Air Computing: A Survey on a New Generation Computation Paradigm

Baris Yamansavascilar[a,*], Atay Ozgovde[a,*], Cem Ersoy[a,*]

[a]*Department of Computer Engineering, Bogazici University, Istanbul, Turkey*

Abstract

There is an ever-growing race between what novel applications demand from the infrastructure and what the continuous technological breakthroughs bring in. Especially after the proliferation of smart devices and diverse IoT requirements, we observe the dominance of cutting-edge applications with ever-increased user expectations in terms of mobility, pervasiveness, and real-time response. Over the years, to meet the requirements of those applications, cloud computing provides the necessary capacity for computation, while edge computing ensures low latency. However, these two essential solutions would be insufficient for next-generation applications since computational and communicational bottlenecks are inevitable due to the highly dynamic load. On the other hand, inadequate infrastructure considering rural areas and disaster sites makes the utilization of those solutions difficult. Therefore, a 3D networking structure using different air layers including Low Altitude Platforms, High Altitude Platforms, and Low Earth Orbits in a harmonized manner for both urban and rural areas should be applied to satisfy the requirements of the dynamic environment. In this perspective, we put forward a novel, next-generation paradigm called Air Computing that presents a dynamic, responsive, and high-resolution computation environment for all spectrum of applications. In this survey, we define the components of air computing, investigate its architecture in detail, and discuss its essential use cases and the advantages it brings for next-generation application scenarios. We provide a detailed and technical overview of the benefits and challenges of air computing as a novel paradigm and spot the important future research directions.

Keywords:
Air Computing, Edge Computing, Unmanned Aerial Vehicle (UAV), Quality of Service (QoS)

1. Introduction

In order to meet the stringent demands of fully connected, intelligent, and computation-intensive applications with low latency support, vertical networking solutions provide many opportunities [1]. Especially, considering the number of Internet of Things (IoT) connections which are estimated as 14.7 billion in 2023, vertical networking would be crucial for the seamless coverage and dense connection capabilities [2, 3]. Moreover, since the connection of devices that have separate requirements must be processed heterogeneously to ensure Quality of Experience (QoE), reliable computation of different task types is critical in the next-generation networking systems [4].

The diverse requirements of new generation applications are hard to satisfy with well-known practices [5]. Therefore, deployment of Unmanned Aerial Vehicles (UAVs) as Low Altitude Platforms (LAP), airplanes as High Altitude Platforms (HAP), and Low Earth Orbit (LEO) satellites are legitimate candidates for future networks in order to satisfy the requirements of different applications since they can provide low latency, high computation capability, reliability, and availability. These properties are specifically important for the processing of the corresponding tasks of those applications in terms of content caching, resource allocation, task offloading, and extreme mobility.

Although edge computing provides promising results in urban settings in the short-run, reaching genuine ubiquitous execution of real-time, computationally intensive novel applications will require further approaches. Air components with computational processing units in this respect harmonize traditional terrestrial edge computing with a wide range of air technologies to obtain a robust, high-capacity computational infrastructure that embraces urban, suburban, and rural scenarios.

1.1. Air Computing as a New Computational Paradigm

In the literature, the organization and orchestration of the 3D networking structure is called under different names such as aerial communication, Space-Air-Ground Integrated Network (SAGIN), airborne networks, and aerial computing [6, 7, 8, 9]. Especially, the aerial term is widely used to point to the utilization of air components in 3D networking. Since air layers and the corresponding air components would be an essential part of the next-generation networking systems rather than an auxiliary, we call this next-generation computation paradigm as air computing. The architecture of air computing and relationship between different components are shown in Figure 1.

The main advantage of air computing as a new computational paradigm is that it can meet the requirements of dynamically

*Corresponding author.
Email addresses: baris.yamansavascilar@boun.edu.tr (Baris Yamansavascilar), ozgovde@boun.edu.tr (Atay Ozgovde), ersoy@boun.edu.tr (Cem Ersoy)

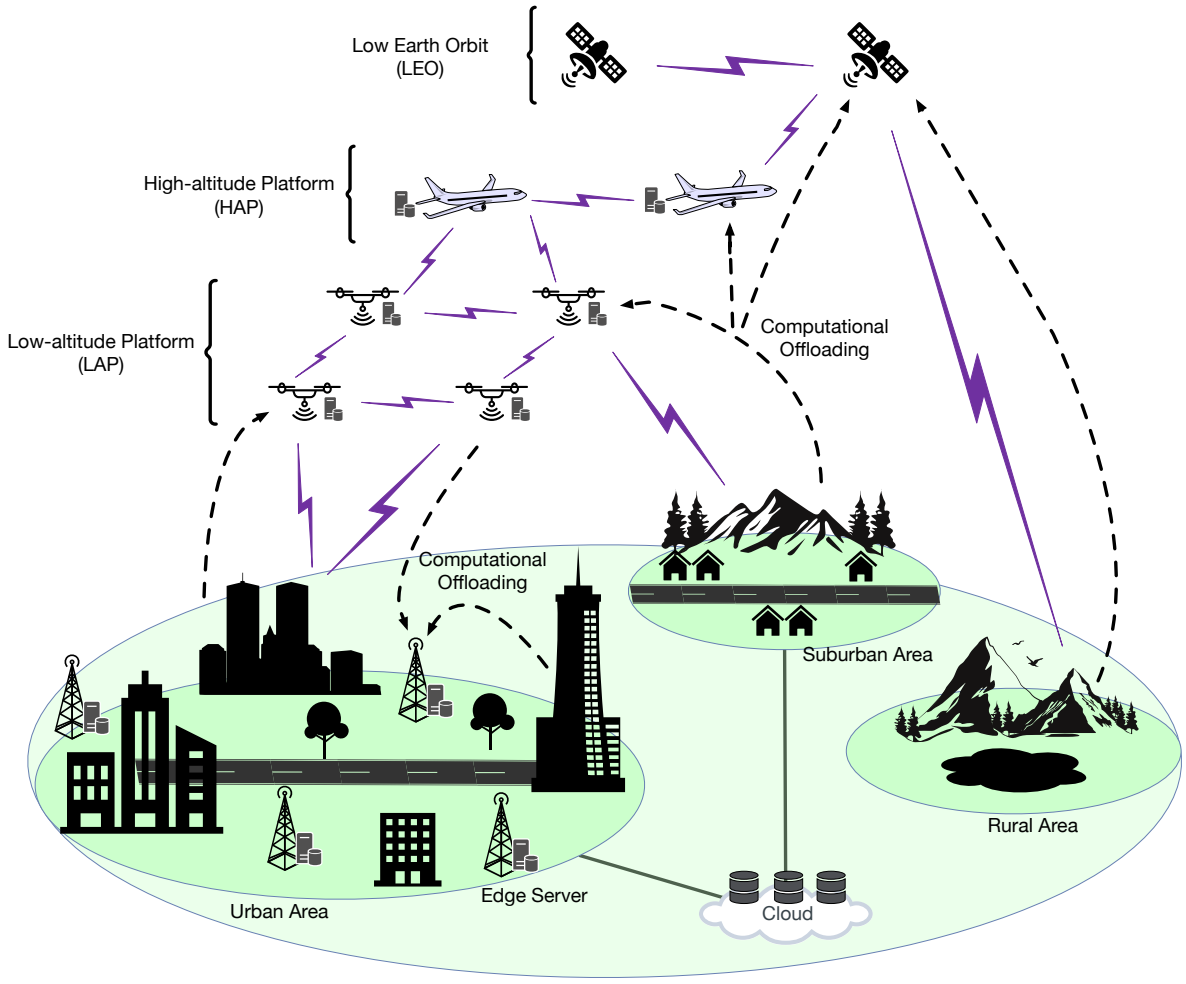


Figure 1: Air Computing Architecture.

changing QoS needs. Since infrastructure-based fixed capacity may not satisfy application requirements during an event, different air platforms can be directed to the corresponding location through air computing. Thus, the QoE of users would not be affected even though the capacity of the infrastructure can be exceeded thanks to the coordination between different components in air computing as shown in Figure 1. This coordination can be through computational offloading, content caching, coverage, and mobility. Moreover, it can be performed in urban, suburban, and rural areas. To this end, we elaborate on different situations, applications, and use cases throughout this study.

1.2. Motivation

A wise deployment of terrestrial servers and air components for the computational needs will change the traditional methods such as edge and cloud computing. Therefore, in this study, we investigate the computing opportunities that would be the result of the intelligent communication between terrestrial servers and air vehicles which we call air computing. These opportunities manifest themselves as an enhancement of QoS and QoE considering both communication requirements and end-user needs [12, 13]. Moreover, as shown in Figure 1, we believe that the coordination between devices and different communication mediums through the air would open new research challenges that will shape the future of the Internet.

Even though there are many use cases for air computing as we investigate throughout this survey, we believe that air computing would be especially useful for dynamic capacity enhancement scenarios considering the battery and energy requirements of air vehicles. Therefore, air computing is beneficial in cases where the service load exceeds the capacity of the fixed infrastructure as we indicated in Section 1.1. This load can be triggered by dynamic events such as sport performances and concerts. Also in the case of a disaster existing infrastructure may be completely destroyed and air vehicles become the only remedy.

We foresee that air computing will be the next generation computation paradigm as a result of the evolution of Multi-Access Edge Computing (MEC) [14]. Since the computation paradigm in MEC is typically restricted with the 2D terrestrial networks in which the resources allocated for the application tasks in either an edge server or cloud server, system bottlenecks can limit meeting the diverse requirements of heterogeneous applications. Since air computing is comprised of different air communication technologies, resource allocation alternatives are considerably increased when compared with the 2D

Table 1: Comparison Between Related Surveys

	Main Theme						
Study	**6G/RAN**	**LAP**	**HAP**	**LEO**	**MEC**	**Architecture and QoE Related Use Cases**	**Contributions**
[10]	✓	✓	✓	✓			Focused on protocols for the design of aerial communication
[8]	✓	✓	✓	✓			Investigated RF wireless technologies for future aerial communications
[11]	✓	✓	✓	✓			Mainly focused on aerial RANs considering 6G access infrastructure
[9]	✓	✓	✓	✓	✓		Mainly focused on RF technologies and examined vertical domain applications
Our study		✓	✓	✓	✓	✓	We mainly focus on detailed scenarios to improve QoS and QoE considering different use-cases of air computing architecture

settings. Vertical networking structure shown in Figure 1 depicts how different applications would use different resources to ensure their QoS requirements.

1.3. Research Scope and Contributions

In this survey, we focus on the computational requirements of the next generation heterogeneous applications and corresponding solutions formulated as the air computing in which terrestrial servers, LAP, HAP, and LEO layers are coordinated intelligently. To put relevant technologies in perspective, we first investigate the differences between edge computing and air computing in terms of the network architecture, challenges, and use cases. Next, we evaluate the advantages of air computing regarding latency, computation capability, storage, mobility, coverage, and reliability.

Since air computing includes both terrestrial servers and air components, we also examine studies on edge computing, UAVs, and other air components in order to show the benefits of air computing more concretely. Furthermore, we detail the possible scenarios in which air computing would be the only valid alternative.

It is important to note that the technical aspects of the Aerial Radio Access Network (ARAN) technologies including 5G and 6G which can be used in air computing are out of the scope of this survey. We only focus on the computation part of the air computing regarding the aforementioned architectural advantages and possible use cases. Regarding our air computing definition, there are four recent survey papers including [10, 8, 11, 9] that investigated the ARAN. In [10], Cao et al. investigated the mechanisms and protocols for airborne communication networks. They detailed LAP-based and HAP-based communication networks regarding their channel models, protocols, and spectrum efficiency. In [8], Baltaci et al. focused on the connectivity requirements and use cases of aerial vehicles considering the challenges of employing wireless communication standards. They introduce the term Future Aerial Communications (FACOM) for aerial connectivity and its use cases. They also examined Radio Frequency (RF) wireless technologies to apply in FACOM. In [11], authors focused on the future network design, system model analysis, and enabling technologies in terms of 6G access infrastructure, transmission propagation, communication latency, and energy consumption. They defined the radio access model as ARAN. On the other hand, the study in [9] is the closest work to our survey paper. However, they mostly focus on 6G and wireless technologies including the frequency spectrum and communication model, while we consider only the computing paradigm along with its advantages. To this end, we investigate thoroughly the paradigm-related scenarios such as energy efficiency, task offloading, and content caching which affect QoS directly, while they did not examine deeply. On the other hand, conceptually, we consider air computing as a next-generation paradigm which is the evolution of edge computing, while they evaluated aerial computing as an amalgamation of ARAN and edge computing. The main differences between those studies and our study are given in Table 1.

The main contributions of this survey are as follows.

- We introduce air computing which is the next-generation computation paradigm. We define its components including terrestrial, LAP, HAP, and LEO layers and investigate them thoroughly.
- We analyze recent studies that focus on edge computing, UAVs, and other air components in the literature and compare them with air computing in order to show its concrete advantages and possible solutions that cannot be offered by traditional networking paradigms.
- We detail the scenarios for air computing that improve the QoS and QoE for end-users. Especially, we focus on several use cases such as natural disasters, real-time video, and outdoor activities. These use cases require seamless connection, intelligent routing, and dynamic capacity enhancement which may not be met by traditional computing schemes regarding Metropolitan Area Network (MAN) and Wide Area Network (WAN).
- We investigate the open research problems and challenges for air computing considering the architecture design, request management, utilization of Artificial Intelligence (AI), a required protocol, energy issues, air regulations, and movement mechanisms. We believe that we point out

important spots in the literature so that readers would expand their studies through one of those areas.

1.4. Methodology

In this study, to survey the literature, we consider the computational features of ground and air resources. Therefore, our fundamental methodology to evaluate the related studies is based on QoS enhancement through the optimization of task offloading, caching, coverage, computational capacity, resource allocation, and energy efficiency schemes. Moreover, we take the deployment and trajectory methods into account for air vehicles which are also crucial in system performance.

Based on our goals for this survey, our inclusion criteria for relevant studies include the publication date of studies, their publisher, corresponding search engines, and keywords. In most cases, we include studies published after 2017. However, there are some exceptions if the number of citations of the corresponding study is high. We mainly evaluated studies published by IEEE, ACM, and Elsevier. Moreover, we used Google Scholar, Scopus, and IEEE Xplore search engines to find relevant studies. In these search engines, we used the following keywords: "edge computing", "task offloading", "resource allocation in edge computing", "UAV-assisted edge computing", "UAV-assisted task offloading", "UAV-assisted resource allocation", "UAV deployment", "UAV trajectory optimization", "High Altitude Platforms", "LEO Satellite-based task offloading", and "UAV Energy Efficiency".

Throughout our survey, we exclude dissertations, theses, and book chapters. Moreover, we do not include studies that focus on 5G, 6G, and medium access control technologies. Furthermore, we exclude studies if there is a newer study with similar goals/methods.

The rest of the paper is organized as follows. In Section II, we introduce air computing considering its advantages, and its differences with edge computing. We show the use cases of air computing in different scenarios in Section III. In Section IV, we investigate edge, LAPs, HAPs, and LEOs which are the main components of air computing. We examine corresponding studies and show the reader that unresolved issues in those papers would be solved by air computing. In Section V, we provide the challenges, opportunities, and future research directions. Finally, we conclude our paper in Section VI. We list the abbreviations used throughout the paper in Table 2.

Table 2: List of abbreviations

Notation	Description
AI	Artificial Intelligence
ARAN	Aerial Radio Access Network
AR	Augmented Reality
CAPEX	Capital Expenditures
DNN	Deep Neural Network
DRL	Deep Reinforcement Learning
FACOM	Future Aerial Communications
FL	Federated Learning
GEO	Geosynchronous Equatorial Orbit
HAP	High Altitude Platform
IoT	Internet of Things
LAN	Local Area Network
LAP	Low Altitude Platform
LEO	Low Earth Orbit
MAN	Metropolitan Area Network
MAR	Mobile Augmented Reality
MDP	Markov Decision Process
MEC	Multi-Access Edge Computing
ML	Machine Learning
NFV	Network Function Virtualization
OPEX	Operating Expenses
QoS	Quality of Service
QoE	Quality of Experience
QoL	Quality of Life
SAGIN	Space-Air-Ground Integrated Network
SDN	Software-Defined Networks
UAV	Unmanned Aerial Vehicle
UE	User Equipment
WAN	Wide Area Network

2. Air Computing

Air computing is a next-generation computational paradigm in which ubiquitous applications with radical networking and computational requirements are satisfied with the help of a family of novel communication opportunities. It provides a highly dynamic, scalable and responsive computational infrastructure in which terrestrial servers are harmonized with various air layers including LAP, HAP, and LEO as shown in Figure 1. Moreover, air computing augments traditional 2D edge computing with a wide spectrum of different computational servers in the air considering a highly dynamic context.

Currently in the urban area, users can enjoy the underlying terrestrial resources to experience seamless connection based on the available infrastructure. With the help of edge computing, mobile devices can reach one of the nearest servers for the execution of their delegated tasks via offloading mechanisms. Despite these advanced approaches, ever-growing application requirements and increasing user mobility patterns push the limits of the fixed infrastructure which led to UAVs being deployed for dynamic capacity enhancement [15, 16]. Adding a vertical dimension to the network greatly enhances the possibilities in terms of the interaction of the users with computational resources. Accordingly, one of the main features of next-generation systems is expected to be their dynamically provisioned 3-Dimensional (3D) structure which leads to many opportunities in terms of QoS and user experience [17]. In a typical edge computing scenario, computational offloading would end either in an edge server or in a cloud server in the 2D terrestrial networks. Cloud servers, although providing seemingly infinite computational capacity, due to latency may be prohibitive or cause low QoE. In that respect, by adding a new dimension using UAVs and other HAP entities, the overall capacity is considerably enhanced and access becomes agile which leads to the server selection process to be more versatile. Since different application types would have instant access to the dynamically arranged computational array of resources in the air, as well as terrestrial ones, this architecture will provide a dramatically increased QoE offerings. Moreover, air computing will also address users or autonomous entities in the air. A user in

an air vehicle can perform computational offloading to other air units and/or terrestrial servers, which manifests itself as a capacity enhancement. As shown in Figure 1, the offloading can be directly from the air vehicle or indirectly via routing over a LAP or a HAP before it is sent to the corresponding server.

A suburban area consists of residential homes and has less population density than the urban area. Hence, the communication infrastructure is not as pervasive as in the urban area which results in fewer resources for the computational needs of the applications. Even though the cloud servers can be reachable via existing infrastructure, the latency would be high for many applications which require low latency for the QoS. Therefore, using LAP and HAP layers as the vertical networking for the suburban areas will increase QoE. The first effect would be on the capacity of the area as new resources can be available for the applications of the users. The second influence would be on the latency since time-critical applications can use the computational sources of LAP vehicles for their corresponding tasks without using the cloud resources that cause high delay. Third, the coverage in the area can be enhanced by placing the UAVs into the corresponding places and the connection would not be interrupted [15, 18].

In the rural area, we assume that there is an environment in which people perform different activities such as sailing, kayaking, climbing, trekking, and camping. The essential fact is that the communication infrastructure may not exist or exist with extremely limited capacity. Moreover, accessing the cloud server is not always possible. In these circumstances, air computing can be used to meet essential QoS requirements since users would utilize fundamental resources for their applications. Even though LAP and HAP platforms are the backbone of the air computing, LEOs are generally used in rural areas since the replacement and energy refilling could be important issues for air vehicles such as UAVs. By deploying LEOs as the computational server or as the relay node that offloads the corresponding tasks to the suitable server, the end-user can enjoy the benefits of the applications also in rural areas.

2.1. Differences Between Air and Edge Computing

Since several application types such as image rendering, video editing, and simulation require intensive computation, cloud computing is proposed as the possible solution regarding CPU and battery constraints of end-devices [19]. However, with the proliferation of versatile devices and corresponding applications known as the Internet of Things (IoT), cloud computing could not meet the low latency requirements [20]. As a result, several computation architectures including Cloudlet [21], MEC [22], and Fog Computing [23] have emerged. In the literature, these architectures are named under the umbrella term edge computing [24] which is also used in this study. The general architecture of edge computing is shown in Figure 2.

The main idea behind the edge computing is that processing the computation-intensive tasks, which cannot be processed in the end-device due to CPU and battery limitation, in the suitable server in the Local Area Network (LAN). Moreover, the concept can be enhanced to Metropolitan Area Network (MAN) due to scarce capacity [20]. In that case, the most suitable server

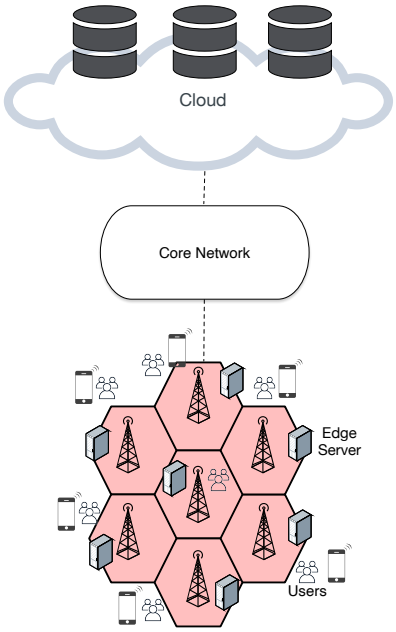


Figure 2: Edge Computing.

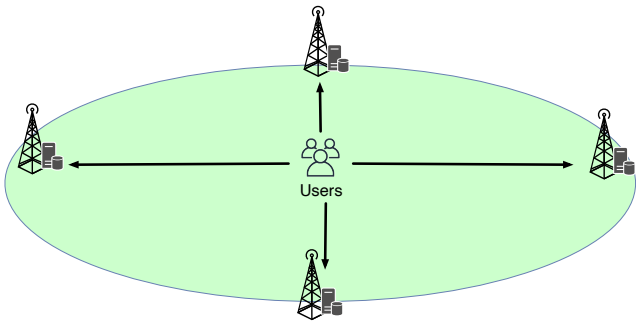
in the MAN is selected for the corresponding application. Furthermore, cloud servers can also be used with edge computing for latency insensitive applications in order to serve more users. In general, edge computing is used for many application domains such as agriculture, healthcare, smart home, robotics, data processing, video analytics, and virtual reality [25, 26].

The capacity of the networks considering 2D terrestrial resources is limited to serve very dense mobile and IoT devices. To solve these issues, air computing provides a third layer which is the air including LAP, HAP, and LEO layer to enhance the 2D computational paradigm into 3D. The 3D structure is also considered as vertical networking and it provides important solutions that cannot be given by traditional edge computing.

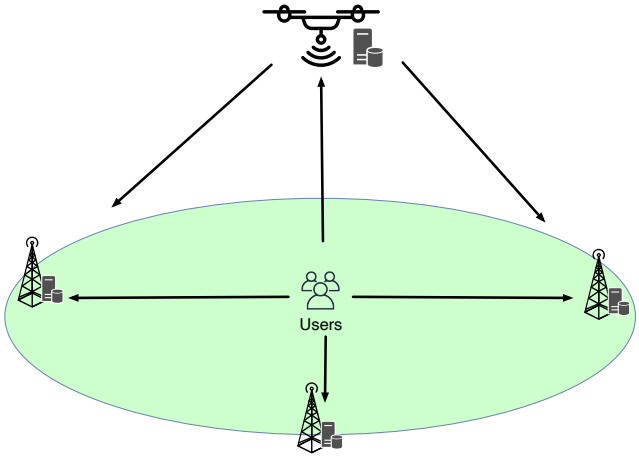
As shown in Figure 3, one of the most important differences between edge and air computing is the direction of the computational task offloading. In edge computing, the direction is horizontal as computational resources are inside the terrestrial 2D area. Moreover, these directions are always one way that is from the user device to the corresponding server. An offloading of a task or a request comes from the user application to the server and then the server process the task in order to give the suitable service. On the other hand, in air computing, the direction of the computational offloading can be both horizontal and vertical. An application would benefit from terrestrial resources and air platforms at the same time. Furthermore, the vertical case of the computational offloading may be in both directions. A typical user in an urban area can use the resources in the air and, conversely, an application in an air vehicle can also benefit from the terrestrial servers. The important differences between edge and air computing are summarized in Table 3.

2.2. Advantages of Air Computing

Air computing has many advantages that can be categorized as *offloading, content caching, latency, computational capabil-*



(a) Horizontal and one-way direction in edge computing



(b) Horizontal, vertical and two-way direction in air computing

Figure 3: Direction of the computational task offloading in air and edge computing.

ity, coverage, and *mobility*. In this subsection, we explain those advantages in detail.

2.2.1. Offloading

The main advantage of air computing regarding offloading is the vertical network opportunities. Unlike traditional edge computing which is based on terrestrial resources, air computing employs a wide variety of computational technologies in air layers each with a different degree of geographical and mobility capabilities. Moreover, air components not only add a new physical dimension to the overall infrastructure but also their ability to be dynamically arranged creates a vision where the system can swiftly adapt itself to the ever-changing conditions of the users, applications, and the network itself. This allows air computing to adequately respond to the full spectrum of application profiles including ones with stringent latency and computational requirements.

We can detail the advantages of air computing for offloading in three different scenarios. In the first scenario, we can assume that a user device in a terrestrial place decides to offload an atomic task of an application which cannot be processed in the device itself. Hence, the task can be offloaded to either terrestrial resources or air components. If terrestrial servers are selected for the offloading, the corresponding procedure would be similar to the edge computing process in which the task is processed on the server and then the results are transmitted to the user. However, in contrast to edge computing, if edge servers in the terrestrial area cannot serve due to their limited capacity, the tasks can be offloaded to air components rather than the cloud servers. This is a crucial advantage of air computing since blockage-free air routes and dynamically provisioned servers would provide lower latency.

Table 3: Important Differences between edge and air computing

Feature	**Edge Computing**	**Air Computing**
Network Architecture	2D terrestrial	3D vertical
Typical Latency	< 10 ms	< 5 ms
Mobility	< 500 km/hr	< 1000 km/hr
Coverage	Cell-based	Cell-less
Offloading Direction	Horizontal one-way	Vertical two-way
Bandwidth	Static	Dynamic

In the second scenario, we can assume that tasks are non-atomic and different parts of the main task can be processed in different resources in an air computing environment as shown in Figure 4. Since coverage is not a problem in vertical networking as in the case of traditional 2D networking, the corresponding partial tasks may be sent to the terrestrial resources in other regional domains if capacity problems occur in the air. As the communication would be in very high bit rates and the air resources would be in the vicinity, the latency may not be an important issue in this situation. Thus, partial offloading can be carried out more efficiently in air computing. On the other hand, if only the terrestrial resources are used for this purpose, this scenario can cause capacity problems regarding the processing if the number of users and their corresponding tasks are high with respect to the resources. Moreover, sending partial tasks to different terrestrial resources may cause congestion in particular parts of the network regarding the link capacity and user density.

The third scenario for task offloading in air computing is that the tasks can be offloaded from the air to the ground. This is crucial since without air computing, users in the air must use the device capabilities, resources of the air vehicle, or relay capabilities of Geosynchronous Equatorial Orbit (GEO) satellites which results in low QoE. Now, through air computing, the tasks can be offloaded to terrestrial sources with low latency and applications can enjoy the benefits of the corresponding resources. Moreover, since air components can cover a large area, different tasks may be offloaded to different terrestrial areas.

2.2.2. Content Caching

Content caching is one of the important practices in order to access the requested pages, tools, and applications with low latency. Therefore, content caching optimization contains three objectives including QoS guarantee, content popularity, and utility maximization [27, 28]. To this end, hit ratio is used as the primary metric to indicate the quality of the content caching optimization. Especially, when the storage capacity of the corresponding servers is insufficient, the quality of the optimization would be more important.

By using vertical networking through air computing, the capacity is enhanced regarding two possible methods. First, stor-

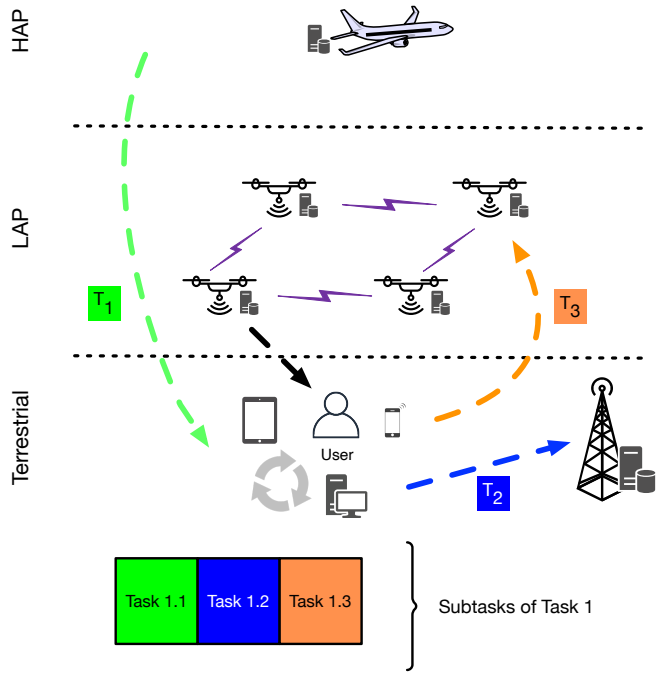


Figure 4: Subtasks of a single task can be processed by different components of an air computing environment.

age of the air vehicles can be used for this purpose. As a result, the capacity can be improved and more content can be reachable by users. In the second method, air components would be used as a relay for the request and the corresponding content can be reached through nearby terrestrial servers. This method can provide lower latency than using WAN.

Even though air computing has important features for content caching in the urban area, its main advantage manifests itself in suburban and rural areas as the communication infrastructure is less developed. Considering the fact that even cloud resources cannot be reachable in rural areas, the importance of air computing would be better recognized. However, since using UAVs would be less efficient as they need corresponding battery charging stations which cannot be found in the rural area, HAPs and LEOs are more suitable to use. As both air platforms can provide content caching, the QoE of users would be enhanced.

2.2.3. *Latency*

Regarding the QoS, one of the most important metrics for a computation paradigm is latency. Since users would like to obtain the corresponding content or result as quickly as possible, providing low latency is critical. Air computing makes use of vertical networking opportunities to provide low latency for specific scenarios. This allows air computing to support diverse application profiles such as remote health, mobile augmented reality, and natural disaster emergency intervention. In traditional terrestrial networking paradigms such as edge computing, the range of the servers is crucial for the latency even though edge servers are located at the LAN or MAN. On the other hand, as air computing allows for the dynamic placement

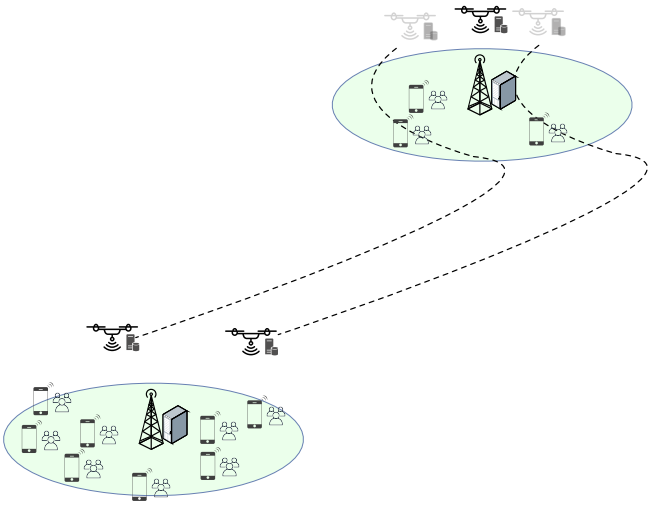


Figure 5: Air components can be replaced in the environment dynamically based on the changing user demands in particular locations.

and provisioning of resources in places needed, typically latency would be independent of the geographical location of the users as shown in Figure 5. This advantage also provides important stability in terms of QoE.

2.2.4. *Computational Capability*

Even though the computational power of the air vehicles is not as powerful as the terrestrial servers, the tasks of applications can be offloaded to multiple sources to enhance the throughput. Moreover, by using both terrestrial servers and air components, a single task may be partitioned to be processed. Since the data rate would be higher and latency would be lower in air computing regarding edge computing, using several sources in different layers for a single task would increase QoE.

2.2.5. *Coverage*

Since air vehicles and their corresponding components are used in air computing, the end devices will not depend on the cell infrastructure in which there is a limited capacity regarding the number of users. As shown in Figure 6, this cell-less structure will provide pervasive connectivity which is crucial for the heterogeneity of future applications.

2.2.6. *Mobility*

As a result of seamless coverage and pervasive connectivity, air computing provides high mobility which is over 1000 km/hr. Moreover, since air components communicate with each other, handover for the processed tasks of users that are in the air or terrestrial vehicle is carried out more smoothly. Furthermore, mobility in air computing can be considered for the users in the air and ground. Accordingly, the processed tasks can also be sent to the terrestrial servers from the air or vice versa.

3. Air Computing Use Cases

The actions that can be taken in air computing are similar to edge computing as they include computation/task offloading, resource allocation, and resource provisioning. However, since air components provide important flexibility regarding vertical networking, the use cases in air computing are more versatile than edge computing. To this end, we elaborate on the potential use cases of air computing in this section.

3.1. Natural Disasters

Natural disasters such as earthquakes, hurricanes, tsunami, and floods wreak havoc on settlements and residential areas. They cause the loss of human lives and the destruction of important resources that people need. Considering communication perspective, there would be two important consequences. First, the communication facilities would be destructed by the natural disaster and as a result people can be deprived of important resources that cause isolation in the disaster site. This is crucial for those people affected by the disaster because they cannot obtain the required aid which must be given to the heavy injury cases and also to humans that expect to be rescued. Without communication resources, the outcome of the disaster would be much worse. Secondly, even in the cases where communication facilities are not seriously damaged, bursty traffic caused by people that want to make an emergency call and to reach their friends, and relatives brings about congestion in the network. Note that this congestion can be also caused by the people outside the disaster site since they also would like to reach the people that are affected by the disaster. Similar to the first case, as a result, people can be isolated in the disaster site so that the outcome of the disaster would be heavier.

Considering both cases, the essential requirement for providing the communication and computation in a disaster site is to enhance the capacity dynamically. This can be carried out by using air components of air computing including UAVs, HAP vehicles, and LEOs [29, 30]. Note that the utilization of those components depends on the disaster type. For example, if the disaster is a hurricane, UAVs and HAP vehicles cannot be used during the disaster. Similarly, if the disaster is an earthquake or a flood, using UAVs would be more effective considering the latency which is a crucial metric for these scenarios.

3.2. Well-being Monitoring

In the event of a medical crisis elderly people who suffer from chronic diseases may require real-time action where high latency would be fatal, especially in rural areas. Therefore, well-being monitoring must be provided such that the latency should not be destructive for the patients. As traditional methods may not ensure low latency, air computing can be used for this purpose [31, 32]. For the urban areas, air computing may be utilized as the complementary resource regarding capacity since the corresponding wireless networks can handle most of the requests. On the other hand, for suburban and rural areas, air computing would be the primary resource for well-being monitoring. By using the coverage, latency, and data rate advantages of air computing, the Quality of Life (QoL) of those patients can be enhanced.

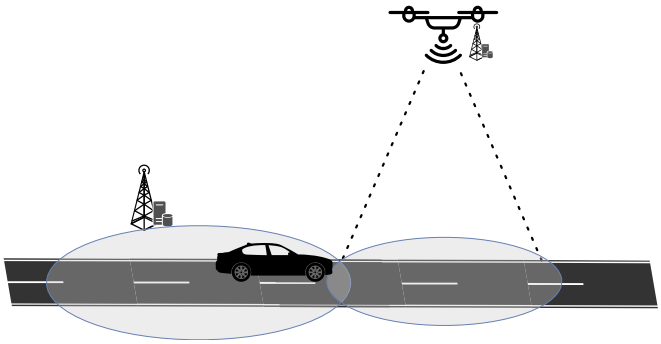


Figure 6: Cell-less structure through air components would provide seamless connection for end-users.

3.3. Remote Health

The issues in remote health are similar to those in well-being monitoring, however they are more critical as medical operations are carried out in this case. The scope of remote health includes the actions of doctors and monitoring the results of those actions on the patients in operations. Moreover, critical life-related metrics such as heart beating, and adrenalin level must be constantly monitored in the operation. To this end, air computing provides important opportunities in this area through its paradigm and corresponding components. For example, if the patient or doctor cannot move from one place to another due to several reasons, the operation can be carried out from a remote area where the doctor leads it.

3.4. Real-time Video

Since real-time video requires a constant bit rate through the lifetime of the video, ensuring the desired QoE is more difficult than the video on demand systems in which the delay can be compensated using dynamically changing buffers. This issue is experienced differently by two main use-cases: (1) real-time conferences/calls such as Zoom and Skype, and (2) watching sports activities.

The fundamental problem in the real-time video for sports events through the Internet is that viewers obtain the content with higher delay regarding terrestrial broadcast. As a result, QoE reduces significantly as viewers may hear the sound of terrestrial broadcast viewers when an important incident has occurred in the event including a football or basketball competition. On the other hand, in conferences and calls, the video can stall or the voice cannot be synchronized with the video due to jitter or congestion in the network.

Air computing can provide a possible solution which proposes a novel architecture for video streaming using air components as shown in Figure 7. In this solution, the video segments are routed via different components including terrestrial servers, UAVs, HAP vehicles, and LEOs based on the current requirements of the network and video. Note that this approach may bring about its own challenges considering scheduling and management of segments and user profiles. However, by using the underlying technology and novel utilization of the air components, the problems above can be solved.

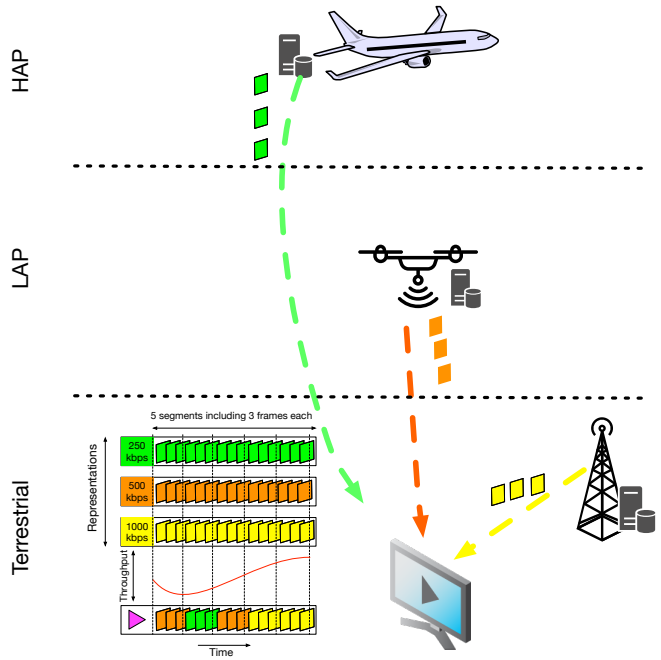


Figure 7: Various representations of the video segments as conveyed by different components of air computing.

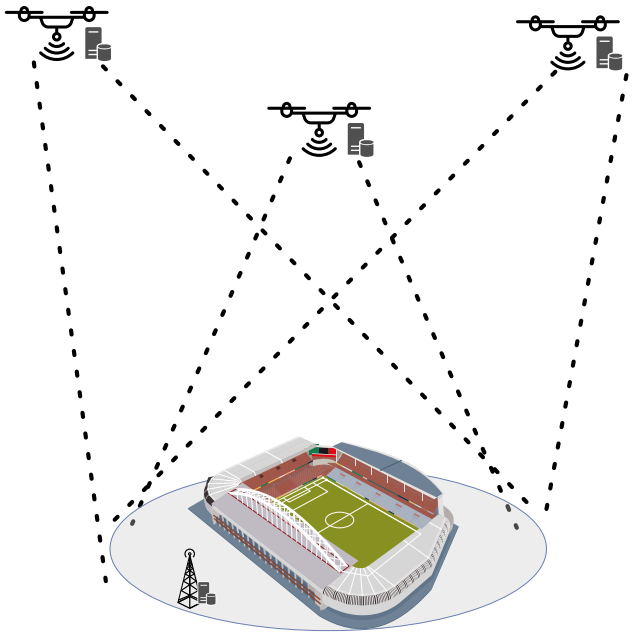


Figure 8: Dynamic network capacity enhancement by air components for organization with intense participation.

3.5. Mobile Augmented Reality

Augmented Reality (AR) in which the real and virtual objects are combined has been used in many application areas including entertainment, healthcare, and education [33]. Moreover, since they perform in real-time, augmented reality applications are delay intolerant in order to provide satisfactory QoE for end-users [34]. Especially, after the widespread usage of smartphones with their expanded capabilities, augmented reality applications are widely deployed. To this end, AR is recently named as Mobile Augmented Reality (MAR).

The most significant difference between AR and MAR is their processing methods. While traditional AR applications perform in-device processing, MAR uses offloading to carry out corresponding computations [35]. The main reason for offloading in MAR is the limited capacity of mobile devices regarding battery and CPU. Therefore, edge and cloud computing have been broadly utilized by MAR applications in recent years. However, since the application requirements of MAR has changed over the years, traditional edge solutions may not be sufficient to provide required latency and capacity.

Air computing can solve the issues related to MAR including scalability, latency, and expected data capacity indicated in [35]. The scalability problem can be handled by multiple components of air computing that can be reachable via seamless connection. Even though users are outside of the urban areas, air components can handle MAR tasks considering latency. Moreover, a huge amount of data can be transferred using the advantages of 3D network structures.

3.6. Sport and Concert Activities

The communication infrastructure in terrestrial areas is built considering fixed resources in which once the facility is deployed, it can be changed with difficulty and significant additional cost. Considering Capital Expenditures (CAPEX) and Operating Expenses (OPEX) of companies, investing on fixed resources was plausible over years. For example, if there are limited resources such as base stations for users that increase in years, improving the capacity which means adding new base stations could be the solution for this problem. However, especially after the proliferation of the versatile application types in smartphones, the fixed infrastructure may not be sufficient for the user needs that change dynamically.

One of the most important examples of this situation is the sport and concert activities where thousands of people rally and use their applications. Statically built infrastructures undergo a significant amount of traffic and computational offloading which may be an order of magnitude higher than the expected requests. Even though network slicing, Network Function Virtualization (NFV), Software-Defined Networks (SDN), and edge computing are used to provide a solution, the issue is still open. On the other hand, by using air components and directing them to the corresponding places where activities occur based on the current network requirements, we can increase the capacity of the network dynamically [36] as shown in Figure 8. Hence, even though the network resources cannot meet the number of application requests, new resources and routes can be created through air computing. As a result, QoS and QoE would be enhanced.

3.7. Outdoor Activities

Since people that perform outdoor activities including sailing, kayaking, climbing, and trekking may not access corresponding resources due to lack of communication infrastructure, they can face isolation in the natural environments. This situation would be crucial considering several issues. In the

Figure 9: Leveraging communication through air computing for outdoor activities.

case of an injury or a health problem such as a heart attack, emergency services must be accessed immediately in order to obtain the aid. Regarding the current networking conditions, the communication and computation from secluded areas including sea, mountain, and forest cannot be provided properly. Thus, even though there would not be such a problem, people carrying out the activities do not feel safe about those possible issues. Apart from the health, daily routines such as social media, communication via chat, and telephone call also cannot be carried out in those conditions. Hence, even though those activities provide important relief for people, their life quality may have deteriorated.

Through air computing, the issues for outdoor activities can be solved as shown in Figure 9. By deploying UAVs, HAP vehicles, and LEOs in suitable places, people in secluded areas can reach important contents and communication infrastructure [37]. We assume that HAP and LEOs would be more useful for such activities as UAVs may face battery reload problems which is similar to the situation in rural areas.

4. Components of Air Computing and Their Applications: Edge, LAP/HAP, LEO

Air computing is made up of a variety of technological elements. In this section, we review individual elements, and discuss their benefits and contributions in the use-cases. Moreover, their open issues are also focused on and how air computing may serve as a remedy is put forward.

4.1. Edge Computing

Even though vertical networking adds another dimension to the current network infrastructure in air computing, the essential element in an urban area would be edge computing as it is deployed widely. As a result, determining a profitable edge strategy is still important [38]. In this section, we investigate the most recent edge computing studies considering resource allocation, task offloading, edge caching, and energy efficiency issues. Moreover, we give an evaluation of selected edge studies in Table 4 and analyze them considering the benefits of the air computing.

4.1.1. Resource Allocation

Along with the computation/task offloading, resource allocation is one of the primary research issues in edge computing

because of the new generation application requirements including transmission bandwidth, latency, energy consumption, and reliability [38, 39]. In [40], authors perform intelligent task execution using Deep Neural Network (DNN) partitioning regarding heterogeneous edge server capacities. They propose a joint method considering cost-effective resource allocation and self-adaptive DNN partition in order to provide collaborative computation between IoT devices and edge servers. Lieang et al. focus on the challenge of handover between base stations considering the management of computation and radio resources [41]. Therefore, the goals in this study consist of maximizing the throughput and minimizing the handover cost. Chen et al. propose a cache-assisted multi-user MEC mechanism considering to cache executive codes of tasks proactively [42]. Their goal was to reduce the task execution delay and energy consumption of users. Xia et al. consider the problems including resource allocation on demand for limited edge servers, and developing heterogeneous task offloading strategies [43]. To this end, they implement an online distributed optimization algorithm based on game theory to perform optimal offloading and energy harvesting decisions. Bahreini et al. develop an auction-based mechanism by addressing the resource allocation and monetization challenges in MEC [44]. They focus on the dynamic provisioning of computing resources since the tasks of users are heterogeneous. On the other hand, Roostaei et al. investigate Stackelberg game based distributed algorithm [45, 46] in order to dynamically allocate and price edge resources [47]. Zhao et al. develop a hybrid system considering beamforming and resource allocation [48]. They benefit from the advantages of mmWave communications indicated in [49] in order to optimize beamforming vectors at users and base stations to minimize the maximum delay. In [50], Wang et al. focus on the challenges of limited capacity of edge servers in a heterogeneous multi-IoT environment. To address this issue, they propose a weighted cost model which is solved by a Deep Reinforcement Learning (DRL)-based algorithm considering dynamic and stochastic edge computing environments.

4.1.2. Task Offloading

Task offloading is the most crucial issue in edge computing regarding QoS of IoT devices and their corresponding applications [20]. The performance of task offloading is generally evaluated with other metrics such as energy efficiency and edge caching hit [56]. Peng et al. used three constrained multiobjective algorithms considering time and energy consumption in order to solve the computation offloading problem in an edge environment [57]. Feng et al. on the other hand focus on different requirements of mission-critical applications regarding their priorities [52]. To this end, they benefit Lyapunov optimization considering the energy consumption of resources [58]. However, they use only a single edge server in their experiments. Chen et al. on the other hand, develop a system that jointly optimizes task assignment and offloading scheduling in order to minimize maximum completion delay [59]. For this system, they also consider different communication and computation capabilities. Xue et al. develop a dynamic incentive mechanism to investigate the problem of the task offloading and resource allocation [53]. They consider a multi-user and multi-vehicle system. They use the Stackelberg game for the interaction between MEC service provider and user equipments (UEs). In [60], authors focus on data caching and computing offloading in a two-tier MEC environment. They consider the constraints of tasks in terms of the delay and the minimization of the network cost at user. However, they do not include mobility for the users and cloud option for offloading. Xu et al. investigate the performance of task offloading in High-Speed Railways (HSRs) considering proper data routing paths for each offloaded task [61]. Since handovers are frequent in HSRs, they focus on how frequent handovers in uplinks and downlinks affect offloading. In [62], Zhang et al. focus on autonomous manufacturing by considering their delay sensitivity. To this end, they propose a risk-aware cloud-edge computing framework by developing a branch-and-check approach for solving the nonlinear programming problem. Yang et al. propose a Machine Learning (ML) solution to solve the offloading problem in MEC [63]. They train and jointly optimize the offloading decisions and resource allocation. Li et al. focus on caching techniques to optimize QoS in MEC [64]. They propose three algorithms to forecast the next executing task. Moreover, they jointly consider cache hit rate and load balance of edge servers.

4.1.3. Energy Efficiency

Since the battery capacity of IoT devices is limited, and service providers would like to lower their expenses regarding power consumption of edge servers, energy efficiency is an important research topic in edge computing. In [54] authors focus on minimizing the energy consumption and task processing delay in this study. For that purpose, they develop an evolutionary algorithm that finds the best trade-offs between energy consumption and processing delay. Song et al. consider minimizing the energy consumption of mobile devices when executing corresponding tasks at satellites [65]. Moreover, their model provides MEC services using LEOs for mobile devices in disaster areas. Chen et al. investigate energy-efficient offloading considering QoS requirements of DNN-based smart IoT systems [55]. To this end, they design a self-adaptive particle swarm optimization algorithm for the corresponding energy-efficient offloading strategy. In [66], authors develop a multi-armed bandit algorithm to provide a solution for the server selection problem in edge computing. They define corresponding reward and cost terms considering the energy and required time in offloading rounds. Zhou et al. focus on energy-efficient service migration in considering MEC-enabled dense cellular networks [67]. They formulate the service migration process as a Mixed-Integer Nonlinear Programming (MINP) problem. They also use the Lyapunov optimization technique to decouple the migration process.

4.2. LAP

Since providing Line of Sight (LoS) links has many advantages in terms of connectivity, service provision, and latency, UAVs have been deployed in many areas especially remote locations. However, this deployment also brings its own issues including mobility management, UAV networking management,

Table 4: Evaluation of Selected Edge Studies

Study	Category	Goal	Solution	Open Issue	Benefits of Air Computing
[40]	Allocation	Providing both computation efficiency and cost effectiveness to accelerate DNN-based task acceleration in the MEC	A joint method by a self-adaptive DNN partition with cost-effective resource allocation	Only a single edge server is used	With multiple components, it can increase the capacity
[41]	Allocation	Maximizing the sum offloading rate, quantifying MEC throughput, and minimizing the migration cost	They relax the corresponding binary variables in the original problem to overcome non-convex issue	Energy efficiency regarding edge servers is not considered	3D network structure would alleviate the handover problem
[42]	Caching	A cache-assisted multi-user MEC mechanism	They formulate a non-linear programming problem which involves a joint optimization	There is no mobility	UAVs can be helpful for caching important contents
[51]	Caching	They investigate the cooperation problem of edge nodes in MEC	They use Lagrangian multipliers and then a distributed optimization algorithm	There is no mobility and handover consideration	Vertical networking would increase capacity for caching based on air components.
[52]	Offloading	A priority-differentiated offloading strategy that considers the stringent QoS requirements of mission-critical services	They use Lyapunov optimization for priority-differentiation	Only a single edge server is used	Mission-critical applications can be handled very well regarding multiple components of air computing.
[53]	Offloading	Task offloading for multi-user and multi-vehicle in vehicular MEC	They propose a dynamic incentive mechanism	Mobility model is not clear	Seamless connectivity can alleviate the problems in vehicular MEC systems
[54]	Energy	Minimizing both the energy consumption and task processing delay of the mobile devices	They propose an evolutionary algorithm that can efficiently find a representative sample of the best trade-offs	There is no mobility and cloud consideration	Multiple air components can alleviate the trade-offs between offloading and energy consumption
[55]	Energy	Energy-efficient offloading for DNN based smart IoT systems	They propose a swarm optimization algorithm for energy-efficient offloading strategy	There is no mobility	With the offloading to different nodes in the air, it may reduce energy consumption

and flight formation [68, 69]. To this end, we categorize and evaluate these issues under trajectory planning, task offloading, placement of UAVs, and energy consumption.

4.2.1. Trajectory Planning

In order to provide efficient on-demand services, trajectory planning is crucial for UAVs [70, 71]. Moreover, optimization of the pre-defined paths based on the dynamic events is also critical for the performance of UAVs in terms of QoS [72]. Zhao et al. investigate a proactive mobility management solution for users' trajectories in order to deploy UAVs dynamically in the network [73]. To this end, they propose a distributed learning framework in which edge servers are considered as local data owners that collect connection data. Wang et al. aim at minimizing the energy consumption of users in the network by considering the resource allocation and trajectory of UAVs [74]. They propose two solutions: (1) convex optimization based trajectory control algorithm to minimize energy consumption, and (2) DRL-based trajectory control algorithm for real-time decisions. Similarly, authors in [75] propose that the trajectory of UAVs can be approximated using traditional convex optimization approaches and discrete variables. Liu et al. optimize UAV trajectories considering energy consumption [76]. They formulated the problem as Markov Decision Process (MDP) and proposed DRL with a double q-network. Wang et al. proposed a multi-UAV communication system for 6G in which they consider UAV trajectories and radio resource scheduling [77].

4.2.2. Task Offloading

One of the most important motivations for the deployment of UAVs is the computation rate maximization of applications by using task offloading [78, 79]. Through task offloading via Line of Sight, the burden on the edge servers would be alleviated and required QoS can be provided. Seid et al. study on minimization of the computation costs in terms of energy consumption and computation delay [80]. They propose a Multi-Agent Deep Reinforcement Learning (MADRL)-based approach in a multi-UAV enabled IoT edge network using a single centralized SDN

controller. In [81], the authors propose an offloading system in which users can perform partial offloading regarding UAVs and MEC servers. They formulate the problem as a maximization problem and use the principles of Prospect Theory [82]. Haber et al. on the other hand focus on mission-critical applications that require ultra-reliable low-latency computation offloading [18]. They use UAVs considering the maximization of served request rate and the optimization of UAVs' positions with the offloading decision. Zhan et al. propose a framework for a multi-UAV enabled MEC system in order to maximize the number of served IoT devices regarding computation offloading and resource allocation [83]. Zhao et al. proposed a collaborative task offloading approach in a multi-UAV multi-MEC system considering energy consumption, and UAV trajectory [84]. For this purpose, they use a cooperative MADRL method in which the policy gradient algorithm is utilized. Diao et al. investigate the usage of UAVs as relay nodes considering emergency conditions [85]. Moreover, they consider energy consumption minimization by optimizing offloading and scheduling. Zeng et al. focus on multi-UAV assisted MEC environment in order to maximize revenue of ground users considering the offloading of multi-user scenarios [86]. To this end, they take different time sensitivity of each user task into account and construct a two-hierarchy Stackelberg game framework model. In this model, UAVs are considered as leaders performing location deployments and users are the followers that perform offloading selections. In [87], Shi et al. propose a model-free DRL offloading scheme based on considering the dynamic channel state, renewable energy utilization, UAVs trajectory, and tasks offloading ratio. To perform this, they use Twin Delayed Deep Deterministic Policy Gradient (TD3) algorithm.

4.2.3. Energy Consumption

Even though energy consumption is considered as a performance metric that is evaluated in studies along with the other issues such as task offloading, and trajectory planning, some studies take energy efficiency into account as the main problem. Ji et al. consider nonorthogonal and orthogonal multiple access modes for a UAV-assisted MEC system and focus on weighted-sum energy consumption [88]. To this end, they propose alternating iterative algorithms in order to optimize UAV trajectory and resource allocation. The goal of Li et al. is to create a model for UAV-assisted MEC by considering energy-efficient UAV trajectory design and optimized computation offloading [89]. They also consider partial offloading in this study. Chen et al. focus on ultra-dense networks considering the resource allocation problem by maximizing energy efficiency [90]. They used UAVs as flying base stations (BS) and utilized DQN as the solution technique. Liu et al. aim to minimize the energy consumption of users by optimizing relay and computing features of UAVs [91]. They use an iterative algorithm to solve the non-convex problem. Li et al. investigate the optimization of energy efficiency in an UAV-assisted network in which UAVs are used for an energy station and MEC server [92]. Their goal is to maximize average energy efficiency in the network by considering user transmit power, user computing frequency, UAV transmit power, bandwidth allocation, and UAV trajectory planning. They use a proximal policy optimization algorithm as a DRL agent.

4.2.4. Placement

The placement of UAVs is substantial as coverage provides connectivity that enhances the network capacity in terms of the data transmission and computation [93]. Moreover, their utilization in poorly covered terrestrial regions would increase end-users QoE [94]. Therefore, placement optimization would be crucial for the performance of the UAV-based network along with the trajectory planning [95, 96].

In [97], Lui et al. use actor-critic methods in DRL in order to provide connectivity between UAVs so that they can cover required areas to improve QoS. Wang et al. optimize the placement of the UAVs considering the offloading decision and resource allocation in a multi-UAV-enabled MEC environment [98]. They propose a two-layer optimization method to solve the problem. On the other hand, Yuan et al. focus on the dynamic placement of UAVs in a vehicular network [99]. They utilize the actor-critic DRL approach to carry out real-time UAV placement. They also consider UAVs' flying range, communicating range, and energy resources. Abdelhakam et al. focus on strong co-channel interference that can be caused by line-of-sight channels between UAVs and the ground terminals [100]. To solve this problem, they propose a Coordinated Multi-Point (CoMP) technique considering the deployment of UAVs in a multi-UAV-assisted IoT network.

4.3. HAP Components and LEO Satellites

Considering the limited battery energy and cell-based coverage of UAVs, HAP and LEO layers provide important advantages regarding long-distance communication, energy consumption, and management opportunities. Moreover, their performance would not be affected by the weather conditions due to their high altitudes as 10 - 30km for HAP components, and 160 - 200km for LEO satellites.

The main goal of the deployment of the HAP components is to provide connectivity such as Internet access, and to perform as a controller node for the UAVs [15, 101]. However, in certain circumstances, such as a congestion in the lower layers, they can be used as an edge computing server or a relay node. Since their coverage is on a regional scale, UAVs that can be considered as dynamic cells can reach the corresponding resources in a particular area via HAP components. Even though this is one of the reasons that the deployment of HAP components is generally in urban and suburban areas, maintenance issues based on energy consumption also cause an important restriction for their deployment in rural areas.

On the other hand, the essential use case of the LEO satellites is to meet the coverage problems of rural areas where infrastructure for the communication is insufficient. They can be deployed for months thanks to their efficient energy consumption, however they are not recoverable after their deployment. Moreover, they cannot be used for low latency applications due to their high propagation delay. Therefore, they can be used as a complementary resource for urban and suburban areas in order

Table 5: The Summary of Main Differences Between Air Layers

Issue	**LAP**	**HAP**	**LEO**
Altitude	Less than 10km	10 - 30km	160 - 2000km
Propagation Delay	10 - 30 μs	50 - 85 μs	1.5 - 3ms
Coverage Scale	Cell	Regional	Continental
Main Deployment	Urban Areas	Urban and Suburban Areas	Rural Areas
Low Latency Apps	It can be provided since underlying wireless communication infrastructure may ensure the corresponding requirements.	It depends on the channel and weather conditions regarding propagation delay.	Because of the high altitude and propagation delay, satellites cannot provide the requirements of low latency applications.
Main Use-cases	Seamless mobility is ensured through UAVs. Moreover, components in this layer provide either edge computing solutions or access to edge servers.	Airplanes and balloons can be used as management nodes for UAVs considering their regional coverage. Furthermore, they can also be used for edge computing purposes.	Satellites can perform edge computing solutions however their service would be limited due to their low on-board capacity. Therefore, they are generally used to access the cloud computing solutions.
Performance	As UAVs can be configurable easily regarding their stationary position, they can use their capacity effectively based on the user density.	Even though their configurability is not flexible as UAVs, balloons and aircrafts may provide a relative stationary position. However, they cannot use their capacity as efficient as UAVs.	Due to their high speeds and their deployment in underpopulated areas, some of the capacity of satellites would be wasted.
Maintenance	Even though UAVs provide important flexibility in dynamic environments, they need charging stations. Moreover, their maintenance would be daily due to their limited battery capacity. They can be reusable multiple times.	Balloons and aircrafts can fly for days based on their fuel capacity. However, they must return to their corresponding bases for maintenance. They can be reusable multiple times.	Satellites are not recoverable after their deployment. However, they can give service for months.
Energy Consumption	The required energy is ensured from batteries. Their energy consumption can be heavily affected by winds and weather conditions especially if they fly against the wind.	The required energy is provided from fuels. The effect of the winds and weather conditions to the energy consumption is limited due to their altitude.	They meet the required energy from the solar power and corresponding batteries. Their energy consumption cannot be affected by weather conditions.

to get access to cloud computing solutions in WAN. Besides, exploiting services in continental or beyond regional distances would be more efficient using LEO satellites since terrestrial nodes may cause high latency [102]. Figure 10 depicts reaching regional and inter-regional resources using HAP and LEO layers.

It is important to note that when the LEO layer is exploited, SAGIN is the term that is generally used by studies to indicate the corresponding communication system [7]. In [103], Tang et al. developed an efficient offloading mechanism for SAGIN. They benefited from the communication between the LEO satellite network, LAP vehicles, and terrestrial resources. To minimize the total delay and to handle the high mobility of nodes, they proposed a deep reinforcement learning traffic offloading approach. Chen et al. [104] benefit from satellite constellations to apply Federated Learning considering communication overhead and privacy issues. They use four different modes including remote cloud learning, onboard satellite learning, federated learning with data sharing, and federated learning with no data sharing to evaluate the performance of their system. On the other hand, Guo et al. focus on service coordination between different air layers in SAGIN [6]. Therefore they separate the requirements of the environment using three service coordination scenarios: (1) fine-grained, (2) medium-grained, and (3) coarse-grained. In the fine-grained scenario, the network in the air is used as complementary regarding the need for the terrestrial network and ubiquitous coverage. Considering the delay-sensitive applications, coordination of the data processing and data communication services is provided by using medium-grained coordination. Finally, mobility and the corresponding service migration are ensured using coarse-grained service coordination.

In [105], Zhou et al. investigate dynamic scheduling problem in task offloading considering SAGIN environment. They deploy UAVs as flying gateways in order to perform the offloading decision. Considering the dynamic environment, they formulate the corresponding problem as an MDP and then apply linear programming. Similarly, Cheng et al. focus on compu-

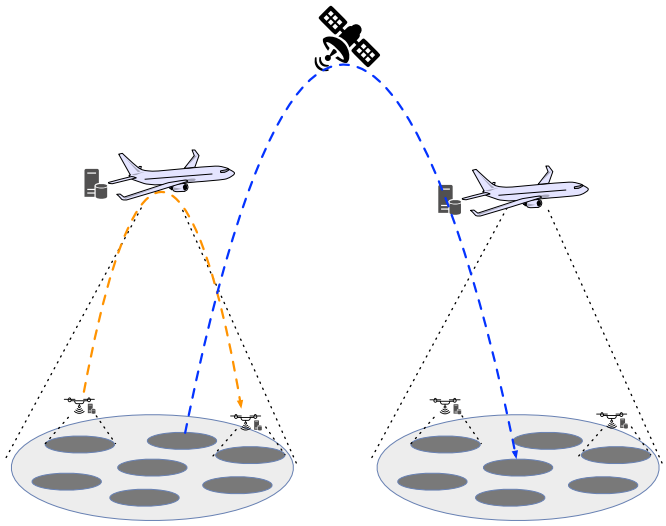


Figure 10: HAPs and LEO satellites provide regional and inter-regional access for LAP components.

tational offloading in SAGIN considering energy and computation constraints [106]. In their design, UAVs provide edge computing while satellites ensure access to cloud computing. To learn the optimal policy in a dynamic environment regarding large action space and mobility, they use an actor-critic DRL algorithm. In [107], Zhang et al. analyze the architecture and corresponding application scenarios of satellite MEC. They propose network function virtualization and cooperative task offloading methods in order to integrate computing resources and improve the efficiency regarding delay and energy consumption.

Resource allocation and controller placement problems are also essential for LEO studies. In [108], Chen et al. examine the dynamic assignment and placement of controllers in SDN-based LEO satellite networks. They consider two challenges including highly dynamic topology, and randomly fluctuating load. To this end, they take propagation and queueing delays into account and then formulate the adaptive controller placement and assignment problem considering management costs. On the other hand, Zhang et al. investigate resource allocation in Non-Orthogonal Multiple Access (NOMA) terrestrial-satellite networks in which terrestrial and satellite components use the same spectrum for the communication [109]. Since the original optimization problem is non-convex, they divided the original problem into three subproblems including user association, bandwidth assignment, and power allocation. In [110], Dahrouj et al. focus on the user scheduling in integrated satellite-HAPs-ground networks considering user-connectivity, backhaul, and power constraints. To this end, they propose a deep neural network driven optimization for the user scheduling policies. Their online approach outperforms traditional model-based optimization methods which fail to meet the QoS requirements.

Since each layer is crucial for the performance of an air computing environment, we also summarize their essential features in Table 5 based on the critical issues. Thus, which layer should be used for particular requirements can be more distinct.

5. Challenges and Future Research Directions

Even though air computing can solve many issues related to the limitation of current networking paradigms, it would face many challenges such as network architecture, regulation of air vehicles, battery issues, coverage, and a communication protocol. Considering the fact that UAV communication between different devices causes many challenges [111], applying different vehicles in the air and providing communication between them is not trivial.

Therefore, all of these issues must be investigated thoroughly in order to apply the air computing paradigm correctly. On the other hand, these challenges open new research areas regarding the provision of QoS, energy efficiency, determining air vehicle placement, and the deployment of AI. In this section, we elaborate on those challenges and corresponding research opportunities.

5.1. Challenges

We evaluate challenges considering the architecture design of air computing, corresponding protocol, and flying vehicle regulations.

5.1.1. Air Computing Architecture Design

Since air computing has a 3D structure with four major layers including terrestrial, LAP, HAP, and LEO, the networking architecture in terms of offloading mechanism and routing is crucial [112]. One of the main concerns in networking is how the requests would be handled considering the mesh connected different nodes in the 3D structure. To investigate this, we propose three candidate design approaches including hierarchical, free, and hybrid designs. Moreover, we also make a comparison between a distributed approach and orchestration in air computing regarding task offloading.

Hierarchical Design - As the name suggests, there is a systematic order between air computing layers in the hierarchical design as shown in Figure 11. If a task is offloaded from a user in a terrestrial network, the selection of the corresponding server for the computation is handled by another entity in the air computing environment, not by the user. For example, if the service for the corresponding application task can be met in one of the HAP vehicles due to network conditions, the task is routed regarding a predefined path rather than directly transmitted to the corresponding HAP component by the user. Therefore, the task is first sent to the nearest edge server in LAN. If the edge server cannot process the task because of its high load or its service incapability, the task is relayed to one of the UAVs in LAP. Note that we assume that the corresponding edge server is in the vicinity of the UAV. Next, if the UAV similarly cannot execute the task, it relays the task to one of the UAVs in its vicinity, one of the available edge servers on the ground, or one of the airplanes in HAP. Since the corresponding service has been given in HAP in this case, the task is pushed to the HAP.

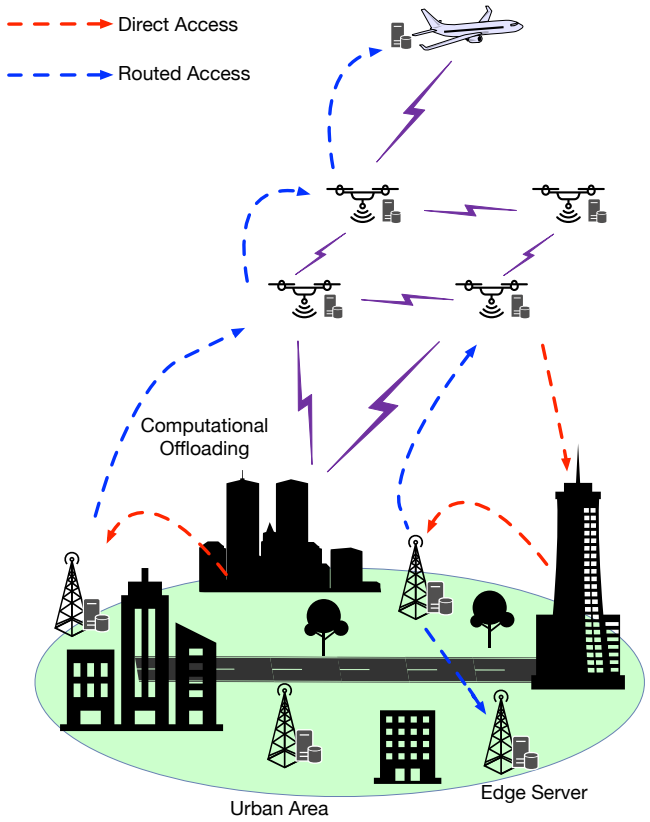


Figure 11: Architecture of the hierarchical design.

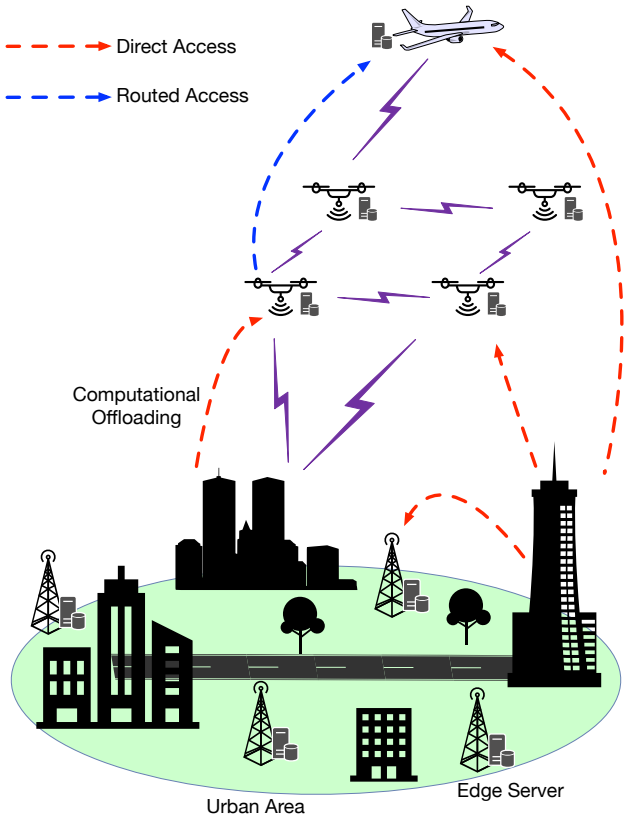


Figure 12: Architecture of the direct access design.

Even though the hierarchical design provides important manageability throughout the air computing environment, it may face high delays due to its layered structure. Based on the use cases and user profiles, it can be applied in the network for specific areas.

Direct Access Design - In contrast to the hierarchical design, in direct access design, IoT devices in an air computing environment can select the corresponding server to meet the requirements of their tasks as shown in Figure 12. As a result, if the corresponding sender knows which server provides the required service, it offloads the task directly to the appropriate layer through the seamless connectivity feature of air computing.

However, this free access for each device in the network would cause congestion and underutilization of the resources. In terms of congestion, if a service is given by a particular server in the network, all devices which require the service may offload their task to that server. As a result, communication links and server capacity would be heavily affected by this situation. On the other hand, if devices use particular servers in the network due to low delay, high processing, and energy efficiency, some resources would be underutilized.

Hybrid Design - Considering the advantages of hierarchical design and direct access design, a hybrid design would be applied in an air computing environment. For the urban areas, in which the network is extremely dense, the application of hierarchical design would be more suitable in order to prevent underutilization and congestion cases. Moreover, the delay issue in hierarchical design can be covered by diverse resources regarding the provision of different services, and seamless connectivity.

On the other hand, considering the suburban and rural areas, the utilization of direct access design would be more convenient since the infrastructure in those areas is limited. Therefore, direct access to the corresponding servers can be of benefit in terms of delay.

5.1.2. Air Computing Protocol

As there are many different entities in the air computing environment, the communication between them should be carried out based on predefined rules, which are defined by a protocol. An air computing protocol should be reliable, secure, and fast so that the entities carry out communication easily [113]. Moreover, it should facilitate the management of the network as it must provide data integrity.

5.1.3. Flying Vehicle Regulations

Since each country has different regulation policies considering flying air vehicles, the entities in an air computing environment should comply with those corresponding rules. Moreover, the air computing protocol should also be in compliance with regulations as reliable and fast communication is vital for flying entities in the air. On the other hand, considering the existing cellular network infrastructure, the standardization between

existing resources and newly deployed air computing devices must be investigated for 6G and beyond [114, 115].

5.2. Future Research Directions

We examine future research direction considering request management, deploying artificial intelligence, energy issue of air vehicles, and movement/coverage of air vehicles. We believe that these directions provide important research opportunities for researchers.

5.2.1. Request Management

In an air computing environment, handling user and IoT requests are crucial to meet the required QoS. Moreover, since there are different layers in the air, the request management would be more complex than other networking paradigms such as edge computing. Even though architecture design is essential for request management, deciding where to offload and when to offload is crucial for the performance. To this end, the benefits of a distributed system and an orchestrator-based system must be investigated thoroughly.

A distributed system in air computing can be described such that the device which offloads the task takes the decision of where to offload. For example, users can select one of the edge servers for offloading, or an edge server relays the offloaded task to one of the UAVs without consulting any intermediate device. One of the most important advantages of a distributed system is its scalability and its low latency. Since there is no intermediate element considering the offloading, the transmission of the request would be faster.

On the other hand, it has a significant drawback considering the current state of the network. An offloading device in the environment cannot be aware of the current condition of the corresponding servers and network elements in terms of the load of the servers, and the number of requests that may affect the communication links. Therefore, if there is no additional communication regarding this information, the decision taken by the offloading device would cause a failed task offloading. Moreover, note that if an additional communication mechanism is deployed in the environment, it must be well-optimized so that the communication links cannot be affected by the transmission regarding congestion.

In contrast to distributed systems, an orchestrator to which the tasks are first sent by the offloading devices can be used in different parts of an air computing environment. In such a system, a user can directly offload tasks to the orchestrator. Here the decision of where to offload is taken by the orchestrator which has full access to the current system state regarding the communication links and corresponding servers.

Even though it has advantages, several points must be investigated and optimized in an orchestrator-based system. First, the deployment of the orchestrator is crucial for the performance of the system since many entities in the network may send their corresponding requests. Therefore, it should be reachable in a suitable time so that it can relay those requests without violating the maximum delay requirement based on application type. Second, the centralized deployment of an orchestrator results in a single point of failure such that the corresponding portion of the network would be heavily affected as users and IoT devices cannot offload their tasks. To this end, a fault management system should be considered with the deployment of an orchestrator. Finally, third, the cost of an orchestrator in terms of new communication links, delay, and congestion must be minimized for the system performance.

5.2.2. Deploying Artificial Intelligence

Considering the diverse requirements of various applications and IoT devices in an air computing environment, traditional optimization-based solutions and heuristics would provide limited solutions. Therefore, the application of AI-based solutions will be inevitable. Especially, regarding edge and UAV-based systems, there are already many studies which benefit from AI-based solutions including ML, Deep Learning (DL), and DRL [116, 117, 118]. As a result, along with the requirements of the air computing paradigm, novel AI solutions should be applied to meet new challenges.

Since there are many resources that produce an enormous amount of data in an air computing environment, processing them and then learning meaningful patterns considering the performance of the system can be feasible using ML techniques. However, in recent years, DL solutions have been preferred rather than traditional ML algorithms since DL is more successful in terms of training, non-linear transformation, efficiency, and required computing power [119].

Even though DL solutions are preferred in recent studies, the data collection phase in an air computing environment would cause serious degradation of system performance. First, the huge amount of data can bring about congestion problems on communication links. Second, for such a heterogeneous environment, providing the privacy of the data would be difficult. Therefore, applying Federated Learning (FL) solutions would be more suitable with air computing [120, 121].

On the other hand, since many decisions must be taken based on dynamic events in the air computing environment, and they may not be labeled due to the nature of the problem, supervised ML and DL based solutions would be inadequate. To this end, recent studies benefit from DRL in which the agent can learn directly from the environment without needing human interaction [122]. The agent in an air computing environment can be the orchestrator, UAV, or edge server since they take actions based on the current state of the system. Thus, we believe that future studies can consider the deployment of DRL solutions along with FL.

5.2.3. Energy Issue of Air Vehicles

The vehicles in air layers use either batteries or fuel. Therefore, their deployment, trajectories, and computing capacities should be well-optimized to ensure energy efficiency. As mentioned in Section 4, energy-related issues are currently evaluated regarding edge and UAV studies. However, considering all layers of air computing, a collaboration between different air vehicles can reduce energy consumption further.

5.2.4. Movement and Coverage of Air Vehicles

Even though deployment is an important research issue for 2D terrestrial networks, this problem is more difficult to manage as air vehicles move. Moreover, since their coverage, transmission quality, and power consumption are heavily affected by their vertical movement, optimization of their altitude and trajectory is crucial [123, 124]. Therefore, request management can be handled along with this optimization in order to provide efficient performance.

6. Conclusion

In this study, we defined a novel, next-generation computational paradigm called air computing. In the face of ever-growing application resource demands, air computing strives to solve bottlenecks and inefficiencies of the computational infrastructure by intelligently harmonizing 2D legacy terrestrial resources with novel 3D vertical networking technologies.

Air computing is indeed based on a family of technologies such as UAV, LAP, HAP, LEO, and edge computing. In order to give a complete overview, we first investigated air computing as a whole regarding its main features and how it contrasts with former systems such as edge and cloud computing. We then described the individual technological components and how they fit in the overall architecture. A detailed literature review for the individual components is also provided to give a full technical overview of air computing in all aspects.

Moreover, we examined the advantages that would be put forward by a possible air computing implementation regarding the QoS requirements of the next-generation applications and QoE expectations of end-users. Then, we elaborated on the potential use cases where the current paradigms experience difficulty in meeting the dynamic user demands. Finally, we analyzed the opportunities and the corresponding challenges in an overall context from the perspectives of both the end users and service providers. Inspired by the challenges involved, we presented a selection of future research directions which we believe have the strong potential to transform the domain.

Acknowledgment

This work is supported by the Turkish Directorate of Strategy and Budget under the TAM Project number 2007K12-873.

References

[1] B. Mao, F. Tang, Y. Kawamoto, N. Kato, Optimizing computation offloading in satellite-uav-served 6g iot: A deep learning approach, IEEE Network 35 (4) (2021) 102–108.

[2] W. Saad, M. Bennis, M. Chen, A vision of 6g wireless systems: Applications, trends, technologies, and open research problems, IEEE network 34 (3) (2019) 134–142.

[3] M. Giordani, M. Polese, M. Mezzavilla, S. Rangan, M. Zorzi, Toward 6g networks: Use cases and technologies, IEEE Communications Magazine 58 (3) (2020) 55–61.

[4] W. Z. Khan, M. Y. Aalsalem, M. K. Khan, M. S. Hossain, M. Atiquzzaman, A reliable internet of things based architecture for oil and gas industry, in: 2017 19th International conference on advanced communication Technology (ICACT), IEEE, 2017, pp. 705–710.

[5] T. Li, Y. Fan, Y. Li, S. Tarkoma, P. Hui, Understanding the long-term evolution of mobile app usage, IEEE Transactions on Mobile Computing (01) (2021) 1–1.

[6] Y. Guo, Q. Li, Y. Li, N. Zhang, S. Wang, Service coordination in the space-air-ground integrated network, IEEE Network 35 (5) (2021) 168–173.

[7] J. Liu, Y. Shi, Z. M. Fadlullah, N. Kato, Space-air-ground integrated network: A survey, IEEE Communications Surveys & Tutorials 20 (4) (2018) 2714–2741.

[8] A. Baltaci, E. Dinc, M. Ozger, A. Alabbasi, C. Cavdar, D. Schupke, A survey of wireless networks for future aerial communications (facom), IEEE Communications Surveys & Tutorials.

[9] Q.-V. Pham, R. Ruby, F. Fang, D. C. Nguyen, Z. Yang, M. Le, Z. Ding, W.-J. Hwang, Aerial computing: A new computing paradigm, applications, and challenges, IEEE Internet of Things Journal 9 (11) (2022) 8339–8363.

[10] X. Cao, P. Yang, M. Alzenad, X. Xi, D. Wu, H. Yanikomeroglu, Airborne communication networks: A survey, IEEE Journal on Selected Areas in Communications 36 (9) (2018) 1907–1926.

[11] N.-N. Dao, Q.-V. Pham, N. H. Tu, T. T. Thanh, V. N. Q. Bao, D. S. Lakew, S. Cho, Survey on aerial radio access networks: toward a comprehensive 6g access infrastructure, IEEE Communications Surveys & Tutorials 23 (2) (2021) 1193–1225.

[12] E. Sisinni, A. Saifullah, S. Han, U. Jennehag, M. Gidlund, Industrial internet of things: Challenges, opportunities, and directions, IEEE transactions on industrial informatics 14 (11) (2018) 4724–4734.

[13] M. A. Siddiqi, H. Yu, J. Joung, 5g ultra-reliable low-latency communication implementation challenges and operational issues with iot devices, Electronics 8 (9) (2019) 981.

[14] H. Wang, G. Ding, F. Gao, J. Chen, J. Wang, L. Wang, Power control in uav-supported ultra dense networks: Communications, caching, and energy transfer, IEEE Communications Magazine 56 (6) (2018) 28–34.

[15] G. K. Kurt, M. G. Khoshkholgh, S. Alfattani, A. Ibrahim, T. S. Darwish, M. S. Alam, H. Yanikomeroglu, A. Yongacoglu, A vision and framework for the high altitude platform station (haps) networks of the future, IEEE Communications Surveys & Tutorials 23 (2) (2021) 729–779.

[16] N. Cheng, W. Xu, W. Shi, Y. Zhou, N. Lu, H. Zhou, X. Shen, Air-ground integrated mobile edge networks: Architecture, challenges, and opportunities, IEEE Communications Magazine 56 (8) (2018) 26–32.

[17] B. Li, Z. Fei, Y. Zhang, Uav communications for 5g and beyond: Recent advances and future trends, IEEE Internet of Things Journal 6 (2) (2018) 2241–2263.

[18] E. El Haber, H. A. Alameddine, C. Assi, S. Sharafeddine, Uav-aided ultra-reliable low-latency computation offloading in future iot networks, IEEE Transactions on Communications 69 (10) (2021) 6838–6851.

[19] P. K. Senyo, E. Addae, R. Boateng, Cloud computing research: A review of research themes, frameworks, methods and future research directions, International Journal of Information Management 38 (1) (2018) 128–139.

[20] M. Laroui, B. Nour, H. Moungla, M. A. Cherif, H. Afifi, M. Guizani, Edge and fog computing for iot: A survey on current research activities & future directions, Computer Communications.

[21] M. Satyanarayanan, V. Bahl, R. Caceres, N. Davies, The case for vm-based cloudlets in mobile computing, IEEE pervasive Computing.

[22] X. Chen, L. Jiao, W. Li, X. Fu, Efficient multi-user computation offloading for mobile-edge cloud computing, IEEE/ACM Transactions on Networking 24 (5) (2015) 2795–2808.

[23] Cisco, Fog computing and the internet of things: Extend the cloud to where the things are, Cisco white paper.

[24] A. C. Baktir, A. Ozgovde, C. Ersoy, How can edge computing benefit from software-defined networking: A survey, use cases, and future directions, IEEE Communications Surveys & Tutorials 19 (4) (2017) 2359–2391.

[25] P. Mach, Z. Becvar, Mobile edge computing: A survey on architecture and computation offloading, IEEE communications surveys & tutorials 19 (3) (2017) 1628–1656.

[26] Y. Mao, C. You, J. Zhang, K. Huang, K. B. Letaief, A survey on mobile edge computing: The communication perspective, IEEE Communications Surveys & Tutorials 19 (4) (2017) 2322–2358.

[27] T. Ouyang, Z. Zhou, X. Chen, Follow me at the edge: Mobility-aware dynamic service placement for mobile edge computing, IEEE Journal

on Selected Areas in Communications 36 (10) (2018) 2333–2345.
[28] X. Zhang, Q. Zhu, Hierarchical caching for statistical qos guaranteed multimedia transmissions over 5g edge computing mobile wireless networks, IEEE Wireless Communications 25 (3) (2018) 12–20.
[29] N. Zhao, W. Lu, M. Sheng, Y. Chen, J. Tang, F. R. Yu, K.-K. Wong, Uav-assisted emergency networks in disasters, IEEE Wireless Communications 26 (1) (2019) 45–51.
[30] M. Erdelj, E. Natalizio, K. R. Chowdhury, I. F. Akyildiz, Help from the sky: Leveraging uavs for disaster management, IEEE Pervasive Computing 16 (1) (2017) 24–32.
[31] P. Dong, X. Wang, S. Wang, Y. Wang, Z. Ning, M. S. Obaidat, Internet of uavs based remote health monitoring: An online ehealth system, IEEE Wireless Communications 28 (3) (2021) 15–21.
[32] S. Ullah, K.-I. Kim, K. H. Kim, M. Imran, P. Khan, E. Tovar, F. Ali, Uav-enabled healthcare architecture: Issues and challenges, Future Generation Computer Systems 97 (2019) 425–432.
[33] F. Manuri, A. Sanna, A survey on applications of augmented reality, ACSIJ Advances in Computer Science: an International Journal 5 (1) (2016) 18–27.
[34] D. Chatzopoulos, C. Bermejo, Z. Huang, P. Hui, Mobile augmented reality survey: From where we are to where we go, Ieee Access 5 (2017) 6917–6950.
[35] Y. Siriwardhana, P. Porambage, M. Liyanage, M. Ylianttila, A survey on mobile augmented reality with 5g mobile edge computing: architectures, applications, and technical aspects, IEEE Communications Surveys & Tutorials 23 (2) (2021) 1160–1192.
[36] E. Montero, C. Rocha, H. Oliveira, E. Cerqueira, P. Mendes, A. Santos, D. Rosário, Proactive radio-and qos-aware uav as bs deployment to improve cellular operations, Computer Networks 200 (2021) 108486.
[37] Z. Zhao, P. Cumino, C. Esposito, M. Xiao, D. Rosário, T. Braun, E. Cerqueira, S. Sargento, Smart unmanned aerial vehicles as base stations placement to improve the mobile network operations, Computer communications 181 (2022) 45–57.
[38] L. J. Horner, Edge strategies in industry: Overview and challenges, IEEE Transactions on Network and Service Management.
[39] Q. Luo, S. Hu, C. Li, G. Li, W. Shi, Resource scheduling in edge computing: A survey, IEEE Communications Surveys & Tutorials.
[40] C. Dong, S. Hu, X. Chen, W. Wen, Joint optimization with dnn partitioning and resource allocation in mobile edge computing, IEEE Transactions on Network and Service Management 18 (4) (2021) 3973–3986.
[41] Z. Liang, Y. Liu, T.-M. Lok, K. Huang, Multi-cell mobile edge computing: Joint service migration and resource allocation, IEEE Transactions on Wireless Communications 20 (9) (2021) 5898–5912.
[42] Z. Chen, Z. Zhou, C. Chen, Code caching-assisted computation offloading and resource allocation for multi-user mobile edge computing, IEEE Transactions on Network and Service Management 18 (4) (2021) 4517–4530.
[43] S. Xia, Z. Yao, Y. Li, S. Mao, Online distributed offloading and computing resource management with energy harvesting for heterogeneous mec-enabled iot, IEEE Transactions on Wireless Communications 20 (10) (2021) 6743–6757.
[44] T. Bahreini, H. Badri, D. Grosu, Mechanisms for resource allocation and pricing in mobile edge computing systems, IEEE Transactions on Parallel and Distributed Systems 33 (3) (2021) 667–682.
[45] J. Zhang, Q. Zhang, Stackelberg game for utility-based cooperative cognitiveradio networks, in: Proceedings of the tenth ACM international symposium on Mobile ad hoc networking and computing, 2009, pp. 23–32.
[46] S. Maharjan, Q. Zhu, Y. Zhang, S. Gjessing, T. Basar, Dependable demand response management in the smart grid: A stackelberg game approach, IEEE Transactions on Smart Grid 4 (1) (2013) 120–132.
[47] R. Roostaei, Z. Dabiri, Z. Movahedi, A game-theoretic joint optimal pricing and resource allocation for mobile edge computing in noma-based 5g networks and beyond, Computer Networks 198 (2021) 108352.
[48] C. Zhao, Y. Cai, A. Liu, M. Zhao, L. Hanzo, Mobile edge computing meets mmwave communications: Joint beamforming and resource allocation for system delay minimization, IEEE Transactions on Wireless Communications 19 (4) (2020) 2382–2396.
[49] C.-X. Wang, F. Haider, X. Gao, X.-H. You, Y. Yang, D. Yuan, H. M. Aggoune, H. Haas, S. Fletcher, E. Hepsaydir, Cellular architecture and key technologies for 5g wireless communication networks, IEEE communications magazine 52 (2) (2014) 122–130.
[50] Z. Wang, M. Goudarzi, M. Gong, R. Buyya, Deep reinforcement learning-based scheduling for optimizing system load and response time in edge and fog computing environments, Future Generation Computer Systems 152 (2023) 55–69.
[51] P. Yuan, S. Shao, L. Geng, X. Zhao, Caching hit ratio maximization in mobile edge computing with node cooperation, Computer Networks 200 (2021) 108507.
[52] L. Feng, Y. Zhou, T. Liu, X. Que, P. Yu, T. Hong, X. Qiu, Energy-efficient offloading for mission-critical iot services using evt-embedded intelligent learning, IEEE Transactions on Green Communications and Networking 5 (3) (2021) 1179–1190.
[53] J. Xue, Q. Hu, Y. An, L. Wang, Joint task offloading and resource allocation in vehicle-assisted multi-access edge computing, Computer Communications 177 (2021) 77–85.
[54] A. Bozorgchenani, F. Mashhadi, D. Tarchi, S. A. S. Monroy, Multi-objective computation sharing in energy and delay constrained mobile edge computing environments, IEEE Transactions on Mobile Computing 20 (10) (2020) 2992–3005.
[55] X. Chen, J. Zhang, B. Lin, Z. Chen, K. Wolter, G. Min, Energy-efficient offloading for dnn-based smart iot systems in cloud-edge environments, IEEE Transactions on Parallel and Distributed Systems 33 (3) (2021) 683–697.
[56] R. A. Dziyauddin, D. Niyato, N. C. Luong, A. A. A. M. Atan, M. A. M. Izhar, M. H. Azmi, S. M. Daud, Computation offloading and content caching and delivery in vehicular edge network: A survey, Computer Networks 197 (2021) 108228.
[57] G. Peng, H. Wu, H. Wu, K. Wolter, Constrained multiobjective optimization for iot-enabled computation offloading in collaborative edge and cloud computing, IEEE Internet of Things Journal 8 (17) (2021) 13723–13736.
[58] M. J. Neely, Stochastic network optimization with application to communication and queueing systems, Synthesis Lectures on Communication Networks 3 (1) (2010) 1–211.
[59] Y. Chen, X. Zhou, W. Wang, H. Wang, Z. Zhang, Z. Zhang, Delay-optimal closed-form scheduling for multi-destination computation offloading, IEEE Wireless Communications Letters 10 (9) (2021) 1904–1908.
[60] H. Feng, S. Guo, L. Yang, Y. Yang, Collaborative data caching and computation offloading for multi-service mobile edge computing, IEEE Transactions on Vehicular Technology 70 (9) (2021) 9408–9422.
[61] J. Xu, Z. Wei, Z. Lyu, L. Shi, J. Han, Throughput maximization of offloading tasks in multi-access edge computing networks for high-speed railways, IEEE Transactions on Vehicular Technology 70 (9) (2021) 9525–9539.
[62] Y. Zhang, H.-Y. Wei, Risk-aware cloud-edge computing framework for delay-sensitive industrial iots, IEEE Transactions on Network and Service Management 18 (3) (2021) 2659–2671.
[63] B. Yang, X. Cao, J. Bassey, X. Li, L. Qian, Computation offloading in multi-access edge computing: A multi-task learning approach, IEEE transactions on mobile computing 20 (9) (2020) 2745–2762.
[64] C. Li, J. Liu, Q. Zhang, Y. Luo, Efficient cooperative cache management for latency-aware data intelligent processing in edge environment, Future Generation Computer Systems 123 (2021) 48–67.
[65] Z. Song, Y. Hao, Y. Liu, X. Sun, Energy-efficient multiaccess edge computing for terrestrial-satellite internet of things, IEEE Internet of Things Journal 8 (18) (2021) 14202–14218.
[66] S. Ghoorchian, S. Maghsudi, Multi-armed bandit for energy-efficient and delay-sensitive edge computing in dynamic networks with uncertainty, IEEE Transactions on Cognitive Communications and Networking 7 (1) (2020) 279–293.
[67] X. Zhou, S. Ge, T. Qiu, K. Li, M. Atiquzzaman, Energy-efficient service migration for multi-user heterogeneous dense cellular networks, IEEE Transactions on Mobile Computing.
[68] J. Lyu, Y. Zeng, R. Zhang, Uav-aided offloading for cellular hotspot, IEEE Transactions on Wireless Communications 17 (6) (2018) 3988–4001.
[69] L. Gupta, R. Jain, G. Vaszkun, Survey of important issues in uav communication networks, IEEE Communications Surveys & Tutorials 18 (2) (2015) 1123–1152.
[70] W. Shi, J. Li, N. Cheng, F. Lyu, S. Zhang, H. Zhou, X. Shen, Multi-drone

3-d trajectory planning and scheduling in drone-assisted radio access networks, IEEE Transactions on Vehicular Technology 68 (8) (2019) 8145–8158.
[71] Y. Zhou, N. Cheng, N. Lu, X. S. Shen, Multi-uav-aided networks: Aerial-ground cooperative vehicular networking architecture, ieee vehicular technology magazine 10 (4) (2015) 36–44.
[72] S. Jeong, O. Simeone, J. Kang, Mobile edge computing via a uav-mounted cloudlet: Optimization of bit allocation and path planning, IEEE Transactions on Vehicular Technology 67 (3) (2017) 2049–2063.
[73] Z. Zhao, L. Pacheco, H. Santos, M. Liu, A. Di Maio, D. Rosário, E. Cerqueira, T. Braun, X. Cao, Predictive uav base station deployment and service offloading with distributed edge learning, IEEE Transactions on Network and Service Management 18 (4) (2021) 3955–3972.
[74] L. Wang, K. Wang, C. Pan, W. Xu, N. Aslam, A. Nallanathan, Deep reinforcement learning based dynamic trajectory control for uav-assisted mobile edge computing, IEEE Transactions on Mobile Computing.
[75] Z. Li, M. Chen, C. Pan, N. Huang, Z. Yang, A. Nallanathan, Joint trajectory and communication design for secure uav networks, IEEE Communications Letters 23 (4) (2019) 636–639.
[76] Q. Liu, L. Shi, L. Sun, J. Li, M. Ding, F. Shu, Path planning for uav-mounted mobile edge computing with deep reinforcement learning, IEEE Transactions on Vehicular Technology 69 (5) (2020) 5723–5728.
[77] J. Wang, Z. Na, X. Liu, Collaborative design of multi-uav trajectory and resource scheduling for 6g-enabled internet of things, IEEE Internet of Things Journal 8 (20) (2020) 15096–15106.
[78] F. Zhou, Y. Wu, R. Q. Hu, Y. Qian, Computation rate maximization in uav-enabled wireless-powered mobile-edge computing systems, IEEE Journal on Selected Areas in Communications 36 (9) (2018) 1927–1941.
[79] N. H. Motlagh, T. Taleb, O. Arouk, Low-altitude unmanned aerial vehicles-based internet of things services: Comprehensive survey and future perspectives, IEEE Internet of Things Journal 3 (6) (2016) 899–922.
[80] A. M. Seid, G. O. Boateng, B. Mareri, G. Sun, W. Jiang, Multi-agent drl for task offloading and resource allocation in multi-uav enabled iot edge network, IEEE Transactions on Network and Service Management 18 (4) (2021) 4531–4547.
[81] P. A. Apostolopoulos, G. Fragkos, E. E. Tsiropoulou, S. Papavassiliou, Data offloading in uav-assisted multi-access edge computing systems under resource uncertainty, IEEE Transactions on Mobile Computing.
[82] D. Kahneman, A. Tversky, Prospect theory: An analysis of decision under risk, in: Handbook of the fundamentals of financial decision making: Part I, World Scientific, 2013, pp. 99–127.
[83] C. Zhan, H. Hu, Z. Liu, Z. Wang, S. Mao, Multi-uav-enabled mobile-edge computing for time-constrained iot applications, IEEE Internet of Things Journal 8 (20) (2021) 15553–15567.
[84] N. Zhao, Z. Ye, Y. Pei, Y.-C. Liang, D. Niyato, Multi-agent deep reinforcement learning for task offloading in uav-assisted mobile edge computing, IEEE Transactions on Wireless Communications.
[85] X. Diao, W. Yang, L. Yang, Y. Cai, Uav-relaying-assisted multi-access edge computing with multi-antenna base station: Offloading and scheduling optimization, IEEE Transactions on Vehicular Technology 70 (9) (2021) 9495–9509.
[86] Y. Zeng, D. Lu, J. Du, Joint optimized multi-user access and uav deployments based on heterogeneous revenue in iot network, Computer Networks 234 (2023) 109919.
[87] J. Shi, C. Li, Y. Guan, P. Cong, J. Li, Multi-uav-assisted computation offloading in dt-based networks: A distributed deep reinforcement learning approach, Computer Communications 210 (2023) 217–228.
[88] J. Ji, K. Zhu, C. Yi, D. Niyato, Energy consumption minimization in uav-assisted mobile-edge computing systems: joint resource allocation and trajectory design, IEEE Internet of Things Journal 8 (10) (2020) 8570–8584.
[89] M. Li, N. Cheng, J. Gao, Y. Wang, L. Zhao, X. Shen, Energy-efficient uav-assisted mobile edge computing: Resource allocation and trajectory optimization, IEEE Transactions on Vehicular Technology 69 (3) (2020) 3424–3438.
[90] X. Chen, X. Liu, Y. Chen, L. Jiao, G. Min, Deep q-network based resource allocation for uav-assisted ultra-dense networks, Computer Networks 196 (2021) 108249.
[91] Z. Liu, X. Tan, M. Wen, S. Wang, C. Liang, An energy-efficient selection mechanism of relay and edge computing in uav-assisted cellular networks, IEEE Transactions on Green Communications and Networking 5 (3) (2021) 1306–1318.
[92] B. Li, W. Liu, W. Xie, X. Li, Energy-efficient task offloading and trajectory planning in uav-enabled mobile edge computing networks, Computer Networks 234 (2023) 109940.
[93] R. Borralho, A. Mohamed, A. U. Quddus, P. Vieira, R. Tafazolli, A survey on coverage enhancement in cellular networks: challenges and solutions for future deployments, IEEE Communications Surveys & Tutorials 23 (2) (2021) 1302–1341.
[94] M. Mozaffari, W. Saad, M. Bennis, M. Debbah, Unmanned aerial vehicle with underlaid device-to-device communications: Performance and tradeoffs, IEEE Transactions on Wireless Communications 15 (6) (2016) 3949–3963.
[95] J. Lyu, Y. Zeng, R. Zhang, T. J. Lim, Placement optimization of uav-mounted mobile base stations, IEEE Communications Letters 21 (3) (2016) 604–607.
[96] M. Alzenad, A. El-Keyi, F. Lagum, H. Yanikomeroglu, 3-d placement of an unmanned aerial vehicle base station (uav-bs) for energy-efficient maximal coverage, IEEE Wireless Communications Letters 6 (4) (2017) 434–437.
[97] C. H. Liu, Z. Chen, J. Tang, J. Xu, C. Piao, Energy-efficient uav control for effective and fair communication coverage: A deep reinforcement learning approach, IEEE Journal on Selected Areas in Communications 36 (9) (2018) 2059–2070.
[98] Y. Wang, Z.-Y. Ru, K. Wang, P.-Q. Huang, Joint deployment and task scheduling optimization for large-scale mobile users in multi-uav-enabled mobile edge computing, IEEE transactions on cybernetics 50 (9) (2019) 3984–3997.
[99] T. Yuan, C. E. Rothenberg, K. Obraczka, C. Barakat, T. Turletti, Harnessing uavs for fair 5g bandwidth allocation in vehicular communication via deep reinforcement learning, IEEE Transactions on Network and Service Management 18 (4) (2021) 4063–4074.
[100] M. M. Abdelhakam, M. M. Elmesalawy, I. I. Ibrahim, S. G. Sayed, Collaborative comp and trajectory optimization for energy minimization in multi-uav-assisted iot networks with qos guarantee, Computer Networks 237 (2023) 110074.
[101] J. Qiu, D. Grace, G. Ding, M. D. Zakaria, Q. Wu, Air-ground heterogeneous networks for 5g and beyond via integrating high and low altitude platforms, IEEE Wireless Communications 26 (6) (2019) 140–148.
[102] O. Kodheli, E. Lagunas, N. Maturo, S. K. Sharma, B. Shankar, J. F. M. Montoya, J. C. M. Duncan, D. Spano, S. Chatzinotas, S. Kisseleff, et al., Satellite communications in the new space era: A survey and future challenges, IEEE Communications Surveys & Tutorials 23 (1) (2020) 70–109.
[103] F. Tang, H. Hofner, N. Kato, K. Kaneko, Y. Yamashita, M. Hangai, A deep reinforcement learning-based dynamic traffic offloading in space-air-ground integrated networks (sagin), IEEE Journal on Selected Areas in Communications 40 (1) (2021) 276–289.
[104] H. Chen, M. Xiao, Z. Pang, Satellite-based computing networks with federated learning, IEEE Wireless Communications 29 (1) (2022) 78–84.
[105] C. Zhou, W. Wu, H. He, P. Yang, F. Lyu, N. Cheng, X. Shen, Delay-aware iot task scheduling in space-air-ground integrated network, in: 2019 IEEE Global Communications Conference (GLOBECOM), IEEE, 2019, pp. 1–6.
[106] N. Cheng, F. Lyu, W. Quan, C. Zhou, H. He, W. Shi, X. Shen, Space/aerial-assisted computing offloading for iot applications: A learning-based approach, IEEE Journal on Selected Areas in Communications 37 (5) (2019) 1117–1129.
[107] Z. Zhang, W. Zhang, F.-H. Tseng, Satellite mobile edge computing: Improving qos of high-speed satellite-terrestrial networks using edge computing techniques, IEEE network 33 (1) (2019) 70–76.
[108] L. Chen, F. Tang, X. Li, Mobility-and load-adaptive controller placement and assignment in leo satellite networks, in: IEEE INFOCOM 2021-IEEE Conference on Computer Communications, IEEE, 2021, pp. 1–10.
[109] Y. Zhang, H. Zhang, H. Zhou, K. Long, G. K. Karagiannidis, Resource allocation in terrestrial-satellite-based next generation multiple access networks with interference cooperation, IEEE Journal on Selected Areas in Communications 40 (4) (2022) 1210–1221.
[110] H. Dahrouj, S. Liu, M.-S. Alouini, Machine learning-based user

scheduling in integrated satellite-haps-ground networks, IEEE Network 37 (2) (2023) 102–109.
[111] A. Fotouhi, H. Qiang, M. Ding, M. Hassan, L. G. Giordano, A. Garcia-Rodriguez, J. Yuan, Survey on uav cellular communications: Practical aspects, standardization advancements, regulation, and security challenges, IEEE Communications Surveys & Tutorials 21 (4) (2019) 3417–3442.
[112] X. Huang, J. A. Zhang, R. P. Liu, Y. J. Guo, L. Hanzo, Airplane-aided integrated networking for 6g wireless: Will it work?, IEEE Vehicular Technology Magazine 14 (3) (2019) 84–91.
[113] V. Hassija, V. Chamola, A. Agrawal, A. Goyal, N. C. Luong, D. Niyato, F. R. Yu, M. Guizani, Fast, reliable, and secure drone communication: A comprehensive survey, IEEE Communications Surveys & Tutorials.
[114] A. S. Abdalla, A. Yingst, K. Powell, A. Gelonch-Bosch, V. Marojevic, Open source software radio platform for research on cellular networked uavs: It works!, IEEE Communications Magazine 60 (2) (2022) 60–66.
[115] I. Bor-Yaliniz, M. Salem, G. Senerath, H. Yanikomeroglu, Is 5g ready for drones: A look into contemporary and prospective wireless networks from a standardization perspective, IEEE Wireless Communications 26 (1) (2019) 18–27.
[116] X. Wang, Y. Han, V. C. Leung, D. Niyato, X. Yan, X. Chen, Convergence of edge computing and deep learning: A comprehensive survey, IEEE Communications Surveys & Tutorials 22 (2) (2020) 869–904.
[117] X. Chen, S. Leng, J. He, L. Zhou, Deep-learning-based intelligent intervehicle distance control for 6g-enabled cooperative autonomous driving, IEEE Internet of Things Journal 8 (20) (2020) 15180–15190.
[118] C. Sun, X. Wu, X. Li, Q. Fan, J. Wen, V. C. Leung, Cooperative computation offloading for multi-access edge computing in 6g mobile networks via soft actor critic, IEEE Transactions on Network Science and Engineering.
[119] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, nature 521 (7553) (2015) 436–444.
[120] D. C. Nguyen, M. Ding, P. N. Pathirana, A. Seneviratne, J. Li, H. V. Poor, Federated learning for internet of things: A comprehensive survey, IEEE Communications Surveys & Tutorials 23 (3) (2021) 1622–1658.
[121] S. Samarakoon, M. Bennis, W. Saad, M. Debbah, Federated learning for ultra-reliable low-latency v2v communications, in: 2018 IEEE Global Communications Conference (GLOBECOM), IEEE, 2018, pp. 1–7.
[122] W. Chen, X. Qiu, T. Cai, H.-N. Dai, Z. Zheng, Y. Zhang, Deep reinforcement learning for internet of things: A comprehensive survey, IEEE Communications Surveys & Tutorials 23 (3) (2021) 1659–1692.
[123] H. He, S. Zhang, Y. Zeng, R. Zhang, Joint altitude and beamwidth optimization for uav-enabled multiuser communications, IEEE Communications Letters 22 (2) (2017) 344–347.
[124] W. Huang, Z. Yang, C. Pan, L. Pei, M. Chen, M. Shikh-Bahaei, M. Elkashlan, A. Nallanathan, Joint power, altitude, location and bandwidth optimization for uav with underlaid d2d communications, IEEE Wireless Communications Letters 8 (2) (2018) 524–527.