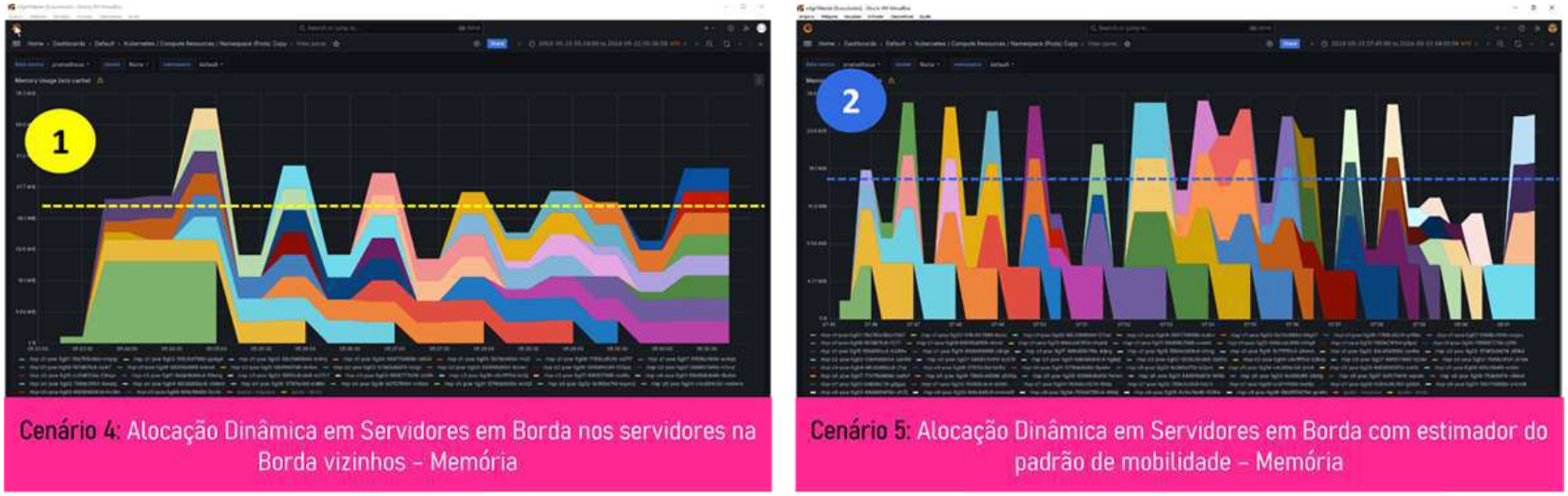
Fonte: Produzida pelo autor.

Figura 7.40: Captura de tela do Grafana para utilização de Memória no Cenário 5.



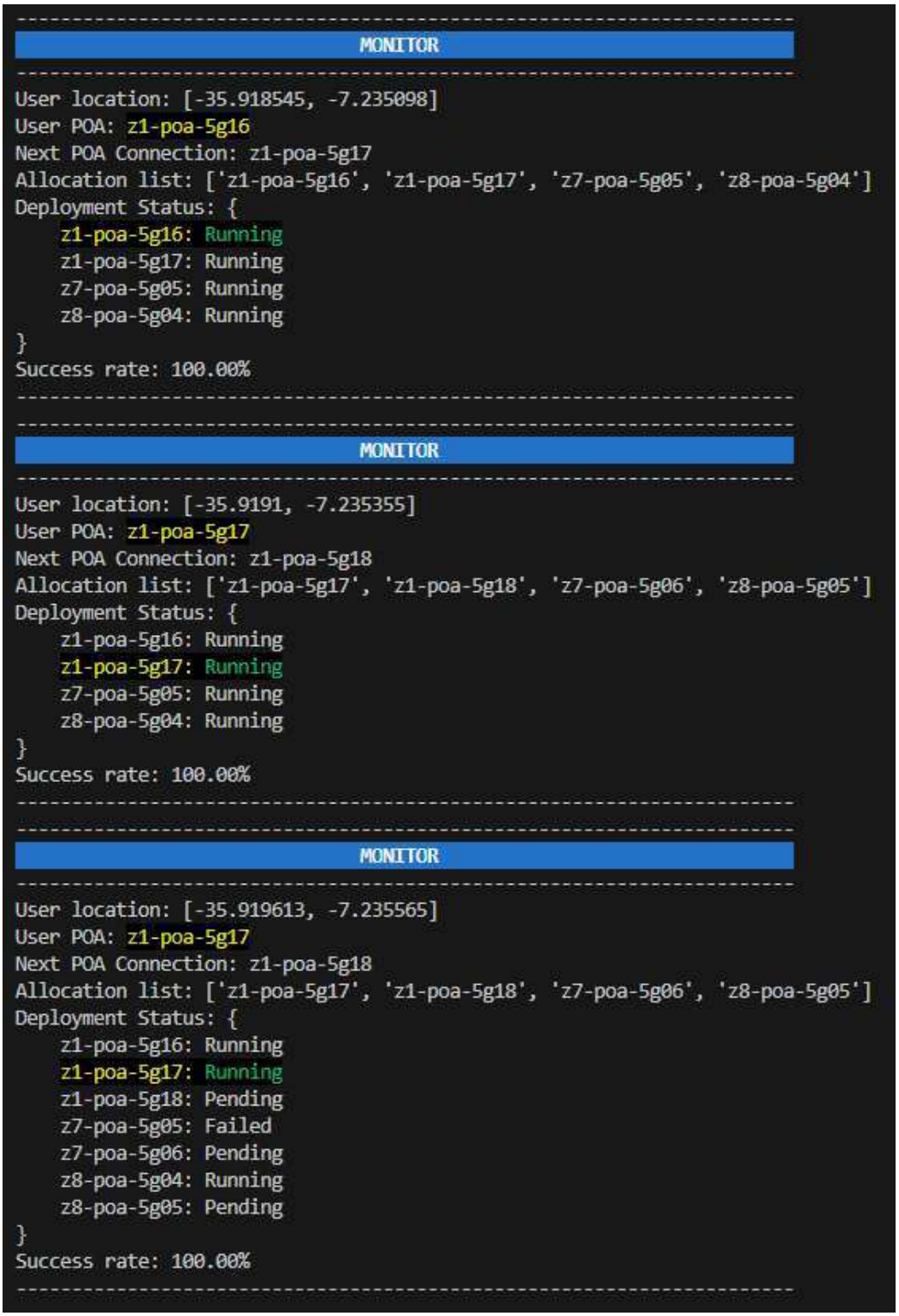
Fonte: Produzida pelo autor.

Figura 7.41: Comparativo de consumo de memória para os cenários 4 e 5.



Fonte: Produzida pelo autor.

Figura 7.42: Captura de tela do terminal para o Cenário 5 - Alocação de Serviços.



Fonte: Produzida pelo autor.

Figura 7.43: Comparativo de consumo de memória para os cinco cenários.

Fonte: Produzida pelo autor.

7.2.4.2 Discussão

A estratégia proposta para a alocação dinâmica de serviços aplica uma abordagem heurística que combina informações geográficas e históricas de movimentação da ambulância para otimizar a disponibilização de serviços críticos. Esta abordagem representa uma contribuição significativa, já que combina a localização geográfica atual e passada com o comportamento de movimentação para prever o percurso futuro da ambulância, permitindo uma gestão mais eficiente dos recursos computacionais distribuídos.

A utilização de um algoritmo heurístico, como detalhado, baseia-se em cálculos de distância entre vetores, que incorporam tanto as distâncias atuais da ambulância até as antenas próximas quanto o histórico de deslocamentos para calcular a trajetória provável da ambulância. Esta metodologia aponta para uma maneira eficaz de reduzir a sobrecarga geral no sistema ao evitar o uso de algoritmos de predição de mobilidade mais complexos e onerosos em termos computacionais.

Além disso, a partir dos resultados obtidos de simulações, observou-se uma redução no consumo de memória, proporcional à não alocação de serviços em servidores fora do padrão de mobilidade da ambulância. Isso não apenas reflete a eficiência da abordagem em termos de uso de recursos, mas também garante a continuidade da disponibilidade dos serviços essenciais ao longo do trajeto previsto.

Portanto, a estratégia implementada em Cenário 5 oferece um equilíbrio entre precisão na previsão e eficiência no uso de recursos, destacando-se como uma solução viável e prática para a gestão de serviços críticos em cenários de mobilidade, como é o caso das 'ambulâncias conectadas'.

7.3 Considerações finais

A partir do Cenário 1, foi possível observar o comportamento da alocação de serviços em Servidores em Nuvem quando implementada para disponibilizar serviços de aplicações críticas de saúde. A partir da análise do tempo de resposta, percebe-se que este tempo não atende aos requisitos de latência exigidos.

Posteriormente, no Cenário 2, uma solução em potencial para o problema apresentado no Cenário 1 foi simulado: a alocação de serviços em Servidores na Borda. Por meio desse

experimento, percebe-se que há uma melhoria no tempo de resposta, dada a redução na latência da rede.

No entanto, em um cenário de mobilidade, alocar os serviços de forma estática pode ser uma solução cara, uma vez que em boa parte de um trajeto, grande parte dos serviços alocados estarão ociosos. Neste sentido, alocar os recursos dinamicamente foi uma alternativa explorada (Cenários 3, 4 e 5) que pode ser utilizada para otimizar a utilização da infraestrutura de computação em borda, ainda atendendo aos requisitos de comunicação para o cenário crítico de saúde explorado.

Desse modo, por meio da atuação proativa e dinâmica, conforme a estratégia utilizada no método proposto *Make Way*, que considerem o padrão de mobilidade da Ambulância Conectada, consegue-se viabilizar a alocação ótima de recursos e serviços em ambientes e infraestrutura de computação em borda, além do atendimento aos requisitos de comunicação para aplicações críticas de saúde, o que promove o desenvolvimento de soluções de ambulâncias conectadas.

Capítulo 8

Conclusão

Nesta tese foi apresentado um método para disponibilização de serviços críticos de saúde em ambientes com computação na borda. Neste método, a alocação de serviços críticos é realizada de forma antecipativa, de acordo com o padrão de mobilidade na unidade de atendimento móvel, de modo que os serviços não-críticos, eventualmente alocados em servidores de borda necessários para atender às aplicações críticas, são migrados para servidores menos críticos ou para a nuvem. Com isso, é possível viabilizar o acesso a servidores de borda mais próximos da ambulância ao longo de um trajeto. O método *Make Way* é a solução adotada para gerenciar eficientemente essa alocação dinâmica dos serviços críticos e a migração dos serviços não-críticos, uma vez que esse meio de troca de informações é intrínseco ao uso da computação na borda, suprindo, assim, os requisitos esperados e desejados.

Foi verificado na revisão bibliográfica que as soluções baseadas em Realidade Aumentada (AR) e Inteligência Artificial (IA) demandam um processamento de dados de alta velocidade e baixa latência para serem eficazes. No cenário de ambulâncias conectadas, a importância de uma resposta rápida e precisa para o tratamento de emergências é crucial, reforçando a necessidade de sistemas robustos e ágeis. Além disso, foi verificado que sistemas de computação em borda podem proporcionar a infraestrutura necessária para suportar essas demandas, aproximando o processamento dos dados dos locais de uso. A partir daí, foram realizados experimentos utilizando um ambiente de simulação controlado, para testar a viabilidade e eficácia dessas tecnologias em um contexto de emergência médica. O mé-

todo foi projetado para permitir a alocação e migração de recursos de maneira dinâmica e proativa, assegurando que os serviços críticos fossem disponibilizados de forma antecipada e contínua. Nos experimentos foi constatado que a abordagem de alocação dinâmica de serviços em ambientes de computação em borda melhora significativamente o desempenho dos sistemas de AR e IA, inclusive na ocorrência de flutuações de rede e picos de demanda.

Foram experimentados cinco cenários para validação experimental do problema e da solução proposta para a Prova de Conceito definida para validação do método proposto, conforme apresentado no Capítulo 7. A partir do Cenário 1, foi possível validar que a latência impacta diretamente na qualidade dos serviços de cenários críticos, principalmente quando as aplicações são disponibilizadas baseadas em ambientes de computação em nuvem. O Cenário 2 representa uma possível solução para o problema de latência, ao alocar os serviços em servidores na borda. Este é um cenário hipotético, pois na grande maioria dos casos não se sabe o caminho a ser percorrido pelas ambulâncias. Além disso, com os resultados de simulação deste cenário foi possível observar um aumento no consumo de memória ao disponibilizar-se todos os serviços de forma estática, entretanto, o funcionamento do cenário não foi prejudicado, onde a latência foi efetivamente reduzida. Visando a otimização de recursos em servidores de computação em borda, ainda garantindo os requisitos de comunicação para aplicações críticas, nos Cenários 3, 4 e 5 a validação da solução proposta foi avaliada, onde os serviços são disponibilizados dinamicamente, conforme o padrão de mobilidade da ambulância.

Quando se analisa as métricas obtidas nas simulações, como CPU e memória, no servidor de borda percebe-se que a média de tempo de uso de CPU está diretamente ligados à necessidade computacional do serviço disponibilizado. Quanto mais soluções inteligentes para assistência aos paramédicos no cenário, maiores os valores para essas métricas, visto que, com mais soluções assistidas por IA, o servidor de borda terá uma quantidade maior de serviços a serem disponibilizados nos servidores em borda, sejam estes para melhoria da experiência dos usuários de óculos inteligentes de AR, seja no processamento de grande volume de dados provenientes de dispositivos médicos inteligente ou até na execução de modelos de IA visando assistir os paramédicos no diagnóstico diferencial e recomendação de procedimentos, provenientes, principalmente de sistemas de visão computacional associados às capturas de quadros através dos óculos inteligentes.

Em relação ao uso de memória dos recursos computacionais da borda, o Cenário 2 apresentou maior consumo de memória, proporcional à quantidade de servidores em borda a serem utilizados ao longo de um determinado trajeto da ambulância conectada. A quantidade de serviços inteligentes a serem disponibilizados também podem ter impacto direto no uso de CPU. No que concerne às métricas associadas aos requisitos de comunicação, a latência e a taxa de transmissão emergem como parâmetros fundamentais para a avaliação da eficácia do sistema na transmissão de dados críticos de maneira eficiente e confiável. A latência, em particular, é um indicador importantíssimo que reflete o tempo necessário para o processamento e a entrega de dados aos pontos finais, enquanto a taxa de transmissão quantifica a velocidade com que os dados são enviados através da rede. Ambas as métricas são influenciadas pela proximidade geográfica dos servidores de borda aos dispositivos finais e pela capacidade desses servidores em administrar e priorizar o tráfego de rede eficazmente. A análise destas métricas é vital para assegurar que os requisitos de aplicações críticas, como os serviços de saúde em ambientes conectados, sejam cumpridos adequadamente, garantindo que a infraestrutura de comunicação possa sustentar elevados padrões de desempenho mesmo sob condições de demanda intensiva.

Na perspectiva da IoT e Cidades Inteligentes, o método proposto pode ser utilizado como uma solução eficaz para gerenciamento e coordenação de serviços críticos, como operações de bombeiros, segurança pública e coordenação de tráfego inteligente. Junto com sistemas de monitoramento e alerta, ele otimiza as respostas a emergências e melhora a eficiência do tráfego urbano. Além disso, tendo sua parte principal funcionando no servidor de borda, a arquitetura pode ser integrada a outros serviços, como, por exemplo, um sistema de comunicação e coordenação entre diferentes forças de segurança durante ocorrências policiais. Considerando ainda o requisito de resposta rápida imposto por Cidades Inteligentes, com o método é possível agilizar as operações de emergência, segurança pública e tráfego, com baixo custo, dado que as configurações principais serão realizadas via software, facilitando a implementação e adaptação em diferentes cenários urbanos.

8.1 Sugestões de trabalhos futuros

Algumas sugestões para continuidade do trabalho, são:

- **Investigação e Otimização de Algoritmos de Estimativa do Padrão de Mobilidade:** Desenvolver e aperfeiçoar algoritmos que prevejam e otimizem o padrão de mobilidade em cenários urbanos, visando melhorar a eficiência do método em situações dinâmicas e em movimento;

- **Simulações e Testes em Ambientes Reais:** Expandir as simulações para incluir testes em ambientes reais, abrindo oportunidade para exploração do estado-da-arte de diferentes vertentes tecnológicas com foco na disponibilização de aplicações médicas inteligentes. A integração de tecnologias de AR, equipamentos médicos inteligentes e sistemas baseado em IA fomenta pesquisas avançadas que abrangem desde o desenvolvimento de infraestruturas de comunicação e computação até a efetiva criação de aplicações;

- **Avaliação de Cenários Híbridos:** Concentrar esforços na avaliação de cenários que integram computação em nuvem e computação de borda, aplicando estas tecnologias em diversos setores como saúde, segurança, entre outros. O objetivo é analisar como diferentes configurações e aplicações podem coexistir e se complementar em ambientes híbridos, visando otimizar o desempenho e a eficiência em várias verticais de mercado. Este estudo contribuirá para melhorar a implementação e a gestão de infraestruturas tecnológicas complexas adaptadas às necessidades específicas de cada setor.

Por fim, os desafios específicos que demandam soluções e investigações incluem a necessidade de desenvolver métodos robustos para a migração de sessões e dados entre diferentes nós de borda sem perda de informação, garantir a segurança e a privacidade dos dados do paciente em todas as etapas do processo e otimizar a gestão de energia dos dispositivos móveis utilizados no atendimento emergencial. A resposta a esses desafios é crucial para a implementação efetiva do *Make Way*, e sua capacidade de atender às necessidades específicas do problema de negócios relacionado à alocação dinâmica de serviços críticos de saúde.

Bibliografia

[1] 3GPP, "Study on communication services for critical medical applications (release 17)," 2021.

[2] S. Shalini and T. Devi, *Digital Transformation*, p. 67–79. Auerbach Publications, Dec. 2022.

[3] M. Z. Yaqub and A. Alsabban, "Industry-4.0-enabled digital transformation: Prospects, instruments, challenges, and implications for business strategies," *Sustainability*, vol. 15, p. 8553, May 2023.

[4] P. Raja, D. S. Kumar, D. S. Yadav, and D. T. Singh, "The internet of things (iot): A review of concepts, technologies, and applications," *International Journal of Information technology and Computer Engineering*, p. 21–32, Mar. 2023.

[5] S. Dhillon, N. Mishra, and D. K. Shakya, *Applications of IoT and Various Attacks on IoT*, p. 705–719. Springer Nature Singapore, 2023.

[6] C. Rocha, C. Quandt, F. Deschamps, S. Philbin, and G. Cruzara, "Collaborations for digital transformation: Case studies of industry 4.0 in brazil," *IEEE Transactions on Engineering Management*, vol. 70, p. 2404–2418, July 2023.

[7] N. Lakemond, G. Holmberg, and A. Pettersson, "Digital transformation in complex systems," *IEEE Transactions on Engineering Management*, vol. 71, p. 192–204, 2024.

[8] M. Carter, J. Stephenson, and S. Carlon, "Data-based decision-making," July 2020.

[9] R. Sala, M. Bertoni, F. Pirola, and G. Pezzotta, "Data-based decision-making in maintenance service delivery: the d3m framework," *Journal of Manufacturing Technology Management*, vol. 32, p. 122–141, Apr. 2021.

[10] A. Klimesch, A. Martinez-Pereira, C. Topf, M. Härter, I. Scholl, and P. Bravo, "Conceptualization of patient-centered care in latin america: A scoping review," *Health Expectations*, vol. 26, p. 1820–1831, July 2023.

[11] G. Eysenbach, "What is e-health?," *Journal of Medical Internet Research*, vol. 3, p. e20, June 2001.

[12] M. Wehde, "Healthcare 4.0," *IEEE Engineering Management Review*, vol. 47, p. 24–28, Sept. 2019.

[13] L. He, M. Eastburn, J. Smirk, and H. Zhao, "Smart chemical sensor and biosensor networks for healthcare 4.0," *Sensors*, vol. 23, p. 5754, June 2023.

[14] J. Al-Jaroodi, N. Mohamed, N. Kesserwan, and I. Jawhar, "Human factors affecting the adoption of healthcare 4.0," in *2023 IEEE International Systems Conference (SysCon)*, IEEE, Apr. 2023.

[15] M. Sony, J. Antony, and G. L. Tortorella, "Critical success factors for successful implementation of healthcare 4.0: A literature review and future research agenda," *International Journal of Environmental Research and Public Health*, vol. 20, p. 4669, Mar. 2023.

[16] M. E. Winters, K. Hu, J. P. Martinez, H. Mallemat, and W. J. Brady, "The critical care literature 2019," *The American Journal of Emergency Medicine*, vol. 39, pp. 197–206, Jan. 2021.

[17] S. A. Fahim Yegane, A. Shahrami, H. R. Hatamabadi, and S.-M. Hosseini-Zijoud, "Clinical information transfer between ems staff and emergency medicine assistants during handover of trauma patients," *Prehospital and Disaster Medicine*, vol. 32, p. 541–547, June 2017.

[18] M. Schinle, I. Papantonis, and W. Stork, "Personalization of monitoring system parameters to support ambulatory care for dementia patients," in *2018 IEEE Sensors Applications Symposium (SAS)*, IEEE, Mar. 2018.

[19] F. Martinez-Suarez and C. Alvarado-Serrano, "Prototype of an ambulatory ECG monitoring system with r wave detection in real time based on FPGA," in *2019 16th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE)*, IEEE, Sept. 2019.

[20] C. R. Gomez, "Editorial: Time is brain!," *Journal of Stroke and Cerebrovascular Diseases*, vol. 3, p. 1–2, Jan. 1993.

[21] M. P. Lin, "Time matters greatly in acute stroke care," *Neurologia i Neurochirurgia Polska*, vol. 54, p. 104–105, Apr. 2020.

[22] H.-A. Chen, S.-T. Hsu, M.-J. Hsieh, S.-S. Sim, S.-E. Chu, W.-S. Yang, Y.-C. Chien, Y.-C. Wang, B.-C. Lee, E. P.-C. Huang, H.-Y. Lin, M. H.-M. Ma, W.-C. Chiang, and J.-T. Sun, "Influence of advanced life support response time on out-of-hospital cardiac arrest patient outcomes in taipei," *PLOS ONE*, vol. 17, p. e0266969, Apr. 2022.

[23] A. Abrardo, M. Marsan, N. Melazzi, and S. Buzzi, "5g trials in italy," *5G Italy White eBook: From Research to Market*, 2019.

[24] G. Yang, Z. Pang, M. Jamal Deen, M. Dong, Y.-T. Zhang, N. Lovell, and A. M. Rahmani, "Homecare robotic systems for healthcare 4.0: Visions and enabling technologies," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, p. 2535–2549, Sept. 2020.

[25] M. Sony, J. Antony, and O. McDermott, "The impact of healthcare 4.0 on the healthcare service quality: A systematic literature review," *Hospital Topics*, vol. 101, p. 288–304, Mar. 2022.

[26] K. Apiratwarakul, L. W. Cheung, and K. Ienghong, "Impact of smart glasses on patient care time in emergency medical services ambulance," *Prehospital and Disaster Medicine*, vol. 38, p. 735–739, Oct. 2023.

[27] C. Thaijiam, "A smart ambulance with information system and decision-making process for enhancing rescue efficiency," *IEEE Internet of Things Journal*, vol. 10, p. 7293–7302, Apr. 2023.

[28] D. Tedesco, A. Capodici, G. Gribaudo, Z. Di Valerio, M. Montalti, A. Salussolia, V. Barbagallo, M. Rolli, M. Fantini, and D. Gori, "Innovative health technologies to improve emergency department performance," *European Journal of Public Health*, vol. 32, Oct. 2022.

[29] A. Kavithamani, R. Vijay, S. Monika, E. Sentamilan, and S. Ramya, *Augmented Reality Based Smart Ambulance System*, p. 635–641. Springer International Publishing, 2020.

[30] K. Apiratwarakul, L. W. Cheung, S. Tiamkao, P. Phungoen, K. Tientanopajai, W. Taweepworadej, W. Kanarkard, and K. Ienghong, "Smart glasses: A new tool for assessing the number of patients in mass-casualty incidents," *Prehospital and Disaster Medicine*, vol. 37, p. 480–484, June 2022.

[31] S. H. A. Shah, D. Koundal, V. Sai, and S. Rani, "Guest editorial: Special section on 5g edge computing-enabled internet of medical things," *IEEE Transactions on Industrial Informatics*, vol. 18, p. 8860–8863, Dec. 2022.

[32] M. Satyanarayanan, "Edge computing," *Computer*, vol. 50, no. 10, p. 36–38, 2017.

[33] S. Lu, J. Lu, K. An, X. Wang, and Q. He, "Edge computing on iot for machine signal processing and fault diagnosis: A review," *IEEE Internet of Things Journal*, vol. 10, p. 11093–11116, July 2023.

[34] K. Cao, S. Hu, Y. Shi, A. Colombo, S. Karnouskos, and X. Li, "A survey on edge and edge-cloud computing assisted cyber-physical systems," *IEEE Transactions on Industrial Informatics*, vol. 17, p. 7806–7819, Nov. 2021.

[35] K. Zen, S. Mohanan, S. Tarmizi, N. Annuar, and N. U. Sama, "Latency analysis of cloud infrastructure for time-critical iot use cases," in *2022 Applied Informatics International Conference (AiIC)*, IEEE, May 2022.

[36] DeepShah, "A comparative study on cloud, fog and edge computing," in *2021 5th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT)*, IEEE, Dec. 2021.

[37] A. K. Saxena, R. Pandey, and N. K. Singh, "Latency analysis and reduction methods for edge computing," in *2023 IEEE World Conference on Applied Intelligence and Computing (AIC)*, IEEE, July 2023.

[38] M. Hunko, V. Tkachov, A. Kovalenko, and H. Kuchuk, "Advantages of fog computing: A comparative analysis with cloud computing for enhanced edge computing capabilities," in *2023 IEEE 4th KhPI Week on Advanced Technology (KhPIWeek)*, IEEE, Oct. 2023.

[39] D. W. Lee, H. J. Moon, and N. H. Heo, "Association between ambulance response time and neurologic outcome in patients with cardiac arrest," *The American Journal of Emergency Medicine*, vol. 37, p. 1999–2003, Nov. 2019.

[40] A. Alumran, H. Albinali, A. Saadah, and A. Althumairi, "The effects of ambulance response time on survival following out-of-hospital cardiac arrest," *Open Access Emergency Medicine*, vol. Volume 12, p. 421–426, Dec. 2020.

[41] Y. Lu, G. Zhao, C. Chakraborty, C. Xu, L. Yang, and K. Yu, "Time-sensitive networking-driven deterministic low-latency communication for real-time telemedicine and e-health services," *IEEE Transactions on Consumer Electronics*, vol. 69, p. 734–744, Nov. 2023.

[42] Y. Zhai, X. Xu, B. Chen, H. Lu, Y. Wang, S. Li, X. Shi, W. Wang, L. Shang, and J. Zhao, "5g-network-enabled smart ambulance: Architecture, application, and evaluation," *IEEE Network*, vol. 35, p. 190–196, Jan. 2021.

[43] Y. Siriwardhana, P. Porambage, M. Liyanage, and M. Ylianttila, "A survey on mobile augmented reality with 5g mobile edge computing: Architectures, applications, and technical aspects," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 2, p. 1160–1192, 2021.

[44] H. Chen, Y. Dai, H. Meng, Y. Chen, and T. Li, "Understanding the characteristics of mobile augmented reality applications," in *2018 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, IEEE, Apr. 2018.

[45] H. Yu, S. Leng, and F. Wu, "Joint cooperative computation offloading and trajectory optimization in heterogeneous uav-swarm-enabled aerial edge computing networks," *IEEE Internet of Things Journal*, vol. 11, p. 17700–17711, May 2024.

[46] O. Tao, X. Chen, Z. Zhou, L. Li, and X. Tan, "Adaptive user-managed service placement for mobile edge computing via contextual multi-armed bandit learning," *IEEE Transactions on Mobile Computing*, pp. 1–1, 2021.

[47] A. Sarkar, M. Guzdial, S. Snodgrass, A. Summerville, T. Machado, and G. Smith, "Procedural content generation via knowledge transformation (pcg-kt)," *IEEE Transactions on Games*, vol. 16, p. 36–50, Mar. 2024.

[48] V. Benavente, L. Yantas, I. Moscol, C. Rodriguez, R. Inquilla, and Y. Pomachagua, "Comparative analysis of microservices and monolithic architecture," in *2022 14th International Conference on Computational Intelligence and Communication Networks (CICN)*, IEEE, Dec. 2022.

[49] M. Uddin, S. Manickam, H. Ullah, M. Obaidat, and A. Dandoush, "Unveiling the metaverse: Exploring emerging trends, multifaceted perspectives, and future challenges," *IEEE Access*, vol. 11, p. 87087–87103, 2023.

[50] J. Ren, Y. He, G. Huang, G. Yu, Y. Cai, and Z. Zhang, "An edge-computing based architecture for mobile augmented reality," *IEEE Network*, vol. 33, p. 162–169, July 2019.

[51] A. Al-Shuwaili and O. Simeone, "Energy-efficient resource allocation for mobile edge computing-based augmented reality applications," *IEEE Wireless Communications Letters*, vol. 6, p. 398–401, June 2017.

[52] J. Ahn, J. Lee, S. Yoon, and J. K. Choi, "A novel resolution and power control scheme for energy-efficient mobile augmented reality applications in mobile edge computing," *IEEE Wireless Communications Letters*, vol. 9, p. 750–754, June 2020.

[53] J. Ahn, J. Lee, D. Niyato, and H.-S. Park, "Novel qos-guaranteed orchestration scheme for energy-efficient mobile augmented reality applications in multi-access edge com-

puting," *IEEE Transactions on Vehicular Technology*, vol. 69, p. 13631–13645, Nov. 2020.

[54] P. Ren, X. Qiao, Y. Huang, L. Liu, C. Pu, S. Dustdar, and J. Chen, "Edge ar x5: An edge-assisted multi-user collaborative framework for mobile web augmented reality in 5g and beyond," *IEEE Transactions on Cloud Computing*, vol. 10, p. 2521–2537, Oct. 2022.

[55] N. M. Quy, L. A. Ngoc, N. T. Ban, N. V. Hau, and V. K. Quy, "Edge computing for real-time internet of things applications: Future internet revolution," *Wireless Personal Communications*, vol. 132, p. 1423–1452, July 2023.

[56] H. N. Qureshi, M. Manalastas, A. Ijaz, A. Imran, Y. Liu, and M. O. A. Kalaa, "Communication requirements in 5g-enabled healthcare applications: Review and considerations," *Healthcare*, vol. 10, p. 293, Feb. 2022.

[57] F. Vhora and J. Gandhi, "A comprehensive survey on mobile edge computing: Challenges, tools, applications," in *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, IEEE, Mar. 2020.

[58] K. Cao, Y. Liu, G. Meng, and Q. Sun, "An overview on edge computing research," *IEEE Access*, vol. 8, p. 85714–85728, 2020.

[59] V. Hayyolalam, M. Aloqaily, O. Ozkasap, and M. Guizani, "Edge-assisted solutions for iot-based connected healthcare systems: A literature review," *IEEE Internet of Things Journal*, vol. 9, p. 9419–9443, June 2022.

[60] S. I. Loutfi, U. Tureli, and I. Shayea, "Augmented reality with mobility awareness in mobile edge computing over 6g network: A survey," in *2023 10th International Conference on Wireless Networks and Mobile Communications (WINCOM)*, IEEE, Oct. 2023.

[61] S. Alkaabi, M. Gregory, and S. Li, "Multi-access edge computing handover strategies, management, and challenges: A review," *IEEE Access*, p. 1–1, 2024.

[62] A. Talpur and M. Gurusamy, "Reinforcement learning-based dynamic service placement in vehicular networks," in *2021 IEEE 93rd Vehicular Technology Conference (VTC2021-Spring)*, IEEE, Apr. 2021.

[63] H. T. Malazi, S. R. Chaudhry, A. Kazmi, A. Palade, C. Cabrera, G. White, and S. Clarke, "Dynamic service placement in multi-access edge computing: A systematic literature review," *IEEE Access*, vol. 10, pp. 32639–32688, 2022.

[64] E. Ahmed, A. Gani, M. Khurram Khan, R. Buyya, and S. U. Khan, "Seamless application execution in mobile cloud computing: Motivation, taxonomy, and open challenges," *Journal of Network and Computer Applications*, vol. 52, p. 154–172, June 2015.

[65] W. Z. Khan, E. Ahmed, S. Hakak, I. Yaqoob, and A. Ahmed, "Edge computing: A survey," *Future Generation Computer Systems*, vol. 97, p. 219–235, Aug. 2019.

[66] L. Huang, P. Wang, and Y. Ji, "An implementation of low-latency and high-reliability mobile edge computing system," in *2020 IEEE MTT-S International Wireless Symposium (IWS)*, IEEE, Sept. 2020.

[67] X. Qiao, P. Ren, G. Nan, L. Liu, S. Dustdar, and J. Chen, "Mobile web augmented reality in 5g and beyond: Challenges, opportunities, and future directions," *China Communications*, vol. 16, p. 141–154, Sept. 2019.

[68] G. Panek, P. Matysiak, N. E.-h. Nouar, I. Fajjari, and H. Tarasiuk, "5g-edge relocator: A framework for application relocation in edge-enabled 5g system," in *ICC 2023 - IEEE International Conference on Communications*, IEEE, May 2023.

[69] X. Jiang, P. Hou, H. Zhu, B. Li, Z. Wang, and H. Ding, "Dynamic and intelligent edge server placement based on deep reinforcement learning in mobile edge computing," *Ad Hoc Networks*, vol. 145, p. 103172, June 2023.

[70] J. L. Vieira, A. L. E. Battisti, E. L. C. Macedo, P. F. Pires, D. C. Muchaluat-Saade, F. C. Delicato, and A. C. B. Oliveira, "Dynamic and mobility-aware vnf placement in 5g-edge computing environments," in *2023 IEEE 9th International Conference on Network Softwarization (NetSoft)*, IEEE, June 2023.

[71] Z. Ji, I. Ganchev, and M. O'Droma, "An iwbc consumer application for "always best connected and best served": design and implementation," *IEEE Transactions on Consumer Electronics*, vol. 57, p. 462–470, May 2011.

[72] S. Ullah, K.-I. Kim, K. H. Kim, M. Imran, P. Khan, E. Tovar, and F. Ali, "Uav-enabled healthcare architecture: Issues and challenges," *Future Generation Computer Systems*, vol. 97, p. 425–432, Aug 2019.

[73] H. Ullah, N. Gopalakrishnan Nair, A. Moore, C. Nugent, P. Muschamp, and M. Cuevas, "5G Communication: An Overview of Vehicle-to-Everything, Drones, and Healthcare Use-Cases," *IEEE Access*, vol. 7, pp. 37251–37268, 2019.

[74] Reply Practice, "Tecnologia 5G: Dominando o Triângulo Mágico." Disponível em: https://www.reply.com/br/industries/telco-and-media/5g-mastering-the-magic-triangle. Acessado em 09 de Julho de 2022.

[75] N. Al-Falahy and O. Y. Alani, "Technologies for 5g networks: Challenges and opportunities," *IT Professional*, vol. 19, no. 1, pp. 12–20, 2017.

[76] ITU-R Rec., ITU-R M. 2083-0, *IMT Vision — Framework and Overall Objectives of the Future Development of IMT for 2020 and Beyond*, 2015. Acesso em: 07/01/2021.

[77] M. Shafi, A. F. Molisch, P. J. Smith, T. Haustein, P. Zhu, P. De Silva, F. Tufvesson, A. Benjebbour, and G. Wunder, "5g: A tutorial overview of standards, trials, challenges, deployment, and practice," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 6, pp. 1201–1221, 2017.

[78] T. Taleb, K. Samdanis, B. Mada, H. Flinck, S. Dutta, and D. Sabella, "On multi-access edge computing: A survey of the emerging 5g network edge cloud architecture and orchestration," *IEEE Communications Surveys Tutorials*, vol. 19, no. 3, pp. 1657–1681, 2017.

[79] B. C. Ooi, G. Chen, D. Loghin, W. Wang, and M. Zhang, "5g: Agent for further digital disruptive transformations," *IEEE Data Eng. Bull.*, vol. 42, pp. 9–12, 2019.

[80] Forbes, R. (ETSI ISG ENI Chairman), "Experiential Networked Intelligence (ENI)." Disponível em: https://www.etsi.org/technologies/experiential-networked-intelligence. Acessado em 16 de Novembro de 2019.

[81] M. J. Shehab, I. Kassem, A. A. Kutty, M. Kucukvar, N. Onat, and T. Khattab, "5g networks towards smart and sustainable cities: A review of recent developments, applications and future perspectives," *IEEE Access*, vol. 10, p. 2987–3006, 2022.

[82] T.L. Inm, "Smart City Technologies Take on Covid-19," *Penang Institute: Making Ideas Work*, March 2020.

[83] S.Latif, J. Qadir, S. Farooq, M.A. Imran , "How 5g wireless (and concomitant technologies) will revolutionize healthcare?," *Future Internet*, vol. 9, no. 93, 2017.

[84] G.C.L.,Programme, "Socio-economic impact of mhealth: An assessment report for the european union," 2013.

[85] B. C. Group, "The socio-economic impact of mhealth," *Boston Consulting Group, Commissioned by Telenor Group: Boston*, 2012.

[86] Research and Markets, "I.t.u. ict facts and figures 2017," 2016.

[87] Ministério da Saúde do Brasil, "Relatório de Gestão." Disponível em: https://www.saude.gov.br/relatorio-de-gestao. Acessado em 01 de Maio de 2020.

[88] B. Pradhan, S. Das, D. S. Roy, S. Routray, F. Benedetto, and R. H. Jhaveri, "An ai-assisted smart healthcare system using 5g communication," *IEEE Access*, vol. 11, p. 108339–108355, 2023.

[89] B. N. de Desenvolvimento Econômico e Social, "Relatório do plano de ação – Iniciativas e Projetos Mobilizadores," 2017.

[90] D. Loghin, S. Cai, G. Chen, T. T. A. Dinh, F. Fan, Q. Lin, J. Ng, B. C. Ooi, X. Sun, Q.-T. Ta, W. Wang, X. Xiao, Y. Yang, M. Zhang, and Z. Zhang, "The Disruptions of 5G on Data-driven Technologies and Applications," 2019.

[91] S. Que, J. Chen, B. Chen, H. Jiang, "The application of 5g technology in logistics information acquisition," *2016 International Conference on Electronic Information Technology and Intellectualization (ICEITI 2016)* , 2016.

[92] J. Choi, V. Va, N. Gonzalez-Prelcic, R. Daniels, C. R. Bhat, and R. W. Heath, "Millimeter-wave vehicular communication to support massive automotive sensing," *IEEE Communications Magazine*, vol. 54, no. 12, pp. 160–167, 2016.

[93] C. Greer, M. Burns, D. Wollman, and E. Griffor, "Cyber-physical systems and internet of things," Mar. 2019.

[94] K. Jewani and S. Abimannan, "Edge intelligence in iot: Architecture and applications," in *2023 8th International Conference on Communication and Electronics Systems (ICCES)*, IEEE, June 2023.

[95] Red Hat, "O que é um container Linux?." Disponível em: https://www.redhat.com/pt-br/topics/containers/whats-a-linux-container. Acessado em 30 de Agosto de 2020.

[96] P. Pace, G. Aloi, R. Gravina, G. Caliciuri, G. Fortino, and A. Liotta, "An edge-based architecture to support efficient applications for healthcare industry 4.0," *IEEE Transactions on Industrial Informatics*, vol. 15, p. 481–489, Jan 2019.

[97] G. Aloi and G. Caliciuri and G. Fortino and R. Gravina and P. Pace and W. Russo and C. Savaglio, "A mobile multi-technology gateway to enable iot interoperability," in *2016 IEEE First International Conference on Internet-of-Things Design and Implementation (IoTDI)*, pp. 259–264, 2016.

[98] G. Aloi, G. Caliciuri, G. Fortino, R. Gravina, P. Pace, W. Russo, and C. Savaglio, "Enabling iot interoperability through opportunistic smartphone-based mobile gateways," *Journal of Network and Computer Applications*, vol. 81, p. 74–84, Mar 2017.

[99] S. Anandasabapathy, S. Michel, and C. Popper, "Mobile clinics." Publication Date: 19 May 2016.

[100] R. A. B. Peri Avitan, "Personal medical device." Publication Date: 05 Oct. 2017.

[101] P. M. Clarence Groff, "Wearable vital sign monitoring system." Granted Date: 15 Aug. 2000, Publication Date: 15 Aug. 2000.

[102] A. Sobrinho, L. D. da Silva, A. Perkusich, M. E. Pinheiro, and P. Cunha, "Design and evaluation of a mobile application to assist the self-monitoring of the chronic kidney disease in developing countries," *BMC Med. Inf. & Decision Making*, vol. 18, no. 1, pp. 7:1–7:14, 2018.

[103] A. K. Kshepakaran, G. H. Pelton, J. J. Jacobsen, Y. J. Fanusie, M. J. McShea, and P. A. O'shea, "Mobile health management database, targeted educational assistance (tea) engine, selective health care data sharing, family tree graphical user interface, and health journal social network wall feed, computer-implemented system, method and computer program product." Publication Date: 21 Dez. 2017.

[104] G. Ross, "Point-of-care workstation/cart with smartphone interface." Publication Date: 20 Apr. 2017.

[105] Santos, D. F. S. , Martins, A. F. , Rodrigues, A. F. A., Nascimento, J. L., Perkusich, A., Almeida, H. O. , "Personal health data hub." Granted Date: 02 Oct. 2018, Publication Date: 10 Jul. 2014.

[106] S. Anandasabapathy, S. Michel, and C. Popper, "Dynamic threading gateway for embedded health management systems." Publication Date: 30 Oct. 2013.

[107] D. F. S. Santos, K. C. Gorgônio, A. Perkusich, and H. O. Almeida, "A standard-based and context-aware architecture for personal healthcare smart gateways," *J. Medical Systems*, vol. 40, no. 10, pp. 224:1–224:14, 2016.

[108] Amy Papadopoulos, Cindy Crump, Bruce Wilson, "Mobile wireless customizable health and condition monitor." Granted Date: 31 Dec. 2013, Publication Date: 31 Dec. 2013.

[109] Stephen Jacobsen, Tomasz Petelenz, Stephen Peterson, Roland Wyatt, "System for remote monitoring of personnel." Granted Date: 06 Mar. 2001, Publication Date: 10 Jun. 1998.

[110] Taiwan Emergency Man Assotiation, "Disaster area medical system." Publication Date: 11 Aug. 2019.

[111] D. F. S. Santos, H. O. Almeida, and A. Perkusich, "A personal connected health system for the internet of things based on the constrained application protocol," *Comput. Electr. Eng.*, vol. 44, pp. 122–136, 2015.

[112] Z. Shelby, K. Hartke, and C. Bormann, "The Constrained Application Protocol (CoAP)," Jun., 2014.

[113] F. A. Kraemer, A. E. Braten, N. Tamkittikhun, and D. Palma, "Fog computing in healthcare–a review and discussion," *IEEE Access*, vol. 5, pp. 9206–9222, 2017.

[114] Red Hat, "Virtualização: o que é, como funciona e quais os seus benefícios." Disponível em: https://www.redhat.com/pt-br/topics/virtualization/what-is-virtualization. Acessado em 30 de Agosto de 2020.

[115] T. Bui, "Analysis of docker security," 01 2015.

[116] D. Bernstein, "Containers and cloud: From lxc to docker to kubernetes," *IEEE Cloud Computing*, vol. 1, no. 3, pp. 81–84, 2014.

[117] N. Naik, "Building a virtual system of systems using docker swarm in multiple clouds," in *2016 IEEE International Symposium on Systems Engineering (ISSE)*, 2016.

[118] Docker, "Docker for the Virtualization Admin." Disponível em: https://goto.docker.com/rs/929-FJL-178/images/docker-for-the-virtualization-admin.pdf. Acessado em 30 de Agosto de 2020.

[119] W. Li and A. Kanso, "Comparing containers versus virtual machines for achieving high availability," in *2015 IEEE International Conference on Cloud Engineering*, 2015.

[120] VM Ware, "Virtualização." Disponível em: https://www.vmware.com/br/solutions/virtualization.htm Acessado em 30 de Agosto de 2020.

[121] A. Przybyłek, "An empirical study on the impact of AspectJ on software evolvability," *Empirical Software Engineering*, vol. 23, pp. 2018–2050, Dec. 2017.

[122] A. Przybyłek, "Where the truth lies: AOP and its impact on software modularity," in *Fundamental Approaches to Software Engineering*, pp. 447–461, Springer Berlin Heidelberg, 2011.

[123] G. Blinowski, A. Ojdowska, and A. Przybylek, "Monolithic vs. microservice architecture: A performance and scalability evaluation," *IEEE Access*, vol. 10, pp. 20357–20374, 2022.

[124] C. M. Aderaldo, N. C. Mendonca, C. Pahl, and P. Jamshidi, "Benchmark requirements for microservices architecture research," in *2017 IEEE/ACM 1st International Workshop on Establishing the Community-Wide Infrastructure for Architecture-Based Software Engineering (ECASE)*, IEEE, May 2017.

[125] Y. Wang, H. Kadiyala, and J. Rubin, "Promises and challenges of microservices: an exploratory study," *Empirical Software Engineering*, vol. 26, no. 4, pp. 1–44, 2021.

[126] M. Viggiato, R. Terra, H. Rocha, M. T. Valente, and E. Figueiredo, "Microservices in practice: A survey study," 2018.

[127] M. Jagiełło, M. Rusek, and W. Karwowski, "Performance and resilience to failures of an cloud-based application: Monolithic and microservices-based architectures compared," in *Computer Information Systems and Industrial Management*, pp. 445–456, Springer International Publishing, 2019.

[128] J. Fritzsch, J. Bogner, S. Wagner, and A. Zimmermann, "Microservices migration in industry: Intentions, strategies, and challenges," in *2019 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, IEEE, Sept. 2019.

[129] A. Poniszewska-Marańda and E. Czechowska, "Kubernetes cluster for automating software production environment," *Sensors*, vol. 21, p. 1910, Mar. 2021.

[130] J. Lewis and M. Fowler, "Microservices: a definition of this new architectural term." Disponível em: https://www.martinfowler.com/articles/microservices.html. Acessado em 02 de Julho de 2022.

[131] J. Ghofrani and A. Bozorgmehr, "Migration to microservices: Barriers and solutions," in *International Conference on Applied Informatics*, pp. 269–281, Springer, 2019.

[132] O. Al-Debagy and P. Martinek, "A comparative review of microservices and monolithic architectures," in *2018 IEEE 18th International Symposium on Computational Intelligence and Informatics (CINTI)*, pp. 000149–000154, IEEE, 2018.

[133] A. de Camargo, I. Salvadori, R. dos Santos Mello, and F. Siqueira, "An architecture to automate performance tests on microservices," in *Proceedings of the 18th International Conference on Information Integration and Web-based Applications and Services*, ACM, Nov. 2016.

[134] V. Lenarduzzi, F. Lomio, N. Saarimäki, and D. Taibi, "Does migrating a monolithic system to microservices decrease the technical debt?," *Journal of Systems and Software*, vol. 169, p. 110710, Nov. 2020.

[135] M. Štefanko, O. Chaloupka, and B. Rossi, "The saga pattern in a reactive microservices environment," in *Proceedings of the 14th International Conference on Software Technologies*, SCITEPRESS - Science and Technology Publications, 2019.

[136] M. Ianculescu, A. Alexandru, G. Neagu, and F. Pop, "Microservice-based approach to enforce an IoHT oriented architecture," in *2019 E-Health and Bioengineering Conference (EHB)*, IEEE, Nov. 2019.

[137] S. Yi, C. Li, and Q. Li, "A survey of fog computing: concepts, applications and issues," in *Proceedings of the 2015 workshop on mobile big data*, pp. 37–42, 2015.

[138] C.-H. Hong and B. Varghese, "Resource management in fog/edge computing: a survey on architectures, infrastructure, and algorithms," *ACM Computing Surveys (CSUR)*, vol. 52, no. 5, pp. 1–37, 2019.

[139] B. Costa, J. Bachiega Jr, L. R. de Carvalho, and A. P. Araujo, "Orchestration in fog computing: A comprehensive survey," *ACM Computing Surveys (CSUR)*, vol. 55, no. 2, pp. 1–34, 2022.

[140] Kubernetes, "Kubernetes." Disponível em: https://kubernetes.io/. Acessado em 02 de Julho de 2022.

[141] Docker Swarm, "Swarm mode overview | Docker Documentation." Disponível em: https://docs.docker.com/engine/swarm/. Acessado em 02 de Julho de 2022.

[142] NOMAD, "Nomad by HashiCorp." Disponível em: https://www.nomadproject.io/. Acessado em 02 de Julho de 2022.

[143] Marathon, "Marathon: A container orchestration platform for Mesos and DC/OS." Disponível em: https://mesosphere.github.io/marathon/. Acessado em 02 de Julho de 2022.

[144] Kubernetes, "Overview." Disponível em: https://kubernetes.io/docs/concepts/overview/. Acessado em 01 de Maio de 2024.

[145] Red Hat, "O que é Kubernetes?." Disponível em: https://www.redhat.com/pt-br/topics/containers/what-is-kubernetes. Acessado em 01 de Maio de 2024.

[146] Kubernetes, "Conceitos." Disponível em: https://kubernetes.io/pt/docs/concepts/. Acessado em 01 de Maio de 2024.

[147] Kubernetes, "Kubernetes Components." Disponível em: https://kubernetes.io/docs/concepts/overview/components/. Acessado em 01 de Maio de 2024.

[148] Kubernetes, *Kubernetes Components*, n.d. Accessed: 2024-06-04.

[149] C. Kotronis, G. Minou, G. Dimitrakopoulos, M. Nikolaidou, D. Anagnostopoulos, A. Amira, F. Bensaali, H. Baali, and H. Djelouat, "Managing criticalities of e-health IoT systems," in *2017 IEEE 17th International Conference on Ubiquitous Wireless Broadband (ICUWB)*, IEEE, Sept. 2017.

[150] E. Liu, E. Effiok, and J. Hitchcock, "Survey on health care applications in 5g networks," *IET Communications*, vol. 14, pp. 1073–1080, Apr. 2020.

[151] D. Soldani, "Fighting pandemics by exploiting 5g, AI and bigdata enabled technologies," *Journal of Telecommunications and the Digital Economy*, vol. 8, pp. 146–158, June 2020.

[152] O. Akrivopoulos, D. Amaxilatis, A. Antoniou, and I. Chatzigiannakis, "Design and evaluation of a person-centric heart monitoring system over fog computing infrastructure," in ., HumanSys'17, (New York, NY, USA), p. 25–30, Association for Computing Machinery, 2017.

[153] E. S. Pramukantoro and A. Gofuku, "Prototype of multi-layer personal cardiac monitoring system for data interoperability problem," in *Proceedings of the 5th International Conference on Sustainable Information Engineering and Technology*, SIET '20, (New York, NY, USA), p. 84–89, Association for Computing Machinery, 2020.

[154] H. Ozkan, O. Ozhan, Y. Karadana, M. Gulcu, S. Macit, and F. Husain, "A portable wearable tele-ECG monitoring system," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, pp. 173–182, Jan. 2020.

[155] F. Concone, G. L. Re, and M. Morana, "A fog-based application for human activity recognition using personal smart devices," *ACM Transactions on Internet Technology*, vol. 19, p. 1–20, Mar. 2019.

[156] M. Inoue, R. Taguchi, and T. Umezaki, "Vision-based bed detection for hospital patient monitoring system," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, IEEE, July 2018.

[157] F. Cardile, G. Iannizzotto, and F. L. Rosa, "A vision-based system for elderly patients monitoring," in *3rd International Conference on Human System Interaction*, IEEE, May 2010.

[158] Y. Kurylyak, F. Lamonaca, and G. Mirabelli, "Detection of the eye blinks for human's fatigue monitoring," in *2012 IEEE International Symposium on Medical Measurements and Applications Proceedings*, IEEE, May 2012.

[159] D. Brulin, Y. Benezeth, and E. Courtial, "Posture recognition based on fuzzy logic for home monitoring of the elderly," *IEEE Transactions on Information Technology in Biomedicine*, vol. 16, pp. 974–982, Sept. 2012.

[160] K. Gu, Y. Zhang, and J. Qiao, "Vision-based monitoring of flare soot," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, pp. 7136–7145, Sept. 2020.

[161] L. Jurišica, F. Duchoň, M. Dekan, A. Babinec, and P. Pásztó, "General concepts of teleoperated systems," in *2018 ELEKTRO*, pp. 1–4, 2018.

[162] N. Ishikawa and G. Watanabe, "Ultra-minimally invasive cardiac surgery: robotic surgery and awake CABG," *Surgery Today*, vol. 45, pp. 1–7, Oct. 2014.

[163] R. W. Dobbs, W. R. Halgrimson, S. Talamini, H. T. Vigneswaran, J. O. Wilson, and S. Crivellaro, "Single-port robotic surgery: the next generation of minimally invasive urology," *World Journal of Urology*, vol. 38, pp. 897–905, Aug. 2019.

[164] L. CHOHAN and J. B. NIJJAR, "Minimally invasive surgery in pregnancy," *Clinical Obstetrics and Gynecology*, vol. 63, pp. 379–391, Mar. 2020.

[165] Mater Private, "Robotic Surgery - da Vinci Surgical System." Accessed: 2021-09-17.

[166] J. L. Martin and J. Barnett, "Integrating the results of user research into medical device development: insights from a case study," *BMC Medical Informatics and Decision Making*, vol. 12, July 2012.

[167] J. L. Martin, D. J. Clark, S. P. Morgan, J. A. Crowe, and E. Murphy, "A user-centred approach to requirements elicitation in medical device development: A case study from an industry perspective," *Applied Ergonomics*, vol. 43, pp. 184–190, Jan. 2012.

[168] Lev Med, "Mobile ECG." Accessed: 2021-09-17.

[169] K. Murugan, S. Murugeswari, J. P. Reddy, M. H. Chandra, and P. V. Reddy, "Smart medical telemetry acquisition system," in *2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC)*, IEEE, Aug. 2021.

[170] O. Oti, I. Azimi, A. Anzanpour, A. M. Rahmani, A. Axelin, and P. Liljeberg, "IoT-based healthcare system for real-time maternal stress monitoring," in *Proceedings of the 2018 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies*, ACM, Sept. 2018.

[171] WHO Global Observatory for eHealth, "Telemedicine: opportunities and developments in member states: report on the second global survey on ehealth," 2010.

[172] P. Ranaweera, M. Liyanage, and A. D. Jurcut, "Novel MEC based approaches for smart hospitals to combat COVID-19 pandemic," *IEEE Consumer Electronics Magazine*, vol. 10, pp. 80–91, Mar. 2021.

[173] G.-Z. Yang, B. J. Nelson, R. R. Murphy, H. Choset, H. Christensen, S. H. Collins, P. Dario, K. Goldberg, K. Ikuta, N. Jacobstein, D. Kragic, R. H. Taylor, and M. McNutt, "Combating COVID-19—the role of robotics in managing public health and infectious diseases," *Science Robotics*, vol. 5, Mar. 2020.

[174] N. Alshurafa, C. Sideris, M. Pourhomayoun, H. Kalantarian, M. Sarrafzadeh, and J.-A. Eastwood, "Remote health monitoring outcome success prediction using baseline and first month intervention data," *IEEE Journal of Biomedical and Health Informatics*, vol. 21, pp. 507–514, Mar. 2017.

[175] N. R. Praneeth, J. Nagaswetha, P. Meghana, M. Sushma, and G. J. Lakshmi, "Smart remote health monitoring system classification using fuzzy c-means," in *2021 2nd International Conference on Smart Electronics and Communication (ICOSEC)*, IEEE, Oct. 2021.

[176] E. Baba, A. Jilbab, and A. Hammouch, "A health remote monitoring application based on wireless body area networks," in *2018 International Conference on Intelligent Systems and Computer Vision (ISCV)*, IEEE, Apr. 2018.

[177] C. Raj, C. Jain, and W. Arif, "HEMAN: Health monitoring and nous: An IoT based e-health care system for remote telemedicine," in *2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, IEEE, Mar. 2017.

[178] Y. Huang, P. Yang, and Z. Zhang, "An ECG acquisition system with piezoelectric energy harvesting for low power healthcare devices," in *2021 IEEE 14th International Conference on ASIC (ASICON)*, IEEE, Oct. 2021.

[179] Z. Lin, Z. Hong, F. Ye, W. Qin, X. Cao, Y. Wang, R. Hu, R. Yan, Y. Qin, and T. Yi, "A low-power, wireless, real-time, wearable healthcare system," in *2016 IEEE MTT-S International Wireless Symposium (IWS)*, IEEE, Mar. 2016.

[180] W. Cai, "Research on remote monitoring method of body state and physiological parameters during fitness," in *2020 International Conference on Wireless Communications and Smart Grid (ICWCSG)*, IEEE, June 2020.

[181] I. P. Korneeva, K. A. Kramar, T. M. Magrupov, and E. A. Semenova, "Using of a portable urine analyzer for remote health assessment," in *2021 International Conference on Information Science and Communications Technologies (ICISCT)*, IEEE, Nov. 2021.

[182] D. Li, "5g and intelligence medicine—how the next generation of wireless technology will reconstruct healthcare?," *Precision Clinical Medicine*, vol. 2, pp. 205–208, Oct. 2019.

[183] S. Hamm, A.-C. Schleser, J. Hartig, P. Thomas, S. Zoesch, and C. Bulitta, "5g as enabler for digital healthcare," *Current Directions in Biomedical Engineering*, vol. 6, pp. 1–4, Sept. 2020.

[184] P. T and S. S. Nayak, "5g technology for e-health," in *2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, IEEE, Oct. 2020.

[185] H. N. Qureshi, M. Manalastas, S. M. A. Zaidi, A. Imran, and M. O. A. Kalaa, "Service level agreements for 5g and beyond: Overview, challenges and enablers of 5g-healthcare systems," *IEEE Access*, vol. 9, pp. 1044–1061, 2021.

[186] G. Cisotto, E. Casarin, and S. Tomasin, "Requirements and enablers of advanced healthcare services over future cellular systems," *IEEE Communications Magazine*, vol. 58, p. 76–81, Mar. 2020.

[187] M. Zanatta, "Pre-hospital ultrasound: Current indications and future perspectives," *International Journal of Critical Care and Emergency Medicine*, vol. 2, Dec. 2016.

[188] R. L. McNamara, Y. Wang, J. Herrin, J. P. Curtis, E. H. Bradley, D. J. Magid, E. D. Peterson, M. Blaney, P. D. Frederick, and H. M. Krumholz, "Effect of door-to-balloon time on mortality in patients with ST-segment elevation myocardial infarction," *Journal of the American College of Cardiology*, vol. 47, pp. 2180–2186, June 2006.

[189] C. P. Cannon, "Relationship of symptom-onset-to-balloon time and door-to-balloon time with mortality in patients undergoing angioplasty for acute myocardial infarction," *JAMA*, vol. 283, p. 2941, June 2000.

[190] J. Park, K. H. Choi, J. M. Lee, H. K. Kim, D. Hwang, T.-M. Rhee, J. Kim, T. K. Park, J. H. Yang, Y. B. Song, J.-H. Choi, J.-Y. Hahn, S.-H. Choi, B.-K. Koo, S. C. Chae, M. C. Cho, C. J. Kim, J. H. Kim, M. H. Jeong, H.-C. Gwon, and H.-S. K. and, "Prognostic implications of door-to-balloon time and onset-to-door time on mortality in patients with ST-segment–elevation myocardial infarction treated with primary percutaneous coronary intervention," *Journal of the American Heart Association*, vol. 8, May 2019.

[191] M. A. Usman, N. Y. Philip, and C. Politis, "5g enabled mobile healthcare for ambulances," in *2019 IEEE Globecom Workshops (GC Wkshps)*, IEEE, Dec. 2019.

[192] S. Yu, F. Yi, X. Qiulin, and S. Liya, "A framework of 5g mobile-health services for ambulances," in *2020 IEEE 20th International Conference on Communication Technology (ICCT)*, IEEE, Oct. 2020.

[193] B. Charyyev, E. Arslan, and M. H. Gunes, "Latency comparison of cloud datacenters and edge servers," in *GLOBECOM 2020 - 2020 IEEE Global Communications Conference*, IEEE, Dec. 2020.

[194] A. Li, X. Yang, S. Kandula, and M. Zhang, "CloudCmp: comparing public cloud providers," in *Proceedings of the 10th annual conference on Internet measurement - IMC '10*, ACM Press, 2010.

[195] T. Vu, C. J. Mediran, and Y. Peng, "Measurement and observation of cross-provider cross-region latency for cloud-based IoT systems," in *2019 IEEE World Congress on Services (SERVICES)*, IEEE, July 2019.

[196] Z. Chen, W. Hu, J. Wang, S. Zhao, B. Amos, G. Wu, K. Ha, K. Elgazzar, P. Pillai, R. Klatzky, D. Siewiorek, and M. Satyanarayanan, "An empirical study of latency in an emerging class of edge computing applications for wearable cognitive assistance,"

in *Proceedings of the Second ACM/IEEE Symposium on Edge Computing*, ACM, Oct. 2017.

[197] T. K. Hoey, G. R. Sinniah, R. Khoshdelniat, A. Abdullah, and S. Subramaniam, "Mobility management schemes and mobility issues in low power wireless sensor network," in *2012 International Symposium on Telecommunication Technologies*, IEEE, Nov. 2012.

[198] C. Meshram, C.-C. Lee, S. G. Meshram, R. J. Ramteke, and A. Meshram, "An efficient mobile-healthcare emergency framework," *Journal of Medical Systems*, vol. 44, Jan. 2020.

[199] H. Rogers, K. C. Madathil, A. Joseph, C. Holmstedt, S. Qanungo, N. McNeese, T. Morris, R. J. Holden, and J. T. McElligott, "An exploratory study investigating the barriers, facilitators, and demands affecting caregivers in a telemedicine integrated ambulance-based setting for stroke care," *Applied Ergonomics*, vol. 97, p. 103537, Nov. 2021.

[200] A. Joseph, K. Chalil Madathil, R. Jafarifiroozabadi, H. Rogers, S. Mihandoust, A. Khasawneh, N. McNeese, C. Holmstedt, and J. T. McElligott, "Communication and teamwork during telemedicine-enabled stroke care in an ambulance," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 64, p. 21–41, Mar. 2021.

[201] J. Kandimalla, A. R. Vellipuram, G. Rodriguez, A. Maud, S. Cruz-Flores, and R. Khatri, "Role of telemedicine in prehospital stroke care," *Current Cardiology Reports*, vol. 23, May 2021.

[202] H. Kim, S.-W. Kim, E. Park, J. H. Kim, and H. Chang, "The role of fifth-generation mobile technology in prehospital emergency care: An opportunity to support paramedics," *Health Policy and Technology*, vol. 9, pp. 109–114, Mar. 2020.

[203] M. Abdeen, M. H. Ahmed, H. Seliem, M. El-Nainay, and T. R. Sheltami, "Improving the performance of ambulance emergency service using smart health systems,"

in *2021 IEEE/ACM Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*, IEEE, Dec. 2021.

[204] M. D. Kamal, A. Tahir, M. B. Kamal, and M. A. Naeem, "Future location prediction for emergency vehicles using big data: A case study of healthcare engineering," *Journal of Healthcare Engineering*, vol. 2020, pp. 1–11, Nov. 2020.

[205] I. U. Rehman, M. M. Nasralla, A. Ali, and N. Philip, "Small cell-based ambulance scenario for medical video streaming: A 5g-health use case," in *2018 15th International Conference on Smart Cities: Improving Quality of Life Using ICT & IoT (HONET-ICT)*, IEEE, Oct. 2018.

[206] A. Bujari, C. E. Palazzi, D. Polonio, and M. Zanella, "Service function chaining: a lightweight container-based management and orchestration plane," in *2019 16th IEEE Annual Consumer Communications & Networking Conference (CCNC)*, IEEE, Jan. 2019.

[207] RedHat, "What is orchestration?." Disponível em: https://www.redhat.com/pt-br/topics/automation/what-is-orchestration. Acessado em 11 de Setembro de 2022.

[208] R. Kamal and S. Agrawal, "A design framework of orchestrator for computing systems," in *2010 International Conference on Computer Information Systems and Industrial Management Applications (CISIM)*, IEEE, Oct. 2010.

[209] K. Velasquez, D. P. Abreu, M. R. M. Assis, C. Senna, D. F. Aranha, L. F. Bittencourt, N. Laranjeiro, M. Curado, M. Vieira, E. Monteiro, and E. Madeira, "Fog orchestration for the internet of everything: state-of-the-art and research challenges," *Journal of Internet Services and Applications*, vol. 9, July 2018.

[210] N. C. Fakude, P. Tarwireyi, M. O. Adigun, and A. M. Abu-Mahfouz, "Fog orchestrator as an enabler for security in fog computing: A review," in *2019 International Multidisciplinary Information Technology and Engineering Conference (IMITEC)*, IEEE, Nov. 2019.

[211] B. Wilson, "16 Best Container Orchestration Tools and Services." Disponível em:

https://devopscube.com/docker-container-clustering-tools/. Acessado em 11 de Setembro de 2022.

[212] L. R. de Carvalho and A. P. F. de Araujo, "Performance comparison of terraform and cloudify as multicloud orchestrators," in *2020 20th IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing (CCGRID)*, IEEE, May 2020.

[213] COLA - Cloud Orchestration at the Level of Application, "D4.2 requirements gathering and performance benchmarking of microservices," 2018.

[214] T. Kudla, M. Fogli, S. Webb, G. Pingen, N. Suri, and H. Bastiaansen, "Quantifying the performance of cloud-oriented container orchestrators on emulated tactical networks," *IEEE Communications Magazine*, vol. 60, pp. 74–80, May 2022.

[215] Kubernetes, "Kubernetes Performance Measurements and Roadmap." Disponível em: https://kubernetes.io/blog/2015/09/kubernetes-performance-measurements-and/. Acessado em: 11 de Setembro de 2022.

[216] N. Arya, "Energy efficient task offloading in mobile based cloud computing environment," in *2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, IEEE, Sept. 2021.

[217] H. Yu, Q. Wang, and S. Guo, "Energy-efficient task offloading and resource scheduling for mobile edge computing," in *2018 IEEE International Conference on Networking, Architecture and Storage (NAS)*, IEEE, Oct. 2018.

[218] P. Kayal and J. Liebeherr, "Autonomic service placement in fog computing," in *2019 IEEE 20th International Symposium on "A World of Wireless, Mobile and Multimedia Networks"(WoWMoM)*, IEEE, June 2019.

[219] F. A. Salaht, F. Desprez, and A. Lebre, "An overview of service placement problem in fog and edge computing," *ACM Computing Surveys*, vol. 53, pp. 1–35, May 2021.

[220] A. Yousefpour, C. Fung, T. Nguyen, K. Kadiyala, F. Jalali, A. Niakanlahiji, J. Kong, and J. P. Jue, "All one needs to know about fog computing and related edge computing paradigms: A complete survey," *Journal of Systems Architecture*, vol. 98, pp. 289–330, Sept. 2019.

[221] R. Mahmud, R. Kotagiri, and R. Buyya, "Fog computing: A taxonomy, survey and future directions," in *Internet of Things*, pp. 103–130, Springer Singapore, Oct. 2017.

[222] Y. Gong, "Optimal edge server and service placement in mobile edge computing," in *2020 IEEE 9th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*, IEEE, Dec. 2020.

[223] Z. Ning, P. Dong, X. Wang, S. Wang, X. Hu, S. Guo, T. Qiu, B. Hu, and R. Y. K. Kwok, "Distributed and dynamic service placement in pervasive edge computing networks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 32, pp. 1277–1292, June 2021.

[224] H. T. Malazi and S. Clarke, "Distributed service placement and workload orchestration in a multi-access edge computing environment," in *2021 IEEE International Conference on Services Computing (SCC)*, IEEE, Sept. 2021.

[225] N. Aljeri and A. Boukerche, "Mobility management in 5g-enabled vehicular networks," *ACM Computing Surveys*, vol. 53, pp. 1–35, Oct. 2020.

[226] T. Bahreini and D. Grosu, "Efficient placement of multi-component applications in edge computing systems," in *Proceedings of the Second ACM/IEEE Symposium on Edge Computing*, ACM, Oct. 2017.

[227] E. Saurez, K. Hong, D. Lillethun, U. Ramachandran, and B. Ottenwälder, "Incremental deployment and migration of geo-distributed situation awareness applications in the fog," in *Proceedings of the 10th ACM International Conference on Distributed and Event-based Systems*, ACM, June 2016.

[228] B. Ottenwälder, B. Koldehofe, K. Rothermel, and U. Ramachandran, "MigCEP," in *Proceedings of the 7th ACM international conference on Distributed event-based systems - DEBS '13*, ACM Press, 2013.

[229] K. Hong, D. Lillethun, U. Ramachandran, B. Ottenwälder, and B. Koldehofe, "Mobile fog," in *Proceedings of the second ACM SIGCOMM workshop on Mobile cloud computing - MCC '13*, ACM Press, 2013.

[230] R. Urgaonkar, S. Wang, T. He, M. Zafer, K. Chan, and K. K. Leung, "Dynamic service migration and workload scheduling in edge-clouds," *Performance Evaluation*, vol. 91, pp. 205–228, Sept. 2015.

[231] K. Velasquez, D. P. Abreu, M. Curado, and E. Monteiro, "Service placement for latency reduction in the internet of things," *Annals of Telecommunications*, vol. 72, pp. 105–115, June 2016.

[232] J. Santos, T. Wauters, B. Volckaert, and F. D. Turck, "Resource provisioning for IoT application services in smart cities," in *2017 13th International Conference on Network and Service Management (CNSM)*, IEEE, Nov. 2017.

[233] A. Chakraborty, S. Misra, and J. Maiti, "Mobility-aware controller orchestration in multi-tier service-oriented architecture for IoT," *IEEE Transactions on Vehicular Technology*, vol. 71, pp. 1820–1831, Feb. 2022.

[234] Z. Zhang, J. Brazil, M. Ozkaynak, and K. Desanto, "Evaluative research of technologies for prehospital communication and coordination: a systematic review," *Journal of Medical Systems*, vol. 44, Apr. 2020.

[235] E. Park, J. H. Kim, H. S. Nam, and H.-J. Chang, "Requirement analysis and implementation of smart emergency medical services," *IEEE Access*, vol. 6, p. 42022–42029, 2018.

[236] S. T. Ahmed, S. M. Basha, M. Ramachandran, M. Daneshmand, and A. H. Gandomi, "An edge-ai-enabled autonomous connected ambulance-route resource recommendation protocol (aca-r3) for ehealth in smart cities," *IEEE Internet of Things Journal*, vol. 10, p. 11497–11506, July 2023.

[237] M. Z. Chowdhury, M. T. Hossan, M. Shahjalal, M. K. Hasan, and Y. M. Jang, "A new 5g ehealth architecture based on optical camera communication: An overview, prospects, and applications," *IEEE Consumer Electronics Magazine*, vol. 9, p. 23–33, Nov. 2020.

[238] R. Singh, K. D. Ballal, S. C. Nwabuona, M. S. Berger, L. Dittmann, S. Ruepp, and T. Wienecke, "Assessment of cellular coverage for a smart ambulance use case," in

2022 IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS), IEEE, Dec. 2022.

[239] C. Comito, D. Falcone, and A. Forestiero, "Current trends and practices in smart health monitoring and clinical decision support," in *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, IEEE, Dec. 2020.

[240] C. G. Fidalgo, Y. Yan, H. Cho, M. Sousa, D. Lindlbauer, and J. Jorge, "A survey on remote assistance and training in mixed reality environments," *IEEE Transactions on Visualization and Computer Graphics*, vol. 29, p. 2291–2303, May 2023.

[241] H.-J. Guo, J. Z. Bakdash, L. R. Marusich, and B. Prabhakaran, "Augmented reality and mixed reality measurement under different environments: A survey on head-mounted devices," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, p. 1–15, 2022.

[242] M. Mehrnoush, C. Hu, and C. Aldana, "Ar/vr spectrum requirement for wi-fi 6e and beyond," *IEEE Access*, vol. 10, p. 133016–133026, 2022.

[243] K. Ishikawa, Y. Yanagawa, S. Ota, K. Muramatsu, H. Nagasawa, K. Jitsuiki, H. Ohsaka, T. Nara, Y. Nishizaki, and H. Daida, "Preliminary study of prehospital use of smart glasses," *Acute Medicine & Surgery*, vol. 9, Jan. 2022.

[244] G. Bansal, K. Rajgopal, V. Chamola, Z. Xiong, and D. Niyato, "Healthcare in metaverse: A survey on current metaverse applications in healthcare," *IEEE Access*, vol. 10, p. 119914–119946, 2022.

[245] X. Qiao, P. Ren, S. Dustdar, L. Liu, H. Ma, and J. Chen, "Web ar: A promising future for mobile augmented reality — state of the art, challenges, and insights," *Proceedings of the IEEE*, vol. 107, p. 651–666, Apr. 2019.

[246] Y. Sun, J. Chen, Z. Wang, M. Peng, and S. Mao, "Enabling mobile virtual reality with open 5g, fog computing and reinforcement learning," *IEEE Network*, vol. 36, p. 142–149, Nov. 2022.

[247] S. Karunarathna, S. Wijethilaka, P. Ranaweera, K. T. Hemachandra, T. Samarasinghe, and M. Liyanage, "The role of network slicing and edge computing in the metaverse realization," *IEEE Access*, vol. 11, p. 25502–25530, 2023.

[248] H. Zhang, S. Mao, D. Niyato, and Z. Han, "Location-dependent augmented reality services in wireless edge-enabled metaverse systems," *IEEE Open Journal of the Communications Society*, vol. 4, p. 171–183, 2023.

[249] P. Rajpurkar, E. Chen, O. Banerjee, and E. J. Topol, "Ai in health and medicine," *Nature Medicine*, vol. 28, p. 31–38, Jan. 2022.

[250] J. Bajwa, U. Munir, A. Nori, and B. Williams, "Artificial intelligence in healthcare: transforming the practice of medicine," *Future Healthcare Journal*, vol. 8, p. e188–e194, July 2021.

[251] M. A. Uddin, A. Stranieri, I. Gondal, and V. Balasubramanian, "Rapid health data repository allocation using predictive machine learning," *Health Informatics Journal*, vol. 26, p. 3009–3036, Sept. 2020.

[252] L. Yue, D. Tian, W. Chen, X. Han, and M. Yin, "Deep learning for heterogeneous medical data analysis," *World Wide Web*, vol. 23, p. 2715–2737, Mar. 2020.

[253] M. Sallam, "Chatgpt utility in healthcare education, research, and practice: Systematic review on the promising perspectives and valid concerns," *Healthcare*, vol. 11, p. 887, Mar. 2023.

[254] K. Tsoi, K. Yiu, H. Lee, H. Cheng, T. Wang, J. Tay, B. W. Teo, Y. Turana, A. A. Soenarta, G. P. Sogunuru, S. Siddique, Y. Chia, J. Shin, C. Chen, J. Wang, and K. Kario, "Applications of artificial intelligence for hypertension management," *The Journal of Clinical Hypertension*, vol. 23, p. 568–574, Feb. 2021.

[255] Z. I. Attia, D. M. Harmon, E. R. Behr, and P. A. Friedman, "Application of artificial intelligence to the electrocardiogram," *European Heart Journal*, vol. 42, p. 4717–4730, Sept. 2021.

[256] N. Mouawad, R. Naja, and S. Tohme, *Quality of Service Provisioning for Ambulance Tele-medicine in a Slice-based 5G Network*, p. 73–90. CRC Press, May 2022.

[257] G.-Y. Cho, S.-J. Lee, and T.-R. Lee, "Research on a solution for efficient ecg data transmission in u-healthcare environment," *Journal of Digital Convergence*, vol. 12, no. 1, pp. 397–403, 2014.

[258] A. Charef, Z. Jarir, and M. Quafafou, "Smart system for emergency traffic recommendations: Urban ambulance mobility," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 10, 2022.

[259] A. Botta, W. de Donato, V. Persico, and A. Pescapé, "Integration of cloud computing and internet of things: A survey," *Future Generation Computer Systems*, vol. 56, pp. 684–700, Mar. 2016.

[260] E. Yaacoub, K. Abualsaud, T. Khattab, and A. Chehab, "Secure transmission of iot mhealth patient monitoring data from remote areas using dtn," *IEEE Network*, vol. 34, p. 226–231, Sept. 2020.

[261] J. W. Barrett, "Ambulance clinicians' perspectives of sharing patient information electronically," *British Paramedic Journal*, vol. 4, p. 49–50, Dec. 2019.

[262] M. E. H. Ong, T. F. Chiam, F. S. P. Ng, P. Sultana, S. H. Lim, B. S. Leong, V. Y. K. Ong, E. C. Ching Tan, L. P. Tham, S. Yap, and V. Anantharaman, "Reducing ambulance response times using geospatial–time analysis of ambulance deployment," *Academic Emergency Medicine*, vol. 17, p. 951–957, Sept. 2010.

[263] R. Sonkin, E. Jaffe, O. Wacht, H. Morse, and Y. Bitan, "Real-time video communication between ambulance paramedic and scene – a simulation-based study," *BMC Health Services Research*, vol. 22, Aug. 2022.

[264] O. Ben-Assuli, I. Shabtai, M. Leshno, and S. Hill, "Ehr in emergency rooms: Exploring the effect of key information components on main complaints," *Journal of Medical Systems*, vol. 38, Apr. 2014.

[265] A. V. de Alencar, M. M. Bezerra, D. C. G. Valadares, D. F. S. Santos, and A. Perkusich, *An Interoperable Microservices Architecture for Healthcare Data Exchange*, p. 193–205. Springer International Publishing, 2023.

[266] A. Abelsson, I. Rystedt, B. Suserud, and L. Lindwall, "Learning by simulation in prehospital emergency care – an integrative literature review," *Scandinavian Journal of Caring Sciences*, vol. 30, p. 234–240, Aug. 2015.

[267] B.-M. Gunnarsson and M. Warrén Stomberg, "Factors influencing decision making among ambulance nurses in emergency care situations," *International Emergency Nursing*, vol. 17, p. 83–89, Apr. 2009.

[268] D. Ravi, C. Wong, F. Deligianni, M. Berthelot, J. Andreu-Perez, B. Lo, and G.-Z. Yang, "Deep learning for health informatics," *IEEE Journal of Biomedical and Health Informatics*, vol. 21, p. 4–21, Jan. 2017.

[269] J. Thevenot, M. B. Lopez, and A. Hadid, "A survey on computer vision for assistive medical diagnosis from faces," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, p. 1497–1511, Sept. 2018.

[270] P. Chen, S. Huang, and Q. Yue, "Skin lesion segmentation using recurrent attentional convolutional networks," *IEEE Access*, vol. 10, p. 94007–94018, 2022.

[271] J. Baek and S. Chang, "Individual thermal comfort prediction based on upper body thermal imaging and computer vision," in *2022 3rd International Conference on Human-Centric Smart Environments for Health and Well-being (IHSH)*, IEEE, Oct. 2022.

[272] A. Boukerche, S. Guan, and R. E. De Grande, "A task-centric mobile cloud-based system to enable energy-aware efficient offloading," *IEEE Transactions on Sustainable Computing*, vol. 3, p. 248–261, Oct. 2018.

[273] H. Wu, Y. Sun, and K. Wolter, "Energy-efficient decision making for mobile cloud offloading," *IEEE Transactions on Cloud Computing*, vol. 8, p. 570–584, Apr. 2020.

[274] J. Mhatre, A. Lee, and T. N. Nguyen, "Towards an optimal latency-energy dynamic offloading scheme for collaborative cloud networks," *IEEE Access*, p. 1–1, 2023.

[275] H. Bornholdt, K. Röbert, M. Breitbach, M. Fischer, and J. Edinger, "Measuring the edge: A performance evaluation of edge offloading," in *2023 IEEE International Con-*

ference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops), IEEE, Mar. 2023.

[276] F. Firouzi, S. Jiang, K. Chakrabarty, B. Farahani, M. Daneshmand, J. Song, and K. Mankodiya, "Fusion of iot, ai, edge–fog–cloud, and blockchain: Challenges, solutions, and a case study in healthcare and medicine," *IEEE Internet of Things Journal*, vol. 10, p. 3686–3705, Mar. 2023.

[277] N. Ding, Y. Qin, G. Yang, F. Wei, Z. Yang, Y. Su, S. Hu, Y. Chen, C.-M. Chan, W. Chen, J. Yi, W. Zhao, X. Wang, Z. Liu, H.-T. Zheng, J. Chen, Y. Liu, J. Tang, J. Li, and M. Sun, "Parameter-efficient fine-tuning of large-scale pre-trained language models," *Nature Machine Intelligence*, vol. 5, p. 220–235, Mar. 2023.

[278] J. Qiu, L. Li, J. Sun, J. Peng, P. Shi, R. Zhang, Y. Dong, K. Lam, F. P.-W. Lo, B. Xiao, W. Yuan, N. Wang, D. Xu, and B. Lo, "Large ai models in health informatics: Applications, challenges, and the future," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, p. 6074–6087, Dec. 2023.

[279] H. M. Mentis, I. Avellino, and J. Seo, "Ar hmd for remote instruction in healthcare," in *2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, IEEE, Mar. 2022.

[280] N. Z. Naqvi, K. Moens, A. Ramakrishnan, D. Preuveneers, D. Hughes, and Y. Berbers, "To cloud or not to cloud: a context-aware deployment perspective of augmented reality mobile applications," in *Proceedings of the 30th Annual ACM Symposium on Applied Computing*, SAC 2015, ACM, Apr. 2015.

[281] W. Zhang, B. Han, and P. Hui, "On the networking challenges of mobile augmented reality," in *Proceedings of the Workshop on Virtual Reality and Augmented Reality Network*, SIGCOMM '17, ACM, Aug. 2017.

[282] M. Bender, A. Sudbring, H. A. Kazwini, D. Richards, K. Dwivedi, M. Nayak, and C. McGuire, "Azure network round-trip latency statistics." https://learn.microsoft.com/en-us/azure/networking/azure-network-latency. [Accessed 06-01-2024].

[283] "AWS latency test." https://aws-latency-test.com/. [Accessed 06-01-2024].

[284] "GCPing.com." https://gcping.com/. [Accessed 06-01-2024].

[285] R. F. Mansour, A. E. Amraoui, I. Nouaouri, V. G. Diaz, D. Gupta, and S. Kumar, "Artificial intelligence and internet of things enabled disease diagnosis model for smart healthcare systems," *IEEE Access*, vol. 9, p. 45137–45146, 2021.

[286] K. Toczé, J. Lindqvist, and S. Nadjm-Tehrani, "Performance study of mixed reality for edge computing," in *Proceedings of the 12th IEEE/ACM International Conference on Utility and Cloud Computing*, UCC '19, ACM, Dec. 2019.

[287] L. Zhang, X. Wu, F. Wang, A. Sun, L. Cui, and J. Liu, "Edge-based video stream generation for multi-party mobile augmented reality," *IEEE Transactions on Mobile Computing*, vol. 23, p. 409–422, Jan. 2024.

[288] G. S. Park, R. H. Kim, and H. Song, "Collaborative virtual 3d object modeling for mobile augmented reality streaming services over 5g networks," *IEEE Transactions on Mobile Computing*, vol. 22, p. 3855–3869, July 2023.

[289] J. Kässinger, H. Trötsch, F. Dürr, and J. Edinger, "Simedge: Towards accelerated real-time augmented reality simulations using adaptive smart edge computing," in *Proceedings of the Int'l ACM Conference on Modeling Analysis and Simulation of Wireless and Mobile Systems*, MSWiM '23, ACM, Oct. 2023.

[290] J. Jerald, M. Whitton, and F. P. Brooks, "Scene-motion thresholds during head yaw for immersive virtual environments," *ACM Transactions on Applied Perception*, vol. 9, p. 1–23, Mar. 2012.

[291] K. Raaen and I. Kjellmo, *Measuring Latency in Virtual Reality Systems*, p. 457–462. Springer International Publishing, 2015.

[292] 3GPP, "Study on communication for automation in vertical domains (release 16)," 2020.

[293] I. F. Akyildiz and H. Guo, "Wireless communication research challenges for extended reality (xr)," *ITU Journal on Future and Evolving Technologies*, vol. 3, p. 273–287, Apr. 2022.

[294] A. L. d. Sousa, O. D. OKey, R. L. Rosa, M. Saadi, and D. Z. Rodriguez, "Unified approach to video-based ai inference tasks in augmented reality systems assisted by mobile edge computing," in *2023 International Conference on Software, Telecommunications and Computer Networks (SoftCOM)*, IEEE, Sept. 2023.

[295] Y.-C. Wang, J. Xue, C. Wei, and C. C. J. Kuo, "An overview on generative ai at scale with edge–cloud computing," *IEEE Open Journal of the Communications Society*, vol. 4, p. 2952–2971, 2023.

[296] M. A. Khan, E. Baccour, Z. Chkirbene, A. Erbad, R. Hamila, M. Hamdi, and M. Gabbouj, "A survey on mobile edge computing for video streaming: Opportunities and challenges," *IEEE Access*, vol. 10, p. 120514–120550, 2022.

[297] M. Mehrabi, H. Salah, and F. H. P. Fitzek, "A survey on mobility management for mec-enabled systems," in *2019 IEEE 2nd 5G World Forum (5GWF)*, IEEE, Sept. 2019.

[298] I. Parvez, A. Rahmati, I. Guvenc, A. I. Sarwat, and H. Dai, "A survey on low latency towards 5g: Ran, core network and caching solutions," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 4, p. 3098–3130, 2018.

[299] G. Di Modica, A. Galletta, L. Carnevale, A. Alkhansa, A. Costantini, D. Cesini, P. Bellavista, and M. Villari, "Orchestration of containerized applications in the cloud continuum," in *2023 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*, IEEE, Mar. 2023.

[300] A. Liutkevičius, N. Morkevičius, A. Venčkauskas, and J. Toldinas, "Distributed agent-based orchestrator model for fog computing," *Sensors*, vol. 22, p. 5894, Aug. 2022.

[301] G. Davoli, W. Cerroni, D. Borsatti, M. Valieri, D. Tarchi, and C. Raffaelli, "A fog computing orchestrator architecture with service model awareness," *IEEE Transactions on Network and Service Management*, vol. 19, p. 2131–2147, Sept. 2022.

[302] O. Tembhurne, S. Milmile, G. R. Pathak, A. O. Thakare, and A. Thakare, "An orchestrator: A cloud-based shared-memory multi-user architecture for robotic process automation," *International Journal of Open Source Software and Processes*, vol. 13, p. 1–17, Sept. 2022.

[303] A. Qadeer, A. Waqar Malik, A. Ur Rahman, H. Mian Muhammad, and A. Ahmad, "Virtual infrastructure orchestration for cloud service deployment," *The Computer Journal*, vol. 63, p. 295–307, Dec. 2019.

[304] A. Pandiaraj, N. Vinothkumar, and R. Venkatesan, "Virtual machine migration for infrastructure service in cloud network," in *2022 Smart Technologies, Communication and Robotics (STCR)*, IEEE, Dec. 2022.

[305] B. Nogales, I. Vidal, D. R. Lopez, J. Rodriguez, J. Garcia-Reinoso, and A. Azcorra, "Design and deployment of an open management and orchestration platform for multi-site nfv experimentation," *IEEE Communications Magazine*, vol. 57, p. 20–27, Jan. 2019.

[306] F. Foresta, W. Cerroni, L. Foschini, G. Davoli, C. Contoli, A. Corradi, and F. Callegati, "Improving openstack networking: Advantages and performance of native sdn integration," in *2018 IEEE International Conference on Communications (ICC)*, IEEE, May 2018.

[307] T. Sechkova, M. Paolino, and D. Raho, "Virtualized infrastructure managers for edge computing: Openvim and openstack comparison," in *2018 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, IEEE, June 2018.

[308] F. Asquini, A. Bujari, D. Munaretto, C. E. Palazzi, and D. Ronzani, "An etsi nfv implementation for automatic deployment and configuration of a virtualized mobile core network," in *2021 17th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*, IEEE, Oct. 2021.

[309] I. Quintana-Ramirez, A. Tsiopoulos, M. A. Lema, F. Sardis, L. Sequeira, J. Arias, A. Raman, A. Azam, and M. Dohler, "The making of 5g: Building an end-to-end 5g-

enabled system," *IEEE Communications Standards Magazine*, vol. 2, p. 88–96, Dec. 2018.

[310] S. R. Basnet, R. S. Chaulagain, S. Pandey, and S. Shakya, "Distributed high performance computing in openstack cloud over sdn infrastructure," in *2017 IEEE International Conference on Smart Cloud (SmartCloud)*, IEEE, Nov. 2017.

[311] B. Jiang, Z. Tang, X. Xiao, J. Yao, R. Cao, and K. Li, "Efficient and automated deployment architecture for openstack in tianhe supercomputing environment," *IEEE Transactions on Parallel and Distributed Systems*, vol. 33, p. 1811–1824, Aug. 2022.

[312] G. Turin, A. Borgarelli, S. Donetti, F. Damiani, E. B. Johnsen, and S. L. Tapia Tarifa, "Predicting resource consumption of kubernetes container systems using resource models," *Journal of Systems and Software*, vol. 203, p. 111750, Sept. 2023.

[313] C. Centofanti, W. Tiberti, A. Marotta, F. Graziosi, and D. Cassioli, "Latency-aware kubernetes scheduling for microservices orchestration at the edge," in *2023 IEEE 9th International Conference on Network Softwarization (NetSoft)*, IEEE, June 2023.

[314] W.-K. Lai, Y.-C. Wang, and S.-C. Wei, "Delay-aware container scheduling in kubernetes," *IEEE Internet of Things Journal*, vol. 10, p. 11813–11824, July 2023.

[315] J. Qian, Y. Wang, X. Wang, P. Zhang, and X. Wang, "Load balancing scheduling mechanism for openstack and docker integration," *Journal of Cloud Computing*, vol. 12, Apr. 2023.

[316] S. Yang, X. Wang, X. Wang, L. An, and G. Zhang, "High-performance docker integration scheme based on openstack," *World Wide Web*, vol. 23, p. 2593–2632, Mar. 2020.

[317] O. Tkachova, M. J. Salim, and A. R. Yahya, "An analysis of sdn-openstack integration," in *2015 Second International Scientific-Practical Conference Problems of Infocommunications Science and Technology (PIC ST)*, IEEE, Oct. 2015.

[318] M. Yang and M. Huang, "An microservices-based openstack monitoring tool," in *2019 IEEE 10th International Conference on Software Engineering and Service Science (ICSESS)*, IEEE, Oct. 2019.

[319] "Kubernetes and openstack at yahoo! japan." https://kubernetes.io/blog/2016/10/kubernetes-and-openstack-at-yahoo-japan/. Acesso em: 29/04/2024.

[320] H. R. Kouchaksaraei and H. Karl, "Service function chaining across openstack and kubernetes domains," in *Proceedings of the 13th ACM International Conference on Distributed and Event-based Systems*, DEBS '19, ACM, June 2019.

[321] C. H. C. Jojoa, S. Svorobej, A. Palade, A. Kazmi, and S. Clarke, "MAACO: A dynamic service placement model for smart cities," *IEEE Transactions on Services Computing*, pp. 1–1, 2022.

[322] W. Wang, X. Duan, W. Sun, and M. AI, "Research on mobility prediction in 5g and beyond for vertical industries," in *2021 IEEE/CIC International Conference on Communications in China (ICCC Workshops)*, IEEE, July 2021.

[323] R. Gazda, M. Roy, J. Blakley, A. Sakr, and R. Schuster, "Towards open and cross domain edge emulation – the advantedge platform," in *2021 IEEE/ACM Symposium on Edge Computing (SEC)*, pp. 339–344, 2021.

[324] InterDigital Communications, Inc., "Architecture overview of advantedge," 2022. Acesso em: 16 de abril de 2024.

[325] P. V. Wadatkar, R. G. Garroppo, and G. Nencioni, *MEC Application Migration by Using AdvantEDGE*, p. 104–118. Springer Nature Switzerland, 2023.

[326] InterDigital Communications, Inc., "Api overview of advantedge," 2022. Acesso em: 16 de abril de 2024.