

NOVOS MODELOS BILP PARA A OTIMIZAÇÃO EFICIENTE DA ALOCUÇÃO DE RECURSOS EM SMART CITIES

Novel BILP Models for Efficient Resource Allocation Optimization in Smart Cities

Bruno L. R. e Silva ^{1*}, Viviane C. Santos ¹ e Samuel M. A. Araújo ²

¹ Curso de Engenharia Mecatrônica - Universidade Federal de São João Del-Rei (UFSJ) - Ouro Branco, MG - Brasil

² Departamento de Tecnologia em Engenharia Civil, Computação, Automação, Telemática e Humanidades - UFSJ

*Autor correspondente: brunoluizsilva@aluno.ufsj.edu.br

RESUMO

A vida urbana pode ser transformada de diversas maneiras, com destaque para a integração digital entre redes de computadores, dispositivos inteligentes e o cotidiano das pessoas em ambientes urbanos. Nesse contexto, as tecnologias baseadas na Internet das Coisas (*Internet of Things*, IoT) têm evoluído consideravelmente, impulsionando o desenvolvimento das Cidades Inteligentes (*Smart Cities*) em várias partes do mundo. Dispositivos de IoT atualmente demandam processamento rápido, sendo as tecnologias de *Edge Computing* (EC) e *Cloud Computing* (CC) frequentemente utilizadas para esse fim. Este estudo aborda o problema de alocação de recursos em sistemas de EC e CC para atender às demandas de dispositivos de IoT em uma *Smart City*. No entanto, o arcabouço teórico apresentado pode ser expandido para outros contextos, como regiões agrícolas e parques industriais. Para tal, é apresentada uma definição formal e três modelos matemáticos para o problema de integração de sensores de IoT, EC e CC em *Smart Cities*. Os modelos consideram restrições reais, como capacidade de processamento e capacidade de memória. Eles possuem objetivos distintos: o Modelo M1 busca uma solução viável para o problema, maximizando o atendimento das demandas; o Modelo M2 maximiza o atendimento das demandas enquanto minimiza os custos operacionais da rede; e o Modelo M3 visa uma solução viável para o problema, maximiza o atendimento das demandas, controlando os custos através de um parâmetro α e minimiza o atraso no atendimento, definido como *makespan*. Os modelos apresentados possuem especificidades relevantes que impactam diferentes métricas. A minimização dos custos é benéfica para os provedores envolvidos, especialmente para as empresas de telecomunicações, enquanto a minimização dos atrasos de atendimento beneficia diretamente os usuários finais, impactando métricas relacionadas à Qualidade da Experiência (*Quality of Experience*, QoE).

Palavras-chave: CC, EC, IoT, Modelos Matemáticos, *Smart Cities*.

ABSTRACT

Urban life can be transformed in various ways, with a notable focus on the digital integration between computer networks, smart devices, and daily life in urban environments. In this context, Internet of Things (IoT) technologies have evolved significantly, driving the development of Smart Cities around the world. IoT devices currently require rapid processing, and technologies such as Edge Computing (EC) and Cloud Computing (CC) are frequently employed to meet this need. This study addresses the resource allocation problem in EC and CC systems to meet the demands of IoT devices in a Smart City. However, the theoretical framework presented can be expanded to other contexts, such as agricultural regions and industrial parks. To this end, a formal definition and three mathematical models are proposed for the problem of integrating IoT sensors, EC, and CC in Smart Cities. The models account for real-world constraints such as processing capacity and memory capacity. They have distinct objectives: Model M1 seeks a viable solution to the problem by maximizing demand fulfillment; Model M2 aims to maximize demand fulfillment while minimizing the network's operational costs; and Model M3 seeks a viable solution to the problem by maximizing demand fulfillment, controlling costs through a parameter α , and minimizing the response delay, defined as *makespan*. The presented models have relevant specifics that impact different metrics. Minimizing costs benefits the involved providers, especially telecom companies, while minimizing response delays directly benefits end users by impacting metrics related to Quality of Experience (QoE).

Keywords: CC, EC, IoT, Mathematical Models, *Smart Cities*.

1. Introdução

A vida urbana pode ser transformada de diversas maneiras, com destaque para integração entre redes de computadores, dispositivos inteligentes e ambientes urbanos [1],[2],[3]. Nesse contexto, as tecnologias baseadas em IoT impactam diretamente o cotidiano das pessoas e, quando integradas às *Smart Cities*, promovem melhorias nas atividades diárias, automatizando rotinas e contribuindo para a qualidade de vida [4],[5],[6],[7]. Tais melhorias no contexto urbano representam uma abordagem inovadora para a autonomia da população e a automação urbana [3].

A IoT Inteligentes (*Intelligent IoT*, *IIoT*), por sua vez, é uma subárea da IoT que se concentra em aplicações e dispositivos do mundo real, que interagem por meio de tecnologias de rede. Esses sistemas coletam dados e empregam técnicas de Inteligência Artificial (*Artificial Intelligence*, *AI*) para tomar decisões precisas em diferentes contextos [8]. No contexto dessas tecnologias, um dos setores que têm atraído grande atenção é o da saúde pública, onde esses dispositivos se destacam como ferramentas eficazes no monitoramento de ambientes e indivíduos, visando à melhoria da qualidade de vida [9]. Um exemplo disso são as aplicações de pré-triagem de pacientes, que permitem o monitoramento dos sintomas e o direcionamento adequado para os locais de atendimento apropriados [10]. Além da saúde, a IoT oferece benefícios em diversos outros setores, como a gestão de sistemas de energia, o monitoramento de infraestruturas (por exemplo, pontes) para prever falhas e aumentar a segurança, e o uso de sensores para análise de solos e clima, o que pode auxiliar na agricultura [9],[10]. Essas aplicações ilustram o vasto potencial da IoT para transformar diferentes áreas.

Na Indústria 4.0, por exemplo, a integração da IoT pode otimizar o controle de qualidade, melhorar a segurança dos produtos, facilitar a gestão de inventário e a logística de abastecimento, além de permitir a detecção precoce de falhas [3]. Outro setor beneficiado é a agricultura. Neste caso, a IoT é utilizada para automatizar processos, reduzir custos e maximizar a produção. Essas são algumas das muitas oportunidades proporcionadas pela IoT, com inúmeras outras vantagens ainda a serem exploradas e aplicadas em diferentes contextos.

Dado o impacto e os benefícios que tais tecnologias promovem, o aumento no número de dispositivos IoT conectados em 2023, que alcançou

aproximadamente 16,7 bilhões, não é surpreendente [6]. Além disso, estima-se que o número de dispositivos IoT no mundo cresça ainda mais, chegando a 22 bilhões em 2025 e podendo ultrapassar 29 bilhões até 2030, evidenciando a adoção e integração dessas soluções em diversos contextos de aplicações [5],[11],[12]. Dentre outros fatores, esse crescimento é impulsionado pela premissa de que as tecnologias e aplicações baseadas em IoT possibilitam a interação inteligente entre os usuários e os dispositivos, facilitando as atividades cotidianas e melhorando a qualidade de vida das pessoas [13].

Corroborando as afirmações anteriores, o relatório *Smart City Index* [14] avalia cidades globalmente com base na eficácia de suas tecnologias inteligentes e soluções digitais para melhorar a qualidade de vida dos cidadãos e a eficiência urbana. O relatório revela que diversas cidades ao redor do mundo já alcançaram um alto nível de desenvolvimento em termos de aplicações e dispositivos de IoT, com o objetivo de aprimorar a vida dos seus habitantes. Outras cidades continuam em processo de desenvolvimento. Globalmente, as cidades mais avançadas em tecnologias de IoT são, na ordem, Zurique, na Suíça; Oslo, na Noruega; e Canberra, na Austrália. No Brasil, as cidades com melhores posições são Brasília, São Paulo e Rio de Janeiro, ocupando, respectivamente, a 128ª, 130ª e 136ª posições no ranking mundial.

Neste contexto, a implementação de tecnologias baseadas em IoT exige que dispositivos em diversas áreas de uma cidade, por exemplo, estejam conectados por uma rede confiável [3],[4]. Para oferecer suporte a essa crescente demanda de aplicações, muitos pesquisadores têm concentrado suas investigações no desenvolvimento e na adoção de servidores de nuvem, paradigma conhecido como CC [1],[2],[3]. Além disso, buscando uma redução no atraso de comunicação, as pesquisas também se voltaram para o aprimoramento de tecnologias baseadas em servidores de borda, denominados de EC [9],[15],[16].

Em conjunto, as tecnologias de IoT, EC e CC, quando integradas de maneira adequada, têm o potencial de transformar a operação das cidades, promovendo maior eficiência, sustentabilidade e qualidade de vida para os cidadãos [3]. A implementação dessas soluções inteligentes não só revoluciona a infraestrutura urbana, mas também impacta os processos industriais e outras áreas, com o

objetivo de fortalecer a conectividade e adaptar-se às necessidades humanas [9].

Ainda no contexto da implementação das tecnologias baseadas em IoT, além dos pesquisadores envolvidos, muitas empresas estão ativamente enfrentando e adotando os desafios associados a essa integração de tecnologias. Neste caso, as empresas buscam otimizar métricas como maximizar ganhos e reduzir custos operacionais [1],[3],[5]. Em uma abordagem de comunicações e redes verdes, métricas ambientais também podem ser consideradas, como consumo de energia, reciclabilidade de componentes, uso de energias renováveis e crédito de carbono. Para tal, de modo a atender demandas de mercado cada vez mais rígidas, as infraestruturas baseadas em servidores de EC e CC devem operar e processar as requisições das aplicações oriundas dos dispositivos IoT na rede em tempo real (ou quase real), assegurando tanto a Qualidade de Serviço (*Quality of Service*, QoS) quanto a QoE [4],[16].

1.1. Background

De maneira formal, a IoT é um paradigma que se refere à interconexão de dispositivos físicos e objetos através da internet, possibilitando a coleta, troca e análise de dados sem a necessidade de intervenção humana direta [6]. As aplicações de dispositivos IoT abrangem desde atividades simples até as mais complexas. Exemplos incluem desde o envio de mensagens e o uso de câmeras de segurança até carros autônomos, realização de exames de pré-triagem de saúde e o uso de dispositivos de realidade virtual e aumentada [1],[10],[17].

Cada dispositivo IoT possui demandas específicas, que devem ser atendidas [3],[4],[9]. Neste artigo, são considerados requisitos como armazenamento, capacidade de processamento, cobertura de conexão, memória e largura de banda. Além de verificar se essas restrições são satisfatoriamente atendidas, garantindo aspectos de QoS, também é analisado o custo operacional da rede, denominado *Operational Expenditure* (OpEx).

Neste cenário de integração entre IoT, EC e CC, as demandas geradas por sensores de IoT são enviadas para serem processadas em servidores da rede. Neste contexto, EC refere-se a uma infraestrutura formada por servidores que processam os dados localmente, próximos à origem dos dados, mitigando, assim, problemas relacionados a atrasos de comunicação

[4],[2]. No entanto, essa abordagem tem limitações e não consegue lidar com o crescimento exponencial das demandas geradas [18]. Outra camada de rede considerada neste artigo é a de CC. Neste caso, tal camada utiliza a internet para hospedar recursos e programas, eliminando a necessidade de servidores EC. No entanto, ela também enfrenta desafios, como atrasos potencialmente significativos no atendimento das demandas geradas pelos dispositivos de IoT [2].

Abordando o conceito de *Smart Cities*, este pode ser explorado em diversos setores, como segurança pública, agricultura familiar, gerenciamento de resíduos, mobilidade urbana, energia e tráfego, sempre com o objetivo de promover maior autonomia e automação para as ações [3]. Contudo, é importante destacar que, além dos aspectos positivos, também existem desafios a serem superados. Assim, conforme mencionado na literatura, questões relacionadas à alocação de recursos utilizando IoT, CC e EC são desafios que requerem atenção [3],[19].

Para que as tecnologias baseadas em IoT funcionem de maneira eficaz, é essencial a interconexão de dispositivos através de uma rede de dados, preferencialmente com altas taxas de transferência e baixa latência [3]. Com o aumento da adoção das tecnologias baseadas em IoT, e o consequente aumento do tráfego de dados, surgiram diversas pesquisas focadas no desenvolvimento de tecnologias de redes avançadas, como o 5G e o 6G [13],[17],[20]. No entanto, embora essas redes tenham evoluído significativamente, elas também trazem novos desafios. Por exemplo, à medida que as redes se tornam mais avançadas, como no caso do 6G, seu alcance tende a diminuir, limitando a cobertura e reduzindo aspectos de ubiquidade, em comparação com redes anteriores, como o 4G, que cobre uma área maior, apesar da menor largura de banda [13],[20],[21]. A Tabela 1 apresenta exemplos de aplicações características das redes sem fio, contextualizadas conforme a evolução das tecnologias do 1G a 6G. Em contraste, a Tabela 2 oferece um resumo da evolução das tecnologias sem fio, abordando conceitos de largura de banda e cobertura, conforme a literatura. Vários fatores, incluindo frequência e interferência, afetam a largura de banda e a cobertura média apresentadas; os valores fornecidos são baseados nos artigos [13],[21],[22],[23],[24].

Para modelar o problema de alocação de recursos no cenário de integração entre dispositivos de IoT, EC e CC, este artigo utiliza Programação Linear

Inteira (*Integer Linear Programming*, ILP). A ILP é uma técnica de otimização matemática empregada para resolver problemas em que a função objetivo e as restrições são lineares, e as variáveis de decisão devem assumir valores inteiros [25]. Mais especificamente, será adotada a ILP Binária (*Binary ILP*, BILP), uma forma de ILP em que as variáveis de decisão são restritas a valores binários [26]. Por fim, este artigo apresenta três modelos matemáticos, com funções objetivo diferentes, baseados em BILP, que podem ser utilizados no futuro para fundamentar algoritmos exatos.

Tabela 1. Evolução das Tecnologias Móveis e Suporte a Aplicações de IoT em Tempo Real

Geração	Aplicações Suportadas
1G	Chamadas de voz analógicas
2G	Mensagens de texto (SMS), chamadas de voz digitais
3G	Navegação na web, e-mail móvel, chamadas de vídeo
4G	Streaming de vídeo em alta definição, redes sociais, jogos online
5G	IoT em tempo quase real, realidade virtual e aumentada, monitoramento remoto de estruturas urbanas
6G (estimado)	IoT em tempo real com latência ultrabaixa, veículos autônomos, comunicações holográficas, Smart Cities, IIoT, indústria 4.0, aplicações avançadas de IA

Tabela 2. Evolução das Tecnologias de Redes Móveis

Geração	Taxa de Transmissão de Dados	Raio de Cobertura
1G	2.4 Kbps	20 km
2G	64 Kpbs	10 km
3G	2 Mpbs	5 km
4G	100 Mpbs - 1 Gbps	3 km
5G	1 - 10 Gpbs	0.6 km
6G (estimado)	100 Gpbs - 1 Tbps	0.32 km

1.2. Objetivos e Contribuições

Este artigo tem como objetivo desenvolver e avaliar modelos para a alocação de recursos em redes que integram tecnologias de IoT com servidores de EC e CC. A pesquisa foca na criação de uma estrutura de rede para oferecer serviços em uma Smart City, como

por exemplo a de Modena, Itália [3]. Adicionalmente, pode-se extrapolar os modelos propostos para outros contextos e topologias de redes, como parques industriais e agrícolas [3],[17].

Diferentemente de outros estudos, como [27], os modelos propostos neste artigo permitem a rejeição de demandas de dispositivos IoT devido a inviabilidades como cobertura de rede insuficiente ou limitações de recursos. Além disso, adicionalmente à literatura revisada, artigos [2],[4],[9], são exploradas restrições reais, tais como capacidade de processamento, cobertura de rede e consumo de armazenamento, memória e largura de banda.

Com base no cenário delineado, este artigo propõe um arcabouço de modelos que, se implementados, poderão dar origem a algoritmos exatos e responder a questões de pesquisa como: i) qual é o impacto das diferentes tecnologias de rede em termos de atendimento às demandas dos dispositivos de IoT? ii) o tempo de resolução do algoritmo exato para o problema de alocação de recursos é afetado pela tecnologia de comunicação utilizada? e iii) ao não se otimizar questões referentes ao custo, tratando o problema apenas como uma dificuldade de viabilidade, qual é o impacto em relação ao tempo de execução e aos custos? Para tal, são propostos:

1. Uma definição formal do problema, com base em restrições reais do mercado de telecomunicações;
2. Três modelos matemáticos em BILP, sendo que:
 - a. o primeiro modelo tem como objetivo assegurar a viabilidade do problema;
 - b. o segundo modelo tem o propósito de minimizar os custos operacionais dos servidores;
 - c. o terceiro modelo busca otimizar o tempo de resposta máximo, definido como makespan.

O restante do artigo está organizado da seguinte forma. A Seção 2 descreve os materiais e métodos adotados na construção deste trabalho. A Seção 3 apresenta uma revisão dos trabalhos existentes na literatura. O problema em questão é formalmente definido na Seção 4, enquanto os modelos em BILP são apresentados na Seção 5. As conclusões e os trabalhos futuros são discutidos na Seção 6.

2. Materiais e Métodos

Similarmente aos trabalhos [3],[9],[27], esta pesquisa adota um design experimental teórico, baseado em modelos matemáticos, para descrever o ambiente a ser implementado em trabalhos futuros.

Para tal, foi realizada uma revisão sistemática e crítica da literatura, além de uma análise conceitual das pesquisas para realizar o desenvolvimento de modelos teóricos. O objetivo é modelar com BILP diferentes configurações de rede para o processamento de demandas originadas de dispositivos IoT. Para embasar este estudo, foi conduzida uma pesquisa bibliográfica abrangente, que incluiu a análise de artigos científicos recentes e relevantes, além de livros nas áreas de redes de computadores, IoT, CC e EC.

No desenvolvimento dos modelos em BILP, primeiro foi feita a definição do problema, juntamente com as variáveis e parâmetros a serem adotados, assim como foram estipuladas as relações entre eles. Esta etapa foi importante para garantir que o modelo refletisse o ambiente de integração de IoT, CC e EC em *Smart Cities*. Em seguida, formalizou-se o problema em termos matemáticos, o que envolveu a definição das funções objetivo e das restrições. Após, foi realizada uma etapa de validação de conceitos e teorias, garantindo que os conceitos utilizados no modelo estivessem alinhados com a literatura e fossem pertinentes aos ambientes reais de *Smart Cities*.

3. Referencial Teórico

O estudo conduzido por [3] aborda a complexidade inerente ao paradigma das infraestruturas de CC e EC. Os autores destacam questões críticas relacionadas à alocação de fluxos de dados provenientes dos sensores para os nós de EC, além da determinação da quantidade e localização ótimas desses nós. Essas decisões são fundamentais, pois influenciam diretamente o desempenho do sistema, especialmente no que tange ao tempo de resposta, além de impactarem os custos operacionais, que podem ser tanto energéticos quanto monetários.

A literatura existente corrobora a relevância desses desafios, enfatizando que a otimização na alocação de serviços e a gestão eficiente dos recursos são problemas centrais em arquiteturas de computação de borda voltadas para a IoT [1],[28]. A necessidade de uma infraestrutura que balanceie adequadamente a latência de rede e o consumo de recursos recebem uma atenção constante em projetos que envolvem sensores distribuídos geograficamente. Neste caso, a eficiência no processamento e transmissão de dados é crucial para o sucesso das aplicações de dispositivos IoT [1],[2],[4]. Nesse contexto, soluções que otimizam a distribuição de nós de borda, bem como a utilização de algoritmos

de otimização e heurísticas avançadas, têm se mostrado fundamentais para atender às exigências de desempenho e custo dos sistemas modernos de computação distribuída.

O artigo [29] oferece uma contribuição significativa ao abordar o problema de alocação de recursos em redes de computação distribuída, por meio de uma abordagem heurística chamada *Fog Node Site Selection* (FNSS). Essa heurística visa otimizar a seleção de locais para servidores, com o objetivo de atender de maneira eficaz às demandas variáveis e à distribuição geográfica dos usuários em uma rede. Ao considerar as dinâmicas de demanda e a localização dos usuários, a heurística FNSS busca maximizar a eficiência operacional e melhorar a QoS oferecido. Os resultados obtidos pela aplicação do FNSS são rigorosamente comparados com uma modelagem matemática baseada em ILP e um algoritmo de aproximação. Nesse contexto, o FNSS se destaca por selecionar um conjunto de servidores que minimiza os custos de comunicação entre as funções de rede, ao mesmo tempo em que realiza o roteamento dos dados de forma a balancear a carga entre os recursos computacionais disponíveis e a largura de banda utilizada. Este estudo destaca a importância de uma seleção criteriosa dos locais para a instalação de servidores, evidenciando o impacto direto dessa decisão na performance da rede.

O trabalho desenvolvido por [30] se insere em um domínio em expansão dentro da computação, que foca na criação de infraestruturas que otimizem a latência, um dos principais critérios para garantir aspectos de QoS em sistemas distribuídos. Nesse contexto, o autor propõe um modelo robusto para o dimensionamento dessas arquiteturas, utilizando a ILP Mista (*Mixed ILP*, MILP). Essa abordagem visa não apenas a otimização da alocação de recursos, mas também o desenvolvimento de um design escalável que atenda às exigências específicas das aplicações sensíveis à latência, assegurando a eficiência operacional em ambientes complexos e dinâmicos.

A computação distribuída, particularmente em contextos que exigem baixa latência e alta eficiência, como é o caso de jogos e aplicações de realidade aumentada, tem encontrado na EC uma solução promissora. De acordo com [31], a proximidade do processamento de dados em relação ao usuário final permite não apenas uma redução significativa no tempo de resposta, mas também uma melhoria substancial na experiência do usuário, ao minimizar a

dependência de uma conectividade contínua e de alta velocidade com a nuvem. Essa eficiência é potencializada quando a EC é integrada a tecnologias emergentes, como o 5G, que, conforme apontado por [32], aprimora ainda mais o desempenho de jogos em dispositivos móveis.

As oportunidades decorrentes dessa integração são amplas e promissoras, com impactos que vão além da área de jogos. O estudo de [31] explora como a infraestrutura de EC pode ser ajustada para acomodar um número crescente de usuários, especialmente em ambientes urbanos com alta densidade populacional, onde a demanda por serviços de baixa latência é crítica. Além disso, essa abordagem pode ser expandida para outras áreas, como a integração da IoT, CC e EC em *Smart Cities*, visando à otimização de processos e serviços urbanos [3],[29]. Este aspecto enfatiza a importância de aplicações de dispositivos IoT como um facilitador essencial para o desenvolvimento sustentável e eficiente de infraestruturas urbanas, potencializando o uso de recursos e melhorando a qualidade de vida nas cidades [13].

4. Definição Formal do Problema

As aplicações oriundas de dispositivos IoT geram demandas que precisam ser atendidas por servidores EC ou CC. Cada uma dessas infraestruturas enfrenta desafios específicos. Portanto, a integração entre a CC e a EC tem se tornado um tópico de crescente interesse na literatura [9], [13]. A Figura 1 ilustra essa integração, mostrando as camadas de dispositivos e suas respectivas relações. Essa visualização facilita a compreensão da interação entre os diferentes componentes e a alocação eficiente dos recursos.

A proposta neste artigo é modelar cenários em que os dispositivos podem ou não ser processados nas camadas de EC e CC, e, caso sejam processados, em quais locais foram designados para essa alocação. O objetivo é gerar um arcabouço teórico que permita calcular o custo associado a essas situações, o tempo de processamento e o consumo de recursos da rede. Tais aspectos ajudarão a compreender melhor os impactos econômicos e operacionais associados à falta de capacidade de processamento ou aos gastos com recursos redundantes.

Para abordar esse problema, é essencial desenvolver uma definição formal que considere as restrições e características do ambiente de telecomunicações. Neste contexto, foi considerada uma

Smart City com diversas aplicações oriundas de dispositivos IoT. Esses dispositivos precisam se conectar a servidores localizados tanto na camada de EC quanto na camada de CC. Caso o servidor da camada de EC não possua recursos residuais, e não consiga atender às demandas, elas podem ser repassadas para a camada de CC. Nesse caso, assume-se que os servidores de EC estão conectados com a camada de servidores de CC.

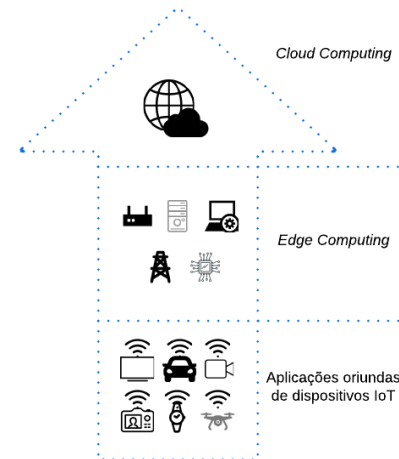


Figura 1. Estrutura hierárquica da integração entre IoT, EC e CC

Para tal, é fundamental verificar se o dispositivo IoT está dentro do raio de cobertura do servidor, seja de EC ou CC, para garantir a conexão e, se possível, o processamento. Caso o dispositivo esteja dentro da cobertura do servidor da rede, a próxima etapa consiste em avaliar se o servidor possui as capacidades necessárias de processamento para atender às demandas de armazenamento, largura de banda e memória exigidas pelas aplicações. Se tal capacidade não estiver disponível, dependendo da cobertura, outros servidores da camada de EC serão avaliados, ou a demanda poderá ser repassada para a camada de CC, onde, a depender dos recursos residuais, poderá ser processada. Neste cenário, diferentes fatores devem ser considerados. Entre eles, estão o custo de operação dos servidores de EC e CC, o tempo gasto para a tomada de decisão sobre o processamento ou não de uma demanda, bem como o local de atendimento.

5. Modelos Matemáticos em BILP

Na sequência desta seção, são descritos os conjuntos (Subseção 5.1), parâmetros (Subseção 5.2),

variáveis (Subseção 5.3) e restrições (Subseção 5.4) utilizadas nos modelos formulados em BILP. Posteriormente, são introduzidos os modelos propostos (Seções 5.5 a 5.7), juntamente com suas respectivas funções objetivo.

5.1. Conjuntos

- D : conjunto de todas as aplicações oriundas de dispositivos de IoT, que devem ser processadas em algum servidor da rede, seja na camada de EC ou de CC;
- V : conjunto de todos os servidores de EC e CC existentes;
- $A^d \subseteq V$: conjunto de servidores de EC e CC $i \in V$ que cobrem o dispositivo de IoT $d \in D$.

Caso uma demanda $d \in D$ não possa ser atendida na camada de EC devido à falta de recursos residuais, ela pode ser encaminhada, através de conexões cabeadas entre dispositivos de EC e a camada de CC, para possível processamento. Assim, assume-se que os servidores da camada de CC também pertencem ao conjunto A^d .

5.2. Parâmetros

- $bw^d \in R_+$: demanda de largura de banda da aplicação $d \in D$ (megabits por segundo, Mbps);
- $c^d \in R_+$: custo de não atendimento da aplicação $d \in D$ (euro, €);
- $m^d \in R_+$: demanda de memória da aplicação $d \in D$ (gigabytes, GB);
- $pcc^d \in R_+$: demanda de ciclos de clock para processamento, em stock¹, da aplicação $d \in D$ (gigahertz, GHz);
- $pcn^d \in R_+$: demanda do número de núcleos para processamento da aplicação $d \in D$ (core);
- $st^d \in R_+$: demanda da capacidade de armazenamento da aplicação $d \in D$ (GB);
- $s^d \in R_+$: demanda da quantidade de dados a serem enviados da aplicação $d \in D$ (Mbps);
- $BW_i \in R_+$: largura de banda oferecida pelo servidor $i \in V$ (Mbps);
- $C_i \in R_+$: custo de operação do servidor $i \in V$ (€);
- $M_i \in R_+$: memória oferecida pelo servidor $i \in V$

(GB);

- $PCC_i \in R_+$: oferta de ciclos de clock do processador, em stock¹, do servidor $i \in V$ (GHz);
- $PCN_i \in R_+$: quantidade de núcleos oferecida pelo processador do servidor $i \in V$ (core);
- $ST_i \in R_+$: armazenamento oferecido pelo servidor $i \in V$ (GB);
- $C_M \in R_+$: custo de operação máximo definido para o modelo M (€);
- $t^p_i \in R_+$: tempo de atraso do processamento do servidor $i \in V$ (milissegundos, ms);
- $t^d_i \in R_+$: tempo de atraso da comunicação da aplicação $d \in D$ para o servidor $i \in V$ (ms). Conforme [33], considera-se o atraso de transmissão, definido em relação a s^d e o atraso de propagação no meio;
- $\alpha \in R_+$: parâmetro experimental, empregado para controlar um aumento ou uma redução de custo;
- $\mathcal{M} \in R_+$: valor alto adotado como penalidade para o não atendimento da aplicação $d \in D$.

5.3. Variáveis

- TR_{max} : *makespan*, definido como o tempo total necessário para processar todo o conjunto de demandas existentes;
- $w^d \in \{0,1\}$: se 1, a aplicação $d \in D$ teve sua requisição atendida;
- $x^d_i \in \{0,1\}$: se 1, a aplicação $d \in D$ envia uma requisição para serem processados no servidor $i \in V$;
- $z_i \in \{0,1\}$: se 1, o servidor $i \in V$ está ativo atendendo a demanda de alguma aplicação $d \in D$.

5.4. Restrições

Os conjuntos de restrições (1) a (6) asseguram que as demandas das aplicações $d \in D$, em termos de QoS, sejam atendidas, sempre que possível. Contudo, caso alguma demanda $d \in D$ não seja viável de atendimento garantindo aspectos de QoS, seu mapeamento será rejeitado. Neste caso, a variável w^d assumirá o valor 0.

As Restrições (1) garantem que, sempre que possível, cada demanda $d \in D$ seja atendida em um servidor $i \in V$, forçando a variável w^d a assumir o valor 1.

$$\sum_{i \in V | i \in A^d} x^d_i = w^d, \forall d \in D \quad (1)$$

¹ *stock*, termo utilizado para processadores com as configurações padrões de fábrica.

O conjunto de Restrições (2) tem como objetivo garantir que a demanda de largura de banda bw^d da aplicação $d \in D$ seja sempre inferior ou, no máximo, igual à largura de banda disponível BW_i no servidor $i \in V$. Isso garante a viabilidade na gestão dos recursos de rede, assegurando que o tráfego de dados das aplicações não sobrecarregue as conexões.

$$\sum_{d \in D | i \in A^d} bw^d x_i^d \leq BW_i z_i, \forall i \in V \quad (2)$$

As Restrições (3) são responsáveis por garantir que a demanda de memória m^d da aplicação $d \in D$ está dentro dos limites da memória disponível M_i no servidor $i \in V$. Essa verificação garante a eficiência no uso da memória, prevenindo falhas ou degradação de desempenho que poderiam ocorrer caso a demanda exceda a quantidade disponível.

$$\sum_{d \in D | i \in A^d} m^d x_i^d \leq M_i z_i, \forall i \in V \quad (3)$$

O conjunto de Restrições (4) assegura que a demanda por núcleos de processamento pcn^d da aplicação $d \in D$ seja inferior ou equivalente ao número de núcleos PCN_i disponíveis no servidor $i \in V$. Essa garantia é fundamental para evitar a sobrecarga do processador, o que poderia resultar em um desempenho reduzido ou até mesmo em falhas no processamento.

$$\sum_{d \in D | i \in A^d} pcn^d x_i^d \leq PCN_i z_i, \forall i \in V \quad (4)$$

As Restrições (5) certificam que a demanda por ciclos de clock para processamento pcc^d da aplicação $d \in D$ seja menor ou igual à capacidade de ciclos de clock PCC_i disponível no servidor $i \in V$. Esse controle assegura que o servidor possa lidar com a carga de trabalho sem comprometer a velocidade de processamento.

$$\sum_{d \in D | i \in A^d} pcc^d x_i^d \leq PCC_i z_i, \forall i \in V \quad (5)$$

As Restrições (6) são implementadas para garantir que a demanda de armazenamento st^d da aplicação $d \in D$ não ultrapasse a capacidade de armazenamento ST_i do servidor $i \in V$.

$$\sum_{d \in D | i \in A^d} st^d x_i^d \leq ST_i z_i, \forall i \in V \quad (6)$$

O conjunto de Restrições (7) garante que o tempo total de atraso na comunicação da aplicação $d \in D$ para o servidor $i \in V$, somado ao tempo de processamento no servidor $i \in V$ para cada demanda, não exceda o tempo de resposta máximo TR_{max} .

$$(t_i^d + t_p^d) x_i^d \leq TR_{max}, \forall d \in D, \forall i \in A^d \quad (7)$$

As Restrições (8) asseguram que o custo total C_i associado à ativação dos servidores z_i deve ser inferior ou igual ao custo total máximo permitido C_M . Dessa forma, busca-se manter os custos dentro dos limites estabelecidos. Neste caso, o parâmetro α é empregado para, experimentalmente, permitir um certo grau de aumento ou redução desse custo. A intenção da adoção de tal parâmetro é possibilitar a análise do impacto de um aumento percentual no custo em relação às métricas de QoS e QoE, explorando mais regiões do espaço de soluções.

$$\sum_{d \in D | i \in A^d} C_i z_i \leq C_M \alpha \quad (8)$$

5.5. Modelo 1: viabilidade do problema (M1)

O modelo proposto utiliza uma função objetivo (Equação (9)) que busca garantir a viabilidade do problema. Neste caso, a variável w^d é forçada a assumir o valor 1, com uma penalidade $\mathcal{M}$ por demanda não atendida.

$$\min \sum_{d \in D} (1 - w^d) \mathcal{M} \quad (9)$$

sujeito a:

Restrições (1) a (6)

5.6. Modelo 2: custo da operação (M2)

O modelo proposto utiliza uma função objetivo (Equação (10)) que busca minimizar o custo de ativação dos servidores, levando em consideração as penalidades associadas às demandas não atendidas. Neste caso, a variável w^d é forçada a assumir o valor 1 para que o custo de não atendimento definido por c^d não seja contabilizado, garantindo o cumprimento dos aspectos do Acordo de Nível de Serviço (*Service Level Agreement*, SLA) entre o provedor e os usuários.

$$\min (\sum_{i \in V} C_i z_i + \sum_{d \in D} (1 - w^d) c^d) \quad (10)$$

sujeito a:

Restrições (1) a (6)

5.7. Modelo 3: *makespan* (M3)

Este modelo busca minimizar o tempo de resposta máximo para o atendimento de uma demanda, definido como TR_{max} (*makespan*). Esse objetivo, mostrado na Equação (11), é alcançado por meio da minimização do tempo necessário para processar as demandas nos servidores disponíveis. Na função objetivo, C_M representa o custo de operação definido pelo modelo M2. Adicionalmente, C_M pode ser obtido por meio de algum algoritmo aproximado ou heurística.

$$\min(TR_{max} + \sum_{d \in D} (1 - w^d)c^d) \quad (11)$$

sujeito a:

Restrições (1) a (8)

6. Considerações finais

A alocação eficiente de recursos em redes é um elemento crucial para o desenvolvimento de soluções que atendam às crescentes demandas por conectividade e processamento de dados em tempo real em *Smart Cities*. No entanto, a complexidade e a interdependência entre EC, CC e IoT requerem uma análise aprofundada e uma modelagem rigorosa, considerando as particularidades de cada ambiente. Neste contexto, este estudo apresentou três modelos matemáticos baseados em BILP, que abordam restrições reais, como capacidade de processamento, cobertura de rede e consumo de armazenamento, largura de banda e memória. Ressalta-se que a cobertura de rede também é um aspecto importante a ser avaliado, impactando diretamente a ubiquidade e o desenvolvimento das *Smart Cities*.

Cada modelo desempenha um papel específico na eficiência da rede, oferecendo uma estrutura para lidar com a complexidade da alocação de recursos. Eles otimizam os recursos disponíveis e garantem um atendimento econômico e eficiente das demandas. Existem ainda desafios relacionados à arquitetura, gerenciamento de infraestrutura, comunicação e alocação de recursos. Neste contexto, este estudo contribui significativamente para o meio acadêmico ao fornecer suporte para a adoção de modelos matemáticos que refletem aspectos reais do ambiente de telecomunicações. Além disso, as modelagens servirão de base para o desenvolvimento de algoritmos exatos e heurísticas em futuras pesquisas.

Os modelos apresentados abrangem a definição de conjuntos, parâmetros, variáveis e restrições. Especificamente, o Modelo M1 (Subseção 5.5) tem como objetivo assegurar a viabilidade do problema, garantindo que todas as demandas sejam atendidas. A hipótese a ser investigada em trabalhos futuros é que o tempo de execução de um algoritmo exato para este modelo é menor em relação aos outros modelos. Em contrapartida, o Modelo M2 (Subseção 5.6) tem o propósito de minimizar os custos operacionais dos servidores, levando em consideração as penalidades associadas às demandas não atendidas. Neste caso, uma hipótese a ser investigada em trabalhos futuros é que existe uma redução de custo a ser considerada entre M1 e M2, aspecto que possui relevância para provedores e demais empresas envolvidas. Por fim, o Modelo M3 (Subseção 5.7) busca minimizar o tempo de resposta máximo (*makespan*), reduzindo o tempo necessário para o processamento de todas as demandas. A hipótese a ser investigada no futuro é que este modelo pode gerar benefícios para o usuário final, onde um atraso de comunicação fim a fim menor impacta diretamente as métricas de QoE, embora possa aumentar o custo de prover o serviço.

Os modelos apresentados abrangem a definição de conjuntos, parâmetros, variáveis e restrições. Especificamente, o Modelo M1 tem como objetivo assegurar a viabilidade do problema, garantindo que todas as demandas sejam atendidas. A hipótese a ser investigada em trabalhos futuros é que o tempo de execução de um algoritmo exato para este modelo é menor em relação aos outros modelos. Em contrapartida, o Modelo M2 tem o propósito de minimizar os custos operacionais dos servidores, levando em consideração as penalidades associadas às demandas não atendidas. Neste caso, uma hipótese a ser investigada em trabalhos futuros é que existe uma redução de custo a ser considerada entre M1 e M2, aspecto que possui relevância para provedores e demais empresas envolvidas. Por fim, o Modelo M3 busca minimizar o tempo de resposta máximo (*makespan*), reduzindo o tempo necessário para o processamento de todas as demandas. A hipótese a ser investigada no futuro é que este modelo pode gerar benefícios para o usuário final, onde um atraso de comunicação fim a fim menor impacta diretamente as métricas de QoE, embora possa aumentar o custo de prover o serviço.

7. Agradecimentos

Este estudo contou com o apoio financeiro da Universidade Federal de São João del-Rei, por meio da bolsa do Programa Institucional de Bolsas de Iniciação Científica (PIBIC). Agradece-se também ao Departamento de Tecnologia em Engenharia Civil, Computação, Automação, Telemática e Humanidades (DTECH) pela infraestrutura e suporte técnico fornecidos ao longo do desenvolvimento da pesquisa.

Referências

- [1] W. Z. Khan, E. Ahmed, S. Hakak, I. Yaqoob, and A. Ahmed, "Edge computing: A survey," *Future Generation Computer Systems*, vol. 97, pp. 219–235, 2019.
- [2] P. Silva, A. Costan, and G. Antoniu, "Investigating edge vs. cloud computing trade-offs for stream processing," in *2019 IEEE International Conference on Big Data (Big Data)*, pp. 469–474, 2019.
- [3] T. A. d. Queiroz, C. Canali, M. Iori, and R. Lancellotti, *An Optimization View to the Design of Edge Computing Infrastructures for IoT Applications*, pp. 1–30. Cham: Springer International Publishing, 2022.
- [4] S. Araújo, M. Moreira, and G. Mateus, "Alocação de recursos em edge e cloud computing para atender dispositivos de iot: Uma análise rumo ao 6g," in *Anais do XLII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*, (Porto Alegre, RS, Brasil), pp. 225–238, SBC, 2024.
- [5] Oracle, "What is iot?," <https://tinyurl.com/248k9t2e>, 1 2024. Accessed on February 08, 2024.
- [6] IoT Analytics, "Iot 2023 in review," 2023. Acesso em: 15 ago. 2024.
- [7] Y. Li and W. Tu, "Traffic modelling for iot networks: A survey," in *Proceedings of the 2020 10th International Conference on Information Communication and Management, ICICM 2020*, (New York, NY, USA), p. 4–9, Association for Computing Machinery, 2020.
- [8] I. Ullah, D. Adhikari, X. Su, F. Palmieri, C. Wu, and C. Choi, "Integration of data science with the intelligent IoT (IIoT): current challenges and future perspectives," *Digital Communications and Networks*, 2024.
- [9] J. Santos, T. Wauters, B. Volckaert, and F. De Turck, "Towards end-to-end resource provisioning in fog computing over low power wide area networks," *Journal of Network and Computer Applications*, vol. 175, p. 102915, 2021.
- [10] T. Araujo, I. Ribeiro, and V. Mota, "Covid19-fast-check: Uma abordagem internet das coisas aplicada à saúde para pré-triagem hospitalar do coronavírus," in *Anais do XLI Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*, (Porto Alegre, RS, Brasil), pp. 518–531, SBC, 2023.
- [11] L. S. Vailshery, "Number of IoT connected devices worldwide 2019-2023, with forecasts to 2030," 2023. [Online]; acessado 20/12/2023, disponível em <https://www.statista.com/statistics/1183457/iot-connected-devices-worldwide>.
- [12] V. Salunke and S. Mahajan, "The evolution and implications of the internet of things," 03 2024.
- [13] M. Alsabah, M. A. Naser, B. M. Mahmmod, S. H. Abdulhussain, M. R. Eissa, A. Al-Baidhani, N. K. Noordin, S. M. Sait, K. A. Al-Utaibi, and F. Hashim, "6G Wireless Communications Networks: A Comprehensive Survey," *IEEE Access*, vol. 9, 2021.
- [14] A. Bris, C. Cabolis, B. Lanvin, O. Madureira, J. Caballero, M. Pistis, M. Zargari, and A. Tozer, "IMD Smart City Index 2023: Index Report," 2023. [Online]; acessado 10/06/2024, disponível em <https://imd.cld.bz/IMD-Smart-City-Index-Report-20231/12/>.
- [15] L. U. Khan, I. Yaqoob, N. H. Tran, S. M. A. Kazmi, T. N. Dang, and C. S. Hong, "Edge-Computing-Enabled Smart Cities: A Comprehensive Survey," *IEEE Internet of Things Journal*, vol. 7, no. 10, pp. 10200–10232, 2020.
- [16] S. N. Srirama, "A decade of research in fog computing: Relevance, challenges, and future directions," *Software: Practice and Experience*, vol. 54, no. 1, pp. 3–23, 2024.
- [17] V. K. Quy, D. C. Nguyen, D. Van Anh, and N. M. Quy, "Federated learning for green and sustainable 6g iiot applications," *Internet of Things*, vol. 25, p. 101061, 2024.
- [18] L. Wang, M. Dolati, and M. Ghaderi, "Change: Delay-aware service function chain orchestration at the edge," in *2021 IEEE 5th International Conference on Fog and Edge Computing (ICFEC)*, pp. 19–28, 2021.
- [19] U. Brezolin, N. P. Jr., A. Vergütz, and M. Nogueira, "Um método para detecção de vulnerabilidades através da análise do tráfego de rede iot," in *Anais do XL Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*, (Porto Alegre, RS, Brasil), pp. 447–460, SBC, 2022.
- [20] C. D. Alwis, A. Kalla, Q.-V. Pham, P. Kumar, K. Dev, W.-J. Hwang, and M. Liyanage, "Survey on 6G Frontiers: Trends, Applications, Requirements, Technologies and Future Research," *IEEE Open Journal of the Communications Society*, vol. 2, pp. 836–886, 2021.
- [21] A. F. M. S. Shah, A. N. Qasim, M. A. Karabulut, H. Ilhan, and M. B. Islam, "Survey and performance evaluation of multiple access schemes for next-generation wireless communication systems," *IEEE Access*, vol. 9, pp. 113428–113442, 2021.
- [22] C. D. Alwis, A. Kalla, Q.-V. Pham, P. Kumar, K. Dev, W.-J. Hwang, and M. Liyanage, "Survey on 6G Frontiers: Trends, Applications, Requirements, Technologies and Future Research," *IEEE Open Journal of the Communications Society*, vol. 2, pp. 836–886, 2021.
- [23] P. Rajkotia, "How 6G Can Transform The World and Technology," article 8, *IEEE Standards Association*, 8 2022. [online] disponível em standards.ieee.org/beyond-standards/how-6g-can

transform-the-world-and-technology.

[24] S. Periannasamy, C. Thangavel, S. Latha, G. Reddy, S. Ramani, P. V. Phad, S. Chandline, and S. Gopalakrishnan, "Analysis of Artificial Intelligence Enabled Intelligent Sixth Generation (6G) Wireless Communication Networks," in 2022 IEEE International Conference on Data Science and Information System (ICDSIS), 2022.

[25] R. Alves and C. Delgado, Programação Linear Inteira. PhD thesis, Universidade do Porto, Porto, Portugal, 1997. Acesso em: 19 ago. 2024.

[26] G. L. Nemhauser and L. A. Wolsey, Integer and Combinatorial Optimization. New York: Wiley, 1988.

[27] M. Moreira, S. Araújo, and G. Mateus, "Definição e modelos para o problema de localização de nós fog e designação de demandas de sensores iot," in Anais do XLII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos, (Porto Alegre, RS, Brasil), pp. 141–154, SBC, 2024.

[28] Z. Chang, S. Liu, X. Xiong, Z. Cai, and G. Tu, "A survey of recent advances in edge-computing-powered artificial intelligence of things," IEEE Internet of Things Journal, vol. 8, no. 18, pp. 13849–13875, 2021.

[29] S. Singh and D. Vidyarthi, "Fnss: A heuristics for fog node site selection," SN Computer Science, vol. 5, 01 2024.

[30] I. Martinez, A. Jarray, and A. S. Hafid, "Scalable design and dimensioning of fog-computing infrastructure to support latency-sensitive iot applications," IEEE Internet of Things Journal, vol. 7, no. 6, pp. 5504–5520, 2020.

[31] G. Robaina and A. Fiorese, "Gaming on the edge: Uma arquitetura de computação na borda para jogos em dispositivos móveis," in Anais do XLI Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos, (Porto Alegre, RS, Brasil), pp. 574–587, SBC, 2023.

[32] N. Panwar, S. Sharma, and A. K. Singh, "A survey on 5g: The next generation of mobile communication," Physical Communication, vol. 18, pp. 64–84, 2016. Special Issue on Radio Access Network Architectures and Resource Management for 5G.

[33] A. S. Tanenbaum and D. Wetherall, Computer networks, 5th Edition. Pearson, 2011.