O-007

Proposal of a workload migration plan method capable of responding to prediction errors

Satoshi Kaneko

1. Introduction

Data centers (DCs) that serve as hubs for knowledge creation through AI utilization consume significant amounts of electricity, making them one of the challenges outlined in the CN2050 initiative. To decarbonize DCs, efforts are underway to improve the energy efficiency of IT equipment and cooling systems, as well as to adopt renewable energy (RE) that does not emit CO_2. Recently, advanced initiatives have begun to align RE power generation and consumption on an hourly basis[1][2].

However, DCs require a constant large amount of electricity, making the geographical and temporal constraints of renewable energy a bottleneck. In response, a method of allocating workloads (WL) across multiple DCs according to the availability of renewable energy in different regions and at different times is gaining attention.

This enables WL to be executed using electricity with lower CO_2 emissions. However, in practice, there are errors in predicting renewable energy supply and DC power consumption, making flexible power demand adjustment a challenge.

In this paper, we propose a WL transition planning technology capable of addressing power supply and demand prediction errors and conduct a comparative evaluation with conventional methods.

2. Impact of prediction error

A technology has been proposed to adjust data center power demand in accordance with renewable energy supply [4], and it is described as a technology for managing and optimizing energy consumption within data centers. This paper assumes a response time of one day in advance for power prediction errors. However, actual power supply and demand conditions are more dynamic, and rapid response to short-term fluctuations is often necessary. In particular, renewable energy supply is highly dependent on weather conditions, so there is a need to develop systems capable of shorter-term prediction and response.

When power demand adjustments are made based on the previous day's supply and demand forecast, the power demand adjustment plan based on the WL transition planned one day in advance does not take into account prediction errors. Therefore, the entire error cannot be adjusted if prediction errors occur.

Figure 1 illustrates the impact of errors. The vertical axis represents renewable energy supply, and the horizontal axis represents time. The black solid line represents the predicted renewable energy supply, with the peaks illustrating the times of day when solar power generation is highest. The black dotted lines at regular intervals represent the timing for creating transition plans. The figure detects an error in the renewable energy forecast at time Te, and the renewable energy supply forecast value at time Te + d is received, indicated by the red solid line. In this case, the timing for revising the next demand adjustment plan is Tp, so the error during the period from time Te + d to time Tp (the shaded area in the figure) is difficult to address using conventional technology. In other words, a timely response to power supply and demand errors is critical.

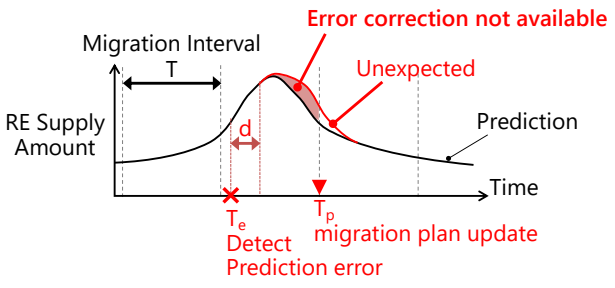

Figure 1 Issue of Prediction error

Furthermore, even if errors are detected, it is not guaranteed that all WLs can be adjusted before the errors occur. In particular, if WLs store data or rely on data access, post-processing to maintain data access is required, and the transition time for these processes must be considered when updating the transition plan on demand.

3. Proposal Method

3.1 Application migration time

The migration time for WL (Workload) depends on the time required for application construction. Here, WL is defined as software applications that include communication with external applications and data access. Application construction consists of three elements: (1) program loading, (2) communication settings with external applications, and (3) data storage access configuration. (1) Programs are loaded from the code management server, and if they have been replicated to the migration destination in advance, the loading time can be ignored. ② Communication reconfiguration can also be considered zero time if pre-designed. ③ Data access reconfiguration has three methods: Remote Access, On-demand Copy, and Replication, each with differences in performance, migration time, and cost. These characteristics must be considered when planning the migration.

3.2 WL migration plan method capable of addressing prediction errors

Figure 1 shows an overview of the proposed method. In the proposed method, the time lag between the detection of prediction errors in renewable energy supply or power consumption and the actual manifestation of the errors is defined as d. If the transition of the workload (WL) is completed within this time, it is considered that the impact of the prediction errors can be reduced.

The proposed method targets WLs that can be migrated within this d range and sets the migration methods “on-demand copy,” “remote access,” and “replication” in order of priority. This

Table 1 Characteristics of data access patterns

	Remote Access	On-demand copy	Replication
data access performance	Poor※	Good	Good
data cost	Good	Good	Poor
migration time	Good	Poor※	Good

※: Depend on network performance

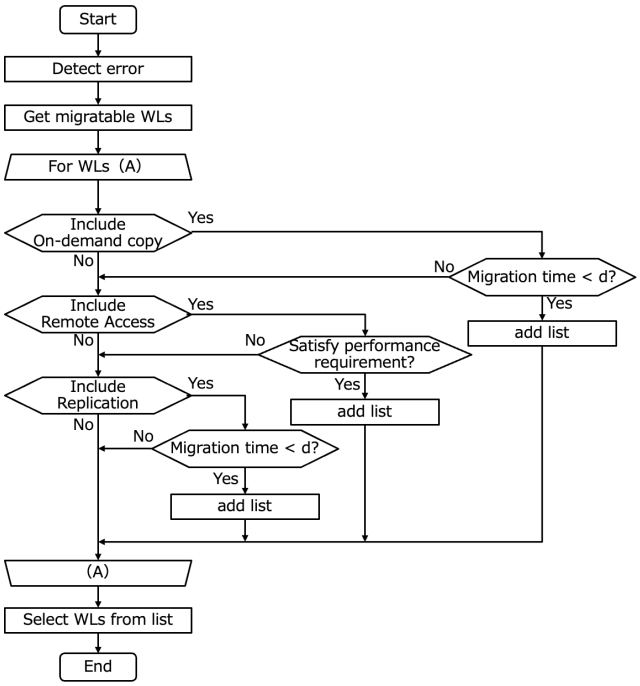

Figure 2 Process overview

enables a migration plan that satisfies performance requirements while minimizing data retention costs.

First, after detecting prediction errors, the migration types allowed for each currently running WL are confirmed based on the execution definition file, and a group of WLs that can be migrated is extracted.

Next, the following steps are performed for each target WL.

If on-demand copy is permitted, the time required to complete the migration is estimated based on the current network performance and data volume. If this is less than d, it is considered a migration candidate. Note that if the data has been duplicated in advance, the copy process is not necessary, and the decision is made based on the switchover time alone.

If on-demand is not suitable or does not meet the conditions, it is confirmed whether remote access is permitted, and performance estimation is performed based on the post-migration configuration. If the value meets the specified performance requirements (e.g., user-defined response time or throughput), it is considered a candidate for migration via remote access.

If remote access is also unsuitable, and if duplication is permitted, calculate the time required for the switchover. If this is less than d, add it to the list of migration candidates.

After completing this process for all WLs, select the optimal WL from the candidate list and generate the final (error correction) migration plan to execute the migration.

In selecting the load to be migrated, the proposed method selects the WL to be migrated as the load to be migrated based on, for example, the estimated power consumption of the WL and the amount of deviation of the error information, so that the total estimated power consumption is equivalent to the amount of deviation. Here, the total estimated power consumption is equivalent to the amount of deviation when the total estimated power consumption is equal to or approximately equal to the amount of deviation.

4. Evaluation

Evaluate the supply-demand adjustment error by comparing it with the conventional method. The supply-demand adjustment error is defined as the difference between the adjustment value resulting from the WL transition and the target power consumption for supply-demand adjustment. A supply-demand adjustment error of 0% indicates that the power consumption matches the target value due to the WL transition.

As a prerequisite for evaluation, we assume a data center that operates by matching the amount of renewable energy generated with the power demand of the data center on an hourly basis. Supply and demand adjustment is performed between two DCs, and when an excess or shortage of renewable energy occurs at one DC and the other DC experiences a shortage, WL transition is used to adjust supply and demand from the excess DC to the shortage DC once every hour. Assume that two-thirds of the WLs in operation during each time period are WLs that retain data. Additionally, assume that errors occur once every two hours, 30 minutes prior to the error, with an error margin of ±20%. The evaluation period is set to any 24-hour period.

First, the conventional technology cannot handle the aforementioned errors. Therefore, it cannot handle the ±20% errors that occur in 12 out of 24 supply-demand adjustment opportunities, resulting in a supply-demand adjustment error of 1 * 12/24 * 20/100 = 10%.

Next, regarding the proposed method, demand-based supply adjustment is possible by switching to the WL in response to errors. However, depending on the data volume, there is a possibility that the switch may not be completed in time for errors detected 30 minutes prior. Assuming a network performance of 1 Gbps between data centers and a data size range of 10–500 GB, the median value of 225 GB would result in a data transfer time of 30 minutes. Therefore, it can be assumed that half of the WLs containing the data can respond to the error. Thus, the demand-supply adjustment error of the proposed method is 1 × 12/24 × 1/4 × 20/100 = 2.5%. This means that the demand-supply adjustment error is reduced to one-quarter of the conventional method. As the proportion of renewable energy increases and prediction errors become more frequent, the effectiveness of the proposed method is expected to grow.

Reference:

[1] Google, “24x7 Carbon Free Energy,” 2021. [Online]. Available: https://www.gstatic.com/gumdrop/sustainability/247-carbon-free-energy.pdf. [Accessed: June 2025].

[2] Digital Realty, 04 February 2025. [Online]. Available: https://www.digitalrealty.com/about/newsroom/press-releases/123307/digital-realty-joins-innovative-24-7-hourly-energy-matching-programs-in-sweden-and-france. [Accessed: June 2025].

[3] J. S. S. N. S. I. A. B. R. C. S. S. a. S. K. Anup Agarwal, “Redesigning Data Centers for Renewable Energy,” (HotNets’21) 2021.