KICS The Korean Institute of Communications and Information Sciences

Available online at www.sciencedirect.com

ScienceDirect

ICT Express 10 (2024) 1151–1173

www.elsevier.com/locate/icte

Edge computing in future wireless networks: A comprehensive evaluation and vision for 6G and beyond

Mustafa Ergen[a], Bilal Saoud[a,b,c,*], Ibraheem Shayea[a], Ayman A. El-Saleh[d,*], Onur Ergen[a,e], Feride Inan[f], Mehmet Fatih Tuysuz[g]

[a] *Electronics & Communications Engineering Department, Faculty of Electrical and Electronics Engineering, Istanbul Technical University (ITU), 34469, Istanbul, Turkey*

[b] *Electrical Engineering Department, Faculty of Sciences and Applied Sciences, University of Bouira, 10000, Bouira, Algeria*

[c] *LISEA Laboratory, Faculty of Sciences and Applied Sciences, University of Bouira, 10000, Bouira, Algeria*

[d] *Department of Electronics and Communication Engineering, College of Engineering, A'Sharqiyah University (ASU), Ibra 400, Oman*

[e] *Electrical and Electronics Engineering, Koc University, 34450, Istanbul, Turkey*

[f] *Ambeent Inc., Yuba, CA 95991, USA*

[g] *Computer and Information Sciences Department, Northumbria University, Newcastle Upon Tyne NE1 8ST, UK*

Received 9 June 2024; received in revised form 22 July 2024; accepted 11 August 2024
Available online 17 August 2024

Abstract

Future internet aims to function as a neutral in-network storage and computation platform, essential for enabling 6G and beyond wireless use cases. Information-Centric Networking and Edge Computing are key paradigms driving this vision by offering diversified services with fast response times across heterogeneous networks. This approach requires effective coordination to dynamically utilize resources like links, storage, and computation in near real-time within a non-homogenous and distributed computing environment. Additionally, networks must be aware of resource availability and reputational information to manage unknown and partially observed dynamic systems, ensuring the desired Quality of Experience (QoE). This paper provides a comprehensive evaluation of edge computing technologies, starting with an introduction to its architectural frameworks. We examine contemporary research on essential aspects such as resource allocation, computation delegation, data administration, and network management, highlighting existing research gaps. Furthermore, we explore the synergy between edge computing and 5G, and discuss advancements in 6G that enhance solutions through edge computing. Our study emphasizes the importance of integrating edge computing in future considerations, particularly regarding sustainable energy and standards.

© 2024 The Author(s). Published by Elsevier B.V. on behalf of The Korean Institute of Communications and Information Sciences. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Keywords: Wireless communication; 5/6G; IoT/ioE; Communication; Latency; Edge computing; CROWN

Contents

1. Introduction 1152
2. Background and literature review 1153
   2.1. Edge computing 1153
      2.1.1. Solutions based on edge 1154
      2.1.2. Content delivery networks (CDNs) 1154
      2.1.3. Ubiquitous computing 1154
      2.1.4. P2P 1154
      2.1.5. Cloud computing 1155
      2.1.6. Cloudlet 1155

* Corresponding authors.
E-mail addresses: mustafaergen@itu.edu.tr (M. Ergen), bilal340@gmail.com (B. Saoud), shayea@itu.edu.tr (I. Shayea), ayman.elsaleh@asu.edu.om (A.A. El-Saleh), oergen@itu.edu.tr (O. Ergen), feride.inan@tepav.org.tr (F. Inan), fatih.tuysuz@northumbria.ac.uk (M.F. Tuysuz).

Peer review under responsibility of The Korean Institute of Communications and Information Sciences (KICS).

https://doi.org/10.1016/j.icte.2024.08.007

2405-9595/© 2024 The Author(s). Published by Elsevier B.V. on behalf of The Korean Institute of Communications and Information Sciences. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

2.1.7. Mobile edge computing 1155
2.1.8. Fog computing 1155
2.2. Literature review 1156
3. Edge computing fundamentals 1157
3.1. Resource management 1157
3.2. Computation offloading 1158
3.3. Data management 1158
3.4. Network management 1158
3.5. Security and privacy 1158
3.6. EC pricing and billing 1158
3.7. Practical implications and potential applications of EC 1159
4. Resource allocation 1159
4.1. Integrated sensing and communications (ISAC) in EC 1164
5. Lack of research 1164
6. Edge computing and 5G/6G networks 1164
6.1. Synergy between edge intelligence and 5G/6G networks 1165
7. Contractual, reputational, and opportunistic wireless networking 1166
8. Discussion of the future EC and MEC 1168
9. Conclusion 1169
CRediT authorship contribution statement 1169
Acknowledgments 1169
References 1169

1. Introduction

Cisco forecasts that the global number of connected devices will surpass 75 billion by the year 2025 (see Fig. 1). Cisco predicts that the amount of devices connected to the Internet will reach 500 billion by 2030 [1]. A centralized cloud system by itself will not have the capability to efficiently handle upcoming networks [2–4]. In this setting, there will be a vast number of devices that are outfitted with sensors and designed to gather massive quantities of data. These devices will then use this data to make inferences and carry out certain actions. The transfer of substantial volumes of data from interconnected devices to the cloud for processing will result in a highly congested network infrastructure [5–9]. Furthermore, the bidirectional movement of data between the cloud and specific devices leads to an increase in latency. A multitude of emerging technologies, including autonomous vehicles, telemedicine procedures, Augmented Reality/Virtual Reality (AR/VR), high-resolution 8 K video, sophisticated manufacturing robots and drone surveillance systems, necessitate instantaneous and extremely little delay in performance [8,10,11]. The current status of the Fifth-Generation (5G) network will become even more important if Sixth-Generation (6G) networks are introduced. Considering the updated requirements, networks (5G/6G) will require bands with higher frequencies and decreased latency in comparison to 5G networks [12–14]. Moreover, the implementation of more resilient Edge Computing (EC) will help address the needs for bandwidth, privacy and autonomy. The rapid digital transformation and increased adoption of cloud and edge technologies are leading to significant changes in WAN (Wide Area Network) edge architecture [15]. These changes are shaping the strategies about infrastructure and operations in the next future networks. Furthermore, EC

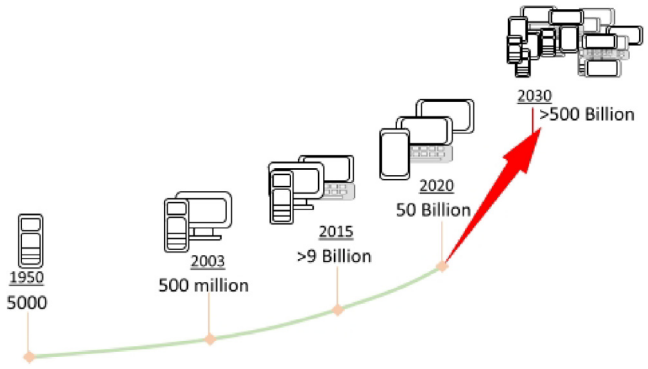


Fig. 1. Expansion of interconnected devices from 1950 to 2050.

has transitioned from being primarily focused on Internet of Things (IoT) to being widely recognized as a valuable addition to the centralized structures of hyper-scale cloud. The relevance of EC in distributed cloud and digital business guarantees its permanence, despite the emerging industry, technologies and architectures. Furthermore, the range of use-cases for EC is extensive, and initial implementations are highly tailored to specific requirements. Hence, it is imperative for the future network to create a long-term EC strategy that encompasses the diverse range of use cases demanded by their company. Finally, the convergence of space/terrestrial networks, particularly with the advent of 5G/6G networks, is becoming increasingly important. This is exemplified in the architecture depicted in Fig. 2 The motivation behind EC primarily revolves around addressing the limitations of traditional centralized cloud computing in an era of exponential data growth and the need for real-time processing [16,17]. EC brings data processing and storage closer to the data source, reducing latency and ensuring quicker response times essential for applications such as autonomous vehicles, augmented reality (AR) and virtual reality (VR). This proximity

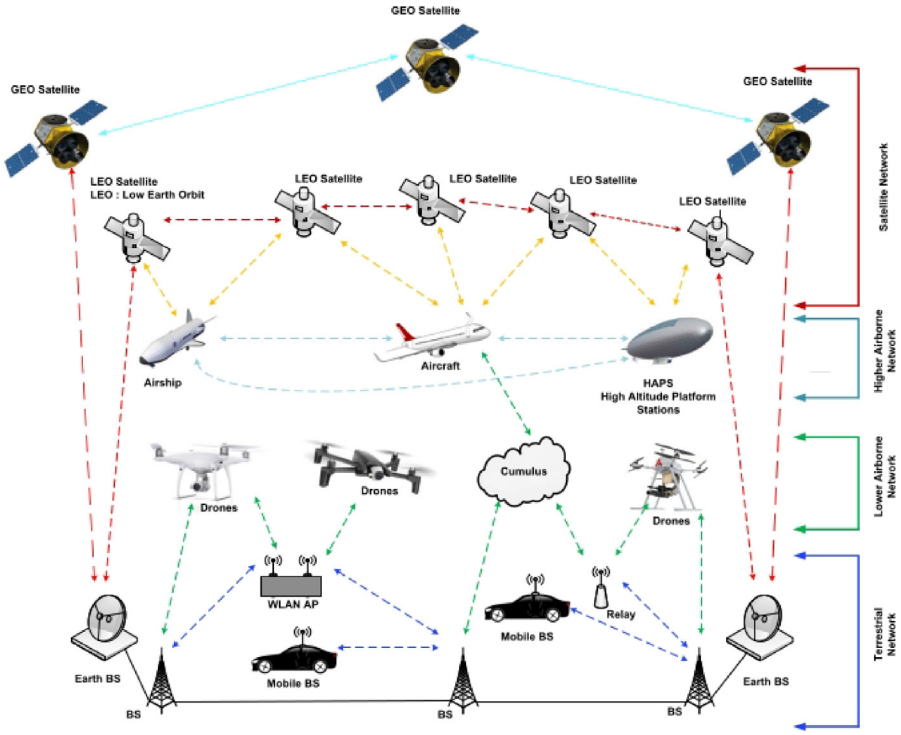


Fig. 2. Integrated space networks architecture in ultra-dense networks combines terrestrial and aerial networks.

also optimizes bandwidth usage by minimizing the volume of data transmitted to central servers, thereby alleviating network congestion. Moreover, EC enhances privacy and security by keeping sensitive data local, reducing the risk of breaches during transmission. The flexibility and scalability of EC allow it to adapt to varying workloads and dynamic network conditions efficiently. Additionally, with the support of advanced 5G/6G networks, which offer high-speed, reliable, and low-latency connectivity, EC is poised to unlock new potentials in smart cities, industrial automation, healthcare, and immersive media, making it a critical component in the future of digital infrastructure.

Considering these new challenges, data center operations need to be relocated to the network edge. However, edges enable specific time-sensitive and security-conscious based on Artificial Intelligence (AI) applications to function either solely on a device or in collaboration with localized data centers [16,17]. Based on the present circumstances, it appears that none of the suggested edge architectures [18–23] are capable of effectively managing the substantial volume of data traffic that will be processed in future networks. The majority of current solutions concentrate on deploying edge devices to fixed sites, such as factories or shopping malls or within certain geographical regions, such as metropolitan centers. This approach incurs the expense of setting up new infrastructure [24–27]. Research on leveraging connection among nearby end-devices to collaboratively perform tasks in local area computing groups is also increasing. Nevertheless, these endeavors are constrained in their extent.

Subsequently, this paper is structured as outlined below, background and literature review about EC is given in Section 2. Then, EC fundamentals are explained in Section 3. Section 4 will provide further elaboration on resource allocation in EC. Some points about the gap in EC research are discussed in Section 5. EC performances between 5G and 6G will be illustrated in Section 6. A new scheme to overcome EC challenges will be illustrated in Section 7. Section 8 will provide some insights into the future of application based on EC. Finally the paper will be closed by a conclusion (Section 9).

2. Background and literature review

This section offers a concise introduction to EC. This encompasses the notion, variety of solutions, and relevant research that has been conducted in the existing body of literature. This section will provide a concise overview of the research difficulties that need to be tackled.

2.1. Edge computing

The growth of Internet traffic has been exponential and is gaining new momentum on account of enhanced mobile broadband, IoT, and mission-critical applications. The number of mobile Internet users is projected to increase from 3.8 billion in 2019 to 5 billion by 2025. According to Cisco's forecast, there will be more than 500 billion devices connected to the Internet by 2030 [1]. Moreover, emerging applications, such as remote surgery, self-driving vehicles, AR/VR, 8 K video, drone surveillance communication and advanced robotics in manufacturing, which require high reliability, ultra-low latency and high bandwidth will add to the burden. Fig. 3 illustrates some tasks based on wireless networks. McKinsey predicts that network performance demands would grow ten times

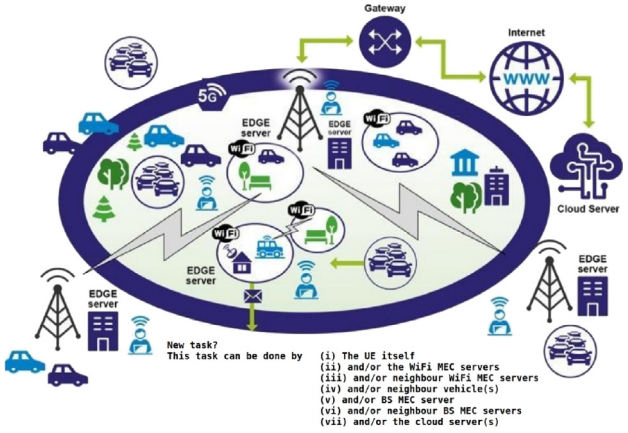


Fig. 3. Framework of distributed Edge Paradigm.

higher than current levels in all network aspects, such as latency, throughput, dependability and scale [2]. Even with 5G and beyond wireless networks – all of which entail costly infrastructure investments [28,29] – the same latency and congestion problems will persist due to the exponential growth in traffic and devices. In order to cope with exponentially growing demand, it is simply not enough to increase Internet and Radio Access Networks' (RAN) capacities. Due to these problems, data center operations are being moved to the periphery of the network. In future wireless networks, all communications' types, computing and caching resources are expected to be distributed throughout the network. Moving intelligence to the edge will dramatically lower latency and backbone network traffic as well as reduce the costs associated with building a network. Cognitive, distributed computing and storage strategies are essential to lift the burden off strained capacity and lower latency; however, allocating resources effectively across networks and devices is a highly challenging task due their heterogeneity and characteristics that change dynamically.

2.1.1. Solutions based on edge

Solutions based on edge enable data processing to be at the gateway or device level. This leads to minimize the necessity of transferring data between data center and/or the cloud, hence reducing bandwidth demands, latency and reliance on connectivity. In addition to technical considerations, decentralized computing offers energy efficiency by reducing the electricity and cooling expenses associated with large-scale data centers. The motivation behind research on EC stems from concerns over security and privacy associated with the centralized cloud, as seen by states, corporations, and consumers [30–35]. Simultaneously, the decreasing costs of compute and storage, along with the increasing prevalence of Machine Learning (ML), are driving the implementation of EC. As per the International Data Corporation (IDC), at least 40% from the generated data by the IoT will be analyzed, processed, acted and kept upon in nearby to, or directly at, the network edge. The systems commonly referred to as EC encompass cloudlets [36], Cyber Foraging [37], Fog Computing [38] and Mobile Edge Computing (MEC).

The advancement of this computing paradigm has undergone multiple phases since 1990. Fig. 4 demonstrates that every step had a direct or indirect impact on the notion and altered the perception of EC [39,40].

2.1.2. Content delivery networks (CDNs)

A team of researchers at MIT developed content delivery networks (CDNs) to address the flash crowd problem, which occurs when a single server is not able to handle a high volume of requests [41,42]. Scientists from MIT suggest duplicating the content on multiple strategically distributed edge servers. Furthermore, CDNs were the initial technology that could provide storage resources at the network edges. Over the past ten years, the utilization of CDNs has gained significant popularity among website owners. They store their data (HTML, images, Java-Scripts) with a CDN provider to enhance the web experience for their consumers.

2.1.3. Ubiquitous computing

Ubiquitous computing, often referred to as "pervasive computing," is a paradigm that envisions the integration of computing capabilities into every aspect of daily life. The fundamental idea is to embed computing devices and technologies seamlessly into the surrounding environment, making them an intrinsic part of our everyday activities. This concept transcends traditional computing models, emphasizing the creation of an interconnected network of smart devices that collaborate to enhance efficiency, convenience, and user experience. In a ubiquitous computing environment, devices communicate with each other, collect data from the surroundings, and adapt their functionality to suit the context. This pervasive integration aims to create a more natural and user-friendly computing experience, where technology fades into the background, becoming an invisible enabler of tasks and interactions in both personal and professional spheres. As technology advances, ubiquitous computing continues to shape the future of human–computer interaction, fostering a world where computing is seamlessly interwoven with the fabric of our daily lives.

2.1.4. P2P

Peer-to-peer (P2P) refers to a decentralized network architecture that enables direct communication and resource sharing between individual computers, or peers, without the need for a central server [16]. In a P2P network, each node has equal status, and participants can both contribute and consume resources. This approach contrasts with traditional client–server models, offering increased resilience, scalability, and reduced dependency on a single point of failure. P2P networks are commonly associated with file sharing, where users can exchange digital files directly between their devices. However, the concept extends to various applications, including communication, distributed computing, and decentralized cryptocurrencies like Bitcoin. P2P systems empower users by fostering a more inclusive and collaborative digital environment, promoting the efficient sharing of information and resources across a distributed network [43].

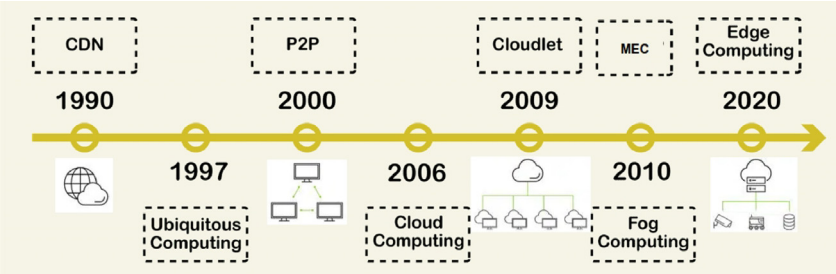


Fig. 4. A chronograph illustrating the evolution of EC.

2.1.5. Cloud computing

Cloud computing is a transformative paradigm in information technology that revolutionizes the way individuals and organizations access and manage digital resources. In essence, it involves the delivery of computing services-such as the capacity of storage, processing power, and applications-over the internet, with users accessing these resources on a pay-as-you-go basis. Cloud computing eliminates the need for organizations to invest in and maintain extensive physical infrastructure, allowing them to leverage the capabilities of remote data centers [44]. This model offers scalability, flexibility, and cost-effectiveness, as users can scale their computing resources up or down based on demand. Key service models within cloud computing include Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS). Whether for individuals storing files in the cloud or enterprises running complex applications, cloud computing has become an integral part of the modern digital landscape, enabling efficient collaboration, innovation, and resource optimization.

2.1.6. Cloudlet

A cloudlet is a concept in cloud computing that refers to a small-scale data center or a specialized computing node located at the edge of a network, closer to end-users and devices [36,45]. Positioned between mobile devices and larger cloud infrastructure, cloudlets aim to reduce latency and enhance the performance of real-time applications by bringing computing resources closer to the point of need. Acting as an intermediary layer, cloudlets facilitate offloading computational tasks from mobile devices to improve efficiency, response time, and conserve device resources. This approach is particularly beneficial for applications requiring low latency, such as AR/VR, and other time-sensitive tasks. Cloudlets represent a distributed computing model that complements the broader cloud infrastructure, providing a more responsive and adaptive computing environment for emerging EC applications [46].

2.1.7. Mobile edge computing

The term MEC has lately been substituted with Multi-Access Edge Computing. The European Telecommunication Standards Institute (ETSI) introduced Multi-Access Edge Computing in 2014, specifically targeting Virtual Machine (VM) technologies and mobile networks [47]. In 2017, its range was broadened to include other virtualization technologies such as non-mobile network needs. The initial concept aimed to extend cloud-computing features to the edges of mobile networks. In addition, the solution is based on RAN by deploying MEC servers at Long-Term Evolution (LTE)/LTE-A macro base station (eNodeB) sites, multi-Radio Access Technology (RAT) sites and 3G Radio Network Controller (RNC) sites. The ETSI project was tasked with defining the Application Programming Interfaces (APIs) that link the mobile edge platform and applications to encourage innovation in an open environment.

2.1.8. Fog computing

The concept of Fog computing was pioneered by Cisco in 2012. It can be seen as expansion of the cloud computing model that encompasses both the central infrastructure and the network's periphery [48]. Therefore, in contrast to MEC, fog computing is closely connected to the cloud and cannot function independently. Consequently, there has been particular emphasis on the communication between the cloud and fog [49–51]. In contrast to MEC, which is often implemented at a fixed location such as a base station, fog nodes have the flexibility to be installed in many locations that have a network connection, such as the factory floor, on top of a power pole, railway tracks, vehicles, and so on. Cisco provides application platforms that streamline the creation of fog applications, along with Cisco Fog Data Services for data analytics [52]. Simultaneously, major Internet corporations are deploying edge infrastructure. Facebook is constructing small-scale data centers tailored for various applications' types and workloads. The company of Amazon has introduced Amazon Web Services (AWS) Greengrass, which enables developers to offload certain operations to the device itself. Additionally, there are companies such as EdgeConneX and vXchnge that identify themselves as edge companies. These companies are actively constructing networks of urban data centers. One example is Vapor IO, a firm that creates micro-centers that can be installed in any location.

The majority of the suggested applications incur the expense of deploying supplementary infrastructure, either by putting edge devices at individual fixed sites (such as factories or shopping malls) or throughout particular geographical regions (such as metropolitan centers). Therefore, the scalability is adversely impacted by the substantial rise in the number of individuals conducting transactions within a certain edge domain. Our proposed method offers increased scalability as adoption increases. Put simply, with a large number of users needing to do storage and computing tasks on the network. A

significant increase in computational speed, which will enable the emergence of a genuinely global network, is anticipated.

2.2. Literature review

The emergence of edge-computing brings up new opportunities to balance the trade-off between distributed network architectures and centralized [53]. The literature on edge systems has largely focused on service time delay and to a lesser extent there is some literature, which takes into consideration application specifications. Research has also highlighted the need of coordination between cloud computing and EC [54–56]. [55] highlights decision-making on what actions to perform locally and what actions to be sent to the cloud. The study centers on utilizing network-wide knowledge and historical data stored at the cloud center to direct EC units in meeting diverse performance needs of different wireless IoT networks. There is a lack of scientific understanding on how different edge architectures impact system efficiency and performance, especially in terms of overall service time. Moreover, there is limited understanding of how the different offloading strategies can impact end-to-end performance of IoT applications and services, especially, when resource patterns and task characteristics are dynamic [57].

A majority of the literature on the EC environment focuses on minimizing service time, in particular, computational delay and communication delay as well as with other objectives such as cost and energy. In [58], an optimization framework of offloading from a single mobile device (MD) has been proposed to several edge devices to minimize the computational latency and power consumption of MDs. [59] propose low-complexity suboptimal algorithm to solve the computation offloading problem in a mixed fog/cloud system by jointly optimizing the offloading decisions and the allocation of computation resource, transmit power, and radio bandwidth while guaranteeing acceptable computational delay. In [60], authors design an optimal task scheduling policy to reduce the overall latency taking into consideration the average each task delay and the power consumption average at the mobile device. The computation tasks are scheduled based on the states of queuing of the task buffer, transmission unit and execution of the local processing unit. [61] proposes a method with a dual focus on computation and communication elements to minimize service delay in a scenario with two cloudlet servers. Processing delay is controlled via VM migration and improving transmission delay. The work proposed in [62] presents algorithms to reduce latency when using unmanned aerial vehicles (UAVs) to connect mobile devices to the fog node and the fog node to the remote data centre. Authors take into consideration both transmission power, processing frequencies of mobile devices and the fog node.

[63] proposes an edge simulated framework approach to allocate sufficient resources to network function virtualization (NFV) enabled nodes on time to support latency-sensitive multimedia applications by guaranteeing latency requirements under changing workload and reduced operational cost. In [64], authors propose a strategy to minimize task completion time with an efficient task scheduling and resource management strategy. A fog computing supported software-defined embedded system is considered where task images lay in the storage server while computations can be conducted on either embedded device or a computation server. [65] proposes a scheduling method to address the multi-workflow scheduling problem with proximity constraint in the EC environment. The primary objective is to reduce expenses. The investigation was conducted using both experimental and real-world datasets. [66] concentrated on dynamic resource allocation in an edge context for IoT systems. Finally, a reinforcement learning has been applied in this study to establish a policy for allocating resources.

There is also some literature focusing on the different applications' types that are supported by edge- cloud system [67]. [68] focuses on the optimal application deployment in term of resource-constrained distributed edges. It has a goal to determine the deployment issues of microservice-based applications in edge networks. The application deployment cost has been minimized based on a new strategy was, while considering resource and performance constraints. [69] proposes an approach taking into consideration the type of application based on allocating the offloading tasks to suitable resources in the edge-cloud system to overcome the issue of computational and communication delay. [45] proposes a strategy of selecting the edge node as a target based on application specification with an eye to reduce power consumption and delay. In [70] authors propose a two-stage approach for identifying the ideal solution for Distributed Stream Processing (DSP) applications in EC. This strategy aims to tackle the joint operator scaling and identify the issue. The goal is to provide increased cost efficiency while taking into account user-specified QoS limitations. The study was conducted using real-world DSP test scenarios as the basis for experiments. The study referenced by [71] examines data-intensive applications that are implemented using numerous service components at the edge. The approach was developed using the Genetic Algorithm (DSEGA). Five algorithms were given to get an ideal deployment plan. Most recently, there has been a growing body of research on edge intelligence. [72] discusses how AI solutions can contribute to address EC problems.

An increasing amount of research is currently centered on utilizing user equipment (UE) as resources and the connectivity among user equipment to process tasks. Therefore, the end-user stratum and edge stratum have been combined. Collaboration is a key aspect in literature. The authors in [73] suggest a concept in which mobile devices collaborate to create "mobile clouds" for doing tasks within a local network. "Transient clouds" were proposed in [74] as a collaborative computing platform in which neighboring devices create an ad-hoc network to offer different capabilities as cloud services. The study in [75] introduced a resource sharing technique that leverages mobile devices via opportunistic interactions to highlight resource aspects of mobile cloud computing. A different publication, in [76], concentrates on VM technology to utilize the whole potential of local hardware at the peripheries of the Internet. Alternatively, an adaptive resource discovery

strategy is suggested by [77] to handle service provisioning in opportunistic computing settings, managing increased requests without generating instability as discussed in [78]. The architecture proposed by [79], known as Vehicular Fog Computing (VFC) for vehicular applications, utilizes vehicles as edge infrastructure.

Caching at the edge: Edge servers have a key advantage of bringing storage capability from the source to the edge. The literature provides various caching strategies at the edge based on popularity or re-play possibility of multimedia streams. In [80], authors propose video segment caching at the network through users' announcements. [81] proposes an edge-based server caching scheme considering context-aware RAN condition, video popularity and the Channel Quality Index (CQI) informed by the UE. These works basically utilize only the edge server the users are associated with. There are also a number of cooperative solutions that utilize more than one edge server. [82] proposes cooperative mobility-aware hierarchical caching among edge servers. [83] presents an optimization for DASH-based QoE-aware edge caching placement through coordination among edges. [84] proposes a joint QoE traffic optimization with collaborative edge caching through adjustments of backhaul and fronthaul traffic ratio parameters. Although an important portion of time, energy saving and backhaul traffic reduction can be achieved via caching, the proposed solutions do not utilize any caching strategy considering popularity or replay possibility of any of the segments.

Pre-fetching at the edge involves many methods for pre-fetching streams on either the user equipment (UE) or the network side. Pre-fetching methods on the UE-side involve social networking suggestions [85], downloading multiple segments/chunks simultaneously using multi-path Transmission Control Protocol (TCP) citeE38 and a recommendation system that forecasts videos likely to be seen by users [86]. User-side pre-fetching techniques may not adequately account for the entire journey from the RAN to the cloud, thereby leading to suboptimal TCP performance. On the network-side, pre-fetching methods involve using pre-fetching proxies at access points (APs) [87] or cellular BSs [88] with specific pre-defined policies, such as segment-based pre-fetching. Another approach is pre-fetching segments from the cloud to the MEC server to stay ahead of the user's request progress [89]. However, these studies do not take into account collaborative pre-fetching among network entities/edges. Therefore, implementing pre-fetching on just one network edge may not function effectively and could potentially harm the QoE in crowded networks because of potential buffer overflows.

Prospects and challenges of information-centric networking (ICN) architectures: In recent years, information-centric networking (ICN) architectures, in particular, Named Data networking (NDN), have emerged to mitigate traffic congestion and prioritize critical messages in both wired and wireless networks [90,91]. NDN has advantages over traditional IP networks due to its delay-tolerant and content-centric features and can deliver traffic over a disrupted and highly dynamic network environment [92]. However, when using NDN, there is a significant overhead and depending on traffic congestion and application type, throughput can be up to 30% lower compared to TCP/IP data [93]. Yet, it is still useful architecture to prioritize critical traffic over the network, especially the multimedia flows. For example, in a network scenario where there is TCP, background and multimedia traffic on the same line, multimedia can be prioritized making use of the NDN. There is a growing body of work that applies the NDN concept to EC due to its intrinsic qualities. Most significantly, NDN allows for computation to be executed remotely by any node in the network that supports the required compute capabilities making the client and the network agnostic of the list of providers and their corresponding IP mapping [94]. However, there are a number of challenges in an NDN design driven communication including enabling compute reuse, mobility management, and ensuring secure and trustworthy task offloading [94].

Reputation-based trust in NDN: Commonly, cryptographic signatures and certificates are utilized to establish trust, also known as credential-based [95] or policy-based [96] trust. Alternatively, a "reputation-based trust" approach has been put forth to mitigate cache- and PIT-centric attacks, and strengthen credential-based NDN security schemes. In a reputation based approach, agents decide whether to trust another agent or not by relying on evidence of past behavior supplied by trusted sources. However, the research focus on reputation-based schemes has been primarily limited to mitigating privacy and security attacks and other use-cases have not been thoroughly explored [97]. This proposal leverages a reputation-based scheme for secure, reliable and low latency routing and distributing computation and storage.

3. Edge computing fundamentals

Building on the historical context provided in the previous section, we now delve into the core concepts and principles of EC. Here, we explain the architectural frameworks, key components, and operational mechanisms that underpin EC systems.

3.1. Resource management

Resource management involves the allocation of appropriate and sufficient edge resources (such as CPU, memory and Input/Output) to meet the needs of edge applications. It also aims to optimize the utilization of available resources pool. A proficient management plan of resource in cloud computing enables the cloud provider to effectively and flexibly oversee their information technology resources, thereby becoming a crucial component of a prosperous cloud business. Resource management is a crucial aspect of EC, since it facilitates the integration of several dynamic, diverse, and scattered edge nodes. Resource management encompasses several interconnected stages, which involve the creation and distribution of a resource pool, the monitoring and proactive supply of resources, and ultimately, the allocation of resources to meet future demands (see Fig. 5). Among the stage of resources management, we can find:

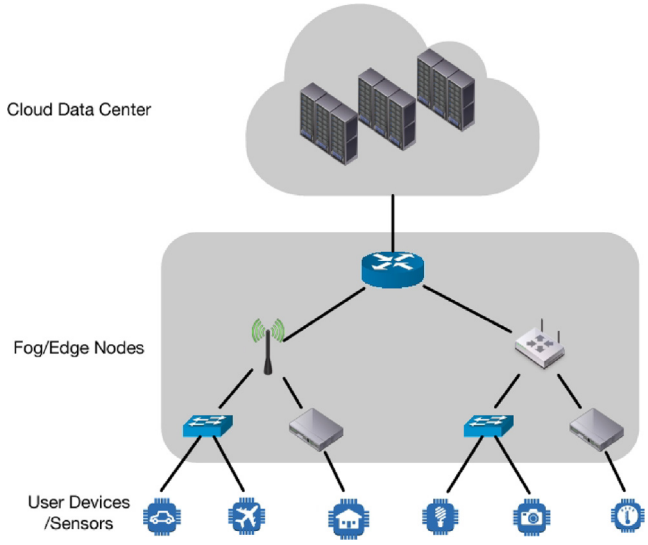


Fig. 5. Resource management.

- Allocation and scheduling of resources,
- Transfer of resources,
- Resources provisioning,
- Pooling of resources.

3.2. Computation offloading

This task is a subfield of computer science that focuses on the decision of whether to execute a task on a local device or transmit it to an external server for processing. The popularity of computational offloading increased with the emergence of mobile cloud computing [98]. A disabled mobile device necessitates additional resources to operate advanced programs such as Google Assistant or Apple Siri. Consequently, those tasks involving voice recognition are transferred to the Cloud. However, with the increasing interest in EC, the issue of offloading has become more prominent and has also evolved in different ways. Through EC, offloading occurs in both vertical and horizontal directions, encompassing transfers between IoT devices, edge servers, and destination servers within the continuum of mist-fog-cloud. EC categorizes offloading directions into four types:

- End-device-to-End-device,
- End-device-to-Cloud,
- End-device-to-Edge-to-Cloud,
- Vertical and horizontal.

3.3. Data management

Management of data includes the processes of obtaining, retaining, disseminating, and utilizing data. Data management facilitates the coordination of edge nodes in order to optimize their overall efficiency and minimize the time it takes to process and exchange data, enabling them to make informed judgments and take appropriate actions. Nevertheless, management of data in EC is inherently more intricate compared to cloud computing. The challenge arises from the extensive quantity of diverse and widely scattered edge nodes that can conform to any geographic topology distribution. Among data aspects that should be taken in consideration, we can find:

- Data dissemination,
- Content caching,
- Data streaming,
- Data placement,
- Data analytics,
- ...

3.4. Network management

The management of network encompasses the progress achieved in network infrastructure and topologies, which adjust the network settings to align with the evolving computer paradigm. The network management function primarily concerns itself with the monitoring, analysis, and dynamic adjustment of the network's state as required. In order for EC to achieve success, it is crucial that it facilitates robust and economical network management techniques, encompassing traffic engineering, access control and the integration of cutting-edge network technology. Publications [99,100] are prominent sources that have extensively examined the EC networking aspects. The subsequent sections will examine significant technologies of network that have aided EC, including network abstraction (SDN, NFV), RANs (C-RAN, F-RAN), and radio resource allocation.

3.5. Security and privacy

Due to the distinct nature of the network edge environment, EC security concerns are seen as one of its most significant obstacles. Data in cloud computing is housed in numerous expansive data centers that are meticulously safeguarded by robust physical security measures, including the presence of numerous guards, fortified fences, and stringent security regulations. Cloud providers are making significant investments in cybersecurity in addition to their focus on physical security. On the other hand, the circumstances in EC are distinct. The physical edge devices are widely spread out and varied, which increases their susceptibility to physical attacks. An example of such attacks is the injection of additional thermal load on the cooling systems, known as cooling system attacks. In addition, the enormous volume of data being offloaded and circulated at the network edge increases the susceptibility of edge servers to cyber vulnerabilities [101,102]. Due to the proximity of data storage to end-users, EC offers superior privacy protection compared to Cloud Computing.

3.6. EC pricing and billing

The pricing and billing of EC services are significant considerations for every Edge Service Provider (ESP). Within the EC market, there are four primary participants: clients, which can be either people or enterprises, ISPs (Internet Service

Providers), clouds (cloud providers) and ESP. These players are interrelated and have the ability to directly or indirectly influence each other [103]. This section will explore the many approaches used to determine the pricing of EC services.

3.7. Practical implications and potential applications of EC

EC is revolutionizing the way data is processed and managed, bringing computation and storage closer to the data source [19,100,104]. This paradigm shift has profound practical implications across various industries, enabling a host of innovative applications that were previously infeasible with traditional centralized cloud computing. Here, we explore some practical implications and potential applications of EC.

EC plays a crucial role in the development and deployment of autonomous vehicles [100]. By processing data locally on the vehicle or at nearby edge nodes, EC reduces latency, ensuring real-time decision-making capabilities. This enables vehicles to respond swiftly to dynamic road conditions, improving safety and efficiency. In addition, EC supports the vast network of IoT devices that monitor and manage urban infrastructure [104]. Applications such as traffic management, environmental monitoring, and smart grid operations benefit from EC's ability to process data close to the source, facilitating real-time analytics and timely responses to changing conditions.

The healthcare sector stands to gain significantly from EC through applications such as remote patient monitoring and telemedicine [105]. EC enables the real-time analysis of health data from wearable devices, allowing for timely interventions and personalized treatment plans. Additionally, EC can support advanced medical imaging and diagnostics by providing rapid processing capabilities at the point of care. Furthermore, EC enhances industrial automation by enabling real-time monitoring and control of manufacturing processes [106]. By processing data at the edge, factories can achieve higher levels of efficiency and productivity. Predictive maintenance, powered by EC, allows for the early detection of equipment failures, reducing downtime and maintenance costs. In addition, the immersive experiences offered by AR and VR applications require low latency and high computational power. EC meets these demands by processing data close to the user, reducing latency and providing smooth, responsive interactions [107]. This is critical for applications in gaming, training, and remote collaboration.

EC can optimize energy management systems by processing data from distributed energy resources (DERs) such as solar panels and wind turbines [108]. By enabling real-time monitoring and control, EC facilitates efficient energy distribution and usage, contributing to the sustainability and resilience of energy grids. Moreover, EC enhances security and surveillance systems by enabling real-time video analytics and anomaly detection [109]. By processing video feeds locally, EC reduces the bandwidth required to transmit data to central servers and allows for faster response times in identifying and addressing security threats. In addition, telecommunications providers can leverage EC to improve network performance and reliability. EC enables edge caching and localized content delivery, reducing latency and enhancing the user experience for services such as video streaming and online gaming [16,19].

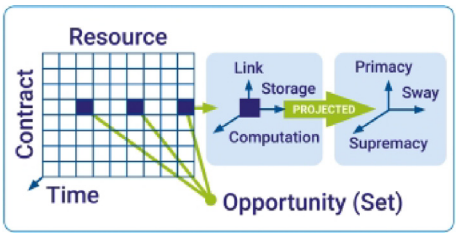


Fig. 6. Definition of Resource.

Finally, EC's practical implications extend across a wide range of industries, driving innovation and efficiency in applications that require real-time data processing and low-latency responses. By bringing computation closer to the data source, EC is poised to transform how we interact with technology, paving the way for new and enhanced services that improve quality of life and operational efficiency. As EC continues to evolve, its potential applications will expand, unlocking further opportunities for growth and innovation across various sectors.

4. Resource allocation

This section will provide a detailed explanation of resource allocation, namely in terms of link, capacity and computation. To ensure optimal user experience (QoE), resources should be provided promptly and in real-time. Extensive research has been conducted on this matter. The following concepts on resource allocation will be presented, along with illustrative case studies.

Within the research, the resource allocation problem in EC/MEC is commonly denoted as the task offloading, task scheduling and joint resource allocation problem (see Fig. 6). The task offloading problem involves determining the optimal decisions for what, how much and where to offload to edge servers or cloud [110]. On the other hand, the task scheduling problem involves choosing the appropriate computational resources, such as servers, that will be utilized to complete each task. The task allocation problem [111] involves the process of ranking and mapping each task to the most suitable computation resource. Furthermore, the term "task scheduling problem" is commonly has a relationship with joint resource allocation problem [112]. The term "joint resource allocation" commonly denotes the integration of offloading and scheduling task difficulties to efficiently offload and schedule an application's jobs to the most suitable cloud/edge servers [113]. Task offloading in MEC is a crucial aspect that involves the distribution of computational tasks between mobile devices and edge servers [110,113]. The goal is to optimize resource utilization, reduce latency, and enhance the overall performance of mobile applications. In addition, MEC extends cloud computing capabilities to the edge of the network, closer to mobile devices. To reach this goal task Offloading involves deciding whether to execute a task locally on the mobile device or offload it

to an edge server. Furthermore, the decision to offload a task depends on factors like task characteristics, device capabilities and network conditions. However, offloading decisions may need to be dynamic, considering the varying workload and network conditions. For instance, let $x_i \in \{0, 1\}$ be a binary variable indicating whether task i is offloaded ($x_i = 1$) or executed locally ($x_i = 0$). Such as the computation resource can be modeled as following:

$$\sum_{i=1}^{N} x_i \cdot C_{task}(i) \leq C_{device} \tag{1}$$

where $C_{task}(i)$ is the computation resource required for task i and C_{device} is the total computation capacity of the mobile device. It is worth to mention that data size has also a huge importance in this context where:

$$D_{offload}(i) \leq D_{max} \tag{2}$$

where $D_{offload}(i)$ is the data size to be offloaded for task i, and D_{max} is the maximum allowed data size for offloading. Moreover, mobile devices often have limited computational power, storage and energy resources, which must be taken in consideration. Based on that, offloading tasks may lead also to increased data transmission, which leads to affect the network bandwidth. Nevertheless, certain applications require low latency, making it essential to minimize the communication delay between the mobile device and the edge server. The latency could be represented as following:

$$l = \sum_{i=1}^{N} T_{local}(i) + T_{offload}(i) \tag{3}$$

where $T_{local}(i)$ is the local execution time of task i, and $T_{offload}(i)$ is the offloading time for task i. Such as $T_{offload}(i) \leq T_{max}$ where T_{max} is the maximum tolerable latency for offloading task i. Finally, the current network must be evaluated in terms of latency, bandwidth and reliability.

In order to improve and optimize offloading task some strategies have been used [104]. For instance, execute tasks on the mobile device itself is suitable for lightweight tasks with minimal processing requirements. Dividing tasks and executing some parts locally while offloading the remaining to the edge server can be seen as an effective strategy. However, offloading the entire task to the edge server could be a necessity in some cases and will be suitable for resource-intensive tasks. Moreover, energy consumption is also a big challenge and must be optimized by Minimizing energy consumption on mobile devices and edge servers. Energy consumption can be seen as:

$$E = \sum_{i=1}^{N} E_{local}(i) + E_{offload}(i) \tag{4}$$

where $E_{local}(i)$ is the energy consumed when task i is executed locally and $E_{offload}(i)$ is the energy consumed when task i is offloaded.

Addressing the task offloading problem in mobile EC involves a multidimensional optimization approach that considers device capabilities, task characteristics and dynamic network conditions to achieve optimal performance. Researchers and practitioners continue to explore innovative solutions to enhance the efficiency and effectiveness of task offloading in MEC. Table 1 illustrates some studies around improving MEC in term of task offloading.

Task scheduling in EC/MEC is a critical aspect that involves the efficient allocation of computational tasks to edge devices within a network. Unlike traditional cloud computing, where tasks are typically executed in centralized data centers, EC leverages the computational capabilities of devices at the network periphery. This paradigm shift aims to reduce latency, enhance responsiveness and optimize resource utilization by pushing computational tasks closer to the data source [19,31,49,99,104]. However, the decentralized nature of EC introduces unique challenges in task scheduling. The task scheduling problem in EC revolves around determining the optimal assignment of tasks to edge devices, considering various constraints and objectives. One of the primary objectives is to minimize the overall makespan, which represents the completion time of all tasks [134]. This minimization contributes to lower latency and improved responsiveness in edge applications, crucial for real-time and interactive scenarios. Let C_i is the completion time of task i, the minimizing makespan will be minimizing the following formula:

$$\sum_{i=1}^{N} C_i \tag{5}$$

Energy efficiency is another significant consideration in task scheduling for EC. Given the resource constraints and limited energy availability of edge devices, scheduling strategies aim to minimize energy consumption while meeting task deadlines. Optimizing energy efficiency is vital for prolonging the operational lifespan of battery-powered edge devices and reducing the environmental impact of EC infrastructure. Furthermore, task scheduling in EC must address the heterogeneity of edge devices and their varying computational capacities. The scheduling algorithm needs to balance the load across devices, ensuring that resource-intensive tasks are allocated to devices with sufficient capabilities (see Eq. (6)). This approach prevents overloading individual devices and maximizes the overall throughput of the EC system.

$$\sum_{i=1}^{N} Resource_i \cdot x_{ij} \leq Capacity_j, \forall j \tag{6}$$

where the total resource requirements of task i assigned to an edge device j do not exceed its capacity.

Task scheduling also involves addressing the dynamic nature of edge environments. The availability and performance of edge devices may change over time due to factors such as device mobility, network conditions and varying workloads. As a result, scheduling algorithms often incorporate adaptability and resilience features to dynamically adjust task assignments based on real-time conditions. Additionally, considerations such as task dependencies, deadlines, and data locality play a crucial role in the task scheduling problem. Precedence constraints ensure that tasks are executed in the correct order and time window constraints guarantee that tasks are completed

Table 1
Studies around improving task offloading.

Ref	Description
[114]	At the edge layer, service providers (SPs) allocate tasks from user equipment to be handled either at the BS or the cloud center.
[115]	This study has addressed the issues of vertical and horizontal offloading, as well as the problem of workload and capacity optimization. The branch and bound method has been utilized.
[113]	The user equipment tasks are allocated to several edge servers using the benders decomposition technique.
[116]	This research involves the coordinated management of communication and computational resources, as well as the assignment of tasks across many layers. The authors employed a latency minimization method to devise the solution.
[117]	The objective of this study is to reduce the energy consumption of all mobile devices while still meeting the time limit for computation execution using an edge server. The solution was derived from the principles of game theory.
[118]	This work has jointly optimized user association, power control, computing capacity allocation, and placement planning. The use of unmanned aerial vehicles was a crucial aspect of this investigation. The solution relied on the principles of compressive sensing and a search algorithm.
[119]	This study has utilized game theory and the Jacobi method. The objective was to represent the computation offloading problem as a potential game on edge services in the local and adjacent regions.
[120]	Optimize the energy usage of mobile devices by choosing the most efficient collection of computational components to transfer to an edge server. The study employed deep learning techniques.
[121]	This paper has developed an offloading approach and resource allocation based on reinforcement learning. The solution was designed for a vehicular edge server and fixed edge server, taking into account the unpredictable nature of vehicle traffic, dynamic computation requests, and changing communication conditions over time.
[122]	An iterative technique was utilized in this study to maximize the transmission powers, processing speeds, bandwidth, and offloading ratio. This study considers the mobile edge relay server and the user.
[123]	The study utilized game theory to optimize energy consumption and computation time for task processing.
[124]	The study examined the computational capabilities of edge servers and edge gateways, as well as the interference between edge gateways.
[125]	This paper offered a method utilizing Deep learning, namely Long Short-Term Memory (LSTM). The objective of the study is to incorporate AI, local computing, EC and cloud computing.
[126]	Service providers deploy resources to the edge layer, specifically micro data centers, to enable a job scheduling system that is sensitive to delay.
[127]	This work presents a method for orchestrating tasks that is based on a Fuzzy Decision Tree. The system employed reinforcement learning to effectively adjust to the unforeseen variations in the surroundings and to offer enhanced assistance for the diversity of devices.
[128]	This work offered task offloading as a means to lower the energy consumption of mobile devices. The result was derived using a decision tree algorithm.
[129]	This study discusses the implementation of cooperative offloading in balloon networks using Support Vector Machines (SVM).
[130]	The study introduced an innovative method that integrates a Niching Genetic Algorithm (NGA) with a Markov Decision Process (MDP) to improve multisite offloading, effectively tackling issues present in current methodologies.
[131]	The study focused on the difficulties associated with energy usage in the context of mobile cloud computing. The proposal suggests employing a hybrid meta-heuristic algorithm, namely the African Wild Dog Algorithm (AWDA), in conjunction with the cellular learning automata technique, to adopt an environmentally friendly approach.
[46]	The study proposed a new cloudlet selection policy to optimize energy efficiency and reduce latency in offloading computations in mobile cloud computing. The proposed technique takes into account the nature of the application and the time it takes to execute. It shows, through experimental analysis, that jobs such as numerical computations, gaming, sorting, and searching can be transferred to suitable cloudlets.
[132]	The study focused on creating a diverse framework to enhance the energy efficiency of mobile devices by incorporating task offloading to cloud servers. The framework utilized in this study is based on a Hybrid Red Fox Flow Direction-based Ensemble SVM Forest Classifier.

within specified deadlines ($Release_i \leq C_i \leq Deadline_i, \forall i$). Moreover, data-intensive tasks may benefit from being scheduled on edge devices closer to the data source, minimizing data transfer latency.

Task scheduling can be formulated as an optimization problem [104]. Where many studies in EC continually explore innovative scheduling algorithms, optimization techniques and heuristics to address the complexity of the task scheduling problem. These efforts aim to unlock the full potential of EC by delivering low-latency, energy-efficient and responsive applications across diverse and dynamic edge environments. Table 2 shows studies about task scheduling.

Joint resource allocation in the context of EC/MEC is a multifaceted challenge that involves optimizing the allocation of various resources across both mobile devices and edge servers [154–156]. This problem arises from the distributed

Table 2
Task scheduling and joint resource allocation studies.

Ref	Description
[133]	The study proposes a method for scheduling tasks in distributed computing systems. This method uses a Markov decision process and ML to adjust to unpredictable patterns of task arrival and execution. The comparative studies clearly show that the suggested approach outperforms classic techniques in terms of decreased scheduling overhead and enhanced average reaction time, thus emphasizing its robustness in both synthetic and practical contexts.
[134]	The work examines a two-phase learning-based scheduling algorithm for data-intensive activities in cluster-based data grids, utilizing a hierarchical multi-agent system comprising a global broker and local agents. The experimental findings clearly illustrate the algorithm's exceptional performance, particularly in dynamic workloads, when compared to standard scheduling methods, thereby highlighting its capacity to adapt to changing environmental conditions.
[135]	The paper presents a novel reinforcement learning-based algorithm for scheduling multiple DAG workflow applications with diverse priorities in cloud computing. Experimental results with real cloud workflows indicate that the proposed scheme not only resolves fairness issues but also enhances resource utilization by effectively managing multiple DAGs with varying priorities.
[136]	This paper addresses the resource management challenges in large-scale, resource-constrained IoT ecosystems by introducing a deep reinforcement learning mechanism. The proposed approach optimizes resource allocation for dynamic user requests with varying requirements, aiming to minimize overall energy consumption and achieve long-term user satisfaction through reduced response time, efficiently handling fluctuations in demand and demonstrating substantial improvements in energy consumption, response time, and resource utilization in simulation analysis.
[137]	This paper introduces Spear, a novel scheduling framework for modern data parallel frameworks like Apache Spark, aiming to minimize the makespan of complex jobs with task dependencies and heterogeneous resource demands. Inspired by AI, Spear combines Monte Carlo Tree Search (MCTS) with deep reinforcement learning.
[138]	This study presents Deep Learning Smart Scheduling (DSS), an automated solution for scheduling big data jobs in cloud computing. It leverages Deep Reinforcement Learning (DRL) and Long Short-Term Memory (LSTM) to forecast the most suitable VMs for incoming large data tasks.
[139]	This study presents QL-HEFT, a new task scheduling algorithm for cloud computing that integrates Q-learning with the heterogeneous earliest finish time (HEFT) technique to minimize makespan. The algorithm uses the ranku value of HEFT as an immediate reward to enhance learning outcomes. It achieves a shorter makespan than three traditional scheduling algorithms and performs well in terms of average response time in experimental tests.
[140]	This study presents a workflow management system (WMS) that uses machine learning to forecast job durations and failure probabilities. This allows for better scheduling decisions to reduce task failures and enhance resource efficiency. The suggested method involves a revised sequence-to-sequence neural network trained with reinforcement learning. It autonomously improves over time and routinely surpasses existing scheduling algorithms in real-world situations involving complex computational and data-heavy tasks.
[141]	This paper addresses the challenge of optimally scheduling interfering links in a dense wireless network with full frequency reuse. By employing a deep learning approach that bypasses channel estimation and relies on geographic locations, the proposed neural network architecture achieves near-optimal performance for sum-rate maximization and demonstrates competitive and generalizable results in network utility maximization, even for larger deployment areas and varying link densities.
[142]	This paper proposes ISVM-Q, a Q-learning algorithm for task scheduling in wireless sensor networks (WSNs), aiming to optimize both application performance and energy consumption. The algorithm utilizes Improved Support Vector Machine (ISVM) to address the "dimensionality disaster" problem of Q-learning, demonstrating its ability to make nodes perform tasks reasonably in dynamic environments and outperforming classic task scheduling algorithms by achieving better application performance with lower energy consumption while maintaining system stability.
[143]	This paper addresses challenges in multi-objective workflow scheduling in Cloud Computing by applying a deep-Q-network model in a multi-agent reinforcement learning setting. The proposed approach, based on a Markov game model, optimizes multi-workflow completion time and user cost without prior expert knowledge, outperforming traditional methods in terms of scheduling plan optimality in dynamic real-time environments, as demonstrated in extensive case studies with scientific workflow templates and Amazon EC2 cloud.
[144]	This paper introduces DeepSoCS, a Deep Reinforcement Learning (DRL)-based scheduler for System-on-Chip (SoC) systems with heterogeneous chip resources, designed to efficiently handle continuously arriving hierarchical jobs represented by directed acyclic graphs. DeepSoCS outperforms traditional HEFT-like algorithms, known for vulnerability to environmental noise, by learning the optimal task ordering under dynamic changes, demonstrated through extensive simulations showing significantly higher job execution performance and robustness under realistic noise conditions.
[145]	This paper addresses the challenges of real-time task scheduling and security in complex embedded systems, particularly focusing on networked embedded devices. The proposed comprehensive task scheduling algorithm dynamically adapts scheduling methods based on task classification and system load changes, while a dynamic integrity measurement model using a star trust chain structure and FPGA-based dynamic reconfigurable hardware design aims to enhance the security of deep network embedded devices.

(*continued on next page*)

Table 2 (*continued*).

Ref	Description
[146]	This paper introduces RLTS, a task scheduling algorithm based on a deep reinforcement learning architecture, designed to address efficiency challenges in cloud manufacturing environments. Utilizing Deep-Q-Network, the algorithm dynamically schedules tasks with precedence relationships, demonstrating effectiveness in minimizing task execution time and outperforming four heuristic algorithms in simulation experiments.
[147]	This paper addresses data scarcity on edge devices and the heavy computation burden of transfer learning by proposing a Data-driven Cooperative Task Allocation (DCTA) approach. DCTA efficiently solves the NP-complete Knapsack problem associated with Multi-task Transfer Learning (MTL).
[148]	This research proposes a Q-learning-based task scheduling framework, QEEC, for energy-efficient cloud computing, addressing the concern of high energy consumption in complex cloud data centers. QEEC utilizes a two-phase approach, incorporating a centralized task dispatcher with an M/M/S queueing model for user request assignment and a Q-learning-based scheduler on each server to prioritize tasks based on task laxity and lifetime, demonstrating reduced average task response time and superior energy efficiency compared to other scheduling policies in simulation experiments.
[149]	This article proposes a two-phase scheduling algorithm for IoT, combining fog and cloud computing to enhance efficiency and reduce consumption costs. The algorithm employs deep learning methods, utilizing the Self-Organizing Map (SOM) clustering technique with hierarchical SOM and Autoencoder for feature extraction, demonstrating improved clustering and reduced missed task rates and costs in cloud and fog environments.
[150]	This paper addresses the joint power control and user association problem in millimeter-wave (mmWave) heterogeneous networks (HetNets) using a reinforcement learning framework. The approach employs a decomposed solution based on the topology of mmWave HetNet, inspired by belief propagation (BP) algorithm, and is evaluated using real-world measurements in the 60 GHz band to handle the complex resource allocation challenges posed by mmWave communications.
[151]	This paper proposes a deep reinforcement learning (DRL)-based online resource scheduling framework for large-scale mobile-edge computing (MEC) systems in the context of IoT. The framework optimizes offloading decisions, transmission power, and resource allocation, incorporating a related and regularized stacked autoencoder for data compression, an adaptive simulated annealing approach for action search, and a preserved and prioritized experience replay to train the policy network, achieving near-optimal performance with reduced computational time compared to existing benchmarks.
[152]	This article introduces an intelligent offloading system for vehicular EC, utilizing deep reinforcement learning to address challenges in executing computing-intensive applications on resource-constrained vehicles. The system represents communication and computation states as finite Markov chains. It creates a joint optimization problem for task scheduling and resource allocation to enhance users' QoE. It utilizes a two-sided matching scheme and deep reinforcement learning for efficient scheduling and network resource allocation.
[153]	This article presents a cooperative computation offloading and resource allocation framework for blockchain-enabled Mobile-edge Computing (MEC) systems, addressing security, privacy, cooperative offloading, and dynamic optimization issues. The framework utilizes blockchain technology for security, and a multi-objective function is designed to jointly optimize offloading decisions, power allocation, block size, and block interval, with a developed asynchronous advantage actor-critic-based algorithm demonstrating fast convergence and significant performance improvements in terms of total reward in simulation results.

and dynamic nature of edge environments, where computational resources, network bandwidth, and energy are limited and need to be efficiently managed. Addressing the joint resource allocation problem is crucial for enhancing the overall performance, reliability, and efficiency of edge and MEC systems. In EC, joint resource allocation encompasses the allocation of computational resources, such as CPU and memory, network resources, and storage capacities across edge devices. The objective is to balance the workload, prevent resource bottlenecks, and ensure that tasks are executed efficiently. This requires adaptive algorithms that can dynamically allocate resources based on the changing demands of applications and the availability of resources in the edge network.

In the realm of MEC, the joint resource allocation problem extends to the coordination of resources between edge servers and mobile devices. Mobile devices typically have limited computational power and battery life, making it imperative to offload certain tasks to more capable edge servers. The challenge lies in determining the optimal allocation strategy, considering factors such as task characteristics, network conditions and the energy constraints of mobile devices. Mathematically, the joint resource allocation problem involves formulating optimization objectives and constraints that capture the trade-offs between various resources and system goals. The objective functions may include minimizing energy consumption, reducing task completion time or optimizing resource utilization. Constraints can include limitations on the computation capacity of edge devices, network bandwidth constraints and latency requirements. Moreover, the joint resource allocation problem often requires collaboration between edge devices and edge servers. This collaborative approach involves making decisions collectively to achieve system-wide optimization. Advanced algorithms, including game theory and distributed optimization techniques are employed to coordinate resource allocation decisions across the decentralized edge environment.

Security and privacy considerations also play a significant role in joint resource allocation. Sensitive data may be processed either on the mobile device or at the edge, and careful allocation decisions need to be made to ensure data privacy and compliance with regulatory requirements. Many studies in the field of EC are actively exploring innovative solutions

to the joint resource allocation problem (see Table 2). ML and AI techniques are being leveraged to develop adaptive algorithms that can learn from historical data and make real-time decisions. As EC continues to evolve, addressing the joint resource allocation problem will be pivotal in unlocking the full potential of edge and MEC systems, enabling efficient and responsive applications in diverse and dynamic environments.

4.1. Integrated sensing and communications (ISAC) in EC

ISAC is a promising technology for B5G/6G networks that enables simultaneous wireless communication and environmental sensing. Recent studies have shown that ISAC can significantly enhance MEC by providing real-time context-aware services and optimizing resource allocation [157–159]. For instance, ISAC can be used to dynamically adjust network resources based on real-time sensing data, improving the efficiency and responsiveness of EC systems. ISAC systems leverage the ability of communication signals to perform dual functions-transmitting data and sensing the environment. This dual functionality allows for the collection of contextual information such as user location, mobility patterns, and environmental conditions. Integrating these sensing capabilities with EC can lead to several key benefits [160,161]. ISAC provides real-time environmental and user context, which can be used by edge computing systems to optimize resource allocation and service delivery. For instance, knowing the precise location and movement patterns of users can help in dynamically adjusting computational resources to meet changing demands. In addition, by using sensing data to monitor network conditions and user activities, ISAC can facilitate more efficient management of network resources. This includes optimizing the placement and migration of EC tasks, leading to reduced latency and improved QoE. ISAC can help in conserving energy by adjusting the operational parameters of edge devices based on real-time sensing information. For example, edge servers can enter low-power states during periods of low activity, or adjust their processing power based on the sensed workload.

5. Lack of research

In order to harvest the full potential of an edge-cloud system, an intelligent breakthrough architecture is required to distribute resources efficiently across cloud, edge servers and other nodes (i.e., user-equipment, vehicles, etc.). Two fundamental challenges emerge:

1. Scalability and complexity in heterogenous and dynamic network environments (computation and communication demands): The growing number and variety of devices and applications have different functional requirements and workflows change dynamically. Different traffic characteristics and required performance of applications i.e., delay sensitive multimedia, need to be considered separately, adding to the complexity. Moreover, dynamicity of connected devices, resource patterns and task characteristics of dynamic network environments leads to unpredictable demands of resource allocation and task offloading over nodes. For instance, tasks dynamically change in real-time and each task's procedure may have different execution times while the number of connected devices in a given area also fluctuate, thus impacting the overall workload.
2. Resource management including connecting everything and efficient task offloading: It is a challenge to achieve optimal offloading in an assembled system comprising nodes over a wide geographic distribution with mobility and resource constraints (including computational and storage capacity) [15].

6. Edge computing and 5G/6G networks

Today, EC and 5G/6G networks are emerging technologies that are becoming more prominent and are frequently addressed together because of their complementary nature. Moreover, EC is of utmost importance in the context of 5G and is anticipated to further develop in tandem with the emergence of 6G. The next of this section will provide some studies focused on utilizing 5G/6G technology to enhance EC. Subsequently, this section will demonstrate the utilization of EC in the situation of both 5G and the forthcoming advancements in 6G. Lastly, this section will conclude with an exploration of how EC is bolstered by the advancements of 5G and potentially 6G networks.

The authors of [162] aim to showcase the practicability of integrating 5G technology, cloud computing and IoT networks. To accomplish this, they utilize a research test-bed to assess the functionality of the system by executing on an edge cloud platform a time-sensitive application. The deployment of a LuMaMi test-bed enabled the creation of a 5G wireless network, which played a vital role in the system. This network efficiently established connections and seamlessly integrated multiple computer nodes. The computation operations in this network were executed using four separate types of nodes: Intel Core i7 desktop for the edge nodes, Raspberry Pi 3Bs to simulate the plant, Intel Xeon VM for AWS and Intel Core i7 VM for the ERDC. The two VMs belonging to Amazon Web Services (AWS) and the Engineering Research and Development Center (ERDC) were connected through a virtual private network (VPN) to the subnet, facilitating direct communication between the computing nodes. The mechanism of ball and beam was employed as supervision method in the test-bed research to evaluate the system's performance. This application was considered to be time-sensitive and crucial to the objective. Testing was conducted to evaluate the system's characteristics, general performance, and the continued operational success of the controller. The evaluation findings indicate that the native control loops created on the cloud platform could efficiently function on the edge cloud infrastructure. Moreover, these control loops can be relocated without jeopardizing the benefits provided cloud nodes or edge and without influencing the stability of the system. By implementing improvements in scheduling, synchronizations,

and message passing, the performance of targeted applications can be greatly enhanced. These upgrades will bring about high adaptability, user-friendliness, little variance, and low latency.

The paper [163] introduces a novel framework that aims to offer on-demand communication services, computation and caching inside a 3D space, regardless of location and time. This framework relies on a close integration between flying nodes and traditional terrestrial base stations. This is especially significant for 6G, which will utilize both terrestrial and non-terrestrial platforms (such as satellites and aerial vehicles) together. The goal is to enhance radio access capability and enable the provision of cloud services in three-dimensional space as needed. This will be achieved by incorporating MEC functions on non-terrestrial platforms, such as low-orbit satellites and aerial vehicles. This could enable the expansion of MEC support to network elements in the sky and devices, creating a space-based MEC that facilitates AI-powered, personalized, and decentralized on-demand services. Consequently, the end users ultimately perceive the sensation of being encompassed by a decentralized computer system that can fulfill all requests with minimal delay. In order to do this, it is crucial to coordinate resources through the use of AI algorithms and the deployment of virtualized network functions that can be dynamically dispersed over both terrestrial and non-terrestrial nodes.

EC is employed in both 5G and the projected advancements in 6G. This section will provide an overview of EC, including examples of its uses in the context of 5G and 6G.

- Low Latency Applications: Example: AR and VR applications benefit from low latency, and EC in 5G helps process the data closer to the end-users, reducing latency and providing a smoother experience.
- IoT: Example: IoT devices generate vast amounts of data. EC in 5G allows processing this data closer to the devices, reducing the need to send everything to centralized cloud servers. This is essential for real-time decision-making in IoT applications.
- MEC: Example: MEC in 5G allows for the deployment of applications and services at the edge of the cellular network. This is beneficial for applications like video streaming, online gaming, and other latency-sensitive services.
- Content Delivery: Example: Edge servers in 5G networks can cache and deliver content closer to the end-users, reducing the load on centralized servers and improving the overall speed and efficiency of content delivery.
- Smart Cities: Example: EC in 5G facilitates the implementation of smart city solutions, such as intelligent traffic management, environmental monitoring, and public safety applications, by processing data locally and in real-time.

Essentially, 5G offers numerous benefits to enable a wide range of applications. Transitioning from 5G to 6G will yield further benefits, resulting in improved EC capabilities and increased reliability of applications. The subsequent application of EC in 6G is shown.

While 6G is still in the early stages of development, several trends and expectations suggest how EC might be utilized:

- Ultra-Low Latency Applications: Applications requiring extremely low latency, such as real-time haptic communication or highly responsive autonomous systems, could leverage the anticipated ultra-low latency capabilities of 6G in conjunction with EC.
- Advanced AR and VR: With the expected advancements in 6G, AR and VR experiences could become even more immersive and interactive, supported by powerful EC capabilities at the network edge.
- Distributed AI and ML: EC in 6G networks may support distributed AI and ML models, allowing devices at the edge to process and analyze data locally without relying heavily on centralized cloud resources.
- Enhanced Network Slicing: The concept of network slicing in 6G could become more sophisticated, allowing for finer-grained control over resources and enabling EC environments tailored to specific application requirements.
- Autonomous Systems: EC in 6G can play a crucial role in supporting autonomous systems, such as self-driving vehicles and drones, by enabling real-time decision-making at the edge.
- Advanced IoT Applications: The combination of 6G and EC can further advance IoT applications, supporting a diverse range of connected devices with varying requirements in terms of data processing, communication, and energy efficiency.

Both 5G and the anticipated 6G networks leverage EC to enhance various applications, ranging from low-latency use cases to distributed computing scenarios. As technology continues to evolve, the collaboration between EC and advanced wireless networks is expected to drive innovation across multiple industries.

The following table (Table 3) presents how EC is enhanced by 5G and potentially 6G networks.

6.1. Synergy between edge intelligence and 5G/6G networks

Edge intelligence, an advanced paradigm within edge computing, integrates AI capabilities directly into edge devices and edge servers [164,165]. This enables real-time data processing, decision-making, and learning at the edge of the network, reducing latency and improving efficiency. The convergence of edge intelligence with 5G/6G networks creates a powerful synergy that enhances the performance and capabilities of both technologies. Moreover, the deployment of 5G and the development of 6G networks bring significant advancements in terms of high data rates, low latency, massive connectivity and enhanced reliability. These features create an ideal environment for the implementation of edge intelligence, enabling a wide range of applications and services. The synergy between edge intelligence and 5G/6G networks can be understood through several aspects, such as low latency and real-time processing; massive connectivity and scalability; enhanced data privacy and security. In addition, the synergy between edge

Table 3
How EC is enhanced by 5G and potentially 6G networks.

Feature	5G	6G
Low Latency	5G networks provide significantly lower latency compared to previous generations. This low latency is crucial for real-time applications, and it complements EC by reducing the time it takes for data to travel between the edge devices and the central cloud.	6G is expected to push the boundaries of low latency even further, potentially enabling new applications and use cases that demand ultra-low latency, such as AR, VR, and haptic communication.
High Bandwidth	With higher bandwidth capabilities, 5G networks can handle more data traffic efficiently. This is particularly beneficial for EC applications that involve the processing of large volumes of data generated by devices at the edge.	6G is expected to further increase bandwidth, enabling even more data-intensive applications. This is essential for emerging technologies like IoT, autonomous vehicles, and high-definition content streaming at the edge.
Distributed Computing	EC involves processing data closer to the source, reducing the need to send all data to centralized cloud servers. 5G networks facilitate distributed computing by providing a robust and high-speed communication infrastructure between edge devices and cloud services.	As 6G networks evolve, they may enhance the capabilities of distributed computing even further. This could lead to more powerful edge devices and increased collaboration between edge nodes.
Enhanced MEC	MEC is a key concept in 5G, allowing computation to take place at the edge of the network, closer to the end-users. This reduces latency and improves the overall user experience.	With the advancements expected in 6G, MEC could see improvements in terms of processing capabilities, making it possible to run more sophisticated applications at the edge.
Ubiquitous Connectivity	5G networks aim to provide ubiquitous connectivity, ensuring that edge devices are always connected to high-speed networks. This is crucial for applications that rely on constant communication with the cloud or other edge devices.	The vision for 6G includes even more pervasive connectivity, potentially extending coverage to remote areas and supporting a wide range of devices with diverse requirements.
Network Slicing	Network slicing in 5G allows the creation of virtualized, dedicated network segments for different applications. This enables efficient resource allocation and customization of network services based on specific EC requirements.	Continued advancements in network slicing may provide even more granular control over network resources, catering to the specific needs of diverse EC applications.

intelligence and 5G/6G networks represents a transformative advancement in edge computing. By combining the real-time processing capabilities of edge intelligence with the enhanced connectivity and performance of 5G/6G networks, a wide range of innovative applications can be realized. Addressing the associated challenges will be crucial for unlocking the full potential of this powerful combination.

7. Contractual, reputational, and opportunistic wireless networking

Addressing the challenges identified in the previous sections, we propose a novel scheme designed to enhance the performance and efficiency of EC. This section outlines the key features, implementation details, and expected benefits of the proposed scheme.

Future imaginative applications of 6G and beyond networks depend on extensive computation resources and ultra-low latency communication (human reaction time about 250 ms, AR/VR < 5 ms, $10-100$ GB data generation per entity [166]) that entail costly infrastructure investments [28,29]. Increasing capacities with more of the same approach will not meet exponentially growing demand. EC or pushing computation and data storage closer to end-users has emerged to meet capacity and low latency requirements of emerging applications. However, edge solutions rely on static and predeployed infrastructure, falling short in harnessing the abundant resources at the network's edge. One key issue is the failure of resource discovery in exascale distributed computing systems with near real-time requirements due to fundamental challenges of:

- Complex entropy mapping of resources in heterogenous and dynamic network environments: Discovering resources and services is a challenging task due to the highly unpredictable nature of the edge ecosystem with heterogeneity of resources, diversity of traffic, and varying link conditions [167].
- Lack of orchestration of resources in a dynamic environment of user/resource mobility, on and off resource availability and accessibility: The decentralized structure of EC requires coordination. However, controllers to assure fault tolerance lead to a high overhead [168]. Edge architectures must rely on available local information and decision making based on partial information leading to sub-optimal resource utilization decisions, which negatively impacting users' QoE [167,168]. Mobility of connected devices also leads to some unpredictable requests, such as offloading task and resource allocation [19].

In order to address these challenges we propose Contractual, Reputational, and Opportunistic Wireless Networking (CROWN). CROWN tackles these problems making based on:

- A resource mapping framework to enable all available resources – with respect to link (l), storage (s) and

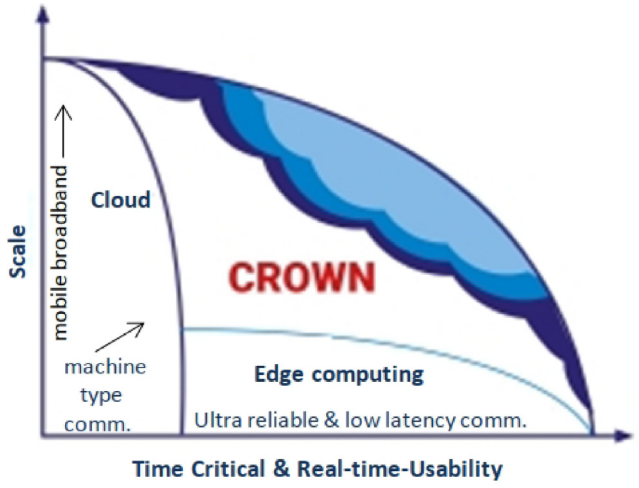


Fig. 7. Impact of CROWN.

computation (c) capacity – to dynamically participate in a large-scale edge ecosystem. Our vision involves the formulating the real-time composition of available edge resources starting with cloud and edge servers and network elements, and eventually end-user devices.

- Logical resource disaggregation in heterogenous compute and storage resources. Rather than aggregating resources into fixed servers, CROWN introduces a flexible system to accommodate varying needs i.e. resource needs vary between jobs and even over time within a single job.

The CROWN framework will play a critical role in realizing the full utilization of edge resources (l, s, c) enabling ultra-low latency ($1000X$ lower task completion time and $<$ 5 millisecond latency) for emerging applications. Our vision it to increase network capacity that will possibly pave the way for free of charge cloud services (see Fig. 7). To enable these advancements, CROWN solves the entropy problem in resource mapping in decentralized networks with near real time optimization providing active orchestration. To achieve this, we will transform the NP-hard problem of resource mapping into a well-known resource allocation optimization problem such as OFDMA (orthogonal frequency-division multiple access). OFDMA is widely used in 4G, 5G and beyond, Wi-Fi 6E as a resource allocation technology to accommodate multiple users at the same time and allocate bandwidth efficiently. The PI is an expert in OFDMA optimization methods [169,170] and will use fundamentals of OFDMA to create OEDMA (orthogonal entropy-division multiple access). PI's experience in leading research and development activities in communication technologies (OFDMA, Wi-Fi, 4G, 5G) in Europe, United States and Turkey; his scientific contributions to the design of PHY/MAC/Routing/Handover/OFDMA/AI protocols ($>$ 36 USPTO patents, $>$ 28 high-impact factors publications, 50+ conferences and three books) will be key to the success of this new scheme. The CROWN research presupposes a multi-disciplinary approach of wireless communication and networking, mathematical modeling, deep learning, classification, prediction, statistical analysis, simulation, and testbed design. Reformulating the resource mapping problem will entail the following steps: We will introduce a three-dimensional resource mapping architecture leveraging NDN. A resource includes contracts (with respect to the desired task/content) and opportunities (grasped with respect to availability of set of nodes). For grasping opportunities, availability is a vector and defined with respect to $\lceil l, s, c \rceil$ capacity per node. We project this vector onto three novel metrics:

- $primacy$ provides the rank when the node will be grasped as a resource,
- $supremacy$ determines with what type of tasks a node will be grasped for quality of experience and,
- $sway$ gives how long a node will be requested to be occupied.

We then develop a practical framework for the above theoretical resource mapping solution introducing three novel schemes — Contractual and Opportunistic Broadcasting (COB), Proof of Reputation (POR) and Proof of Consensus (POC). COB scheme is a proactive protocol that enables each edge unit to broadcast its context information and available resources defined as device capacity, CPU, memory, usage rates, speed, direction, preferences, past and future actions/behaviors, per node. COB uses contracts and requires information about node capability to build the opportunity map. Proof of Reputation (POR) is another parameter in the decision to predict which node forwards the traffic; each node has a reputation score based on historical COB data including tasks it has completed, whether it successfully routed traffic, etc. POR uses reinforcement learning to determine the $primacy$, $supremacy$ and $sway$ metrics (PASS) as seen in Fig. 6. POC mechanism over blockchain provides a consensus and coordination among all nodes for PASS values. Hence, PASS with POR becomes the practical projection of the network capacity (as the Shannon capacity formula provides for subcarriers of OFDMA) and is broadcasted among nodes (e.g. embedded in the NDN Interests as .../PASS/...).

An opportunity will be under the sway of the node in a contract. Based on the above, we then define the resource allocation problem as a OEDMA convex optimization problem (we consider Lagrange dual decomposition method) and further introduce a heuristic and deep learning assisted framework to speed up the process since complexity increases exponentially with the number of resources and contracts. Finally, we will use the time projection function of sway and apply network coding theory to project opportunities over time. The latter will form the basis of our novel pre-computing, pre-fetching and compute re-use systems based on COB and POR predictions to improve network speed, reduce latency and congestion and avoid repeat computations. The system architecture is presented in Fig. 8. The new scheme CROWN is high gain introducing a novel way of solving the complex resource mapping problem by transforming it to a well-known convex optimization problem. The literature fell short of addressing the resource mapping problem due to the nature of exascale distributed computing paradigms with near real-time resource allocation needs. CROWN solves this bottleneck. This is a

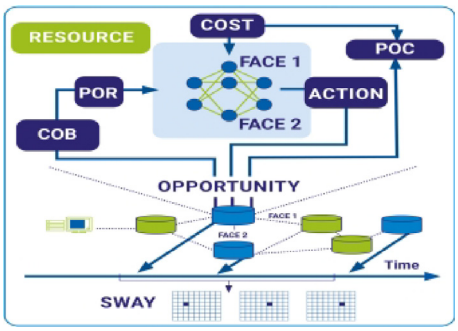


Fig. 8. Overview of the CROWN.

high-risk endeavor and the complexity lies in defining the mathematical foundation of this transformation. The CROWN also has to combine networking paradigms (NDN, Resource Allocation, EC, Network Coding) which have not previously brought together. Our ultimate vision is to transform the cloud into a completely decentralized, federated and ubiquitous environment similar to the historical transformation of the Internet. CROWN has the potential to democratize storage and computation resources by reducing costs tremendously and can be seen as the first step towards cost free cloud services.

8. Discussion of the future EC and MEC

In recent years, the landscape of computing has undergone a transformative shift with the advent of EC and MEC. As these technologies redefine the boundaries of traditional cloud computing, we find ourselves at the forefront of a new era where processing power is distributed closer to the data source, promising unprecedented gains in efficiency, latency reduction and improved user experiences. EC, with its emphasis on localized data processing and reduced latency, has become a cornerstone in the era of IoT and 5G connectivity. MEC, an extension of this paradigm, takes computing capabilities to the every edge of the mobile network, offering a fertile ground for innovation in applications ranging from AR to autonomous systems. However, as we stand at the intersection of theoretical potential and practical implementation, numerous research avenues await exploration. One crucial area that requires focus is the smooth incorporation of EC technologies into current network topologies. Researchers are encouraged to delve into the intricacies of interoperability, standardization and the development of robust frameworks that ensure the smooth coexistence of edge devices in diverse ecosystems.

The future integration of vehicles and mobile devices into EC and MEC is poised to revolutionize various industries, offering enhanced connectivity, real-time processing and novel applications. For instance, Autonomous vehicles generate vast amounts of data that require rapid processing for real-time decision-making. Vehicular Edge Computing (VEC) will facilitate onboard processing of critical data, enabling quicker responses to changing road conditions, traffic patterns and ensuring the safety of passengers and pedestrians. In addition, Integration with EC enables vehicles to communicate with each other and with roadside infrastructure to optimize traffic flow, reduce congestion, and enhance overall transportation efficiency. Furthemore, mobile devices will leverage MEC to offload resource-intensive AR and VR processing tasks to edge servers, ensuring a seamless and immersive user experience without straining the device's hardware. MEC empowers mobile devices to access ultra-low latency applications, such as real-time language translation, gesture recognition and interactive gaming. These applications benefit from the proximity of edge servers, minimizing latency and improving responsiveness.

Mobile devices will act as gateways to the IoT, connecting a myriad of smart devices. MEC facilitates the processing of IoT-generated data at the edge, reducing the need to send vast amounts of data to centralized cloud servers. In addition, vehicles can leverage edge servers to enhance navigation systems. Real-time updates on road conditions, weather, and potential hazards can be processed at the edge, providing drivers with more accurate and timely information. Furthermore, fleet management systems integrated with EC can optimize delivery routes, monitor vehicle health in real time, and enhance overall logistics efficiency. Mobile devices carried by drivers act as nodes in this network, facilitating seamless communication with edge servers.

Essentially, the merging of vehicles and mobile devices with EC and MEC signifies a fundamental change towards a connected, intelligent, and efficient future. The interdependent connection between these technologies is anticipated to generate advancements that enhance safety, elevate user experiences and revolutionize our perception and interaction with our mobile surroundings.

The convergence of EC and 5G/6G networks brings several critical issues that need thorough examination, particularly regarding energy consumption, energy efficiency design and security. As edge devices and servers perform increasingly complex computations locally, their energy consumption rises significantly, necessitating the development of energy-efficient architectures and algorithms to ensure sustainable operations. This includes optimizing resource allocation, implementing power-saving techniques, and leveraging AI to predict and manage energy usage dynamically. Additionally, the integration of edge computing with 5G/6G networks introduces heightened security challenges, such as safeguarding data privacy, preventing cyber-attacks and ensuring secure communication channels. Robust security frameworks must be designed to protect sensitive information processed at the edge, incorporating advanced encryption methods, secure authentication protocols and real-time threat detection mechanisms. Addressing these energy and security concerns is crucial for the successful deployment and operation of EC within the 5G/6G ecosystem, ensuring both efficiency and protection in next-generation networks.

The integration of advanced ML and AI techniques with EC remains a fertile ground for innovation, particularly in improving resource allocation and predictive maintenance. Moreover, addressing security and privacy concerns is paramount, necessitating the development of robust frameworks to safeguard data at the edge. Additionally, energy efficiency continues to

be a significant challenge, highlighting the need for research into low-power edge devices and sustainable architectures. The deployment of EC in 6G networks also presents new challenges and opportunities, such as enhancing network reliability and reducing latency through novel communication protocols and infrastructure. Lastly, the scalability of EC solutions across diverse and large-scale environments requires innovative approaches to ensure seamless operation and management. By focusing on these areas, future research can drive the evolution of edge computing, unlocking its full potential in various applications and industries.

9. Conclusion

EC is a very promising technology that proves valuable in addressing the numerous daily issues encountered by real-time sensitive applications and smart services. This paper provides a thorough examination of EC technology, covering recent research topics. Regarding recent research findings, the authors discovered a lack of comprehensive surveys on various subjects related to EC. These include data security, strategic server positioning, hybrid edge-cloud systems, methods for enhancing edge networks, simulation platforms and the application of EC to 5G/6G networks and the industrial IoT. Therefore, the authors offer the primary accomplishments related to the previously specified study subjects, which indicate the current path in which creative solutions are being created and implemented.

In the foreseeable future, as Internet, networks (5G/6G) and IoT applications continue to advance, EC will assume a crucial role in driving the progress of many industries. The success of the system will rely heavily on its capacity to provide decision support to businesses and effectively utilize the benefits of hybrid edge-cloud architectures. These designs can handle both real-time small-scale intelligent analysis and centralized processing on a large scale.

CRediT authorship contribution statement

Mustafa Ergen: Writing – original draft, Validation, Methodology. **Bilal Saoud:** Writing – original draft, Methodology, Investigation, Conceptualization. **Ibraheem Shayea:** Writing – original draft, Investigation, Formal analysis. **Ayman A. El-Saleh:** Writing – review & editing, Validation, Formal analysis. **Onur Ergen:** Writing – review & editing, Formal analysis. **Feride Inan:** Writing – review & editing, Formal analysis. **Mehmet Fatih Tuysuz:** Writing – review & editing, Validation, Formal analysis.

Acknowledgments

This research has been supported by the Ambeent.ai. Meanwhile, this work is stemming from the following patent: Method and system for distribution of computational and storage capacity using a plurality of moving nodes in different localities: a new decentralized edge architecture. Also, this paper work was partly supported by the Ministry of Higher Education, Research and Innovation (MoHERI) in the Sultanate of Oman under the Block Funding Program with agreement no. MoHERI/BFP/ASU/2022. This research was also partially sponsored by the Internal Research Grant (IRG) Program of A'Sharqiyah University (ASU), with the project no. ASU/IRG/22/23/03.

References

[1] How the next evolution of the internet is changing everything, 2023, https://www.cisco.com/c/dam/en_us/about/ac79/docs/innov/IoT_IBSG_0411FINAL.pdf. (26 December 2023).
[2] M.H. Ghahramani, M. Zhou, C.T. Hon, Toward cloud computing QoS architecture: Analysis of cloud systems and cloud services, IEEE/CAA J. Autom. Sin. 4 (1) (2017) 6–18.
[3] S. Sasikala, S. Prema, Massive centralized cloud computing (MCCC) exploration in higher education, Adv. Comput. Sci. Technol. 3 (2) (2010) 111–118.
[4] B.P. Rimal, D.P. Van, M. Maier, Mobile-edge computing versus centralized cloud computing over a converged FiWi access network, IEEE Trans. Netw. Serv. Manage. 14 (3) (2017) 498–513.
[5] K. Zhang, Y. Mao, S. Leng, Y. He, Y. Zhang, Mobile-edge computing for vehicular networks: A promising network paradigm with predictive offloading, IEEE Veh. Technol. Mag. 12 (2) (2017) 36–44.
[6] Z. Ning, X. Wang, J. Huang, Mobile edge computing-enabled 5G vehicular networks: Toward the integration of communication and computing, IEEE Veh. Technol. Mag. 14 (1) (2019) 54–61.
[7] J. Feng, Z. Liu, C. Wu, Y. Ji, Mobile edge computing for the internet of vehicles: Offloading framework and job scheduling, IEEE Veh. Technol. Mag. 14 (1) (2019) 28–36.
[8] S.Y. Cheung, S.C. Ergen, P. Varaiya, Traffic surveillance with wireless magnetic sensors, in: Proc. 12th ITS World Congr., vol. 1917, 2005, 173181.
[9] S.Y. Cheung, S. Coleri, B. Dundar, S. Ganesh, C.W. Tan, P. Varaiya, Traffic measurement and vehicle classification with a single magnetic sensor, Transp. Res. Rec. 1917 (1) (2005) 173–181.
[10] N. Akhtar, S.C. Ergen, O. Ozkasap, Vehicle mobility and communication channel models for realistic and efficient highway VANET simulation, IEEE Trans. Veh. Technol. 64 (1) (2015) 248–262.
[11] A. Ferdowsi, U. Challita, W. Saad, Deep learning for reliable mobile edge analytics in intelligent transportation systems: An overview, IEEE Veh. Technol. Mag. 14 (1) (2019) 62–70.
[12] I. Tomkos, D. Klonidis, E. Pikasis, S. Theodoridis, Toward the 6G network era: Opportunities and challenges, IT Prof. 22 (1) (2020) 34–38.
[13] L. Zhang, Y.-C. Liang, D. Niyato, 6G visions: Mobile ultrabroadband, super internet-of-things, and artificial intelligence, China Commun. 16 (8) (2019) 1–14.
[14] M. Giordani, M. Polese, M. Mezzavilla, S. Rangan, M. Zorzi, Toward 6G networks: Use cases and technologies, IEEE Commun. Mag. 58 (3) (2020) 55–61.
[15] N. Rosele, K.M. Zaini, N.A. Mustaffa, A. Abrar, S.I. Fadilah, M. Madi, Digital transformation in wireless networks: A comprehensive analysis of mobile data offloading techniques, challenges, and future prospects, J. King Saud Univ.-Comput. Inf. Sci. 36 (5) (2024) 102071.
[16] H. Yao, M. Li, J. Du, P. Zhang, C. Jiang, Z. Han, Artificial intelligence for information-centric networks, IEEE Commun. Mag. 57 (6) (2019) 47–53.
[17] I. Atov, K. Chen, S. Yu, Data science and artificial intelligence for communications, IEEE Commun. Mag. 57 (5) (2019) 56.
[18] C.K. Dominicini, G.L. Vassoler, L.F. Meneses, R.S. Villaca, M.R.N. Ribeiro, M. Martinello, VirtPhy: Fully programmable NFV orchestration architecture for edge data centers, IEEE Trans. Netw. Serv. Manage. 14 (4) (2017) 817–830.
[19] T. Taleb, K. Samdanis, B. Mada, H. Flinck, S. Dutta, D. Sabella, On multi-access edge computing: A survey of the emerging 5G network edge cloud architecture and orchestration, IEEE Commun. Surv. Tuts. 19 (3) (2017) 1657–1681.

[20] M. Marjanovic, A. Antonic, I.P. Zarko, Edge computing architecture for mobile crowdsensing, IEEE Access 6 (2018) 10662–10674.
[21] P.K. Sharma, S. Rathore, Y.-S. Jeong, J.H. Park, SoftEdgeNet: SDN based energy-efficient distributed network architecture for edge computing, IEEE Commun. Mag. 56 (12) (2018) 104–111.
[22] Q. Chen, H. Wang, N. Liu, Integrating networking, storage, and computing for resilient battlefield networks, IEEE Commun. Mag. 57 (8) (2019) 56–63.
[23] Y. Zhou, L. Tian, L. Liu, Y. Qi, Fog computing enabled future mobile communication networks: A convergence of communication and computing, IEEE Commun. Mag. 57 (5) (2019) 20–27.
[24] H. El-Sayed, S. Sankar, M. Prasad, D. Puthal, A. Gupta, M. Mohanty, C.-T. Lin, Edge of things: The big picture on the integration of edge, IoT and the cloud in a distributed computing environment, IEEE Access 6 (2018) 1706–1717.
[25] M. Chiang, T. Zhang, Fog and IoT: An overview of research opportunities, IEEE Internet Things J. 3 (6) (2016) 854–864.
[26] J. Pan, J. McElhannon, Future edge cloud and edge computing for internet of things applications, IEEE Internet Things J. 5 (1) (2018) 439–449.
[27] P. Corcoran, S.K. Datta, Mobile-edge computing and the internet of things for consumers: Extending cloud computing and services to the edge of the network, IEEE Consum. Electron. Mag. 5 (4) (2016) 73–74.
[28] The mobile economy - GSMA, 2023, https://www.gsma.com/mobileeconomy/wp-content/uploads/2020/03/GSMA_MobileEconomy2020_Global.pdf. (Accessed 26 December 2023).
[29] F. Grijpink, A. Ménard, H. Sigurdsson, N. Vucevi, The Road To 5G: The Inevitable Growth of Infrastructure Cost, McKinsey Telecommunications, 2018.
[30] M. Mukherjee, R. Matam, L. Shu, L. Maglaras, M.A. Ferrag, N. Choudhury, V. Kumar, Security and privacy in fog computing: Challenges, IEEE Access 5 (2017) 19293–19304.
[31] J. Zhang, B. Chen, Y. Zhao, X. Cheng, F. Hu, Data security and privacy-preserving in edge computing paradigm: Survey and open issues, IEEE Access 6 (2018) 18209–18237.
[32] Z. Wang, Security and privacy issues within the cloud computing, in: Proc. Int. Conf. Comput. Inf. Sci., 2011, pp. 175–178.
[33] T. Mather, S. Kumaraswamy, S. Latif, Cloud Security and Privacy: An Enterprise Perspective on Risks and Compliance, O'Reilly Media, Sebastopol, CA, USA, 2009.
[34] Y. Xiao, Y. Jia, C. Liu, X. Cheng, J. Yu, W. Lv, Edge computing security: State of the art and challenges, Proc. IEEE 107 (8) (2019) 1608–1631.
[35] H. Zhao, S. Deng, Z. Liu, J. Yin, S. Dustdar, Distributed redundancy scheduling for microservice-based applications at the edge, IEEE Trans. Serv. Comput. (2020) early access.
[36] Q. Fan, N. Ansari, On cost aware cloudlet placement for mobile edge computing, IEEE/CAA J. Autom. Sin. 6 (4) (2019) 926–937.
[37] M. Satyanarayanan, Pervasive computing: Vision and challenges, IEEE Pers. Commun. 8 (4) (2001) 10–17.
[38] P. Zhang, M. Zhou, G. Fortino, Security and trust issues in fog computing: A survey, Future Gener. Comput. Syst. 88 (2018) 16–27.
[39] J. Dilley, B. Maggs, J. Parikh, H. Prokop, R. Sitaraman, B. Weihl, Globally distributed content delivery, IEEE Internet Comput. 6 (5) (2002) 50–58.
[40] M. Pathan, R. Buyya, A. Vakali, Content delivery networks: state of the art, insights, and imperatives, in: Content Delivery Networks, in: Lecture Notes Electrical Engineering, vol. 9, 2008, pp. 3–32.
[41] Z. Ma, M. Yan, R. Wang, S. Wang, Performance analysis of P2p network content delivery based on queueing model, Cluster Comput. 27 (3) (2024) 2901–2915.
[42] Y. Kyung, E. Kim, T. Song, Opportunistic offloading scheme for content delivery service using electro-mobility networks, IET Intell. Transp. Syst. 18 (4) (2024) 591–598.
[43] G.P. Mattia, R. Beraldi, P2PFaaS: A framework for faas peer-to-peer scheduling and load balancing in fog and edge computing, SoftwareX 21 (2023) 101290.
[44] B. Cinar, J.P. Bharadiya, Cloud computing forensics; challenges and future perspectives: A review, Asian J. Res. Comput. Sci. 16 (1) (2023) 1–14.
[45] D.G. Roy, D. De, A. Mukherjee, R. Buyya, Application-aware cloudlet selection for computation offloading in multi-cloudlet environment, J. Supercomput. 73 (4) (2016) 1672–1690.
[46] P. Sen, R. Pandit, D. Sarddar, App-exec cloudlet based computation offloading in mobile cloud computing, in: International Conference on Applied Computational Intelligence and Analytics, ACIA-2022, 2023.
[47] Y.C. Hu, M. Patel, D. Sabella, N. Sprecher, V. Young, Mobile edge computing a key technology towards 5G, in: Eur. Telecommun. Standards Inst, 1st ed., ETSI, Sophia Antipolis, France, White Paper 11, 2015.
[48] Fog Computing and the Internet of Things: Extend the Cloud to where the Things are, Cisco, San Jose, CA, USA, 2015.
[49] C. Mouradian, D. Naboulsi, S. Yangui, R.H. Glitho, M.J. Morrow, P.A. Polakos, A comprehensive survey on fog computing: State-of-the-art and research challenges, IEEE Commun. Surv. Tuts. 20 (1) (2018) 416–464.
[50] Y. Mao, C. You, J. Zhang, K. Huang, K.B. Letaief, A survey on mobile edge computing: The communication perspective, IEEE Commun. Surv. Tuts. 19 (4) (2017) 2322–2358.
[51] F. Bonomi, R. Milito, J. Zhu, S. Addepalli, Fog computing and its role in the internet of things, in: Proc. 1st Ed. MCC Workshop Mobile Cloud Comput. (MCC), 2012, pp. 13–16.
[52] Solutions, Cisco Fog Computing. Unleash the power of the Internet of Things, Cisco Systems Inc, 2015.
[53] S.K. Sharma, X. Wang, Live data analytics with collaborative edge and cloud processing in wireless IoT networks, IEEE Access 5 (2017) (2017) 4621–4635.
[54] X. Masip-Bruin, E. Marn-Tordera, G. Tashakor, A. Jukan, G.J. Ren, Foggy clouds and cloudy fogs: A real need for coordinated management of fog-to-cloud computing systems, IEEE Wirel. Commun. 23 (5) (2016) 120–128.
[55] S.H. Park, O. Simeone, S. Shamai (Shitz), Joint optimization of cloud and edge processing for fog radio access networks, IEEE Trans. Wirel. Commun. 15 (11) (2016) 7621–7632.
[56] M. Chiang, T. Zhang, Fog and IoT: An overview of research opportunities, IEEE Internet Things J. 3 (6) (2016) 854–864.
[57] J. Almutairi, M. Aldossary, Modeling and analyzing offloading strategies of IoT applications over edge computing and joint clouds, Symmetry 13 (3) (2021) 402.
[58] T.Q. Thinh, J. Tang, Q.D. La, T.Q. Quek, Offloading in mobile edge computing: Task allocation and computational frequency scaling, IEEE Trans. Commun. (2017) 1–1.
[59] J. Du, L. Zhao, J. Feng, X. Chu, Computation offloading and resource allocation in mixed fog/cloud computing systems with min-max fairness guarantee, IEEE Trans. Commun. 66 (4) (2018) 1594–1608.
[60] J. Liu, Y. Mao, J. Zhang, K.B. Letaief, Delay-optimal computation task scheduling for mobile-edge computing systems, in: 2016 IEEE International Symposium on Information Theory, ISIT, 2016.
[61] T.G. Rodrigues, K. Suto, H. Nishiyama, N. Kato, Hybrid method for minimizing service delay in edge cloud computing through VM migration and transmission power control, IEEE Trans. Comput. 66 (5) (2017) 810–819.
[62] X. Wei, C. Tang, J. Fan, S. Subramaniam, Joint optimization of energy consumption and delay in cloud-to-thing continuum, IEEE Internet Things J. 6 (2) (2019) 2325–2337.
[63] B. Yang, W.K. Chai, G. Pavlou, K.V. Katsaros, Seamless support of low latency mobile applications with NFV-enabled mobile edge-cloud, in: 2016 5th IEEE International Conference on Cloud Networking, Cloudnet.
[64] D. Zeng, L. Gu, S. Guo, Z. Cheng, S. Yu, Joint optimization of task scheduling and image placement in fog computing supported software-defined embedded system, IEEE Trans. Comput. 65 (12) (2016) 3702–3712.

[65] Y. Ma, J. Zhang, S. Wang, Y. Xia, P. Chen, L. Wu, W. Zheng, A novel approach to cost-efficient scheduling of multi-workflows in the edge computing environment with the proximity constraint, in: Proc. Int. Conf. Algorithms Archit. Parallel Process, Springer, Cham, Switzerland, 2019, pp. 655–668.

[66] S. Deng, Z. Xiang, P. Zhao, J. Taheri, H. Gao, J. Yin, A.Y. Zomaya, Dynamical resource allocation in edge for trustable internet-of-things systems: A reinforcement learning method, IEEE Trans. Ind. Inform. 16 (9) (2020) 6103–6113.

[67] J. Almutairi, M. Aldossary, Modeling and analyzing offloading strategies of IoT applications over edge computing and joint clouds, Symmetry 13 (3) (2021) 402.

[68] S. Deng, Z. Xiang, J. Taheri, K.A. Mohammad, J. Yin, A. Zomaya, S. Dustdar, Optimal application deployment in resource constrained distributed edges, IEEE Trans. Mob. Comput. (2020) early access.

[69] Q. Fan, N. Ansari, Application aware workload allocation for edge computing-based IOT, IEEE Internet Things J. 5 (3) (2018) 2146–2153.

[70] Q. Peng, Y. Xia, Y. Wang, C. Wu, X. Luo, J. Lee, Joint operator scaling and placement for distributed stream processing applications in edge computing, in: Proc. Int. Conf. Service-Oriented Comput., Springer, Cham, Switzerland, 2019, pp. 461–476.

[71] Y. Chen, S. Deng, H. Ma, J. Yin, Deploying data-intensive applications with multiple services components on edge, Mob. Netw. Appl. (2019).

[72] S. Deng, H. Zhao, W. Fang, J. Yin, S. Dustdar, A.Y. Zomaya, Edge intelligence: The confluence of edge computing and artificial intelligence, IEEE Internet Things J. 7 (8) (2020) 7457–7469.

[73] E. Miluzzo, R. Cáceres, Y.-F. Chen, Vision: Mclouds-computing on clouds of mobile devices, in: Proc. 3rd ACM Workshop Mobile Cloud Comput. Services, MCS, 2012, pp. 9–14.

[74] T. Penner, A. Johnson, B. Van Slyke, M. Guirguis, Q. Gu, Transient clouds: Assignment and collaborative execution of tasks on mobile devices, in: Proc. IEEE Global Commun. Conf., 2014, pp. 2801–2806.

[75] W. Liu, R. Shinkuma, T. Takahashi, Opportunistic resource sharing in mobile cloud computing: The single-copy case, in: Proc. 16th Asia-Pacific Netw. Oper. Manage. Symp., 2014, pp. 1–6.

[76] M. Satyanarayanan, S. Smaldone, B. Gilbert, J. Harkes, L. Iftode, Bringing the cloud down to earth: transient pcs everywhere, in: M. Gris, G. Yang (Eds.), Mobile Computing, Applications, and Services, Springer, Berlin, Germany, 2012, pp. 315–322.

[77] W. Liu, T. Nishio, R. Shinkuma, T. Takahashi, Adaptive resource discovery in mobile cloud computing, Comput. Commun. 50 (2014) 119–129.

[78] D. Mascitti, M. Conti, A. Passarella, L. Ricci, Service provisioning through opportunistic computing in mobile clouds, Procedia Comput. Sci. 40 (2014) 143–150.

[79] X. Hou, Y. Li, M. Chen, D. Wu, D. Jin, S. Chen, Vehicular fog computing: A viewpoint of vehicles as the infrastructures, IEEE Trans. Veh. Technol. 65 (6) (2016) 3860–3873.

[80] M. Claeys, N. Bouten, D. DeVleeschauwer, W. VanLeekwijck, S. Latr, F. De Turck, An announcement-based caching approach for video- on-demand streaming, in: 2015 11th International Conference on Network and Service Management, CNSM, 2015, pp. 310–317.

[81] J.O. Fajardo, I. Taboada, F. Liberal, Improving content delivery efficiency through multi-layer mobile edge adaptation, IEEE Netw. 29 (6) (2015) 40–46.

[82] K. Zhang, S. Leng, Y. He, S. Maharjan, Y. Zhang, Cooperative content caching in 5 g networks with mobile edge computing, IEEE Wirel. Commun. 25 (3) (2018) 80–87.

[83] C. Li, L. Toni, J. Zou, H. Xiong, P. Frossard, Qoe-driven mobile edge caching placement for adaptive video streaming, IEEE Trans. Multimed. 20 (4) (2018) 965–984.

[84] A. Mehrabi, M. Siekkinen, A. Yl-Jski, Qoe-traffic optimization through collaborative edge caching in adaptive mobile video streaming, IEEE Access 6 (2018) 52261–52276.

[85] X. Wang, T. Kwon, Y. Choi, H. Wang, J. Liu, Cloud-assisted adaptive video streaming and social-aware video prefetching for mobile users, IEEE Wirel. Commun. 20 (3) (2013) 72–79.

[86] S. Wilk, D. Schreiber, D. Stohr, W. Effelsberg, On the effectiveness of video prefetching relying on recommender systems for mobile devices, in: CCNC, 2016, pp. 429–434.

[87] K. Dong, J. He, W. Song, Qoe-aware adaptive bitrate video streaming over mobile networks with caching proxy, in: ICNC, 2015, pp. 737–741.

[88] V. Krishnamoorthi, N. Carlsson, D. Eager, A. Mahanti, N. Shahmehri, Helping hand or hidden hurdle: Proxy-assisted http-based adap- tive streaming performance, in: Mascot 2013, 2013, pp. 182–191.

[89] C. Ge, N. Wang, G. Foster, M. Wilson, Toward qoe-assured 4k video-on-demand delivery through mobile edge virtualization with adaptive prefetching, IEEE Trans. Multimed. 19 (10) (2017) 2222–2237.

[90] A. Abrar, A.S.C.M. Arif, K.M. Zaini, M.H. Omar, Y. Meng, Advancing producer mobility management in named data networking: A comprehensive analytical model, J. King Saud Univ.-Comput. Inf. Sci. 36 (4) (2024) 102045.

[91] A. Abrar, A.S.C.M. Arif, K.M. Zaini, A systematic analysis and review on producer mobility management in named data networks: Research background and challenges, Alex. Eng. J. 69 (2023) 785–808.

[92] Y. Liu, A. Dowling, L. Huie, Benchmarking network performance in named data networking (NDN), in: 2020 29th Wireless and Optical Communications Conference, WOCC, 2020.

[93] B. Rainer, D. Posch, H. Hellwagner, Investigating the performance of pull-based dynamic adaptive streaming in NDN, IEEE J. Sel. Areas Commun. 34 (8) (2016) 2130–2140.

[94] A. Mtibaa, R. Tourani, S. Misra, J. Burke, L. Zhang, Towards edge computing over named data networking, in: 2018 IEEE International Conference on Edge Computing, EDGE, 2018, pp. 117–120.

[95] V. Shmatikov, C. Talcott, Reputation-based trust management, J. Comput. Secur. 13 (2005) 167–190.

[96] Piero Bonatti, et al., An integration of reputation-based and policy-based trust management, in: W9: The Semantic Web and Policy Workshop (SWPW), vol. 2, (14) 2007.

[97] I.A. Kapetanidou, C.-A. Sarros, V. Tsaoussidis, Reputation-based trust approaches in named data networking, Future Internet 11 (11) (2019) 241, http://dx.doi.org/10.3390/fi11110241.

[98] A.S. Alahmad, H. Kahtan, Y.I. Alzoubi, O. Ali, A. Jaradat, Mobile cloud computing models security issues: A systematic review, J. Netw. Comput. Appl. 190 (2021) 103152.

[99] Y. Mao, C. You, J. Zhang, K. Huang, K.B. Letaief, A survey on mobile edge computing: The communication perspective, IEEE Commun. Surv. Tuts. 19 (4) (2017) 2322–2358.

[100] L. Liu, C. Chen, Q. Pei, S. Maharjan, Y. Zhang, Vehicular edge computing and networking: A survey, Mobile Netw. Appl. 26 (3) (2020) 1145–1168.

[101] S. Forti, G.-L. Ferrari, A. Brogi, Secure cloud–edge deployments, with trust, Future Gener. Comput. Syst. 102 (2020) 775–788.

[102] Z. Shao, M.A. Islam, S. Ren, Heat behind the meter: A hidden threat of thermal attacks in edge colocation data centers, in: Proc. Int. Symp. High-Performance Comput. Archit., 2021, pp. 318–331.

[103] Z. Wang, L. Gao, T. Wang, J. Luo, Monetizing edge service in mobile internet ecosystem, IEEE Trans. Mob. Comput. 21 (5) (2022) 1751–1765.

[104] H. Djigal, J. Xu, L. Liu, Y. Zhang, Machine and deep learning for resource allocation in multi-access edge computing: A survey, IEEE Commun. Surv. Tutor. 24 (4) (2022) 2449–2494.

[105] A.K. Alnaim, A.M. Alwakeel, Machine-learning-based IoT-edge computing healthcare solutions, Electronics 12 (4) (2023) 1027.

[106] B. Babayigit, M. Abubaker, Industrial internet of things: A review of improvements over traditional scada systems for industrial automation, IEEE Syst. J. 18 (1) (2023) 120–133.

[107] A. Toma, S. Samara, M. Qamhieh, Edge computing systems: Modeling and resource optimization for augmented reality and soft real-time applications, J. Netw. Syst. Manage. 31 (4) (2023) 79.

[108] L. Strezoski, Distributed energy resource management systems-DERMS: State of the art and how to move forward, Wiley Interdiscip. Rev.: Energy Environ. 12 (1) (2023) e460.

[109] M. Ali, L. Goyal, C.M. Sharma, S. Kumar, Edge-computing-enabled abnormal activity recognition for visual surveillance, Electronics 13 (2) (2024) 251.

[110] P. Mach, Z. Becvar, Mobile edge computing: A survey on architecture and computation offloading, IEEE Commun. Surv. Tuts. 19 (3) (2017) 1628–1656.

[111] B. Taylor, V.S. Marco, W. Wolff, Y. Elkhatib, Z. Wang, Adaptive deep learning model selection on embedded systems, ACM SIGPLAN Notices 53 (6) (2018) 31–43.

[112] M.A. Rodriguez, R. Buyya, A taxonomy and survey on scheduling algorithms for scientific workflows in iaas cloud computing environments, Concurr. Comput. Pract. Exp. 29 (8) (2017) e4041.

[113] H.A. Alameddine, S. Sharafeddine, S. Sebbah, S. Ayoubi, C. Assi, Dynamic task offloading and scheduling for low-latency IoT services in multi-access edge computing, IEEE J. Sel. Areas Commun. 37 (3) (2019) 668–682.

[114] C. Zhang, H. Du, Q. Ye, C. Liu, H. Yuan, Dmra: A decentralized resource allocation scheme for multi-sp mobile edge computing, in: 2019 IEEE 39th International Conference on Distributed Computing Systems, ICDCS, IEEE, 2019, pp. 390–398.

[115] M.-T. Thai, Y.-D. Lin, Y.-C. Lai, H.-T. Chien, Workload and capacity optimization for cloud–edge computing systems with vertical and horizontal offloading, IEEE Trans. Netw. Serv. Manag. 17 (1) (2019) 227–238.

[116] P. Wang, Z. Zheng, B. Di, L. Song, Hetmec: Latency-optimal task assignment and resource allocation for heterogeneous multi-layer mobile edge computing, IEEE Trans. Wireless Commun. 18 (10) (2019) 4942–4956.

[117] H. Guo, J. Liu, Collaborative computation offloading for multiaccess edge computing over fiber-wireless networks, IEEE Trans. Veh. Technol. 67 (5) (2018) 4514–4526.

[118] Z. Yang, C. Pan, K. Wang, M. Shikh-Bahaei, Energy efficient resource allocation in uav-enabled mobile edge computing networks, IEEE Trans. Wireless Commun. 18 (9) (2019) 4576–4589.

[119] C. Wang, C. Dong, J. Qin, X. Yang, W. Wen, Energy-efficient offloading policy for resource allocation in distributed mobile edge computing, in: 2018 IEEE Symposium on Computers and Communications, ISCC, IEEE, 2018, pp. 00366–00372.

[120] Z. Ali, L. Jiao, T. Baker, G. Abbas, Z.H. Abbas, S. Khaf, A deep learning approach for energy efficient computational offloading in mobile edge computing, IEEE Access 7 (2019) 149623-149633.

[121] Y. Liu, H. Yu, S. Xie, Y. Zhang, Deep reinforcement learning for offloading and resource allocation in vehicle edge computing and networks, IEEE Trans. Veh. Technol. 68 (11) (2019) 11158-11111.

[122] X. Chen, Y. Cai, Q. Shi, M. Zhao, B. Champagne, L. Hanzo, Efficient resource allocation for relay-assisted computation offloading in mobile-edge computing, IEEE Internet Things J. 7 (3) (2019) 2452–2468.

[123] Z. Hong, W. Chen, H. Huang, S. Guo, Z. Zheng, Multi-hop cooperative computation offloading for industrial iot-edge-cloud computing environments, IEEE Trans. Parallel Distrib. Syst. 30 (12) (2019) 2759–2774.

[124] W. Na, S. Jang, Y. Lee, L. Park, N.-N. Dao, S. Cho, Frequency resource allocation and interference management in mobile edge computing for an internet of things system, IEEE Internet Things J. 6 (3) (2018) 4910–4920.

[125] Y. Miao, G. Wu, M. Li, A. Ghoneim, M. Al-Rakhami, M.S. Hossain, Intelligent task prediction and computation offloading based on mobile-edge cloud computing, Future Gener. Comput. Syst. 102 (2020) 925–931.

[126] J. Xu, B. Palanisamy, H. Ludwig, Q. Wang, Zenith: Utility-aware resource allocation for edge computing, in: 2017 IEEE International Conference on Edge Computing, EDGE, IEEE, 2017, pp. 47–54.

[127] C. Mechalikh, H. Taktak, F. Moussa, A fuzzy decision tree based tasks orchestration algorithm for edge computing environments, in: L. Barolli, F. Amato, F. Moscato, T. Enokido, M. Takizawa (Eds.), Advanced Information Networking and Applications, Springer, Cham, Switzerland, 2020, pp. 193–203.

[128] H. Guo, J. Liu, J. Lv, Toward intelligent task offloading at the edge, IEEE Netw. 34 (2) (2020) 128–134.

[129] S. Wang, et al., Federated learning for task and resource allocation in wireless high-altitude balloon networks, IEEE Internet Things J. 8 (24) (2021) 17460–17475.

[130] M.S. Zalat, S.M. Darwish, M.M. Madbouly, An adaptive offloading mechanism for mobile cloud computing: A niching genetic algorithm perspective, IEEE Access 10 (2022) 76752–76765.

[131] A. Almadhor, A. Alharbi, A.M. Alshamrani, W. Alosaimi, H. Alyami, A new offloading method in the green mobile cloud computing based on a hybrid meta-heuristic algorithm, Sustain. Comput.: Inform. Syst. 36 (2022) 100812.

[132] E.V. Subramaniam, V. Krishnasamy, Hybrid optimal ensemble SVM forest classifier for task offloading in mobile cloud computing, Comput. J. (2023).

[133] Z. Tong, Z. Xiao, K. Li, K. Li, Proactive scheduling in distributed computing-a reinforcement learning approach, J. Parallel Distrib. Comput. 74 (7) (2014) 2662–2672.

[134] M.H. Moghadam, S.M. Babamir, Makespan reduction for dynamic workloads in cluster-based data grids using reinforcement-learning based scheduling, J. Comput. Sci. 24 (2018) 402–412.

[135] D. Cui, W. Ke, Z. Peng, J. Zuo, Multiple dags workflow scheduling algorithm based on reinforcement learning in cloud computing, in: Communications in Computer and Information Science, Springer, Singapore, 2016, pp. 305–311.

[136] A. Chowdhury, S.A. Raut, H.S. Narman, DA-DRLS: Drift adaptive deep reinforcement learning based scheduling for IoT resource management, J. Netw. Comput. Appl. 138 (2019) 51–65.

[137] Z. Hu, J. Tu, B. Li, Spear: Optimized dependency-aware task scheduling with deep reinforcement learning, in: Proc. IEEE 39th Int. Conf. Distrib. Comput. Syst., ICDCS, 2019, pp. 2037–2046.

[138] G. Rjoub, J. Bentahar, O. Abdel Wahab, A.S. Bataineh, Deep smart scheduling: A deep learning approach for automated big data scheduling over the cloud, in: Proc. 7th Int. Conf. Future Internet Things Cloud, FiCloud, 2019, pp. 189–196.

[139] Z. Tong, X. Deng, H. Chen, J. Mei, H. Liu, QL-HEFT: A novel machine learning scheduling scheme base on cloud computing environment, Neural Comput. Appl. 32 (2020) 5553–5570.

[140] A.M. Kintsakis, F.E. Psomopoulos, P.A. Mitkas, Reinforcement learning based scheduling in a workflow management system, Eng. Appl. Artif. Intell. 81 (2019) 94–106.

[141] W. Cui, K. Shen, W. Yu, Spatial deep learning for wireless scheduling, IEEE J. Sel. Areas Commun. 37 (6) (2019) 1248–1261.

[142] Z. Wei, F. Liu, Y. Zhang, J. Xu, J. Ji, Z. Lyu, A Q-learning algorithm for task scheduling based on improved SVM in wireless sensor networks, Comput. Netw. 161 (2019) 138–149.

[143] Y. Wang, et al., Multi-objective workflow scheduling with deep-q-network-based multi-agent reinforcement learning, IEEE Access 7 (2019) 39974–39982.

[144] T.T. Sung, J. Ha, J. Kim, A. Yahja, C.-B. Sohn, B. Ryu, DeepSoCS: A neural scheduler for heterogeneous system-on-chip (SoC) resource scheduling, Electronics 9 (6) (2020) 936.

[145] J. Zhou, Real-time task scheduling and network device security for complex embedded systems based on deep learning networks, Microprocess. Microsyst. 79 (2020) 103282.

[146] T. Dong, F. Xue, C. Xiao, J. Li, Task scheduling based on deep reinforcement learning in a cloud manufacturing environment, Concurr. Comput. Pract. Exp. 32 (11) (2020) e5654.

[147] Q. Chen, Z. Zheng, C. Hu, D. Wang, F. Liu, On-edge multi-task transfer learning: Model and practice with data-driven task allocation, IEEE Trans. Parallel Distrib. Syst. 31 (6) (2020) 1357–1371.

[148] D. Ding, X. Fan, Y. Zhao, K. Kang, Q. Yin, J. Zeng, Q-learning based dynamic task scheduling for energy-efficient cloud computing, Future Generat. Comput. Syst. 108 (2020) 361–371.

[149] S. Shadroo, A.M. Rahmani, A. Rezaee, The two-phase scheduling based on deep learning in the internet of things, Comput. Netw. 185 (2021) 107684.
[150] Y. Fan, Z. Zhang, H. Li, Message passing based distributed learning for joint resource allocation in millimeter wave heterogeneous networks, IEEE Trans. Wirel. Commun. 18 (5) (2019) 2872–2885.
[151] F. Jiang, K. Wang, L. Dong, C. Pan, K. Yang, Stacked autoencoder-based deep reinforcement learning for online resource scheduling in large-scale MEC networks, IEEE Internet Things J. 7 (10) (2020) 9278–9290.
[152] Z. Ning, P. Dong, X. Wang, J.J.P.C. Rodrigues, F. Xia, Deep reinforcement learning for vehicular edge computing: An intelligent offloading system, ACM Trans. Intell. Syst. Technol. 10 (6) (2019) 1–24.
[153] J. Feng, F. Richard Yu, Q. Pei, X. Chu, J. Du, L. Zhu, Cooperative computation offloading and resource allocation for blockchain-enabled mobile-edge computing: A deep reinforcement learning approach, IEEE Internet Things J. 7 (7) (2020) 6214–6228.
[154] J. Zhang, W. Xia, F. Yan, L. Shen, Joint computation offloading and resource allocation optimization in heterogeneous networks with mobile edge computing, IEEE Access 6 (2018) 19324–19337.
[155] Q. Luo, S. Hu, C. Li, G. Li, W. Shi, Resource scheduling in edge computing: A survey, IEEE Commun. Surv. Tutor. 23 (4) (2021) 2131–2165.
[156] M. Raeisi-Varzaneh, O. Dakkak, A. Habbal, B.-S. Kim, Resource scheduling in edge computing: Architecture, taxonomy, open issues and future research directions, IEEE Access 11 (2023) 25329–25350.
[157] N. Huang, T. Wang, Y. Wu, Q. Wu, T.Q. Quek, Integrated sensing and communication assisted mobile edge computing: An energy-efficient design via intelligent reflecting surface, IEEE Wirel. Commun. Lett. 11 (10) (2022) 2085–2089.
[158] N. Huang, C. Dou, Y. Wu, L. Qian, R. Lu, Energy-efficient integrated sensing and communication: A multi-access edge computing design, IEEE Wirel. Commun. Lett. (2023).
[159] C. Kaijun, X. Fang, X. Wang, Optimized resource allocation and time partitioning for integrated communication, sensing, and edge computing network, Comput. Commun. 194 (2022) 240–249.
[160] H. Ning, C. Dou, Y. Wu, L. Qian, B. Lin, H. Zhou, X. Shen, Mobile edge computing aided integrated sensing and communication with short-packet transmissions, IEEE Trans. Wireless Commun. (2023).
[161] W. Dingzhu, X. Li, Y. Zhou, Y. Shi, S. Wu, C. Jiang, Integrated sensing-communication-computation for edge artificial intelligence, IEEE Internet Things Mag. 7 (4) (2024) 14–20.
[162] P. Skarin, W. Tärneberg, K.E. Arzen, M. Kihl, Towards mission-critical control at the edge and over 5G, in: Proceedings of the 2018 IEEE International Conference on edge Computing, EDGE, San Francisco, CA, USA, 2018, pp. 50–57.
[163] E.C. Strinati, S. Barbarossa, T. Choi, A. Pietrabissa, A. Giuseppi, E. De Santis, I. Kim, 6G in the sky: on-demand intelligence at the edge of 3D networks, 2020, arXiv arXiv:2010.09463.
[164] Z. Fanqin, W. Li, Y. Yang, L. Feng, P. Yu, M. Zhao, X. Yan, J. Wu, Intelligence-endogenous networks: Innovative network paradigm for 6G, IEEE Wirel. Commun. 29 (1) (2022) 40–47.
[165] C. Saurabh, R. Arya, M.P. Singh, Intelligent resource management in 5G/6G network by adopting edge intelligence for higher education systems, e-Prime-Adv. Electr. Eng., Electron. Energy 8 (2024) 100517.
[166] N. Mohan, L. Corneo, A. Zavodovski, S. Bayhan, W. Wong, J. Kangasharju, Pruning edge research with latency shears, in: HotNets '20: Proceedings of the 19th ACM Workshop on Hot Topics in Networks, 2020, pp. 182–189.
[167] R. Tourani, S. Srikanteswara, S. Misra, R. Chow, L. Yang, X. Liu, Y. Zhang, Democratizing the edge: A pervasive edge computing framework, Open Access (2020).
[168] D. Kondo, T. Ansquer, Y. Tanigawa, H. Tode, Resource discovery for edge computing over named data networking, in: Proceedings of IEEE Annual International Computer Software and Applications Conference, COMPSAC 2021, 2021.
[169] Mustafa Ergen, Mobile broadband: including WiMAX and LTE, Springer Science & Business Media, 2009.
[170] B. A. R. S., B.R. Saltzberg, M. Ergen, Multi-Carrier Digital Communications: Theory and Applications of OFDM, Springer, New York, 2011.