The Power Control Method of Data Center Based on Cloud-Edge Collaboration

Junjie Zhao¹, Yucheng Hou²*, Liu Liu¹, Feng Xie¹, Xiang Mao¹, Pan Luo¹, Jinxin Wang¹, Zhiqiang Zhang¹, Jiang Deng¹

¹Science and Technology on Reactor System Design Technology Laboratory
Nuclear Power Institute of China, Chengdu, 610213, China
²Key Laboratory of the Ministry of Education on Smart Power Grids (Tianjin University), Tianjin, 300072, China
*Corresponding author’s e-mail: houyucheng@tju.edu.cn

Abstract. Data centers have the characteristics of huge energy consumption and strong power fluctuations. To solve the problem of unstable supply and demand such as frequent power fluctuations in the micro-grid of the data center using renewable energies, this paper establishes a controllable load model of server clusters based on the task migration mechanism by analyzing the various tasks and characteristics of large cloud data centers, as well as edge computing nodes. Through the collaborative optimization control of edge and cloud computing tasks, and dynamically adjusting and migrating the server cluster load, a new data center tie-line power control method is proposed. The results of calculation examples show that the method can achieve power smoothing of large-scale cloud data center tie lines through cloud-edge collaborative control.

1. Introduction
With the rapid development of new-generation information technologies such as cloud computing, edge computing, big data, and artificial intelligence, data has exploded. In order to better support high-density, large-bandwidth, and low-latency business scenarios, researchers have proposed to transform from the "cloud computing" model to the “edge computing” model to solve network transmission problems. Currently, edge data centers are in the early stage of development, but have shown the characteristics of rapid expansion of industrial scale, further acceleration of technological development, continuous enhancement of industrial relevance, and have gradually become an engine to promote local economic growth. Cloud-side collaboration between large data centers and edge computing nodes can greatly improve user experiences such as games and videos, and drive the development of new businesses. At present, marginalized and miniaturized edge data centers have gradually emerged and assumed important roles. The market has shown a development situation in which large-scale data centers and miniaturized edge computing environments coexist. However, data centers have huge energy consumption and strong power fluctuations. In 2019, data centers consumed 1% to 3% of the world's total electrical energy [1-2]. High energy consumption and power fluctuations in data centers have become a major problem of the power grid. The data center using renewable energy sources reduces the burden of energy consumption and operating pressure, but due to the randomness and intermittent of renewable energies and fluctuations in data center power consumption, the impact of the power of the tie line between the data center and the grid on the stability of the main grid is increased.
In recent years, in order to meet the trend of high penetration of renewable energy and promote the development of low energy consumption technologies in data centers, low-carbon data center microgrids integrating renewable energy generation have been rapidly developed [3-10]. The microgrid can realize the collaborative optimization and management of renewable energy power tracking, which is an important way to realize low-carbon data center. Much progress has been made in research related to energy optimization and management of large-scale data center microgrids and power tracking of green energy. In [4], a decision model based on two-stage stochastic planning is developed to improve energy efficiency and reduce costs and CO2 emissions, taking into account the synergistic optimization of server configuration and energy storage procurement. In [5], it is described that data centers can participate in proactive demand response activities, such as playing an important role in the power and ancillary services markets. The above-mentioned studies focus on reducing the operating costs of large data centers. However, the intermittent, unpredictable, and uncontrollable nature of renewable energy generation can lead to an imbalance between supply and demand in low-carbon data center microgrids, resulting in frequent power fluctuations in microgrid contact lines. Existing data center microgrid power smoothing methods over-utilize energy storage batteries and fail to fully exploit the flexibility of cloud-edge data centers [8] [9].

To this end, this paper proposes a new method for power smoothing control of micro network contact lines in large data centers. By analyzing the various tasks and characteristics of large-scale cloud data center and miniaturized edge computing, a server cluster controllable load model based on task migration mechanism is established. Through the collaborative optimized control of edge computing and cloud computing, the load of the server cluster is dynamically adjusted to effectively smooth out the power fluctuation of the contact line in the data center while ensuring the quality of service and reliability of the power supply.

2. Load control model of data center

From the perspective of power generation and utilization, the data center and its connected primary and renewable energy systems can be modeled as an autonomous smart micro-grid system, and from the perspective of computing and communication, multiple geographically distributed edge data center stations create a resource load balancing system over the communication network. This chapter considers data center computing energy consumption, uncontrolled energy consumption such as cooling and lighting, and distributed power generation units to model data center energy consumption. The studied data center architecture consists of the main network and distributed energy supply system, the data center energy management system, and the data center power consumption load. The power balance equation for this microgrid, when power losses are not considered, is as follows.

$$P_{TL,i} + (P_{W,i} + P_{V,i}) = P_{cs,i} + P_{UL,i}$$  \(1\)

where \(P_{TL,i}\) is the power injected into the data center campus by the main network through the contact lines in the \(i\)-th control period; \(P_{W,i}\) and \(P_{V,i}\) are the real-time output of wind and photovoltaic power generation in the \(i\)-th control period; \(P_{cs,i}\) is the server cluster load in the data center campus in the \(i\)-th control period; and \(P_{UL,i}\) is the other loads such as cooling and lighting in the data center campus.
For a single server, the load can be represented by the following model.

\[ P_{\text{server}} = P_{\text{idle}} + \alpha_{\text{CPU}}(P_{\text{max}} - P_{\text{idle}}) \]  

(2)

where \( P_{\text{server}} \) is the power consumption of the server, \( P_{\text{max}} \) and \( P_{\text{idle}} \) are the power consumption of the server at full load and no load respectively, and \( \alpha_{\text{CPU}} \) is the CPU utilization rate of the server. PUE (Power Usage Efficiency) [10] is generally used to measure uncontrollable loads such as air conditioning and cooling, communication and storage, etc., which is defined as the ratio of the total energy consumption of the data center and the energy consumption of the server cluster, and the power consumption of the data center in the \( i \)-th control period can be obtained as follows.

\[ P_{\text{DC},i} = \text{PUE} \cdot M_0 \cdot [P_{\text{idle}} + \alpha_{\text{CPU},i} \cdot (P_{\text{max}} - P_{\text{idle}})] \]  

(3)

where \( M_i \) is the number of servers turned on during the \( i \)-th control period in the data center, \( \alpha_{\text{CPU},i} \) is the CPU utilization of the servers during the \( i \)-th control period, and PUE is the power usage efficiency. \( \alpha_{\text{CPU},i} \) is the average CPU occupancy of the server cluster in the \( i \)-th control period. The average CPU occupancy of the server cluster is the sum of the CPU occupancy of the currently running tasks, as shown in the following equation.

\[ \alpha_{\text{CPU},i} = \frac{\sum_{k=1}^{K} \alpha_{\text{CPU},k}}{M_0} \]  

(4)

In this paper, each computing task is designed to include the following parameters: \([\alpha_{\text{CPU},k}, H_k, T_{\text{ar},k}, t_{p,k,T}, T_{d,k}]\), where \( \alpha_{\text{CPU},k} \) represents the CPU occupancy of a single server when task \( k \) is processed; \( H_k \) represents the data size of task \( k \); \( T_{\text{ar},k} \) represents the time for task \( k \) to reach the data center; \( t_{p,k,T} \) represents the time left for task \( k \) to be processed at time \( T \); and \( T_{d,k} \) represents the latest time for task \( k \) to be completed. According to the Service-Level Agreement, computing tasks in the data center need to be completed within the time specified by the user, and thus computing requests can be divided into delay-sensitive and delay-tolerant requests. For delay-sensitive tasks, \( T_{\text{ar},k} + t_{p,k} = T_{d,k} \) (where \( t_{p,k} \) represents the total processing time of the task), i.e., the task needs to be processed immediately upon arrival at the data center without interruption to ensure that the computation is completed before the specified deadline; for delay-tolerant tasks, at any moment in time, \( T_{\text{ar},k} + t_{p,k} = T_{d,k} \) (where \( t_{p,k} \) represents the total processing time of the task), \( T \), whose maximum delay time is \( t_{\text{delay,k,T}} = T_{\text{deadline,k,T}} - T - t_{p,k,T} \). The task can only be optimized if \( t_{\text{delay,k,T}} > 0 \), and must be activated within the time period \([T, T + t_{\text{delay,k,T}}]\) to avoid violating the service-level agreement (Service-level agreement). Level Agreement, SLA). Delay-tolerant tasks can not only be processed in the current data center, but can also be packaged for cross-data center migrations.
$t_{\text{travel},k}$, and can be migrated only if $t_{\text{delay},k} \geq 0$.

### 3. Power smoothing method for tie line

In this paper, we consider the smoothing of contact line power in a large cloud data center, and set the contact line power control step to $\Delta t$. In each $\Delta t$ period, we first obtain the initial contact line power of the data center, and according to the contact line target power $P_{TL}^\text{set}$, we obtain the contact line optimization target power $P_{TL0,j}$ for a large data center.

$$
\min \sum \left| P_{TL,i} - P_{TL,i}^\text{set} \right|^2 
$$

(5)

Then the target power of the server cluster in the data center is

$$
P_{\text{co-target},i,j} = \frac{P_{TL0,j} + (P_{W,j,i} + P_{V,j,i})}{\text{PUE}}
$$

(6)

This target power is in steps of $\Delta t$, and the server cluster power control step is $\Delta t$. The actual target power of the data center server cluster in each time period $\Delta t$ is

$$
P_{\text{co-target},i,j} = P_{\text{co-target},i,j}, \quad T_i \in [T_j, T_j + \Delta T)
$$

(7)

where $T_i$ and $T_j$ are the initial moments of the $i$-th control period $t$ and the $j$-th control period $\Delta T$, respectively.

The constraints for data center server power optimization are as follows.

$$
\sum_{k \in S} \frac{H_k}{\Delta t} \leq B,
$$

$$
\alpha_{\text{CPU,min}} \leq \alpha_{\text{CPU},i,j} \leq \alpha_{\text{CPU,max}},
$$

$$
t_{\text{delay},k} \geq 0,
$$

(8)

where $P_{\text{DC-max}}$ is the maximum power consumption of data center [9], which is defined as follows.

$$
P_{\text{DC-max}}^i = P_{\text{max}}^i \cdot M_0^i \cdot \text{PUE}
$$

(9)

$S$ represents the set of tasks migrated by the data center in the $i$-th control period, $B$ is the maximum bandwidth between data centers. This constraint means that when migrating tasks across data centers, the amount of data transferred at any time shall not exceed the maximum bandwidth. The second constraint is the upper and lower bounds of the average CPU utilization of the server cluster in the data center, respectively, to prevent excessive waste and overuse of computing resources in the data center. The third constraint means that the delayable time of any task $k$ at any time is not negative to ensure that the task can be processed within the specified time.

The mathematical model of the optimization problem in this paper is a multivariate nonlinear planning problem, and the variables in the server cluster optimization problem are the number of pending tasks in different control periods and the number of tasks migrated between cloud data centers and edge data centers (all are integer values), which can be solved by a genetic algorithm.

### 4. Simulation and analysis

The performance evaluation of the contact line control method proposed in this paper is conducted using a large cloud data center microgrid campus and four edge data centers in a city as an example. Table 1 lists the specific system parameters for each data center.

| Table 1. Parameter setting |
|---------------------------|
| Parameter | Cloud data Center 1 | Edge data Center 1 | Edge data Center 2 | Edge data Center 3 | Edge data Center 4 |
|-----------|---------------------|---------------------|---------------------|---------------------|---------------------|
| $M_0$     | 40000               | 2000                | 2000                | 1500                | 3000                |
| $P_{\text{idle}}$ (kW) | 0.1                 | 0.1                 | 0.1                 | 0.1                 | 0.1                 |
| $P_{\text{max}}$ (kW) | 0.5                 | 0.4                 | 0.4                 | 0.4                 | 0.4                 |
| PUE       | 1.3                 | 1.47                | 1.68                | 1.53                | 1.7                 |
The simulation results are shown in Fig.3. The orange curve is the initial tie line power curve, the blue curve is the tie line stabilized power after the individual cloud data center needs response, and the green curve is the edge data center and the cloud data center after coordinated control. The smoothing result of the line power shows that under the premise of ensuring the data center and user delay tolerance, through the implementation of the cloud-edge collaborative optimal control strategy, the data center microgrid tie line power fluctuation has been effectively stabilized, and the curve standard. The difference is changed from the initial 0.534MW to 0.402MW, which can effectively reduce the power fluctuation by 24.7%.

5. Conclusion
The tunable and complementary nature of data center computing tasks provides sufficient control margin for achieving contact line power balance. This paper establishes a server cluster load model based on the task delay mechanism and task migration across data centers, leverages the real-time complementarity between cloud data centers and edge data centers in terms of new energy output and load fluctuation, and utilizes the flexible time domain migration of computing tasks and the space migration mapping of edge-cloud data centers to the server to achieve effective response to the power fluctuation of the contact line in large cloud data centers.

Acknowledgments
This paper was supported by the National Key Research and Development Program of China (2018YFB0905000); Headquarters of state grid corporation of science and technology project (SGTJDK00DWJS1800232).
References

[1] M. Anastasopoulos, A. Tzanakaki, and D. Simeonidou, “Stochastic energy efficient cloud service provisioning deploying renewable energy sources,” IEEE Journal on Selected Areas in Communications, vol. 34, no. 12, pp. 3927–3940, 2016.

[2] X. Deng, D. Wu, J. Shen, and J. He, “Eco-aware online power management and load scheduling for green cloud datacenters,” IEEE Systems Journal, vol. 10, no. 1, pp. 78–87, 2016.

[3] Y. Xia, M. Zhou, X. Luo, S. Pang, and Q. Zhu, “A stochastic approach to analysis of energy-aware dvs-enabled cloud datacenters,” IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 45, no. 1, pp. 73–83, 2015.

[4] S. Kwon, L. Ntaimo and N. Gautam, "Demand Response in Data Centers: Integration of Server Provisioning and Power Procurement," in IEEE Transactions on Smart Grid, vol. 10, no. 5, 2019, pp. 4928-4938.

[5] M Jawad, M B Qureshi, U Khan, et al. “A robust Optimization Technique for Energy Cost Minimization of Cloud Data Centers,” IEEE Transactions on Cloud Computing, 2018.

[6] C. Li, Y. Hu, J. Gu, J. Yuan, and T. Li, “Oasis: Scaling out datacenter sustainably and economically,” IEEE Transactions on Parallel and Distributed Systems, vol. 28, no. 7, pp. 1960–1973, 2017.

[7] D. W. Wang and X. M. Liu, “A Resource Scheduling Strategy for Cloud Computing Platform of Power System Simulation Based on Dynamic Migration of Virtual Machine,” Automation of Electric Power Systems(in Chinese), vol. 12, pp. 97-105, Jun. 2015.

[8] N Kumar, G S Aujla, S Garg, et al. "Renewable Energy-Based Multi-Indexed Job Classification and Container Management Scheme for Sustainability of Cloud Data Centers", IEEE Transactions on Industrial Informatics, vol. 15, no. 5, pp. 2947 – 2957, 2019.

[9] Nguyen H. Tran et al., "How Geo-Distributed Data Centers Do Demand Response: A Game-Theoretic Approach", IEEE Transactions on Smart Grid, vol. 7.2, pp. 937-947, 2016.

[10] Z. Hu, B. Ld and I. Luo, "Tdm- and Cost-Efficient Task Scheduling across Geo-Distributed Data Centers," in IEEE Transactions on Parallel and Distributed Systems, vol. 29, no. 3, pp. 705-718, 1 March 2018.