A Roadmap Towards Optimal Resource Allocation Approaches in the Internet of Things

Jiyin Zhou
College of Big Data, Chongqing Vocational College of Transportation Chongqing 402247, Chongqing, China

***Abstract*—Introducing new technologies has facilitated people's lives more than ever. As one of these emerging technologies, the Internet of Things (IoT) enables objects we handle daily to interact with each other or humans and exchange information through the Internet by being equipped with sensors and communication technologies. IoT turns the physical world into a virtual world where heterogeneous objects and devices can be interconnected and controlled. IoT-based networks face numerous challenges, including energy and sensor transmission limitations. New technologies are needed to spread the IoT platform, optimize costs, cover heterogeneous connections, reduce power consumption, and diminish delays. Users of IoT-based systems typically use services that are integrated into these networks. Service providers provide users with on-demand services. The interrelationship between this request and response must be managed in a way that is done using a resource allocation strategy. Therefore, resource allocation plays a major role in these systems and networks. The allocation of resources involves matters such as how much, where, and when available resources should be provided to the user economically. The allocation of resources in the IoT environment is also subject to various challenges, including maintaining the quality of service, achieving a predetermined level of service, storing power, controlling congestion, and reducing costs. As the IoT resource allocation problem is an NP-Hard one, many research efforts have been conducted on this topic, and various algorithms have been developed. This paper reviews published publications on IoT resource allocation, outlining the underlying principles, the latest developments, and current trends.**

***Keywords*—Internet of things; resource utilization; resource allocation; systematic review**

I. Introduction

Over the last decade, the Internet of Things [1] has acquired popularity as a worldwide network, allowing physical objects to be controlled, monitored, and managed via the Internet and communication networks [2]. The integration of 5G connectivity [3], cloud computing [4], smart grids [5], and plasma sources [6] in IoT plays a vital role in enabling high-speed, scalable, reliable, and energy-efficient communication, storage, energy management, and advanced sensing capabilities, paving the way for a new era of interconnected devices and intelligent systems. Physical objects become smart by integrating radio frequency identification tags, sensors, actuators, etc. The most important characteristic of IoT objects is the ability to sense the surroundings, communicate, and interact with other objects. IoT devices operate on rechargeable batteries and non-renewable energy sources [7]. Whenever this equipment is deployed for permanent purposes, like ongoing environmental monitoring, it becomes imperative to extend the life of the network. Wireless communication is the main source of energy consumption in IoT devices, which must operate over an extended time with limited resources [8]. The development of IoT devices requires both hardware and software to facilitate access at any place and at any time, and they also need to use the most efficient method for communication and resource allocation [9].

The optimum allocation of resources has always been of significant importance in improving the efficiency of the network. Resource allocation in IoT networks is challenging due to complex and large-scale communication between objects, heterogeneity, and the traditional characteristics of Wireless Sensor Networks (WANs) [10]. In addition, the IoT is not characterized by a uniform and fixed correlation. Therefore, in this environment, where the order and location of the nodes change regularly and quickly, resource allocation should be done in a distributed manner. Also, as mentioned above, owing to the limitations of energy sources, low power processors, limited memory capacity, wireless communication range, and communication bandwidth, the allocation of resources in this environment should bear lower overhead, reducing communication costs [11].

Two general aspects of resource allocation should be discussed and investigated to ensure high performance and create resource allocation strategies within an acceptable time frame. Several factors influence resource allocation, including cost, energy, response time, and security. The problem of reducing delay in resource allocation is also a critical and fundamental issue facing researchers in this research area [12]. The IoT has become an important and fundamental issue because of the connection of countless objects, the high volume of traffic, the storage of data and information, and the need for high-speed resources [13].

Machine learning and Artificial Intelligence (AI) have emerged as indispensable tools in the field of IoT resource allocation, revolutionizing the way resources are managed and utilized in complex and dynamic IoT ecosystems. The sheer volume and heterogeneity of IoT devices, coupled with the varying resource requirements and dynamic nature of IoT applications, pose significant challenges for efficient resource allocation. Machine learning algorithms, coupled with AI techniques, offer a powerful solution by leveraging data-driven insights and intelligent decision-making capabilities [14, 15]. One of the primary benefits of machine learning and AI in IoT resource allocation is their ability to analyze vast amounts of data collected from IoT devices and sensors. These technologies can extract meaningful patterns, detect anomalies, and predict resource demands with high accuracy. By leveraging this data-driven intelligence, resource allocation algorithms can dynamically adapt and optimize resource allocation strategies based on real-time conditions and demands [16].

Moreover, machine learning algorithms can learn from historical resource allocation patterns and optimize resource utilization, leading to enhanced efficiency and cost-effectiveness [17]. These algorithms can identify resource bottlenecks, predict resource congestion, and dynamically allocate resources to alleviate these issues. By intelligently managing resource allocation, machine learning and AI techniques can improve system performance, reduce energy consumption, and enhance the overall quality of service in IoT networks [18, 19]. Furthermore, machine learning and AI can enable proactive resource allocation by considering contextual information, user preferences, and application requirements [20]. These technologies can learn from past user behavior and application performance to make intelligent predictions and allocate resources accordingly [21]. This personalized resource allocation approach can lead to enhanced user satisfaction, improved application performance, and efficient resource utilization.

This paper makes several significant contributions to the field of resource allocation in the IoT. First and foremost, it provides a comprehensive review of existing literature on IoT resource allocation, outlining the underlying principles, latest developments, and current trends. Furthermore, the paper analyzes the characteristics of datasets used in the evaluation of resource allocation algorithms. It considers factors such as the heterogeneity of IoT devices, data traffic patterns, and resource requirements. This analysis provides insights into how these dataset characteristics can impact the performance and suitability of resource allocation algorithms. It highlights the need for tailored approaches that consider the specific requirements of different types of data and IoT scenarios. The paper also conducts a detailed analysis of comparative results obtained from different datasets. It identifies patterns, trends, and discrepancies in algorithm performance, shedding light on the strengths and weaknesses of various resource allocation approaches. This analysis enhances our understanding of which algorithms may be better suited for specific types of data or IoT deployments.

The rest of the paper is organized as follows. Section II will provide a comprehensive background on the IoT and its significance in enabling objects to interact and exchange information through the Internet. It will highlight the challenges faced by IoT-based networks, particularly in terms of resource allocation, and emphasize the need for efficient resource utilization and allocation strategies. In Section III, we will conduct a systematic review of published publications on IoT resource allocation. We will discuss the underlying principles and concepts related to resource allocation in IoT systems. The review will encompass various algorithms, techniques, and methodologies proposed in the literature, highlighting their strengths, limitations, and applicability in different scenarios. We will analyze the latest developments and trends in resource allocation approaches, considering factors such as quality of service, power consumption, congestion control, and cost reduction. Section IV will provide a comprehensive discussion and analysis of the reviewed resource allocation approaches. We will compare the different strategies, identify common challenges and emerging trends, and discuss their implications for future research and practical implementations. Additionally, we will address any gaps or limitations in the existing approaches and propose potential avenues for further exploration and improvement. Section 5 outlines V the potential research directions. The final section of the paper will summarize the key findings and contributions of the research.

II. Background

In this section, general information about IoT, resource allocation, and the challenges of resource allocation is given.

A. Internet of Things

The rapid development of hardware and network technology has allowed a variety of smart devices to connect to the Internet and exchange data. This has led to the development of a new technology called the IoT [22]. IoT has evolved into a worldwide network of physical objects that can be controlled, monitored, and managed through the Internet and communication networks [23]. In IoT-based networks, physical objects are transformed into smart objects using radio frequency identification tags, sensors, actuators, and other components and then managed and controlled via mobile applications. With these capabilities embedded in the IoT, many applications have been developed, including home automation, industrial automation, the medical and healthcare industry, energy management, traffic management, and many more. By connecting physical objects, such as sensors, with the Internet, IoT technology can capture real-time data and then process and analyze it to create insights that can decide and take action. This makes it possible to automate various processes and create intelligent systems that can respond to events in real-time [24].

The IoT devices communicate with each other, share information, and take coordinated actions by sharing their vision, hearing, and thinking. IoT also faces challenges, such as technical difficulties, standardization, security, efficient resource utilization, and privacy concerns. With the significant growth of IoT resources in recent years, much data has been produced, which requires storage, processing, security, and management. To handle and manage such a large amount of data, it is necessary to use new technologies [25]. IoT-based networks and their nodes are challenging to manage due to the increase in data volume, the diversity of the nodes, and the requirement for resource allocation. Because countless objects in this network generate data in real time [26], IoT is characterized by the architecture shown in Fig. 1. Several factors have been considered and addressed in this architecture, including scalability, interoperability, reliability, and quality of service. This architecture is composed of five layers, which are described as follows.

- Perception layer: This layer, called the objects layer, represents the lowest layer of physical or hardware components. Data is collected at this layer and converted into signals that can be transmitted over networks and accessed by applications.

- Network layer: This layer facilitates the exchange of information between objects by connecting them. This layer ensures the secure transfer of information from sensor devices to the information processing system.

- Service management layer: This layer allows IoT programmers to work on heterogeneous devices regardless of their hardware configuration. Also, this layer handles the management of services, gathers data from the network layer, and archives it in the database.

- Application layer: This layer is typically responsible for providing services and applications that enable the integration and analysis of information received. This layer is crucial to providing high-quality intelligent IoT services.

- Business layer: This layer manages IoT services and activities and creates flowcharts, graphs, and business models based on the information derived from the application layer.

IoT capabilities are based on six key elements. These six key elements are essential for maximizing the potential of IoT technology.

- Identification: Identification is crucial in determining the IoT's purpose and providing services under customer demand. Identification can be accomplished using a variety of methods. An electronic product code is a permanent tagging technology containing a unique code for everything anywhere in the world.

- Sensing: Sensing in the IoT involves capturing data generated by sensors located on the network and forwarding it to a central repository, database, or the cloud. The collected data is processed to perform specific tasks in accordance with the requested services. This data can identify patterns, anticipate future needs, and improve the efficiency of the IoT system.

- Connectivity: This technology enables heterogeneous objects to communicate with each other to provide intelligent services. IoT nodes should be capable of operating with limited power supplies and weak or noisy communication links. Bluetooth, IEEE 802.15.4, and Wi-Fi are common IoT communication protocols.

- Computing: The computing ability of the IoT is represented by the processing unit (for example, a microcontroller) and application software. A variety of software platforms provide IoT capabilities. For instance, Amazon Web Services (AWS) provides users with a range of tools for developing and deploying applications for the IoT, such as AWS Greengrass, AWS IoT Core, and AWS IoT Analytics.

- Services: IoT services are classified into four groups, namely ubiquitous, collaborative, information aggregation, and identity-related. Identity-related services provide valuable services that can be incorporated into other services. Aggregation services summarize or aggregate sensed data. Collaborative services use information obtained by information-gathering services to react and decide. Ubiquitous services make collaborative services available to anyone, anywhere, at any time.

- Semantics: to provide services, knowledge must be intelligently extracted from different devices. Knowledge extraction includes discovery, resource utilization, and information modeling. It involves identifying and analyzing data to provide accurate services.

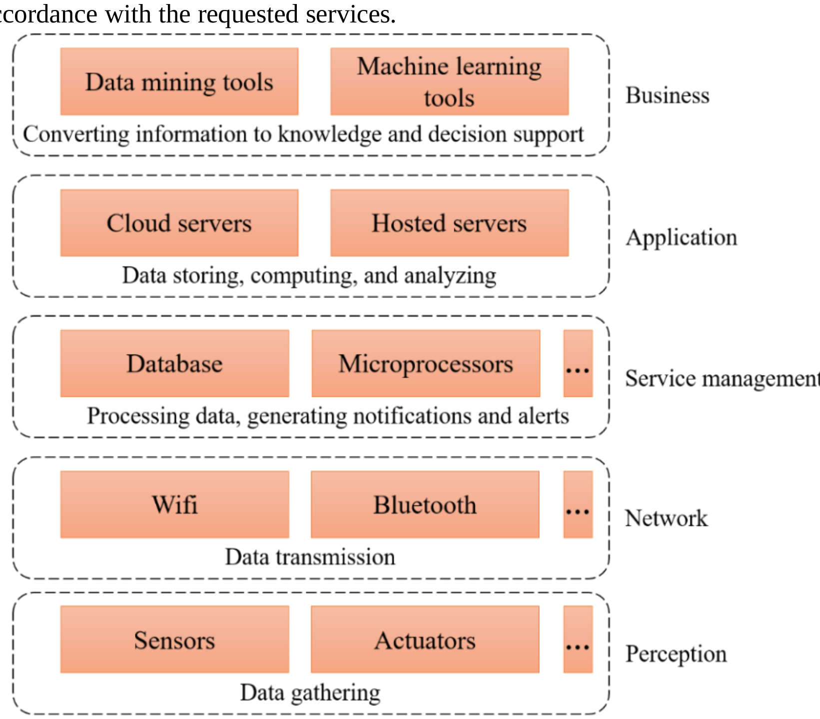

Fig. 1. IoT architecture.

B. Resource Allocation

Resource allocation refers to allocating available resources to users efficiently regarding CPU, memory, network, bandwidth, etc. Also, resource allocation in IoT environments is subject to several challenges, including maintaining the quality of service, ensuring an agreed service level, conserving energy, handling congestion, and reducing costs. These challenges arise from the need to balance the conflicting objectives of allocating resources fairly and equitably to all users, while also optimizing the utilization of those resources. Resource allocation must consider that IoT environments are often characterized by heterogeneity, dynamism, and unpredictability [27]. Key considerations in allocating IoT resources are briefly stated in the following.

- Heterogeneity: The main idea of IoT is the widespread presence of objects in human life. Such an environment contains a wide range of different hardware and devices in diverse shapes and sizes. These devices can be very diverse. Computers, mobile phones, personal digital assistants, and wireless devices, such as radio frequency identification chips, have various communication ranges. Therefore, resource allocation should be conducted by considering the different computing capabilities of objects.

- High adaptability: IoT objects operate in unpredictable patterns and establish ad hoc connections with other nearby objects. Therefore, the movement of nodes cannot be ignored, and the network is expected to be highly dynamic. Centralized resource allocation is inappropriate in a dynamic structure, where nodes are moved around regularly and frequently. The resource allocation process must be flexible to respond to the dynamics of network connectivity.

- Scalability: Another challenge in IoT resource allocation is the need for interaction between thousands of devices at any time and place. As the network dimensions increase, creating a logical structure, such as a tree structure, becomes more challenging. It is also difficult to allocate and manage resources because of a lack of uniform coordination.

- Energy awareness: Energy consumption for communication and computing is a significant constraint for various IoT entities. Since IoT devices are powered by batteries for a long period, resource allocation should be done in a manner that minimizes energy consumption so that the network's energy status is used when allocating resources.

- Load balancing: The system load must be balanced for optimal resource utilization. During the resource allocation process, the workload, which includes energy consumption and communication activities, should be distributed throughout the network so that no part of the network runs out of resources more quickly than others.

C. Mathematical Formulations of the Problem

It is first necessary to define the parameters, conditions, and characteristics of resources and tasks to explain the resource allocation problem. Virtual Resources (VRs) are expressed as follows.

$$VR = \{CP_r . SR_r . SW_r . CO_r . ER_r\} \quad (1)$$

In Eq.1, CP_r refers to the number of computing resources, SR_r denotes the memory space, SW_r represents the set of software supported by VR, CO_r signifies the cost of VR per unit of time, and ER_r reflects the energy consumption of VR. Since certain tasks may require more than one machine, each task is divided into sub-tasks based on priority and execution order. Subtasks are defined as follows:

$$ST = \{CP_t . SR_t . SW_t\} \quad (2)$$

STs are measured in terms of three parameters: number of computing resources required (CP_t), memory space required (SR_t), and software type required (SW_t). Therefore, the following tasks are included in each task.

$$T = \{ST_1 . ST_2 . \dots . ST_{nT}\} \quad (3)$$

In the above definition, nT indicates the number of sub-tasks related to task T. Assuming that there are N tasks and M VR nodes, the problem of assigning tasks to resources involves assigning all the sub-tasks of these N tasks to M VRs. Allocation is subject to several basic conditions as follows.

- Software limitation: Each task with its sub-tasks must be assigned to a machine that can execute the software requirements of the corresponding task.

- The number of resources: The number of resources required for the task with its sub-tasks should not exceed the number of VR resources to which the task is allocated.

- Order and priority of execution of sub-tasks: The order of execution of sub-tasks should be based on the sequence of its execution process, and all the sub-tasks of the sequence should be assigned to a VR for execution.

- Time: Tasks must be assigned to resources executed and completed within a specified time frame. The execution of all sub-tasks must start and end within the specified time interval.

III. Review of Resource Allocation Approaches

In this section, previous methods in IoT resource allocation are examined in five different categories.

A. QoS-aware Approaches

Two key factors in an IoT ecosystem are providing different service quality requirements and quick access to resources. These parameters represent resource distribution across different layers of the IoT environment, encompassing context awareness, performance, response time, and availability. Fog computing is one of the effective techniques to increase service quality, reduce network latency and energy consumption for IoT devices.

In [28], to minimize the network overhead, including delay and energy consumption, while meeting the quality of service requirements, a quality of service-aware resource allocation approach is proposed that jointly communicates between fog nodes and IoT nodes. Transfer and allocation of computing resources are considered optimizing allocation decisions and minimize network overhead. First, an evaluation framework based on a hierarchical process is developed to prioritize the QoS parameters and different IoT tasks. Then, a resource block allocation algorithm is proposed to allocate resources to IoT devices based on device priority, satisfaction level, and resource quality. In addition, a QoS-aware bilateral matching game is introduced to optimize communication between cloud computing nodes and IoT devices. The simulation results show that the proposed method effectively ensures network load balance, improves resource utilization, and reduces network overhead.

New-generation networks have become increasingly important in the transfer of heterogeneous data streams. In fact, besides typical data and multimedia traffic, intelligent IoT applications create new types of traffic and interactions among millions of devices. This traffic creates a scalability issue, particularly regarding resource management and decision-making. The method presented in [29] aims to synthesize a flexible 5G radio frame by packing heterogeneous streams from multiple users into rectangular grids of time-frequency resources. The proposed approach includes the classification of flows based on quality of service, followed by the development of two offline databases.

In [30], a new QoS-aware resource allocation policy based on users' implicit feedback ratings is proposed for mobile edge computing for IoT to overcome service delays. The proposed method selects the user based on previous purchase preferences and implicit feedback from the cluster generated using similarity calculation. Using time-based collaborative filtering, resources are recommended for qualified users according to users' implicit feedback. The selected user receives the resource according to the minimum distance between the user and the resource. In [31], an effective algorithm is presented to solve the resource allocation problem in an IoT environment to minimize communication costs. A multi-layered resource allocation strategy is presented based on the meta-heuristic and data clustering paradigm for reducing latency in IoT applications. Several aspects of the proposed algorithm are discussed in depth, such as the coding of solutions for resource allocation, transitive operators, an objective function, and an analysis of the algorithm's time complexity.

In [32], a cellular-based frequency resource sharing (CFRM) approach is introduced for multi-cell device-to-device connectivity. Each cell consists of two zones, each with a different spectral source assigned to it to minimize interference between neighboring cells. The uplink resources of cellular users are shared between D2D users, thereby reducing the interference from device-to-device users to cellular users. The proposed design of the paper is a frequency multiplexer based on cross-cell FFR in D2D communication. It distributes different frequency resources to system users and cells in the local area while also investigating how well each cell can multiplex resources to a system that maximizes system/network performance. This paper provides the following key contributions: 1) In the cross-cell and FFR-based D2D communication context, an optimization scheme is planned to optimize the network capacity, which supports the SINRs of cell users and devices, and the interference with cell users and devices in comments 2) A Cross-Cell Frequency Resource Sharing (CFRM) strategy for multi-cell D2D communications is developed to enhance network throughput and reduce interference with cell and system users.

B. Context-aware Approaches

Context awareness in ICT refers to considering the state of entities, such as users or devices. Context-aware computing performs well in the IoT due to its importance in handling very large data sets. In [33], In 2, the optimization of channel selection is discussed, which is critical for reliable and efficient task delivery. The proposed approach's primary goal is to maximize the long-term throughput of energy and service reliability limitations. A combination of machine learning, matching theory, and Lyapunov optimization proposes a learning-based channel selection approach with conflict awareness, energy awareness, and service reliability awareness.

In [34], resource allocation for wireless IoT networks with short packet communications is discussed. A wireless IoT network with short packet communications is studied and investigated, where a hybrid access point first wirelessly supplies power to IoT devices. Then the devices transmit their short data, which has disadvantages: the destruction rate and packet error. To increase the efficiency and reliability of transmission, first, the efficiency values and the effective amount of information are defined as parameters to ensure a balance between transmission rate and packet error rate, and then the transmission time and packet error for each user is maximized so that the total throughput is Maximize or minimize the total transfer time that is effective for the user and this is achieved through the algorithms used in this paper.

In [35], the authors presented a method for allocating resources in edge computing for the IoT. Connecting objects to the Internet to make them intelligent is very important, but massive connectivity, big data processing, and significant energy consumption limit the use of the IoT. To address these challenges, a novel architecture for resource allocation at the edge is presented in this paper. Radio resource management and computing resource management in the IoT have also been evaluated in edge computing to improve system performance. The evaluation and review of the conducted studies show that the proposed resource allocation through this method can improve the system's efficiency and minimize the delay.

In [36], the authors have taken advantage of the satisfaction level of the quality of experiments (QoE) to obtain smart center-based services. The innovation of this paper is that it uses reinforcement learning (RL) to achieve superior accuracy in resource allocation. Two algorithms based on reinforcement learning are presented to obtain the cost of mapping tables and optimal resource allocation. The proposed method of the article is entitled "Smart Center-Centered Services for IoT (SCCS-IoT)," which uses an address and route-oriented architecture for communication. The term intelligent in the proposed model is intelligent operation capable of providing the most optimal solution using low-level or cost-effective computing resources. Furthermore, the proposed strategy focuses on two aspects. The first area of interest is the IoT resource allocation problem, and the paper has proposed a new method that uses the RL mechanism to construct the resource allocation strategy. Implementing the RL mechanism is aimed at avoiding conflicts and intelligent resource allocation operations. As a second focus, QoE is considered when constructing RL value functions. A method that incorporates the quality of experience with learning power is being proposed. The cost of each type of task for a computing node comprises a state. The response given by the user determines the status value.

C. SLA-aware Approaches

The heterogeneity and dynamic nature of IoT have made service-level agreements a key component of consumer-provider relationships. Ongoing monitoring of quality-of-service features shall be performed actively to provide this agreement. Furthermore, customers should consider several factors, such as trust (in the provider). In [1], a new mechanism for resource allocation considering buffering, scheduling, and rate limiting methods is proposed to address service level agreement problems.

In [37], the execution time constraint is considered the service level agreement constraint in the hybrid auction system. In the proposed approach, to optimally allocate resources and reduce costs, the winners in each tender round are determined according to the urgency of the tasks and based on the execution time deadline. To evaluate the performance of the suggested mechanism, the resource provider's profit and task completion success rate is compared to existing mechanisms using real workload data.

Collaboration between cloud computing and fog computing has proven extremely effective for resource allocation modeling and service delivery. The authors in [38] have developed a new approach for resource allocation to provide QoS requirements and service level agreement. This algorithm considers three parameters of completion time, service size and virtual machine capacity to manage user requests.

D. Resource Utilization-aware Approaches

These approaches focus on the optimal use of IoT resources. Optimal resource utilization impacts profits and revenue in the IoT. For this purpose, usage-aware resource allocation methods are essential for reducing energy consumption, optimizing resources, and distributing resources fairly. In [39], a hierarchical architecture using a gateway connects IoT devices to eNB to use network resources optimally, and a multi-class resource allocation algorithm is presented for LTE-based IoT communications. The simulation outcomes indicate that the proposed algorithm performs well regarding latency and data rate.

Software-oriented networks are a promising technology for simplifying network management due to the provision of reconfigurable network elements; therefore, integrating this approach and the IoT provides a potentially practical solution to enhance the management and control capabilities of the IoT network. By using software-based networking technology, resource efficiency in the IoT network can be further improved. In [40], the authors have proposed a new architecture for allocating IoT resources based on software-oriented networks. In this article, the resource allocation problem is planned as a Markov decision process, and the optimal solution is obtained using the relative value iteration algorithm.

In [41], a new solution to the resource allocation problem by adopting cooperative game values is presented. In the proposed game model, the concept of Shipley value is developed and used to design a bandwidth allocation algorithm. The results demonstrate that the proposed approach enhances the optimal use of network resources while ensuring performance balance compared to previous methods.

In [42], the fifth-generation mobile communication system is mentioned as an important factor in increasing the importance, value, and increasing use of the Internet of Things, which requires high speed in data transmission, permanent and uninterrupted connection, and very low delay. Therefore, to achieve these requests, this paper presents a new biologically inspired resource allocation method for network slices in the 5th generation IoT with activation capability. Personal service allocation and users' evolutionary interest relationships are used to model the dynamic and complex network, and a biological allocation strategy inspired by nature is used, which is continuously updated. By observing the proposed results, the evidence shows that the proposed method has improved efficiency and flexibility in resource allocation.

E. Energy-aware Approaches

Energy consumption and heat production in data centers are important factors in dealing with challenges related to energy consumption-aware resource allocation techniques. The primary cause of energy consumption and unnecessary heat generation is the increase in the number of servers, rapid data center growth, power loss, high load, and high demand.

Integrating mobile edge computing and IoT enables IoT devices with limited computational and energy capabilities to offload their latency-sensitive computing tasks to the edge of the network, providing quality services to the devices. In [43], a non-orthogonal multiple access technique is used to enable widespread connectivity, and how to use this technique to achieve efficient mobile edge computing in IoT networks is investigated. To maximize the energy efficiency for loading and the maximum tolerable delay limits of IoT devices, the radio and computing resource allocation problem is formulated in which intra- and extra-cellular interferences are considered.

In [44], a 5G-based communication framework supports the physical-cyber deployment of the IoT in a centrally controlled manner. Based on this structure, several actuators and sensors can communicate with the central controller in a two-way fashion. The resource allocation problem is formulated as an exact non-convex programming problem aiming to maximize energy consumption within the available channel band.

Cognitive radio can reduce the spectrum scarcity issue of IoT applications, and wireless energy harvesting eliminates the need for battery charging or replacement for IoT and cognitive radio networks. For this purpose, in [45], the authors have used wireless energy harvesting for cognitive radio, where cognitive radio devices cannot only cooperatively detect available radio frequencies but also collect the wireless energy transmitted by an access point. An optimization framework is proposed to strike a balance between the energy efficiency and spectral efficiency of the network.

In [46], an energy-aware and network density-aware approach to address the resource allocation problem in IoT networks based on hybrid optimization strategies is presented. This paper uses data clustering and meta-heuristic algorithms to reduce congestion between IoT devices and gateways. Furthermore, this paper contributes to the path discovery mechanism by proposing a queue-based collective intelligence optimization algorithm that selects the optimal path for the future path based on multiple constraints.

IV. Discussion

The processing of an application generates a workload on the IoT system. The workload refers to the number of resources required to complete the tasks required by the program. Workload includes the amount of bandwidth and measurement devices consumed by the program and the amount of memory and processing power consumed. In the previous section, the existing methods for the resource allocation problem in the IoT were examined in 5 categories: service quality-aware, context-aware, service-level agreement-aware, resource-use optimization-aware, and energy-aware approaches. In this section, the main features of the approaches are studied, and the important factors that have been improved in the mechanisms are briefly shown in Table I. Increasing the performance of IoT resource allocation is one of the most important goals in the reviewed articles. In addition, the increase in service quality and its use has been examined in some studies. The reduction of energy and time consumption has been investigated in these articles. Optimizing resource allocation approaches can be obtained by guaranteeing availability and reliability and reducing the failure rate in the IoT ecosystem. The objective functions presented in the reviewed works can be stated as follows:

- Reducing the cost of the network: due to the existence of a large volume of services, users, information, and resources in networks based on the Internet of Things, the cost of creating this type of network has become a fundamental challenge; therefore, minimizing network cost is one of the important functions in IoT-based networks.
- Reducing the number of base stations: with the increase in the number of base stations, the consumption cost of the entire network will increase; therefore, adopting the efficient number of base stations is one of the other objective functions.
- Improving network lifespan: In these models, improving network lifespan is achieved by reducing energy consumption and increasing the lifespan of IoT devices.
- Increasing network coverage: In this model, the area covered by base stations increases.
- Increasing the network output: By maximizing the total transmission rate by the base stations, the output, which is one of the most basic parameters of the network, increases.
- Reducing the total mobile power consumption of base stations: if mobile base stations are used, the lifespan of these stations is improved by minimizing their mobile power consumption.
- Reducing the number of stopping points of base stations: in the models that use mobile base stations, telecommunication communication is considered only if the stations are stationary to reduce the complexity. As the number of static points increases, the delay in the network increases.

According to the previously reviewed articles and works, in most of the presented articles, two types of nodes are studied in the proposed system model for resource allocation in the Internet of Things. The first type of these nodes are resource nodes that provide appropriate services. The second node is the gateways that are connected to the resources. Gateways connect different parts of the IoT system. Each gateway is responsible for controlling the path of several resources. Resources can be connected to different gateways. The communication cost between a gateway and all resources is predetermined. One aspect of this problem is that the resources are distributed among the gateways to create the lowest communication cost. Cost between gateways is also an important part of the total cost, so another aspect of this issue is how the gateways are connected.

Due to the high communication cost between gateways, each solution to the problem connects the gateways with the least number of connections. This connection can be linear or loop. Fig. 2 shows an example of connecting resources and gateways. The communication cost is also different with the change in the communication model. One of the goals of the problem is to find a model for resource allocation with the lowest communication cost. This figure can be considered as an outline of the proposed system model. Of course, parts of this system can be slightly changed depending on the designed scenario.

According to Fig. 2, the considered environment includes IoT resources, gateways, and connections between them. These components are deployed in the environment to respond to users' needs and create integrated services. A composite service includes instances of a service formed by establishing connections between IoT resources to provide capabilities and respond to users' needs. Data must be exchanged between IoT resources to create a suitable hybrid service and perform interaction between services.

A service gateway is a computing node that provides computing capabilities to implement service instances and communication capabilities to enhance IoT resources. Each gateway has a resource-binding capacity that can limit the number of IoT resources allocated to the gateway. IoT resources assigned to different gateways can communicate with each other through gateway connections. Possible connections between gateways are determined when they are deployed. Although not all gateways are directly connected to each other, all gateways must be reachable from any other gateway in the IoT environment. Each resource must be assigned a gateway to pass data to other resources to create a hybrid service. The connection between a source and a gateway creates this allocation. Each resource has a list of connectable gateways, and the number of connectable gateways varies by resource type. Connections between gateways should also be specified. Gateways can be connected directly or indirectly through other gateways. Each gateway can directly connect with other gateways within the range of connection capacity. To overcome the limitation of connection capacity, redundancy or ring connections between gateways should be avoided, and connections between gateways should form a tree structure.

TABLE I. A SIDE-BY-SIDE COMPARISON IF IOT RESOURCE ALLOCATION APPROACHES

Category	Reference	Service quality	Delay	Energy consumption	Reliability	Availability	Overhead	Scalability
QoS-aware	[28]	↑	↓	↓			↓	
	[47]	↑	↓				↓	↑
	[30]	↑	↓			↓		
	[31]	↑	↓					
	[32]	↑						↑
Context-aware	[33]		↓	↓	↑			
	[34]			↓	↑	↑	↑	
	[35]	↑	↓	↓				
	[36]	↑		↓				
SLA-aware	[1]						↓	
	[37]	↑						
	[38]		↓		↑			
Resource utilization-aware	[39]	↑	↓					
	[40]				↑			↓
	[41]	↑						↑
	[42]	↑	↓		↑			
Energy-aware	[43]		↓	↓				
	[44]	↑		↓				
	[45]			↓				
	[27]		↓	↓			↓	↑

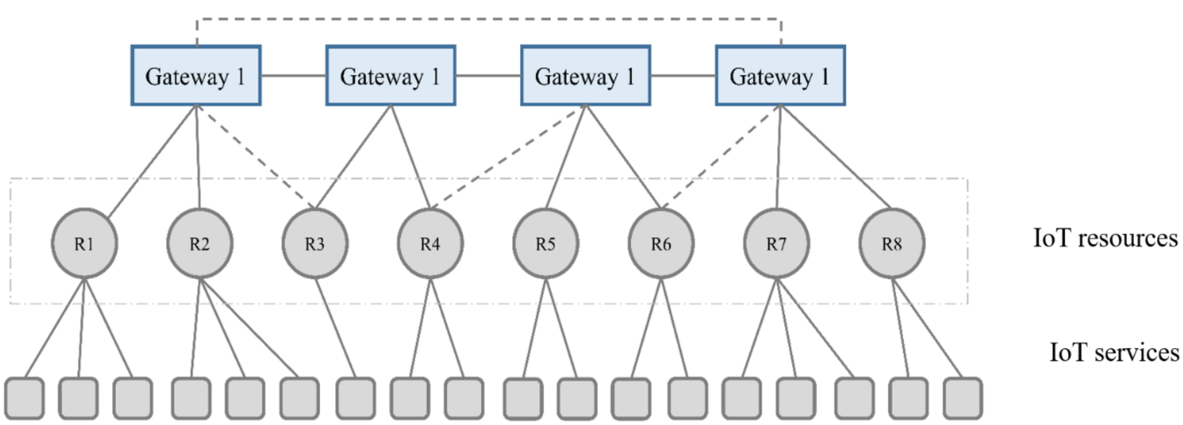

Fig. 2. Resource allocation model.

In the evaluation of the existing resource allocation approaches, it is essential to consider their limitations and assess their suitability for addressing the resource allocation problem in the IoT. While many of the reviewed methods have shown promising results, certain limitations need to be acknowledged. One common limitation is the assumption of homogeneous IoT devices in some approaches, which may not accurately reflect the real-world scenarios where devices exhibit varying capabilities and resource requirements. This limitation can impact the performance and effectiveness of resource allocation algorithms in heterogeneous IoT environments. Another limitation is the neglect of dynamic data traffic patterns in certain methods. The resource allocation algorithms that do not consider the varying data traffic characteristics may not be optimal in scenarios where the data flow changes dynamically, leading to suboptimal resource utilization and potential congestion issues.

Furthermore, some existing methods may lack scalability and fail to handle the increasing number of IoT devices and the growing complexity of IoT deployments. As the IoT ecosystem continues to expand, resource allocation approaches should be capable of accommodating the larger scale and diverse requirements of IoT networks. Additionally, the trade-off between energy consumption and service quality is a critical aspect that needs to be addressed. Some methods may prioritize energy efficiency at the cost of compromising service quality, while others may focus more on providing high-quality services but consume excessive energy. Striking a balance between energy conservation and maintaining satisfactory service quality remains a challenge in resource allocation.

V. FUTURE RESEARCH DIRECTIONS

Future research directions in IoT resource allocation can focus on several key areas to address emerging challenges and further enhance the efficiency and effectiveness of resource allocation strategies. Some potential research directions include:

- Dynamic resource allocation: Develop adaptive resource allocation algorithms that can dynamically adjust resource allocation based on changing network conditions, varying demands, and evolving IoT environments. This can improve resource utilization and optimize the allocation of resources in real-time.
- Energy-efficient resource allocation: Design energy-aware resource allocation techniques that consider the limited energy resources of IoT devices. Explore energy-efficient algorithms that minimize energy consumption while maintaining desired levels of service quality and network performance.
- Security-aware resource allocation: Investigate resource allocation approaches that integrate security considerations into the allocation process. Develop mechanisms to allocate resources in a way that ensures data privacy, confidentiality, and integrity within the IoT ecosystem.
- Multi-objective optimization: Explore multi-objective optimization techniques that consider multiple performance metrics simultaneously, such as energy consumption, network latency, resource utilization, and quality of service. Develop resource allocation algorithms that strike a balance between conflicting objectives to achieve optimal resource allocation outcomes.
- Edge and fog computing: Investigate resource allocation strategies that leverage edge and fog computing capabilities to enhance the efficiency of IoT systems. Explore how resources can be allocated across edge devices and fog nodes to minimize latency, improve response times, and optimize overall system performance.
- Machine learning and AI-based approaches: Explore the use of machine learning and artificial intelligence techniques to optimize resource allocation in IoT networks. Develop intelligent algorithms that can learn from historical data, predict resource demands, and dynamically allocate resources based on real-time conditions.
- Scalability and heterogeneity: Address the challenges of resource allocation in large-scale IoT deployments with heterogeneous devices and varying resource requirements. Develop scalable and adaptive resource allocation algorithms that can handle the complexities of diverse IoT environments.

VI. CONCLUSION

The IoT provides an environment in which physical and digital devices, equipped with identification, processing, diagnosis, and network functions, can communicate through the Internet to achieve a specific goal. In fact, the IoT turns everyday devices into a diverse set of smart objects in order to realize useful applications such as traffic management, energy management, education, finance, and smart transportation. Generally, in IoT-based systems, IoT users use the services available in these types of networks. Service providers offer services on demand to users. Resource allocation strategy as an effective method is used to manage requests and responses. In line with the allocation of resources in the IoT environment, this article presented a background of the IoT and the resource allocation issue. Then the available methods were divided into five categories and examined. Finally, a proposed model for the allocation of IoT resources was presented. The IoT network is often used as a special purpose network according to the application that the network has; therefore, in this network, some nodes (network resources) have a high priority according to the type of data they produce, and therefore these resources should have a high priority in network activities, including resource allocation.

In most of the methods used for resource allocation, all network nodes and their resources are considered with the same priority. This causes network power, network delay, and network resources to be used in a non-optimal way in these networks. Resources with high priority are considered the same as resources with low priority, and optimal resource allocation is not done. Therefore, in the IoT network, for the users of this network to have optimal access to the network resources and to somehow allocate the resources according to the wishes and expectations of the network users and the network requirements, it is better to use the network and also prioritizes the data that the network nodes produce and sent and processed to be determined. For example, in networks used for real-time applications, a series of resources are more important than other network resources in terms of time and delay, so an optimal allocation should be made for these types of resources.

REFERENCES

[1] A. Singh and Y. Viniotis, "Resource allocation for IoT applications in cloud environments," in 2017 International Conference on Computing, Networking and Communications (ICNC), 2017: IEEE, pp. 719-723.

[2] B. Pourghebleh, N. Hekmati, Z. Davoudnia, and M. Sadeghi, "A roadmap towards energy-efficient data fusion methods in the Internet of Things," Concurrency and Computation: Practice and Experience, p. e6959, 2022.

[3] C. Casetti, "5G Standalone Deployments Are on the Rise [Mobile Radio]," IEEE Vehicular Technology Magazine, vol. 18, no. 1, pp. 5-11, 2023.

[4] G. Saravanan, S. Neelakandan, P. Ezhumalai, and S. Maurya, "Improved wild horse optimization with levy flight algorithm for effective task scheduling in cloud computing," Journal of Cloud Computing, vol. 12, no. 1, p. 24, 2023.

[5] S. H. Haghshenas, M. A. Hasnat, and M. Naeini, "A Temporal Graph Neural Network for Cyber Attack Detection and Localization in Smart Grids," arXiv preprint arXiv:2212.03390, 2022.

[6] A. Niazi, R. Amrollahi, and H. Sadeghi, "Design of a High-Efficiency Dual-Helical Antenna for Microwave Plasma Sources," IEEE Transactions on Plasma Science, vol. 50, no. 2, pp. 203-209, 2022.

[7] A. Peivandizadeh and B. Molavi, "Compatible authentication and key agreement protocol for low power and lossy network in IoT environment," Available at SSRN 4194715, 2022.

[8] M. Mohseni, F. Amirghafouri, and B. Pourghebleh, "CEDAR: A cluster-based energy-aware data aggregation routing protocol in the internet of things using capuchin search algorithm and fuzzy logic," Peer-to-Peer Networking and Applications, pp. 1-21, 2022.

[9] V. Srinadh and P. N. Rao, "Implementation of Dynamic Resource Allocation using Adaptive Fuzzy Multi-Objective Genetic Algorithm for IoT based Cloud System," in 2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT), 2022: IEEE, pp. 111-118.

[10] B. Cao, Z. Sun, J. Zhang, and Y. Gu, "Resource allocation in 5G IoV architecture based on SDN and fog-cloud computing," IEEE Transactions on Intelligent Transportation Systems, vol. 22, no. 6, pp. 3832-3840, 2021.

[11] A. K. Sangaiah, A. A. R. Hosseinabadi, M. B. Shareh, S. Y. Bozorgi Rad, A. Zolfagharian, and N. Chilamkurti, "IoT resource allocation and optimization based on heuristic algorithm," Sensors, vol. 20, no. 2, p. 539, 2020.

[12] Y. Kumar, S. Kaul, and Y.-C. Hu, "Machine learning for energy-resource allocation, workflow scheduling and live migration in cloud computing: State-of-the-art survey," Sustainable Computing: Informatics and Systems, vol. 36, p. 100780, 2022.

[13] B. Pourghebleh and N. J. Navimipour, "Data aggregation mechanisms in the Internet of things: A systematic review of the literature and recommendations for future research," Journal of Network and Computer Applications, vol. 97, pp. 23-34, 2017.

[14] M. Sarbaz, M. Manthouri, and I. Zamani, "Rough neural network and adaptive feedback linearization control based on Lyapunov function," in 2021 7th International Conference on Control, Instrumentation and Automation (ICCIA), 2021: IEEE, pp. 1-5.

[15] H. Kosarirad, M. Ghasempour Nejati, A. Saffari, M. Khishe, and M. Mohammadi, "Feature Selection and Training Multilayer Perceptron Neural Networks Using Grasshopper Optimization Algorithm for Design Optimal Classifier of Big Data Sonar," Journal of Sensors, vol. 2022, 2022.

[16] M. Sadi et al., "Special Session: On the Reliability of Conventional and Quantum Neural Network Hardware," in 2022 IEEE 40th VLSI Test Symposium (VTS), 2022: IEEE, pp. 1-12.

[17] S. Yumusak, S. Layazali, K. Oztoprak, and R. Hassanpour, "Low-diameter topic-based pub/sub overlay network construction with minimum–maximum node degree," PeerJ Computer Science, vol. 7, p. e538, 2021.

[18] H. Seraji, R. Tavakkoli-Moghaddam, S. Asian, and H. Kaur, "An integrative location-allocation model for humanitarian logistics with distributive injustice and dissatisfaction under uncertainty," Annals of Operations Research, vol. 319, no. 1, pp. 211-257, 2022.

[19] H. Kashgarani and L. Kotthoff, "Is algorithm selection worth it? Comparing selecting single algorithms and parallel execution," in AAAI Workshop on Meta-Learning and MetaDL Challenge, 2021: PMLR, pp. 58-64.

[20] S. Aghakhani, A. Larijani, F. Sadeghi, D. Martín, and A. A. Shahrakht, "A Novel Hybrid Artificial Bee Colony-Based Deep Convolutional Neural Network to Improve the Detection Performance of Backscatter Communication Systems," Electronics, vol. 12, no. 10, p. 2263, 2023.

[21] B. M. Jafari, M. Zhao, and A. Jafari, "Rumi: An Intelligent Agent Enhancing Learning Management Systems Using Machine Learning Techniques," Journal of Software Engineering and Applications, vol. 15, no. 9, pp. 325-343, 2022.

[22] S. Habib, S. Aghakhani, M. G. Nejati, M. Azimian, Y. Jia, and E. M. Ahmed, "Energy management of an intelligent parking lot equipped with hydrogen storage systems and renewable energy sources using the stochastic p-robust optimization approach," Energy, p. 127844, 2023.

[23] F. Kamalov, B. Pourghebleh, M. Gheisari, Y. Liu, and S. Moussa, "Internet of Medical Things Privacy and Security: Challenges, Solutions, and Future Trends from a New Perspective," Sustainability, vol. 15, no. 4, p. 3317, 2023.

[24] B. Sellami, A. Hakiri, S. B. Yahia, and P. Berthou, "Energy-aware task scheduling and offloading using deep reinforcement learning in SDN-enabled IoT network," Computer Networks, vol. 210, p. 108957, 2022.

[25] B. Pourghebleh and V. Hayyolalam, "A comprehensive and systematic review of the load balancing mechanisms in the Internet of Things," Cluster Computing, pp. 1-21, 2019.

[26] B. Pourghebleh, V. Hayyolalam, and A. A. Anvigh, "Service discovery in the Internet of Things: review of current trends and research challenges," Wireless Networks, vol. 26, no. 7, pp. 5371-5391, 2020.

[27] K. Praveen and P. J. Prathap, "Energy efficient congestion aware resource allocation and routing protocol for IoT network using hybrid optimization techniques," Wireless Personal Communications, vol. 117, no. 2, pp. 1187-1207, 2021.

[28] X. Huang, Y. Cui, Q. Chen, and J. Zhang, "Joint task offloading and QoS-aware resource allocation in fog-enabled Internet-of-Things networks," IEEE Internet of Things Journal, vol. 7, no. 8, pp. 7194-7206, 2020.

[29] Y. Boujelben, "Scalable and QoS-Aware Resource Allocation to Heterogeneous Traffic Flows in 5G," IEEE Internet of Things Journal, vol. 8, no. 20, pp. 15568-15581, 2021.

[30] P. Das, A. R. Jamader, B. R. Acharya, and H. Das, "HMF Based QoS aware Recommended Resource Allocation System in Mobile Edge Computing for IoT," in 2019 International Conference on Intelligent Computing and Control Systems (ICCS), 2019: IEEE, pp. 444-449.

[31] C.-W. Tsai, "SEIRA: An effective algorithm for IoT resource allocation problem," Computer Communications, vol. 119, pp. 156-166, 2018.

[32] Y. Li, Y. Liang, Q. Liu, and H. Wang, "Resources allocation in multicell D2D communications for Internet of Things," IEEE Internet of Things Journal, vol. 5, no. 5, pp. 4100-4108, 2018.

[33] H. Liao et al., "Learning-based context-aware resource allocation for edge-computing-empowered industrial IoT," IEEE Internet of Things Journal, vol. 7, no. 5, pp. 4260-4277, 2019.

[34] J. Chen, L. Zhang, Y.-C. Liang, X. Kang, and R. Zhang, "Resource allocation for wireless-powered IoT networks with short packet communication," IEEE Transactions on Wireless Communications, vol. 18, no. 2, pp. 1447-1461, 2019.

[35] S. Li et al., "Joint admission control and resource allocation in edge computing for internet of things," IEEE Network, vol. 32, no. 1, pp. 72-79, 2018.

[36] K. Gai and M. Qiu, "Optimal resource allocation using reinforcement learning for IoT content-centric services," Applied Soft Computing, vol. 70, pp. 12-21, 2018.

[37] Y. Choi and Y. Lim, "Optimization approach for resource allocation on cloud computing for iot," International Journal of Distributed Sensor Networks, vol. 12, no. 3, p. 3479247, 2016.

[38] A. A. Alsaffar, H. P. Pham, C.-S. Hong, E.-N. Huh, and M. Aazam, "An architecture of IoT service delegation and resource allocation based on collaboration between fog and cloud computing," Mobile Information Systems, vol. 2016, 2016.

[39] J. Li, Q. Sun, and G. Fan, "Resource allocation for multiclass service in IoT uplink communications," in 2016 3rd International Conference on Systems and Informatics (ICSAI), 2016: IEEE, pp. 777-781.

[40] X. Xiong, L. Hou, K. Zheng, W. Xiang, M. S. Hossain, and S. M. M. Rahman, "SMDP-based radio resource allocation scheme in software-defined Internet of Things networks," IEEE Sensors Journal, vol. 16, no. 20, pp. 7304-7314, 2016.

[41] S. Kim, "Asymptotic shapley value based resource allocation scheme for IoT services," Computer Networks, vol. 100, pp. 55-63, 2016.

[42] D. Wu, Z. Zhang, S. Wu, J. Yang, and R. Wang, "Biologically inspired resource allocation for network slices in 5G-enabled Internet of Things," IEEE Internet of Things Journal, vol. 6, no. 6, pp. 9266-9279, 2018.

[43] B. Liu, C. Liu, and M. Peng, "Resource allocation for energy-efficient MEC in NOMA-enabled massive IoT networks," IEEE Journal on Selected Areas in Communications, vol. 39, no. 4, pp. 1015-1027, 2020.

[44] S. Li, Q. Ni, Y. Sun, G. Min, and S. Al-Rubaye, "Energy-efficient resource allocation for industrial cyber-physical IoT systems in 5G era," IEEE Transactions on Industrial Informatics, vol. 14, no. 6, pp. 2618-2628, 2018.

[45] A. Shahini, A. Kiani, and N. Ansari, "Energy efficient resource allocation in EH-enabled CR networks for IoT," IEEE Internet of Things Journal, vol. 6, no. 2, pp. 3186-3193, 2018.

[46] K. Praveen and P. Prathap, "Energy Efficient Congestion Aware Resource Allocation and Routing Protocol for IoT Network using Hybrid Optimization Techniques," Wireless Personal Communications, vol. 117, no. 2, pp. 1187-1207, 2021.