

Alocação de Recursos em *Edge* e *Cloud Computing* com comunicação D2D em *Smart Cities*: modelos e análises

Alex Vidigal Bastos^{a,b,1}, Samuel Moreira Abreu Araújo^{b,2}, Diógenes Cecílio da Silva Jr.^{c,3}

^aPrograma de Pós-Graduação em Engenharia Elétrica

Universidade Federal de Minas Gerais - UFMG

Av. Antônio Carlos 6627, 31270-901, Belo Horizonte, MG, Brasil

^bDepartamento de Tecnologia em Engenharia Civil, Computação, Automação, Telemática e

Humanidades - Universidade Federal de São João del-Rei - UFSJ

Rodovia MG 443, km 7, 36490-972, Ouro Branco, MG, Brasil

^cDepartamento de Engenharia Elétrica

Universidade Federal de Minas Gerais - UFMG

Av. Antônio Carlos 6627, 31270-901, Belo Horizonte, MG, Brasil

¹alexvbh@ufsj.edu.br, ²sabreu@ufsj.edu.br, ³diogenes@ufmg.br

RESUMO

Smart Cities utilizam tecnologias digitais para gerenciar recursos, visando a benefícios econômicos e sociais. Este artigo aborda o problema de alocação de recursos de *Edge* e *Cloud Computing* para atender as demandas de dispositivos de IoT de uma *Smart City*. O modelo de redes tradicional resulta em problemas de atendimento devido à baixa cobertura das tecnologias de quinta e sexta geração (5G e 6G). Neste artigo, é apresentada uma definição formal do problema e um algoritmo exato, baseado em Programação Linear Inteira (PLI), que utiliza a comunicação *Device-to-Device* (D2D) para ampliar a cobertura da rede. Os algoritmos propostos são comparados com outros da literatura que não utilizam comunicação D2D. Experimentos computacionais realizados com instâncias da literatura indicam que, apesar de um aumento do atraso fim a fim, o uso da comunicação D2D pode melhorar a taxa de aceitação em até 486%, se considerada a tecnologia 6G.

PALAVRAS CHAVE. Comunicação D2D, Alocação de Recursos, Computação de Borda.

PO em Telecomunicações, Programação Matemática, Otimização Combinatória.

ABSTRACT

Smart Cities employ digital technologies to manage resources, aiming for economic and social benefits. This article addresses the resource allocation problem in *Edge* and *Cloud Computing* to meet the demands of IoT devices in a *Smart City*. The current network model results in service issues due to the low coverage of fifth and sixth-generation technologies (5G and 6G). In this article, a formal definition of the problem is presented along with an exact algorithm based on Integer Linear Programming (ILP), which utilizes *Device-to-Device* (D2D) communication to expand network coverage. The proposed algorithms are compared with others in the literature that do not employ D2D communication. Experiments with instances from the literature indicate that, despite the increase in end-to-end delay, the use of D2D communication can improve the acceptance rate by up to 486%, considering 6G technology.

KEYWORDS. D2D communication, Resource Allocation, Edge Computing.

OR in Telecommunications, Mathematical Programming, Combinatorial Optimization.

1. Introdução

A conectividade digital está se tornando cada vez mais integrada à nossas vidas. O uso de dispositivos móveis e sensores conectados à internet, também chamados de Equipamentos de Usuários (*User Equipment*, UE), como dispositivos de *streaming* de mídia e *smartphones*, tornou-se comum e ubíquo. Em um contexto atual, tais UEs, não se limitam a comunicação entre pessoas. Eles também desempenham um papel vital em uma variedade de serviços, desde navegar na internet e realizar transações como operações bancárias, até conceitos mais atuais envolvendo a Internet das Coisas (*Internet of Things*, IoT), direção colaborativa e Indústria 5.0 [Alwis et al., 2021].

Corroborando as afirmações anteriores, a era das tecnologias inteligentes é uma realidade e representa um avanço tecnológico significativo e atual [Murrioni et al., 2023; Araújo et al., 2024]. Neste panorama, a IoT refere-se à interconexão de objetos à internet, ampliando suas possibilidades de uso. Com o desenvolvimento da IoT, as cidades e corporações passaram a empregar tal tecnologia, afluindo assim o conceito de Cidades Inteligentes (*Smart Cities*) [Santos et al., 2021].

Conceitualmente, a ideia de uma *Smart City* envolve um município no qual as redes de comunicação e os serviços se tornam mais eficientes por meio de soluções digitais, frequentemente com tomadas de decisão em tempo real. Esse conceito implica em fornecer automação em áreas como indústria, saúde e segurança [Queiroz et al., 2022]. Consequentemente, tal conceito acarreta benefícios técnicos, econômicos e sociais para residentes e empresas [Khan et al., 2020].

Neste contexto, torna-se necessário que as redes de comunicações apresentem características de disponibilidade e confiabilidade para atender às demandas geradas pelos UEs. Como dificultador, observa-se uma forte tendência por aplicações de IoT que requerem um baixo atraso de comunicação. Por exemplo, aplicações de *Smart Grid* 2.0 e Indústria 5.0 toleram um atraso máximo de comunicação da ordem de milissegundos [Alwis et al., 2021]. Como um desafio adicional está o crescimento da adoção da IoT no mundo. Prevê-se que até 2025 mais de 27 bilhões de dispositivos estarão conectados [TGT Consult, 2022]. Além disso, espera-se que o tráfego global de dados móveis aumente 55% ao ano, alcançando 5016 exabytes em 2030 [Banafaa et al., 2023].

Percebida a crescente demanda por Serviços de Rede (*Network Service*, NS) e o aumento do número de UEs conectados, os atuais sistemas de comunicação sem fio estão sendo impulsionados a expandir-se para atender às necessidades dos usuários em relação às métricas de Qualidade de Serviço (*Quality of Service*, QoS) como largura de banda e atrasos [Alwis et al., 2021]. Para isso, as tecnologias móveis 5G e 6G propõem mudanças radicais em relação às de quarta geração (4G), principalmente em relação latência, possibilitando uma ampla conectividade, com atendimento aos requisitos de conexão onipresente, operando em frequências altas e bandas amplas (Tabela 1).

Com suas características de baixo atraso de comunicação, as redes baseadas em tecnologias 5G e 6G, em contraste com a 4G, mostram-se ideais para suportar NS que exigem conexões de alta velocidade (Tabela 1). Por esses fatores, este trabalho concentra-se nessas tecnologias. Adicionalmente, para alcançar os objetivos relacionados a custos, atrasos e ubiquidade, diversas tecnologias estão sendo desenvolvidas ou aprimoradas, incluindo Computação de Borda (*Edge Computing*, EC), Computação em Nuvem (*Cloud Computing*, CC) e comunicação D2D.

Tabela 1: Comparação dos sistemas de computação móvel 4G, 5G e 6G [Banafaa et al., 2023].

Característica	4G (2010)	5G (2020)	6G (2030)
Largura de banda teórica	1Gbps	10Gbps	1Tbps
Latência ¹	100ms	1ms	0.1ms
Densidade Conexão	$2 \cdot 10^3$ UEs/km ²	10^6 UEs*/km ²	10^7 UEs*/km ²
Raio de cobertura ²	até 3 quilômetros	até 600 metros	até 320 metros
Frequência máxima	6 GHz	90 GHz	10 THz

¹ Considera-se o tempo médio que leva para os dados viajarem de um ponto a outro em uma rede.

² A cobertura é influenciada por diversos fatores. Os valores fornecidos são aproximações baseadas em Araújo et al. [2024].

1.1. Background

A comunicação D2D pode ser aplicada em diversas áreas, como segurança pública, indústrias e jogos. Para atender aos rigorosos requisitos de NS atuais, essa tecnologia vem sendo amplamente utilizada para o aprimoramento das tecnologias 5G e 6G. A comunicação D2D oferece diversas vantagens para as redes de comunicação, tais como o auxílio ao tráfego de dados móveis e a ampliação da cobertura da rede [Gismalla et al., 2022], aspectos estes investigados neste artigo.

Por sua vez, a CC é um modelo de computação que permite o acesso a recursos computacionais contratados sob demanda. Tal tecnologia centraliza o processamento de dados em servidores robustos e fornece uma ampla gama de serviços virtualizados [Silva et al., 2019]. No entanto, em sua concepção, tal modelo não considerou a grande quantidade de dados que vários UEs estão gerando atualmente. Esse grande volume de dados gera problemas de largura de banda, tornando um desafio oferecer suporte a novos NS de baixa e ultrabaixa latência. Assim, dado que todos esses dispositivos estarão conectados à rede, enviando e recebendo dados para a CC, os modelos de soluções de CC centralizadas se tornarão cada vez mais impraticáveis [Santos et al., 2021].

Diante desses desafios, o modelo de EC, em contrapartida à CC, busca processar os dados em servidores próximos dos UEs, mitigando aspectos como altos volumes de dados trafegando na rede e reduzindo os atrasos de comunicação [Xing et al., 2019]. A EC refere-se à inserção de uma camada de servidores intermediária entre os UEs e a CC [Khan et al., 2020]. Tal modelo torna-se ideal para suporte a UEs que requerem aspectos de baixa ou ultrabaixa latência.

Se combinados os modelos de CC e EC, estes resultam em transferências de dados quase instantâneas e aplicações em tempo real. Nesse cenário, a EC é ideal para processar dados em tempo real, enquanto a CC é mais adequada para lidar com requisitos menos sensíveis a atrasos. Tal integração também pode se beneficiar com aspectos das tecnologias de comunicação móvel (Tabela 1), fornecendo os requisitos necessários para os NS demandados em *Smart Cities*.

Contudo, conforme mostrado na Tabela 1, o raio de cobertura do sinal diminui à medida que as tecnologias avançam. Isso implica na necessidade de um maior número de estações base para cobrir uma área com a tecnologia 5G em comparação com o 4G. Nessas circunstâncias, o uso da comunicação D2D surge como uma alternativa viável para estender a cobertura da rede e permitir que os UEs sejam atendidos, mesmo em áreas onde não há cobertura de rede tradicional. A comunicação D2D possibilita que UEs se conectem sem a necessidade de uma estação base, melhorando assim o desempenho dos sistemas tradicionais de redes [Gismalla et al., 2022]. No entanto, a coexistência da tecnologia D2D na rede tradicional pode acarretar diversos problemas e desafios, incluindo a alocação de recursos e a definição de roteamentos entre os UEs envolvidos.

Considerando o arcabouço tecnológico apresentado, Araújo et al. [2024] abordam o problema de alocação de recursos em redes baseadas em tecnologias móveis, integrando EC e CC em uma *Smart City*. No entanto, no referido artigo, os autores concluem que o raio de cobertura dos UEs que utilizam as tecnologias 5G e 6G é inferior ao do 4G, gerando uma alta taxa de rejeição no atendimento das demandas de NS dos UEs. Diante dessa observação, este artigo investiga o problema proposto em Araújo et al. [2024], mas, complementarmente à literatura, propõe e examina especificamente a expansão da cobertura da rede com base em tecnologias D2D, 5G e 6G.

Em um levantamento da literatura sobre o tema, e fazendo uma análise dos pontos apresentados anteriormente, surgem algumas questões de pesquisa a serem investigadas, dentre elas: a) O uso da tecnologia D2D pode reduzir a taxa de rejeição no atendimento das demandas originadas de UEs? b) A adoção da tecnologia D2D aumenta o atraso fim a fim gerado no atendimento das demandas originadas de UEs? e c) A adoção da tecnologia D2D para aumentar a cobertura, aumenta o tempo de processamento do algoritmo de alocação? Para investigar as questões de pesquisa levantadas, apresentamos as seguintes contribuições: i) Uma definição formal do problema,

abrangendo alocação de recursos e roteamento em UEs habilitados com D2D; *ii*) Um modelo matemático utilizando PLI, com restrições de alocação de recursos, capacidade e fluxo (roteamento); *iii*) Um algoritmo exato, baseado no modelo em PLI; e *iv*) Uma análise de experimentos, empregando o algoritmo proposto e outro da literatura, em instâncias reais, considerando variações entre as tecnologias de redes e diferentes NS de uma *Smart City*. A escolha de utilizar um algoritmo exato é justificada pela sua capacidade de fornecer soluções com garantias de otimalidade. Tal escolha proporciona análises assertivas, eliminando possíveis erros de decisão e vieses que poderiam surgir ao utilizar heurísticas.

O restante do artigo está organizado como segue. Na Seção 2 é apresentado uma revisão dos trabalhos da literatura. O problema é definido na Seção 3, e o modelo em PLI apresentado na Seção 4. Os cenários e experimentos computacionais são demonstrados na Seção 5. Por fim, as conclusões e trabalhos futuros são discutidos na Seção 6.

2. Trabalhos Relacionados

A comunicação D2D está atraindo atenção crescente tanto da indústria quanto da academia. Na literatura podemos encontrar vários estudos revisando diferentes aspectos da comunicação D2D. Aqui, apresentamos uma seleção de assuntos relacionados ao trabalho, abordando *Smart Cities*, alocação de recursos nas comunicações D2D e PLI. Os trabalhos de Queiroz et al. [2022] e Araújo et al. [2024] foram utilizados como referência para este artigo, ambos abordando a alocação de recursos em redes. Ambos propõem um modelo em PLI e um algoritmo exato, explorando conceitos de IoT, *Smart Cities* e EC. Queiroz et al. sugerem que servidores de EC não utilizados podem ser desligados para minimizar os custos operacionais, princípios adotados como função objetivo neste artigo. No entanto, Queiroz et al. não consideram o roteamento das conexões, nem aplicam conceitos de consumo de memória ou avaliam diferentes tecnologias de comunicação móvel, aspectos estes tratados neste estudo.

Araújo et al. [2024] analisam as tecnologias de rede atuais e demonstram que, para as tecnologias 5G e 6G, apesar da significativa redução no atraso fim a fim em relação à 4G, há uma alta taxa de rejeição devido à falta de cobertura de vários UEs. Este ponto é abordado neste artigo, onde as tecnologias D2D são consideradas para expandir a cobertura da rede e reduzir tal alta taxa de rejeição. Assim, as tecnologias de comunicação D2D são incorporadas tanto na definição do problema quanto nas restrições do modelo apresentado neste artigo. Como outra limitação, Araújo et al. utilizam apenas um tipo de demanda, o que não reflete adequadamente a realidade de uma *Smart City*. Este aspecto é tratado no presente estudo, que considera diversas demandas reais de aplicações de IoT, extraídas de Santos et al. [2021].

Em relação aos modelos de otimização para comunicação D2D, em Zhou et al. [2020], é proposto um modelo baseado em *big data* baseado em dados reais, utilizando os dados reais de tráfego urbano para explorar e resolver o problema de planejamento das estações e estabelecer o sistema de otimização correspondente, a fim de injetar novo poder científico na modernização das cidades tráfego verde.

Em Pawar e Trivedi [2021] abordam a comunicação através de uma estratégia conjunta de alocação de recursos *uplink-downlink*, chamada JUDRA. Tal estratégia busca maximizar o rendimento da rede, e simultaneamente, garantir a QoS para os UEs e os usuários de D2D (*D2D users*, DUs). Em particular, a abordagem JUDRA permite que os DUs reutilizem os recursos das UEs de *uplink* ou *downlink*. Para isso, o problema de otimização é formulado como um problema de Programação Não Linear Inteiro Misto (PNLIM). O problema é resolvido em duas etapas: a etapa de atribuição de subcanais e a etapa de alocação de potência, fazendo uso das características da programação geométrica sequencial.

Em Xing et al. [2019], é avaliada a EC Móvel (*Mobile EC*, MEC), para fornecer recursos

de computação e/ou armazenamento semelhantes à CC na borda da rede. Essa estratégia busca reduzir atrasos de computação, e conservar energia para os dispositivos sem fio. A proposta de Xing et al. avalia um novo sistema MEC multi-ajudante habilitado para D2D. Neste caso, um usuário local solicita que seus dispositivos móveis próximos sirvam como ajudantes para computação cooperativa. O problema é formulado como PNLIM. Com base no modelo em PNLIM, foi proposto um algoritmo eficiente relaxando o problema original em um problema convexo e, em seguida, construindo uma solução viável com base na relaxação. Foi considerado um esquema de *benchmark* que adota UEs de suas capacidades computacionais máximas. Para fins de comparação, também foi desenvolvido um algoritmo heurístico baseado em uma atribuição gulosa de tarefas.

3. Definição Formal do Problema

A definição do problema descreve a interação entre UEs, EC e de CC em uma *Smart City*. É assumido que os UEs são capazes de estabelecer conexões D2D, ampliando assim a cobertura de rede na região. Nesse contexto, os UEs produzem dados que precisam ser enviados e processados em algum servidor do Substrato de Rede (*Substrate Network*, SN). Sejam $\mathcal{S}$, $\mathcal{E}$ e $\mathcal{C}$ respectivamente os conjuntos de UEs, servidores de EC integrados com estação base e servidores CC. Assume-se que cada UEs $s \in \mathcal{S}$ pode se conectar, com tecnologias sem fio, a uma estação base associada a um servidor $e \in \mathcal{E}$ se s estiver dentro do raio de cobertura de e . Este comportamento é definido pelo parâmetro c_{se} , que possui o valor 1 se ambos podem se conectar, e 0, caso contrário. Cada $s \in \mathcal{S}$ pode se conectar diretamente a outro $s' \in \mathcal{S}$ com tecnologias sem fio D2D se s estiver dentro do raio de cobertura de s' ($c_{ss'} = 1$). Adicionalmente, cada servidor $e \in \mathcal{E}$ pode se conectar de modo cabeado a um servidor $c \in \mathcal{C}$, e cada servidor $c \in \mathcal{C}$ pode ainda se conectar de modo cabeado a outro servidor $c' \in \mathcal{C}$. Em todos os casos, a conexão só pode acontecer se existir largura de banda residual suficiente. A cada conexão realizada existe uma latência relacionada à transmissão dos dados.

A Figura 1 mostra uma infraestrutura de um SN tradicional, sem a comunicação D2D. Neste caso, os UE1 e UE2 não estão dentro do raio de cobertura de nenhuma estação base, resultando na impossibilidade de atendimento. No entanto, considerando a comunicação D2D (Figura 2), os dispositivos UE1 e UE2 poderiam estabelecer conexão com a rede. Por exemplo, o UE1 poderia se conectar à estação base utilizando o UE4 como ponto de acesso. Um comportamento similar poderia ser observado com o UE2, que se conectaria à rede através dos UE1 e UE4.

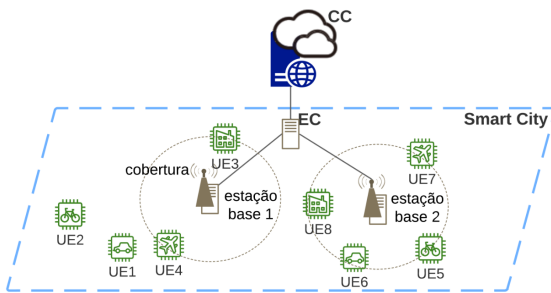


Figura 1: SN sem comunicação D2D.

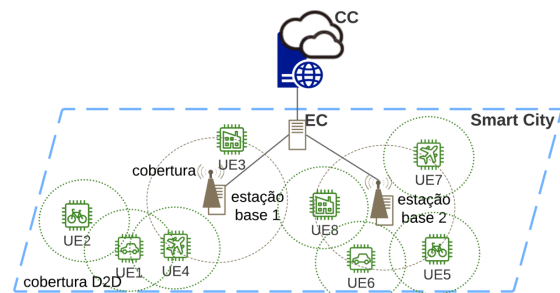


Figura 2: SN com comunicação D2D.

Cada UE $s \in \mathcal{S}$ gera uma quantidade de dados e possui demandas de processamento e memória, representadas respectivamente por $p_s \in \mathbb{R}_+$ (*cores*) e $m_s \in \mathbb{R}_+$ (MB), que precisam ser atendidas em algum servidor do SN com capacidades suficientes. O SN possui nós de EC e nós de CC. Cada nó $e \in \mathcal{E}$ e $c \in \mathcal{C}$ possui, respectivamente, capacidades de processamento P_e e P_c (*cores*), e memória M_e e M_c (MB). Tanto os $s \in \mathcal{S}$ quanto os $c \in \mathcal{C}$ e $e \in \mathcal{E}$ possuem posicionamento geográfico. Cada servidor $c \in \mathcal{C}$ e $e \in \mathcal{E}$ possui ainda um atraso de processamento da demanda solicitada ou, caso a demanda não seja atendida no nó em questão, um atraso para examinar os dados

e repassá-los para um próximo nó na rede. Seja $b_s \in \mathbb{R}_+$ a largura de banda demandada por cada $s \in \mathcal{S}$ (Mb/s). Sejam ainda $B_{ij} \in \mathbb{R}_+$ parâmetros, em Mb/s, que representam, respectivamente, a largura de banda disponível para a conexão entre os dispositivos $i, j \in \mathcal{E} \cup \mathcal{C}$.

4. Modelo em Programação Linear Inteira

A PLI é uma técnica de otimização utilizada para resolver problemas nos quais as variáveis de decisão são restritas a valores inteiros. As variáveis, restrições e funções utilizadas no modelo são apresentadas na sequência desta seção. Os parâmetros e terminologias utilizadas são as mesmas apresentadas na Seção 3.

4.1. Variáveis

- $y^s \in \{0, 1\}$, igual a 1 se as demandas de cada UE $s \in \mathcal{S}$ foram mapeadas com sucesso no SN;
- $w_i^s \in \{0, 1\}$, igual a 1 se o nó $i \in \mathcal{S} \cup \mathcal{E}$ do SN está sendo utilizado para a conexão do UE $s \in \mathcal{S}$;
- $z_i^s \in \{0, 1\}$, igual a 1 se o nó $i \in \mathcal{E} \cup \mathcal{C}$ do SN foi alocado para atender as demandas de p_s e m_s da UE $s \in \mathcal{S}$. Referidas demandas podem ser atendidas tanto na camada de EC, quanto de CC;
- $x_{ij}^s \in \{0, 1\}$, igual a 1 se uma conexão virtual originada do UE $s \in \mathcal{S}$, entre o ponto de conexão e o local de processamento no SN, foi alocada em conexões $(i, j) \in \mathcal{S} \cup \mathcal{E} \cup \mathcal{C}$;
- $v_i \in \{0, 1\}$, igual a 1 se o nó $i \in \mathcal{E} \cup \mathcal{C}$ do SN está ativo para processar as demandas de algum UE $s \in \mathcal{S}$, ou se é usado como nó de roteamento para algum outro nó do SN.

4.2. Restrições

$$\sum_{s' \in \mathcal{S}: s' \neq s} w_{s'}^s c_{ss'} + \sum_{i \in \mathcal{E}} w_i^s c_{si} = y^s, \quad \forall s \in \mathcal{S} \quad (1)$$

$$\sum_{i \in \mathcal{E}} z_i^s + \sum_{i \in \mathcal{C}} z_i^s = y^s, \quad \forall s \in \mathcal{S} \quad (2)$$

$$\sum_{s \in \mathcal{S}} z_i^s p_s \leq v_i P_i, \quad \forall i \in \mathcal{E} \cup \mathcal{C} \quad (3)$$

$$\sum_{s \in \mathcal{S}} z_i^s m_s \leq v_i M_i, \quad \forall i \in \mathcal{E} \cup \mathcal{C} \quad (4)$$

$$\sum_{s \in \mathcal{S}} x_{ij}^s c_{ij} b_s \leq B_{ij}, \quad \forall (i, j) \in \mathcal{E} \cup \mathcal{C} \quad (5)$$

$$x_{ij}^s c_{ij} \leq v_i, \quad \forall s \in \mathcal{S}, \forall (i, j) \in \mathcal{E} \cup \mathcal{C} \quad (6)$$

$$x_{ij}^s c_{ij} \leq v_j, \quad \forall s \in \mathcal{S}, \forall (i, j) \in \mathcal{E} \cup \mathcal{C} \quad (7)$$

$$\sum_{j \in \mathcal{S} \cup \mathcal{E} \cup \mathcal{C}} x_{ij}^s c_{ij} - \sum_{h \in \mathcal{S} \cup \mathcal{E} \cup \mathcal{C}} x_{hi}^s c_{hi} = w_i^s - z_i^s, \quad \forall i \in \mathcal{S} \cup \mathcal{E} \cup \mathcal{C}, \forall s \in \mathcal{S} \quad (8)$$

As Restrições (1) garantem que cada UEs $s \in \mathcal{S}$ deve se conectar a um nó $e \in \mathcal{E}$ do SN, ou a outro $s' \in \mathcal{S}$, estabelecendo uma conexão D2D. Observa-se que para a conexão existir, os UEs devem estar dentro de um raio de cobertura, definido pelo parâmetro c_{ij} . As Restrições (2) asseguram que as demandas de cada $s \in \mathcal{S}$ devem ser atendidas por um nó $e \in \mathcal{E}$ ou $c \in \mathcal{C}$. Nestas restrições, as variáveis y^s são utilizadas em conjunto com a função objetivo para garantir o atendimento de cada demanda originada por cada $s \in \mathcal{S}$. As Restrições (3) e (4) garantem, respectivamente, o atendimento das capacidades de memória e processamento demandadas pelos UEs $s \in \mathcal{S}$ nos nós de EC e CC. As Restrições (5) garantem o atendimento à largura de banda demandada por cada UEs $s \in \mathcal{S}$ ao utilizar as conexões cabeadas entre componentes de EE e CC do SN. Por fim, as Restrições (6) e (7) garantem que cada nó $i \in \mathcal{E} \cup \mathcal{C}$ seja determinado como ativo, seja para processamento ou para encaminhar o fluxo para outro nó do SN. As Restrições (8) são clássicas e são chamadas na literatura de *Capacitated Multi-Commodity Flow*. Elas aplicam os princípios de conservação de fluxo para garantir a conectividade entre os UEs $s \in \mathcal{S}$, $e \in \mathcal{E}$ e $c \in \mathcal{C}$. Cada UEs $s \in \mathcal{S}$ pode se conectar a um nó $e \in \mathcal{E}$, ou, quando apropriado, a um nó $s' \in \mathcal{S}$, utilizando conexões D2D. Esses nós podem encaminhar a demanda para um nó de $e \in \mathcal{E}$ ou $c \in \mathcal{C}$. Um nó $c \in \mathcal{C}$ pode, por sua vez, redirecionar o fluxo para outro nó $c' \in \mathcal{C}$.

4.3. Função Objetivo

Duas funções objetivo são formuladas: uma considera os custos relacionados aos servidores ativos, e outra o atraso fim a fim. A função expressa pela Equação (9) indica o gasto associado ao processamento das demandas dos UEs, buscando minimizar os custos dos servidores ativos no SN. O parâmetro α_i representa o custo de utilização de um servidor $i \in \mathcal{E} \cup \mathcal{C}$ do SN em €. Tal parâmetro é definido com base na relação kWh e no custo da energia elétrica tarifado na Itália. Cada requisição gerada por um UE acarreta em um atraso fim a fim, decorrente da alocação de recursos do SN, processamento e transmissão de dados (roteamento). A minimização desse atraso é crucial e pode afetar diretamente a qualidade de experiência do serviço. Dentro desse contexto, uma função é formulada com o objetivo de reduzir esse atraso, conforme representado pela Equação (10). As equações (9) e (10) incluem uma constante $\mathcal{M}$ com um valor elevado, que atua como penalidade para cada demanda de dispositivo não atendida, sendo adotado $\mathcal{M} = 10^6 \times |\mathcal{S}|$.

$$\text{Minimizar: } \sum_{i \in \mathcal{E} \cup \mathcal{C}} \alpha_i v_i + \mathcal{M} \cdot \sum_{s \in \mathcal{S}} (1 - y^s) \quad (9)$$

$$\text{Minimizar: } \sum_{s \in \mathcal{S}} \left(\sum_{(i,j) \in \mathcal{S} \cup \mathcal{E} \cup \mathcal{C}} x_{ij}^s c_{ij} d_{atraso}(i,j) + \sum_{i \in \mathcal{E} \cup \mathcal{C}} d_{proc}(i) z_i^s + \mathcal{M} \cdot (1 - y^s) \right) \quad (10)$$

Assim como Araújo et al. [2024], a função $d_{proc}(i)$ representa o tempo necessário para processar os dados em um servidor. Caso o nó i seja utilizado apenas como um nó de encaminhamento dos dados no SN, o atraso considerado é da ordem de microssegundos (adotado 0.002 ms). No entanto, se o nó i estiver efetivamente processando as demandas do UE $s \in \mathcal{S}$, o tempo considerado é de 5 ms. A função $d_{atraso}(i, j)$ pode ser decomposta em atraso de transmissão e propagação, estes calculados da mesma forma e com os mesmos parâmetros de Araújo et al. [2024].

5. Experimentos Computacionais

Os experimentos foram conduzidos em um computador Intel i5-10210, com 8GB de RAM e SO Ubuntu 14.04.2. O simulador empregado, implementado em C++, foi adaptado do código disponibilizado em Araújo et al. [2024]. Os modelos foram resolvidos utilizando o solver CPLEX.

As métricas utilizadas foram extraídas de Queiroz et al. [2022]; Araújo et al. [2024], e sendo: a) taxa de aceitação, definida como a proporção de demandas atendidas em relação às solicitadas; b) tempo de processamento, medido em milissegundos (ms); c) custo, Equação (9); e d) atraso, Equação (10). Para a avaliação das métricas c e d, a penalidade $\mathcal{M}$ é desconsiderada.

Assim como em Queiroz et al. [2022]; Araújo et al. [2024], a topologia do SN é baseada no projeto de uma *Smart City* em Modena, na Itália. No SN, os servidores de EC e CC estão localizados em edifícios municipais. Do mesmo modo que em Araújo et al. [2024], é considerado que, se necessário, o servidor de CC municipal pode enviar dados para serem processados em um servidor de CC da *Amazon Elastic Compute Cloud* (Amazon EC2) em Milão¹. O SN é composto por 10 servidores de EC, 1 de CC municipal e 1 de CC da EC2. As configurações adotadas são mostradas na Tabela 2. Considera-se que os servidores de EC do SN podem comunicar-se diretamente com o servidor de CC municipal com largura de banda de 1 Gbps. Em modo D2D, os UEs também possuem tal largura de banda. Por sua vez, o servidor de CC municipal pode comunicar-se diretamente com o servidor de CC EC2 com largura de banda de 10 Gbps. A distribuição física dos servidores é mostrada na Figura 3. Assim como em Queiroz et al. [2022], o servidor de CC municipal está localizado no prédio da prefeitura (ponto vermelho na Figura 3). Os pontos pretos representam as estações base equipadas com servidores de EC.

¹Serviço que oferece recursos computacionais sob demanda, disponível em <http://tiny.cc/AMZEC2>

Tabela 2: Configuração dos servidores¹.

Modelo	P_i (cpu)	M_i (GB)	α_i (€\$ ⁴)	Tipo
HX-CPU-I4309Y	8	128	0.0112	EC
HX-CPU-I5315Y	8	128	0.0151	EC
HX-CPU-I4310T	10	128	0.0112	EC
HX-CPU-I4316	20	128	0.0161	CC municipal
m6g.12xlarge	48	192	2.0100	CC Amazon EC2

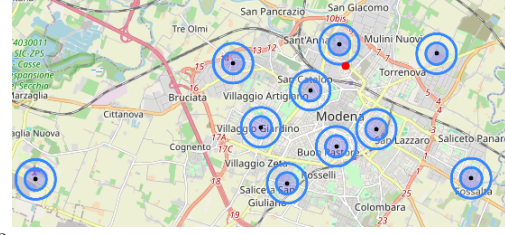
¹O custo α_i de cada servidor da CISCO [2023] é definido com base no custo da energia elétrica na Itália (€\$0.107/kWh). Os valores da Amazon são definidos pela própria empresa.

Tabela 3: Parâmetros de comunicação.

Tipo	Largura de banda teórica	Largura de banda média ²	Cobertura
4G	1 Gbps	10 Mbps	1 - 3 Km
5G	10 Gbps	120 Mbps	até 600
6G	até 1Tbps	1Gbps	até 320
D2D	—	até 13.5 Mbps	até 500

²A largura de banda e a cobertura são influenciadas por diversos fatores. Os valores fornecidos são aproximações baseadas em Haus et al. [2017]; Guo et al. [2020]; Araújo et al. [2024].

Figura 3: Distribuição de servidores de EC³.



³Os círculos mostram a cobertura de cada estação base, sendo o interno com parâmetros da tecnologia 6G e o externo considerando a 5G (Tabela 3).

Tabela 4: Característica dos UEs⁴.

Dispositivo	p_i (cpu)	m_i (GB)	Quantidade de dados (bits)
Lixeira inteligente	1.25	2.25	29600
Qualidade do ar	1.25	1.75	74400
Câmera de vigilância	2.00	3.00	1200000

⁴Configurações baseadas em aplicações retiradas do trabalho de Santos et al. [2021].

Em relação à conexão D2D, vários aspectos podem ser considerados, como a qualidade do canal, o ganho da antena e as interferências. Neste trabalho, esses aspectos são abstraídos, assumindo-se que esses dispositivos são capazes de fornecer comunicação D2D de médio alcance, utilizando a tecnologia *Long-Term Evolution (LTE) Direct* [Haus et al., 2017]. Para avaliar a cobertura e expansão das tecnologias de redes sem fio com uso de tecnologias D2D, a posição física dos UEs segue a mesma utilizada em Queiroz et al. [2022], mas considerando dispositivos que estejam a até 1.3 km de alguma estação base. A configuração dos UEs utilizados segue uma distribuição uniforme, com parâmetros mostrados na Tabela 4. O número de UEs avaliados varia em $|S| \in \{20, 40, 60, 80, 100\}$. Os parâmetros gerais de comunicação são mostrados na Tabela 3. Os demais parâmetros seguem os mesmos definidos em Araújo et al. [2024].

Os algoritmos propostos são desenvolvidos com base no modelo em PLI (Seção 4)². A variação dos modelos é definida como: $PLItGf$, onde t indica a tecnologia de rede em questão, podendo variar entre 5G e 6G. Adicionalmente, f indica a função objetivo, podendo variar entre minimizar custos (c , Equação 9), e minimizar o atraso (a , Equação 10). As abordagens utilizados para comparação (*baseline*, b) seguem o modelo proposto por Araújo et al. [2024], que não utiliza comunicação D2D. Neste caso, a variação das abordagens é definida como: $btGf$, onde t e f são definidos seguindo a notação das abordagens $PLItGf$. O solver utilizado resolve cada modelo, com um limite de tempo de até 300 segundos por execução. Tal limitação pode gerar soluções viáveis, mas não ótimas. Caso alguma solução não ótima seja gerada, ela é indicada nos experimentos.

5.1. Análise de Desempenho

Em cada gráfico apresentado, o número de UEs processadas varia, permitindo observar a métrica em questão no eixo y e as abordagens no eixo x . Os valores indicados em vermelho representam uma solução viável, porém não ótima, devido ao limite de tempo máximo de processamento estipulado. Em termos de taxa de aceitação (Figura 4), em conformidade com os resultados da literatura [Araújo et al., 2024], mesmo quando as UEs processadas estão distribuídas até distância de apenas 1.3 km das estações base, a taxa de aceitação das demandas processadas foi baixa. Tal comportamento ocorre devido à conexão com as tecnologias 5G e 6G possuir um raio de cobertura

²O código dos algoritmos, bem como as instâncias utilizadas, estão disponíveis no link <https://tinyurl.com/SBPocode2024>

baixo, resultando em uma alta taxa de rejeição. Este comportamento ocorre independentemente do número de UEs processadas, sendo mais evidente nos experimentos com o 6G, devido à maior limitação de cobertura. Como exemplo, o algoritmo b5gc com 100 UEs tem uma taxa de aceitação de apenas 46% e o b6GC de apenas 13%. Em contrapartida, corroborando a hipótese investigada, ao utilizar a comunicação D2D, a cobertura da rede se estende, reduzindo o número de rejeições. Tal melhora é promissora ao comparar os modelos com e sem D2D, neste caso, para as tecnologias 6G, considerando o cenário com 100 UEs este valor aumenta de 13% com o algoritmo b6Ga para 76% com o algoritmo PLI6Ga, indicando uma melhora de $\approx 486.6\%$.

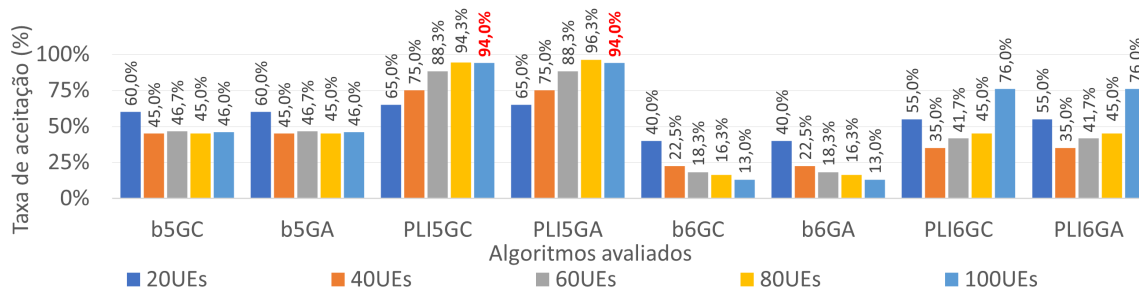


Figura 4: Taxa de aceitação das abordagens, variando número de UEs processados

Ainda em relação à taxa de aceitação (Figura 4), percebe-se nas abordagens que fazem uso das tecnologias D2D que, quanto mais UEs são processadas juntas, menor é a rejeição por inviabilidades de conexão. Como exemplo, com o algoritmo PLI5Ga e 20 UEs processadas, a taxa de aceitação é de 65%, mas, ao aumentar o número de UEs para 100, tal taxa de aceitação aumenta aproximadamente 44.6%, chegando a 94% (mesmo sendo uma solução viável, mas não ótima). Este fato indica que, com poucas UEs, a conexão D2D estende pouco a cobertura da rede. Com mais UEs sendo processadas, a cobertura é maior, visto que elas são utilizadas como ponto de acesso para a rede. Em relação à taxa de aceitação, não existe diferença entre minimizar o custo ou o atraso fim a fim, indicando que esses aspectos não são cruciais para definir a taxa de aceitação neste cenário. Contudo, ao analisar a métrica de custo, percebe-se que sim, existe uma variação significativa entre as diferentes funções objetivo utilizadas. Neste caso, o custo chega a ser de 0.094€\$ para o algoritmo PLI6Gc e 20 UEs, e de 2.120€\$ para o algoritmo PLI6Ga e 20 UEs, um aumento de $\approx 2155\%$. Este fato ocorre porque o algoritmo que minimiza custos efetua a alocação de recursos em dispositivos com menor custo, evitando, por exemplo, o uso de servidores de CC da EC2, que possuem um custo alto em relação aos dispositivos de EC municipais.

Assim como em relação à otimização do custo, que gerou uma variação de $\approx 2155\%$ (figura omitida por questões de espaço), a escolha da função objetivo também pode gerar variações significativas em relação ao atraso (Figura 5). Como exemplo, observando os algoritmos PLI5Ga e PLI5Gc, mesmo com uma taxa de aceitação próxima (Figura 4), estes possuem respectivamente um atraso médio fim a fim de 20.8 ms e 218.1 ms, uma diferença $\approx 948\%$. Tal comportamento é decorrente de aspectos como, ao minimizar o atraso, o algoritmo prioriza a alocação de recursos em dispositivos de EC, mesmo com um custo mais alto, pois estão mais próximos e geram um atraso menor. Por outro lado, ao minimizar os custos, os roteamentos possuem mais saltos em termos de conexão entre nós do SN, priorizando a redução de custos gerada pelo atendimento centralizado em um número menor de servidores ativos.

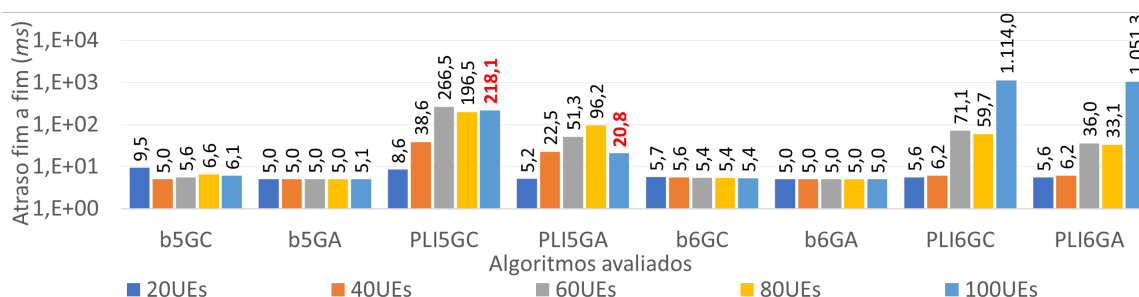


Figura 5: Atraso fim a fim médio (Equação 10), variando número de UEs processados

Ao analisar a Figura 5, percebe-se que, assim como constatado em Araújo et al. [2024], devido às altas taxas de transferência, as tecnologias 5G e 6G são promissoras para reduzir os atrasos fim a fim. Como exemplo, as abordagens b5Ga e b6Ga possuem um atraso de comunicação de aproximadamente 5 ms, habilitando uma comunicação rápida de dados entre os UEs e servidores. Os tempos próximos entre as abordagens b5Ga e b6Ga se devem a fatores como a alta taxa de rejeição (Figura 4) e a baixa quantidade de dados demandados para serem trafegados e processados em algum servidor da rede. Como poucos dados são demandados, o tempo de transmissão é baixo e próximo entre as abordagens. Como a taxa de rejeição é alta, os dados podem ser processados perto das UEs que originam as demandas, com roteamentos curtos em relação ao número de saltos, visto que existe uma baixa disputa por recursos do SN. Nestes casos, existem vários recursos ociosos na rede e as restrições ficam com folgas em relação à capacidade residual. Contudo, ao considerar a comunicação D2D, os atrasos fim a fim tendem a subir, devido a tal tecnologia possuir uma velocidade de transmissão inferior às tecnologias 5G e 6G (Tabela 3). Como exemplo, o algoritmo b6Ga gera um tempo de ≈ 5 ms para 100 UEs, e o algoritmo PLI6Ga de ≈ 1051 ms, um aumento de $\approx 20000\%$. Isso evidencia um *trade-off* entre adotar a comunicação D2D e aumentar a taxa de aceitação; ou reduzir o atraso de comunicação, mas impactando no atraso fim a fim.

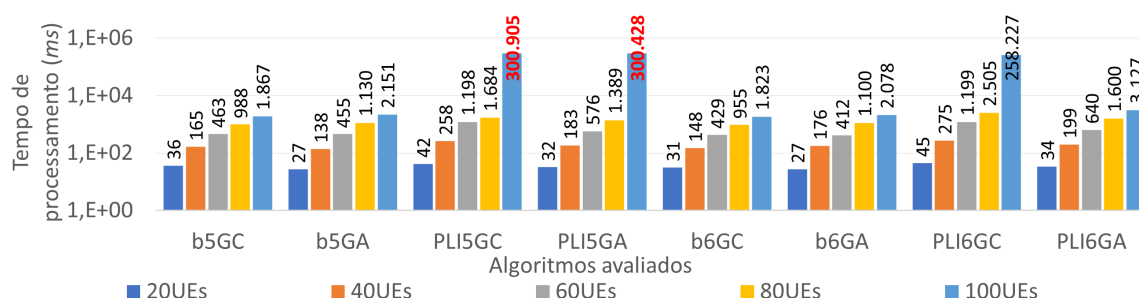


Figura 6: Tempo de processamento, variando número de UEs processados

Em relação ao tempo de processamento (Figura 6), este é mais alto para cenários com mais UEs processadas. Ao aumentar linearmente o número de UEs processadas, o tempo de processamento aumentou significativamente, principalmente em relação algoritmo que realiza a comunicação D2D. Como exemplo, considerando o algoritmo PLI6Gc e os cenários com 20 UEs e 100 UEs, esse tempo aumentou, respectivamente, de $\approx 45ms$ para $\approx 258s$. Tal fato é justificado por um aumento no número de variáveis e restrições analisadas, sendo 42728 restrições e 21813 variáveis para o algoritmo PLI6Gc e 20 UEs processadas; e 2532968 restrições e 1277013 variáveis para 100 UEs processadas. Um aumento de aproximadamente $\approx 5828\%$ no número de restrições e $\approx 5754\%$ de variáveis geradas pelo modelo. Este comportamento indica que um aumento linear

no número de componentes processados na rede, gera um aumento com características exponenciais no tempo de execução. Como hipótese a ser investigada em trabalhos futuros, supõe-se que esse comportamento é característico de problemas pertencentes à classe NP-difícil em sua versão de otimização. Também como hipótese a ser investigada, considera-se que o problema tratado é uma variação do *Location Routing Problem* (LRP), que é conhecido por pertencer à classe NP-difícil.

6. Conclusões e Trabalhos Futuros

Neste estudo foi abordado problema de alocação de recursos em EC e CC com comunicação D2D para atender demandas de UEs, como dispositivos de IoT. Uma definição formal do problema foi discutida, a qual, com base na literatura, incorpora características realistas do cenário de *Smart Cities*. O problema foi formulado utilizando PLI, com restrições de alocação de recursos, capacidade e roteamento entre os componentes da rede, incluindo a comunicação direta entre UEs, conhecida como D2D. Seguindo a literatura, duas funções objetivo foram consideradas: minimizar os custos dos servidores utilizados e minimizar os atrasos fim a fim. Para evitar erros de decisão e possíveis vieses provenientes de heurísticas, os experimentos foram conduzidos utilizando algoritmos exatos com simulações. Experimentos detalhados foram realizados em instâncias da literatura, considerando aspectos contemporâneos das tecnologias de comunicação, incluindo o 6G.

Os experimentos foram conduzidos comparando abordagens baseadas no modelo em PLI e outros da literatura. Observou-se que a escolha entre minimizar o custo e o atraso fim a fim gera um *trade-off*. Ao minimizar o atraso, os roteamentos são mais curtos e os dados são processados próximos dos UEs, porém o número de servidores ativos aumenta, resultando em um aumento no custo de operação. Por outro lado, ao minimizar o custo, o número de servidores ativos é reduzido, levando ao processamento centralizado em menos componentes do SN, mas com um atraso fim a fim médio maior. Percebeu-se que ao fornecer comunicação D2D, a melhoria percebida foi de até aproximadamente $\approx 486\%$ na taxa de aceitação. Nos cenários avaliados, notou-se que quanto mais UEs são processadas, maior é o raio de cobertura, reduzindo a taxa de rejeição. Por outro lado, quanto mais demandas de UEs são processadas, maior também é o tempo de processamento, podendo aumentar em até $\approx 20000\%$. Este comportamento indica que, embora o algoritmo exato gere uma análise assertiva dos resultados com base em soluções ótimas, tende a ser inviável de ser adotado na prática. Na vida real, a quantidade de componentes processados na rede é alta, o que aumentaria ainda mais o número de variáveis e restrições a serem analisadas, e consequentemente o tempo de execução. Assim, em trabalhos futuros, pretende-se fazer uma análise de complexidade do problema e propor algoritmos heurísticos e metaheurísticos que, apesar de abrirem mão da otimalidade da solução, gerem uma tomada de decisão em tempo viável para o cenário analisado.

Agradecimentos

Este trabalho foi realizado com o apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) por meio do Programa de Excelência Acadêmica (PROEX). Os autores agradecem também ao CNPq e FAPEMIG.

Referências

- Alwis, C. D., Kalla, A., Pham, Q.-V., Kumar, P., Dev, K., Hwang, W.-J., e Liyanage, M. (2021). Survey on 6G Frontiers: Trends, Applications, Requirements, Technologies and Future Research. *IEEE Open Journal of the Communications Society*, 2:836–886.
- Araújo, S., Moreira, M., e Mateus, G. (2024). 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*, p. 225–238, Porto Alegre, RS, Brasil. SBC. URL <https://sol.sbc.org.br/index.php/sbrc/article/view/29795>.

- Banafaa, M., Shayea, I., Din, J., Azmi, M. H., Alashbi, A., Daradkeh, Y. I., e Alhammadi, A. (2023). 6G Mobile Communication Technology: Requirements, Targets, Applications, Challenges, Advantages, and Opportunities. ISSN 11100168.
- CISCO (2023). Cisco HyperFlex - All Flash and Hybrid Server Nodes Spec Sheet. Spec Sheet REV A.25, CISCO SYSTEMS.
- Gismalla, M. S. M., Azmi, A. I., Salim, M. R. B., Abdullah, M. F. L., Iqbal, F., Mabrouk, W. A., Othman, M. B., Ashyap, A. Y., e Supa'at, A. S. M. (2022). Survey on Device to Device (D2D) Communication for 5GB/6G Networks: Concept, Applications, Challenges, and Future Directions. *IEEE Access*, 10:30792–30821. ISSN 21693536.
- Guo, T., Schormans, J., Xu, L., Wu, J., e Cao, Y. (2020). Proximity as a Service for the Use Case of Access Enhancement Via Cellular Network-Assisted Mobile Device-to-Device. *IEEE Access*, 8:31562–31573.
- Haus, M., Waqas, M., Ding, A. Y., Li, Y., Tarkoma, S., e Ott, J. (2017). Security and Privacy in Device-to-Device (D2D) Communication: A Review. *IEEE Communications Surveys e Tutorials*, 19(2):1054–1079.
- Khan, L. U., Yaqoob, I., Tran, N. H., Kazmi, S. M. A., Dang, T. N., e Hong, C. S. (2020). Edge-Computing-Enabled Smart Cities: A Comprehensive Survey. *IEEE Internet of Things Journal*, 7 (10):10200–10232.
- Murroni, M., Anedda, M., Fadda, M., Ruiiu, P., Popescu, V., Zaharia, C., e Giusto, D. (2023). 6G—Enabling the New Smart City: A Survey. ISSN 14248220.
- Pawar, P. e Trivedi, A. (2021). Joint uplink-downlink resource allocation for d2d underlaying cellular network. *IEEE Transactions on Communications*, 69(12):8352–8362.
- Queiroz, T. A. d., Canali, C., Iori, M., e Lancellotti, R. (2022). *An Optimization View to the Design of Edge Computing Infrastructures for IoT Applications*, p. 1–30. Springer International Publishing, Cham. ISBN 978-3-030-80821-1. URL https://doi.org/10.1007/978-3-030-80821-1_1.
- Santos, J., Wauters, T., Volckaert, B., e De Turck, F. (2021). Towards end-to-end resource provisioning in Fog Computing over Low Power Wide Area Networks. *Journal of Network and Computer Applications*, 175:102915.
- Silva, P., Costan, A., e Antoniu, G. (2019). Investigating Edge vs. Cloud Computing Trade-offs for Stream Processing. In *2019 IEEE International Conference on Big Data (Big Data)*, p. 469–474.
- TGT Consult (2022). Provider Lens™ Internet das Coisas (IoT) – Serviços e Soluções 2022. <https://tinyurl.com/TGTIoT2022>. Acessado: 2024-04-03.
- Xing, H., Liu, L., Xu, J., e Nallanathan, A. (2019). Joint Task Assignment and Resource Allocation for D2D-Enabled Mobile-Edge Computing. *IEEE Transactions on Communications*, 67(6): 4193–4207.
- Zhou, J., Wang, P., e Xie, L. (2020). Research on resource allocation optimization of smart city based on big data. *IEEE Access*, 8:158852–158861. ISSN 21693536.