A Survey of QoS Optimization and Energy Saving in Cloud, Edge and IoT

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

[1] Jiangsu Collaborative Innovation Center of Atmospheric Environment and Equipment Technology (CICAEET), Nanjing, China
sunle2009@gmail.com

[2] Nanjing University of Information Science and Technology, Nanjing 210044, China

Abstract. Since the emergence of cloud computing, it has been serving people as an important way of data processing. Later, with the development of computer and people's demand for higher service quality, fog computing, edge computing, mobile edge computing (MEC), Internet of Things (IoTs) and other models gradually appeared. They are developed step by step to bring better service to people. In recent years, IoTs technology has also been developed rapidly. This paper firstly gives a brief overview of cloud computing, fog computing, edge computing, MEC and IoTs. Then, we investigated the important papers related to these technologies, classify and compared the papers, so as to have a deeper understanding of these technologies.

Keywords: Cloud computing · Fog computing · Edge computing · MEC · IoT

1 Introduction

With the development of the Internet and the maturity of various processing and storage technologies, more and more computing resources emerge, and a large amount of data needs to be processed. New computing models emerge when users' requirements on resources increase, which include cloud computing, fog computing, edge computing, mobile edge computing (MEC) and Internet of Things (IoTs).

This paper investigates the important papers related to these computing models, and divides them into five categories based on the problems solved in these papers. On the basis of the problem classification in each section, further division is made according to the model used. For each paper, we point out the problem it aims to solve and introduce the methods it uses to solve the problem. In addition, in each problem category, we compare the methods proposed by the reviewed papers to solve the problem.

Supported by organization CloudComp.

X. Zhang et al. (Eds.): CloudComp 2019/SmartGift 2019, LNICST 322, pp. 33–46, 2020.
https://doi.org/10.1007/978-3-030-48513-9_3

2 Computing Models

2.1 Cloud Computing

With the rapid development of the Internet, more and more computing resources emerge, so a new computing model is needed to manage these resources, and cloud computing comes into being. The National Institute of Standards and Technology defined "cloud computing" as follows: it can provide convenient and quick network access to shared configurable resources, such as networks and servers. In addition, the provisioning and publishing of these resources does not require much administration and interaction of service providers [51].

2.2 Fog Computing

Due to the development of the IoT and the increasing needs of people, the IoT system based on cloud computing faces some limitations, such as the failure of playing a good role in large-scale or heterogeneous conditions [18]. 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 can extend cloud resources to the network edge to facilitate the management of resources and services [26].

2.3 Edge Computing

Edge computing is defined as a technology that allows computing to be performed on the edge of a network [32]. 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 the 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 [32].

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 by 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 that provides the capabilities of information technology and cloud computing at the network edge closing to mobile customers.

2.4 IoT

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 development of the IoT, its application has gradually expanded to cover increasingly wider domains, but its goal has always been to make computers perceive information [38].

3 Energy Saving Techniques in Different Computing Models

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

3.1 QoS Guaranteeing or SLA Assurance

Cloud Computing. Mazzucco et al. [24] 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 to optimize users' experience while consuming the least amount of power. Mazzucco and Dyachuk [23] were also committed to making cloud service providers obtain the largest profits, and proposed the dynamic distribution strategy of powering server switch, which not only enables users to get good service, but also reduces power consumption. In order to make users have a good experience, this paper further uses a forecasting method to accurately predict the users' needs at different times. Mustafa et al. [25] 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. Bi et al. [5] established an architecture that can administrate itself in cloud data centers firstly, which 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. Finally, a problem of misalignment restrained optimization and a heuristic mixed optimization algorithm are proposed to make more revenues and meet different requirements of customers. Singh et al. [34] proposed a technology named "STAR" which can manage resources itself in the cloud computing environment and aims at reducing SLA violations, so that payment efficiency of cloud services can be improved. Beloglazov and Buyya [4] proposed a system to manage energy in 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. [14] proposed an automatic management system for resources to provide certain QoS and reduce energy consumption. Different from the traditional static method, this method can not only fit the changing workloads dynamically, but also achieve remarkable results in reducing QoS violations. Sun et al. [36] established a model to simplify the decision of cloud resource allocation and realize the independent allocation of resources. According to the description of advanced application and requirements of QoS and the performance of this method under different loads and resources, the optimal resource configuration can be obtained, so the QoS requirements can be well met. Siddesh and Srinivasa [33] paid close attention

to the dynamic resource allocation and risks which meet the SLA and proposed a framework which can deal with workload types that are heterogeneous by dynamically planning capacity and assessing risks. The framework considers not only scheduling methods to reduce SLA, but also risks in resource allocation to maximize revenues on the cloud. Garg et al. [11] proposed a resource allocation strategy for VM dynamic allocation, which can improve resource utilization and increase providers' profits while reducing SLA violation. Jing et al. [6] proposed a new dynamic allocating technique using mixed queue model. Meeting customers' different requirements of performance at different levels by providing virtualized resources to each layer of virtualized application services. All these methods can reasonably configure the resources in the cloud data center, improve the system performance, reduce the additional cost of using resources and meet the required QoS.

Fog Computing. Gu et al. [13] 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), and in order to reduce the cost of FCMCPS, they did researches on the joint of base station, task assignment and VM layout. In addition, the problem is modeled as a mixed integer linear programming (MILP), and a two-stage heuristic algorithm based on linear programming (LP) is proposed to help solve the problem of large computational amount. Ni et al. [27] 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.

Edge Computing. Wei et al. [43] 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 computing, which can make full use of renewable energy and provide better QoS.

IoT. Rolik et al. [30] proposed a method to build a framework of IoT infrastructure based on microcloud, which can help use resources rationally, reduce the cost of management infrastructure, and improve the quality of life of consumers. Yao and Ansari [48] proposed an algorithm to determine the number of VMs to be rented and to control the power supply, thus the cost of system can be minimized and the QoS can be improved.

The ultimate goal of these papers is to improve QoS or reduce SLA violations, but the methods used are different and the computing modes are also different. The paper [24] allocates the switch of power servers dynamically to minimize the power cost, so as to solve the contradiction that the high QoS cannot coexist with the low power consumption. The allocation strategy of the paper [23] also manages the opening and closing of the server, and at the same time satisfies a certain QoS and low energy consumption. The paper [4] manages energy

and integrates VMs according to the real-time utilization of resources, so as to improve QoS. The method of improving QoS in the paper [27] is making reasonable allocation of resources, but this method allows users to select resources independently. The algorithm in the paper [48] can effectively guarantee QoS in IoT network by controlling VMs and power supply.

The paper [13] introduces fog computing to ensure that Medical Cyber-Physical System (MCPS) provides high QoS and connects medical devices and data centers stably and with short delay at the same time. The framework in the paper [43] can manage energy uniformly, so as to meet the energy demand of equipment and QoS. The paper [30] establishes a framework for the management of IoT infrastructure to make rational use of resources, so as to improve QoS to a certain extent. The system in the paper [14] can self-manage the resources of cloud infrastructures to provide appropriate QoS. The model that can realize the independent allocation of resources established in the paper [36] considers many factors, so it can well meet QoS requirements. The paper [6] introduces a mixed queue model to allocate virtual resources, which can greatly reduce the extra cost of using resources and guarantee QoS.

The paper [25] uses techniques for consolidating workloads to achieve energy savings while reducing SLA violations. The paper [5] minimizes SLA by the reasonable allocation of VMs. The paper [34] takes advantage of technologies that enable self-administration of cloud services to effectively reduce SLA violations. The characteristics of the resource management model in the paper [33] are that it can meet the requirements of heterogeneous load types' resource management and SLA. The measurement and outline of the paper [11] can realize the dynamic allocation of VMs in heterogeneous environment and guarantee the SLA requirement.

3.2 Resource Management and Allocation

Cloud Computing. Wang et al. [40] introduced an allocation method for VM based on distributed multi-agent to allocate VMs to physical machines, which 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. [15] 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. [44] proposed a scheduling algorithm based on the technology that can scale the voltage frequency dynamically in cloud computing, through which resources can be allocated for performing tasks and low power consumption network infrastructure can be realized. Compared with other schemes, this scheme not only sacrifices the performance of execution operations, but also saves more energy. Sarbazi and Zomaya [31] 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 focus on energy consumption which is active and idle. Using

these two methods can promote the concurrent execution of multiple tasks and improve the energy efficiency. Hsu et al. [17] proposed a job consolidation technique aiming at energy saving, which can consume the least energy. In addition, the technology will limit the CPU usage and merge tasks in the virtual cluster. Once the task migration happens, the energy cost model will take into account the latency of the network. Hsu et al. [16] proposed a task intergration 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 [28] proposed an algorithm with several criteria to combine tasks, which not only considers the time needed for processing jobs, but also considers the utilization rate of VMs. The algorithm is more energy efficient because it takes into account not only the processing time but also the utilization rate of VMs. Wang et al. [42] 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. [21] proposed a dynamic auction approach for resource allocation, which can ensure that even if there are many users and resources, the providers will have reasonable profits and the computing resources will be allocated correctly. Yazir et al. [49] 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 [19] 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 reducing the time of processing tasks. Lin et al. [22] 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.

Fog Computing. Yin et al. [50] established a new model of scheduling jobs, which applies containers. And in order to make sure that jobs can be finished on time, a job scheduling algorithm is developed, which can also optimize the number of tasks that can be performed together on the nodes in fog computing. Moreover according to the specialties of the containers, this paper proposes a redistribution mechanisms to shorten the delay of tasks. These methods are very effective in reducing task delays. Aazam and Huh [1] 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 and they are subject to exit the use of resources. So a method to predict and administrate resources is proposed. The method considers that any objects or devices can quit using resources at anytime, so it can provide effective

management. Cuong et al. [9] studied the problem of allocating resources jointly and carbon footprint minimization in fog data center. In addition, a distributed algorithm is proposed to solve the problem of wide range optimization.

Edge Computing. Tung et al. [37] proposed a new framework for resource allocation based on market. The allocated resources come from edge nodes (ENs) with limited heterogeneous capabilities and are allocated to multiple competing services on the edge of the network. The advantages of generating a market equilibrium solution by reasonably pricing ENs is that not only the maximum utilization of marginal computing resources can be obtained, but also the optimal solution can be achieved.

MEC. Chen et al. [8] studied the problem of computing unloading with several users in the environment of mobile edge cloud computing with wireless interference which have many channels. In addition, a distributed algorithm for computing unload is developed, which can perform the unloading well even when there are a large number of users. Gao et al. [10] 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.

IoT. Barcelo et al. [3] expressed the problem of service allocation as a mixed flow problem with minimum cost which can be solved effectively by using LP. Solving this service allocation problem can solve the problems of unbalanced network load, delay of end-to-end service and excessive total consumption of electricity brought by the architecture of centralized cloud. Angelakis et al. [2] assigned the requirements of services' resources to heterogeneous network interface of equipment, and a MILP is given, so that more heterogeneous network interfaces of equipment can be used efficiently by a large amount of services. Song et al. [20] proposed a framework for communication used in 5G and the problem of resource allocation is transformed into the problem of power and channel allocation for making the signal data in the channel to be available and make the total energy efficiency of the system maximum.

In order to find better ways to manage computing resources, these papers pay special attention to the allocation of resources. The allocation of VMs or services and so on can make the system more energy efficient and contribute to the realization of green cloud computing. The paper [40] proposes a VM allocation method to efficiently allocate VM resources, and also considers the migration cost of VM migration, so it can consume less energy. The dynamic allocation of resources in the paper [15] is based on QoS requirements of applications, which can effectively reduce the waste of resources. In the paper [10], the tasks under edge computing are assigned scientifically and managed reasonably. The paper [3] studies the distribution of service resources in IoTCloud networks, which can effectively solve the defects of architecture of centralized cloud and bring people

better experience. In the paper [2], the allocation strategy of service resource requirements makes the server use network interface efficiently. The method proposed in the paper [19] can effectively improve the efficiency of resource allocation when submitting large tasks by dynamically allocating VMs.

The paper [44] uses scheduling algorithm to allocate resources for executing tasks. The advantages of this scheduling method is to ensure the performance of executing jobs while implementing green computing. The job scheduling algorithm in the paper [50] can realize redistribution, which enables timely response of tasks.

The paper [31] manages resources by integrating tasks to improve resource utilization and reduce energy consumption. The paper [17] also integrates tasks to manage resources, and the methods used limit the use of CPU, which is very helpful for resource saving. The paper [28] also integrates tasks to facilitate efficient management of resources. In particular, this integration method takes into account the time needed to process tasks and the use of VMs.

The paper [1] proposes a framework to administrate resources in fog computing mode, which is characterized by the ability to deal with the phenomenon that objects or devices withdraw from resource utilization at any time. The paper [37] proposes a framework that allocates resources of EN with limited heterogeneous capability. The paper [20] is suitable for 5G communication framework to decompose the problem of resource allocation, promoting the development of 5G.

The paper [8] realizes the effective management of resources by solving the problem of user computing unloading. The paper [16] uses the principle of using 70% CPU to combine tasks, which can also manage resources and improve the utilization of resources. The algorithm in the paper [42] can greatly improve the communication performance in a large range and reduce the communication traffic in allocating resources. The mechanism of the paper [21] can deal with resource allocation in large-scale user and resource situations. The method of resource management in the paper [49] is characterized by its decomposition, and each job is executed by autonomous nodes, which is more flexible and convenient. The paper [22] uses the threshold value to reconfigure the resources, which can optimize the results of the allocation. The paper [9] solves the problem of joint allocation of resources to minimize the footprint of carbon.

3.3 Scientific Workflow Execution

Cloud Computing. Xu et al. [45] proposed a resource allocation method based on energy perception called "EnRealan" to solve the problem of energy consumption caused by the extension of cloud platform, and the dynamic deployment of VMs is generally adopted to obtain executions of scientific workflow. Bousselmi et al. [7] proposed a scheduling method based on energy perception for scientific workflows in cloud computing. Firstly, algorithm of spliting workflow for energy minimization is presented to divide workflow, 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, which can minimize

the total consumption of energy and execution time of workflows. Sonia et al. [35] proposed a workflow scheduling method with several objects and hybrid particle swarm optimization algorithm. In addition, a technology for scaling voltage and frequency dynamically is proposed, which can make the processors work at any voltage level, so as to minimize the energy consumption in the process of work flow scheduling.

The purpose of these papers is to obtain the implementation of scientific workflow, which is also very conducive to the least energy consumption. A scientific workflow can be achieved through resource allocation or scheduling and so on. The paper [45] uses a resource allocation method based on energy perception and dynamic deployment of VMs to obtain executions of scientific workflow. The paper [7] uses scientific workflows' scheduling method based on energy to generate the workflow and partition, so as to get the scientific workflows. The method of obtaining scientific workflows in the paper [35] is to study the scheduling problem of workflows on heterogeneous systems, not only to optimize events and cost constraints, but also to reduce energy consumption as much as possible.

3.4 Servers Optimization

Cloud Computing. Yang et al. [12] proposed a game-theoretic method and transformed the problem of minimizing energy into a congestion game. All the mobile devices in this method are participants in the game, and then it chooses a server to unload the computation, which can optimize the system and save energy. Wang et al. [41] proposed a MapReduce-based multi-task scheduling algorithm to achieve the objective of energy saving. This model is a two-layer model, which not only considers the impact of servers' performance changes on energy consumption, but also considers the limitation of network bandwidth. In addition, a local search operator is designed, and on this basis, a two-layer genetic algorithm is proposed to schedule tens of thousands of tasks in the cloud, so as to achieve large-scale optimization. Yanggratoke et al. [47] proposed a general generic gossip protocol, aiming at allocating resources in cloud environment which is on a large scale. An instantiation of this protocol was developed to enable server consolidation to allocate resources to more servers while meeting changing load patterns.

These papers are all devoted to the optimization of servers. Through the unloading or integration of servers can optimize the number of servers, which can also save the energy. The paper [12] uses a game-theoretic method to unload the server, which can not only optimize the system but also save energy. The task scheduling model and task scheduling method proposed in the paper [41] can effectively improve the energy efficiency of the server and thus reduce the energy consumption of the data center. The protocol proposed in the paper [47] can integrate servers and optimize the number of servers, so that more servers can be allocated to resources and reduce the total energy consumption.

3.5 Load Balancing

Cloud Computing. Paya and Marinescu [29] introduced an operation model that can balance cloud computing load and expand application, which aims at saving energy. The principle of this model is to define an operating system that optimizes its energy, makes as many servers as possible run in the system, and adjusts to sleep when no tasks are being performed or when the server is light, thus saving energy.

Fog Computing. Xu et al. [46] proposed a method that is called "DRAM" to dynamically allocate resources in fog computing environment, which can avoid both too high load and too low load. Through analyzing different kinds of computing nodes' load balance firstly, then allocating resources statically and migrating service dynamically in fog environment to design the relevant resource allocation method, so as to achieve load balance.

IoT. Wang et al. [39] established the architecture of the energy-saving targeted system, which is based on the industrial IoT. And due to its three levels the traffic load can be balanced. In addition, in order to predict sleep intervals, a sleep scheduling as well as a wake protocol are developed, which can save energy better.

All these papers adopt certain methods to achieve load balance, achieve green cloud computing, improve resource utilization and reduce energy consumption. The paper [29] proposes a model to manage the number of servers running in the system and achieve load balance. The approach to make the system's load balance in the paper [46] is to make reasonable allocation of resources in fog environment. The three-tier architecture constructed in the paper [39] can ensure the load balance of traffic.

4 Conclusion

Nowadays, cloud computing, fog computing, edge computing, MEC, IoT and other technologies are developing rapidly, and they are rapidly improving the development of information technology. Although there will be many challenges in this field, their development can greatly change people's lives and bring convenience to people. This paper investigates and classifies the papers related to these five technologies to facilitate people's understanding of them.

Acknowledgement. This work is supported by the National Natural Science Foundation of China (Grants No 61702274) and the Natural Science Foundation of Jiangsu Province (Grants No BK20170958), and PAPD.

References

1. Aazam, M., Huh, E.N.: Dynamic resource provisioning through fog micro datacenter. In: 2015 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops), pp. 105–110. IEEE (2015)
2. Angelakis, V., Avgouleas, I., Pappas, N., Yuan, D.: Flexible allocation of heterogeneous resources to services on an IoT device. In: 2015 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), pp. 99–100. IEEE (2015)
3. Barcelo, M., Correa, A., Llorca, J., Tulino, A.M., Vicario, J.L., Morell, A.: IoT-cloud service optimization in next generation smart environments. IEEE J. Sel. Areas. Commun. **34**(12), 4077–4090 (2016)
4. Beloglazov, A., Buyya, R.: Energy efficient resource management in virtualized cloud data centers. In: Proceedings of the 2010 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing, pp. 826–831. IEEE Computer Society (2010)
5. Bi, J., Yuan, H., Tie, M., Tan, W.: SLA-based optimisation of virtualised resource for multi-tier web applications in cloud data centres. Enterp. Inf. Syst. **9**(7), 743–767 (2015)
6. Bi, J., Zhu, Z., Yuan, H.: SLA-aware dynamic resource provisioning for profit maximization in shared cloud data centers. In: Wu, Y. (ed.) ICHCC 2011. CCIS, vol. 163, pp. 366–372. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-25002-6_52
7. Bousselmi, K., Brahmi, Z., Gammoudi, M.M.: Energy efficient partitioning and scheduling approach for scientific workflows in the cloud. In: 2016 IEEE International Conference on Services Computing (SCC), pp. 146–154. IEEE (2016)
8. Chen, X., Jiao, L., Li, W., Fu, X.: Efficient multi-user computation offloading for mobile-edge cloud computing. IEEE/ACM Trans. Netw. **24**(5), 2795–2808 (2016)
9. Do, C.T., Tran, N.H., Pham, C., Alam, M.G.R., Son, J.H., Hong, C.S.: A proximal algorithm for joint resource allocation and minimizing carbon footprint in geo-distributed fog computing. In: 2015 International Conference on Information Networking (ICOIN), pp. 324–329. IEEE (2015)
10. Gao, B., He, L., Lu, X., Chang, C., Li, K., Li, K.: Developing energy-aware task allocation schemes in cloud-assisted mobile workflows. In: 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 (2015)
11. Garg, S.K., Gopalaiyengar, S.K., Buyya, R.: SLA-based resource provisioning for heterogeneous workloads in a virtualized cloud datacenter. In: Xiang, Y., Cuzzocrea, A., Hobbs, M., Zhou, W. (eds.) ICA3PP 2011. LNCS, vol. 7016, pp. 371–384. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-24650-0_32
12. Ge, Y., Zhang, Y., Qiu, Q., Lu, Y.H.: A game theoretic resource allocation for overall energy minimization in mobile cloud computing system. In: Proceedings of the 2012 ACM/IEEE International Symposium on Low Power Electronics and Design, pp. 279–284. ACM (2012)
13. Gu, L., Zeng, D., Guo, S., Barnawi, A., Xiang, Y.: Cost efficient resource management in fog computing supported medical cyber-physical system. IEEE Trans. Emerg. Top. Comput. **5**(1), 108–119 (2015)
14. Guazzone, M., Anglano, C., Canonico, M.: Energy-efficient resource management for cloud computing infrastructures. In: 2011 IEEE Third International Conference on Cloud Computing Technology and Science, pp. 424–431. IEEE (2011)

15. Hassan, M.M., Song, B., Hossain, M.S., Alamri, A.: Efficient resource scheduling for big data processing in cloud platform. In: Fortino, G., Di Fatta, G., Li, W., Ochoa, S., Cuzzocrea, A., Pathan, M. (eds.) IDCS 2014. LNCS, vol. 8729, pp. 51–63. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-11692-1_5
16. Hsu, C.H., et al.: Energy-aware task consolidation technique for cloud computing. In: 2011 IEEE Third International Conference on Cloud Computing Technology and Science, pp. 115–121. IEEE (2011)
17. Hsu, C.H., Slagter, K.D., Chen, S.C., Chung, Y.C.: Optimizing energy consumption with task consolidation in clouds. Inf. Sci. **258**(3), 452–462 (2014)
18. Iorga, M., Feldman, L., Barton, R., Martin, M.J., Goren, N.S., Mahmoudi, C.: Fog computing conceptual model. Technical report, Recommendations of the National Institute of Standards and Technology (2018)
19. Krishnajyothi, K.: Parallel data processing for effective dynamic resource allocation in the cloud. Int. J. Comput. Appl. **70**(22), 1–4 (2013)
20. Li, S., Ni, Q., Sun, Y., Min, G., Al-Rubaye, S.: Energy-efficient resource allocation for industrial cyber-physical IoT systems in 5G era. IEEE Trans. Industr. Inf. **14**(6), 2618–2628 (2018)
21. Lin, W.Y., Lin, G.Y., Wei, H.Y.: Dynamic auction mechanism for cloud resource allocation. In: Proceedings of the 2010 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing, pp. 591–592. IEEE Computer Society (2010)
22. Lin, W., Wang, J.Z., Chen, L., Qi, D.: A threshold-based dynamic resource allocation scheme for cloud computing. Procedia. Eng. **23**(5), 695–703 (2011)
23. Mazzucco, M., Dyachuk, D.: Optimizing cloud providers revenues via energy efficient server allocation. Sustain. Comput. Inform. Syst. **2**(1), 1–12 (2012)
24. Mazzucco, M., Dyachuk, D., Deters, R.: Maximizing cloud providers' revenues via energy aware allocation policies. In: 2010 IEEE 3rd International Conference on Cloud Computing, pp. 131–138. IEEE (2010)
25. Mustafa, S., Bilal, K., Malik, S.U.R., Madani, S.A.: SLA-aware energy efficient resource management for cloud environments. IEEE Access. **6**, 15004–15020 (2018)
26. Nebbiolo: Fog vs edge computing. Technical report, Nebbiolo Technologies Inc. (2018)
27. Ni, L., Zhang, J., Jiang, C., Yan, C., Yu, K.: Resource allocation strategy in fog computing based on priced timed petri nets. IEEE Internet Things J. **4**(5), 1216–1228 (2017)
28. Panda, S.K., Jana, P.K.: An efficient task consolidation algorithm for cloud computing systems. In: Bjørner, N., Prasad, S., Parida, L. (eds.) ICDCIT 2016. LNCS, vol. 9581, pp. 61–74. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-28034-9_8
29. Paya, A., Marinescu, D.C.: Energy-aware load balancing and application scaling for the cloud ecosystem. IEEE Trans. Cloud Comput. **5**(1), 15–27 (2015)
30. Rolik, O., Zharikov, E., Telenyk, S.: Microcloud-based architecture of management system for IoT infrastructures. In: 2016 Third International Scientific-Practical Conference Problems of Infocommunications Science and Technology (PIC S&T), pp. 149–151. IEEE (2016)
31. Sarbazi-Azad, H., Zomaya, A.Y.: Energy-efficient resource utilization in cloud computing. In: Large Scale Network-Centric Distributed Systems, pp. 377–408. Wiley-IEEE Press (2014)
32. Shi, W., Jie, C.: Edge computing: vision and challenges. IEEE Internet Things J. **3**(5), 637–646 (2016)

33. Siddesh, G.M., Srinivasa, K.G.: SLA - driven dynamic resource allocation on clouds. In: Thilagam, P.S., Pais, A.R., Chandrasekaran, K., Balakrishnan, N. (eds.) ADCONS 2011. LNCS, vol. 7135, pp. 9–18. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-29280-4_2
34. Singh, S., Chana, I., Buyya, R.: STAR: SLA-aware autonomic management of cloud resources. IEEE Trans. Cloud Comput., 1 (2017)
35. Sonia, Y., Rachid, C., Hubert, K., Bertrand, G.: Multi-objective approach for energy-aware workflow scheduling in cloud computing environments. Sci. World. J. **2013**(3–4), 350934 (2013)
36. Sun, Y., White, J., Eade, S.: A model-based system to automate cloud resource allocation and optimization. In: Dingel, J., Schulte, W., Ramos, I., Abrahão, S., Insfran, E. (eds.) MODELS 2014. LNCS, vol. 8767, pp. 18–34. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-11653-2_2
37. Tung, N.D., Bao, L.L., Vijay, B.: Price-based resource allocation for edge computing: a market equilibrium approach. IEEE Trans. Cloud Comput., 1 (2018)
38. Vashi, S., Ram, J., Modi, J., Verma, S., Prakash, C.: Internet of Things (IoT): a vision, architectural elements, and security issues. In: 2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), pp. 492–496. IEEE (2017)
39. Wang, K., Wang, Y., Sun, Y., Guo, S., Wu, J.: Green industrial internet of things architecture: an energy-efficient perspective. IEEE Commun. Mag. **54**(12), 48–54 (2016)
40. Wang, W., Jiang, Y., Wu, W.: Multiagent-based resource allocation for energy minimization in cloud computing systems. IEEE Trans. Syst. Man Cybern. Syst. **47**(2), 1–16 (2016)
41. Wang, X., Wang, Y., Yue, C.: An energy-aware bi-level optimization model for multi-job scheduling problems under cloud computing. Soft. Comput. **20**(1), 303–317 (2016)
42. Wang, Z., Su, X.: Dynamically hierarchical resource-allocation algorithm in cloud computing environment. J. Supercomput. **71**(7), 2748–2766 (2015)
43. Wei, L., Yang, T., Delicato, F.C., Pires, P.F., Tari, Z., Khan, S.U., Zomaya, A.Y.: On enabling sustainable edge computing with renewable energy resources. IEEE Commun. Mag. **56**(5), 94–101 (2018)
44. Wu, C.M., Chang, R.S., Chan, H.Y.: A green energy-efficient scheduling algorithm using the DVFS technique for cloud datacenters. Future Gener. Comput. Syst. **37**(7), 141–147 (2014)
45. Xu, X., Dou, W., Zhang, X., Chen, J.: EnReal: an energy-aware resource allocation method for scientific workflow executions in cloud environment. IEEE Trans. Cloud Comput. **4**(2), 166–179 (2016)
46. Xu, X., et al.: Dynamic resource allocation for load balancing in fog environment. Wirel. Commun. Mob. Comput. **2018**(2), 1–15 (2018)
47. Yanggratoke, R., Wuhib, F., Stadler, R.: Gossip-based resource allocation for green computing in large clouds. In: 2011 7th International Conference on Network and Service Management, pp. 1–9. IEEE (2011)
48. Yao, J., Ansari, N.: QoS-aware fog resource provisioning and mobile device power control in IoT networks. IEEE Trans. Netw. Serv. Manage. **16**(1), 167–175 (2018)
49. Yazir, Y.O., et al.: Dynamic resource allocation in computing clouds using distributed multiple criteria decision analysis. In: 2010 IEEE 3rd International Conference on Cloud Computing, pp. 91–98. IEEE (2010)

50. Yin, L., Juan, L., Haibo, L.: Tasks scheduling and resource allocation in fog computing based on containers for smart manufacture. IEEE Trans. Industr. Inf. **14**(10), 4712–4721 (2018)
51. Zhang, Q., Cheng, L., Boutaba, R.: Cloud computing: state-of-the-art and research challenges. J. Internet Serv. Appl. **1**(1), 7–18 (2010)