IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 17, 2018, accepted December 18, 2018, date of publication December 20, 2018, date of current version January 11, 2019.

Digital Object Identifier 10.1109/ACCESS.2018.2888976

AI-Powered Green Cloud and Data Center

JUN YANG[1], WENJING XIAO[1], CHUN JIANG[2], M. SHAMIM HOSSAIN[3], (Senior Member, IEEE), GHULAM MUHAMMAD[4], AND SYED UMAR AMIN[4]

[1]School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China
[2]School of Mechanical and Automotive Engineering, South China University of Technology, Guangzhou 510641, China
[3]Department of Software Engineering, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia
[4]Department of Computer Engineering, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia

Corresponding author: M. Shamim Hossain (mshossain@ksu.edu.sa)

This work was supported by the Deanship of Scientific Research at King Saud University, Riyadh, Saudi Arabia, under Project RGP-228.

ABSTRACT As the scale of cloud computing expands, its impact on energy and the environment is becoming more and more prominent. According to statistics, data centers' energy consumption has accounted for 50% of operating costs of the data centers. The rising energy consumption not only needs energy in large quantity but also imposes heavy pressure on the environment. The high energy consumption of cloud data center has become an issue, people pay close attention to, in the information technology field. It is also a problem to be solved urgently. At present, high energy consumption is caused by two reasons. First, a resource scheduling mechanism with the priority of completion time causes low server use ratio, and it is a pervasive phenomenon that small tasks take high consumption. Second, the current refrigerating system of the data center is based on peak value strategy, which causes excessive cooling supply, increases operation cost, and leads to huge waste of energy. In this paper, considering the reason of high energy consumption of the data center, a new framework of green cloud data center is put forward. Using the techniques relevant to artificial intelligence, we put forward a scheduling control engine and an intelligent refrigerating engine aiming at reduction of energy consumption. In addition, we build a green cloud data center platform, realize the scheduling control engine, and verify the feasibility of the framework. It indicates that the framework can realize a cloud platform with low power consumption and a high-energy-efficient data center operation.

INDEX TERMS Green cloud and data center, energy optimizing, deep learning, particle swarm optimization.

I. INTRODUCTION

Since the concept of cloud computing was put forward, it has been developing rapidly in the latest years [1]. Depending on its computing power with high reliability, storage ability with high feasibility and service ability with high efficiency, it has become hot in Internet of Things (IoT) [2], smart applications [3], and et al. The construction of cloud computing centers and cloud platforms is accelerating in the whole world. However, with the increase of cloud computing-oriented data centers, the energy consumption that supports data center becomes larger and larger [4]. Energy consumption is a nonnegligible part of cloud computing operation costs. Besides, high energy consumption also puts tremendous pressure on the environment and energy. Statistically, the power consumption of cloud computing has been more than 1% total power consumption of the global. Energy consumption of data centers includes: continuous power supply to servers, memorizers and other basic equipment, the energy consumption for cooling infrastructure and the unavoidable small energy consumption in operation. Idle servers and high energy consumption of small tasks are ubiquitous in data center. It can be seen that the problem of energy optimization has become a hot topic [5].

Based on the background, the concept of green cloud data center is put forward, the inevitable trend of the development of cloud data centers [6]. Green data center optimizes the IT equipment, refrigeration equipment, illuminating system and electric system in machine room, maximizes energy efficiency and minimizes the influence on environment. Cloud computing has become development direction of the new generation of data center. Green cloud data center has become the cutting edge of energy-saving research. Up to now, some relevant technical researches have been applied to the energy consumption optimization strategy for cloud computing [7]–[9]. Dynamic voltage and frequency scaling (DVFS) adjusts voltage of power supply and clock frequency dynamically based

on the current energy consumption of CPU [10]. Virtualization technology is used in cloud platform to create several virtual machines in host, which is for decreasing the use of hardware resources and improving resource utilization rate [11], [12].

At present, research on energy consumption in data centers is mainly divided into two aspectsčž

- the research on resource allocation and scheduling for raising resource utilization rate [13];
- the optimization of power supply and refrigerating system of data center for lowering the operation and maintenance cost of data center.

The former research is based on traditional scheduling for improvement and optimization. Scheduling control problem is a common research. In many application scenarios, resource scheduling and task allocation issues need to be considered. For example, literature [14] puts forward a dynamic scheduling mechanism based on game theory. In [15], a dynamic programming-based solution has been proposed to schedule the computation offloading from vehicles to clouds. In [16], the iterative energy minimum algorithm is used to make scheduling decisions for mobile agent paths. Literature [17] puts forward a new dynamic resource scheduling mechanism in IoT factories. In [18], the resource scheduling is carried out by using the method of dual-target mixing. Literature [19] uses particle swarm optimization and K-means algorithm to obtain the best service combination. The quantity of researches on optimization of power supply system and heat-removal system is small. Besides, power supply system is relevant to the quantity of equipment operated. The main researches are relevant to the optimization of refrigerating system. The control strategy of the cooling system in the data center is an important direction of energy optimization [20]. The literature [21] addresses the technical and economic issues associated with refrigeration system in data centers through the use of absorption cooling machines. Heuristic optimization algorithm is used. In cloud data center base, electric equipment, refrigerating system and power distribution constitute a mutually related system. Involving complex dynamics principle, traditional method is incapable.

At present, it is a hotspot to solve complex problems by introducing AI technology. The author of literature [22] used deep learning to explore the occurrence regularity of chronic diseases and predicted risks of disease based on convolutional neural network. A new cache mode proposed in [23] proposes an intelligent content distribution mechanism in high-density networks. In literature [24], Lu *et al.* pointed out that the AI-aware technologies will be the main trends in future network applications. In literature [25], deep learning and machine learning are introduced in the Internet of Vehicles(IoV), and the concept of cognitive IoV is proposed. Therefore, to realize cloud computing of low energy consumption and high-efficiency operation of data center, AI technology is applied for improving the energy consumption of cloud data center in this paper. An architecture of AI-enable green cloud is put forward, including scheduling control engine and intelligent refrigeration engine. Concretely, aiming at the reasons of high energy consumption, AI technology is applied to the resources scheduling and refrigerating system of cloud data center. By using reinforcement learning and combinatorial optimization, engine can learn and judge in complex environment by itself. The intelligent situation recognition, demand prediction and scheduling control in the learning methods improve decision-making ability of the system. AI technology is introduced in cloud computing and an architecture of green cloud data center is put forward in this paper. To sum up, the main contributions of the paper are listed below:

- We put forward scheduling control engine and intelligent refrigerating engine.
- AI technology is introduced to scheduling engine and refrigerating engine in this paper, which considers complex and dynamic resource environment and optimizes the energy consumption of data centers.
- We try to build an AI-enable green cloud data center, deploy scheduling control engine to the platform and verify feasibility of scheme in experiment.

Other parts of this paper are arranged below. Section II introduces architecture of AI-Enable Green Cloud, describes the structure of scheduling control engine and intelligent refrigerating engine as well as defines the functions of each module in details. Section III expounds the prediction model and resource scheduling model used by scheduling control engine. Section IV introduces the testbed built for experiment and evaluates the scheduling delay of system and the optimization effect of energy consumption. Finally, Section V makes a summary for this paper.

II. ARCHITECTURE OF AI-ENABLE GREEN CLOUD

Starting from request workflow of users, traditional cloud computing is comprised of application layer, middleware layer, virtual layer and resource layer [26]. Based on the original architecture of cloud computing, we put forward AI-enable green cloud, including scheduling control engine and intelligent refrigerating engine. Fig 1 shows the architecture of AI-enable green cloud. Existing advanced communications and networking technologies such as [27] can be deployed for connecting massive mobile terminals and clouds in the system. The two engines are deployed at the middleware layer and resource layer of cloud computing respectively. The functional modules of scheduling control engine and intelligent refrigerating engine are described in details.

A. SCHEDULING CONTROL ENGINE

Resource scheduling [28] is the key issue of data center. It directly influences the efficacy and cost of cloud system. To cope with the uncertain request of users and burst of networking flow in cloud computing, the current resource scheduling makes decision aiming at high efficiency of completion or maximizing resource utilization [29], [30], which incurs low efficiency of energy

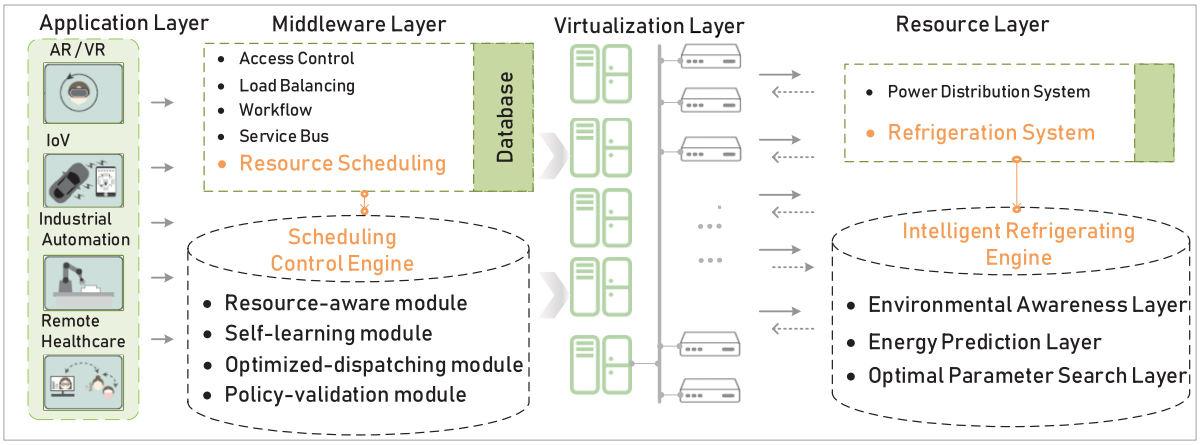


FIGURE 1. The architecture of ai-enable green cloud.

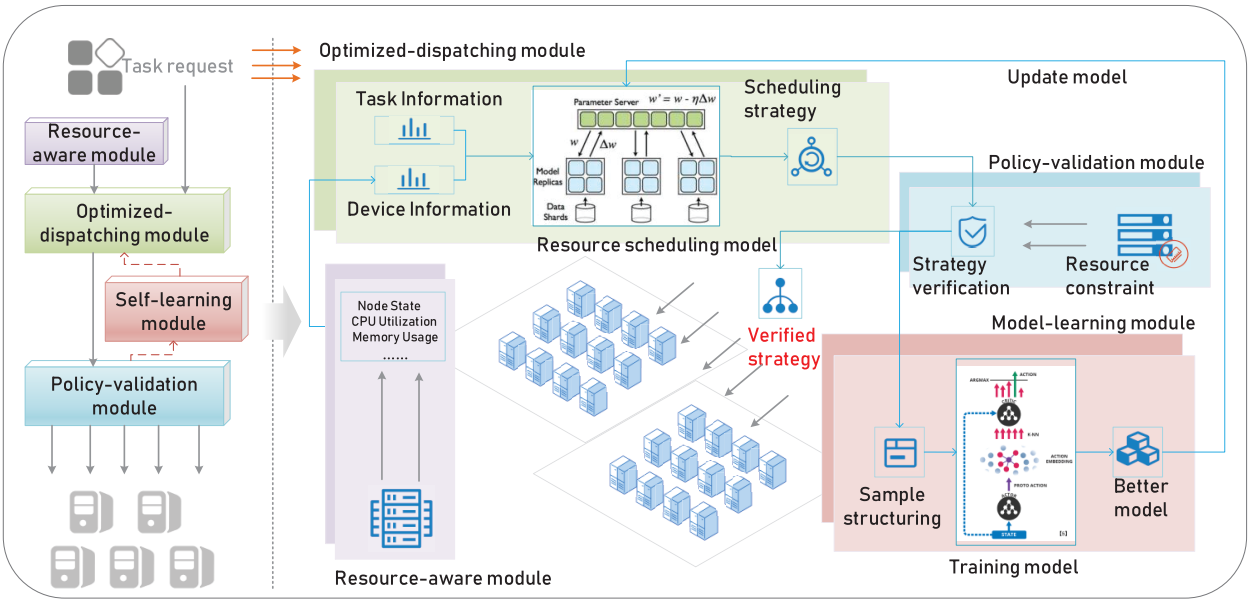


FIGURE 2. The structure of scheduling control engine.

use and low resource utilization rate. Energy is wasted seriously. Physical servers in large quantity for storing and processing data are not used reasonably.

The scheduling control engine in this paper aims at the optimization of both completion time and energy consumption. In multi-target resource scheduling, usersąŕ request can be responded timely, resource utilization rate can be raised as well as energy consumption is decreased. Fig 2 shows the structure of the scheduling control engine. It can be seen from the figure that when a computing task comes in for requesting resources directly from the optimization scheduling module, the optimization scheduling module first acquires information of each device in data center from the resource sensing layer. On the basis of obtaining the above information from the resource-aware layer, the optimization scheduling module uses AI related technology to combine the current task and resource state to realize resource allocation, and gives a task scheduling scheme to make smart decisions [31]. The scheduling policy is then transmitted to the policy verification module for verification. Finally, the scheduling policy that will be verified after verification is executed, to enable the cloud data center to complete rationalized resource scheduling, maximize the use of server resources, and then close idle resources to reduce power consumption. This is the complete working process of the dispatch control engine.

Besides, self-learning module tracks the load of IT equipment in data center, collects relevant information and adds them to experience base. After the experience base constantly trains and optimizes perception model, resource perception model becomes stronger and stronger. Just as that mentioned above, scheduling control engine is comprised of 4 modules: resource perception module, scheduling optimization

module, strategy verification module and self-learning module. According to Fig 2, we will describe each functional module in detail below.

- To realize resource scheduling, the status of each equipment or application in data center shall be known overall firstly. Resource perception module monitors and manages the resource information in the whole data center on the internet. The resources nodes transmit status to the module in real time. The information includes status of each physical resource (open or closed) and the usable resource of each server or virtual node (such as use rate of CPU, current occupation rate of random access memory, the residual capacity of magnetic disk and predicted running time). Deep learning is used to predict environment of the next time slot. The scheduling control engine allocates resources reasonably according to the current and predicted environment [32]–[34].
- The prediction model in resource perception module can predict the load of data center in future short time. On the basis, scheduling optimization module can reserve resources reasonably when allocating resources. It can be seen that it is very important to predict the status of resource equipment. Prediction model directly influences the effect of resource scheduling. Self-learning module and resource perception module interact frequently. The main functions include: i) recording the status of equipment in data center; ii) relearning the prediction model of resource perception module. Because the response to users request is in real time, in-service learning of prediction model is not realistic. Therefore, learning of prediction model is in self-learning module. Resource perception module records the current resource information and predicted load information as data. The actual load information can be got within short time as label. The data group containing data and corresponding label is a sample. Further, with the increase of sample data, experience base can be build. The prediction model is constantly trained using the samples to make the model have more and more experience. In case of sparse request task, the model in scheduling optimization module can be updated. Namely, better model is sent to scheduling module to replace the existing model.
- Traditional scheduling algorithms include First Come First Served (FCFS), priority of short-time operation etc. Shift in turn and priority are the resource planning methods targeting time. Scheduling optimization module allocates resources to data center based on completion time and multi-target scheduling scheme for energy consumption. Besides, the task scheduling of cloud platform is complex and frequent. Heuristic algorithm can offer a feasible solution to living example of combination optimization problem to be solved at reasonable cost [35]. However, owing to the complex and dynamic environment of cloud platform, it is very slow to search the optimal combination. Reinforced learning is used to learn each scheduling. Accumulate experience constantly and use reinforced learning to initialize combination optimization [36]. In this way, the heuristic algorithm can accelerate its search and make a wise decision.
- Strategy verification module is the last step before scheduling strategy is executed. Before scheduling optimization module is trained to be strong enough, as for the whole data center, it is extremely dangerous to directly use it. Therefore, it shall be verified before executing scheduling scheme. The verification module uses basic limitation conditions to verify the scheduling strategy and guarantee the reliability of scheduling engine. For example, the total computing power consumed for the verification module to check existing tasks and preassigned tasks of computing node shall not be more than the current usable computing power of the resource node. Before scheduling strategy controls resources, verify it to guarantee reliable work of data center. If the scheduling optimization model is not strong enough, the traditional scheduling algorithm shall be used firstly until the scheduling optimization model becomes strong enough.

B. INTELLIGENT REFRIGERATING ENGINE

Data center centrally processes, stores, and calculates data through the cluster server. These nodes generate a lot of heat when performing computing tasks. If the data center does not dissipate heat in time, the high-temperature environment of the equipment room will reduce the computing power of the physical equipment. As a high-density equipment cluster, cloud data center produces heat at high level. To guarantee the timely service completion of cloud computing, refrigerating system is crucial for the whole data center. In the whole energy consumption of data center, the energy consumption of refrigerating system accounts for 1/3 of the total energy consumption. Refrigeration is related to the heat dissipation of server and the environment of machine room. Energy consumption of data center cannot be further optimized by only depending on hardware optimization and traditional cooling control strategy. Considering the mutual relevance and complex coupling of the refrigeration control system in cloud data center, the intelligent refrigerating engine in this paper uses deep learning to explore key elements of energy consumption, predict energy consumption and build intelligent cloud refrigeration engine. The detailed structure is shown in Fig 3, and every functional layer is described in details below.

- Before intelligent refrigeration of the intelligent refrigerating engine for data center, it needs to ascertain the surrounding environmental information. Environment perception layer gets environmental information from data center in real time, including temperature, moisture, airflow, the size of machine room etc. At the same time, it gets the status of resources and equipment from the resource perception module in scheduling engine. Then all information is integrated as the input data at

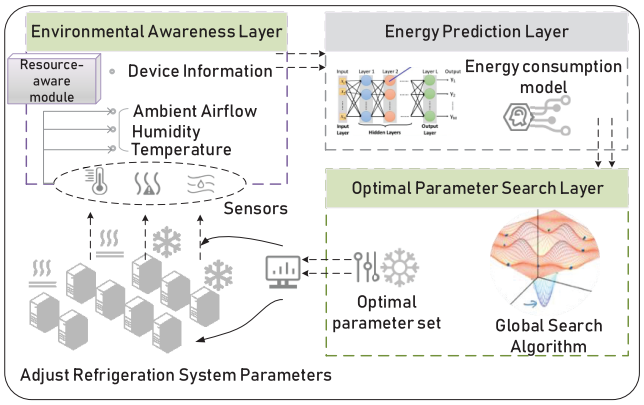


FIGURE 3. The architecture of cloud refrigeration engine.

optimal parameter search layer. The size and structure of machine room is fixed and can be set statically during deployment. As for temperature, moisture and other dynamic information can be got after corresponding sensors are installed in machine room. Environment perception layer gets the dynamic data from sensor and resource scheduling module in real time, integrates them and transmits them to energy consumption prediction layer.

- Energy consumption prediction layer uses the information of environment and resource status to predict the current energy consumption of data center. Traditional algorithm takes long time to observe data and analyze the degree of relevance between data and the energy consumption of data center. Perhaps the acute insight of experts is needed. Deep learning is used in this paper for energy consumption modeling without model learning. Concretely, we firstly use the data of data center to train deep neural network model. Above relevant data is same as the information got from resource perception module. The purpose of model is to get the relation between energy consumption and environment information. Energy consumption of the next time slot can be predicted based on the current environment. For example, Power Usage Effectiveness (PUE) is used as the evaluation indicator of energy consumption efficiency. Energy consumption layer is the model of PUE prediction according to the current environment. Previously, PUE model uses deep learning to learn and analyze large quantity of data, explore the key relations between data and energy consumption, extract the important features of influencing energy consumption and build the mapping relation between environmental data and energy consumption indicator PUE.
- The parameter optimization issue in refrigerating system is a nonlinear issue of global optimization and combined optimization. Intelligent optimization algorithm can solve the issue in short time. Therefore, this paper uses intelligent optimization algorithm to solve the optimal parameter group in refrigerating system. In optimal parameter search, the data at environment perception layer is the input, energy consumption prediction model is the objective function of optimization algorithm, and the natural process is simulated to get the optimal parameter group. Then the optimal parameter group as control command is sent to refrigeration control system. The control system adjusts refrigeration equipment using commands. It shall be noted that refrigeration engine is different from scheduling engine and it is not motivated by computing task. Therefore, the engine uses time polling mechanism for intelligent refrigeration control. When timer is activated, environment perception layer gets the current data and predicted information from data center and resource perception layer. Then energy consumption prediction model is regarded as the objective function at the optimal parameter layer. Optimization search method is used for global optimization and getting group of optimal parameters. At last, the group of optimal parameters is transferred to control system. The control system adjusts the parameters of refrigeration equipment correspondingly.

III. RESOURCE SCHEDULING OF CLOUD DATA CENTER

In this section, we mainly introduce the LSTM-based (i.e. long short-term memory-based) predictive model and RL-based (i.e. reinforcement learning-based) decision model related to the Scheduling control engine.

A. LSTM-BASED PREDICTIVE MODEL

For global resource scheduling of cloud data center, it is necessary to estimate the load of resource node in the next time slot. The prediction of the load of data center can be simplified to be the prediction of every resource node. It is to predict the load of single resource node in the next time slot. Therefore, this paper puts forward the prediction model of cloud data center. It can be seen as a time recurrent neural network (RNN) model. RNN has strong ability of deep semantic expression and exploring the timing sequence information in data. RNN has poor prediction effect as for large change gradient of resource load. To maintain the long-term memory of RNN, we use LSTM to add some structures to RNN and avoid the dependence of prediction model on abnormal data. At the moment t, the prediction model based on LSTM:

$$\begin{aligned} h_t &= \sigma\left(W_h x_t + U_h h_{t-1} + b_h\right) \\ o_t &= \sigma_y(W_y h_t + b_y) \end{aligned} \tag{1}$$

x_t is the input unit at moment t and o_t is the output unit at moment t. The current status h_t of RNN is determined by status of the last moment h_{t-1} and current input x_t. The output unit o_t is determined by the current status h_t. W_h is the weight from input layer to hidden layer. U is the weight of self-circulation at hidden layer. b_h means the deviations, σ_h and σ_y are activation functions. Time series data is the input data of RNN and the output data of network output layer is the prediction of series at the next moment.

B. RL-BASED DECISION MODEL

Discrete Particle Swarm Optimization (DPSO) algorithm is used in this paper to search optimal solution of resource scheduling. It is assumed that there are M resource nodes and N tasks to be allocated. Resource scheduling is to allocate N tasks to M resource nodes. This minimizes the time of completing tasks and the consumption of energy. The distribution matrix and velocity matrix can be expressed as follows:

$$\begin{aligned} X_i &= [(x_{11}, \cdots, x_{1m}), \cdots, (x_{1m}, \cdots, x_{nm})] \\ V_i &= [(v_{11}, \cdots, v_{1m}), \cdots, (v_{1m}, \cdots, v_{nm})] \end{aligned} \quad (2)$$

x_{ij} is the allocation of the i^{th} task to the j^{th} server and $x_{ij} \in \{0, 1\}$.v_{ij} is the probability of the value 0 or 1 of the particle x_{ij} in the position. To make the speed of particle v_{ij} as probability, $Sig(v_{ij})$ function is used to map v_{ij} to the interval [0, 1]. Therefore, the corresponding fitness function and the position & speed update formula are:

$$\begin{aligned} \text{minimize:} \quad \Phi &= T_{cost} + P_{cost} \\ T_{cost} &= \alpha \sum_{i=1}^{m} \sum_{j=1}^{n} time_{ij} x_{ij} \\ P_{cost} &= (1-\alpha) \sum_{i=1}^{m} \sum_{j=1}^{n} power_{ij} x_{ij} \end{aligned} \quad (3)$$

$$v_{ij}^{k+1} = \omega v_{ij}^{k} + c_1(p_{ij}^{k} - x_{ij}^{k}) + c_2(p_{gj}^{k} - x_{ij}^{k}) \quad (4)$$

In the above, Φ is objective function; $time_{ij}$ is the time needed for the i_{th} task to be executed in the j_{th} server; $power_{ij}$ is the energy needed for the i_{th} task to be executed in the j_{th} server; p_{gj} is the optimal position matrix of particle swarm; k is the times of iteration; α is weight factor; c_1 and c_2 are learning factors; w is learning factor. It shall be noted that particle velocity v_{ij}^{k+1} is limited in $[-v_{max}, v_{max}]$. When $v_{ij}^{k+1} < -v_{max}$, $v_{ij}^{k+1} = -v_{max}$; when$v_{ij}^{k+1} > v_{max}$, $v_{ij}^{k+1} = v_{max}$.

DPSO algorithm has great search ability in combination optimization, but the search process from particle initialization to particle optimization takes long time because of dynamic environment. Therefore, reinforcement learning is used in this paper to learn each scheduling and accumulate experience. The particle swarm initialized has been the optimized solution in motion space. This greatly shortens search time. Q-learning is to learn to get certain reward after taking specific action in a certain state. By constantly trying to update reward value, choose the motion with reward maximization finally as the optimal motion output of state. A table is used to record state-action pairs. It is noted using Qtable. Qtable is updated in each execution. Use Bellman Equation to update it. The formula is:

$$Q(s, a) = r + \gamma(max(Q(s', a'))) \quad (5)$$

$$Q(s, a) \leftarrow (1-\alpha)Q(s, a) + \alpha[r + \gamma max(Q(s', a'))] \quad (6)$$

In the formula, s is the current state, α is the action taken in the current state, $\acute{s}$ is the next state incurred by the action as well as $\acute{\alpha}$ is the action taken in new state. r is the reward got by taking the action (noted with Qvalue) and γ is loss factor. The environment of data center is complex and cannot reappear, the experience of all states cannot be got by learning. Therefore, we match the current state with the memorized state using the estimation strategy based on value function. If their similarity reaches certain degree, corresponding action will be taken. Otherwise, the new state will be added to Qtable. This makes Q-Learning have prediction ability and generalization ability. The corresponding fitting function is $Q(s, \alpha; \theta) \approx Q^*(s, \alpha)$. θ is the parameter of model. Deep Q Network (DQN) is used to set up Qtable. End-to-end Q fitting can be realized based on convolutional neural network.

The current environmental state, the LSTM predicted environmental state, and the requested task are used as input data of the convolutional network in the RL-based decision model to deeply mine the intrinsic nonlinear mapping relationship. The output value is the Qvalue that performs each action in this state, and the size is a vector of 1ąÁK, where K represents the number of actions in Q-Learning. The action strategy with large action difference in which Qvalue is arranged in the front is chosen as the initial position of particles. The initialized particles are good solutions with mutual difference. Therefore, different particle corresponds to local optimal solution in different position. Global optimal solution can be found by searching particles. At the same time, search is accelerated.

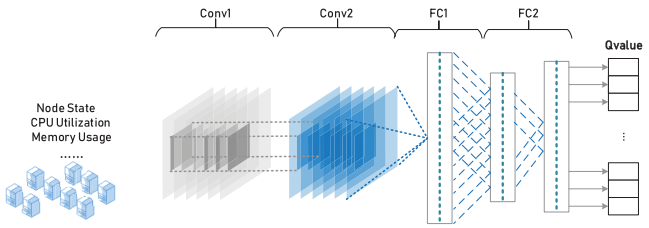


FIGURE 4. The network structure of DQN.

Fig 4 shows the network structure of DQN. Two-layer convolution and fully-connected convolutional neural network are used in this paper.

IV. TESTBED

In this paper, the AI-enable green cloud architecture is proposed and the related algorithms of the scheduling control engine are introduced in detail. In order to verify the RL-based decision model. This paper deploys the proposed scheduling control engine in the CloudSim cloud computing simulation environment. CloudSim is a cloud computing simulation software written in Java language. It provides cloud computing features that support cloud computing resource management and scheduling simulation. It enables researchers to circumvent the inconvenience of actual deployment, and can simulate large-scale cloud clusters and test corresponding algorithms on a single machine. Therefore,

TABLE 1. Cloud computing simulation environment setting.

	Values	
SCHEDULING INTERVAL	300	
VIRTUAL MACHINE	TYPES	50
	MIPS	2500
	SIZE	2.5GB
	BANDWIDTH	100 Mbit/s
HOST MACHINE	TYPES	50
	MIPS	2500
	BANDWIDTH	1 Gbit/s
	STORAGE	1GB

Experiment 1 conducted experiments on the CloudSim simulation environment. It is assumed that the request mission arrives is subject to Poisson distribution. The simulation environment parameters of Experiment 1 are shown in Table 1. There are a total of 50 physical machines and 50 virtual machines. The configuration of each physical machine is the same, and the MIPS (i.e., Million Instructions Per Second) is 2,500.

Besides, the timeliness and stability of system are very important because request is dynamic and changeable in real environment. To verify the reliability of model really, we establish cloud data center platform in Inspur data center to test its stability in real environment. Inspur big data center has 2 management nodes and 7 data nodes in total. 253 TB data can be stored. It can be regarded as a small data center. Therefore, the experiment in Inspur big data center can effectively indicate availability of our model.

Based on the above environment configuration, we first simulate the data center task on the simulation platform with the parameters set. The sample data set is trained to obtain a trained prediction model. To predict the number of tasks at the next point in time. Specifically, the data of the first 5 stages is used to predict the number of tasks at the next point in time, i.e., each sample size is a 6-dimensional vector. Table.2 lists the main parameters of DPSO algorithm used in the two experiments.

TABLE 2. The main parameter settings of the DPSO algorithm.

DPSO Algorithm	Parameter Settings
Particle swarm size P	P=30
Weight factor α	$\alpha = 0.5$
Learning factors c_1,c_2	$c_1 = 2$,$c_2 = 2$
Inertia factor w	$w = 1$
Maximum iterations $Maxiter$	$Maxiter = 1000$

We use cloud computing simulation tool to compare RL-based decision model with the resource algorithm based on Load-Aware. This is to evaluate the energy consumption optimization performance of the resource scheduling based on RL-based decision model. Fig 5 shows the energy

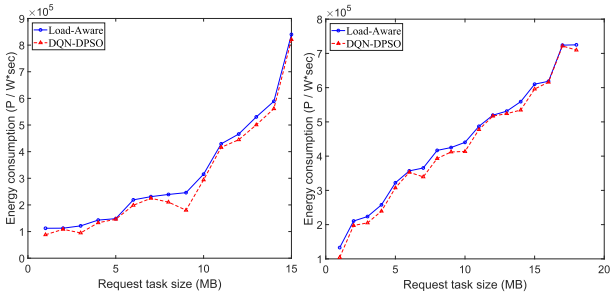


FIGURE 5. Experimental results for the performance evaluation of the energy consumption in cloudsim.

consumption of CloudSim cloud platform in the two resource scheduling algorithms.

Fig 5 shows the result of experiment 1. The x-coordinate is the size of request and the y-coordinate is the average power consumption for executing requested task. For the convenience of observing the energy consumption corresponding to the size of requested task, we arrange the size of tasks in ascending order and operate two tests. In the figure, our model has better effect of energy consumption optimization comparing with the resource algorithm based on load-aware.

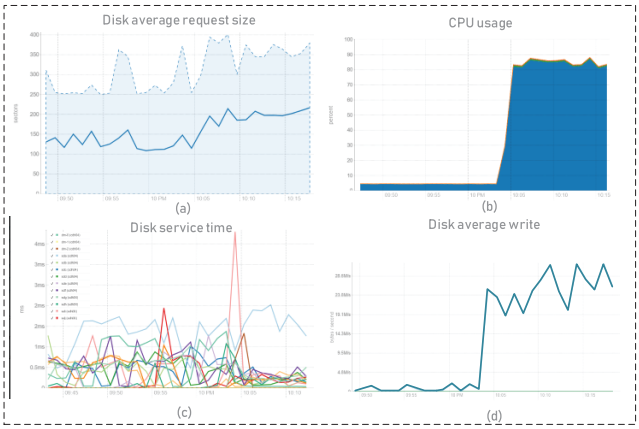


FIGURE 6. Real-time situation of wave data platform deployed with scheduling control engine.

Subsequently, we test the usability of model on the Testbed established in lab, load 5 computing tasks in data center and monitor disk accessing, service time and the CPU use rate. Fig 6 shows the real-time state of CPU and disk in Inspur data center. Throughout the experiment, we gradually increased the amount of computing tasks. As can be seen from Fig 6(a d), the average number of disk requests has steadily increased. There is no downtime under high CPU usage, each disk is load balanced and there are no persistent free disks. In result of experiment, the data center deployed with scheduling control engine can run stably in case of request task.

V. CONCLUSION

AI technology is introduced into cloud data center and an architecture of AI-enable green cloud is put forward in

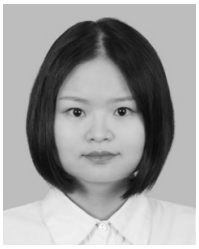
this paper. Besides, this paper introduces scheduling control engine and intelligent refrigeration engine. In one aspect, resources are scheduled reasonably to raise resource utilization rate, decrease the use of physical machine and decrease energy consumption. In another aspect, our work realizes the intelligent adjustment of refrigerating system considering environment and resources, which decrease the operation and maintenance cost of cloud data center. Besides, the models and algorithms relevant to scheduling control engine are introduced in details. Testbed is established and the theoretical algorithm model is verified. In experiment result, our architecture can effectively decrease energy consumption of data center. This paper introduces AI technology in cloud data center to make the cloud data center be a green cloud data center platform with ''low consumption'' of energy, ''green'' energy supply, ''intelligent'' energy use and ''high efficiency'' of energy conversion. In the future, we will enhance the consolidation of mobile networking with AI regarding edge computing, caching and offloading [36]–[38].

REFERENCES

[1] L. Hou *et al.*, ''Internet of Things cloud: Architecture and implementation,'' *IEEE Commun. Mag.*, vol. 54, no. 12, pp. 32–39, Dec. 2016.
[2] M. Chen, J. Yang, X. Zhu, X. Wang, M. Liu, and J. Song, ''Smart home 2.0: Innovative smart home system powered by botanical IoT and emotion detection,'' *Mobile Netw. Appl.*, vol. 22, no. 6, pp. 1159–1169, 2017.
[3] M. Chen, J. Zhou, G. Tao, J. Yang, and L. Hu, ''Wearable affective robot,'' *IEEE Access*, vol. 6, pp. 64766–64776, 2018.
[4] W. Xiang, N. Wang, and Y. Zhou, ''An energy-efficient routing algorithm for software-defined wireless sensor networks,'' *IEEE Sensors J.*, vol. 16, no. 20, pp. 7393–7400, Oct. 2016.
[5] L. Zhou, D. Wu, J. Chen, and Z. Dong, ''Greening the smart cities: Energy-efficient massive content delivery via D2D communications,'' *IEEE Trans. Ind. Informat.*, vol. 14, no. 4, pp. 1626–1634, Apr. 2018.
[6] K. Bilal, S. U. Khan, and A. Y. Zomaya, ''Green data center networks: Challenges and opportunities,'' in *Proc. Int. Conf. Frontiers Inf. Technol.*, Dec. 2013, pp. 229–234.
[7] Z. Zhou, H. Zhang, X. Du, P. Li, and X. Yu, ''Prometheus: Privacy-aware data retrieval on hybrid cloud,'' in *Proc. IEEE INFOCOM*, Turin, Italy, Apr. 2013, pp. 2643–2651.
[8] Z. Guan, J. Li, L. Wu, Y. Zhang, J. Wu, and X. Du, ''Achieving efficient and secure data acquisition for cloud-supported Internet of Things in smart grid,'' *IEEE Internet Things J.*, vol. 4, no. 6, pp. 1934–1944, Dec. 2017.
[9] Y. Xiao, X. Du, J. Zhang, F. Hu, and S. Guizani, ''Internet protocol television (IPTV): The killer application for the next-generation Internet,'' *IEEE Commun. Mag.*, vol. 45, no. 11, pp. 126–134, Nov. 2007.
[10] Q. Deng, D. Meisner, and A. Bhattacharjee, ''CoScale: Coordinating CPU and memory system DVFS in server systems,'' in *Proc. IEEE/ACM Int. Symp. Microarchitecture*, Dec. 2012, pp. 143–154.
[11] A. Di Costanzo, M. D. De Assuncao, and R. Buyya, ''Harnessing cloud technologies for a virtualized distributed computing infrastructure,'' *IEEE Internet Comput.*, vol. 13, no. 5, pp. 24–33, Sep. 2009.
[12] Y. Li and M. Chen, ''Software-defined network function virtualization: A survey,'' *IEEE Access*, vol. 3, pp. 2542–2553, 2015.
[13] Y. Kessaci *et al.*, ''Parallel evolutionary algorithms for energy aware scheduling,'' in *Intelligent Decision Systems in Large-Scale Distributed Environments*. Berlin, Germany: Springer, 2011, pp. 75–100.
[14] G. Wei, A. V. Vasilakos, Y. Zheng, and N. Xiong, ''A game-theoretic method of fair resource allocation for cloud computing services,'' *J. Supercomput.*, vol. 54, no. 2, pp. 252–269, 2010.
[15] J. Zhou, D. Tian, Y. Wang, Z. Sheng, X. Duan, and V. C. M. Leung, ''Reliability-oriented optimization of computation offloading for cooperative vehicle-infrastructure systems,'' *IEEE Signal Process. Lett.*, vol. 26, no. 1, pp. 104–108, Jan. 2019, doi: 10.1109/LSP.2018.2880081.
[16] M. Chen, L. T. Yang, T. Kwon, L. Zhou, and M. Jo, ''Itinerary planning for energy-efficient agent communications in wireless sensor networks,'' *IEEE Trans. Veh. Technol.*, vol. 60, no. 7, pp. 3290–3299, Sep. 2011.
[17] J. Wan, B. Chen, M. Imran, and F. Tao, ''Toward dynamic resources management for IoT-based manufacturing,'' *IEEE Commun. Mag.*, vol. 56, no. 2, pp. 52–59, 2018.
[18] M. Mezmaz *et al.*, ''A parallel bi-objective hybrid metaheuristic for energy-aware scheduling for cloud computing systems,'' *J. Parallel Distrib. Comput.*, vol. 71, no. 11, pp. 1497–1508, 2011.
[19] M. Shamim Hossain, M. Moniruzzaman, G. Muhammad, A. Ghoneim, and A. Alamri, ''Big data-driven service composition using parallel clustered particle swarm optimization in mobile environment,'' *IEEE Trans. Services Comput.*, vol. 9, no. 5, pp. 806–817, Sep. 2016.
[20] H. Cheung, S. Wang, C. Zhuang, and J. Gu, ''A simplified power consumption model of information technology (IT) equipment in data centers for energy system real-time dynamic simulation,'' *Appl. Energy*, vol. 222, pp. 329–342, Jul. 2018.
[21] K. Ebrahimi, G. F. Jones, and A. S. Fleischer, ''Thermo-economic analysis of steady state waste heat recovery in data centers using absorption refrigeration,'' *Appl. Energy*, vol. 139, pp. 384–397, Feb. 2015.
[22] M. Chen, Y. Hao, H. Kai, L. Wang, and L. Wang, ''Disease prediction by machine learning over big data from healthcare communities,'' *IEEE Access*, vol. 5, pp. 8869–8879, 2017.
[23] L. Zhou, D. Wu, Z. Dong, and X. Li, ''When collaboration hugs intelligence: Content delivery over ultra-dense networks,'' *IEEE Commun. Mag.*, vol. 55, no. 12, pp. 91–95, Dec. 2017.
[24] H. Lu, Y. Li, M. Chen, H. Kim, and S. Serikawa, ''Brain intelligence: Go beyond artificial intelligence,'' *Mobile Netw. Appl.*, vol. 23, pp. 368–375, Apr. 2018.
[25] M. Chen, Y. Tian, G. Fortino, J. Zhang, and I. Humar, ''Cognitive Internet of vehicles,'' *Comput. Commun.*, vol. 120, pp. 58–70, May 2018.
[26] K. Hwang and M. Chen, *Big Data Analytics for Cloud/IoT and Cognitive Computing*. London, U.K.: Wiley, 2017.
[27] D. Tian, J. Zhou, Z. Sheng, M. Chen, Q. Ni, and V. C. M. Leung, ''Self-organized relay selection for cooperative transmission in vehicular ad-hoc networks,'' *IEEE Trans. Veh. Technol.*, vol. 66, no. 10, pp. 9534–9549, Oct. 2017.
[28] J. Wan, B. Chen, S. Wang, M. Xia, D. Li, and C. Liu, ''Fog computing for energy-aware load balancing and scheduling in smart factory,'' *IEEE Trans. Ind. Informat.*, vol. 14, no. 10, pp. 4548–4556, Mar. 2018.
[29] Y. Li, F. Zheng, M. Chen, and D. Jin, ''A unified control and optimization framework for dynamical service chaining in software-defined NFV system,'' *IEEE Wireless Commun.*, vol. 22, no. 6, pp. 15–23, Dec. 2015.
[30] M. Chen, Y. Hao, M. Qiu, J. Song, D. Wu, and I. Humar, ''Mobility-aware caching and computation offloading in 5G ultradense cellular networks,'' *Sensors*, vol. 16, no. 7, pp. 974–987, 2016.
[31] M. Chen and Y. Hao, ''Task offloading for mobile edge computing in software defined ultra-dense network,'' *IEEE J. Sel. Areas Commun.*, vol. 36, no. 3, pp. 587–597, Mar. 2018.
[32] X. Jian, Y. Liu, Y. Wei, X. Zeng, and X. Tan, ''Random access delay distribution of multichannel slotted ALOHA with its applications for machine type communications,'' *IEEE Internet Things J.*, vol. 4, no. 1, pp. 21–28, Feb. 2017.
[33] X. Jian, X. Zeng, Y. Jia, L. Zhang, and Y. He, ''Beta/M/1 model for machine type communication,'' *IEEE Commun. Lett.*, vol. 17, no. 3, pp. 584–587, Mar. 2013.
[34] M. Chen, Y. Hao, K. Lin, Z. Yuan, and L. Hu, ''Label-less learning for traffic control in an edge network,'' *IEEE Netw.*, vol. 32, no. 6, pp. 8–14, Nov. 2018.
[35] F. Sardari and M. E. Moghaddam, ''A hybrid occlusion free object tracking method using particle filter and modified galaxy based search meta-heuristic algorithm,'' *Appl. Soft Comput.*, vol. 50, pp. 280–299, Jan. 2017.
[36] X. Wang, Y. Han, C. Wang, Q. Zhao, X. Chen, and M. Chen. (2018). ''In-edge AI: Intelligentizing mobile edge computing, caching and communication by federated learning.'' [Online]. Available: https://arxiv.org/abs/1809.07857
[37] X. Li, X. Wang, P.-J. Wan, Z. Han, and V. C. M. Leung, ''Hierarchical edge caching in device-to-device aided mobile networks: Modeling, optimization, and design,'' *IEEE J. Sel. Areas Commun.*, vol. 36, no. 8, pp. 1768–1785, Jun. 2018.
[38] X. Wang, Y. Zhang, V. C. M. Leung, N. Guizani, and T. Jiang, ''D2D big data: Content deliveries over wireless device-to-device sharing in large-scale mobile networks,'' *IEEE Wireless Commun.*, vol. 25, no. 1, pp. 32–38, Feb. 2018.

JUN YANG received the bachelor's and master's degrees in software engineering from the Huazhong University of Science and Technology (HUST), China, in 2008 and 2011, respectively, and the Ph.D. degree from the School of Computer Science and Technology, HUST, in 2018. He is currently a Post-Doctoral Fellow with the Embedded and Pervasive Computing Lab, School of Computer Science and Technology, HUST. His research interests include cognitive computing, software intelligence, the Internet of Things, cloud computing, and big data analytics.

WENJING XIAO received the bachelor's degree in network engineering from North China Electric Power University, China, in 2018. She is currently a bachelor-straight-to-doctorate student with the Embedded and Pervasive Computing Lab, School of Computer Science and Technology, Huazhong University of Science and Technology. Her research interests include cloud computing, the Internet of Things, and cognitive computing.



CHUN JIANG received the B.Eng. degree in measuring and controlling technology and instrument from the North University of China, Taiyuan, China, in 2015. She is currently pursuing the Ph.D. degree with the South China University of Technology, Guangzhou. Her research interests include edge computing and cyber-physical systems.

M. SHAMIM HOSSAIN (SM'09) received the Ph.D. degree in electrical and computer engineering from the University of Ottawa, Canada. He is currently a Professor with the Department of Software Engineering, College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia. He is also an Adjunct Professor with the School of Electrical Engineering and Computer Science, University of Ottawa. He has authored and co-authored approximately 200 publications in refereed journals and conferences and, books, and book chapters. His research interests include cloud networking, smart environment (smart city, smart health), social media, the IoT, edge computing and multimedia for health care, deep learning approach to multimedia processing, and multimedia big data. He is a Senior Member of the IEEE and ACM. He has served as a member of the organizing and technical committees of several international conferences and workshops. He was a recipient of a number of awards, including the Best Conference Paper Award, the 2016 *ACM Transactions on Multimedia Computing, Communications and Applications* Nicolas D. Georganas Best Paper Award, and the Research in Excellence Award from the College of Computer and Information Sciences, King Saud University (three times in a row). He has served as a co-chair, general chair, workshop chair, publication chair, and TPC for over 12 IEEE and ACM conferences and workshops. He is currently the Co-Chair of the second IEEE ICME workshop on Multimedia Services and Tools for smart-health (MUST-SH 2019). He is on the Editorial Boards the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE NETWORK, the IEEE MULTIMEDIA IEEE WIRELESS COMMUNICATIONS, the IEEE ACCESS, the *Journal of Network and Computer Applications* (Elsevier), *Computers and Electrical Engineering* (Elsevier), *Human-centric Computing and Information Sciences* (Springer), *Games for Health Journal*, and the *International Journal of Multimedia Tools and Applications* (Springer). He also serves as a Lead Guest Editor for the IEEE NETWORK, *Future Generation Computer Systems* (Elsevier), and the IEEE ACCESS. Previously, he served as a Guest Editor of the *IEEE Communications Magazine*, the IEEE TRANSACTIONS ON INFORMATION TECHNOLOGY IN BIOMEDICINE (currently JBHI), the IEEE TRANSACTIONS ON CLOUD COMPUTING, the *International Journal of Multimedia Tools and Applications* (Springer), *Cluster Computing* (Springer), *Future Generation Computer Systems* (Elsevier), *Computers and Electrical Engineering* (Elsevier), *Sensors* (MDPI), and the *International Journal of Distributed Sensor Networks*.

GHULAM MUHAMMAD received the Ph.D. degree in electrical and computer engineering from the Toyohashi University of Technology, Japan, in 2006. He is currently a Professor with the Department of Computer Engineering, College of Computer and Information Sciences, King Saud University (KSU), Riyadh, Saudi Arabia. He has supervised more than 10 Ph.D. and Master Theses. He is involved in many research projects as a principal investigator and a co-principal investigator. He has authored and co-authored more than 200 publications, including the IEEE/ACM/Springer/Elsevier journals, and flagship conference papers. He has a U.S. patent on audio processing. His research interests include image and speech processing, cloud and multimedia for healthcare, biometrics, and security. He was a recipient of the Japan Society for Promotion and Science Fellowship from the Ministry of Education, Culture, Sports, Science and Technology, Japan. He received the Best Faculty Award of Computer Engineering Department, KSU, from 2014 to 2015.

SYED UMAR AMIN received the master's degree in computer engineering from Integral University, India, in 2013. He is currently a Researcher with the Department of Computer Engineering, College of Computer and Information Sciences, King Saud University. His research interests include deep learning, brain–computer interface, and cloud and multimedia for healthcare.