Hindawi
Complexity
Volume 2020, Article ID 8964165, 16 pages
https://doi.org/10.1155/2020/8964165

Review Article

Study QoS Optimization and Energy Saving Techniques in Cloud, Fog, Edge, and IoT

Zhiguo Qu,[1,2] Yilin Wang,[1,2] Le Sun,[1,2] Dandan Peng,[1,2] and Zheng Li[1,2]

[1]*Engineering Research Center of Digital Forensics, Ministry of Education, Nanjing, China*
[2]*Jiangsu Collaborative Innovation Center of Atmospheric Environment and Equipment Technology (CICAEET), Nanjing University of Information Science & Technology, 210044 Nanjing, China*

Correspondence should be addressed to Le Sun; sunle2009@gmail.com

Received 6 December 2019; Revised 22 January 2020; Accepted 7 February 2020; Published 16 March 2020

Guest Editor: Xuyun Zhang

With an increase of service users' demands on high quality of services (QoS), more and more efficient service computing models are proposed. The development of cloud computing, fog computing, and edge computing brings a number of challenges, e.g., QoS optimization and energy saving. We do a comprehensive survey on QoS optimization and energy saving in cloud computing, fog computing, edge computing, and IoT environments. We summarize the main challenges and analyze corresponding solutions proposed by existing works. This survey aims to help readers have a deeper understanding on the concepts of different computing models and study the techniques of QoS optimization and energy saving in these models.

1. Introduction

With the development of the Internet, more and more computing techniques are developed. In this situation, an increasing amount of data needs to be processed. The increase of users' requirements causes the development of different types of computing models, such as cloud computing, fog computing, and edge computing.

Cloud computing is an early computing model that has made great contributions to data processing. It provides convenient and quick network access to shared configurable resources, such as networks and servers. In addition, provisioning and publishing these resources do not require much administration and interaction of service providers [1]. The structure of cloud computing is shown in Figure 1.

Due to the development of the IoT and the increasing needs of people, the IoT system based on cloud computing faces some limitations. In this situation, cloud computing cannot play a good role in large-scale or heterogeneous conditions [3]. Therefore, a new computing model called *fog computing* is developed on the basis of cloud computing. Compared with cloud computing, the main advantage of fog computing is that it extends cloud resources to the network edge. Therefore, fog computing can facilitate the management of resources and services [4]. The structure of fog computing is shown in Figure 2.

Edge computing allows operations to be performed on the edge of a network [2]. Edge computing refers to all the resources of computing and network from data sources to cloud data centers. In edge computing, the flow of computing is bidirectional and things in edge computing can both consume data and produce data. That is, they can not only ask the cloud for services but also carries out computing jobs in the cloud [2]. The structure of edge computing is shown in Figure 3.

The most popular embodiment of edge computing is the MEC, which refers to the technology of performing computation-intensive and delay-sensitive tasks for mobile devices. And its theory is collecting a large amount of free computing power and storage resources located at the edge of a network. The European Telecommunication Standards Institute was the first to define it as a computing model. MEC provides the capabilities of information technology and cloud computing at the network edge.

The IoT is created by the diffusion of sensors, actuators, and other devices in the communication driven network. The development of wireless technologies, such as the wireless sensor network technology and actuator nodes, promotes the development of the IoT technology. With the

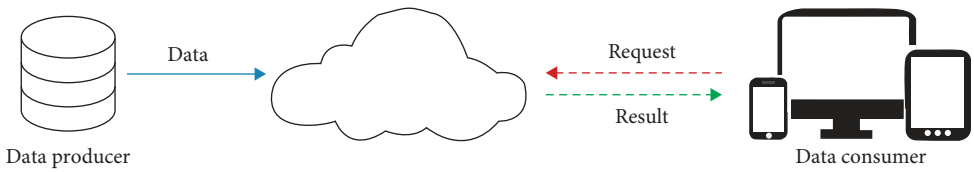

Figure 1: Structure of cloud computing [2].

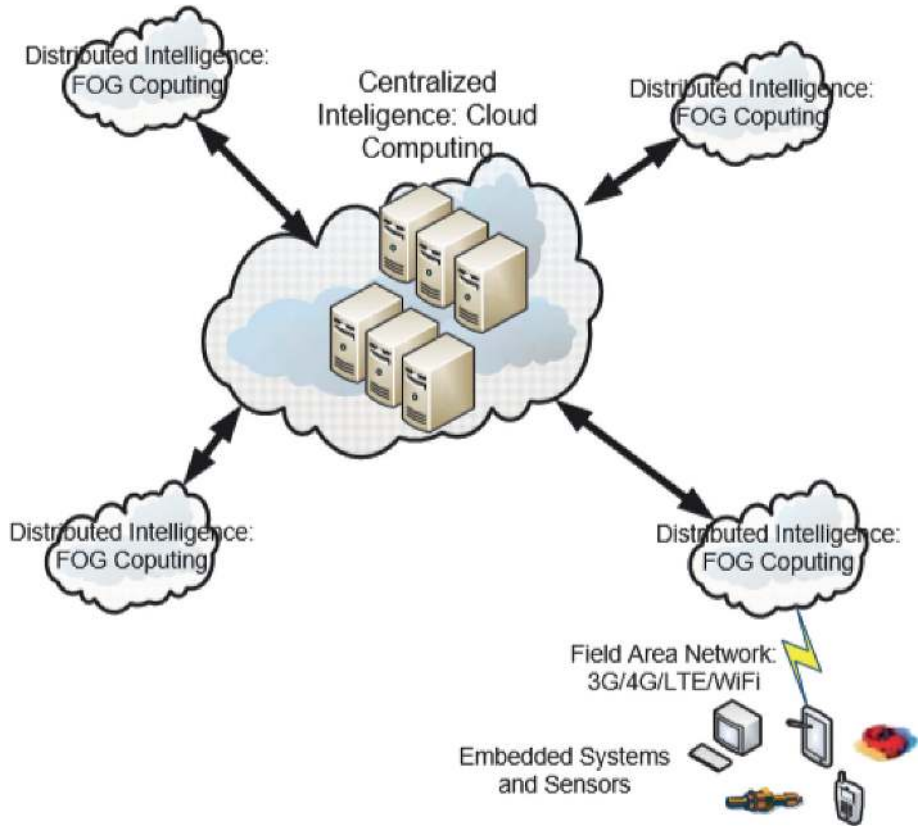

Figure 2: Structure of fog computing [5].

development of the IoT, its application has gradually expanded to cover increasingly wider domains. However, it always aims to make computers perceive information [6].

This paper investigates the important papers related to these computing models. For each paper, we point out the problems it aims to solve and introduce the solutions it proposes. The main contribution of this paper is as follows: (1) do a comprehensive survey on the techniques of QoS optimization and energy saving in different computing models, (2) classify papers according to the problems solved by the reviewed works, and (3) compare and summarize the main features of each type of paper. The structure of this paper is Section 2 studies five energy saving techniques under different computing models, and Section 3 concludes this paper.

2. QoS Optimization and Energy saving Techniques in Different Computing Models

In this section, we introduce the main works of QoS optimization and energy saving techniques in different computing models. We categorize these works in terms of the means they use to achieve the objective of QoS optimization and energy saving, which are (1) quality of service (QoS) guarantee or service-level agreement (SLA) assurance, (2) resource management and allocation, (3) scientific workflow execution, (4) server optimization, and (5) load balancing.

2.1. QoS Guarantee or SLA Assurance. Improving QoS or reducing SLA violations can effectively guarantee the transmission bandwidth, reduce the transmission delay, and reduce the packet loss rate of data. Striking a balance between QoS and limited resources can achieve energy saving.

2.1.1. Cloud Computing. Mazzucco et al. [7] let cloud service providers get the maximum benefit by reducing power consumption. In addition, they introduced and evaluated the policy of dynamic allocation of powering servers' switches. It can optimize users' experience while consuming

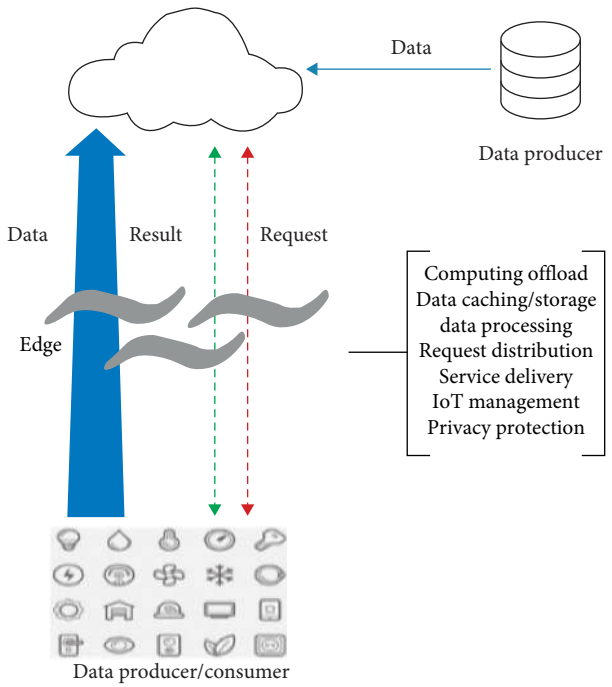

Figure 3: Structure of edge computing [2].

the least amount of power. He et al. [8] proposed a service-based system supporting keyword search, in which different search keywords represent different tasks. This method can help unprofessional service users build service-oriented systems. Sun et al. [9] proposed a cloud service selection method to measure and aggregate the nonlinear relationship between standards. And a framework based on priority is designed to determine the criteria relationships and weights when historical information is insufficient. Mazzucco and Dyachuk [10] were also committed to making cloud service providers obtain the largest profits. They proposed the dynamic distribution strategy of powering server switch. The strategy not only enables users to get good service but also reduces power consumption. The number of live servers determines the state of the system, but running or closing a server cannot be done in a flash. So it is important to take into account of the time. Given the short time required for the server switch, formula (1) [10] represents the cost of changing the number of servers running per unit time. In order to make users have a good experience, this paper further uses a forecasting method to accurately predict the users' time-changing needs:

$$Q = \frac{\Delta n}{t}\left(\sum_{i=1}^{l} d_i + kre_3\right), \tag{1}$$

where t represents the observation time, Δn represents the number of servers whose state change over time, d_i represents the cost of the state change of a hardware component, e_3 represents the energy consumed in a unit time to change the state, k represents the average time to change the state of a server, and l represents the amount of component.

Mazzucco et al. [7] and Mazzucco and Dyachuk [10] both explore strategies of reducing the power cost of running a data center and changing the on-off state of the servers. The two strategies can maximize the users' experience and save energy at the same time. Their difference is that Mazzucco and Dyachuk [10] believes that it is impossible to accurately predict the changes of users' needs over time. So, compared with paper [7], the strategy proposed in paper [10] is fault-tolerant. He et al. [11] proposed three service selection methods that support QoS and can combine multitenant service-based systems. These three methods can achieve three degrees of multitenant maturity, which is more efficient than the traditional single-user approach. Sun et al. [12] proposed a unified semantic model to describe cloud service. This model expands the basic structure of unified service description language. And it defines a transaction module to model the rating system for cloud services from various perspectives. So, it can improve the ability of the model on service ranking. In addition, an annotation system is put forward to enrich the language expression. Wang et al. [13] proposed a fault-tolerant strategy based on multitenant service criticality, which can provide redundancy for key component services, evaluating the criticality of each component service to determine the optimal fault-tolerant policy. Therefore, the quality of the multitenant based service system can be guaranteed. Mustafa et al. [14] leveraged the notion of workload consolidation to improve energy efficiency by putting incoming jobs on as few servers as possible. The concept of SLA is also imported to minimize the total SLA violations. Given that a change in workload changes the utilization of CPU required over time. So, an integral function (formula (2) [14]) is used to represent the total energy (E) consumed by a server (S) operation:

$$E = \int_{t0}^{t1} P(u(t))\mathrm{d}t, \tag{2}$$

where P is the amount of power consumed by the server in terms of CPU utilization (u) in time t.

Bi et al. [15] established an architecture that can administrate itself in cloud data centers firstly. The architecture is suitable for web application services with several levels and has virtualization mechanism. Then, a mixed queuing model is proposed to decide the number of virtual machines (VMs) in each layer of application service environments. Formula (3) [15] is used to represent the local profit that can be made by the ith virtualization application service environment. Finally, a problem of misalignment restrained optimization and a heuristic mixed optimization algorithm are proposed. Both of them can make more revenues and meet requirements of different customers:

$$P_i(E) = \mathrm{Revenue}(E) + \mathrm{Penalty}(E) + \mathrm{Loss}(E) + \mathrm{Cost}(E), \tag{3}$$

where $\mathrm{Revenue}(E)$, $\mathrm{Penalty}(E)$, $\mathrm{Loss}(E)$, and $\mathrm{Cost}(E)$, respectively, represent the total benefit, penalty, loss, and cost of VMs.

Singh et al. [16] proposed a technology named STAR which can manage resources itself in the cloud computing

environment and reduce SLA violations. So, the payment efficiency of cloud services can be improved. Beloglazov and Buyya [17] proposed a system to manage energy in the cloud data center. By continuously integrating VMs and dynamically redistributing VMs, the system can achieve the goal of saving energy and providing a high QoS level at the same time. Guazzone et al. [18] proposed an automatic management system (see Figure 4) for resources to provide certain QoS levels and reduce energy consumption. Resource manager of the framework in Figure 4 combines virtualization technologies and control-theoretic technologies. Virtualization technologies deploy each application to independent VM. And control-theoretic technologies realize the automatic management of computer performance and energy consumption. In addition, the resource manager consists of several independent components named Application Manager, Physical Machine Manager, and Migration Manager. Different from traditional static methods, this method can both fit the changing workloads dynamically and achieve remarkable results in reducing QoS violations. Sun et al. [19] established a model to simplify the decision of cloud resource allocation and realize the independent allocation of resources. The optimal resource configuration can be obtained, so the QoS requirements can be well met. Siddesh and Srinivasa [20] explored the problems of dynamic resource allocation and SLA assurance. They proposed a framework to deal with heterogeneous workload types by dynamically planning computing capacity and assessing risks. The framework uses scheduling methods to reduce SLA violation risks and maximize revenues in resource allocation.

Garg et al. [21] proposed a resource allocation strategy for VM dynamic allocation. The strategy can improve resource utilization, increase providers' profits, and reduce SLA violations. Jing et al. [22] proposed a new dynamic allocating technique using the mixed queue model, meeting customers' different requirements of performance by providing virtualized resources to each layer of virtualized application services. All these methods can reasonably configure resources in the cloud data center, improve system performance, reduce additional costs of using resources, and meet the required QoS.

Qi et al. [23] proposed a QoS-aware VM scheduling strategy named QVMS to satisfy QoS. Firstly, the scheduling problem is transformed into a problem with several objectives. And then the optimal VM migration method is found according to the genetic algorithm. The scheduling strategy can effectively manage resources in the network physical system, thus reducing the energy consumption and improving QoS levels. Qi et al. [25] proposed a service recommendation strategy by considering the time factor to improve the traditional location-sensitive hash technology. The policy emphasizes the influence of dynamic factors on QoS and the protection of user privacy.

Table 1 shows a summary of the abovementioned works. The solution of the problems in Table 1 can improve QoS in cloud computing environment. Server management refers to dynamically allocating powering servers' switches. Workloads consolidation refers to combining work to save energy. VM management refers to reasonable scheduling or integration of VMs to achieve better performance. Self-management refers to the realization of self-management of resources, which can achieve higher efficiency. Resource management refers to the correct allocation of resources to reduce waste. Service management is about making reasonable service choices [26].

2.1.2. Fog Computing. Gu et al. [27] used fog computing to process a large amount of data generated by medical devices and built Fog Computing Supported Medical Cyber-Physical System (FC-MCPS). In order to reduce the cost of FC-MCPS, research studies were carried out on the joint of base station, task assignment, and VM layout. The problem is modeled as a mixed integer linear programming (MILP). A two-stage heuristic algorithm based on linear programming (LP) is proposed to solve the problem. Ni et al. [28] proposed a resource allocation approach based on fog computing, which enables users to select resources independently. In addition, this approach takes into account the price and time required to finish the job. Formula (4) [28] is used to define the credibility $\mathrm{BC}_r e_{ij}$ of Resource R_j received from user_i, when the user interacts with Resource R_i:

$$\mathrm{BC}_r e_{ij} = \omega_1 \lambda_r \mathrm{esp} + \omega_2 \gamma_e \mathrm{xec} + \omega_3 \left(1 - \eta_r \mathrm{eboot}\right) + \omega_4 \mu_r \mathrm{el}, \tag{4}$$

where the value of $\omega_k \epsilon [0, 1]$, $\sum_{k=1}^{4} \omega_k = 1$, which can be determined by the user or the actual situation, $\lambda_r$esp, $\gamma_e$xec, $\eta_r$eboot, and $\mu_r$el are the response speed of the corresponding index service, the efficiency of execution, the speed of restart, and the reliability, respectively.

2.1.3. Edge Computing. Wei et al. [29] proposed a unified framework in the sustainable edge computing to save energy, including the energy that is distributed and renewable. And the architecture can combine the system that supply energy and edge services, which can make full use of renewable energy and provide better QoS. Lai et al. [30] proposed an optimized allocation method for edge users. The method can not only maximize the amount of resources allocated to users but also consider the dynamic QoS level of users. So, edge user allocation problem can be made more general and improving the quality of experience.

2.1.4. MEC. Xu et al. [31] used block chain to improve the traditional crowdsourcing technology. Firstly, they proposed a mobile crowdsourcing framework using block chain technology to protect user privacy. Then, they used dynamic programming strategy of clustering algorithm to classify requesters. Finally, they generated service policies to balance profits and energy consumption.

2.1.5. IoT. Rolik et al. [32] proposed a method to build a framework of IoT infrastructure based on microcloud. The method can help users use resources rationally, reduce the cost of managing infrastructure, and improve the quality of

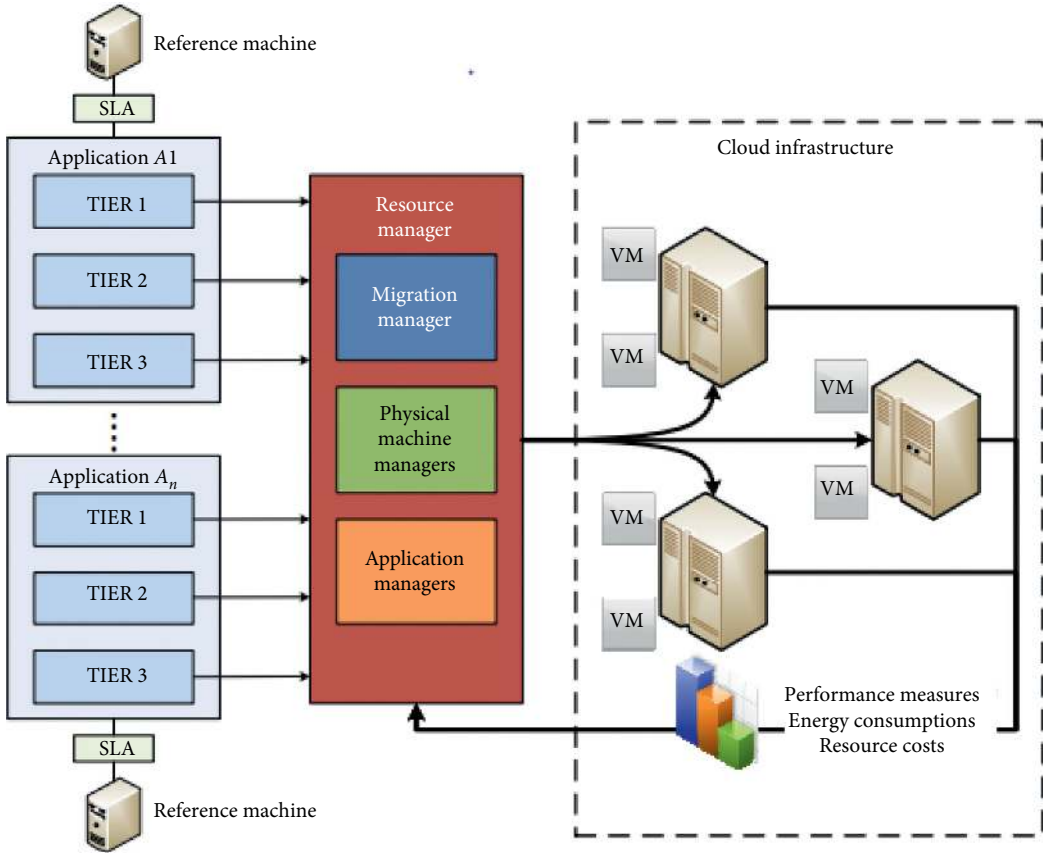

Figure 4: Framework of the proposed three-fold system [18].

life of consumers. He et al. [33] proposed a dynamic network slice strategy. The network slice can be dynamically adjusted according to the time-varying resource demands. This method can improve the utilization of the underlying resources and better meet different QoS demands. Yao and Ansari [34] proposed an algorithm to determine the number of VMs to be rented and to control the power supply. Thus, the cost of the system can be minimized and the QoS can be improved. Formula (5) [34] is used to limit the delay requirement of QoS. The total delay must not exceed the computation deadline of each task, and the total delay is composed of wireless transmission delay and fog processing delay:

$$t_i^c + t_i^w \le D_i, \quad \forall i \epsilon N, \tag{5}$$

where c and w, respectively, represents fog processing and wireless transmission, i denotes a location, t_i^c represents the delay of processing, t_i^w represents the delay of wireless transmission, D_i denotes the deadline, and N denotes different locations.

2.2. Resource Management and Allocation. Rational allocation of resources is an effective means to save energy.

2.2.1. Cloud Computing. Wang et al. [35] introduced an allocation method based on distributed multiagent to allocate VMs to physical machines. The method can realize VM consolidation and consider the migration costs simultaneously. In addition, a VM migration mechanism based on local negotiation is proposed to avoid unnecessary VM migration costs. Hassan et al. [37] established a formulation of universal problem and proposed a heuristic algorithm which has optimal parameters. Under this formulation, dynamic resource allocation can be made to meet the QoS requirements of applications. And the cost needed for dynamic resource allocation can be minimized with this algorithm. Wu et al. [38] proposed a scheduling algorithm based on the technology that can scale the voltage frequency dynamically in cloud computing. The algorithm can allocate resources for performing tasks and realize low power consumption network infrastructure. Compared with other schemes, this scheme not only sacrifices the performance of execution operations but also saves more energy.

Sarbazi and Zomaya [45] used two job consolidation heuristic methods to save energy. One is MaxUtil to better utilize resources and the other is Energy-Conscious Task Consolidation to reduce energy consumption. These two methods can promote the concurrent execution of multiple tasks and improve the energy efficiency. Hsu et al. [46] proposed a job consolidation technique to minimize energy consumption. Formula (6) [46] defines the energy consumption of VM V_i from time t_0 to t_m in the cluster is defined. And formula (7) [46] defines the total energy consumption in a virtual cluster VC_k, in the period. In addition, the proposed technique limits the CPU usage and

Table 1: Work summary of QoS guaranteeing or SLA assurance in cloud computing.

Subproblems	Solutions	Literatures	Advantages
Server management	A policy of dynamic allocation of powering servers switches	[7, 10]	Maximizes benefits, improves QoS, and minimizes power consumption
Workloads consolidation	A technique for consolidating workloads	[14]	Achieves energy savings by using fewest servers while reducing SLA violations
VM management	An architecture that can administrate itself and a mixed queuing model	[15]	Makes more revenues and meets different requirements of customers and decides the number of VMs for each layer of a virtual application
	A system to integrate and dynamically redistribute VMs	[17]	Achieves the goal of saving energy through VM integration and provides a high QoS level at the same time
	A QoS-aware VM scheduling strategy named QVMS	[23]	Effectively manages resources in the network physical system to reduce the energy consumption and improves QoS
	A VM integration method with several targets	[24]	Saves energy and reduces SLA violations by applying different strategies to different load states of the host
Self-management	A technology named STAR	[16]	Reduces SLA violations and improves payment efficiency of cloud services
	A dynamic resource management system	[18]	Self-manages the resources of cloud infrastructures to provide appropriate QoS and fits the changing workloads dynamically
	A model that can realize the independent allocation of resources	[19]	Obtains the optimal resource configuration, meets the QoS requirements, and provides economical cloud resources
Resource management	A dynamic resource allocation strategy	[20, 21]	Reduces SLA and maximizes revenues and resource utilization on the cloud
	A mixed queue model	[22]	Reasonably configures the resources in the cloud data center, improves the system performance, reduces the additional cost of using resources, meets the required QoS, and provides virtual resources to each layer of virtual application services
Service management	A unified semantic model that can describe cloud service	[12]	Improves the ability of model on service ranking and enriches the language expression
	A recommendation service strategy	[25]	Emphasizes the influence of time factors on QoS and improves the traditional location-sensitive hash technology to protect users' privacy
	A cloud service selection method using fuzzy measure and Choquet integral and a framework based on priority	[9]	Selects service when historical information is insufficient to determine the criteria relationships and weights
	Three service selection methods that support QoS and can combine multitenant service-based systems	[11]	Achieves three degrees of multitenant maturity, which is more efficient than the traditional single-user approach
	A fault-tolerant strategy based on multitenant service criticality	[13]	Guarantees the quality of the multitenant-based service system
	A service-based system supporting keyword search	[8]	Effectively helps system engineers who are not familiar with service-oriented architecture technology to build service-oriented systems

merges tasks in virtual clusters. Once a task migration happens, the energy cost model will take into account the network latency. Sarbazi-Azad and Zomaya [45] and Hsu et al. [46] both maximize the benefit of cloud resources by using task merging techniques. Sarbazi-Azad and Zomaya [45] uses a greedy algorithm called MaxUtil. While, Hsu et al. [46] takes into account the network latency associated with task migration. So, in [46], a 17% improvement is achieved over MaxUtil:

$$E_{0,m}(V_i) = \sum_{t=0}^{m} E_t(V_i), \tag{6}$$

$$E_{0,m}(\mathrm{VC}_k) = \sum_{i=0}^{n} E_{0,m}(V_i), \tag{7}$$

where E_t is the energy consumption in unit time and n is the number of VMs in the cluster.

Hsu et al. [47] proposed a task integration technology based on the energy perception. According to the characteristics of most cloud systems, the principle of using 70% CPU is proposed to administrate job integration among virtual clusters. This technology is very effective in reducing the amount of energy consumed in cloud systems by merging tasks. Panda and Jana [48] proposed an algorithm with several criteria to combine tasks. The algorithm not only considers the time needed for processing jobs but also considers the utilization rate of VMs. And the algorithm is more energy efficient because it takes into account not only the processing time but also the utilization rate of VMs. Wang and Su [39] proposed a resource allocation algorithm to deal with wide range of communication between nodes in cloud environment. This algorithm uses recognition technology to dynamically distribute jobs and nodes according to computing ability and factors of storage. And it can reduce the traffic when allocating resources because it uses dynamic hierarchy. Lin et al. [40] proposed a dynamic auction approach for resource allocation. The approach can ensure that even if there are many users and resources, providers will have reasonable profits and computing resources will be allocated correctly. Yazir et al. [41] proposed a new method to manage resources dynamically and autonomously. Firstly, resource management is split into jobs and each job is executed by autonomous nodes. Second, autonomous nodes use the method called PROMETHEE to configure resources. Krishnajyothi [36] proposed a framework which can implement parallel task processing to solve the problem of low efficiency when submitting large tasks. Compared with the static framework, this framework can dynamically allocate VMs, thus reducing costs and the time of processing tasks. Lin et al. [42] proposed a method to allocate resources dynamically by using thresholds. Because this method uses the threshold value, it can optimize the reallocation of resources, improve the usage of resources, and reduce the cost. Xu et al. [43] proposed a data placement strategy named IDP for the data generated by IoTs devices to achieve reasonable data placement. In this way, the privacy of these data can be protected while resources are allocated reasonably. Jo et al. [44] proposed a computing offload framework under 5G network. The framework transfers the computing burden to the cloud, thus reducing the computing load of clients and the communication cost.

Table 2 shows a summary of the abovementioned works. The problem of resource allocation and management in cloud computing can be divided into problems in Table 2. VM management is about a reasonable configuration of VMs. Resource allocation represents the dynamic and flexible allocation of resources. Task integration refers to combining tasks to save energy and improve efficiency.

2.2.2. Fog Computing. Yin et al. [49] established a new model of scheduling jobs, which applies containers. In order to make sure that jobs can be finished on time, a job scheduling algorithm is developed. The algorithm can also optimize the number of tasks that can be performed together on the nodes in fog computing. And this paper proposes a redistribution mechanism to shorten the delay of tasks. These methods are very effective in reducing task delays. Aazam and Huh [50] established a framework to administrate resources effectively in the mode of fog computing. Considering that there are various types of objects and devices, the connection between them may be volatile. So, a method for predicting and administrating resources is proposed. The method considers that any object or device can quit using resources at anytime. Cuong et al. [5] studied the allocating resources jointly problem and the problem of carbon footprint minimization in fog data center. Formula (8) [5] is used to denote the energy consumption of servers. In addition, a distributed algorithm is proposed to solve the problem of wide range optimization:

$$P(y) = C \cdot P_{\mathrm{idle}} + \left(P_{\mathrm{peak}} - P_{\mathrm{idle}}\right) \cdot y \cdot \kappa, \tag{8}$$

where $P(y)$ represents the power supply required by the servers in a data center; y represents the video stream; κ denotes a conversion factor that converts the video stream into workload; C represents the data center's load capacity; and P_{idle} and P_{peak}, respectively, represent the idle power and peak power of the servers.

Jia et al. [51] studied the problem of computing resource allocation in fog computing network with three levels. Firstly, the problem of resource allocation is transformed into a bilateral matching optimal problem. And then a bimatching approach is proposed for this problem, which can improve the performance of the system and obtain higher cost efficiency. Zhang et al. [52] proposed a framework for joint optimization under fog computing to allocate fog nodes' finite computing resources. The framework can achieve the best allocation and effectively improve the networks' performance. Tan et al. [53] presented a method to allocate computing and communication resources. The method transfers computing jobs to remote cloud and nodes and simplifies edge nodes' computing and computing energy. Vasconcelos et al. [54] developed a platform to allocate resources accessible to client devices in fog computing environment, allocating the resources of devices near the host to meet the applications needs for rapid response to computing resources. Aazam et al. [55] presented a method

Table 2: Work summary of resource allocation and management in cloud computing.

Subproblems	Solutions	Literatures	Advantages
VM management	A VM allocation method based on several distributed agents	[35]	Centralizes VMs to physical machines and reduces the overall energy cost
	A framework which can implement parallel task processing	[36]	Dynamically allocates VMs for large tasks
Resource allocation	A general problem formula and a heuristic algorithm with optimization parameters	[37]	Dynamically allocates resources according to QoS requirements and realizes energy saving by optimizing the number of servers
	A scheduling algorithm based on the technology that can scale the voltage frequency dynamically	[38]	Ensures the performance of executing jobs while implementing green computing
	A fuzzy pattern recognition technology	[39]	Reduces the traffic when allocating resources and uses dynamic hierarchy
	A dynamic auction approach	[40]	Guarantees profits of providers and allocates resources correctly when there are a large amount of users and resources
	PROMETHEE	[41]	Allows node agents to decompose and execute tasks autonomously to improve the flexibility of resource allocation
	A method using thresholds	[42]	Optimizes resource reallocation, improves resource usage, reduces cost, and studies resource allocation strategies at the application level
	IoT-oriented data placement (IDP) method	[43]	Protects data privacy, allocates resources reasonably, and focuses on the placement method of IoT data
	A computing offload framework under 5G network	[44]	Reduces the computing load of clients and the communication cost
Task consolidation	Two job consolidation heuristics named "MaxUtil" and energy-conscious task consolidation	[45]	Promotes concurrent execution of multiple tasks and improves energy efficiency
	A job consolidation technique aiming at energy saving	[46, 47]	Reduces the power consumption of cloud system, protects data privacy, allocates resources reasonably, limits CPU usage, and merges tasks in virtual clusters
	An algorithm with several criteria	[48]	Takes into account the job processing time and the VM utilization rate simultaneously. Dramatically reduces energy consumption compared with state-of-the-art works

to estimate and manage resource in fog computing. The method is based on the fluctuation of customer abandonment probability, type and price of service, and so on.

Table 3 shows a summary of the abovementioned works. The problems in Table 3 are also derived from resource allocation and management problems. Task allocation represents the scheduling and redistribution of tasks. Resource allocation is still about the dynamic and flexible allocation of resources. Low latency refers to taking short time to configure and manage resources, which can improve efficiency.

2.2.3. Edge Computing. Tung et al. [56] proposed a new framework for resource allocation based on market needs. The resources come from edge nodes (ENs) with limited heterogeneous capabilities and are allocated to multiple competing services on the network edge. Generating a market equilibrium solution by reasonably pricing ENs can obtain the maximum utilization of marginal computing resources. Xu et al. [57] proposed a strategy to optimize offloading and privacy protection. This strategy shifts tasks firstly to improve the resource utilization of resource-limited edge cells. And then it balances QoS performance and privacy protection to achieve joint optimization.

Xu et al. [58] proposed an offload strategy for edge computing under 5G network, which uses block chain technology. The optimal strategy is further obtained by using the balanced offloading method. It solves the problem of data loss under the condition of transmission delay, which is caused by the uneven requirements of user equipments on resources. Xu et al. [59] proposed a computational offloading method named EACO to reduce the energy consumption in smart computing models. Figure 5 shows architecture of smart edge computing, where the shortest path is used to unload tasks. EACO uses genetic algorithms to reduce the energy consumption for operating edge computing nodes and improve the efficiency of performing complex computing tasks. Xu et al. [60] proposed a computational offloading strategy for edge computing to protect the privacy of interconnected vehicle networks. They firstly analyzed privacy conflicts of tasks. And then they designed the communication route to obtain routing vehicles, which can achieve the optimization of several objectives. Yeting et al. [61] proposed a unique resource allocation mechanism. The mechanism takes each individual task as the basis for

Table 3: Work summary of resource allocation and management in fog computing.

Problems	Solutions	Literatures	Advantages
Task allocation	A new model of scheduling jobs and a redistribution mechanism	[49]	Finishes jobs on time, optimizes the number of tasks, and shortens the delay of tasks
	A framework to administrate resources and a method to predict and administrate resources	[50]	Assists service providers to predict the amount of available resources based on different types of service customers and deals with the phenomenon that objects or devices withdraw from resource utilization at any time
	A bi-matching approach	[51]	Improves the performance of the system and obtains higher cost efficiency
Resource allocation	A framework for joint optimization	[52]	Optimizes resource allocation and improves network performance
	A method to allocate computing and communication resources	[53]	Allocates computing and communication resources, transfers computing jobs to remote cloud and nodes, simplifies edge nodes' computing, and saves computing energy
	A platform to allocate resources accessible to client devices	[54]	Enables rapid response to computing resources and allocates the resources of devices near the host
	A framework to manage resources	[55]	Considers the fluctuation of the customer abandonment probability and service types
Low latency	A distributed algorithm	[5]	Solves the problem of wide range optimization and allocates resources jointly

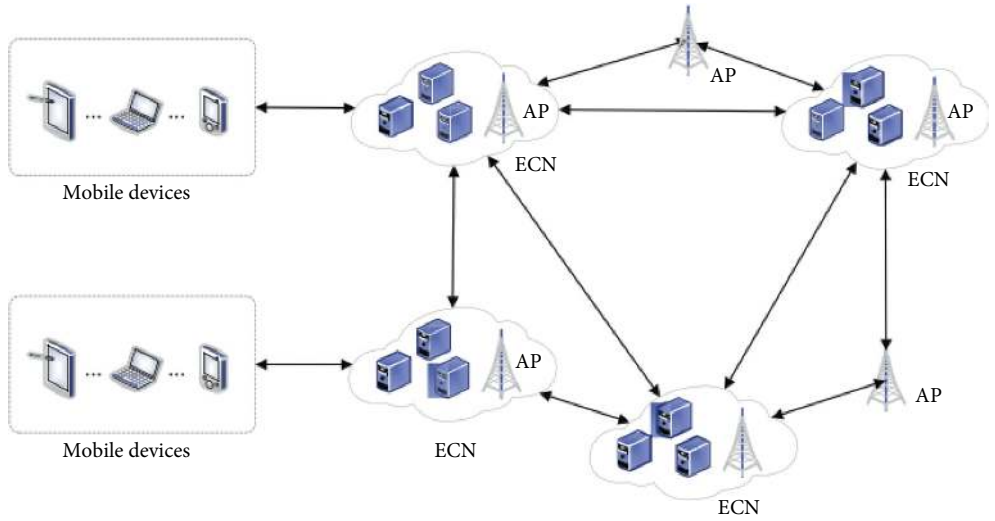

Figure 5: An architecture of smart edge computing [59].

resource allocation, rather than for the whole service. It reduces the packet loss rate and saves energy by unloading services.

2.2.4. MEC. Chen et al. [62] studied the problem of computing unloading with several users in the environment of MEC with wireless interference which have many channels. In addition, a distributed algorithm for computing unloading is developed. The algorithm can perform the unloading well even when there are a large number of users. Gao et al. [63] built a quadratic binary program, which is able to assign tasks in mobile cloud computing environment. Two algorithms are presented to obtain the optimal solution. Both of these heuristic algorithms can effectively solve the task assignment problem. Xu et al. [64] proposed an offloading method using block chain technology. It can guarantee the loss of data in offloading tasks under edge computing. And it can solve the problem of resource requests out of proportion due to the limited load of edge computing equipment during task transfer. Yifei et al. [65] proposed a model-free reinforcement learning framework to solve the problem of computational unloading. This model can be applied to the computational unloading with time-changing computing requests.

2.2.5. IoT. Barcelo et al. [66] expressed the problem of service allocation [67] as a mixed flow problem with minimum cost which can be solved by LP, solving this service allocation problem can solve the problems of unbalanced network load and delay of end-to-end service. And it can also figure out the problem of excessive consumption of electricity brought by the architecture of centralized cloud. Angelakis et al. [68] assigned the requirements of services resources to heterogeneous network interfaces of equipments. So, more heterogeneous network interfaces can be used by a large amount of services.

Li et al. [69] proposed communication framework in 5G and studied the problem of allocating power and channels. So, the signal data in the channel can be available and the total energy efficiency can be maximum. Formula (9) [69] shows how to calculate the energy efficiency of a system:

$$U = \sum_{k=1}^{K} \sum_{i=1}^{M} \mathrm{EE}_{i,K}^{S} + \sum_{K=1}^{K} \sum_{i=1}^{N} \mathrm{EE}_{j,k}^{A}, \tag{9}$$

where $\mathrm{EE}_{i,K}^{S}$ and $\mathrm{EE}_{j,k}^{A}$, respectively, denote the energy efficiency of sensor S and actuator A on channels. The sets of sensors, actuators, and channels are, respectively, represented as $s = \{S_1, S_2, \ldots, S_M\}$, $A = \{A_1, A_2, \ldots, A_N\}$, and $C = \{C_1, C_2, \ldots, C_K\}$.

Liu et al. [70] studied the problem of allocating resources efficiently on IoT that supplies wireless power. In this method, users are first grouped into accessible channels. And then power distribution of users grouped in the same channel is studied to improve throughput of the network. This method can allocate finite resources to a large group of users. Ejaz and Ibnkahla [71] proposed the resource allocation framework with several bands under cognitive 5G IoT. In the highly dynamic environment of the IoT, multiband method can manage resources more flexibly and reduce more energy consumption. In addition, a reconstruction approach with several levels is proposed to allocate resources reasonably for applications with different needs of QoS. Colistra et al. [72] proposed a protocol which is distributed and optimal to allocate resources in heterogeneous IoT. Because this protocol has excellent adaptability when changing topology of network, it can distribute resources evenly among nodes. Jian et al. [73] proposed a multilevel allocating resources algorithm for IoT communication using advanced technology. The algorithm uses hierarchical structure and has fast data processing rate and very low latency in a saturated or not saturated environment.

Zheng and Liu [74] proposed a new algorithm to allocate bandwidth dynamically for controlling remote computers in the IoT. This method can reduce the error of signal reconstruction under the same bandwidth and make the bandwidth allocation of IoT more reasonable. Gai and Qiu [75] used reinforcement learning mechanisms to allocate resources to achieve high Quality of Experience. This method can effectively solve the resource allocation problems caused by the mismatch of service quality and complex service providing condition in the IoT.

Table 4 shows a summary of the above works. The problem in Table 4 represents the realization of dynamic and flexible allocation of resources. The resources here can represent channels, bandwidth, and power.

2.3. Scientific Workflow Execution. Implementing scientific workflows, especially in heterogeneous environments, can reduce resource waste and reduce energy costs. Scientific workflow can be obtained by reasonably allocating resources and dynamically deploying VMs.

2.3.1. Cloud Computing. Xu et al. [76] proposed a resource allocation method called EnRealan to solve the problem of energy consumption. The dynamic deployment of VMs is generally adopted to execute scientific workflows. Bousselmi et al. [77] proposed a scheduling method based on energy perception for executing scientific workflows in cloud computing. At first, an algorithm of splitting workflows for energy minimization is presented, which can achieve a high parallelism without huge energy consumption. Then, a heuristic algorithm used to optimize cat swarm is proposed for the created partitions. The algorithm can minimize the total consumption of energy and the execution time of workflows. Sonia et al. [78] proposed a workflow scheduling method with several objects and hybrid particle swarm optimization algorithm. In addition, a method for dynamically scaling voltage and frequency is proposed. The method can make the processors work at any voltage level, so as to minimize the energy consumption in the process of workflow scheduling. Both Bousselmi et al. [77] and Sonia et al. [78] use scheduling method to achieve scientific workflows and study the problem of energy consumption. The difference is that Bousselmi et al. [77] focuses on intensive computing tasks, while Sonia et al. [78] focuses on workflow scheduling on heterogeneous computing systems.

Cao [79] established a scheduling algorithm of scientific workflows with an objective of energy saving. This algorithm can enable service providers to gain high profits and reduce users' overhead at the same time. Li et al. [80] proposed a scheduling algorithm based on cloud computing, which can minimize cost of performing workflows within a specified time. In addition, the rented VM was modified to save cost further. Khaleel and Zhu [81] proposed a scheduling algorithm and took scientific workflows as a model to make full use of cloud resources and save energy. Shi et al. [82] designed a flexible resource allocation and job scheduling mechanism to implement scientific workflows. Because this mechanism can implement scientific workflows within prescribed budgets and deadlines, so it can work better than other mechanisms.

Table 5 shows a summary of the abovementioned works. The problems in Table 5 are derived from the implementation of the scientific workflow. VM deployment refers to the rational allocation of VMs. Workflow scheduling refers to reducing the scheduling energy and time. In addition, it refers to the scheduling of the workflow on heterogeneous systems. Cost reduction refers to reducing the cost of workflow execution. Effective implementation is about scientific workflow execution within a specified budget and time.

2.4. Server Optimization. Server optimization is also a good way to save energy. The goal of optimizing the server can be achieved by uninstalling unnecessary servers or consolidating servers, as well as by reasonably scheduling tasks. Unlike QoS optimization, server optimization aims to optimize the number of used servers, improves the energy efficiency of servers, and consolidates servers. However, QoS optimization studies how to make users get better experience and meet their needs.

Table 4: Work summary of resource allocation and management in IoT.

Problems	Solutions	Literatures	Advantages
Resource management	A distributed cloud network framework	[66]	Replaces the centralized architecture with a distributed cloud architecture, solves the defects of the centralized cloud architecture, and brings people better experience
	A MILP model	[68]	Assigns the requirement of services' resource to heterogeneous network equipment interface
	A framework for communication used in 5G	[69]	Transforms the resource allocation problem into a power and channel allocation problem, minimizes the total energy consumption, and improves QoS levels
	A low complexity channel allocation algorithm	[70]	Improves throughput of the network and allocates finite resources to a large group of users
	A resource allocation framework with several bands under cognitive 5G IoT	[71]	Manages resources more flexibly and reduces energy consumption than common single-band approach
	A protocol which is distributed and optimal to allocate resources	[72]	Has excellent adaptability when changing topology of network and dynamically manages resources in the heterogeneous IoT environment
	A multilevel allocating resources algorithm for IoT communication	[73]	Has fast data processing rate and very low latency in both saturated and nonsaturated environment
	A new algorithm to allocate bandwidth dynamically	[74]	Reduces the error of signal reconstruction under the same bandwidth and makes the bandwidth allocation of IoT more reasonable
	A reinforcement learning mechanism	[75]	Effectively solves the resource allocation problems caused by the mismatch of service quality and complex service providing condition in the IoT

Table 5: Work summary of scientific workflow execution in cloud computing.

Problems	Solutions	Literatures	Advantages
VM deployment	A resource allocation method named EnRealan	[76]	Performs scientific workflows based on energy perception across cloud platforms
Workflow scheduling	A scheduling method based on energy perception	[77]	Achieves a high parallelism without huge energy consumption and minimizes the total consumption of energy and execution time of workflows
	A workflow scheduling method with several objects and hybrid particle swarm optimization algorithm	[78]	Makes the processors work at any voltage level, minimizes the energy consumption in the process of workflow scheduling, and studies the scheduling problem of workflows on heterogeneous systems
	A scheduling algorithm based on various applications	[79]	Enables service providers to gain high profits and reduces user overhead at the same time
Cost reduction	A scheduling algorithm based on energy perception	[80, 81]	Minimizes the cost of performing workflows while meeting the time constraint
Effective implementation	A flexible resource allocation and job scheduling mechanism	[82]	Implements scientific workflows within prescribed budgets and deadlines

2.4.1. Cloud Computing. Ge et al. [83] proposed a game-theoretic method and transformed the problem of minimizing energy into a congestion game. All mobile devices in this method are participants in the game. The method chooses a server to unload the computation tasks to optimize QoS levels and save energy, which can optimize the system and save energy. Wang et al. [84] proposed a MapReduce-based multitask scheduling algorithm to achieve the objective of energy saving. This model is a two-layer model, which considers the impact of server performance changes on energy consumption, and the limitation of network bandwidth. In addition, a local search operator is designed, based on which a two-layer genetic algorithm is proposed. The algorithm can schedule tens of thousands of tasks in cloud and achieve large-scale optimization. Yanggratoke et al. [85] proposed a general generic gossip protocol, aiming at allocating resources in cloud environment. An instantiation of this protocol was developed to enable server consolidation to allocate resources to more servers to meet changing load patterns.

2.5. Load Balancing. Load balancing can help save energy by managing the number of servers and allocating resources.

2.5.1. Cloud Computing. Paya and Marinescu [86] introduced an operation model that balances cloud computing load and expands applications to save energy. The principle of this model is to define an operating system. The system should make servers run in the system as many as possible.

Table 6: Work summary of load balancing in cloud computing.

Problems	Solutions	Literatures	Advantages
Server management	An operation model that can balance cloud computing load and expand application	[86]	Saves energy by managing the number of servers running in the system
Workload management	A hybrid approach	[87]	Reduces the footprint of carbon and allocates workload across cloud computing
	An algorithm to dynamically manage the load	[88]	Manages the load, evens the load distribution between servers, and allocates tasks between VMs
	A green power administration mechanism	[89]	Monitors and jointly allocates VM resources
VM management	An optimization system	[90]	Migrates VMs to adjust high and low loads without interrupting services and balances the load of VMs running on multiple physical machines

When no tasks are being performed, the system should adjust servers to sleep, thus energy consumption can be reduced. Justafort et al. [87] mainly studied the problem of workload distribution across cloud computing environment and proposed a method to solve the problem of the VM layout. So, the footprint of carbon can be effectively reduced. Panwar and Mallick [88] proposed an algorithm to dynamically manage the load and effectively distribute the total incoming requests between VMs. Through efficient and uniform utilization of resources, this algorithm can achieve uniform distribution of load between servers. Yang et al. [89] proposed a power management mechanism to balance the load. The system can monitor VMs and dynamically allocate the resources. Yang et al. [90] proposed an optimization system to better allocate resources dynamically, which can balance the load of VMs running on multiple physical machines. Under this system, VMs can be migrated automatically to adjust high and low loads without interrupting services. Yang et al. [89, 90] manage VMs to achieve load balancing. They allocate resources dynamically to migrate VMs, which can balance workloads on different physical machines. The difference is that Yang et al.[89] integrate a dynamic resource allocation approach with OpenNebula. While, Yang et al. [90] focuse on avoiding service outages during VM migration.

Table 6 shows a summary of the abovementioned works. The problems in Table 6 are from the load balancing problem. Server management is about the control of the number of servers running in the system. Workload management is the rational allocation of workload or tasks. VM management refers to configuring VM resources and migrating VMs to adjust loads.

2.5.2. Fog Computing. Xu et al. [91] proposed a method called "DRAM" to dynamically allocate resources in fog computing environment, which can avoid both too high and too low loads. The method first analyzes the load balance of different kinds of computing nodes. And then it designs a fog resource allocation method to achieve load balance, which allocates resources statically and migrates services dynamically. Oueis et al. [92] studied the load balance problem in fog computing. A custom fog clustering algorithm is proposed to solve the problem. In the problem, several users need to offload computations and all of their demands need to be handled by local computing cluster.

2.5.3. IoT. Wang et al. [93] established architecture of the energy saving targeted system in industrial IoT. In addition, in order to predict sleep intervals, they developed a sleep scheduling and a wake protocol, which provide a better way for energy saving.

3. Conclusion

This paper did a comprehensive study of QoS optimization and energy saving in cloud computing, edge computing, fog computing, and IoT models. We summarized five main problems and analyzed their solutions proposed by existing works. By conducting this survey, we aim to help readers have a deeper understanding on the concepts of different computing models and the techniques of QoS optimization and energy saving in these models.

The investigated papers focus on issues about ensuring QoS and reducing SLA violations and resource management. In the case of QoS assurance and SLA violations reduction, the main solution of QoS assurance is efficient VM management. This solution can meet customers' requirements through reasonable scheduling and integration of VMs. Most of resource management techniques are realized by reasonable scheduling of resources, which can reduce the waste of VMs, servers, and traffic.

Disclosure

This manuscript is an extension of A Survey of QoS Optimization and Energy Saving in Cloud, Edge, and IoT in The 9th EAI International Conference on Cloud Computing.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant no. 61702274), Natural Science Foundation of Jiangsu Province (Grant no. BK20170958), and PAPD.

References

[1] Q. Zhang, L. Cheng, and R. Boutaba, "Cloud computing: state-of-the-art and research challenges," *Journal of Internet Services and Applications*, vol. 1, no. 1, pp. 7–18, 2010.

[2] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge computing: vision and challenges," *IEEE Internet of Things Journal*, vol. 3, no. 5, pp. 637–646, 2016.

[3] M. Iorga, L. Feldman, R. Barton, M. J. Martin, N. S. Goren, and C. Mahmoudi, "Fog computing conceptual model," Tech. Rep., Recommendations of the National Institute of Standards and Technology, Gaithersburg, MD, USA, 2018.

[4] Nebbiolo, "Fog vs edge computing," Tech. Rep., Nebbiolo Technologies Inc., Milpitas, CA, USA, 2018.

[5] C. T. Do, N. H. Tran, C. Pham, M. G. R. Alam, J. H. Son, and C. S. Hong, "A proximal algorithm for joint resource allocation and minimizing carbon footprint in geo-distributed fog computing," in *Proceedings of the 2015 International Conference on Information Networking (ICOIN)*, pp. 324–329, IEEE, Siem Reap, Cambodia, January 2015.

[6] S. Vashi, J. Ram, J. Modi, S. Verma, and C. Prakash, "Internet of things (IoT): a vision, architectural elements, and security issues," in *Proceedings of the 2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)*, pp. 492–496, IEEE, Coimbatore, India, February 2017.

[7] M. Mazzucco, D. Dyachuk, and R. Deters, "Maximizing cloud providers' revenues via energy aware allocation policies," in *Proceedings of the 2010 IEEE 3rd International Conference on Cloud Computing*, pp. 131–138, IEEE, Miami, FL, USA, July 2010.

[8] Q. He, R. Zhou, X. Zhang et al., "Keyword search for building service-based systems," *IEEE Transactions on Software Engineering*, vol. 43, no. 7, pp. 658–674, 2016.

[9] L. Sun, H. Dong, O. K. Hussain, F. K. Hussain, and A. X. Liu, "A framework of cloud service selection with criteria interactions," *Future Generation Computer Systems*, vol. 94, pp. 749–764, 2019.

[10] M. Mazzucco and D. Dyachuk, "Optimizing cloud providers revenues via energy efficient server allocation," *Sustainable Computing: Informatics and Systems*, vol. 2, no. 1, pp. 1–12, 2012.

[11] Q. He, J. Han, F. Chen et al., "Qos-aware service selection for customisable multi-tenant service-based systems: maturity and approaches," in *Proceedings of the 2015 IEEE 8th International Conference on Cloud Computing*, pp. 237–244, IEEE, New York, NY, USA, July 2015.

[12] L. Sun, J. Ma, H. Wang, Y. Zhang, and J. Yong, "Cloud service description model: an extension of usdl for cloud services," *IEEE Transactions on Services Computing*, vol. 11, no. 2, pp. 354–368, 2015.

[13] Y. Wang, Q. He, D. Ye, and Y. Yang, "Formulating criticality-based cost-effective fault tolerance strategies for multi-tenant service-based systems," *IEEE Transactions on Software Engineering*, vol. 44, no. 3, pp. 291–307, 2017.

[14] S. Mustafa, K. Bilal, S. U. R. Malik, and S. A. Madani, "Sla-aware energy efficient resource management for cloud environments," *IEEE Access*, vol. 6, pp. 15004–15020, 2018.

[15] J. Bi, H. Yuan, M. Tie, and W. Tan, "Sla-based optimisation of virtualised resource for multi-tier web applications in cloud data centres," *Enterprise Information Systems*, vol. 9, no. 7, pp. 743–767, 2015.

[16] S. Singh, I. Chana, and R. Buyya, "Star: sla-aware autonomic management of cloud resources," *IEEE Transactions on Cloud Computing*, p. 1, 2017.

[17] A. Beloglazov and R. Buyya, "Energy efficient resource management in virtualized cloud data centers," in *Proceedings of the 2010 10th IEEEACM International Conference on Cluster, Cloud and Grid Computing*, pp. 826–831, IEEE Computer Society, Melbourne, Australia, May 2010.

[18] M. Guazzone, C. Anglano, and M. Canonico, "Energy-efficient resource management for cloud computing infrastructures," in *Proceedings of the 2011 IEEE Third International Conference on Cloud Computing Technology and Science*, pp. 424–431, IEEE, Athens, Greece, November 2011.

[19] Y. Sun, J. White, and S. Eade, "A model-based system to automate cloud resource allocation and optimization," in *Proceedings of the International Conference on Model Driven Engineering Languages and Systems*, pp. 18–34, Springer, Valencia, Spain, October 2014.

[20] G. Siddesh and K. Srinivasa, "Sla-driven dynamic resource allocation on clouds," in *Proceedings of the International Conference on Advanced Computing, Networking and Security*, pp. 9–18, Springer, Surathkal, India, December 2011.

[21] S. K. Garg, S. K. Gopalaiyengar, and R. Buyya, "Sla-based resource provisioning for heterogeneous workloads in a virtualized cloud datacenter," in *Proceedings of the International Conference on Algorithms and Architectures for Parallel Processing*, pp. 371–384, Springer, Melbourne, Australia, October 2011.

[22] J. Bi, Z. Zhu, and H. Yuan, "Sla-aware dynamic resource provisioning for profit maximization in shared cloud data centers," in *Proceedings of the International Conference on High Performance Networking, Computing and Communication Systems*, pp. 366–372, Springer, Singapore, May 2011.

[23] L. Qi, Y. Chen, Y. Yuan, S. Fu, X. Zhang, and X. Xu, "A Qos-aware virtual machine scheduling method for energy conservation in cloud-based cyber-physical systems," *World Wide Web*, vol. 4, no. 3, pp. 1–23, 2019.

[24] A. Beloglazov, J. Abawajy, and R. Buyya, "Energy-aware resource allocation heuristics for efficient management of data centers for cloud computing," *Future Generation Computer Systems*, vol. 28, no. 5, pp. 755–768, 2012.

[25] L. Qi, R. Wang, C. Hu, S. Li, Q. He, and X. Xu, "Time-aware distributed service recommendation with privacy-preservation," *Information Sciences*, vol. 480, pp. 354–364, 2019.

[26] Z. Zhai, B. Cheng, Y. Tian, J. Chen, L. Zhao, and M. Niu, "A data-driven service creation approach for end-users," *IEEE Access*, vol. 4, pp. 9923–9940, 2016.

[27] L. Gu, D. Zeng, S. Guo, A. Barnawi, and Y. Xiang, "Cost efficient resource management in fog computing supported medical cyber-physical system," *IEEE Transactions on Emerging Topics in Computing*, vol. 5, no. 1, pp. 108–119, 2015.

[28] L. Ni, J. Zhang, C. Jiang, C. Yan, and K. Yu, "Resource allocation strategy in fog computing based on priced timed petri nets," *Ieee Internet of Things Journal*, vol. 4, no. 5, pp. 1216–1228, 2017.

[29] L. Wei, T. Yang, F. C. Delicato et al., "On enabling sustainable edge computing with renewable energy resources," *IEEE Communications Magazine*, vol. 56, no. 5, pp. 94–101, 2018.

[30] P. Lai, Q. He, G. Cui et al., "Edge user allocation with dynamic quality of service," in *Proceedings of the International Conference on Service-Oriented Computing*, pp. 86–101, Springer, Toulouse, France, October 2019.

[31] X. Xu, Q. Liu, X. Zhang, J. Zhang, L. Qi, and W. Dou, "A blockchain-powered crowdsourcing method with privacy

preservation in mobile environment," *IEEE Transactions on Computational Social Systems*, vol. 6, no. 6, pp. 1407–1419, 2019.

[32] O. Rolik, E. Zharikov, and S. Telenyk, "Microcloud-based architecture of management system for IoT infrastructures," in *Proceedings of the 2016 Third International Scientific-Practical Conference Problems of Infocommunications Science and Technology (PIC S&T)*, pp. 149–151, IEEE, Kharkiv, Ukaraine, October 2016.

[33] W. He, S. Guo, Y. Liang, R. Ma, X. Qiu, and L. Shi, "Qos-aware and resource-efficient dynamic slicing mechanism for internet of things," *Computers, Materials & Continua*, vol. 61, no. 3, pp. 1345–1364, 2019.

[34] J. Yao and N. Ansari, "Qos-aware fog resource provisioning and mobile device power control in IoT networks," *IEEE Transactions on Network and Service Management*, vol. 16, no. 1, pp. 167–175, 2018.

[35] W. Wang, Y. Jiang, and W. Wu, "Multiagent-based resource allocation for energy minimization in cloud computing systems," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 47, no. 2, pp. 1–16, 2016.

[36] K. Krishnajyothi, "Parallel data processing for effective dynamic resource allocation in the cloud," *International Journal of Computer Applications*, vol. 70, no. 22, pp. 1–4, 2013.

[37] M. M. Hassan, B. Song, M. S. Hossain, and A. Alamri, "Efficient resource scheduling for big data processing in cloud platform," in *Proceedings of the International Conference on Internet and Distributed Computing Systems*, pp. 51–63, Springer, Calabria, Italy, September 2014.

[38] C.-M. Wu, R.-S. Chang, and H.-Y. Chan, "A green energy-efficient scheduling algorithm using the dvfs technique for cloud datacenters," *Future Generation Computer Systems*, vol. 37, no. 7, pp. 141–147, 2014.

[39] Z. Wang and X. Su, "Dynamically hierarchical resource-allocation algorithm in cloud computing environment," *The Journal of Supercomputing*, vol. 71, no. 7, pp. 2748–2766, 2015.

[40] W.-Y. Lin, G.-Y. Lin, and H.-Y. Wei, "Dynamic auction mechanism for cloud resource allocation," in *Proceedings of the 2010 10th IEEEACM International Conference on Cluster, Cloud and Grid Computing*, pp. 591-592, IEEE Computer Society, Victoria, Australia, May 2010.

[41] Y. O. Yazir, C. Matthews, R. Farahbod et al., "Dynamic resource allocation in computing clouds using distributed multiple criteria decision analysis," in *Proceedings of the 2010 IEEE 3rd International Conference on Cloud Computing*, pp. 91–98, IEEE, Washington, DC, USA, July 2010.

[42] W. Lin, J. Z. Wang, C. Liang, and D. Qi, "A threshold-based dynamic resource allocation scheme for cloud computing," *Procedia Engineering*, vol. 23, no. 5, pp. 695–703, 2011.

[43] X. Xu, S. Fu, L. Qi et al., "An IoT-oriented data placement method with privacy preservation in cloud environment," *Journal of Network and Computer Applications*, vol. 124, pp. 148–157, 2018.

[44] J. Bokyun, P. Md. Jalil, L. Daeho, and S. Doug Young, "Efficient computation offloading in mobile cloud computing for video streaming over 5 g," *Computers, Materials & Continua*, vol. 61, no. 2, pp. 439–463, 2019.

[45] H. Sarbazi-Azad and A. Y. Zomaya, "Energy-efficient resource utilization in cloud computing," in *Large Scale Network-Centric Distributed Systems*, pp. 377–408, Wiley-IEEE Press, Hoboken, NJ, USA, 2014.

[46] C.-H. Hsu, K. D. Slagter, S.-C. Chen, and Y.-C. Chung, "Optimizing energy consumption with task consolidation in clouds," *Information Sciences*, vol. 258, no. 3, pp. 452–462, 2014.

[47] C.-H. Hsu, S.-C. Chen, C.-C. Lee et al., "Energy-aware task consolidation technique for cloud computing," in *Proceedings of the 2011 IEEE Third International Conference on Cloud Computing Technology and Science*, pp. 115–121, IEEE, Athens, Greece, 2011.

[48] S. K. Panda and P. K. Jana, "An efficient task consolidation algorithm for cloud computing systems," in *Proceedings of the International Conference on Distributed Computing and Internet Technology*, pp. 61–74, Springer, Bhubaneswar, India, January 2016.

[49] L. Yin, J. Luo, and H. Luo, "Tasks scheduling and resource allocation in fog computing based on containers for smart manufacturing," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 10, pp. 4712–4721, 2018.

[50] M. Aazam and E.-N. Huh, "Dynamic resource provisioning through fog micro datacenter," in *Proceedings of the 2015 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops)*, pp. 105–110, IEEE, St. Louis, MO, USA, March 2015.

[51] B. Jia, H. Hu, Y. Zeng, T. Xu, and Y. Yang, "Double-matching resource allocation strategy in fog computing networks based on cost efficiency," *Journal of Communications and Networks*, vol. 20, no. 3, pp. 237–246, 2018.

[52] H. Zhang, Y. Xiao, S. Bu, D. Niyato, F. R. Yu, and Z. Han, "Computing resource allocation in three-tier IoT fog networks: a joint optimization approach combining stackelberg game and matching," *IEEE Internet of Things Journal*, vol. 4, no. 5, pp. 1204–1215, 2017.

[53] J. Tan, T.-H. Chang, and T. Q. Quelc, "Minimum energy resource allocation in fog radio access network with fronthaul and latency constraints," in *Proceedings of the 2018 IEEE 19th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, pp. 1–5, IEEE, Kalamata, Greece, June 2018.

[54] D. R. de Vasconcelos, R. M. de Castro Andrade, and J. N. de Souza, "Smart shadow–an autonomous availability computation resource allocation platform for internet of things in the fog computing environment," in *Proceedings of the 2015 International Conference on Distributed Computing in Sensor Systems*, pp. 216-217, IEEE, Fortaleza, Brazil, June 2015.

[55] M. Aazam, M. St-Hilaire, C.-H. Lung, and I. Lambadaris, "Pre-fog: IoT trace based probabilistic resource estimation at fog," in *Proceedings of the 2016 13th IEEE Annual Consumer Communications & Networking Conference (CCNC)*, pp. 12–17, IEEE, Las Vegas, NV, USA, January 2016.

[56] N. D. Tung, L. L. Bao, and B. Vijay, "Price-based resource allocation for edge computing: a market equilibrium approach," *IEEE Transactions on Cloud Computing*, p. 1, 2018.

[57] X. Xu, C. He, Z. Xu, L. Qi, S. Wan, and M. Z. A. Bhuiyan, "Joint optimization of offloading utility and privacy for edge computing enabled IoT," *IEEE Internet of Things Journal*, 2019.

[58] X. Xu, Y. Chen, X. Zhang, Q. Liu, X. Liu, and L. Qi, "A blockchain-based computation offloading method for edge computing in 5G networks," *Software: Practice and Experience*, 2019.

[59] X. Xu, Y. Li, T. Huang et al., "An energy-aware computation offloading method for smart edge computing in wireless metropolitan area networks," *Journal of Network and Computer Applications*, vol. 133, pp. 75–85, 2019.

[60] X. Xu, Y. Xue, L. Qi et al., "An edge computing-enabled computation offloading method with privacy preservation for internet of connected vehicles," *Future Generation Computer Systems*, vol. 96, pp. 89–100, 2019.

[61] G. Yeting, L. Fang, X. Nong, and C. Zhengguo, "Task-based resource allocation bid in edge computing micro datacenter," *Computers, Materials & Continua*, vol. 61, no. 2, pp. 777–792, 2019.

[62] X. Chen, L. Jiao, W. Li, and X. Fu, "Efficient multi-user computation offloading for mobile-edge cloud computing," *IEEE/ACM Transactions on Networking*, vol. 24, no. 5, pp. 2795–2808, 2016.

[63] B. Gao, L. He, X. Lu, C. Chang, K. Li, and K. Li, "Developing energy-aware task allocation schemes in cloud-assisted mobile workflows," in *Proceedings of the 2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing*, pp. 1266–1273, IEEE, Liverpool, UK, October 2015.

[64] X. Xu, X. Zhang, H. Gao, Y. Xue, L. Qi, and W. Dou, "Become: blockchain-enabled computation offloading for IoT in mobile edge computing," *IEEE Transactions on Industrial Informatics*, 2019.

[65] W. Yifei, W. Zhaoying, G. Da, and Y. F. Richard, "Deep q-learning based computation offloading strategy for mobile edge computing," *Computers, Materials & Continua*, vol. 59, no. 1, pp. 89–104, 2019.

[66] M. Barcelo, A. Correa, J. Llorca, A. M. Tulino, J. L. Vicario, and A. Morell, "IoT-cloud service optimization in next generation smart environments," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 12, pp. 4077–4090, 2016.

[67] B. Cheng, M. Wang, S. Zhao, Z. Zhai, D. Zhu, and J. Chen, "Situation-aware dynamic service coordination in an IoT environment," *IEEE/ACM Transactions On Networking*, vol. 25, no. 4, pp. 2082–2095, 2017.

[68] V. Angelakis, I. Avgouleas, N. Pappas, and D. Yuan, "Flexible allocation of heterogeneous resources to services on an IoT device," in *Proceedings of the 2015 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pp. 99-100, IEEE, Hong Kong, China, April 2015.

[69] 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.

[70] X. Liu, Z. Qin, Y. Gao, and J. A. McCann, "Resource allocation in wireless powered IoT networks," *IEEE Internet of Things Journal*, vol. 6, no. 3, pp. 4935–4945, 2019.

[71] W. Ejaz and M. Ibnkahla, "Multi-band spectrum sensing and resource allocation for IoT in cognitive 5g networks," *IEEE Internet of Things Journal*, vol. 5, no. 1, pp. 150–163, 2017.

[72] G. Colistra, V. Pilloni, and L. Atzori, "Task allocation in group of nodes in the IoT: a consensus approach," in *Proceedings of the 2014 IEEE International Conference on Communications (ICC)*, pp. 3848–3853, IEEE, Sydney, Australia, June 2014.

[73] J. Li, Q. Sun, and G. Fan, "Resource allocation for multiclass service in IoT uplink communications," in *Proceedings of the 2016 3rd International Conference on Systems and Informatics (ICSAI)*, pp. 777–781, IEEE, Shanghai, China, November 2016.

[74] L. I. Zheng and K. H. Liu, "Dynamic bandwidth resource allocation algorithm in internet of things and its application," *Computer Engineering*, vol. 38, no. 17, pp. 16–19, 2012.

[75] 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.

[76] X. Xu, W. Dou, X. Zhang, and J. Chen, "Enreal: an energy-aware resource allocation method for scientific workflow executions in cloud environment," *IEEE Transactions on Cloud Computing*, vol. 4, no. 2, pp. 166–179, 2016.

[77] K. Bousselmi, Z. Brahmi, and M. M. Gammoudi, "Energy efficient partitioning and scheduling approach for scientific workflows in the cloud," in *Proceedings of the 2016 IEEE International Conference on Services Computing (SCC)*, pp. 146–154, IEEE, San Francisco, CA, USA, June 2016.

[78] Y. Sonia, C. Rachid, K. Hubert, and G. Bertrand, "Multi-objective approach for energy-aware workflow scheduling in cloud computing environments," *The Scientific World Journal*, vol. 2013, Article ID 350934, 13 pages, 2013.

[79] F. Cao, *Efficient Scientific Workflow Scheduling in Cloud Environment*, Southern Illinois University, Carbondale, IL, USA, 2014.

[80] Z. Li, J. Ge, H. Hu, W. Song, H. Hu, and B. Luo, "Cost and energy aware scheduling algorithm for scientific workflows with deadline constraint in clouds," *IEEE Transactions on Services Computing*, vol. 11, no. 4, pp. 713–726, 2015.

[81] M. Khaleel and M. M. Zhu, "Energy-aware job management approaches for workflow in cloud," in *Proceedings of the 2015 IEEE International Conference on Cluster Computing*, pp. 506-507, IEEE, Chicago, IL, USA, September 2015.

[82] J. Shi, J. Luo, F. Dong, and J. Zhang, "A budget and deadline aware scientific workflow resource provisioning and scheduling mechanism for cloud," in *Proceedings of the 2014 IEEE 18th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, pp. 672–677, IEEE, Hsinchu, Taiwan, January 2014.

[83] Y. Ge, Y. Zhang, Q. Qiu, and Y.-H. Lu, "A game theoretic resource allocation for overall energy minimization in mobile cloud computing system," in *Proceedings of the 2012 ACMIEEE International Symposium on Low Power Electronics and Design*, pp. 279–284, ACM, Redondo Beach, CA, USA, August 2012.

[84] X. Wang, Y. Wang, and C. Yue, "An energy-aware bi-level optimization model for multi-job scheduling problems under cloud computing," *Soft Computing*, vol. 20, no. 1, pp. 303–317, 2016.

[85] R. Yanggratoke, F. Wuhib, and R. Stadler, "Gossip-based resource allocation for green computing in large clouds," in *Proceedings of the 2011 7th International Conference on Network and Service Management*, pp. 1–9, IEEE, Paris, France, October 2011.

[86] A. Paya and D. C. Marinescu, "Energy-aware load balancing and application scaling for the cloud ecosystem," *IEEE Transactions on Cloud Computing*, vol. 5, no. 1, pp. 15–27, 2015.

[87] V. D. Justafort, R. Beaubrun, and S. Pierre, "A hybrid approach for optimizing carbon footprint in intercloud environment," *IEEE Transactions on Services Computing*, vol. 12, no. 2, pp. 186–198, 2016.

[88] R. Panwar and B. Mallick, "Load balancing in cloud computing using dynamic load management algorithm," in *Proceedings of the 2015 International Conference on Green Computing and Internet of Things (ICGCIoT)*, pp. 773–778, IEEE, Noida, India, October 2015.

[89] C.-T. Yang, K.-C. Wang, H.-Y. Cheng, C.-T. Kuo, and W. C. C. Chu, "Green power management with dynamic resource allocation for cloud virtual machines," in *Proceedings of the 2011 IEEE International Conference on High Performance Computing and Communications*, pp. 726–733, IEEE, Alberta, Canada, September 2011.

[90] C.-T. Yang, H.-Y. Cheng, and K.-L. Huang, "A dynamic resource allocation model for virtual machine management

on cloud," in *Proceedings of the International Conference on Grid and Distributed Computing*, pp. 581–590, Springer, Gangneug, Korea, December 2011.

[91] X. Xu, F. Shucun, C. Qing et al., "Dynamic resource allocation for load balancing in fog environment," *Wireless Communications & Mobile Computing*, vol. 2018, no. 2, Article ID 6421607, 15 pages, 2018.

[92] J. Oueis, E. C. Strinati, and S. Barbarossa, "The fog balancing: load distribution for small cell cloud computing," in *Proceedings of the 2015 IEEE 81st Vehicular Technology Conference (VTC Spring)*, pp. 1–6, IEEE, Glasgow, UK, May 2015.

[93] K. Wang, Y. Wang, Y. Sun, S. Guo, and J. Wu, "Green industrial internet of things architecture: an energy-efficient perspective," *IEEE Communications Magazine*, vol. 54, no. 12, pp. 48–54, 2016.